Università di Padova – Dipartimento di Matematica “Tullio Levi-Civita”

Scuole di Dottorato in Matematica Pura e Computazionale

Seminario Dottorato 2022/23

Preface

This document offers an overview of the activity of Seminario Dottorato 2022/23.

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, 2023

Corrado Marastoni, Tiziano Vargiolu

Abstracts (from Seminario Dottorato's webpage)

Wednesday 5 October 2022

Interest rate market: how economic changes affect mathematical modeling

Giacomo LANARO (Padova, Dip. Mat.)

In the last fifteen years many changes in the economic and financial world have caused a great modification in the way in which every financial market is studied by the analysists. One market that has been most affected by these changes is the interest-rate market. We will present how this market worked before the 2007 crisis and what has no longer hold after that date. Moreover, we will show how the mathematical modeling has been adapted to correctly represent the new tasks that have arisen in those years. Finally, we will present how the problem of parameters recalibration for an interest-rate model can be faced in a geometric framework through the computation of the Lie algebra generated by a suitable set of vector fields in an infinite-dimensional setting.

Wednesday 16 November 2022

Shear flows and viscoelastic fluids

Muhanna Ali H. ALRASHDI (Padova, Dip. Mat.)

In this seminar I will give a brief overview on viscoelastic fluids and rheology. Rheology is the science that deals with the way materials deform when forces are applied to them. To learn anything about the rheological properties of a material, we must either measure the deformation resulting from a given force or measure the force required to produce a given deformation. So, we study rheological properties through some experiments such ad steady shear, Small amplitude oscillatory shear, stress growth upon inception of steady shear flow and stress relaxation after a sudden shearing displacement. Moreover, we present a new model that includes logarithmic strains. Finally, comparisons between the new model and a classical one (the upper-convected Maxwell model) will be given to illustrate similarities and differences.

Wednesday 30 November 2022

Moving Least Square approximation using variably scaled discontinuous weight function

Mohammad KARIMNEJAD ESFAHANI (Padova, Dip. Mat.)

Functions with discontinuities appear in many application such as image reconstruction, interface problems, and etc. Accurate approximation and interpolation of these functions are therefore of great importance. After giving an introduction on the required notions, we present an approximation method of discontinuous function f, which incorporates the discontinuities into the approximant s. In a nutshell, the idea is to control the influence of the data on the approximant, not only with regards to their distance, but also with regards to the discontinuities of the underlying function. The numerical experiments show an improvement on the accuracy of the approximation compared with the conventional schemes.

Wednesday 14 December 2022

Binomial coefficients in modular arithmetic - A mathematical solution to musical questions

Riccardo GILBLAS (Padova, Dip. Mat.)

In this seminar we are linking modular binomial coefficients to periodic sequences with modular integer values. In the context of serialism, the romanian composer Anatol Vieru used such sequences to compose several musical pieces. We will introduce the main properties of periodic sequences and we will explain the link between them and the binomial coefficients in $\mathbb{Z}/m\mathbb{Z}$. This allows to use tools such as Kummer's Theorem to answer some questions arisen from Vieru's observations.

Wednesday 18 January 2023

Complex Networks: a highly interdisciplinary field. Theory and Applications

Sara VENTURINI (Padova, Dip. Mat.)

Networks and graph models have become a nearly ubiquitous abstraction and an extremely useful tool to represent a variety of real systems in different fields. They can help us to better understand and analyze different types of interactions and dynamics. Recent researches have shown that real world interactions, in many cases, cannot be fully described by standard graphs. Therefore, there is the need to study more complex structures such as multilayer networks, which enable to take into account different types of information, as well as simplicial complexes and hypergraphs, which consider group interactions. We will give a brief introduction to modern mathematical and computational tools for complex networks, their applications, and their extension to multilayer and hypergraphs.

Wednesday 8 February 2023

A brief introduction to Bloch-Kato conjecture

Shilun WANG (Padova, Dip. Mat.)

The Bloch-Kato conjecture (BKC) is an important and difficult problem in number theory and arithmetic geometry, which came up with Bloch and Kato in 90s, and was then refined by Fontaine and Perrin-Riou later. It is a natural generalization of the Birch and Swinnerton-Dyer conjecture (BSD), which is one of Millennium Problems. However, we only know BKC is true in few cases. The seminar will give a brief introduction to the motivation and historical development of BKC by some explicit examples and explain its relation with the BSD conjecture. In the second part, we will talk about some progresses in BKC.

Wednesday 15 February 2023

A duality based DMK approach to the L^1-norm and Total Variation regularization in optimization problems

Nicola SEGALA (Padova, Dip. Mat.)

Optimization Problems and Inverse Problems are nowadays very popular with several applications in engineering and data analysis. Typically, inverse problems are ill-posed and admit infinite solutions. Analogously, non convex objective functions may admit multiple local extremal points drastically making their identification difficult or even impossible. A common strategy to overcome this problem is to add regularization terms to the objective function. The aim of the regularization term is to gain convexity to improve the identifiability of local minimizers and improving the well-conditioning of the problem. A very effective and challenging choice of regularizers are based on L1-norms (compressed sensing, LASSO) or Total variation of the optimization design parameters. The use of such regularization strategies is hampered by the difficulty in finding efficient and robust numerical solution algorithms.

In this seminar we will introduce the problem of regularization in a very broad sense and describe a Legendre duality based approach to the L^1-norm and TV regularization. We will discuss numerical solution approaches based on the Dynamic Monge Kantorovich (DMK) scheme, originally developed for the numerical solution of the L^1 Optimal Transport problem. We will discuss a few equivalent approaches based on extending the results presented by Bouchitte' and Buttazzo for the L^1 Optimal Transport problem, for which DMK-like gradient flows can be envisaged leading to convergent numerical schemes. We will show a few interesting applications to classical examples, such as 1-Harmonic functions, 1-D signal TV-Denoising, and compressed sensing of graph Laplacian partial eigenproblem..

Wednesday 1 March 2023

Micro-Macro limit: from the Follow-the-Leader model to the Lighthill-Whitham-Richards model

Mohamed BENTAIBI (Padova, Dip. Mat.)

The Lighthill-Whitham-Richards (LWR) model is a hyperbolic conservation law where the solution is a macroscopic density that typically represents the average spatial concentration of vehicles. The Follow-the-Leader model (FtL) instead can be thought as a dynamical system of N cars in which each car travels with a velocity that depends on its relative distance with respect to the car immediately in front. With the FtL model we build a microscopic density which approximates the macroscopic one. After briefly introducing both the LWR and the FtL models, we prove that the microscopic density converges to the macroscopic one in a suitable topology. We also present new results regarding the microscopic stability of the FtL model and its applications to the analysis of the micro-macro limit problem.

Wednesday 15 March 2023

On the volume of (half-)tubular neighborhoods of surfaces in sub-Riemannian geometry

Tania BOSSIO (Padova, Dip. Mat.)

In 1840 Steiner proved that the volume of the tubular neighborhood of a convex set in $\mathbb{R}^n$ is a polynomial of degree n in the size of the tube. The coefficients of such a polynomial carry information about the curvature of the set. In this talk we present Steiner-like formulas in the framework of sub-Riemannian geometry. In particular, we introduce the three-dimensional sub-Riemannian contact manifolds, which the first Heisenberg group is a special case of. Then, we show the asymptotic expansion of the volume of the half-tubular neighborhood of a surface and provide a geometric interpretation of the coefficients in terms of sub-Riemannian curvature objects.

Wednesday 29 March 2023

Ergodic Mean-Field Games with Riesz-type aggregation

Chiara BERNARDINI (Padova, Dip. Mat.)

In this seminar we introduce second-order ergodic Mean-Field Games systems defined in the whole space $\mathbb{R}^n$ with coercive potential and aggregating nonlocal coupling, defined in terms of a Riesz interaction kernel. From a PDE viewpoint, equilibria of the differential game solve a system of PDEs where an Hamilton-Jacobi-Bellman equation is combined with a Kolmogorov-Fokker-Planck equation for the mass distribution. Due to the interplay between the strength of the attractive term and the behavior of the diffusive part, we obtain three different regimes for existence and nonexistence of classical solutions to the MFG system. After briefly introducing the model, we present the main ideas underlying the proof of our results. Finally, we study the behavior of solutions in the vanishing viscosity limit, namely when the diffusion becomes negligible.

Wednesday 19 April 2023

Reciprocity laws

Eduardo ROCHA WALCHEK (Padova, Dip. Mat.)

In this seminar, we give an introduction to reciprocity laws in number theory, since its origins in solving quadratic equations over finite fields to how it evolved – like ever more complex variations on the original theme – to the almost unrecognizable, yet still somehow related, modern explicit reciprocity laws. Our focus will be on the succession of the many results tied by this same name, introducing the relevant concepts and ideas along the way.

Wednesday 3 May 2023

Optimal control with a stochastic switching time: introduction and solution approaches

Maddalena MUTTONI (Padova, Dip. Mat.)

When planning an optimal policy, a farsighted decision-maker should account for the possibile occurrence of disruptive events over the course of the time horizon. For example, when planning the optimal emission abatement policy, account for a possible climate catastrophe; when planning industrial production, account for an unpredictable disruption that may affect the producer?s profit. In the optimal control framework, a stochastic switching time is a random instant, modeled as a positive random variable, which marks a regime shift – i.e., an abrupt and irreversible change in the system – which splits the planning horizon into two stages. The shift may affect the payoff and/or the state trajectory in several ways, all of which are included in the analysis of the most general scenario. In search for the optimal policy under this kind of uncertainty, two methods are featured in the literature: the "backward" approach and the "heterogeneous" one. The two approaches will be described, compared, and then applied to a marketing toy model.

Wednesday 17 May 2023

Symmetries, groups and graphs: from the origins to today's research

Daniele NEMMI (Padova, Dip. Mat.)

Symmetries are everywhere: we can find them in nature, art, music, poetry... we can find them in equations, geometrical objects, mechanical systems, molecules and more generally, all over mathematics and science. But what are they? Why are they so important? Every mathematician, even without noticing it, has used symmetries to solve problems which otherwise would have been more difficult or even impossible to solve. The talk will be a journey into group theory: the branch of mathematics which studies the concept of symmetries and how they relate to one another. We will focus in particular on finite groups and how they are built by fundamental blocks: the finite simple groups, whose classification is considered one of the most remarkable achievements in the mathematics of the last century. We will talk about how some of today's research problems, such as generation problems, can be encoded in the language of graphs which help us to better understand the structure of finite groups.

Wednesday 31 May 2023

Cutting pattern optimization in sawmill

Enrico VICARIO (Padova, Dip. Mat. with MICROTEC Srl, Venezia, Italy)

Optimizing the cutting of wood from log to board is a step within the lumber production chain that has great potential for optimization. This processing in industry is carried out in lines with machines that are increasingly automated and have varying degrees of flexibility in cutting execution. In this work, a specific problem of generating optimal cutting patterns was modeled in order to address it as a Mixed Integer Linear Problem (MILP). The mathematica formulation has been tested on real instances and the result was compared, in terms of obtained value and computation time, with an ad-hoc greedy heuristics.

Wednesday 14 June 2023

Hopf algebras and Kaplansky's sixth conjecture

Elisabetta MASUT (Padova, Dip. Mat.)

Hopf algebras were discovered in 1941 by Heinz Hopf in algebraic topology, but a general theory of these algebras began in the late 1960s. In 1975 Kaplansky listed 10 conjectures on Hopf algebras, which have been the focus of a great deal of research. Some of these conjectures are still open. The one we are interested in is the sixth one, which is about representation theory for Hopf algebras.
The aim of this talk is to present the research work around this conjecture. In order to reach this goal, we will firstly recall what a Hopf algebra is. We will focus on some examples, especially on group algebras. Moreover, we will explain Drinfeld twist, that is a way to deform Hopf algebras, by means of an invertible element.

Interest rate market: how economic changes affect mathematical modeling

Giacomo Lanaro [(*)]

Abstract. In the last fifteen years many changes in the economic and financial world have caused a great modification in the way in which every financial market is studied by the analysists. One market that has been most affected by these changes is the interest-rate market. We will present how this market worked before the 2008 crisis and what has no longer hold after that date. Moreover, we will show how the mathematical modeling has been adapted to correctly represent the new tasks that have arisen in those years. Finally, we will present how the problem of parameters recalibration for an interest-rate model can be faced in a geometric framework through the computation of the Lie algebra generated by a suitable set of vector fields in an infinite-dimensional setting.

1 Introduction

The interest-rate market, also called fixed-income market, is the set of all the traded financial products that guarantee a fixed stream of payments at pre-specified dates in the future. As an example, we introduce the simplest fixed-income product, that will play a crucial role in the construction of the model that we are going to study. It is called is Zero-coupon bond:

Definition 1.1 A Zero-coupon bond (ZCB) is a contract that guarantees to the holder a single payment of one euro at the maturity time T. It is also called *T-bonds.* Its price at time $t < T$ is denoted with $B_t(T)$.

It is the most important interest-rate product, because through the analysis of the price of this contract we answer the following question:

Which is the value today of a payment of 1 euro at time T?

Indeed, the price at time t of a ZCB with maturity T represents the market value at time t of 1 euro received at time $T > t$. Facing the problem of change in time of money value

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: `glanaro@math.unipd.it` . Seminar held on 5 October 2022.

is quite difficult, because, like the price of every fixed-income contract, $B_t(T)$ depends on a stochastic factor called *interest-rate*.

We are going to analyse the most common techniques used to model the dynamics of the interest rates that determine the ZCB, such as the *Heath-Jarrow-Morton* class of models. Moreover, we will see how these techniques must have been modified in order to face the new tasks brought by the financial crisis. In this more recent framework, we will study the problem of parameter recalibration of a Heath-Jarrow-Morton model, through the notion of consistency between a given parameterized manifold (used to describe the term structure at the first day of the analysis) and the Heath-Jarrow-Morton model taken into account.

2 Pre-crisis modeling

2.1 Zero coupon bonds and LIBOR

We are going to exploit the properties of the price of a Zero-coupon bond to introduce the LIBOR interest rate. Indeed, before the crisis of 2008, the ZCBs prices were linked to the most liquid interest rate in the market, the LIBOR. As a first step, we observe that a ZCB price is supposed to be a two variable function satisfying the following condition:

- $B_T(T) = 1, \quad T \geq 0$;
- By empirical observation of the prices, it holds:
  (a) fixed t, the function $T \to B_t(T)$ is a very smooth function, often differentiable. Its Image $\Gamma_t := \{B_t(T) : \ T \in [t, T^*]\}$ is called *term structure at time t*.
  (b) fixed T the function $t \to B_t(T)$ is a stochastic process, whose trajectories are very irregular.

In particular, $B_T(T) = 1$ is necessary to guarantee that the market is well defined, i.e. a class of financial strategies, called *arbitrages* is not allowed. Here, a financial strategy is a linear combination of instruments traded in the market. It is an arbitrage if the probability the its value at the end of the observation is greater or equal to zero is one and moreover its value is possibly strictly greater than zero. Clearly, the absence of this strategies guarantees that the market is fair.

Since the price of a T-bonds represents the evolution in time of the money value, it is convenient to introduce an interest rate as a combination of the T-bonds prices for different maturity. We construct a contract at time t, whose final value (at time T) is given by an investment of 1 euro at time $S > t$ for the time interval $[S, T]$, where $T > S$. I.e., we want to compute the interest-rate at t for an investment during the time interval $[S, T]$ in ZCBs.

First, let us assume that the T-bonds are quoted in the market for each maturity $T > t$ (clearly, this is not a realistic condition, because the bonds actually traded in the market are associated with a finite set of maturities). Then, we consider the following strategy:

- At time t:

(a) We sell one S-bond, to obtain $B_t(S)$ euros. This implies that we must pay 1 euro to the counter party at S;

(b) With the $B_t(S)$ euros obtained by selling the S-bond, we buy $\frac{B_t(S)}{B_t(T)}$ units of T-bonds.

Then, the cash flow is zero.

- At time S: as we said, we must pay 1 euro;
- At time T: the $\frac{B_t(S)}{B_t(T)}$ units of T-bonds mature and we obtain $\frac{B_t(S)}{B_t(T)}$ euros.

In conclusion, at time t, we know that investing one euro in ZCBs at time S, the value of the investment at T will be $\frac{B_t(S)}{B_t(T)}$ euros. Therefore, in analogy with the definition of continuously compounded interest rate, we introduce the *continuously compounded forward rate at time t for the time interval* $[S,T]$ that is the solution $R(t;S,T)$ to the equation $1 \cdot e^{(T-S)R(t;S,T)} = \frac{B_t(S)}{B_t(T)}$. It is fully determined by the price of S-bonds and T-bonds as follows:

$$R(t;S,T) := -\frac{1}{T-S}(\log B_t(T) - \log B_t(S)). \tag{1}$$

The goal of financial analysts is to model interest rates like $R(t;S,T)$ in order to study the market behaviour. Actually, usually the market analysts aim at modelling the following version of the interest rate, called *instantaneous forward rate* at time t:

$$f_t(T) := -\frac{d}{dT}\log B_t(T). \tag{2}$$

It represents the continuously compounded forward interest, contracted at time t over the infinitesimal interval $[T, T+dT]$.

We recall also the simple version of $R(t;S,T)$, which is called *simple forward rate* contracted at time t for the time interval $[S,T]$. It is the solution $F(t;S,T)$ to the equation $(T-S)F(t;S,T)+1 = \frac{B_t(S)}{B_t(T)}$. It can be rewritten as:

$$F(t;S,T) := -\frac{1}{T-S}\frac{B_t(T)-B_t(S)}{B_t(T)}. \tag{3}$$

Both $R(t;S,T)$ and $F(t;S,T)$ are considered risk-less because at time t their value is fully determined by something measurable at time t: $B_t(S)$ and $B_t(T)$.

2.2 Forward-rate modelling - Heath-Jarrow-Morton approach

Until 2008 financial crisis it was possible to obtain every interest-rate instrument with a linear combination of T-bonds for different maturities T. This implied that, modelling the instantaneous forward rate we could determine every fixed-income instrument. Heath-Jarrow-Morton (HJM) is a very common class of models devoted to this purpose. It is defined imposing that the process $(f_t(T))_{t\in[0,T]}$, introduced in equation (2), satisfies the following stochastic differential equation:

$$\begin{cases} df_t(T) = \alpha(t,T)dt + \sigma(t,T)dW_t, \\ f_0(T) = f^(T). \end{cases} \tag{4}$$

In system (4) the two function α and σ associated respectively with the drift and the volatility of the model are specified. We then have an infinite-dimensional system of stochastic differential equations, once for each T. Moreover, the initial term structure $\Gamma_0 := \{f^*(T) :\ T > t\}$ is constructed by the quotations of instruments actually traded in the market, following a procedure called *Bootstrapping.* In particular, Γ_0 is the graph of a function, that is supposed to be smooth.

Remark 2.1 Often, it is not convenient to model the instantaneous forward rates with HJM models using the parameterization described by the dynamics in system (4). Indeed, the domain of the solution $f(T)$ is $[0,T]$ which is different for each maturity T. Therefore, we should parameterize the HJM dynamics in terms of the *time to maturity* $x := T - t$, instead of considering the maturity T. Through this change of variable, the continuous forward rate is represented by the the couple of parameters (t,x): $f_t(T) \equiv f_t(t{+}x) =: r_t(x)$. In particular, for each $x \geq 0$, the domain of the curve $(r_t(x))_{t\in[0,\infty)}$ is the same for each x. $r_t(x)$ is called *Musiela parameterization* of the instantaneous forward rate $f_t(T)$.

As described in Remark 2.1 adopting the Musiela parameterization we have that the instantaneous forward-rate satisfies the relation $r_t(x) = -\frac{\partial B_t(t+x)}{\partial x}$ and the dynamics of $(r_t(x))_{t\in[0,\infty)}$ is given by:

$$\begin{cases} dr_t(x) = \left(\frac{\partial}{\partial x} r_t(x) + \alpha(t,t+x)\right)dt + \sigma(t,t+x)dW_t, \\ r_t(0) = r^M, \end{cases} \tag{5}$$

where $r_0(x)$ is obtained via bootstrapping technique from the market data at time $t = 0$.

Finally, in order to guarantee the absence of arbitrage strategies, the drift component $\alpha(t,t+x)$ must satisfy the so called *Heath-Jarrow-Morton condition*:

$$\alpha(t,t+x) \equiv \sigma(t,t+x)\int_0^x \sigma^*(t,t+u)du, \tag{6}$$

where A^* denotes the transpose of A.

3 Post-crisis modeling

In this section we are going to see how the financial crisis affected the interest-rate market and how the Heath-Jarrow-Morton class of models must have been adapted to consider the new features that must be taken into account.

Starting with the US credit crisis in 2007 that is linked with the sovereign debt crisis in the Eurozone market during 2011, the whole financial market was affected by the worst crisis of the last decades. Two of the most delicate tasks to handle in those years were:

- The credit risk, that is the risk that a counterparty in a financial contract will not fulfill its obligation;
- The liquidity risk, that is the risk of excessing costs of funding a position in a financial market, due to the lack of liquidity.

The behaviour of the most liquid interest rates in the fixed-income market (such as the LIBOR and the EURIBOR) were particularly affected by this kind of risks. In order to understand why, let us give a brief description on how these rates are built.

3.1 The LIBOR

The LIBOR had been the benchmark rate for the interest-rate market until the end of 2021 in the London stock exchange. LIBOR stands for *London InterBank Offered Rate* and it is an interbank interest rates. This means that it is chosen by a panel of private banks acting in the London market as the interest rate for interbank operations. The Libor rate is a polled rate. It is indeed determined by an average of the rates proposed by every bank of the benchmark panel in answer the following question: *At what rate could you borrow funds by asking and then accepting inter-bank offers in a reasonable market?.*

As a consequence, the rates determined by every bank of the panel must not be too high (because otherwise when a bank needs a capital, it has to pay an higher interest rate) but it must not be too low too (because otherwise it's not convenient to lend money to the other banks of the panel).

For the Eurozone market there exists a *Euribor rate*, which is computed by a panel of European banks in a similar way. Clearly, only the biggest and most stable banks of the respective market can be part to LIBOR or EURIBOR panel. In particular, before 2008 they were considered *too big to default.* As a consequence, lending money to a bank in the Libor rate was considered risk-less. Therefore, the LIBOR at time T for the time interval $[T, T+\delta]$ (denoted with $L(T; T, T+\delta)$) was supposed to be equal to the simple forward rate $F(T; T, T+\delta)$ described in equation (3). Indeed, we recall that $F(T; T, T+\delta)$ is supposed to be risk-less by construction. As a consequence, all the derivatives that depend on the LIBOR rate could be determined by a portfolios of Zero Coupon Bonds until 2008.

After the crisis, due to the collapse of several huge banks in the US such as Lehman Brothers, lending money inside the LIBOR panel was no longer considered risk-less. In particular, the banks started to add an additional components to the interest-rate in order to protect themselves from the risk that the counter party (which was another bank of the panel) failed before the complete payment of the loan. As a consequence, after the crisis spreads have emerged between LIBOR rates and the risk less forward rate:

$$L(T; T, T+\delta) > F(T; T, T+\delta) \tag{7}$$

Moreover, from market data, analysts observed that these spreads were bigger when the length of the time interval δ (called tenor) before the maturity of the contract was bigger. In Figure 1, it's possible to notice this phenomenon. In particular, before 2007 the spreads between EURIBOR associated with different tenors were considered negligible. After that time, spreads emerged between the EURIBOR rate associated with the smallest tenor (one day), which is called *EONIA* and the other rates.

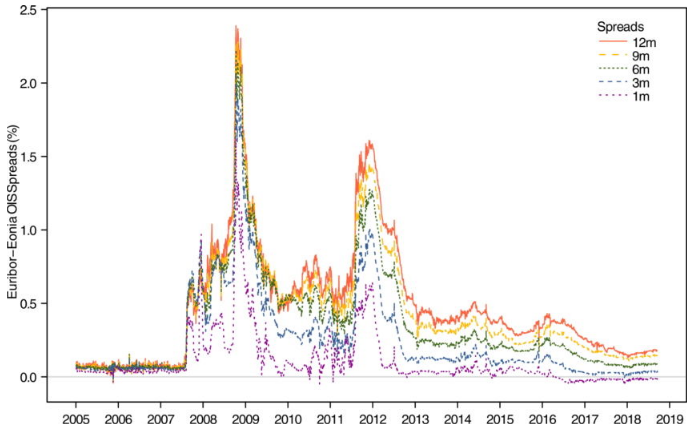


Figure 1: In the figure there are the spreads between the Euribor observed for different lengths of the time interval and the Eonia rates, which is the overnight rate (Euribor OverNight Index Average). Eonia is the rate computed for the smallest possible time interval, one day.

In conclusion, since the relation between the LIBOR rate and the simple forward rate obtained through ZCB prices does not hold anymore, it has been no longer possible to model the entire fixed-income market with the instantaneous forward rates.

3.2 A solution: The multi-curve approach

We saw that the LIBOR depends on the particular tenor it is associated with, then we can't obtain the value of an interest rate derivative depending on the LIBOR using a portfolio of ZCBs (because relation (7) holds). Therefore, a possible way to have a complete description of the market is to model separately the instantaneous forward-rates associated with every LIBOR computed for a benchmark of tenors. This method is called *multi-curve approach* and it is based on the following procedure:

- Let us consider a finite set of positive numbers $\{\delta_0, \dots, \delta_m\}$ from one day to one year. The goal is to model separately the forward-rates associated with tenor equal to δ_i.
- If δ_0 = one day, the counter party risk can be considered negligible, because the probability that a bank in LIBOR or EURIBOR panel will not fulfill an obligation with maturity one day is still supposed to be zero. Then, we can use the EONIA rate as a benchmark on which the spreads related the other Euribor rates are defined. Moreover, for the EONIA rate, the pre-crisis relation still holds:

  $$L^{\delta_0}(T, T, T + \delta_0) \equiv F(T; T, T + \delta_0), \tag{8}$$

  due to the negligibility of the credit risk.

- Finally, we introduce multiplicative spreads $S_t^j(T)$ between the EONIA and the generic EURIBOR associated with tenor δ_j. We adopted the structure described in [2] for the spreads:

  $$S_t^j(T) := \frac{1+\delta_i L^{\delta_j}(t;T,T+\delta_j)}{1+\delta_j L^{\delta_0}(t;T,T+\delta_j)}, \quad j=1,\dots,m, \tag{9}$$

  where $L^{\delta_0}(t;T,T+\delta_j) = F(t;T,T+\delta_j)$ but $L^{\delta_j}(t;T,T+\delta_j)$ can't be modelled using ZCB:

  $$L^{\delta_j}(t;T,T+\delta_j) \neq -\frac{1}{\delta_j}\frac{B_t^0(T+\delta_j)-B_t^0(T)}{B_t^0(T+\delta_j)}.$$

The goal is then to provide a model for the EONIA and the spreads between the EONIA and the generic EURIBOR. Actually, even if there are not bonds associated with the generic LIBOR rate $L^{\delta_j}(t;T,T+\delta_j)$ traded in the market, we introduce fictitious bonds $((B_t^j(T))_{t\in[0,T]})$ associated with every tenor δ_j. These fictitious bonds are defined as follows:

$$S^j(t,T) := S^j(t,t)\frac{B_t^j(T)}{B_t^0(T)}. \tag{10}$$

We interpret $B_t^j(T)$ as bonds because they are positive by construction and the bond relation at the maturity, $B_T^j(T)=1$, is satisfied by definition. The introduction of these new processes is crucial in order to adapt the Heath-Jarrow-Morton models class to the multi-curve approach. Indeed, by $B_t^j(T)$, we can define *instantaneous forward rates* associated with tenor δ_j as in equation (5):

$$r_t^j(x) := -\frac{\partial}{\partial x}\log B_t^j(x+t). \tag{11}$$

In conclusion, the interest-rate market is described through the multi-curve approach as follows:

- The forward-rate for each tenor δ_j is modeled separately by a HJM model:

  $$dr_t^j(x) = \mu_t^j(x)dt + \sigma_t^j(x)dW_t. \tag{12}$$

- The logarithm of the spot spread process between the EONIA and the EURIBOR associated with tenor δ_j, $Y_t^j := \log S^j(t,t)$ is assumed to be described by an Itô process:

  $$dY_t^j = \gamma_t^j dt + \beta_t^j dW_t. \tag{13}$$

Then, differently from the pre-crisis setting, we now have a system of $2m+1$ stochastic differential equations: the first $m+1$ are infinite-dimensional (because they depend on the time to maturity parameter x). The last m are finite-dimensional (associated with each

Y^j). Moreover, a Heath-Jarrow-Morton condition on the drift still holds in this framework. Indeed, the model we are going to study is:

$$\begin{cases} dr_t^0(x) = \big[\mathbf{F}r_t^0(x) + \sigma_t^0(x)\mathbf{H}\sigma_t^0(x)\big]dt + \sigma_t^0(x)dW_t; \\ dr_t^j(x) = \big[\mathbf{F}r_t^j(x) + \sigma_t^j(x)\mathbf{H}\sigma_t^j(x) - \beta_t^j\sigma_t^{j*}(x)\big]dt + \sigma_t^j(x)dW_t; \\ dY_t^j = \big(\mathbf{B}r_t^0(x) - \mathbf{B}r_t^j(x) - \frac{1}{2}||\beta_t^j||^2\big)dt + \beta_t^j dW_t, \end{cases} \tag{14}$$

where $j = 1, \dots, m$, $(\mathbf{H}f)(x) := \int_0^x f^*(z)dz$, f^* is the transpose of f and $\mathbf{B}f(x) = f(0)$. Denoting the solution of the model as $\hat{r}_t = \begin{pmatrix} r_t^0(x) & \cdots & r_t^m(x) & Y_t^1 & \cdots & Y_t^m \end{pmatrix}$, the dynamics in system 14 is compactly denoted with:

$$d\hat{r}_t = \hat{\mu}_t dt + \hat{\sigma}_t dW_t, \tag{15}$$

As we can see, the drift term $\hat{\mu}$ defined by:

$$\hat{\mu}_t := \begin{pmatrix} \mathbf{F}r_t^0(x) + \sigma_t^0(x)\mathbf{H}\sigma_t^0(x) \\ \mathbf{F}r_t^1(x) + \sigma_t^1(x)\mathbf{H}\sigma_t^1(x) - \beta_t^1\sigma_t^{1*}(x) \\ \vdots \\ \mathbf{F}r_t^m(x) + \sigma_t^m(x)\mathbf{H}\sigma_t^m(x) - \beta_t^m\sigma_t^{m*}(x) \\ \mathbf{B}r_t^0(x) - \mathbf{B}r_t^1(x) - \frac{1}{2}||\beta_t^1||^2 \\ \vdots \\ \mathbf{B}r_t^0(x) - \mathbf{B}r_t^m(x) - \frac{1}{2}||\beta_t^m||^2 \end{pmatrix}$$

is fully-determined by the volatility term$\hat{\sigma} := \begin{pmatrix} \sigma^0 & \cdots & \sigma^m & \beta^1 & \cdots & \beta^m \end{pmatrix}^*$ by non-arbitrage constraints (this is the generalization of condition (6) to post-crisis framework).

4 Consistency condition

4.1 The recalibration problem

One of the main tasks that a financial analyst has to deal with when working with interest rate models $\mathcal{M}$ is the parameter recalibration problem. Indeed, a model $\mathcal{M}$ like the one described in system 14 is dependent on several parameters (that determine the volatility terms σ^j and β^h for each j and h). The choice of these parameters is made using the most recent available market data. The usual procedure is the following one:

(a) at $t = 0$,

(i) market data are used to fit the term structure $\Gamma_0 = \{r^M(x) : x \geq 0\}$ to observed market prices for each curve (it represents the initial value of the model we are considering: $\hat{r}_0$). In this framework Γ_0 is $2m+1$-dimensional vector, whose first $m+1$ components are made by a curve in $\mathbb{R}$. In particular, each curve in the first components of Γ_0 is obtained fitting values present in the market through parameterized families, such as the Nelson Siegel family:

$$G^{NS}(\bar{z}, x) = z_1 + (z_2 + z_3 x)e^{-z_4 x}. \tag{16}$$

  (ii) calibrate the model $\mathcal{M}$ to Γ_0 to get the prices of interest-rate derivatives.

(b) at the following day ($t = 1$), the procedure of step 1. has to be reproduced to recalibrate the parameters of the model.

The question one may ask now is: **Why is the recalibration procedure necessary?**

- The first main motivation is that a model $\mathcal{M}$ is an approximation of the reality. Therefore, after few days, the realizations of $\mathcal{M}$ will be no longer coherent with the market. Then, recalibration is useful to add new information given by the most recent market data.
- For interest-rate models $\mathcal{M}$ there are other details to handle carefully. Indeed, even with the most sophisticated model, the image $\mathcal{G} := \{G(\bar{z}, x) : \ x \geq 0\}$ of the parameterized family used to produce the initial term structure can be not well-fitted with the model $\mathcal{M}$. Indeed, $\mathcal{M}$ could not belongs to the family $\mathcal{G}$ after the first day.

To catch the intuition behind this second point, we introduce the following definition:

Definition 4.1 We say the model $\mathcal{M}$, determined by a forward-rate dynamics $(\hat{r}_t(x))_{t\in[0,T]}$, is consistent with a family G of parameterized curves determining a surface $\mathcal{G}$, if the rates $r_t(x)$ produced by $\mathcal{M}$ belong to $\mathcal{G}$ for each $t \in [0, t^*]$ for a positive time t^*.

We aim at conditions that guarantee a model to be consistent with a parameterized family G used to calibrate the initial term structure Γ_0.

4.2 The Geometric Approach

At the end of the previous subsection, we introduced the consistency problem between a given model $\mathcal{M}$ determined by the volatility terms $\sigma^0, \ldots, \sigma^m, \beta^1, \ldots, \beta^m$ and a given parameterized family G. To be more rigorous, we adopt the geometric approach, described for the pre-crisis setting in [1]. We interpret the realization of a forward rate model $(r_t(x))_{t\in[0,t*]}$ as a (single) curve living on a suitable functional space:

$$\mathcal{H} := \{r \in \mathcal{C}^\infty(\mathbb{R}_+; \mathbb{R}) : \ r \text{ is infinite-times differentiable and } ||r||_\gamma < \infty \ \}, \tag{17}$$

where the norm $|| \cdot ||_\gamma$ is defined as:

$$||r||_\gamma^2 = \sum_{n=0}^{\infty} 2^{-n} \int_0^\infty \Big(\frac{\partial^n}{\partial x^n} r(x)\Big)^2 e^{-\gamma} dx, \quad \gamma > 0. \tag{18}$$

In equation (18), we can fix any $\gamma > 0$ in order to have a norm. Then, the dynamics of each forward-rate component of system 14 can be interpreted as a SDE on $\mathcal{H}$:

$$dr_t^j = (\mathbf{F} r_t^j + \sigma_t^j \mathbf{H} \sigma_t^j) dt + \sigma_t^j dW_t, \tag{19}$$

where $\mathbf{F} f(x) := \frac{\partial}{\partial x} f(x)$ and $\mathbf{H} f(x) := \int_0^x f^*(u) du$, for each $f \in \mathcal{H}$.

We interpreted the realization of $(\hat{r}_t(x))_{t\in[0,t^*]}$ as a curve living on the Hilbert space $\hat{\mathcal{H}} := \mathcal{H}^{m+1} \times \mathbb{R}^m$. The first $m + 1$ components of $\hat{r}_t$ are associated with the forward-rate curves $r_t^j(x)$, each of them living in $\mathcal{H}$. The last m, associated with the logarithm of the spreads Y^j, live in $\mathbb{R}$.

4.3 The result

As a first step to understand the kind of result we are looking for, it is convenient to give an example in the simplest case of a sigle curve model whose forward-rate component is finite dimensional:

Example 4.2 We consider a model $\mathcal{M}$ determined by the realization $(r_t)_{t\in[0,t^*]}$, that is a $\mathbb{R}^d$-process given the equation:

$$dr_t = \mu(r_t)dt + \sigma(r_t)dW_t. \tag{20}$$

In this case the goal is to determine conditions such that the model $\mathcal{M}$ is consistent with:

$$\mathcal{G} := Im[G] = \{G(z) : \; z \in \mathcal{Z}\}, \quad \mathcal{Z} \subset \mathbb{R}^n,$$

where $G : \mathbb{R}^n \to \mathbb{R}^d$. In order to exploit the geometric approach we rewrite equation (20) as follows:

$$\frac{dr_t}{dt} = \mu(r_t) + \sigma(r_t)u_t,$$

with $u_t \in \mathbb{R}$ is heuristically interpreted as white noise and it is defined only formally. Intuitively, we have that there exists a positive time interval $[0, t^*]$ such that $(r_t)_{t\in[0,t^*]}$ stays in $\mathcal{G}$ for all $u_t \in \mathbb{R}$ if and only if:

$$\frac{dr_t}{dt} \text{ is tangent to } \mathcal{G}, \quad \forall u_t \in \mathbb{R}, \; t \in [0, t^*]. \tag{21}$$

Then, if r belongs to Im$[G]$ we can represent it as $r \equiv G(z)$, with $z \in \mathbb{R}^n$. Moreover, let us recall that the tangent space of $\mathcal{G}$ in the point $r \equiv G(z)$ is given by:

$$T_r\mathcal{G} = \mathrm{Im}[G'_z(z)] = \Big\langle \frac{\partial G(z)}{\partial z^i}, \quad i = 1, \ldots, n \Big\rangle,$$

where G'_z is the Fréchét derivative (Jacobian) of G.

Thus, in order to have consistency, we must require that the differential of r, heuristically represented by the right member of equation (21), belongs to $T_r\mathcal{G}$. Therefore, a natural guess for the consistency condition should be:

$$\mu(G(z)) + \sigma(G(z))u \in \mathrm{Im}[G'_z(z)], \quad z \in \mathbb{R}^n, \quad \text{for every } u \in \mathbb{R}.$$

In this setting we must take into account the fact that the chain rule doesn't hold, so we can't differentiate in time as we did in equation (21). Indeed, we observe that if F is differentiable, by Itô formula we have that:

$$dF(t, r_t) = \frac{\partial}{\partial t}F(t, r_t)dt + \frac{\partial}{\partial r}F(t, r_t)dr_t + \frac{1}{2}\frac{\partial^2}{\partial r^2}F(t, r_t)d\langle r\rangle_t.$$

To overpass this problem, we change the definition of stochastic integral. We pass from the Itô integral to the Stratonovich stochastic integral, defined as follows:

$$\int_0^t X_s \circ dY_s := \int_0^t X_s dY_s + \frac{1}{2}d\langle X, Y\rangle_t.$$

If we consider a Stratonovich dynamics $dr_t = \widetilde{\mu}(r_t)dt + \widetilde{\sigma}(r_t) \circ dW_t$, then the chain rule holds:

$$dF(t, r_t) = \frac{\partial}{\partial t} F(t, r_t)dt + \frac{\partial}{\partial r} F(t, r_t) \circ dr_t. \tag{22}$$

Since, $d\langle \sigma(r)_\cdot, W\rangle_t = \frac{\partial \sigma}{\partial r}(r_t)\sigma(r_t)dt$, then let's consider:

$$dr_t = \Big(\mu_t - \frac{1}{2}\frac{\partial \sigma}{\partial r}(r_t)\sigma(r_t)\Big)dt + \sigma(r_t) \circ dW_t. \tag{23}$$

In conclusion, renaming the coefficients of dynamics of model $\mathcal{M}$ as in equation (23):

$$dr_t = \widetilde{\mu}(r_t)dt + \sigma(r_t) \circ dW_t, \tag{24}$$

the guess for consistency condition is:

$$\widetilde{\mu}(r) + \sigma(r)u \in \mathrm{Im}[G'_z(z)], \quad \forall u \in \mathbb{R},\ r = G(z), \tag{25}$$

that is equivalent to:

$$\begin{cases} \widetilde{\mu}(G(z)) \in \mathrm{Im}[G'_z(z)], \\ \sigma(G(z)) \in \mathrm{Im}[G'_z(z)], \end{cases}$$

for every $z \in \mathcal{Z}$.

In order to fully exploit the geometric approach, we introduce conditions that allows us to adopt the framework of Example 4.2 in the general multi-curve setting:

Assumption 4.3 Suppose that σ^j and β^j are smoothly defined in the functional form:

$$\sigma^j_t(x) \equiv \sigma^j(\hat{r}_t; x) \Rightarrow \sigma^j : \hat{\mathcal{H}} \times \mathbb{R}_+ \to \mathcal{H} \text{ is a smooth function,}$$
$$\beta^j_t \equiv \beta^j(\hat{r}_t) \Rightarrow \beta^j : \hat{\mathcal{H}} \to \mathbb{R}^d \text{ is a smooth function .}$$

Moreover, we assume that the following functions from $\hat{\mathcal{H}}$ are smooth:

$$\hat{r} \to \sigma^0(\hat{r})\mathbf{H}\sigma^0(\hat{r}) - \frac{1}{2}\frac{\partial}{\partial \hat{r}}\sigma^0(\hat{r}),$$
$$\hat{r} \to \sigma^j(\hat{r})\mathbf{H}\sigma^j(\hat{r}) - \frac{1}{2}\frac{\partial}{\partial \hat{r}}\sigma^j(\hat{r}) - \beta^j(\hat{r})\sigma^{j*}(\hat{r}).$$

Assumption 4.3 allows us to rewrite model 14 using the Stratonovich dynamics on $\hat{\mathcal{H}}$ in the functional form:

$$d\hat{r}_t = \hat{\mu}(\hat{r}_t)dt + \hat{\sigma}(\hat{r}_t) \circ dW_t. \tag{26}$$

Moreover, thanks to the assumptions, the coefficients of the model $\hat{\mu}, \hat{\sigma}$ are smooth functions from $\hat{\mathcal{H}}$ to itself. They can be interpreted as vector fields on $\hat{\mathcal{H}}$.

In this framework it is possible to prove that the notion of consistency between a model $\mathcal{M}$ given by a process $\hat{r}_t \in \hat{\mathcal{H}}$ and a manifold $\mathcal{G}$ determined by a parameterized family $G : \mathbb{R}^n \to \hat{\mathcal{H}}$ is equivalent to the following relation between $\hat{r}_t$ and G:

Definition 4.4 We say that $\mathcal{G} = Im[G]$ is locally $\hat{r}$-invariant under the action of the forward rate process $\hat{r}$ if for each $\hat{r}_0 \in \mathcal{G}$ there exist a positive time t^* and a stochastic process $(Z_t)_t$ on $\mathbb{R}^n$, whose dynamics is

$$dZ_t = a(Z_t)dt + b(Z_t) \circ dW_t, \tag{27}$$

such that for each $t \in [0, t^*]$, $\quad \hat{r}_t(x) = G(x, Z_t)$ for each x.

Exploiting this characterization of the consistency condition, a result analogous with the one presented in the finite dimensional case of Example 4.2 can be proved:

Theorem 4.5 *Consider a function $G : \mathcal{Z} \subset \mathbb{R}^n \to \hat{\mathcal{H}}$ such that:*

- *G is injective;*
- *$G'_z(z) \equiv dG|_z : \mathbb{R}^n \to \hat{\mathcal{H}}$ is injective.*

We introduce $\mathcal{G} := Im[G]$ and a model $\mathcal{M}$, given by the solution of the SDE on $\hat{\mathcal{H}}$:

$$\hat{r}_t = \hat{\mu}(\hat{r}_t)dt + \hat{\sigma}(\hat{r}_t) \circ dW_t. \tag{28}$$

Then, $(\mathcal{M}, \mathcal{G})$ form an consistent couple if and only if:

$$\begin{cases} \hat{\mu}(G(z)) \in Im[G'_z(z)] \equiv T_{G(z)}\mathcal{G}; \\ \hat{\sigma}^j(G(z)) \in Im[G'_z(z)] \equiv T_{G(z)}\mathcal{G}. \end{cases} \tag{29}$$

This result can be interpreted as follows: the couple $\mathcal{M}$ and $\mathcal{G}$ is consistent if and only if the coefficients of the model $\hat{\mu}$ and $\hat{\sigma}$ are tangent vector fields to the manifold $\mathcal{G}$.

5 Existence of finite dimensional realizations

In this section we exploit the concept of invariance developed Section 4, in order to understand if the solution to system 14 can be described as the image of a process whose dynamics is given by a finite-dimensional SDE. Moreover, if it is the case, we provide a strategy to construct the finite-dimensional process and the mapping which associates it with the forward rate $\hat{r}$. This task is called *existence of the finite-dimensional realizations problem.* The purpose is to find the conditions such that a model $\mathcal{M}$ determined by a forward-rate process $\hat{r}_t$ possesses an n-dimensional realization as in the following definition:

Definition 5.1 A model $\mathcal{M}$, whose trajectory is given by the forward-rate process $(\hat{r}_t)_{t\in[0,t^*]}$, solution to system 14, has an n-dimensional realization if there exists a strictly positive time t^*, a process $(Z_t)_{t\in[0,t^*]}$ taking values on $\mathcal{Z} \subseteq \mathbb{R}^n$ defined by the following SDE:

$$dZ_t = a(Z_t)dt + b(Z_t) \circ dW_t, \tag{30}$$

and a smooth injective immersion $G : \mathcal{Z} \to \hat{\mathcal{H}}$ such that:

$$\hat{r}_t(x) = G(Z_t, x), \quad t \in [0, t^*), \quad \forall x \in \mathbb{R}_+. \tag{31}$$

We say that $\mathcal{M}$ has finite-dimensional realizations (FDR in the following) if it has a n-dimensional realization for a suitable $n \in \mathbb{N}$.

5.1 Construction of finite-dimensional realization

The existence of an n-dimensional realizations as in Definition 5.1 is equivalent to the existence of a manifold $\mathcal{G}$ that is $\hat{r}$-invariant with the model $\mathcal{M}$ as in Definition 4.4. Since we stated that the notion of $\hat{r}$-invariance is equivalent to the notion of consistency, we can exploit Theorem 4.5 in order to say that the existence of FDR is equivalent to the existence of a function $G : \mathbb{R}^n \to \hat{\mathcal{H}}$ such that condition (29) is satisfied. Finally, if Assumption 4.3 holds, we can interpret $\hat{\mu}$ and $\hat{\sigma}$ as vector fields on $\hat{\mathcal{H}}$ which implies that condition (29) is equivalent to ask that $\hat{\mu}$ and $\hat{\sigma}$ are tangent to the unknown manifold $\mathcal{G} := \text{Im}[G]$.

As discussed, in order to solve the problem of existence of FDRs, we should be able to construct the tangential manifold $\mathcal{G}$ to the vector space generated by $\hat{\mu}(\hat{r})$ and $\hat{\sigma}(\hat{r})$ in each point $\hat{r} \in \text{Im}[G]$. By standard notions of differential geometry (we recall [3] as main reference) the vector space generated by a set of vector fields $\nu_1, \dots, \nu_m$ is called *distribution generated by* $\nu_1, \dots, \nu_m$. As a consequence, to solve the problem of existence of FDRs, we must construct a manifold $\mathcal{G}$ such that:

$$F(\hat{r}) \leq T_{\hat{r}}\mathcal{G}, \quad \forall \hat{r} \in \mathcal{U} \subseteq \hat{\mathcal{H}}, \text{ open.}$$

where F is the distribution generated by $\hat{\mu}$ and $\hat{\sigma}$. Therefore, we must find conditions that guarantee the existence of the tangential manifold $\mathcal{G}$ of a distribution F. To characterize it, we can exploit a version of the Frobenius theorem, that holds for general Hilbert spaces:

Theorem 5.2 (Frobenius) *Let F be an n-dimensional distribution and $\hat{r}^M \in \hat{\mathcal{H}}$. Then, there exists an n-dimensional tangential manifold to F through each $\hat{r} \in \mathcal{U}_{\hat{r}^M}$ if and only if F is involutive.*

We recall that F is involutive if for each couple of vector fields ξ and η in F, also their *Lie Bracket*:

$$[\xi, \eta](\hat{r}) := \xi'(\hat{r})\eta(\hat{r}) - \eta'(\hat{r})\xi(\hat{r}) \tag{32}$$

is in F.

In conclusion, in order to guarantee the existence of finite dimensional realizations we have to check if the smallest involutive distribution containing $\hat{\mu}$ and $\hat{\sigma}$, also called *Lie algebra $\mathcal{L}$ generated by $\hat{\mu}$ and $\hat{\sigma}$*, is finite-dimensional. If this holds, it is possible to construct a procedure to provide the finite dimensional realizations Z and the embedding G that links Z with $\hat{r}$. In this case, we built a strategy based on the results proposed in [4]:

(a) Find n generators $\{\xi^1, \dots, \xi^n\}$ of the Lie algebra $\mathcal{L}$.

(b) Define the tangential manifold of $\mathcal{L}$ as follows:

$$G(z^1, \dots, z^n) = e^{\xi_n z_n} \cdots e^{\xi_1 z_1} \hat{r}^M, \tag{33}$$

where $e^{\xi_i z_i} \hat{r}$ denotes the integral curve of the vector field ξ_i at time z_i passing through $\hat{r}$ at time 0. Since the Lie algebra is an involutive distribution, the tangential manifold G in equation (33) is invariant with respect to permutations of the integral curves ξ^h.

(c) The coefficients of $(Z_t)_{t \in [0,T]}$ are obtained inverting the consistency condition for G:

$$\begin{aligned} G'_z(z)a(z) &= \hat{\mu}(G(z)); \\ G'_z(z)b(z) &= \hat{\sigma}(G(z)). \end{aligned}$$

By Assumptions 4.3 $G'_z(z)$ is invertible, then we can define:

$$\begin{aligned} a(z) &= (G'_z)^{-1} \hat{\mu}(G(z)), \\ b(z) &= (G'_z)^{-1} \hat{\sigma}(G(z)), \end{aligned}$$

for each z. In particular, we obtained two vector fields on $\mathbb{R}^n$. Since they are smooth, the solution of the sde:

$$dZ_t = a(Z_t)dt + b(Z_t)dW_t.$$

exists uniquely around each point $z = Z_0$. By construction, it holds that $\hat{r}_t = G(Z_t)$ around the initial point $\hat{r}^M$.

6 Conclusions

In this document, we presented how the interest-rate market was modelled before the 2007 crisis with particular interest in the Heath-Jarrow-Morton class of models. Therefore, we described what happened during the global financial crisis of 2007 and how this crisis affected the interest-rate market. As a consequence, the classical pre-crisis framework was no longer appropriate to describe the market. To overpass this problem, we introduced the multi-curve setting adapting the Heath-Jarrow-Morton approach to it.

Then, we described the parameter recalibration problem, focusing on the main tasks that a financial analyst must solve to deal with interest-rate models in a post crisis framework. To attack this problem, we proposed a geometric approach, based on the notion of consistency between a model $\mathcal{M}$ and a parameterized family G. We provided a characterization of the consistency conditions in terms of the coefficients of the model, and we used it to construct a strategy that allows to determine the finite-dimensional realizations of a given model $\mathcal{M}$.

References

[1] T. Björk, *On the geometry of interest rate models.* In: Paris-Princeton Lectures on Mathematical Finance 2003. Springer, 2004, pp. 133–215.

[2] C. Cuchiero, C. Fontana, and A. Gnoatto, *A general HJM framework for multiple yield curve modelling.* In: Finance and Stochastics 20.2 (2016), pp. 267–320.

[3] S. Lang, “Fundamentals of differential geometry”. Vol. 191. Springer Science & Business Media, 2012.

[4] I. Slinko, *On finite dimensional realizations of two country interest rate models.* In: Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics 20.1 (2010), pp. 117–143.

Shear flows and viscoelastic fluids

Muhanna Ali H. Alrashdi [(*)]

1 Introduction

Rheology is the science that deals with the way materials deform when forces are applied to them. The term is most commonly applied to the study of liquids and liquid-like materials such as paint, blood, polymer solutions and molten plastics, i.e., materials that flow [1, 2]. Rheology also includes the study of the deformation of solids such as occurs in metal forming and the stretching of rubber. To learn anything about the rheological properties of a material, we must either measure the deformation resulting from a given force or measure the force required to produce a given deformation.

Steady simple shear is of central importance in applied rheology for two reasons. First, it is the flow that is by far the easiest to generate in the laboratory. Therefore, the data most often reported are based on this flow. Secondly, a number of processes of industrial importance, particularly extrusion and flow in many types of die, approximate steady simple shear flow. Moreover, simple shear is a solution of the flow equation irrespectively of the nature of the material. As such, it can be used to test material properties in a model-independent way.

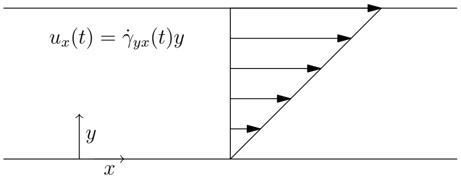
Simple-shear rheometry is essential for the measurement of viscoelastic shear properties. Shear stress is given by the ratio of the shear force F to the sample area A, whereas shear strain is defined as the ratio of the displacement x to the gap size h. Simple shear flow can be depicted as layers of fluid sliding over one another with each layer moving faster than the one beneath it. The uppermost layer has maximum velocity while the bottom layer is stationary.

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: `alrashdi@math.unipd.it`. Seminar held on 16 November 2022.

2 Shear flows and rheological properties

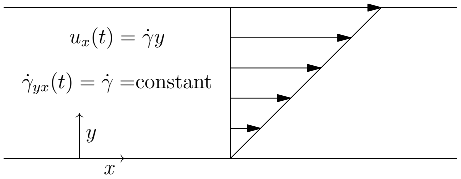
We nae consider some important flows.

- General shear flow.



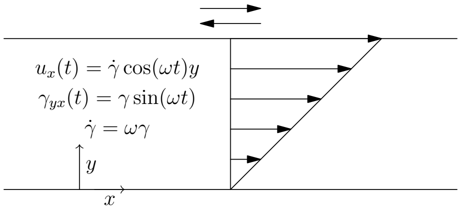
  We have two parallel plates. The top plate is moving with velocity u, it moves only in direction x. The intensity of the velocity grows linearly and $\dot\gamma$ is the slope. when velocity is 0, $\dot\gamma$ is 0.

- Simple steady shear.



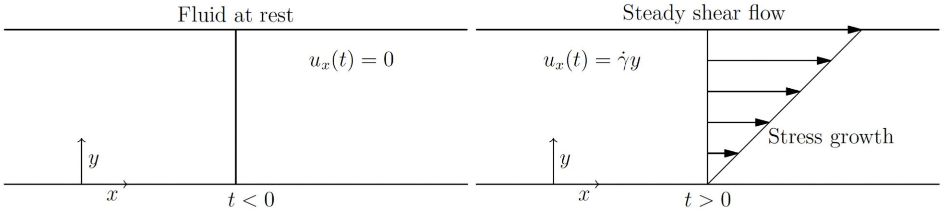
  A simple shear flow is easily generated between parallel plates with $\dot\gamma$ constant.

- Small amplitude oscillatory shear (SAOS).


$u_x(t) = \dot\gamma\cos(\omega t)y$
$\gamma_{yx}(t) = \gamma\sin(\omega t)$
$\dot\gamma = \omega\gamma$
y
x

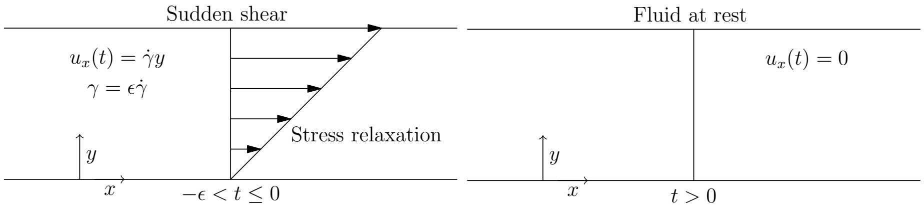
  The shear rate is a periodic function of time with aumplitude $\dot\gamma$ and frequency ω.

- Stress growth upon inception of steady shear flow.



  In a stress growth experiment, the fluid sample is presumed to be at rest for all times previous to $t = 0$; all components of stress are thus zero when the steady shearing starts at time $t = 0$. For times $t \geq 0$ we denote the constant rate as $\dot\gamma$. The objective of this experiment is to observe the approach of the stresses to their steady shear flow values.

- Stress relaxation after a sudden shearing displacement.



A fast deformation produces stress that subsequently relax while the fluid is at rest.

3 A new model

We present a new model that includes logarithmic strains. The placement $\varphi(X,t)$ gives the position at time t of a material point labelled with X and it solves $\frac{\partial\varphi(X,t)}{\partial t} = u(\varphi(X,t),t)$. We set $\hat{\mathsf{F}}(X,t) = \nabla\varphi$ and $\mathsf{F}(x,t) = \hat{\mathsf{F}}(\varphi^{-1}(x,t),t)$. The evolution equation for F is

$$\frac{\partial F}{\partial t} + (u\cdot\nabla)F = (\nabla u)\mathsf{F}$$

and the equation for left Cauchy-Green tensor $B = FF^T$ is

$$\frac{\partial B}{\partial t} + (u\cdot\nabla)B = (\nabla u)B + B(\nabla u)^T.$$

B is symmetric and positive definite with $\det B = 1$. Then $\log B$ is well defined and we have $\log B = \beta_1(t)[\vec{b_1}\otimes\vec{b_1} - \vec{b_2}\otimes\vec{b_2}]$. Where $\vec{b_1}, \vec{b_2}$ are eigenvectors of $\log B$ and $\beta_1, -\beta_1$ are its eigenvalues.

We introduce $\log B_{ref} = \gamma_1(t)[\vec{b_1}\otimes\vec{b_1} - \vec{b_2}\otimes\vec{b_2}]$, with γ_1 solution of

$$\dot\gamma_1(t) = \frac{1}{\lambda}(\beta_1(t) - \gamma_1(t)).$$

We define, with $K > 0$ and $\eta > 0$ constants, the elastic stress

$$\tau_{el} = K(\log B - \log B_{ref}) = K(\beta_1(t) - \gamma_1(t))[\vec{b_1}\otimes\vec{b_1} - \vec{b_2}\otimes\vec{b_2}]$$

and the viscous stress $\tau_{vi} = \eta_v\frac{D}{|D|}\dot\gamma_1(t)$ or $\tau_{vi} = \eta_v\frac{D}{|D|}\dot\xi$ with $\xi = \frac{1}{2}(B_{ref})_{yx}$.

4 Results

In Figure 1 we present some results based on the new model. Data labelled with UCM refer to the prediction of the Upper Convected Maxwell model.

The material function η^+ for stress growth upon inception of steady shear flow [2] is give by

$$\tau_{yx}(t,\dot\gamma) = \eta^+(t,\dot\gamma)\dot\gamma,$$

For the upper-convected Maxwell model [2] we have

$$\eta^+(\dot\gamma) = \eta(1 - e^{-\frac{t}{\tau_r}})$$

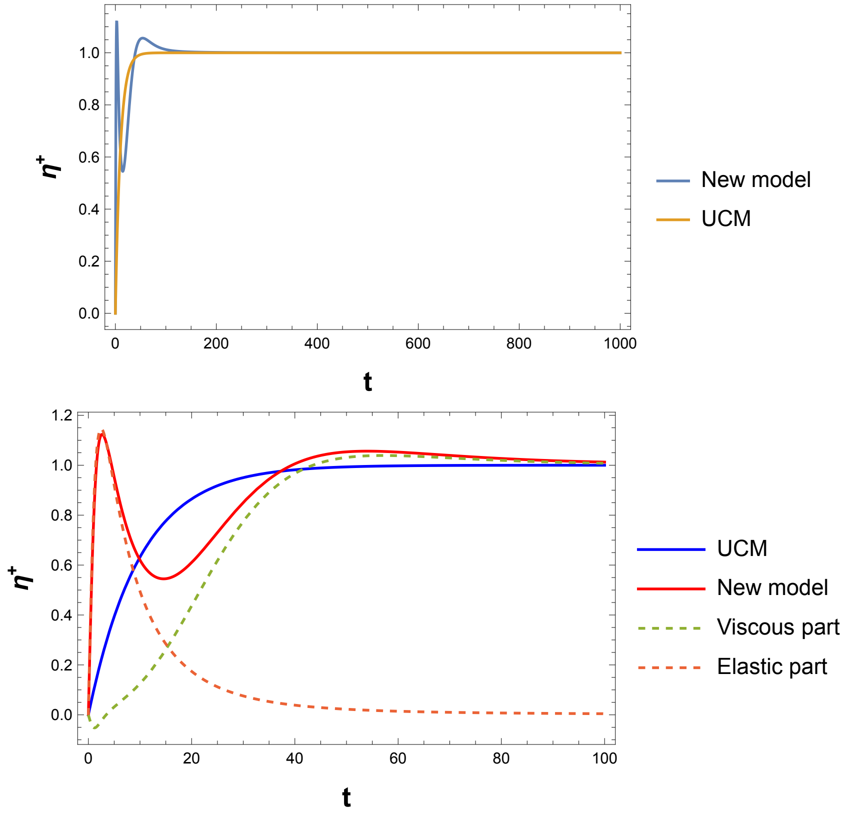


Figure 1: Comparison of the new model and (UCM).

References

[1] Nhan Phan-Thien, Nam Mai-Duy, "Understanding viscoelasticity: an introduction to rheology". Springer Graduate Texts in Physics, 2013.

[2] Bird, R.B.; Armstrong, R.C.; Hassager, O., "Dynamics of polymeric liquids. Vol. 1: Fluid mechanics". John Wiley and Sons Inc., New York, NY, 1987.

Moving Least Square approximation using variably scaled discontinuous weight function

Mohammad Karimnejad Esfahani [(*)]

Abstract. Functions with discontinuities appear in many applications such as image reconstruction, signal processing, optimal control problems, interface problems, engineering applications and so on. Accurate approximation and interpolation of these functions are therefore of great importance. In this paper, we design a moving least-squares approach for scattered data approximation that incorporates the discontinuities in the weight functions. The idea is to control the influence of the data sites on the approximant, not only with regards to their distance from the evaluation point, but also with respect to the discontinuity of the underlying function. We also provide an error estimate on a suitable *piecewise* Sobolev Space. The numerical experiments are in compliance with the convergence rate derived theoretically.

1 Introduction

In practical applications, over a wide range of studies such as surface reconstruction, numerical solution of differential equations and kernel learning, one has to solve the problem of reconstructing an unknown function $f : \Omega \longrightarrow \mathbb{R}$ sampled at some finite set of data sites $X = \{\mathbf{x}_i\}_{1\leq i\leq N} \subset \Omega \subset \mathbb{R}^d$ with corresponding data values $f_i = f(\mathbf{x}_i),\ 1 \leq i \leq N$. Since in practice the function values f_i are sampled at scattered points, and not at a uniform grid, Meshless (or meshfree) Methods (MMs) are used as an alternative of numerical mesh-based approaches, such as Finite Elements Method (FEM) and Finite Differences (FD). The idea of MMs could be traced back to [1]. Afterwards, multivariate MMs existed under many names and were used in different contexts; interested readers are referred to [2] for an overview over MMs. In a general setting, MMs are designed, at least partly, to avoid the use of an underlying mesh or triangulation. The approximant of f at X can be expressed in the form

$$s_{f,X}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i(\mathbf{x}) f_i. \tag{1}$$

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: karimnej@math.unipd.it. Seminar held on 30 November 2022.

One might seek a function $s_{f,X}$ that interpolates the data, i.e. $s_{f,X}(\mathbf{x}_i) = f_i,\ 1 \leq i \leq N$, and in this case $\alpha_i(\mathbf{x})$ will be the *cardinal functions.* However, one might consider a more generalized framework known as *quasi-interpolation* in which $s_{f,X}$ only approximates the data, i.e., $s_{f,X}(\mathbf{x}_i) \approx f_i$. The latter case means that we prefer to let the approximant $s_{f,X}$ only nearly fits the function values. This is useful, for instance, when the given data contain some noise, or the number of data is too large. The standard approach to deal with such a problem is to compute the Least-Squares (LS) solution, i.e., one minimizes the error (or cost) function

$$\sum_{i=1}^{N}[s_{f,X}(\mathbf{x}_i) - f_i]^2. \tag{2}$$

A more generalized setting of LS is known as the weighted LS, in which (2) turns to

$$\sum_{i=1}^{N}[s_{f,X}(\mathbf{x}_i) - f_i]^2 w(\mathbf{x}_i), \tag{3}$$

which is ruled by the *weighted* discrete ℓ_2 inner product. In practice the role of $w(\mathbf{x}_i)$ is to add more flexibility to the LS formulation for data f_i that influence the approximation process, which are supposed, for example, to be affected by some noise. However, these methods are global in the sense that all data sites have influence on the solution at any evaluation point $\mathbf{x} \in \Omega$. Alternatively, for a fixed evaluation point $\mathbf{x}$, one can consider only n-th closest data sites $\mathbf{x}_i$, $i = 1, \ldots, n$ of $\mathbf{x}$ such that $n \ll N$. The *Moving Least-Squares* (MLS) method, which is a *local* variation of the classical weighted least-squares technique, has been developed following this idea. To be more precise, in the MLS scheme, for each evaluation point $\mathbf{x}$ one needs to solve a *weighted least-squares* problem, minimizing

$$\sum_{i=1}^{N}[s_{f,X}(\mathbf{x}_i) - f_i]^2 w(\mathbf{x}, \mathbf{x}_i) \tag{4}$$

by choosing the weight functions $w(\mathbf{x}, \mathbf{x}_i) : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ to be localized around $\mathbf{x}$, so that *few* data sites are taken into account. The key difference with respect to (3) is that the weight function is indeed *moving* with the evaluation point, meaning that it depends on both the $\mathbf{x}_i$ and $\mathbf{x}$. Consequently, for each evaluation point $\mathbf{x}$, a small linear system needs to be solved. Also, one can let $w(\cdot, \mathbf{x}_i)$ be a radial function i.e., $w(\mathbf{x}, \mathbf{x}_i) = \varphi(\|\mathbf{x} - \mathbf{x}_i\|_2)$ for some non-negative univariate function $\varphi : [0, \infty) \longrightarrow \mathbb{R}$. Doing in this way, $w(\cdot, \mathbf{x}_i)$ inherits the translation invariance property of radial basis functions. We mention that (4) could be generalized as well by letting $w_i(\cdot) = w(\cdot, \mathbf{x}_i)$ moves with respect to a *reference* point $\mathbf{y}$ such that $\mathbf{y} \neq \mathbf{x}$.

The earliest idea of MLS approximation technique can be traced back to Shepard's seminal paper [3], and later on developed in [4]. For more details and error analysis we refer the readers to [5, Chap 3-4].

The MLS method has rarely been used for approximating piecewise-continuous functions, i.e, functions that possess some discontinuities or jumps. The idea that is followed in this work is to consider the weight functions in a way that, they take the jumps into account with hope to produce more accurate approximation. In the following, we recall necessary notions of the MLS, Variably Scaled Discontinuous Kernels, and Sobolev spaces. Then we presents the original contribution of this work, both from theoretical and practical point of view.

2 Preliminaries on MLS and VSKs

2.1 Moving Least Squares (MLS) approximation

Let Ω be a non-empty and bounded domain in $\mathbb{R}^d$ and X be the set of N distinct data sites (or centers). We consider the target function f, and the corresponding function values f_i as defined above. Moreover, $\mathcal{P}_\ell^d$ indicates the space of d-variate polynomials of degree at most $\ell \in \mathbb{N}$, with basis $\{p_1, ..., p_Q\}$ and dimension $Q = \binom{\ell+d}{d}$.

Several equivalent formulations exist for the MLS approximation scheme. As the standard formulation, the MLS approximant looks for the best weighted approximation to f at the evaluation point $\mathbf{x}$ in $\mathcal{P}_\ell^d$ (or any other linear space of functions $\mathcal{U}$), with respect to the discrete ℓ_2 norm induced by the weighted inner product $\langle f, g\rangle_{w_\mathbf{x}} = \sum_{i=1}^N w(\mathbf{x}_i, \mathbf{x}) f(\mathbf{x}_i) g(\mathbf{x}_i)$. Mathematically speaking, the MLS approximant will be the linear combination of the polynomial basis i.e.,

$$s_{f,X}(\mathbf{x}) = \sum_{j=1}^{Q} c_j(\mathbf{x}) p_j(\mathbf{x}), \tag{5}$$

where the coefficients are obtained by locally minimizing the weighted least square error in (4), which is equivalent to minimizing $\|f - s_f\|_{w_\mathbf{x}}$. We highlight that the local nature of the approximant is evident from the fact that the coefficient $c_j(\mathbf{x})$ must be computed for each evaluation point $\mathbf{x}$.

In another formulation of MLS approximation known as the *Backus-Gilbert* approach, one considers the approximant $s_{f,X}(\mathbf{x})$ to be a *quasi interpolant* of the form (1). In this case, one seeks the values of the basis functions $\alpha_i(\mathbf{x})$ (also known as generating or shape functions) as the minimizers of

$$\frac{1}{2} \sum_{i=1}^{N} \alpha_i^2(\mathbf{x}) \frac{1}{w(\mathbf{x}_i, \mathbf{x})} \tag{6}$$

subject to the polynomial reproduction constraints

$$\sum_{i=1}^{N} p(\mathbf{x}_i) \alpha_i(\mathbf{x}) = p(\mathbf{x}), \quad \text{for all } p \in \mathcal{P}_\ell^d.$$

Such a constrained quadratic minimization problem can be converted to a system of linear equations by introducing Lagrange multipliers $\boldsymbol{\lambda}(\mathbf{x}) = [\lambda_1(\mathbf{x}), ..., \lambda_Q(\mathbf{x})]^T$. Consequently (e.g. see [5, Corollary 4.4]), the MLS basis function α_i evaluated at $\mathbf{x}$ is given by

$$\alpha_i(\mathbf{x}) = w(\mathbf{x}, \mathbf{x}_i) \sum_{k=1}^{Q} \lambda_k(\mathbf{x}) p_k(\mathbf{x}_i), \quad 1 \leq i \leq N, \tag{7}$$

such that $\lambda_k(\mathbf{x})$ are the unique solution of

$$\sum_{k=1}^{Q} \lambda_k(\mathbf{x}) \sum_{i=1}^{N} w(\mathbf{x}, \mathbf{x}_i) p_k(\mathbf{x}_i) p_s(\mathbf{x}_i) = p_s(\mathbf{x}), \quad 1 \leq s \leq Q. \tag{8}$$

We observe that the weight function $w_i(\mathbf{x}) = w(\mathbf{x}, \mathbf{x}_i)$ controls the influence of the center $\mathbf{x}_i$ over the approximant, so it should be *small* when evaluated at a point that is far from $\mathbf{x}$, that is it should decay to zero fast enough. To this end we may let $w_i(\mathbf{x})$ be positive on a ball centered at $\mathbf{x}$ with radius r, $B(\mathbf{x}, r)$, and zero outside. For example, a compactly supported radial kernel satisfies such a behaviour. Thus, let $I(\mathbf{x}) = \{i \in \{1, \dots, N\}, \|\mathbf{x} - \mathbf{x}_i\|_2 \leq r\}$ be the family of indices of the centers X, for which $w_i(\mathbf{x}) > 0$, with $|I| = n \ll N$. Only the centers $\mathbf{x}_i \in I$ influence the approximant $s_{f,X}(\mathbf{x})$. Consequently, the matrix representation of (7) and (8) is

$$\boldsymbol{\alpha}(\mathbf{x}) = W(\mathbf{x}) P^T \boldsymbol{\lambda}(\mathbf{x}),$$
$$\boldsymbol{\lambda}(\mathbf{x}) = (P W(\mathbf{x}) P^T)^{-1} \mathbf{p}(\mathbf{x}),$$

where $\boldsymbol{\alpha}(\mathbf{x}) = [\alpha_1(\mathbf{x}), ..., \alpha_n(\mathbf{x})]^T$, $W(\mathbf{x}) \in \mathbb{R}^{n \times n}$ is the diagonal matrix carrying the weights $w_i(\mathbf{x})$ on its diagonal, $P \in \mathbb{R}^{Q \times n}$ such that its k-th row contains p_k evaluated at data sites in $I(\mathbf{x})$, and $\mathbf{p}(\mathbf{x}) = [p_1(\mathbf{x}), ..., p_Q(\mathbf{x})]^T$. More explicitly the basis functions are given by

$$\boldsymbol{\alpha}(\mathbf{x}) = W(\mathbf{x}) P^T (P W(\mathbf{x}) P^T)^{-1} \mathbf{p}(\mathbf{x}). \tag{9}$$

Moreover, it turns out that the solution of (5) is identical to the solution offered by the Backus-Gilbert approach (see e.g. [5, Chap. 3-4]).

In the MLS literature, it is known that a local polynomial basis shifted to the evaluation point $\mathbf{x} \in \Omega$ leads to a more stable method (see e.g. [5, Chap. 4]). Accordingly, we let the polynomial basis to be $\{1, (\cdot - \mathbf{x}), \dots, (\cdot - \mathbf{x})^\ell\}$, meaning that different bases for each evaluation point are employed. In this case, since with standard monomials basis we have $p_1 \equiv 1$ and $p_k(0) = 0$ for $2 \leq k \leq Q$, then $\mathbf{p}(\mathbf{x}) = [1, 0, ..., 0]^T$.

To ensure the invertibility of $PW(\mathbf{x})P^T$ in (9), X needs to be $\mathbb{P}^d_\ell$-unisolvent. Then as long as $w_i(\mathbf{x})$ is positive, $PW(\mathbf{x})P^T$ will be a positive definite matrix, and so invertible; more details are available in [12, Chap. 22].

Furthermore, thanks to equation (7), it is observable that the behaviour of $\alpha_i(\mathbf{x})$ is heavily influenced by the behaviour of the weight functions $w_i(\mathbf{x})$, in particular it includes continuity and the support of the basis functions $\alpha_i(\mathbf{x})$. Another significant feature is that the weight functions $w_i(\mathbf{x})$ which are singular at the data sites lead to cardinal basis functions i.e., $\alpha_i(\mathbf{x}_j) = \delta_{i,j}$ $i, j = 1, ..., n$, meaning that MLS scheme interpolates the data (for more details see [6, Theorem 3]).

We also recall the following definitions that we will use for the error analysis.

(a) A set $X = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ with $Q \leq N$ is called $\mathbb{P}_\ell^d$-unisolvent if the zero polynomial is the only polynomial from $\mathbb{P}_\ell^d$ that vanishes on X.

(b) The **fill distance** is defined as

$$h_{X,\Omega} = \sup_{\mathbf{x}\in\Omega} \min_{1\leq j\leq N} \|\mathbf{x} - \mathbf{x}_j\|_2.$$

(c) The **separation distance**

$$q_X = \frac{1}{2}\min_{i\neq j} \|\mathbf{x}_i - \mathbf{x}_j\|.$$

(d) The set of data sites X is said to be **quasi-uniform** with respect to a constant $c_{qu} > 0$ if

$$q_X \leq h_{X,\Omega} \leq c_{qu} q_X.$$

2.2 Sobolev spaces and error estimates for MLS

Assume $k \in \mathbb{N}_0$ and $p \in [1, \infty)$, then the *integer-order* Sobolev space $W_p^k(\Omega)$ consists of all u with distributional (weak) derivatives $D^{\boldsymbol{\delta}} u \in L^p$, $|\boldsymbol{\delta}| \leq k$. The semi-norm and the norm associated with these spaces are

$$(10) \qquad |u|_{w_p^k(\Omega)} := \Big(\sum_{|\boldsymbol{\delta}|=k} \|D^{\boldsymbol{\delta}} u\|_{L^p(\Omega)}^p\Big)^{1/p}, \qquad \|u\|_{w_p^k(\Omega)} := \Big(\sum_{|\boldsymbol{\delta}|\leq k} \|D^{\boldsymbol{\delta}} u\|_{L^p(\Omega)}^p\Big)^{1/p}.$$

Moreover, letting $0 < s < 1$, the *fractional-order* Sobolev space $W_p^{k+s}(\Omega)$ is the space of the functions u for which semi-norm and norm are defined as

$$|u|_{w_p^{k+s}(\Omega)} := \Big(\sum_{|\boldsymbol{\delta}|=k} \int_\Omega \int_\Omega \frac{|D^{\boldsymbol{\delta}} u(\mathbf{x}) - D^{\boldsymbol{\delta}} u(\mathbf{y})|^p}{|\mathbf{x} - \mathbf{y}|^{d+ps}}\Big)^{1/p}$$

$$\|u\|_{W_p^{k+s}(\Omega)} := \Big(\|u\|_{W_p^k(\Omega)} + |u|_{W_p^{k+s}(\Omega)}\Big)^{1/p}.$$

Consider certain Sobolev spaces $W_p^k(\Omega)$ with the condition that $1 < p < \infty$ and $k > m+d/p$ (for $p = 1$ the equality is also possible), then according to [8, Theorem 2.12] the sampling inequality

$$\|u\|_{W_p^m(\Omega)} \leq C h_{X,\Omega}^{k-m-d(1/p-1/p)_+} \|u\|_{W_p^k}$$

holds for a function u that satisfies $u(X) = 0$, with $h_{X,\Omega}$ being the *fill distance* associated with X and $(\mathbf{y})_+ = \max\{0, \mathbf{y}\}$. For more information regarding Sobolev Spaces and sampling inequalities we refer the reader to [13] and [14], respectively.

Getting back to the MLS scheme, let $D^{\boldsymbol{\delta}}$ be a derivative operator such that $|\boldsymbol{\delta}| \leq \ell$ (we recall that ℓ is the maximum degree of the polynomials). Under some mild conditions regarding the weight functions, [7, Theorem 3.11] shows that $\{D^{\boldsymbol{\delta}} \alpha_i(\mathbf{x})\}_{1\leq i\leq n}$ forms a *local polynomial reproduction* in a sense that there exist constants h_0, $C_{1,\boldsymbol{\delta}}$, C_2 such that for every evaluation point $\mathbf{x}$

- $\sum_{i=1}^{N} D^{\boldsymbol{\delta}}\alpha_i(\mathbf{x})p(\mathbf{x}_i) = p(\mathbf{x})$ for all $p \in \mathbb{P}_\ell^d$
- $\sum_{i=1}^{N}|D^{\boldsymbol{\delta}}\alpha_i(\mathbf{x})|\leq C_{1,\boldsymbol{\delta}} h_{X,\Omega}^{-|\boldsymbol{\delta}|}$
- $D^{\boldsymbol{\delta}}\alpha_i(\mathbf{x}) = 0$ provided that $\|\mathbf{x}-\mathbf{x}_i\|_2 \geqslant C_2 h_{X,\Omega}$

for all X with $h_{X,\Omega} \leq h_0$.

The particular case of $|\boldsymbol{\delta}|= 0$ was previously discussed in [5, Theorem 4.7] in which it is shown that $\{\alpha_i(\mathbf{x})\}_{1\leq i\leq n}$ forms a local polynomial reproduction. However in this case the basis functions $\{\alpha_i(\cdot)\}_{1\leq i\leq n}$ could be even discontinuous but it is necessary that $w_i(\mathbf{x})$ are bounded (for more details see [5, Chap 3-4]). Consequently we restate the the MLS error bound in Sobolev Spaces developed in [7].

Theorem 1 [7, Theorem 3.12] *Suppose that $\Omega \subset \mathbb{R}^d$ is a bounded set with a Lipschitz boundary. Let ℓ be a positive integer, $0 \leq s < 1$, $p \in [1,\infty)$,$q \in [1,\infty]$ and let $\boldsymbol{\delta}$ be a multi-index satisfying $\ell > |\boldsymbol{\delta}|+d/p$ for $p > 1$ and $\ell \geqslant |\boldsymbol{\delta}|+d$ for $p = 1$. If $f \in W_p^{\ell+s}(\Omega)$, there exist constants $C > 0$ and $h_0 > 0$ such that for all $X = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \subset \Omega$ which are quasi-uniform with $h_{X,\Omega} \leq \min\{h_0, 1\}$, the error estimate holds*

$$\|f - s_{f,X}\|_{W_q^{|\boldsymbol{\delta}|}(\Omega)} \leq C h_{X,\Omega}^{\ell+s-|\boldsymbol{\delta}|-d(1/p-1/q)_+} \|f\|_{W_p^{\ell+s}(\Omega)}. \tag{11}$$

when the polynomial basis, are shifted to the evaluation point $\mathbf{x}$ and scaled with respect to the fill distance $h_{X,\Omega}$, and $w_i(\cdot)$ is positive on $[0, 1/2]$, supported in $[0, 1]$ such that its even extension is non negative and continuous on $\mathbb{R}$.

Remark 1 The above error bounds holds also when $s = 1$. However, recalling the definition of (semi-)norms in *fractional-order* Sobolev space, we see that in this case we reach to an *integer-order* Sobolev space of $\ell+1$. Therefore, it requires that $\ell+1 > |\boldsymbol{\delta}|+d/p$ for $p > 1$ or $\ell + 1 \geqslant |\boldsymbol{\delta}|$ for $p = 1$ in order that (11) holds true. The key point is that in this case, the polynomial space is still $\mathcal{P}_\ell^d$ and not $\mathcal{P}_{\ell+1}^d$.

2.3 Variably Scaled Discontinuous Kernels (VSDKs)

Variably Scaled Kernels (VSKs) were firstly introduced in [10]. The basic idea behind them is to map the data sites from $\mathbb{R}^d$ to $\mathbb{R}^{d+1}$ via a scaling function $\psi : \Omega \longrightarrow \mathbb{R}$ and to construct an augmented approximation space in which the data sites are $\{(\mathbf{x}_i, \psi(\mathbf{x}_i))\ i = 1, ..., N\}$ (see [10, Def. 2.1]). Though the first goal of doing so was getting a *better* nodes distribution in the augmented dimension, later on in [9] the authors came up with the idea of also encoding the behaviour of the underlying function f inside the scale function ψ. Precisely, for the target function f that possesses some jumps, the key idea is the following.

Definition 1 Let $\mathcal{P} = \{\Omega_1, ..., \Omega_n\}$ be a partition of Ω and let $\boldsymbol{\beta} = (\beta_1, ..., \beta_n)$ be a vector of real distinct values. Moreover, assume that all the jump discontinuities of the underlying function f lie on $\bigcup_{j=1}^n \partial\Omega_j$. The piecewise constant scaling function $\psi_{\mathcal{P},\boldsymbol{\beta}}$ with respect to the partition $\mathcal{P}$ and the vector $\boldsymbol{\beta}$ is defined as

$$\psi_{\mathcal{P},\boldsymbol{\beta}}(\mathbf{x})|_{\Omega_j} = \beta_j,\ \mathbf{x} \in \Omega.$$

Successively, let Φ^{ε} be a positive definite radial kernel on $\Omega \times \Omega$ that depends on the shape parameter $\varepsilon > 0$. A variably scaled discontinuous kernel on $(\Omega \times \mathbb{R}) \times (\Omega \times \mathbb{R})$ is defined as

$$\Phi_{\psi}^{\varepsilon}(\mathbf{x}, \mathbf{y}) = \Phi^{\varepsilon}\big(\Psi(\mathbf{x}), \Psi(\mathbf{y})\big), \quad \mathbf{x}, \mathbf{y} \in \Omega. \tag{12}$$

such that $\Psi(\mathbf{x}) = (\mathbf{x}, \psi(\mathbf{x}))$.

Moreover, we point out that if Φ^{ε} is (strictly) positive definite then so is $\Phi_{\psi}^{\varepsilon}$, and if Φ^{ε} and ψ are continuous then so is $\Phi_{\psi}^{\varepsilon}$ [10, Theorem 2.2].

3 MLS-VSDKs

Let f be a function with some jump discontinuities defined on Ω, $\mathcal{P}$ and $\psi_{\mathcal{P},\beta}$ as in Definition 1. We look for the MLS approximant with *variably scaled discontinuous weight function* such that

$$w_{\psi}(\mathbf{x}, \mathbf{x}_i) = w(\Psi(\mathbf{x}), \Psi(\mathbf{x}_i)). \tag{13}$$

Above all, we point out that in this case the diagonal matrix $W(\mathbf{x})$ in (9) still carries only positive values by assumption, and therefore the equation (9) is still solvable meaning that the basis functions $\alpha(\mathbf{x})$ uniquely exist. However, with new weight functions, from (13) also $\alpha(\mathbf{x})$ might be continuous or discontinuous regarding to the given data values f_i. Therefore our basis functions are indeed data-dependent thanks to (13). From now on, we call this scheme MLS-VSDK, and we will denote the corresponding approximant as $s_{f,X}^{\psi}$.

Since the basis functions are data dependent, one might expect that the space in which we express the error bound should be data dependent as well. Towards this idea, for $k \in \mathbb{Z}$, $0 \le k$, and $1 \le p \le \infty$, we define the *piecewise* Sobolev Spaces

$$\mathcal{W}_p^k(\Omega) = \{f : \Omega \longrightarrow \mathbb{R} \text{ s.t. } f_{|\Omega_j} \in W_p^k(\Omega_j), \quad j \in \{1, ..., n\}\},$$

where $f_{|\Omega_j}$ denotes the restriction of f to Ω_j, and $W_p^k(\Omega_j)$ denote the Sobolev space on Ω_i. We endow $\mathcal{W}_p^k(\Omega)$ with the norm

$$\|f\|_{\mathcal{W}_p^k(\Omega)} = \sum_{j=1}^{n} \|f\|_{W_p^k(\Omega_j)}. \tag{14}$$

When $k = 0$ we simply denote $\mathcal{W}_p^0(\Omega)$ by $\mathcal{L}^p(\Omega)$. Moreover, it could be shown that for any partition of Ω the standard Sobolev space $W_p^k(\Omega)$ is contained in $\mathcal{W}_p^k(\Omega)$ (see [11] and reference therein). We assume that every set $\Omega_j \in \mathcal{P}$ satisfies Lipschitz boundary conditions which will be essential for our error analysis.

Proposition 1 *Let $\mathcal{P}$ be as in Definition 1 and set the derivative order $\boldsymbol{\delta} = 0$. Then, by using Theorem 1, the error satisfies the inequality*

$$\|f - s_{f,X}^{\psi}\|_{L^2(\Omega_j)} \le C_j h_{\Omega_j}^{\ell+1-d(1/p-1/2)_+} \|f\|_{W_p^{\ell+1}(\Omega_j)}, \qquad \textit{for all } \Omega_j \in \mathcal{P} \tag{15}$$

with h_{Ω_j} the fill distance with respect to Ω_j.

Proof. Recalling Definition 1 we know that the discontinuities of f and subsequently $w_i(\cdot)$ are located only at the boundary and not on the domain Ω_j, meaning that $w_i(\cdot)$ is continuous inside Ω_j. Furthermore, the basis $\{\alpha_i(\mathbf{x})\}_{1\leq i\leq n}$ forms a local polynomial reproduction i.e., there exists a constant C such that $\sum_{i=1}^{N}|\alpha_i|\leq C$. Letting $s = 1$ and $q = 2$, by noticing that $W_q^{0'}(\Omega_j) = L^q(\Omega_j)$, then the error bound (15) is an immediate consequence of Theorem 1. □

From the above proposition, it could be understood that $s_{f,X}^{\psi}$ behaves similarly to $s_{f,X}$ in the domain Ω_j, where there is no discontinuity. This is in agreement with Definition 1. Consequently, it is required to extend the error bound (15) to the whole domain Ω.

Theorem 2 *Let f, $\mathcal{P}$, $\psi_{\mathcal{P},\beta}$ be as before, and the weight functions as in (13). Then, for $\ell > |\boldsymbol{\delta}|+d/p$ (equality also holds for $p = 1$), and $f \in \mathcal{W}_p^{\ell+1}(\Omega)$, for the MLS-VSDK approximant $s_{f,X}^{\psi}$ the error can be bounded as follows:*

$$\|f - s_{f,X}^{\psi}\|_{\mathcal{L}^2(\Omega)}\leq Ch^{\ell+1-d(1/p-1/2)_+}\|f\|_{\mathcal{W}_p^{\ell+1}(\Omega)} \tag{16}$$

Proof. By Proposition 1, we know that (15) holds for each Ω_j. Let h_{X,Ω_i} and C_i be the fill distance and a constant associated with each Ω_i, respectively. Then, we have

$$\sum_{j=1}^{n}\|f - s_{f,X}^{\psi}\|_{L^2(\Omega_j)}\leq \sum_{j=1}^{n} C_j h_{X,\Omega_i}^{\ell+1-d(1/p-1/2)_+}\|f\|_{W_p^{\ell+1}(\Omega_j)}.$$

By definition we get $\sum_{j=1}^{n}\|f - s_{f,X}^{\psi}\|_{L^2(\Omega_j)}= \|f - s_{f,X}^{\psi}\|_{\mathcal{L}^2(\Omega)}$. Moreover, letting $C = \max\{C_1, ..., C_n\}$ and $h = max\{h_{X,\Omega_1}, ..., h_{X,\Omega_n}\}$ then the right hand side can be bounded by

$$Ch^{\ell+1-d(1/p-1/2)_+}\|f\|_{\mathcal{W}_p^{\ell+1}(\Omega)}.$$

Putting these together we conclude. □

Some remarks are in order.

(a) One might notice that the error bound in (11) is indeed local (the basis functions are local by assumption), meaning that if f is less smooth in a subregion of Ω, say it possesses only $\ell' \leq \ell$ continuous derivatives there, then the approximant (interpolant) has order $\ell' + 1$ in that region and this is the best we can get. On the other hand according to (16), thanks to the definition of piecewise Sobolev space, the regularity of the underlying function in the interior of the subdomain Ω_j matters. In other words, as long as f possesses regularity of order ℓ in subregions, say Ω_j and Ω_{j+1}, the approximant order of $\ell + 1$ is achievable, regardless of the discontinuities on the boundary of Ω_j and Ω_{j+1}.

(b) Another interesting property of the MLS-VSDK scheme is that it is indeed data dependent. To clarify, for the evaluation point $\mathbf{x} \in \Omega_j$ take two data sites $\mathbf{x}_i$, $\mathbf{x}_{i+1} \in B(\mathbf{x}, r)$ with the same distance from $\mathbf{x}$ such that $\mathbf{x}_i \in \Omega_j$ and $\mathbf{x}_{i+1} \in \Omega_{j+1}$. Due to the definition (12), $w_\psi(\mathbf{x}, \mathbf{x}_{i+1})$ decays to zero faster than $w_\psi(\mathbf{x}, \mathbf{x}_i)$ i.e., the data sites from the same subregion Ω_j pay more contribution to the approximant (interpolant) $s^{\psi}_{f,X}$, rather than the one from another subregion Ω_{j+1} beyond a discontinuity line On the other hand in the classical MLS scheme, this does not happen as the weight function gives the same value to both $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$.

We end this section by recalling that the MLS approximation convergence order is achievable only in the *stationary setting*, i.e., the shape parameter ε must be scaled with respect to the fill distance. It leads to *peaked basis functions* for densely spaced data and *flat basis function* for coarsely spaced data. In other words, the local support of the weight functions $B(\mathbf{x}, r)$, and subsequently basis functions must be tuned with regards to the $h_{X,\Omega}$ using the shape parameter ε. Consequently, this holds also in MLS-VSDK scheme, meaning that after scaling w_i we still need to take care of ε. This is different with respect to VS(D)Ks interpolation where $\varepsilon = 1$ was kept fixed [10, 9].

4 Numerical experiments

In this section, we compare the performance of the MLS-VSDK with respect to the classical MLS method. In all numerical tests we fix the polynomials space up to degree 1. Considering the evaluation points as $Z = \{z_1, ..., z_s\}$ we compute root mean square error and maximum error by

$$RMSE = \sqrt{\frac{1}{s}\sum_{i=1}^{s}(f(z_i) - s_{f,X}(z_i))^2}, \quad MAE = \max_{z_i \in Z} |f(z_i) - s_{f,X}(z_i)|.$$

We consider four different weight functions to verify the convergence order of $s^{\psi}_{f,x}$ to a given f, as presented in Theorem 2.

(a) $w^1(\mathbf{x}, \mathbf{x}_i) = (1 - \varepsilon\|\mathbf{x} - \mathbf{x}_i\|)^4_+ \cdot (4\varepsilon\|\mathbf{x} - \mathbf{x}_i\| + 1)$, which is the well-known C^2 *Wendland* function. Since each w^1_i is locally supported on the open ball $B(0, 1)$ then it verifies the conditions required by Theorem 2.

(b) $w^2(\mathbf{x}, \mathbf{x}_i) = \exp(-\varepsilon\|\mathbf{x} - \mathbf{x}_i\|^2)$, i.e. the Gaussian RBF. We underline that when Gaussian weight functions are employed, with decreasing separation distance of the approximation centers, the calculation of the basis functions in (9) can be badly conditioned. Therefore, in order to make the computations stable, in this case we regularize the system by adding a small multiple, say $\lambda = 10^{-8}$, of the identity to the diagonal matrix W.

(c) $w^3(\mathbf{x}, \mathbf{x}_i) = \exp(-\varepsilon\|\mathbf{x} - \mathbf{x}_i\|)(15 + 15\|\mathbf{x} - \mathbf{x}_i\| + 6\|\mathbf{x} - \mathbf{x}_i\|^2 + \|\mathbf{x} - \mathbf{x}_i\|^3)$, that is a C^6 *Matérn* function.

(d) $w^4(\mathbf{x}, \mathbf{x}_i) = (\exp(\varepsilon\|\mathbf{x} - \mathbf{x}_i\|)^2 - 1)^{-1}$, suggested in [6], which enjoys an additional feature which leads to interpolatory MLS, since it possesses singularities at the centers.

One might notice that w^2, w^3 and w^4 are not locally supported. However, the key point is that they are all decreasing with the distance from the centers and so, in practice, one can overlook the data sites that are so far from the center $\mathbf{x}$. As a result, one generally considers a *local stencil* containing n nearest data sites of the set Z of evaluation points. While there is no clear theoretical background concerning the stencil size, in MLS literature, one generally lets $n = 2 \times Q$ (see e.g. [15]). However, it might be possible that in some special cases one could reach a better accuracy using different stencil sizes. This aspect is covered by our numerical tests, which are outlined in the following.

(a) In Section 4.1, we move to the two-dimensional framework and we keep the same stencil size. Here, we restrict the test to the weight function w^1 and verify Theorem 2.

(b) In Section 4.2, we remain in the two-dimensional setting but the best accuracy is achieved with $n = 20$. Moreover, in addition to w^2 and w^3, we test the interpolatory case by considering w^4 as weight function.

(c) In Section 4.3, we present a two-dimensional experiments where the data sites have been perturbed via some white noise. We fix $n = 25$ and w^2, w^3 are involved.

4.1 Example 1

Consider on $\Omega = (-1, 1)^2$ the discontinuous function

$$f_2(x, y) = \begin{cases} \exp(-(x^2 + y^2)), & x^2 + y^2 \leq 0.6 \\ x + y, & x^2 + y^2 > 0.6 \end{cases}$$

and the discontinuous scale function

$$\psi(x, y) = \begin{cases} 1, & x^2 + y^2 \leq 0.6 \\ 2, & x^2 + y^2 > 0.6 \end{cases}$$

As evaluation points, we take the grid of equispaced points with mesh size $1.00e-2$. Figure 1 shows both the *RMSE* and *absolute error* for the classical *MLS* and *MLS-VSDK* approximation of f_2 sampled from $1089 = 33^2$ uniform data sites taking w^1 as the weight function. Figure 1 shows that using classical MLS, the approximation error significantly increases near the discontinuities, while using MLS-VSDK the approximant can overcome this issue. In order to investigate the convergence rate, we consider increasing sets of $\{25, 81, 289, 1089, 4225, 16641\}$ Halton and uniform points as the data sites. To find an appropriate value for the shape parameter, we fix an initial value and we multiply it by a factor of 2 at each step. Thus, let $\boldsymbol{\varepsilon} = [0.25, 0.5, 1, 2, 4, 8]$ be the vector of shape parameter which is modified with respect to the number of the centers in both cases of uniform and Halton data sites. The left plot of Figure 2 shows a convergence rate of 2.58 for the MLS-VSDK and only 0.66 for classical MLS methods, while these values are 2.04 and 0.70 in the right plot.

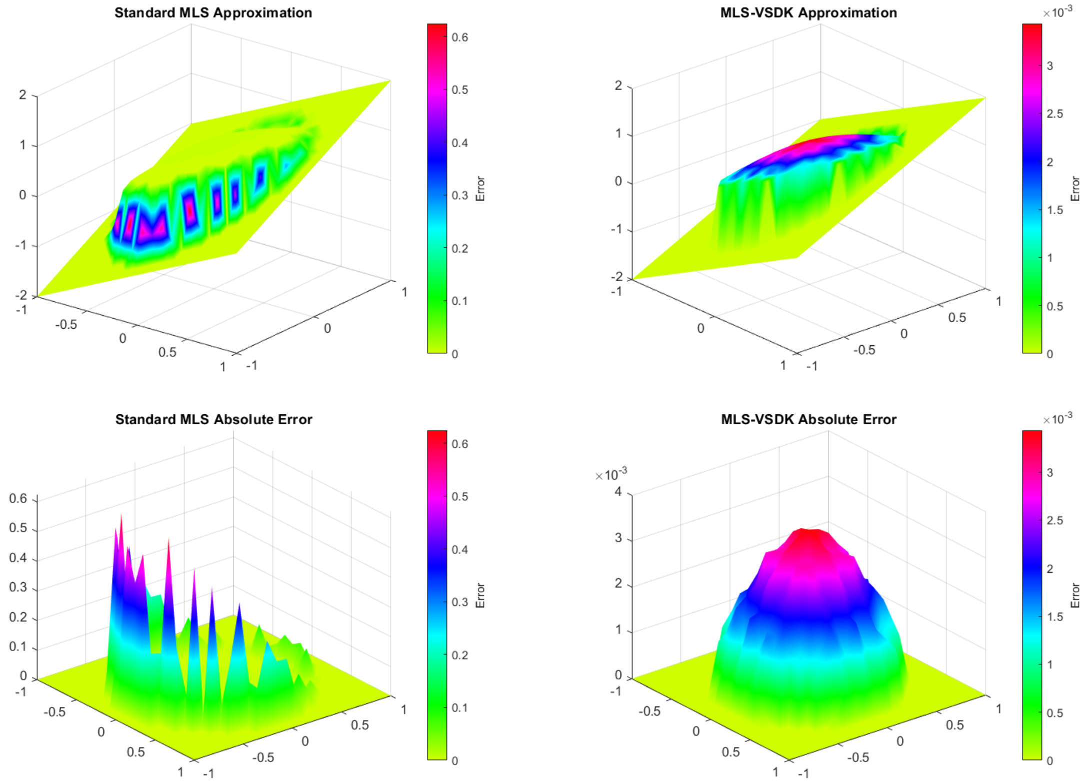


Figure 1: RMSE and abs-error of f_2 MLS (left) and MLS-VSDK (right) aproximation schemes using w^1 weight function.

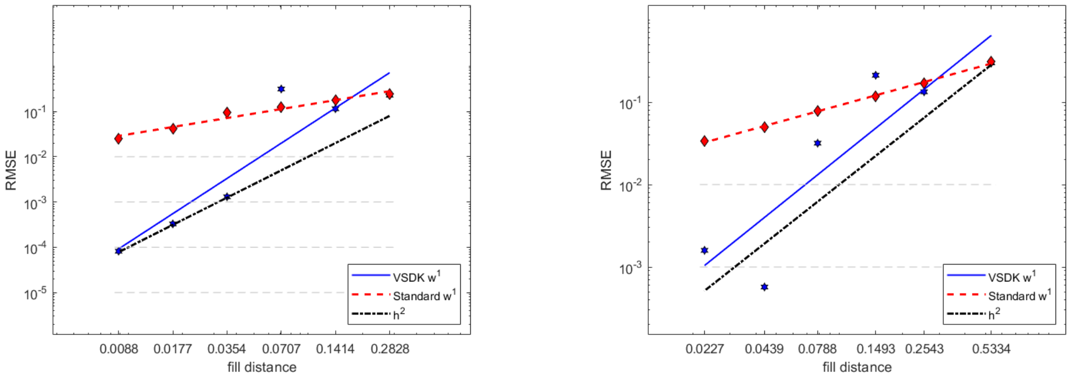


Figure 2: Convergence rates for approximation of function f_2 with MLS-VSDK and MLS standard schemes using *Uniform* data sites (left) and *Halton* data sites (right).

4.2 Example 2

Consider the following function

$$f_3(x,y) = \begin{cases} 2\big(1-\exp(-(y+0.5)^2)\big), & |x|\leq 0.5, \ , |y|\leq 0.5. \\ 4(x+0.8), & -0.8 \leq x \leq -0.65, |y|\leq 0.8. \\ 0.5, & 0.65 \leq x \leq 0.8, |y|\leq 0.2 \\ 0, & \text{otherwise.} \end{cases}$$

defined on $\Omega = (-1,1)^2$. Regarding the discontinuities of f_3, the scale function is considered to be

$$\psi(x,y) = \begin{cases} 1, & |x|\leq 0.5, \ , |y|\leq 0.5. \\ 2, & -0.8 \leq x \leq -0.65, |y|\leq 0.8. \\ 3, & 0.65 \leq x \leq 0.8, |y|\leq 0.2 \\ 0, & \text{otherwise.} \end{cases}$$

Moreover, let the centers and evaluation points be the same as the Example 1. Table 1 and 2 shows RMSE of MLS-VSDK and conventional MLS approximation of f_3 using w^4 which interpolates the data. We underline that our experiments show that the stencil of size $n = 20$ leads to the best accuracy.

number of centers	ε **value**	**RMSE MLS-VSDK**	**RMSE classic MLS**
25	1	3.67e-1	1.47e+0
81	2	3.68e-1	8.86e-1
289	4	1.49e-2	7.44e-1
1089	8	4.23e-3	7.72e-1
4225	16	1.06e-3	6.64e-1
16641	32	2.65e-4	5.25e-1

Table 1: RMSE of f_3 interpolation with *uniform* data sites.

number of centers	ε **value**	**RMSE MLS-VSDK**	**RMSE classic MLS**
25	1	8.84e-1	1.53e+0
81	2	8.95e-2	1.05e+0
289	4	1.42e-2	8.74e-1
1089	8	4.18e-3	6.48e-1
4225	16	1.09e-3	6.68e-1
16641	32	3.02e-4	7.07e-1

Table 2: RMSE of f_3 interpolation with *Halton* data sites.

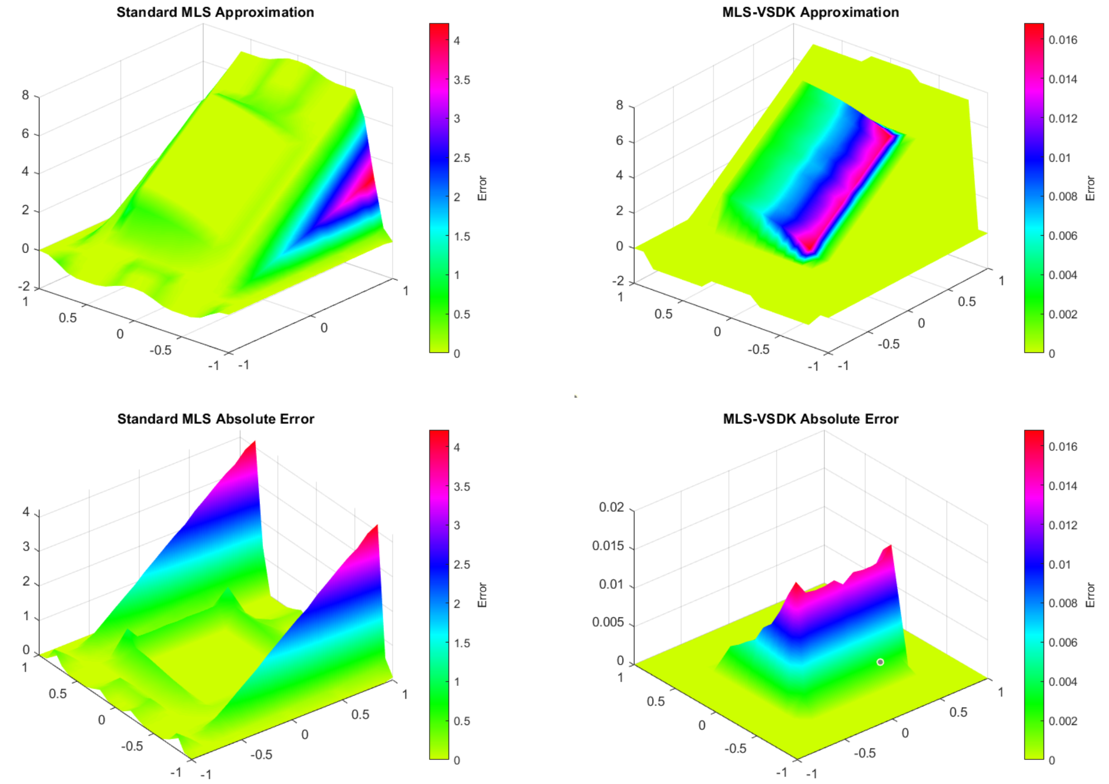


Figure 3: RMSE and abs-error of f_3 MLS(left) and MLS-VSDK(right) aproximation(interpolation) schemes using w^4 weight function.

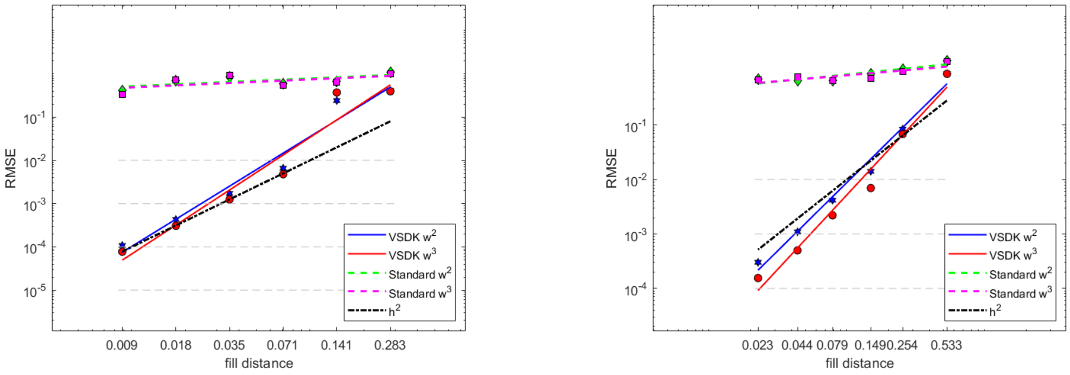


Figure 4: Convergence rates for approximation of function f_3 with MLS-VSDK and MLS standard schemes using *Uniform* data sites (left) and *Halton* data sites (right).

Figure 3 shows **RMSE** and **Absolute Error** for *standard MLS* and *MLS-VSDK approximation* of f_3 sampled from 1089 uniform points using w_4 as weight function. Once again, Figure 3 shows how MLS-VSDK scheme can improve the accuracy by reducing the error near the jumps.

Eventually, letting $\boldsymbol{\varepsilon}^U_{GA} = [2, 4, 8, 16, 32, 64]$ and $\boldsymbol{\varepsilon}^U_{Mat} = [10, 20, 40, 80, 160, 320]$, Figure 4 shows that h^2 convergence is achievable. To be more precise, the rate of convergence in the left plot is 2.54 and 2.69 for w_2 and w_3, respectively. On the other hand, letting $\boldsymbol{\varepsilon}^H_{GA} = [1, 2, 4, 8, 16, 32]$ and $\boldsymbol{\varepsilon}^H_{Mat}$ as the Uniform case, convergence rates of 2.50 and 2.73 is achievable when *Halton* data sites are employed.

4.3 Example 3

In applications, the discontinuities are likely to be unknown. To overcome this problem, one can consider edge detector method to extract the discontinuities. However, in this way the approximation depends also on the performance of the edge detector method as well [11]. In this direction, in this final experiment the location of the discontinuities are not exact. This is modeled by adding some noise drawn from the standard normal distribution multiplied by 0.01 to the edges of $\Omega_i \in \mathcal{P}$. We take the test function f_2 and the data sites in Section 4.1. We fix $n = 25$, and $\boldsymbol{\varepsilon}_{GA} = [0.25, 0.5, 1, 2, 4, 8]$, $\boldsymbol{\varepsilon}_{Mat} = [1, 2, 4, 816, 32]$ for both *Halton* and *uniform* centers. Figure 5 shows that the suggested MLS-VSDK is still able to obtain a good convergence rate when compared to classical MLS even when the discontinuities are nor known exactly.

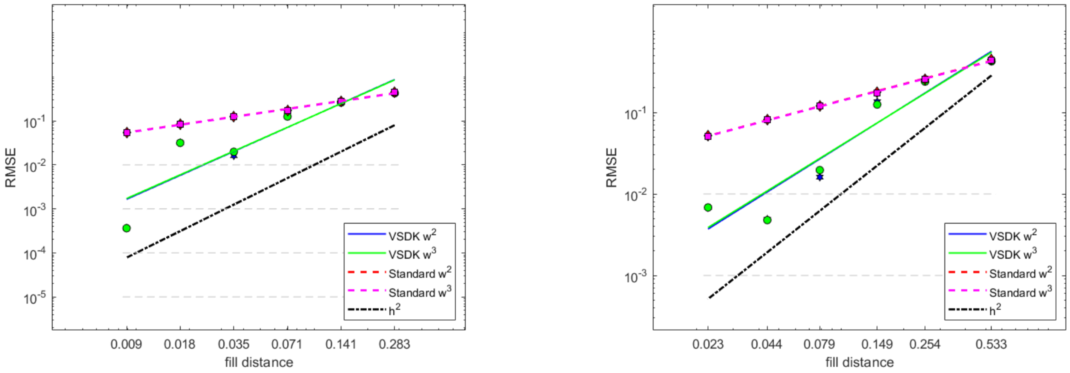


Figure 5: Convergence rates for approximation of function f_2, based on noisy given data values, with MLS-VSDK and MLS standard schemes using *Uniform* data sites (left) and *Halton* data sites (right).

5 Conclusions

To approximate a discontinuous function using scattered data values, we studied a new technique based on the use of discontinuously scaled weight functions, that we called the MLS-VSDK scheme, that is the application of discontinuous scaled weight functions to the MLS. It enabled us to move toward a data-dependent scheme, meaning that MLS-VSDK is able to encode the behavior of the underlying function. We obtained a theoretical Sobolev-type error estimate which justifies why MLS-VSDK can outperform conventional MLS. The numerical experiments confirmed the theoretical convergence rates. Besides, our numerical tests showed that the suggested scheme can reach high accuracy even if the position of the data values are slightly perturbed.

Acknowledgments. This research has been accomplished within the Rete ITaliana di Approssimazione (RITA) and the thematic group on Approximation Theory and Applications of the Italian Mathematical Union. We also received the support of GNCS-INδAM.

References

[1] Lucy, L.B., *A numerical approach to the testing of the fission hypothesis.* The Astronomical Journal 82 (1977), 1013–1024.

[2] Nguyen, V.P. et al., *Meshless methods: a review and computer implementation aspects.* Math. Comput. Simul. 79/3 (2008), 763–813.

[3] Shepard, D., "A two-dimensional interpolation function for irregularly-spaced data". Proceedings of the 1968 23rd ACM national conference, New York, U.S.A., 27-29 August 1968.

[4] Lancaster, P.; Salkuaskas, K., *Surfaces generated by moving least squares methods.* Math. Comput. 37 (1981), 141–158.

[5] Wendland, H., "Scattered Data Approximation (1st edition)". Cambridge University Press: Cambridge, UK, 2005; p. 336..

[6] Levin, D., *The approximation power of moving least-squares.* Math. Comp. 67 (1998), 1517–1531.

[7] Mirzaei, D., *Analysis of Moving Least Square Approximation revisited.* J. Comput. Appl. Math. 282 (2015), 237–250.

[8] Narcowich, F.J; Ward, J.D.; Wendland. H., *Sobolev Bounds On Functions With Scattered Zeros, With Applications To Radial Basis Function Surface Fitting.* Math. Comput. 78 (2005), 743–763.

[9] De Marchi, S.; Marchetti, F.; Perracchione, E., *Jumping with Variably Scaled Discontinuous Kernels.* BIT Numer. Math. 60 (2019), 441–463.

[10] Bozzini, M.; Lenarduzzi, L.; Rossini, M.; Schaback, R., *Interpolation with Variably Scaled Kernels.* SIAM J. Numer. Anal. 35 (2015), 199–219.

[11] De Marchi, S.; Erb. W.; Marchetti, F.; Perracchione, E.; Rossini, M., *Shape-Driven Interpolation with discontinuous Kernels: Error Analysis, Edges Extraction and Application in Magnetic Particle Imaging.* J. Sci. Comput. 42 (2020), 472–491.

[12] Fasshauer, G.E., "Meshfree Approximation Methods (1st edition)". World Scientific Publishing: Singapore, 2007; p. 500.

[13] Adams R.A.; Fournier, J., "Sobolev Spaces (2nd edition)". Elsevier: London, U.K, 2003; p. 305.

[14] C. Rieger, B. Zwicknagl, *Sampling inequalities for infinitely smooth functions, with applications to interpolation and machine learning.* Advances in Computational Mathematics 32.1 (2010), 103–129.

[15] Bayona, V., *Comparison of Moving Least Squares and RBF+poly for Interpolation and Derivative Approximation.* J. Sci. Comput. 81 (2019), 486–512.

[16] Bernard, S.C.; Scott, L.R., "The Mathematical Theory of Finite Element Methods (3rd edition)". Springer: 2003; p. 397.

[17] Fasshauer, G.E.; McCourt, M.J., "Kernel Based Approximation Methods Using MATLAB (1st edition)". World Scientific Publishing: Singapore, 2015; p. 537.

Binomial coefficients in modular arithmetic

A mathematical solution to musical questions

Riccardo Gilblas [(*)]

Abstract. In this seminar we are linking modular binomial coefficients to periodic sequences with modular integer values. In the context of serialism, the romanian composer Anatol Vieru used such sequences to compose several musical pieces. We will introduce the main properties of periodic sequences and we will explain the link between them and the binomial coefficients in $\mathbb{Z}/m\mathbb{Z}$. This allows to use tools such as Kummer's Theorem and the generalisation of Lucas's Theorem to answer some questions arisen from Vieru's observations.

Musical motivation

In his *Book of Modes*, the romanian composer Anatol Vieru (1926–1998) collects periodic sequences by iteratively applying a finite sum operator to a periodic sequence with values in $\mathbb{Z}_{12}$.

He assigns a musical meaning to each obtained sequence and uses the obtained values to compose music. So for example a sequence encodes the theme of a certain instrument, another sequence encodes the rythm, etc. In [12] he explores some mathematical properties of periodic sequences and the finite sum operator, observing some peculiarities: repeatedly applying the sum operator increases the period of the sequences and often some values proliferate among the coefficients.

These observations were developed in several math-music research article, such as [13], [1], [2], [3], [8]. In these works, the authors faced the questions with a computational approach and obtained some statistical results.

Thanks to a new formalisation of the questions and a clearer link between such periodic sequences and binomial coefficients, we managed to completely solve the problems using purely algebraic tools. More, we obtained a useful formula to compute the p-adic valuation of a sequence of binomials.

Here we present a quick introduction to modular periodic sequences, pointing out their main properties. Then we introduce the finite sum operator, which is the main focus of our work. Finally, we give the main ideas of our solution of Vieru's problems.

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: `riccardo.gilblas@math.unipd.it` . Seminar held on 14 December 2022.

1 Periodic sequences and musical interpretation

We fix the following notation:

- The $\mathbb{Z}_m$-module $S_m := \mathbb{Z}_m^{\mathbb{N}}$ of sequences with values in $\mathbb{Z}_m$ for a given positive integer m.

- The shift endomorphism θ of S_m acting on $f \in S_m$ as:
$$\theta(f)(n) = f(n+1) \quad \forall n \in \mathbb{N}.$$

- A sequence $f \in S_m$ is *periodic* if there exists $j \geq 1$ such that $\theta^j(f) = f$. The minimal j with this property is *period* of f and it is denoted $\tau(f)$.

- The $\mathbb{Z}_m$-module $P_m \subset S_m$ of all periodic sequences. Formally:
$$P_m = \bigcup_{j \geq 1} \ker(\theta^j - \mathrm{id}).$$

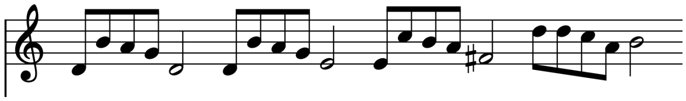
Example 1.1 Consider the following sequences:

$$\begin{aligned} j =& [2, 11, 9, 7, 2, 2, 11, 9, 7, 4, 4, 0, 11, 9, 6, 2, 2, 0, 9, 11] \in P_{12} \\ r =& [1, 1, 1, 1, 4] \in P_5. \end{aligned}$$

With the correspondence:

$$\begin{aligned} \mathbb{Z}_{12} \longleftrightarrow & \{C, C\sharp, D, D\sharp, E, F, F\sharp, G, G\sharp, A, A\sharp, B\} \\ \mathbb{Z}_5 \longleftrightarrow & \{0, \text{♪}, \text{♩}, \text{♩.}, \text{𝅗𝅥}\} \end{aligned}$$

we obtain the Jingle Bells theme:



Definition 1.2 Consider on P_m the *difference operator* $\Delta := \theta - \mathrm{id}$ and take $f \in P_m$.

- f is *nilpotent* if there exists $\eta \geq 1$ such that $\Delta^\eta f = 0$. The minimal η with this property is called the *nilpotency index* of f.

- f is *idempotent* if there exists $\eta \geq 1$ such that $\Delta^\eta f = f$. The minimal η with this property is called the *idempotency index* of f.

Remark 1.3 The operator Δ is the discrete analogous of the usual derivation of derivable functions. For example, $\Delta(f) = 0$ if and only if f is a constant sequence. Nilpotent sequences could correspond to polynomials, with the nilpotency index behaving similarly to the degree of polynomials. Idempotent sequences correspond to functions like e^x, which is idempotent of index 1.

Definition 1.4 We define the *sum operator* Σ as follows: for every periodic sequence $f \in P_m$,

$$\Sigma f(n) := \begin{cases} 0 \text{ if } n = 1 \\ f(n-1) + \Sigma f(n-1) \text{ if } n > 1. \end{cases}$$

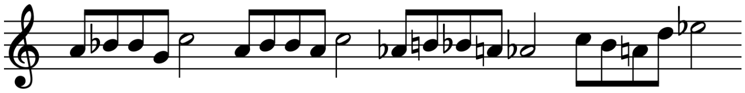
Σf is automatically periodic and it is called the *primitive* of the sequence f.

In analogy with results from differential calculus: $\Delta g = \Delta\Sigma f$ if and only if $g = \Sigma f + [c]$.

Example 1.5 If j is the sequence from the previous example that describes the pitches of Jingle Bells, its derivative is:

$$\Delta j = [9, 10, 10, 7, 0, 9, 10, 10, 9, 0, 8, 11, 10, 9, 8, 0, 10, 9, 2, 3]$$

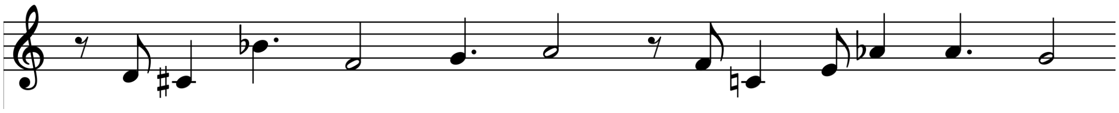
Now if we take this sequence as the pitches and we keep the original rythm of Jingle Bells (in the previous example was r), we get:



Now if we apply Σ to both j and r we get:

$$\Sigma j = [0, 2, 1, 10, 5, 7, 9, 8, 5, 0, 4, 8, 8, 7, \dots]$$
$$\Sigma r = [0, 1, 2, 3, 4, 3, 4, 0, 1, 2, 1, 2, 3, 4, 0, 4, 0, 1, 2, 3, 2, 3, 4, 0, 1]$$

where Σj has period equal to 120. Now if we use Σj for the pitches and Σr for the rythm, we get:



2 Decompositions

We show how it is possible to reduce to study periodic sequences on $\mathbb{Z}_{p^\ell}$, so with coefficients in the integers modulo a power of a prime. Then we show that every periodic sequence uniquely decomposes as a sum of a nilpotent sequence and an idempotent sequence.

We fix $\mathbb{N} \ni m = \prod p_i^{\ell_i}$. The group isomorphism $\mathbb{Z}_m \to \bigoplus \mathbb{Z}_{p_i^{\ell_i}}$ gives rise to an isomorphism of abelian groups

$$\begin{aligned} P_m &\xrightarrow{\sim} \bigoplus P_{p_i^{\ell_i}} \\ f &\longmapsto (f_{p_i})_{1\leq i\leq t} \end{aligned}$$

where $f_{p_i}(j) = f(j) \mod p_i^{\ell_i}$. One has $\tau(f) = \mathrm{lcm}\{\tau(f_{p_i})\}_{1\leq i\leq t}$.

Lemma 2.1 *We can uniquely write $f \in P_m$ as a sum of an idempotent sequence f_I and a nilpotent sequence f_N. Thus*

$$P_m = \bigoplus_{i=1}^{t} I^{\Delta}_{p_i^{\ell_i}} \oplus N^{\Delta}_{p_i^{\ell_i}} \qquad I^{\Delta}_m = \bigoplus_{i=1}^{t} I^{\Delta}_{p_i^{\ell_i}} \qquad N^{\Delta}_m = \bigoplus_{i=1}^{t} N^{\Delta}_{p_i^{\ell_i}}$$

where N^{Δ}_m (resp. I^{Δ}_m) denote the submodule of P_m of nilpotent (resp. idempotent) sequences. Moreover, $\tau(f) = \mathrm{lcm}\{\tau(f_I), \tau(f_N)\}$.

Hence we can study separately nilpotent sequences and idempotent sequences. Also, in the following we will study just the case of sequences with coefficients in $\mathbb{Z}_{p^\ell}$. Regarding nilpotent sequences, the following theorem provides a useful criterion:

Theorem 22 (D.T. Vuza) *Let $f \in P_{p^\ell}$ be a periodic sequence. Then:*

- *$f \in N^{\Delta}_{p^\ell}$ if and only if $\tau(f) = p^t$ for $t \in \mathbb{N}$;*
- *if $f \in N^{\Delta}_{p^\ell}$ with period p^t and nilpotency index η, then $\eta \leq \ell p^t$.*

For idempotent sequences, we have a less powerful result:

Corollary 2.3 *For any integer t, consider the $\mathbb{Z}_{p^\ell}$-module $I^t_{p^\ell}$ of idempotent sequences with period dividing t. Then:*

- *$I^t_{p^\ell} = 0$ if $t = p^u$.*
- *$\mathrm{rk}(I^t_{p^\ell}) = t - 1$ if t is prime to p.*
- *$\mathrm{rk}(I^t_{p^\ell}) = p^u(q-1)$ if $t = p^u q$.*

Example 2.4 We provide an example of decomposition using the sequences used previously to describe the rythm and the theme of Jingle Bells. The rythm corresponds to the sequence $r = [1, 1, 1, 1, 4] \in P_5$. It is nilpotent of index 5 and has constant decomposition:

$$r = [1] + \Sigma^4[3].$$

The theme corresponds to the sequence

$$j = [2, 11, 9, 7, 2, 2, 11, 9, 7, 4, 4, 0, 11, 9, 6, 2, 2, 0, 9, 11] \in P_{12}.$$

It decompose as:

$$\begin{aligned} j_2 =& [2,3,1,3,2,2,3,1,3,0,0,0,3,1,2,2,2,0,1,3] \in P_4 \\ j_3 =& [2,2,0,1,2,2,2,0,1,1,1,0,2,0,0,2,2,0,0,2] \in P_3. \end{aligned}$$

The sequence j_3 has period 20 (prime to 3), thus it is idempotent up to a constant (in fact $j_3 + [1]$ is idempotent of index 80).

The sequence j_2 has decomposition:

$$\begin{aligned} (j_2)_N =& [0,2,3,1,0,2,3,1,0,2,3,1,0,2,3,1,0,2,3,1] \\ (j_2)_I =& [2,1,2,2,2,0,0,0,3,2,1,3,3,3,3,3,1,2,2,2,2] \text{ of index 120.} \end{aligned}$$

Reducing to primitives of constant sequences

The following lemmas show how it is possible to reduce the study of primitives of a sequence to constant sequences and their primitives. For nilpontent sequences, one has:

Lemma 2.5 *Sequences in N_m^{Δ} are finite sums of primitives of constant sequences.*

Hence if we want to study the primitives of a nilpotent sequence, it is enough to look at the primitives of the constants in its decomposition. A similar result can be obtained, surprisingly, also for idempotent sequences:

Lemma 2.6 *Consider $f \in I_m^{\Delta}$ of index η and for every $0 \leq i < \eta$ denote $\delta^i = \Delta^i f(0)$ the first coefficient of the sequence $\Delta^i f$. Then for every $s < \eta$ one has:*

$$\Sigma^s f = \Delta^{\eta-s} f + \sum_{i=1}^{s} \Sigma^{s-i} \delta^{\eta-i}.$$

This roughly says that the primitives of an idempotent sequence f have an idempotent part that corresponds to a suitable derivative of f, and a nilpotent part which is given by a sum of primitives of a fixed number of constants. The period of the idempotent part remains equal to the period of f, while for the nilpotent part we need to study the primitives of constant sequences.

3 Constant sequences and their primitives

The following lemma finally introduces the link with binomial coefficients modulo an integer:

Lemma 3.1 *If $[c]$ is a constant sequence in P_m, then*

$$\Sigma^s[c](n) \equiv_m c\binom{n}{s}.$$

This means that studying constant sequences and their primitives is equivalent to study modular binomial coefficients. In particular, the starting problem (of studying the period and the values of primitives of periodic sequences) can be rephrased in terms of binomial coefficients: fixed a positive integer s, we are to study the periodicity and the values of the function:

$$\mathbb{N} \longrightarrow \mathbb{Z}_m\,, \qquad n \longmapsto \binom{n}{s} \mod m.$$

This perspective suggested to resort to two main results: Kummer's Theorem ([7]) and the generalised version of Lucas's Theorem ([4]). Kummer's Theorem states that the p-adic valuation of the binomial $\binom{n}{s}$ is equal to the number of borrows in performing the operation $n - s$ in base p. The generalisation of Lucas's Theorem provides an explicit formula to compute $\binom{n}{s} \mod p^\ell$. Both theorems require to consider the expression of n and s in base p; we will denote it by $n = \lfloor a_k \cdots a_0 \rfloor_p$ with $a_k \neq 0$.

Using Kummer's Theorem, we provided a new proof of the main theorem ([9]) on the period of constant sequences:

Theorem 3.2 *Let $[c]$ be a non zero constant sequence in P_{p^ℓ} with $\nu_p(c) = t < \ell$. Let $s \in \mathbb{N}$ and $\lfloor a_k a_{k-1} \cdots a_1 a_0 \rfloor_p$. Then the sequence $\Sigma^s[c]$ has period $p^{\ell - t + k}$.*

In binomial terms, this statement corresponds to the function $n \to \binom{n}{s}$ on $\mathbb{Z}_{p^\ell}$ being periodic of period $p^{\ell - t + k}$.

Hence the period of the s-th primitive of $[c]$ depends on the p-adic valuation $\nu_p(c)$ and on $\lfloor \log_p s \rfloor$, the floor of the p-base logarithm of s.

The period of primitives of generic sequences.

Given the decomposition in sum of primitives of constants for both idempotent and nilpotent sequences, finding a formula for the period of primitives of generic sequences is equivalent to study how the sum of sequences affects the period. Unfortunately, this is not trivial: for example, the period of the sum is not necessarily equal to the least common multiple of the periods. This remains false even if we restrict to primitives of constants: the sequences $\Sigma[2]$ and $\Sigma^2[4]$ in P_8 have period 4 while their sum is $[0, 2]$ which has period 2.

Nonetheless, we were able to prove results that hold *definitively*:

Theorem 3.3 *Given $f \in N^{\Delta}_{p^n}$, there is a constant c in the decomposition of f such that $\tau(\Sigma^s f) = \tau(\Sigma^s[c])$ for $s \gg 0$.*

Similarly for idempotent sequences:

Theorem 3.4 *Consider $f \in I^{\Delta}_{p^\ell}$ with idempotency index η. Then there exist a constant ε and a natural number $1 \leq w \leq \eta$ such that $\tau(\Sigma^s f) = \operatorname{lcm}(\tau(f), \tau(\Sigma^{s-w}[\varepsilon]))$ for $s \gg 0$.*

With these results we are able to completely describe the period of (definitive) primitives of a generic sequence.

4 Proliferation of values

The next focus of our interest is the proliferation of certain values among the primitives of certain sequences. Vieru's observation regarded the number of zeros in the primitives of the sequence $V_2 = [2, 1, 2, 0, 0, 1, 0, 0] \in P_4$: this number (over the period) increases when taking higher index primitives, reaching for some indices the 98% of the period. From the musical point of view, this corresponds to a theme that gradually (i.e. with repeated integration) becomes trivial. We provide here a brief idea of how we were able to explain this behaviour and the recursive formula on modular binomials that we came up with.

Given a sequence $f \in P_{p^\ell}$, we denote:

- $Z(f^s)$: the number of zeroes inside the period of $\Sigma^s f$;
- $\Pi_i(f^s)$: the number of coefficients inside the period of $\Sigma^s f$ with p-adic valuation $0 \leq i < \ell$.

The recursive formula for the constant sequence $f = [1] \in P_{p^\ell}$ is structured in the following way: given $2^k \leq s < 2^{k+1}$, we consider the expression of s in base p: $\lfloor a_k \cdots a_0 \rfloor_p$. If s has some peculiarities, we are able to link the numbers $Z(f^s), \Pi_i(f^s)$ to the numbers $Z(f^{s'}), \Pi_i(f^{s'})$ where $s' = \lfloor a_k \cdots \hat{a}_i \cdots a_0 \rfloor_p$ is obtained by removing the coefficient a_i for a suitable $0 < i < k$.

More precisely, we proved the following lemmas:

Lemma 4.1 *With the notation above, suppose that $k > \ell, 0 \leq m \leq k - \ell - 1$ and:*

$$\begin{aligned} s =& \lfloor b_k \cdots b_{k-m} \underbrace{(p-1) \cdots (p-1)}_{\ell} b_{k-m-\ell-1} \cdots b_0 \rfloor_p \\ s' :=& \lfloor b_k \cdots b_{k-m} \underbrace{(p-1) \cdots (p-1)}_{\ell-1} b_{k-m-\ell-1} \cdots b_0 \rfloor_p . \end{aligned}$$

Then $\Pi_i(f^s) = \Pi_i(f^{s'})$ and $Z(f^s) = Z(f^{s'}) + (p-1)p^{k+\ell-1}$.

Lemma 4.2 *With the notation above, suppose that $k > \ell, 0 \leq m \leq k - \ell - 1$ and the expression of s in base p is:*

$$\begin{aligned} s =& \lfloor b_k \cdots b_{k-m} \underbrace{0 \cdots 0}_{\ell} b_{k-m-\ell-1} \cdots b_0 \rfloor_p \\ s' :=& \lfloor b_k \cdots b_{k-m} \underbrace{0 \cdots 0}_{\ell-1} b_{k-m-\ell-1} \cdots b_0 \rfloor_p . \end{aligned}$$

Then $\Pi_i(f^s) = p \cdot \Pi_i(f^{s'})$ and $Z(f^s) = p \cdot Z(f^{s'})$.

Lemma 4.3 *With the notation above, suppose that $k > \ell, 0 \leq m \leq k - \ell - 1$ and the* *expression of s in base p is:*

$$\begin{aligned} s =& \lfloor b_k \cdots b_{k-m} \underbrace{(p-1)\; 0 \cdots 0}_{\ell}\; b_{k-m-\ell-1} \cdots b_0 \rfloor_p \\ s' :=& \lfloor b_k \cdots b_{k-m} \underbrace{(p-1)\; 0 \cdots 0}_{\ell-1}\; b_{k-m-\ell-1} \cdots b_0 \rfloor_p . \end{aligned}$$

Then if $\gamma_s := p^{\ell-1}(p-1)b_{k-m-\ell-1}\prod_{i=0}^{k-\ell-m-2}(p-b_i)\prod_{j=k-m}^{k}(p-b_j)$, *one has:*

$$\begin{aligned} \Pi_i(f^s) &= p \cdot \Pi_i(f^{s'}) \qquad 0 \le i \le \ell - 2 \\ \Pi_{\ell-1}(f^s) &= p \cdot \Pi_{\ell-1}(f^{s'}) + \gamma_s \\ Z(f^s) &= p \cdot Z(f^{s'}) - \gamma_s . \end{aligned}$$

These lemmas provide quantitative information on the zeros and the p-valuation of the coefficients of a primitive of a constant sequence. Notice that in the case of our initial study, i.e. when $p = 2$ and $\ell = 2$, for every $s \ge 8$ the hypotheses of at least one of the lemmas are satisfied, thus for sequences in P_4 it is always possible to use this recursive reduction and to trace back the numbers $Z(f^s)$ and $\Pi_i(f^s)$ for $0 \le i \le 1$ to the numbers $Z(f^j)$ and $\Pi_i(f^j)$ with $0 \le j < 8$.

To get the precise values of the coefficients, we can use the generalisation of Lucas's Theorem; this allows to compute binomial coefficients modulo a power of a prime. We did the explicit computation for the sequence $V_2 = [2,1,2,0,0,1,0,0] \in P_4$, the one that motivated Vieru's interest, to show the recursive peaks of numbers of zeros in its primitives.

Definition 4.4 For every $s \ge 3$, define $g^s := \Sigma^s V_2$ and denote by $z(s)$ the number of zeros among the coefficients in a period of g^s.

Then we proved the following:

Theorem 4.5 *For every $k \ge 3$ we have:*

$$2^k + 1 = z(2^k - 4) < z(2^k - 5) = 2^{k+1} - 8.$$

More precisely, one has:

$$g^{2^k-5} = [\underbrace{0,\ldots,0}_{2^k-5}, 2, 3, 1, \underbrace{0,\ldots,0}_{2^{k-1}-2}, 2, 2, \underbrace{0,\ldots,0}_{2^{k-1}-3}, 2, 1, 3, 0, 0].$$

Hence for any $k \ge 3$, the sequence $\Sigma^{2^k-5}V_2$ has $2^{k+1} - 8$ zeros among the coefficients of its period, which is equal to 2^{k+1}. So the ratio between the number of zeros and the period tends to 1:

$$\frac{z(2^k-5)}{\tau(\Sigma^{2^k-5}V_2)} = \frac{2^{k+1}-8}{2^{k+1}} \overset{k\to\infty}{\longrightarrow} 1.$$

This explains the computational results previously obtained ([13], [1], [2], [3], [8]).

References

[1] M. Andreatta, D.T. Vuza, *On some properties of periodic sequences in Anatol Vieru's modal theory.* Tatra Mountains Mathematical Publications 23 (2001), 1–15.

[2] M. Andreatta, C. Agon and D.T. Vuza, *Analyse et implémentation de certaines techniques compositionnelles chez Anatol Vieru.* Actes des Journées d'Informatique Musicale, Marseille (2002), 167–176.

[3] M. Andreatta, D.T. Vusa and C. Agon, *On some theoretical and computational aspects of Anatol Vieru's periodic sequences.* Soft Computing 8, no. 9 (2004), 588–596.

[4] K.S. Davis, W.A. Webb, *Lucas' theorem for prime powers.* Europ. J. Combinatorics 11 (1990), 229–233.

[5] L.E. Dickson, "History of the theory of numbers". Vol. 1, Chelsea, New York, 1952.

[6] N.J. Fine, *Binomial coefficients modulo a prime.* Am. Math. Monthly 54 (1947), 589–592.

[7] E. Kummer, *Über die Ergänzungssätze zu den allgemeinen Reciprocitätsgesetzen.* Journal für die reine und angewandte Mathematik 44 (1852), 93–146.

[8] P. Lanthier, C. Guichaoua and M. Andreatta, *Reinterpreting and Extending Anatol Vieru's Periodic Sequences Through the Cellular Automata Formalisms.* Proceedings MCM (2019), Springer, 261–272.

[9] C.J. Lu, S.C. Tsai, *The periodic property of binomial coefficients modulo m and its applications.* 10th SIAM Conference on Discrete Mathematics, Minneapolis, Minnesota, USA, 2000.

[10] C. Mariconda, A. Tonolo, *Discrete Calculus - Methods for Counting.* Springer (2016), 52–54.

[11] M.P. Saikia, J. Vogrinc, *Binomial symbols and prime moduli.* J. Indian Math. Soc. (N.S.) 78 (2011), no. 1-4, 137–143.

[12] A. Vieru, "The Book of Modes". Editura Muzicala, Bucharest, 1993.

[13] D.T. Vuza, "Aspects mathématiques dans la théorie modale d'Anatol Vieru". Editura Academiei Republicii Socialiste România (1982).

[14] Ś. Ząbek, *Sur la périodicité modulo m des suites de nombres $\binom{n}{k}$.* Ann. Univ. Marie Curie-Skłodowska, Sect. A, 10 (1956), 37-47 (1958).

Complex Networks: a highly interdisciplinary field. Theory and Applications

Sara Venturini [(*)]

Abstract. Networks and graph models have become a nearly ubiquitous abstraction and an extremely useful tool to represent a variety of real systems in different fields. They can help us to better understand and analyze different types of interactions and dynamics. Recent researches have shown that real world interactions, in many cases, cannot be fully described by standard graphs. Therefore, there is the need to study more complex structures such as multilayer networks, which enable to take into account different types of information, as well as simplicial complexes and hypergraphs, which consider group interactions. We will give a brief introduction to modern mathematical and computational tools for complex networks, their applications, and their extension to multilayer and hypergraphs. In particular, we will focus on the multiplex community detection problem and the application to the science of science.

1 Introduction

The scientific study of networks, such as computer networks, biological networks, and social networks, is an interdisciplinary field that combines ideas from mathematics, physics, biology, computer science, statistics, the social sciences, and many other areas [1].

A network (or graph) is a collection of points (nodes) joined together in pairs by lines (edges). It can be fully described by an adjacency matrix:

$$A_{\mathrm{ij}} = \begin{cases} 1 & \text{if there is an edge between nodes i and j} \\ 0 & \text{otherwise.} \end{cases}$$

However, in real-world applications, we need to describe interactions in detailed and varied ways, like for example directed edges to describe the direction of a message or edge weights to consider the intensity of an interaction.

Some examples of area of applications of complex networks are:

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: sara.venturini.3@studenti.unipd.it . Seminar held on 18 January 2023.

- Technological networks, like the Internet, i.e., the computer data network in which the nodes are computers, and the edges are data connections between them, the telephone network, the transportation networks like networks of roads, rail lines, airline routes. Used for instance to better understand the flow of data traffic.
- Information networks, like the World Wide Web, where the nodes are web pages and the edges are the links between them, the citation network between academic journal articles.
- Social networks, like Facebook or Twitter, where the nodes are people and the edges between them are social connections of some kind like friendship, communication, collaboration. Used for instance to better understand the nature of social interactions, the spread of disease, the structure of society, the spread of disinformation.
- Biological networks, like neural networks, connection between neurons in the brain or networks of macroscopic functional connectivity between large-scale regions of the brain where nodes are entire brain regions that are already known to perform some function such as vision, motor control or learning and memory, ecological networks, where the nodes are species in an ecosystem and the edges represent predator-prey relationships, biochemical networks like metabolic networks, protein-protein interaction networks, genetic regulatory networks. Used for instance to better understand the complex chemical processing in the cell and perhaps even to new therapies for disease or injury.

Networks capture the pattern of interactions between the parts of a system, which can have a big effect on the behavior of the system. There are measures and metrics for quantify network structures, for instance:

- Centrality measures, which quantify how important nodes are in a network.
- Percolation and network resilience, which leaf to a theory of the robustness of networked systems to the failure of their components.
- Epidemics on networks, among which the most popular study of social networks is their connection to the spread of disease.
- Community structure, since a frequent network phenomenon is the occurrence of communities or clusters in networks.

2 The science of science

Among the many applications of complex networks, we focus a quite recent but really promising one: the science of science. The increasing availability of digital data on scholarly output offers unprecedented opportunities to explore patterns characterizing the structure and evolution of science. Science can be described as a complex evolving multiscale network. The science of science aims to find patterns characterizing the structure and evolution of science, with the ultimate goal to accelerate science [2]. Some branches of research are regarding:

- the evolution of fields, modeling how disciplines arise, evolve, disappear [3];
- the gender disparities in science [4];
- the faculty hiring process [5];
- the proposal of new valuation indices [6];
- the interactions between collaborators [7].

Collaboration is a key driver of science and innovation. Mainly motivated by the need to leverage different capacities and expertise to solve a scientific problem, collaboration is also an excellent source of information about the future behavior of scholars. In particular, it allows us to infer the likelihood that scientists choose future research directions via the intertwined mechanisms of selection and social influence. In our paper [8], we thoroughly investigate the interplay between collaboration and topic switches. We find that the probability for a scholar to start working on a new topic increases with the number of previous collaborators, with a pattern showing that the effects of individual collaborators are not independent. The higher the productivity and the impact of authors, the more likely their coworkers will start working on new topics. The average number of coauthors per paper is also inversely related to the topic switch probability, suggesting a dilution of this effect as the number of collaborators increases.

3 Higher order interactions

Networks have become a nearly ubiquitous abstraction and an extremely useful tool to represent a wide variety of real systems in different fields. They can help us to better understand and analyze different type of interactions and dynamics. On the structural side, recent researches have shown that real world interactions, in many cases, cannot be fully described by just standard (single-layer) graphs. Therefore, there is the need to study more complex structures such as multilayer networks, which allow to take into account different types of information [9], simplicial complexes and hypergraphs, which take into account group interactions [10]. In practice, multilayer networks can be described by a tensor, containing the adjacency matrices of each layer and the interactions between nodes of different layers; hypergraphs can be described by an incident matrix, explaining the nodes belonging to each hyperedge.

4 Community detection

In the last century, networks have become extremely useful as representation of a wide variety of real systems in different fields, like sociology, biology, technology, and so on. One of the most important aspects is community structure, or clustering, that is the organization of vertices in clusters, with more edges connecting vertices of the same group and fewer edges linking vertices of different groups. Social communities have been studied for a long time, but they also occur in networks from other fields. For example, in biology, in protein interaction networks, communities correspond to proteins with the same function within the cell, or in the World Wide Web they may represent pages dealing with the same topics.

The first problem in community detection is to look for a quantitative definition of community. No definition is universally accepted. From intuition we get the notion that there must be more edges "inside" the community than edges linking vertices of the community with the rest of the graph. The goal of community detection is to find a partition of the graph, that is a division of the graph in clusters. Therefore, we need a quantitative criterion to measure the goodness of a graph partition. A quality function is a function that assigns a number to each partition of a graph, so we can then rank them. The most popular quality function is the modularity of Newman and Girvan, which implicitly defines communities in based on the observation that random networks are not expected to exhibit a modular structure. Modularity can be written as follows:

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - P_{ij})\delta(C_i, C_j) \tag{1}$$

with m total number of edges, A adjacency matrix, P_{ij} number of edges between nodes i and j in the null model, C_i community of node i, δ Kronecker delta. A null model is used as term of comparison, to verify if a graph shows a community structure. There are several possibilities to choose a null model, but select one with the same degree distribution of the original graph. So, the final expression of modularity is

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j) \tag{2}$$

with k_i node i degree. Large positive values of modularity indicate good partitions.

Recent researches have shown that the interconnected world is composed of networks that are coupled to one another through different layers, where each layer represents one of many types of interactions. Therefore, analysing multi-layer networks is of great importance because interesting patterns cannot be obtained by just analysing single-layer networks.

Many community detection approaches have been proposed for single-layer graphs. Representative algorithms include graph partitioning algorithms, which divide the nodes such that the cut size is minimal modularity-based algorithms, which find partitions with maximum modularity, spectral algorithms, which use the eigenvectors of graph matrices, such as the Laplacian one, and finally structure definition algorithms, which discover communities with strict properties, such as they find k-cliques, r-quasi cliques or s-plex.

New challenges arise for community detection in multi-layer graphs. It is natural to detect multi-layer communities by extending the algorithms for single-layer community detection. We can employ two strategies. The first one reduces the multi-layer networks into a single-layer network and then applies single-layer network algorithms to this graph. Meanwhile the second strategy applies single-layer network algorithms to each layer, and then combines the communities using consensus clustering. However, these algorithms are criticized for their low accuracy because they ignore the connection among various layers. To overcome these problems, algorithms must simultaneously take into account multiple layers.

In our paper [11], we propose a Multiobjective Louvain-like Method for Multiplex. The multiobjective optimization problem is defined as:

$$\max_{\{\text{partitions of } V\}} (Q_1, \dots, Q_k)$$

There is not a unique way to define optimality, since there is no a-priori total order for $\mathbb{R}^k$. Therefore, we use the following definition:

Definition 1 Given two vectors $z^1, z^2 \in \mathbb{R}^k$, we write $z^1 \succeq_P z^2$ if z^1 dominates z^2 according to Pareto, that is:

$$\begin{aligned} z_i^1 &\geq z_i^2 \quad \text{for each index } i = 1,..,k \text{ and} \\ z_j^1 &> z_j^2 \quad \text{for at least one index } j = 1,..,k\,. \end{aligned}$$

A vector $z^* \in \mathbb{R}^k$ is Pareto optimal if there is no other vector $z \in \mathbb{R}^k$ such that $z \succeq_P z^*$. Moreover, the Pareto front is the set of all Pareto optimal points.

The method:

- Start with an initial partition where each node represents a community, to which corresponds the initial modularity vector $Q = (Q_1, \dots, Q_k)$.
- Proceed with a two-phase scheme which generates a list L of community assignments and corresponding modularity vectors such that no one is Pareto-dominated by the others.

In particular:

- Consider each node one by one and insert it in all communities of its neighbors.
- Consider only new modularity vectors that:
  - yield a “strict improvement” of scalar function F;
  - are not dominated according to Pareto.

  Insert the new modularity vector and remove from L all the partitions whose modularity vectors are dominated by the newly inserted one.
- Final control on the length of L by filtering out the elements of L that have small value of F, maintaining only h partitions.
- Repeated until no further improvement in the Pareto sense is possible.

- Selects from L the best partition with respect to F and uses this as new starting point for the second phase.

We explore different choices of F:

- Modularity average on the layers

$$M_Q = \frac{\sum_{s=1}^{k} Q_s}{k} \tag{3}$$

  where k is the number of layers and Q_s is the modularity of layer s.

- Convex combination of average and variance for Informative case

$$F_{\text{-}} = (1-\gamma)M_Q - \gamma V_Q \tag{4}$$

- Convex combination of average and variance for Noisy case

$$F_{+} = (1-\gamma)M_Q + \gamma V_Q \tag{5}$$

  where V_Q is the variance of modularity on the layers and $\gamma \in [0,1]$.

We highlight that these functions can be evaluated via an inexpensive iterative update.

$$\Delta M_Q^{(i\to j)} = \frac{1}{k}\sum_{s=1}^{k} \Delta Q_s^{(i\to j)} \tag{6}$$

$$\Delta F_{\pm}^{(i\to j)} = (1-\gamma)\Delta M_Q^{(i\to j)} \pm \gamma R_Q^{(i\to j)} \tag{7}$$

where

$$R_Q^{(i\to j)} = V_{\Delta Q}^{(i\to j)} + \frac{2}{k-1}(Q - M_Q\mathbf{1})^\top(\Delta Q^{(i\to j)} - \Delta M_Q^{(i\to j)}\mathbf{1})$$

$$V_{\Delta Q}^{(i\to j)} = \frac{1}{k-1}\sum_{s=1}^{k}(\Delta Q_s^{(i\to j)} - \Delta M_Q^{(i\to j)})^2$$

We implemented the methods using Matlab and we tested them both on artificial and real world networks. We did experiments on synthetic and real-world networks, showing the effectiveness and the robustness of the proposed strategies both in the informative case, where all layers show the same community structure, and in the noisy case, where some layers represent only noise.

5 Network semi-supervised learning

Graph Semi-Supervised learning is an important data analysis tool, where given a graph and a set of labeled nodes, the aim is to infer the labels to the remaining unlabeled nodes.

The problem statement is:

- G a graph
- K classes $C_1, C_2, \ldots, C_K$
- $T = \{v_1, v_2, \ldots, v_t\} \subset V$ with $|T| \ll |V|$ labeled nodes.

We represent $T = \{v_1, v_2, \ldots, v_t\}$ by an input assignment matrix Y with K columns $y_1, y_2, \ldots, y_K \in \mathbb{R}^{\mathrm{V}}$ s.t.

$$Y_{\mathrm{ij}} = \begin{cases} 1 & \text{if node } i \text{ is assigned the label } j, \\ 0 & \text{otherwise.} \end{cases}$$

For each community $k \in K$, we consider the function:

$$\varphi_k(z) = \|z - y_k\|_2^2 + \frac{\lambda}{2} \sum_{i,j} A^G_{ij} \left(\frac{z_i}{\sqrt{\delta_i}} - \frac{z_j}{\sqrt{\delta_j}} \right)^2$$

where:

- y_k is the indicator vector of class k
- A^G adjacency matrix of the graph
- $\lambda \geq 0$
- $\delta_i = \sum_j A^G_{ij}$ = weighted degree of i in G

We solve the following optimization problems: $\min_z \varphi_k(z) \quad \forall k \in K$ obtaining the solutions $(z^*_k)_i = \mathrm{Prob}(i \in \text{class } k)$. Therefore, at the end we have a solution vector for each calss $z^*_1, \ldots, z^*_K$, and we insert node i in class k with maximum probability.

In our paper [12], we start by considering the optimization-based formulation of the problem above for an undirected graph, and then we extend this formulation to multilayer hypergraphs.

$$\varphi_k(z) = \|z - y_k\|_2^2 + \sum_{\ell \text{ layer}} \frac{\lambda^\ell}{2} \sum_{i,j} A^{H^\ell}_{ij} \left(\frac{z_i}{\sqrt{\delta^\ell_i}} - \frac{z_j}{\sqrt{\delta^\ell_j}} \right)^2$$

where:

- Multilayer: $A^{G^1}, \dots, A^{G^L}$. The regularization term is summed up across the layers
- Hypergraphs: hyperedges substituted by cliques. The ajacency matrix A^H is obtained by $II^T - D$ where I incident matrix and D degree diagonal matrix.

This extension brings additional complexity and a harder tractability. Therefore, we solve the problem using different coordinate descent approaches, and compare the results with the ones obtained by the classic gradient descent method. We took into considerations:

- Cyclic Coordinate Descent. At every iteration, a variable index is chosen in a cyclic fashion.
- Random Coordinate Descent. At every iteration a variable index is randomly chosen from a uniform distribution.
- Gauss-Southwell Coordinate Descent. At every iteration k a variable index is chosen for every class c as $i_c^k \in \operatorname{Argmax}_{i=1,\dots,n} |\nabla_{ic}\varphi(z^k)|$

Experiments on synthetic and real-world datasets show the potential of using coordinate descent methods with suitable selection rules for the problems at hand.

Clustering (or community detection) on multilayer graphs poses several additional complications with respect to standard graphs as different layers may be characterized by different structures and types of information. One of the major challenges is to establish the extent to which each layer contributes to the cluster assignment in order to effectively take advantage of the multilayer structure and improve upon the classification obtained using the individual layers or their union. However, making an informed a-priori assessment about the clustering information content of the layers can be very complicated. In [13], we assume a semi-supervised learning setting, where the class of a small percentage of nodes is initially provided, and we propose a Laplacian-regularized model that learns an optimal nonlinear combination of the different layers from the available input labels. In particular we consider:

$$\varphi(z, y_c, \alpha, \theta) = \|z - y_c\|_2^2 + \sum_{i,j=1}^{N} g_{\alpha,\theta}(A^{(1)}, \dots, A^{(K)})_{ij} \left(\frac{z_i}{\sqrt{\delta_i^\ell}} - \frac{z_j}{\sqrt{\delta_j^\ell}} \right)^2 \tag{8}$$

where $g_{\alpha,\theta}$ is the generalized mean:

$$g_{\alpha,\theta}(A^{(1)}, \dots, A^{(K)})_{ij} = \left(\sum_{\ell \text{ layer}} \theta_\ell A_{ij}^{(\ell)\alpha} \right)^{\frac{1}{\alpha}} \quad \text{with } \alpha \in \mathbb{R}, \theta \geq 0, e^T\theta = 1$$

We can then formulate, for each community, the following bilevel optimization problem:

$$
\begin{aligned}
\min_{\alpha,\theta} \quad & H(y^{te}, z_{\alpha,\theta,y^{tr}}) \\
\text{s.t.} \quad & z_{\alpha,\theta,y^{tr}} = \operatorname*{Argmin}_x \varphi(z,\alpha,\theta,y^{tr}) \\
& \alpha \in \mathbb{R} \\
& \theta \geq 0, e^T\theta = 1
\end{aligned} \tag{9}
$$

with H is a general loss function.

The lower level problem can be solved explicitly using Label Spreading.
We fix α, θ and minimize φ in the z variable.

$$
\nabla_z \varphi(z,\alpha,\theta) = 2\{(z-y) + \lambda(D(\alpha,\theta) - M(\alpha,\theta))z\} \tag{10}
$$

$$
\begin{aligned}
\nabla_x \varphi(z,\alpha,\theta) = 0 \iff & (z-y) + \lambda(D(\alpha,\theta) - M(\alpha,\theta))z = 0 \\
\iff & (I + \lambda D(\alpha,\theta) - \lambda M(\alpha,\theta))z = y \\
\iff & z = (I + \lambda D(\alpha,\theta) - \lambda M(\alpha,\theta))^{-1} y
\end{aligned} \tag{11}
$$

In practice, we solve this system of linear equations, running the fixed point iteration:

$$
z_{r+1} = \lambda M(\alpha,\theta)(I + \lambda D(\alpha,\theta))^{-1} z_r + (I + \lambda D(\alpha,\theta))^{-1} y \tag{12}
$$

for $r = 0, 1, 2, 3, \ldots$ and starting with $z_0 = \mathbf{0}$.

The feasible region of our problem is:

$$
S = \begin{cases} \alpha \in [-20, 20] \\ \theta \geq 0 \\ e^T\theta = 1 \end{cases} \tag{13}
$$

Since the feasible region S is a compact convex set and H is a differentiable real-valued function, we solve the problem using the Frank Wolfe algorithm (or conditional gradient method).
In each iteration we solve the linearized problem:

$$
(\hat{\alpha}_n, \hat{\theta}_n) = \min_{(\alpha,\theta)\in S} \nabla H(\alpha_n, \theta_n)^T((\alpha,\theta) - (\alpha_n, \theta_n))
$$

which can be solved separately in the the variable $\alpha \in [-20, 20]$ and the set of variables $\theta \in \mathbb{R}^K$.

For variable α, we need to solve

$$
\begin{aligned}
\min_{\alpha} \quad & \nabla_\alpha H(\alpha_n, \theta_n)(\alpha - \alpha_n) \\
\text{s.t.} \quad & \alpha \in [-20, 20]
\end{aligned} \tag{14}
$$

Therefore, since we aim to minimize a linear function over a box constraint, the solution would be:

$$\hat{\alpha}_n = \begin{cases} -20 & \text{if } \nabla_\alpha H(\alpha_n, \theta_n) > 0 \\ 20 & \text{otherwise} \end{cases} \tag{15}$$

For the set of variables $\theta \in \mathbb{R}^K$, we need to solve

$$\begin{aligned} \min_{\theta} \quad & \nabla_\theta H(\alpha_n, \theta_n)^T (\theta - \theta_n) \\ \text{s.t.} \quad & \theta \geq 0 \\ & e^T \theta = 1 \end{aligned} \tag{16}$$

Therefore, since we aim to minimize a linear function over the unit simplex, the solution would be:

$$\hat{\theta}_n = e_{\hat{\jmath}} \tag{17}$$

where $\hat{\jmath} = \text{Argmin}_{j=1,\ldots,K} [\nabla_\theta H(\alpha_n, \theta_n)]_j$ and $e_{\hat{\jmath}}$ is a vector of the canonical basis of $\mathbb{R}^K$.

In the paper we provide a detailed convergence analysis of the algorithm and extensive experiments on synthetic and real-world datasets, showing that the proposed method compares favourably with a variety of baselines and outperforms each individual layer when used in isolation.

References

[1] M. Newman, “Networks”. Oxford university press, 2018.

[2] S. Fortunato, C.T. Bergstrom, K. Börner, J.A. Evans, D. Helbing, S. Milojevíc, A.M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, et al., “Science of science”. Science, vol. 359, no. 6379, p. eaao0185, 2018.

[3] X. Sun, J. Kaur, S. Milojevíc, A. Flammini, and F. Menczer, “Social dynamics of science”. Scientific reports, vol. 3, no. 1, p. 1069, 2013.

[4] V. Larivière, C. Ni, Y. Gingras, B. Cronin, and C.R. Sugimoto, *Bibliometrics: Global gender disparities in science.* Nature, vol. 504, no. 7479, pp. 211–213, 2013.

[5] A. Clauset, S. Arbesman, and D.B. Larremore, *Systematic inequality and hierarchy in faculty hiring networks.* Science advances, vol. 1, no. 1, p. e1400005, 2015.

[6] R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási, *Quantifying the evolution of individual scientific impact.* Science, vol. 354, no. 6312, p. aaf5239, 2016.

[7] L. Wu, D. Wang, and J.A. Evans, *Large teams develop and small teams disrupt science and technology.* Nature, vol. 566, no. 7744, pp. 378–382, 2019.

[8] S. Venturini, S. Sikdar, F. Rinaldi, F. Tudisco, and S. Fortunato, *Collaboration and topic switches in science.* arXiv preprint arXiv:2304.06826, 2023.

[9] M. Kivelä, A. Arenas, M. Barthelemy, J.P. Gleeson, Y. Moreno, and M.A. Porter, *Multilayer networks.* Journal of complex networks, vol. 2, no. 3, pp. 203–271, 2014.

[10] F. Battiston and G. Petri, “Higher-Order Systems”. Springer, 2022.

[11] S. Venturini, A. Cristofari, F. Rinaldi, and F. Tudisco, *A variance-aware multiobjective Louvain-like method for community detection in multiplex networks.* Journal of Complex Networks, vol. 10, 11 2022.

[12] S. Venturini, A. Cristofari, F. Rinaldi, and F. Tudisco, *Laplacian-based semi-supervised learning in multilayer hypergraphs by coordinate descent.* arXiv preprint arXiv:2301.12184, 2023.

[13] S. Venturini, A. Cristofari, F. Rinaldi, and F. Tudisco, *Learning the right layer: a data-driven layer-aggregation strategy for semi-supervised learning on multilayer graphs.* Accepted to International Conference on Machine Learning (ICML), 2023.

A brief introduction to Bloch-Kato conjecture

Shilun Wang [(*)]

This is a basic introduction to the Bloch–Kato conjecture. This conjecture appeared in [1]. It generalizes some important part of the Birch and Swinnerton-Dyer conjecture.

In characteristic 0, the Bloch–Kato conjecture relates two objects attached to a geometric Galois representation. A geometric Galois representation V is a semisimple continuous representation of the absolute Galois group G_K of a number field K on a finite dimensional vector space V over $\mathbb{Q}_p$. It has certain properties that is satisfied by the Galois representations that appears in the étale cohomology $H^i_{\text{ét}}(X, \mathbb{Q}_p)$ of proper and smooth variety X over K. To a geometric representation V of G_K, one can attach two objects, one analytic, and one algebraic. The Bloch–Kato conjecture is a mysterious relation between those objects. The analytic object is an analytic L-function $L(V, s)$ of a complex variable s with possibly some poles. The algebraic object is called the Bloch–Kato Selmer groups and denoted by $H^1_f(G_K, V)$. It is a $\mathbb{Q}_p$-vector space. We also give the motivic interpolation of Bloch–Kato conjecture and under some assumptions, we can have an explicit form of the conjecture.

1 Terminology and convention

We always let a p-adic representation V of G be a finite-dimensional vector space over $\mathbb{Q}_p$, with a continuous linear action of a topological group G. If V is a p-adic representation, $V(n)$ is V tensor the cyclotomic character to the power n. We let K be a number field, and denote by G_K its absolute Galois group. We use v to denote a place of K, and G_v will denote G_{K_v}. There is a natural morphism $G_v \longrightarrow G_K$ well defined up to conjugacy, so we can define the restriction $V|_{G_v}$ to G_v of a representation of G_K.

2 Galois representation

2.1 Representations coming from geometry

We let X be a proper and smooth variety over K of dimension n, i be an positive integer and p be a prime number. One sets

$$H^i(X, \mathbb{Q}_p) = (\varprojlim H^i_{\text{ét}}(X \times_K \overline{K}, \mathbb{Z}/p^n\mathbb{Z})) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p.$$

The $\mathbb{Q}_p$-linear space $H^i(X, \mathbb{Q}_p)$ has a natural $\mathbb{Q}_p$-linear action of the Galois group G_K. We have the following properties about the $\mathbb{Q}_p$-linear space.

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

1. The space $H^i(X,\mathbb{Q}_p)$ is finite dimensional and of dimension independent of p. The action of G_K is continuous.
2. $X \longmapsto H^i(X,\mathbb{Q}_p)$ is a contravariant functor from the category of proper and smooth varieties over K to the category of p-adic representations of G_K.
3. We have $H^i(X,\mathbb{Q}_p)=0$ for $i<0$ and $i>2\dim X$. If X is geometrically connected, then $H^0(X,\mathbb{Q}_p)=\mathbb{Q}_p$ and $H^{2n}(X,\mathbb{Q}_p)=\mathbb{Q}_p(-n)$.
4. There is a functorial cup product map of G_K-representations $H^i(X,\mathbb{Q}_p)\otimes H^j(X,\mathbb{Q}_p)\longrightarrow H^{i+j}(X,\mathbb{Q}_p)$. When $i+j=2n$, it is a perfect pairing. In particular, $H^i(X,\mathbb{Q}_p)^*\cong H^{2n-i}(X,\mathbb{Q}_p)(-n)$.
5. Let v be a finite place of K prime to p. If X has good reduction at v, then the representation $H^i(X,\mathbb{Q}_p)$ is unramified at v. The characteristic polynomial of Frob_v acting on $H^i(X,\mathbb{Q}_p)$ has its coefficients in $\mathbb{Z}$, and is independent of p. We call it $P_v(X)\in\mathbb{Z}[X]$. Its roots all have complex absolute value equal to $q_v^{-i/2}$, where q_v is the cardinality of the residue field k_v.
6. If v be a place of K dividing p, then as a representation of G_v, $H^i(X,\mathbb{Q}_p)$ is de Rham. If X has good reduction at v, $H^i(X,\mathbb{Q}_p)$ is even crystalline.

Definition 2.1 Let V be an irreducible p-adic representation of G_K. We say that V comes from geometry if there is an integer i, an integer n and a proper and smooth variety X over K such that V is isomorphic to a subquotient of $H^i(X,\mathbb{Q}_p)(n)$.

If V is a semi-simple representation of G_K, we say V comes from geometry if every irreducible component of V comes from geomtry.

2.2 Geometric Galois representation

Definition 2.2 Let V be a p-adic semi-simple representation of G_K. We say that V is geometric if it is unramified at almost all places and de Rham at all places dividing p.

2.3 Algebraicity and purity

Let V be a representation of G_K that is unramified outside a finite set of places Σ.

Definition 2.3 We say that a representation is algebraic if there is a finite set of places Σ' containing Σ such that the characteristic polynomial of Frob_v on V has coefficients in $\overline{\mathbb{Q}}$ when $v\notin\Sigma'$. When one wants to precise the set Σ', we say Σ'-algebraic.

For $\omega\in\mathbb{Z}$, we say that a representation is pure of weight ω if there is a finite set of places Σ' containing Σ such that V is Σ'-algebraic and all the roots of the characteristic polynomial of Frob_v have complex absolute values $q_v^{-\omega/2}$. When we want to specify the set Σ', one says Σ'-pure.

When V is pure of weight ω, we call ω the motivic weight of V, or simply its weight.

2.4 Motivic weight and Hodge–Tate weights

A geometric representation V of dimension d of G_K which is pure has exactly one motivic weight. But each of its restriction to G_v for v dividing p has d Hodge–Tate weights.

Definition 2.4 For a geometric representation V of G_K, and for each $k \in \mathbb{Z}$, we denote by $m_k(V)$, the sum

$$m_k(V) = \sum_{v|p} [K_v : \mathbb{Q}_p] m_k(V_{|G_v})$$

where $m_k(V_{|G_v})$ is the multiplicity of the Hodge–Tate weight k for the representation $V_{|G_v}$ of G_v. We call $m_k(V)$ the total multiplicity of k as a Hodge–Tate weight of V.

Obviously, the $m_k(V)$ are almost all 0, and we have

$$\sum_{k \in \mathbb{Z}} m_k(V) = [K : \mathbb{Q}] \dim V.$$

Lemma 2.5 *If K_0 is a subfield of K, and $W = \operatorname{Ind}_{G_K}^{G_{K_0}} V$, then $m_k(V) = m_k(W)$.*

Proposition 2.6 *Let V be a p-adic representation of G_K that is Hodge–Tate at all places dividing p and pure of weight ω. Then we have*

$$\omega [K : \mathbb{Q}] \dim V = 2 \sum_{k \in \mathbb{Z}} m_k(V) k.$$

3 Bloch–Kato Selmer groups

3.0.1 Kummer morphism

Let K be a field and A be a commutative group scheme over K, such that the map 'multiplication by p' which we denote by $[p]\colon A \longrightarrow A$ is finite and surjective. Let N be an integer, the kernel of the map $[p^n]\colon A \longrightarrow A$ denoted by $A[p^n]$ is a finite abelian group scheme over K, and $A[p^n](\overline{K})$ is a finite abelian group with a continuous action of G_K. The multiplication by p induces surjective homomorphisms $A[p^{n+1}] \longrightarrow A[p^n]$ of group scheme over K, hence surjective morphism $A[p^{n+1}](\overline{K}) \longrightarrow A[p^n](\overline{K})$ compatible with the action of G_K. We set $T_p(A) = \varprojlim A[p^n](\overline{K})$ and $V_p(A) = T_p(A) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$. The space $V_p(A)$ is a p-adic representation of G_K.

Example 3.1

1. If $A = \mathbb{G}_m$, then $V = \mathbb{Q}_p(1)$.

2. If A is an abelian variety, then $V_p(A)$ is the usual Tate module of A. It satisfies $V_p(A)^*(1) \cong V_p(A)$.

We can construct injective maps

$$\kappa_n\colon A(K)/p^nA(K) = A(K)\otimes_{\mathbb{Z}}\mathbb{Z}/p^n\mathbb{Z} \longrightarrow H^1(G_K, A[p^n]) \longrightarrow H^1(G_K, A[p^n](\overline{K}))$$

for n are compatible, so they define a map

$$\varprojlim A(K)\otimes_{\mathbb{Z}}\mathbb{Z}/p^n\mathbb{Z} \longrightarrow \varprojlim H^1(G, A[p^n](\overline{K})).$$

The left hand side is the p-adic completion of $A(K)$, and we will denote by $\widehat{A(K)}$. There is a natural map from $A(K)\otimes_{\mathbb{Z}}\mathbb{Z}_p$ to $\widehat{A(K)}$ which is an isomorphism if $A(K)$ is finitely generated. The right hand side is $H^1(G_K, T_p(A))$. Tensoring by $\mathbb{Q}_p$, we finally get an injective map

$$\kappa\colon \widehat{A(K)}\otimes_{\mathbb{Z}_p}\mathbb{Q}_p \longrightarrow H^1(G_K, V_p(A)).$$

3.0.2 Results in local Galois cohomology

Let K be a finite extension of $\mathbb{Q}_\ell$, and V be a p-adic representation of G_K. We have the following properties.

Proposition 3.2

1. *We have $H^i(G_K, V) = 0$ if $i > 2$.*
2. *We have a canonical isomorphism $H^2(G_K, \mathbb{Q}_p(1)) = \mathbb{Q}_p$ and the pairing*

   $$H^i(G_K, V)\times H^{2-i}(G_K, V^*(1)) \longrightarrow H^2(G_K, \mathbb{Q}_p(1)) = \mathbb{Q}_p$$

   given by the cup product is a perfect pairing for $i = 0,\ 1,\ 2$.
3. *We have the following formula on the dimension*

   $$\dim H^0(G_K, V) - \dim H^1(G_K, V) + \dim H^2(G_K, V) = \begin{cases} 0, \textit{ if } \ell \neq p \\ [K:\mathbb{Q}_p]\dim V, \textit{ if } \ell = p. \end{cases}$$

3.0.3 The unramified H^1

Definition 3.3 We denote the unramified H^1 by $H^1_{\mathrm{ur}}(G_K, V)$, which equals

$$\ker(H^1(G_K, V) \longrightarrow H^1(I_K, V)).$$

We have the following properties.

Proposition 3.4

1. *We have* $\dim H^1_{\mathrm{ur}}(G_K, V) = \dim H^0(G_K, V)$.

2. *An element of $H^1(G_K, V)$ that corresponds to an extension*

$$0 \longrightarrow V \longrightarrow W \longrightarrow \mathbb{Q}_p \longrightarrow 0$$

is in $H^1_{\mathrm{ur}}(G_K, V)$ if and only if the sequence

$$0 \longrightarrow V^{I_K} \longrightarrow W^{I_K} \longrightarrow \mathbb{Q}_p \longrightarrow 0$$

is still exact.

3. *Assume $\ell \neq p$. Then from the duality between $H^1(G_K, V)$ and $H^1(G_K, V^*(1))$, the orthogonal of $H^1_{\mathrm{ur}}(G_K, V)$ is $H^1_{\mathrm{ur}}(G_K, V^*(1))$.*

Example 3.5 We consider the case that the representation $V = \mathbb{Q}_p(1)$. The Kummer map is an isomorphism $\kappa \colon \widehat{K}^\times \otimes_{\mathbb{Z}_p} \mathbb{Q}_p \xrightarrow{\sim} H^1(G_K, \mathbb{Q}_p(1))$. When $p \neq \ell$, the isomorphism κ identifies the subspace $\widehat{\mathcal{O}_K}^\times \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$ of $\widehat{K}^\times \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$ with the subspace $H^1_{\mathrm{ur}}(G_K, \mathbb{Q}_p(1))$ of $H^1(G_K, \mathbb{Q}_p(1))$.

3.0.4 Results in Global Galois cohomology and Selmer group

Let K be a number field and p be a prime number. We will always let Σ denote a finite set of primes of K containing all primes above p and ∞. We let $G_{K,\Sigma}$ is the absolute Galois group of the maximal extension K_Σ of K which is unramified at all the places in Σ. Let V be a p-adic representation of $G_{K,\Sigma}$.

For Global Galois cohomology, we still have a simple Euler–Poincaré formula.

Proposition 3.6

$$\dim H^0(G_{K,\Sigma}, V) - \dim H^1(G_{K,\Sigma}, V) + \dim H^2(G_{K,\Sigma}, V) = \sum_{v|\infty} H^0(G_v, V) - [K : \mathbb{Q}] \dim V.$$

Definition 3.7 Let V be a p-adic representation of G_K unramified almost everywhere. A Selmer structure $\mathcal{L} = (L_v)$ for V is the data of a family of subspaces L_v of $H^1(G_v, V)$ for all finite places v of K such that for almost all v, $L_v = H^1_{\mathrm{ur}}(G_v, V)$.

Definition 3.8 The Selmer group attached to $\mathcal{L}$ is the subspace $H^1_{\mathcal{L}}(G_K, V)$ of elements $x \in H^1(G_K, V)$ such that for all finite places v, we have

$$H^1_{\mathcal{L}}(G_K, V) = \ker(H^1(G_K, V) \longrightarrow \prod_{v \text{ finite place of } K} H^1(G_v, V)/L_v).$$

Definition 3.9 If $\mathcal{L}$ is a Selmer structure for V, we define a Selmer structure $\mathcal{L}^\perp$ for $V^*(1)$ by taking for $L_v^\perp$ the orthogonal of L_v in $H^1(G_v, V^*(1))$.

We have the following duality result.

Proposition 3.10

$$\begin{aligned}\dim H^1_{\mathcal{L}}(G_K, V) = {} & \dim H^1_{\mathcal{L}^\perp}(G_K, V^*(1)) \\ & + \dim H^0(G_K, V) - \dim H^0(G_K, V^*(1)) \\ & + \sum_{v \text{ place of } K} \dim L_v - \dim H^0(G_v, V).\end{aligned}$$

3.1 The local Bloch–Kato Selmer groups at places dividing p

Now K is a finite extension of $\mathbb{Q}_p$.

3.1.1 The local Bloch and Kato's finite spaces

If V is a p-adic representation of G_K, we have the following definition.

Definition 3.11 We set the finite space is

$$H^1_f(G_K, V) = \ker(H^1(G_K, V) \longrightarrow H^1(G_K, V \otimes_{\mathbb{Q}_p} B_{\text{cris}})).$$

We have a more concrete description of H^1_f: An element of $H^1(G_K, V)$ that corresponds to an extension

$$0 \longrightarrow V \longrightarrow W \longrightarrow \mathbb{Q}_p \longrightarrow 0$$

is in $H^1_f(G_K, V)$ if and only if the sequence

$$0 \longrightarrow D_{\text{cris}}(V) \longrightarrow D_{\text{cris}}(W) \longrightarrow D_{\text{cris}}(\mathbb{Q}_p) \longrightarrow 0$$

is still exact. In particular, if V is crystalline, then the extension W is in $H^1_f(G_K, V)$ if and only if it is crystalline.

When V is de Rham, we can compute the dimension of the local H^1_f as following.

Proposition 3.12 *Let $D^+_{dR}(V) = (V \otimes B^+_{dR})^{G_K} \subset D_{dR}(V) = (V \otimes B_{dR})^{G_K}$. Then we have*

$$\dim_{\mathbb{Q}_p} H^1_f(G_K, V) = \dim_{\mathbb{Q}_p}(D_{dR}(V)/D^+_{dR}(V)) + \dim_{\mathbb{Q}_p} H^0(G_K, V).$$

We can also find that H^1_f behaves well under the duality. The orthogonal of $H^1_f(G_K, V)$ is $H^1_f(G_K, V^*(1))$ for the duality between $H^1(G_K, V)$ and $H^1(G_K, V^*(1))$.

The next proposition shows that H^1_f is an analogy of H^1_{ur} when $\ell = p$.

Proposition 3.13 *The Kummer map $\kappa\colon \widehat{K}^\times \otimes_{\mathbb{Z}_p} \mathbb{Q}_p \longrightarrow H^1(G_K, \mathbb{Q}_p(1))$ identifies $\mathcal{O}_K^\times \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$ with $H^1_f(G_K, \mathbb{Q}_p(1))$.*

When E is an elliptic curve over K, the Kummer isomorphism κ for E is an isomorphism $E(K) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p \cong H^1_f(G_K, V_p(E))$.

3.1.2 The variants H^1_g and H^1_e

Bloch and Kato also define two variants of $H^1_f(G_K, V)$, one smaller $H_e(G_K, V)$ and one larger $H_g(G_K, V)$.

They are defined as

$$H^1_g(G_K, V) = \ker(H^1(G_K, V) \longrightarrow H^1(G_K, V \otimes B_{\mathrm{dR}}))$$
$$H^1_e(G_K, V) = \ker(H^1(G_K, V) \longrightarrow H^1(G_K, V \otimes B_{\mathrm{cris}}^{\phi=1})).$$

Since $B_{\mathrm{cris}}^{\phi=1} \subset B_{\mathrm{cris}} \subset B_{\mathrm{dR}}$, we have

$$H^1_e(G_K, V) \subset H^1_f(G_K, V) \subset H^1_g(G_K, V).$$

We also have a concrete description of H^1_g and H^1_e: An element of $H^1(G_K, V)$ that corresponds to an extension

$$0 \longrightarrow V \longrightarrow W \longrightarrow \mathbb{Q}_p \longrightarrow 0$$

is in $H^1_g(G_K, V)$(resp. in $H^1_e(G_K, V)$)if and only if the sequence

$$0 \longrightarrow D_{\mathrm{cris}}(V) \longrightarrow D_{\mathrm{cris}}(W) \longrightarrow D_{\mathrm{cris}}(\mathbb{Q}_p) \longrightarrow 0$$

(resp.

$$0 \longrightarrow D_{\mathrm{cris}}(V)^{\phi=1} \longrightarrow D_{\mathrm{cris}}(W)^{\phi=1} \longrightarrow D_{\mathrm{cris}}(\mathbb{Q}_p) \longrightarrow 0$$

) is still exact. In particular, if V is de Rham, then the extension W is in $H^1_g(G_K, V)$ if and only if it is de Rham.

We have some dimensional relations between H^1_f, H^1_g as following.

$$\dim(H^1_g(G_v, V)/H^1_f(G_v, V)) = \dim D_{\mathrm{cris}}((V|_{G_v})^*(1))^{\phi=1}, \text{ if } v \mid p$$
$$\dim(H^1_g(G_v, V)/H^1_f(G_v, V)) = \dim H^0(G_v, V^*(1)), \text{ if } v \nmid p$$

Assume that V is de Rham. From the duality between $H^1(G_K, V)$ and $H^1(G_K, V^*(1))$, we know that the orthogonal $H^1_e(G_K, V)$ is $H^1_g(G_K, V^*(1))$ and the orthogonal of $H^1_g(G_K, V)$ is $H^1_e(G_K, V^*(1))$.

3.1.3 Analogies

For K a finite extension of $\mathbb{Q}_\ell$ and V a p-adic representation, we have the following natural subspaces of $H^1(G_K, V)$

$$\begin{cases} \text{case } l \neq p & (0) \subset H^1_{\mathrm{ur}}(G_K, V) \subset H^1(G_K, V) \\ \text{case } l = p & (0) \subset H^1_e(G_K, V) \subset H^1_f(G_K, V) \subset H^1_g(G_K, V) \subset H^1(G_K, V). \end{cases}$$

So when $\ell \neq p$, we set $H^1_e(G_K, V) = 0$, $H^1_f(G_K, V) = H^1_{\mathrm{ur}}(G_K, V)$ and $H^1_g(G_K, V) = H^1(G_K, V)$.

3.2 Global Bloch–Kato Selmer group

Now we assume K is a number field and V is a geometric p-adic representation of G_K.

3.2.1 Definitions

We directly give the definition of the global Bloch–Kato Selmer group.

Definition 3.14 The global Bloch–Kato Selmer group $H^1_f(G_K, V)$ is the subspace of elements $x \in H^1(G_K, V)$ such that for all finite places v of K, the restriction x_v of x belongs to $H^1_f(G_v, V)$.

More generally, if S is any finite set of finite places of K, we define $H^1_{f,S}(G_K, V)$ as the subspace of elements $x \in H^1(G_K, V)$ such that for all finite places v of K, the restriction x_v of x belongs to $H^1_f(G_v, V)$ if $v \notin S$ and to $H^1_g(G_K, V)$ if $v \in S$. We call $H^1_g(G_K, V)$ the union of all $H^1_{f,S}(G_K, V)$ when S runs among finite sets of primes of K.

We now know that the Bloch–Kato Selmer group $H^1_f(G_K, V)$ is a Selmer group in the sense of Definition 3.8: it is the Selmer groups $H^1_{\mathcal{L}_f}(G_K, V)$ attached to the Selmer structure $\mathcal{L}_f = (L_v)$, where $L_v = H^1_f(G_v, V)$ for all v. So is $H^1_{f,S}(G_K, V) = H^1_{\mathcal{L}_{f,S}}(G_K, V)$, where $\mathcal{L}_{f,S} = (L_v)$ is the Selmer structure (L_v) with $L_v = H^1_f(G_v, V)$ for $v \notin S$ and $L_v = H^1_g(G_v, V)$ if $v \in S$. They are finite dimensional over $\mathbb{Q}_p$. The Selmer structure $\mathcal{L}_f$ is self-dual, which means that the structure $\mathcal{L}_f^{\perp}$ of $V^*(1)$ is the same as $\mathcal{L}_f$. So we have following dimension formula for Bloch–Kato Selmer group.

Theorem 3.15

$$\begin{aligned}\dim H^1_f(G_K, V) &= \dim H^1_f(G_K, V^*(1)) \\ &= \dim H^0(G_K, V) - \dim H^0(G_K, V^*(1)) \\ &+ \sum_{v|p} \dim D_{dR}(V_{|G_v})/D^+_{dR}(V_{|G_v}) \\ &- \sum_{v|\infty} \dim H^0(G_v, V).\end{aligned}$$

3.3 Some examples

Here we talk about two important examples for Bloch–Kato Selmer groups: $V = \mathbb{Q}_p(1)$ and $V = V_p(E)$ for E an elliptic curve.

3.3.1 The case $V = \mathbb{Q}(1)$

Proposition 3.16 *The Kummer map κ realizes an isomorphism*

$$\mathcal{O}_K^\times \otimes_{\mathbb{Z}} \mathbb{Q}_p \longrightarrow H^1_f(G_K, \mathbb{Q}_p(1)).$$

This result relating with the Bloch–Kato Selmer group of $\mathbb{Q}_p(1)$ is a classical object of interest in arithmetic.

3.3.2 The case $V = V_p(E)$ for E an elliptic curve

Now let E be an elliptic curve over K. We recall that the classical p-adic Selmer group $\mathrm{Sel}_p(E)$ of E is defined as the subspace of $H^1(G_K, V_p(E))$ whose elements are x whose restriction x_v at every finite place v belongs to the image of $E(K_v)$ by local Kummer map in $H^1(G_v, V_p(E))$. It is known that the Kummer map induces an injection $\kappa\colon E(K) \otimes_{\mathbb{Z}} \mathbb{Q}_p \hookrightarrow \mathrm{Sel}_p(E)$ which is an isomorphism if and only if the p-primary component $Ш(E)[p^\infty]$ of the Tate–Shafarevich group $Ш(E)$ of E is finite.

Proposition 3.17 *As subspace of $H^1(G_K, V_p(E))$, we have*

$$\mathrm{Sel}_p(E) = H^1_f(G_K, V_p(E)).$$

In particular, the Kummer map induces an injection $E(K) \otimes_{\mathbb{Z}} \mathbb{Q}_p \longrightarrow H^1_f(G_K, V_p(E))$ which is isomorphism if and only if $Ш(E)[p^\infty]$ is finite.

4 L-function

4.1 Definition of L-functions

4.1.1 Euler factors

Let V be a p-adic geometric representation of G_K. We always fix an embedding of $\mathbb{Q}_p$ into $\mathbb{C}$.

For every finite place v of K that does not divides p, we set

$$L_v(V, s) = \det(\mathrm{id} - (\mathrm{Frob}_v^{-1}\, q_v^{-s})\; |_{V^{I_v}})^{-1}.$$

Here s is a complex variable, q_v is the cardinality of the residue field of K at v, and the matrix of Frob_v is seen as a complex matrix using our embedding. The function $s \longmapsto L_v(V, s)$ is called an Euler factor, it is clearly a rational function from $\mathbb{C}$ to $\mathbb{C}$ with only a finite number of poles. It is formally a power series in the variable q_v^{-s}.

For places v of V dividing p, we set

$$L_v(V, s) = \det((\mathrm{id} - \varphi^{-1} q_v^{-1})_{|D_{\mathrm{cris}}(V|_{G_v})})^{-1},$$

where $\varphi = \phi^{f_v}$, ϕ is the crystalline Frobenius and f_v is the integer such that $q_v = p^{f_v}$.

Definition 4.1 We set formally

$$L(V, s) = \prod_{v \text{ finite place of } K} L_v(V, s).$$

More generally, for S any finite set of finite places of K, we set

$$L_S(V, s) = \prod_{v \text{ finite place of } K \text{ not in } S} L_v(V, s)$$

The product of Euler factors defining the L-function is called an Euler product.

Note that if $V = V_1 \oplus V_2$ as G_K representation, then $L(V,s) = L(V_1,s)L(V_2,s)$. Since geometric representations are semi-simple, it is often enough to consider irreducible V.

We have some basic properties.

1. $L(V(n),s) = L(V,s+n)$.
2. Let V be a p-adic representation of a number field K, K_0 be a subfield of K, and $W = \mathrm{Ind}_{G_K}^{G_{K_0}} V$. Then
$$L(V,s) = L(W,s).$$
If S a finite set of finite places of K_0 and S is the set of places of K that lies above some place of S_0, then
$$L_S(V,s) = L_{S_0}(W,s).$$

4.1.2 Convergence

Let V be a p-adic representation that is pure of weight $\omega \in \mathbb{Z}$. Assume that it is Σ-pure, where Σ is a finite set of finite places containing all places above p, and all places where V is ramified. Then for $v \notin \Sigma$, we have

$$L_v(V,s) = \prod_{i=1}^{\dim V} (1 - \alpha_{i,v}^{-1} q_v^{-s})^{-1}$$

where $\alpha_{1,v}, \cdots, \alpha_{\dim V,v}$ are the roots of the characteristic polynomials of Frob_v in V, and we see that $L_v(V,s)$ have no zero, and only a finite number of poles, all on the line $\Re s = \omega/2$.

Proposition 4.2 *Let V be a representation that is Σ-pure of weight ω. Then the Euler product defining $L_\Sigma(V,s)$ converges absolutely and uniformly on all compact on the domain $\Re s > \omega/2 + 1$.*

So we know that $L_\Sigma(V,s)$ is a well defined holomorphic function with no zero on the domain $\Re s > \omega/2 + 1$. The function $L(V,s)$ are well-defined meromorphic functions with no zero on the domain $\Re s > \omega/2 + 1$.

Example 4.3

1. If $V = \mathbb{Q}_p$, the function $L(V,s)$ is the Dedekind zeta function $\zeta_K(s)$. It is well known it has an analytic continuation to $\mathbb{C}$ with only one simple pole at $s = 1$. If $V = \mathbb{Q}_p(n)$, then $L(V,s) = \zeta_K(s+n)$.
2. If $V = V_p(E)$ for E an elliptic curve over K, then $V_p(E) = H^1(E,\mathbb{Q}_p)^* = H^1(E,\mathbb{Q}_p)(1)$, hence $L(V_p(E),s) = L(H^1(E,\mathbb{Q}_p)(1),s) = L(E,s+1)$, where $L(E,s)$ is the usual L-function of the elliptic curve.

4.1.3 Analytic continuation and zeros

We have the following conjecture.

Conjecture 4.4 *Assume that V is a geometric p-adic representation of G_K, which is pure of weight ω. Then the function $L(V,s)$ admits a meromorphic continuation on $\mathbb{C}$. The function $L(V,s)$ has no zero on the domain $\Re s \geq \omega/2+1$. If V is irreducible, $L(V,s)$ has no poles, except if $V \cong \mathbb{Q}_p(n)$, in which case $L(V,s)$ has a unique simple pole at $s=n+1$.*

This conjecture is known to be true if V is automorphic, which means that it is attached to a cuspidal automorphic representation π of GL_d/K, where $d=\dim V$. We also have $L(V,s)=L(\pi,s)$ where $L(\pi,s)$ is the L-function attached to π.

4.2 The functional equation

4.2.1 The completed L-function

To state the functional equation of $L(V,s)$, we need to complete the Euler product that defines it by adding 'Euler factors at infinity'.

Now we have the total multiplicity $m_k(V)$ of the Hodge–Tate weight $k\in\mathbb{Z}$ of V and also two natural integers $a^{\pm}(V)$ which add up to $[K:\mathbb{Q}]\dim V$. We set $m_{<\omega/2}=\sum_{k<\omega/2} m_k$. Then we define

$$L_\infty(V,s)=\prod_{k\in\mathbb{Z},k<\omega/2}\Gamma_{\mathbb{C}}(s-k)^{m_k},$$

if ω is odd and when ω is even, we define a sign $\varepsilon=(-1)^{\omega/2}$ and

$$L_\infty(V,s)=\prod_{k\in\mathbb{Z},k<\omega/2}\Gamma_{\mathbb{R}}(s-\omega/2)^{a^{\varepsilon}-m_{<\omega/2}}\Gamma_{\mathbb{C}}(s-k)^{m_k}\Gamma_{\mathbb{R}}(s-\omega/2+1)^{a^{-\varepsilon}-m_{<\omega/2}}.$$

We have the following properties.

Lemma 4.5 *We have $L_\infty(V(n),s)=L_\infty(V,s+n)$ for all $n\in\mathbb{Z}$.*

Lemma 4.6 *If $\omega<0$, the function $L_\infty(V,s)$ has no pole at $s=0$. If $\omega\geq 0$ is odd, then $L_\infty(V,s)$ has a pole at $s=0$ of order $\sum_{0\leq k<\omega/2} m_k$. If $\omega\geq 0$ is even, then $L_\infty(V,s)$ has a pole at $s=0$ of order $\sum_{0\leq k<\omega/2} m_k+a^{+}-m_{\omega/2}$.*

We can set

$$\Lambda(V,s)=L(V,s)L_\infty(V,s).$$

This is the completed L-function of V.

Example 4.7 Assume $V=\mathbb{Q}_p$. Then we have $\omega=0$, $m_0=[K:\mathbb{Q}]$ and $m_{<\omega/2}=0$. Furthermore, $\varepsilon=1$, $a^{+}=r_1+r_2$ and $a^{-}=r_2$, where r_1 and r_2 are the number of real and complex places of K. We thus have

$$L_\infty(V,s)=\Gamma_{\mathbb{R}}(s)^{r_1+r_2}\Gamma_{\mathbb{R}}(s+1)^{r_2}$$

and

$$\Lambda(V,s) = \zeta_K(s)\Gamma_{\mathbb{R}}(s)^{r_1+r_2}\Gamma_{\mathbb{R}}(s+1)^{r_2} = \zeta_K(s)\Gamma_{\mathbb{R}}(s)^{r_1}\Gamma_{\mathbb{C}}(s)^{r_2}$$

This is equivalent to the work of Dedekind.

4.2.2 The functional equation

Assume that V comes from geometry, then so does $V^*(1)$. Assume Conjecture 4.4, so $L(V,s)$ and $L(V^*(1),s)$ are well-defined meromorphic function on $\mathbb{C}$. Then we have the following conjecture.

Conjecture 4.8 *There exists an entire function with no zero $\epsilon(V,s)$ such that the following holds*

$$\Lambda(V^*(1),-s) = \epsilon(V,s)\Lambda(V,s).$$

Furthermore, $\epsilon(V,s) = AB^s$ for A a complex constant and B a positive real constant.

Example 4.9 Using the functional equation above in the case $K = \mathbb{Q}$, $V = \mathbb{Q}_p$, one sees that the only zeros of $\zeta_{\mathbb{Q}}$ at integers are simple zeros at -2, -4, -6, -8, $\cdots$. This is the work of Riemann. If K is a general number field, using the functional equation above again, one sees that ζ_K has a zero at $s = 0$ of order $r_1 + r_2 - 1$. This is the work of Hecke.

4.2.3 The sign of the functional equation for a polarized representation

The functional equation given above relates two different L-functions, $L(V,s)$ and $L(V^*,s+1)$. When those function are equal, or translation of each other, things become more interesting. We will discuss cases where this happens.

Let V be a geometric and pure p-adic representation of G_K. For τ any automorphism of the field K, we denote V^τ the representation of G_K over the same space V, and an element $g \in G_K$ acts on V^τ as $\sigma g \sigma^{-1}$, where σ is a fixed element of $G_{\mathbb{Q}}$ whose restriction to K is τ. The representation V^τ only depens on τ up to isomorphism and we have $L(V,s) = L(V^\tau,s)$. Also V^τ is pure of the same weight as V.

Definition 4.10 We shall say that V is polarized if for some integer ω and some $\tau \in \mathrm{Aut}(K)$, we have $V^\tau(\omega) \cong V^*$. The integer ω is called the weight of the polarization.

If V is pure and polarized, then the weight of polarization ω is the motivic weight of V.

If V is polarized, geometric and pure of weight ω, we have $\Lambda(V^*(1),s) = \Lambda(V^\tau(1+\omega),s) = \Lambda(V,s+1+\omega)$. Therefore assuming Conjecture, the functional equation becomes

$$\Lambda(V,-s+1+\omega) = \epsilon(V,s)\Lambda(V,s).$$

It involves only one L-function $L(V,s)$ and we can talk about the center of the functional equation at $s = (\omega+1)/2$. Since $L(V,s)$ is not identically 0, one sees that $\epsilon(V,(\omega+1)/2) = \pm 1$. This sign is called the sign of the functional equation of $L(V,s)$ or simply the sign of $L(V,s)$. We have the following elementary but important relation.

Theorem 4.11 (Shimura) *The order of the zero of $L(V,s)$ at $s = (\omega+1)/2$ is odd if the sign of $L(V,s)$ is -1 and even if it is 1.*

This is especially interesting when the wight ω of V is odd, because the center of the functional equation $(\omega+1)/2$ is an integer.

5 Motivic interpretation

5.1 Motives and their realizations

5.1.1 Pure motives

Let K be a number field and write $\mathcal{V}_K$ for the category of smooth projective schemes over K. Given an object X of $\mathcal{V}_K$ and $d \in N$, denote by $\mathcal{Z}_d(X)$ the free abelian group of cycles of codimension d on X. Let $\sim$ be an adequate equivalence relation on cycles and let R be a commutative ring. Set

$$\mathcal{Z}^d_\sim(X)_R\colon = \mathcal{Z}^{d+r}_\sim(X \times_K Y)_R$$

We put $\mathcal{Z}^d_\sim(X)_R = 0$ for $d \in \mathbb{Z}_{<0}$. Let X, Y be objects of $\mathcal{V}_K$, we define the group of correspondences modulo $\sim$ of degree r from X to Y with coefficients in R is

$$\mathrm{Corr}^r_\sim(X,Y)_R\colon = \prod_{X_i} \mathcal{Z}^{\dim X_i + r}_\sim(X_i \times_K Y)_R.$$

Here X_i is the irreducible component of X. Via intersection theory, we have a product structure

$$\mathrm{Corr}^r_\sim(X,Y)_R \times \mathrm{Corr}^s_\sim(Y,Z)_R \longrightarrow \mathrm{Corr}^{r+s}_\sim(X,Z)_R.$$

Now let $\mathcal{V}^0_{K,R}$ be the category whose objects are those of $\mathcal{V}_K$ and whose morphisms are given by degree 0 correspondences modulo $\sim$ with coefficients in R.

Definition 5.1 The category $\mathscr{M}_\sim(K)_R$ of pure $\sim$-motives over K with coefficients in R is the pseudo-abelian completion of $\mathcal{V}^0_{K,R}$.

More explicitly, a pure $\sim$-motive over K with coefficients in R is a triple

$$\mathcal{M} = (X, q, r)$$

where X is an object of $\mathcal{V}_K$, $q \in \mathrm{Corr}^0_\sim(X,X)_R$ is an idempotent and $r \in \mathbb{Z}$. Furthermore, if $\mathcal{M}_1 = (X_1, q_1, r_1)$ and $\mathcal{M}_2 = (X_2, q_2, r_2)$ are motives, then

$$\mathrm{Hom}\,\mathscr{M}_\sim(K)_R(\mathcal{M}_1, \mathcal{M}_2) = q_2 \cdot \mathrm{Corr}^{r-s}_\sim(X,Y)_R \cdot q_1 \subset \mathrm{Corr}^{r-s}_\sim(X,Y)_R.$$

Given motives $\mathcal{M}_i = (X_i, q_i, r_i)$ for $i = 1$, 2, the product of $\mathcal{M}_1$ and $\mathcal{M}_2$ is

$$\mathcal{M}_1 \otimes_K \mathcal{M}_2\colon = (X_1 \times_K X_2, q_1 \times_K q_2, r_1 + r_2),$$

We know $\mathcal{M}_\sim(K)_R$ is an additive, R-linear, pseudo-abelian category. For a motive $\mathcal{M} = (X, q, r)$, we define its n-th Tate twist to be $\mathcal{M}(n)\colon = (X, q, r+n)$. The dual motive $\mathcal{M}^\vee$ is

$$\mathcal{M}^\vee\colon = (X, q^t, \dim X - r)$$

where q^t is the transpose of the projector q. A motive $\mathcal{M}$ is self-dual if $\mathcal{M} \cong \mathcal{M}^\vee(1)$.

5.1.2 Some special categories of motives

1. Chow motives: we take $\sim$ to be rational equivalence to obtain the category $\mathscr{M}_{\text{rat}}(K)_R$ of Chow motives over K with coefficients in R.
2. Grothendieck motives: we take $\sim$ to be homological equivalence to get the category of Grothendieck motives over K with coefficients in R.

5.2 Realization of $\mathcal{M}$

Now we fix two number field K and E. We have several realizations associated to $\mathcal{M}(X, q, r)$ which is a object in $\mathscr{M}_{\text{rat}}(K)_E$.

1. $\mathcal{M}_{\text{dR}} \colon = \bigoplus_{\omega\in\mathbb{Z}} q^* H^{\omega+2r}_{\text{dR}}(X/K)(r) \otimes_{\mathbb{Q}} E$ the de Rham realization equipped with its Hodge filtration $\text{Fil}^i \mathcal{M}_{\text{dR}} \colon = \bigoplus_{\omega\in\mathbb{Z}} q^* \operatorname{Fil}^{i+r} H^{\omega+2r}_{\text{dR}}(X/K) \otimes_{\mathbb{Q}} E$.
2. $\mathcal{M}_{\text{B}} \colon = \bigoplus_{\omega\in\mathbb{Z}} q^* H^{\omega+2r}_{\text{B}}(X \times_K \mathbb{C}, (2\pi i)^r E)$ the Betti realization each summand is equipped with a pure $E \otimes \mathbb{R}$-Hodge structure over $\mathbb{R}$ on $\mathcal{M}_{\text{B}} \otimes_{\mathbb{Q}} \mathbb{R}$.
3. $\mathcal{M}_p \colon = \bigoplus_{\omega\in\mathbb{Z}} q^* H^{\omega+2r}_{\text{ét}}(X \times_K \overline{K}, \mathbb{Q}_p(r)) \otimes_{\mathbb{Q}} E$ the p-adic étale realization with its $\text{Gal}(\overline{K}/K)$ action, where p is a prime number.

Denote $CH^r(X)_0 \otimes E$ the subgroup of the Chow group of X consisting of cycles homologically equivalent to 0. We define the following objects for $M = (X, q, r)$:

1. $\tan_M \colon = M_{\text{dR}}/Fil^0 M_{\text{dR}}$.
2. $H^0(K, M) \colon = (q^* CH^r(X) \otimes E)/(q^* CH^r(X)_0 \otimes E)$.
3. $H^1(K, M) \colon = q^* CH^r(X)_0 \otimes E \oplus (\bigoplus_{\omega\in\mathbb{Z}, \omega\neq -1} q^*(K_{-\omega-1}(X) \otimes E)^{(r)})$, which is motivic cohomology.

Recall that one has Chern class maps defined by Soulé

$$(K_{-\omega-1}(X) \otimes E)^{(r)} \longrightarrow H^1_{\text{cont}}(K, H^{\omega+2r}_{\text{ét}}(X \times_K \overline{K}, \mathbb{Q}_p(r))),$$

where $H^1_{\text{cont}}(K, \cdots)$ denotes the continuous Galois cohomology. Together with the Abel–Jacobi map

$$CH^r(X)_0 \longrightarrow H^1_{\text{cont}}(K, H^{-1+2r}_{\text{ét}}(X \times_K \overline{K}, \mathbb{Q}_p(r))),$$

we get

$$r_p \colon H^1(K, M) \longrightarrow H^1_{\text{cont}}(K, M_p)$$

for all p. Consider for all finite primes $\mathfrak{p}$ of K the group $H^1_f(K_{\mathfrak{p}}, M_p)$ as before.

On the other hand, there should be the following relation of $H^1_f(K, M)$ and $H^0(K, M$ to the Betti and de Rham realization: Consider the comparison isomorphism

$$I_\infty \colon M_B \otimes_{\mathbb{Q}} \mathbb{C} \cong M_{\text{dR}} \otimes_{\mathbb{Q}} \mathbb{C},$$

this induces the period map

$$\alpha_M \colon M_B^+ \otimes_{\mathbb{Q}} \mathbb{R} \longrightarrow \tan_M \otimes_{\mathbb{Q}} \mathbb{R},$$

where $M_B^+\colon = \oplus_{\mathfrak{p}|\infty} H^0(K_{\mathfrak{p}}, M_B)$ are the invariants under complex conjugation on M_B.

5.3 Motives of modular forms

Let ≥ 3 be an integer and $k \geq 4$ be an even integer. We want to introduce the Grothendieck motive of a fixed modular form of weight k and level N.

5.3.1 Anaemic Hecke algebras

We write $\mathcal{H}_k(\Gamma(N)) \subset \mathrm{End}_{\bar{\mathbb{Q}}}(S_k(\Gamma(N)), \bar{\mathbb{Q}})$ for the $\mathbb{Z}$-algebra generated by the Hecke operators T_n with $(n, N) = 1$. It is often called the anaemic Hecke algebra of weight k and level $\Gamma(N)$. We can also define $\mathcal{H}_k(\Gamma_0(N))$. There is a natural surjection

$$\mathcal{H}_k(\Gamma(N)) \twoheadrightarrow \mathcal{H}_k(\Gamma_0(N))$$

of $\mathbb{Z}$-algebras that is induced by the inclusion $S_k(\Gamma_0(N)) \subset S_k(\Gamma(N))$. Finally, for any $\mathbb{Z}$-algebra A, set $\mathcal{H}_k(\Gamma_0(N))_A\colon = \mathcal{H}_k(\Gamma_0(N)) \otimes_{\mathbb{Z}} A$.

5.3.2 The motive of modular forms of weight k and level N

Denote by $\widetilde{\mathcal{E}}_N^{k-2}$ the Kuga–Sato variety of level N and weight k, i.e., the smooth projective $\mathbb{Q}$-scheme defined as the canonical desingularization of the $(k-2)$-fold product $\mathcal{E}_N^{k-2}$ of the universal generalized elliptic curve $\pi\colon \mathcal{E}_N \longrightarrow X(N)$ over the compact modular curve $X(N)$ of level $\Gamma(N)$. If we set

$$\Gamma_{k-2}\colon = ((\mathbb{Z}/N\mathbb{Z})^2 \rtimes \{\pm 1\})^{k-2} \rtimes \mathcal{S}_{k-2},$$

where $\mathcal{S}_{k-2}$ is the symmetric group on $k-2$ letters, then there is a canonical action of Γ_{k-2} on $\widetilde{\mathcal{E}}_N^{k-2}$; we define Π_ϵ to be the projector associated with the character $\epsilon\colon \Gamma_{k-2} \longrightarrow \{\pm 1\}$ that is the sign character on $\mathcal{S}_{k-2}$, the trivial character on $(\mathbb{Z}/N\mathbb{Z})^{2(k-2)}$ and the product character on $\{\pm 1\}^{k-2}$. Moreover, write Π_B for the idempotent attached to the quotient $\Gamma_0(N)/\Gamma(N)$, whose order t_N we need to invert in order to define Π_B. Let

$$\mathcal{M}_k(N)\colon = (\widetilde{\mathcal{E}}_N^{k-2}, \Pi_B \Pi_\epsilon, k/2)$$

be the Chow motive of modular forms of weight k and level N.

5.3.3 The motive of a modular form

Let $f \in S_k(\Gamma_0(N))$ be a normalized newform of weight k and level $\Gamma_0(N)$, whose q-expansion will be denoted by $f(q) = \sum\limits_{n \geq 1} a_n(f) q^n$. Let F be the number field generated by all the eigenvalues of f, and $\mathcal{O}_F$ be its ring of integers. So the motive $\mathcal{M}(f)$ attached to f is given by

$$\mathcal{M}(f)\colon = (\widetilde{\mathcal{E}}_K^{k-2}, (1 - \Psi_f) \circ (\Pi_B \Pi_\epsilon \otimes 1), k/2).$$

Here the projector Ψ_f associated with f in [2, Section 4.2.0].

5.3.4 Realizations of $\mathcal{M}$

Let $\theta_f\colon \mathcal{H}_k(\Gamma_0(N)) \longrightarrow \mathcal{O}_F$ be the ring homomorphism associated with f given by $\theta_f(T_n)\colon = a_n(f)$. For every $\mathcal{H}_k(\Gamma(N))$-module M, let us set

$$M[\theta_f]\colon = \{m \in M \mid T \cdot M = 0 \text{ for all } T \in \ker(\theta_f)\}.$$

We call $M[\theta_f]$ the f-isotypic submodule of M.

We have the following realizations of the $\mathcal{M}$.

1. The Betti realization is the F-vector space $V_B\colon = (\Pi_B\Pi_\epsilon \cdot H^{k-1}(X(\mathbb{C}, F(k/2))))[\theta_f]$.
2. The étale realization at p is the F_p-module $V_p\colon = (\Pi_B\Pi_\epsilon \cdot H^{k-1}_{\text{ét}}(\overline{X}, F(k/2)))[\theta_f]$.
3. We have
$$V_{\text{dR}}\colon = (\Pi_B\Pi_\epsilon \cdot H^{k-1}_{\text{dR}}(X) \otimes_{\mathbb{Z}[1/N]} F))[\theta_f]$$
Here $H^i_{\text{dR}}(X)\colon = \mathbb{H}^i(\Omega^\bullet_X)$ is the i-th hypercohomology group of the de Rham complex $\Omega^\bullet_X$ of X. Since $H^i_{\text{dR}}(X)$ is equipped with the filtration
$$\text{Fil}^n(H^i_{\text{dR}}(X))\colon = \text{Im}(\mathbb{H}^i(\Omega^{\geq n}_X \longrightarrow H^i_{\text{dR}}(X))).$$
We have a filtration $\text{Fil}^n(V_{\text{dR}})$ given by
$$\text{Fil}^n(V_{\text{dR}})\colon = (\Pi_B\Pi_\epsilon \cdot \text{Fil}^{n+k+2}(H^{k-1}_{\text{dR}}(X)) \otimes_{\mathbb{Z}[1/N]} F)[\theta_f].$$
Then the de Rham realization of $\mathcal{M}$ is the filtered F-vector space $(V_{\text{dR}}, (\text{Fil}^n(V_{\text{dR}})))$.

We have a complex conjugation τ induces involutions $\iota_\infty\colon X(\mathbb{C}) \longrightarrow X(\mathbb{C})$ and $F(k/2) \longrightarrow F(k/2)$, which is given by multiplication by $(-1)^{k/2}$. We denote by $\phi_\infty\colon V_B \longrightarrow V_B$ the composition of these two involutions. We then write

$$V_B^+\colon = V_B^{\phi_\infty=1}$$

for the F-subspace of V_B on which ϕ_∞ acts trivially. We also write

$$t(\mathcal{M})\colon = V_{\text{dR}}/\,\text{Fil}^0(V_{\text{dR}}).$$

for the tangent space of $\mathcal{M}$ which is a F-vector space.

6 The Bloch–Kato conjecture

6.1 Refinement for the motive of modular forms

6.1.1 The period map

We have a comparison isomorphism of $(F \otimes_{\mathbb{Q}} \mathbb{C})$-modules

$$\text{Comp}_{B,\text{dR}}\colon V_B \otimes_{\mathbb{Q}} \mathbb{C} \xrightarrow{\sim} V_{\text{dR}} \otimes_{\mathbb{Q}} \mathbb{R}.$$

Furthermore, we have an isomorphism of $\mathbb{R}$-vector spaces

$$\mathrm{Comp}_{B,\mathrm{dR}}\colon (V_B \otimes_{\mathbb{Q}} \mathbb{C})^{\varphi_\infty \otimes \tau = 1} \xrightarrow{\sim} V_{\mathrm{dR}} \otimes_{\mathbb{Q}} \mathbb{R}.$$

Let $F_\infty\colon = F \otimes_{\mathbb{Q}} \mathbb{R}$ and set

$$V^+_{B,\infty}\colon = V^+_B \otimes_{\mathbb{Q}} \mathbb{R} = V^+_B \otimes_F F_\infty, \qquad V_{\mathrm{dR},\infty}\colon = V_{\mathrm{dR}} \otimes_{\mathbb{Q}} \mathbb{R}.$$

We also write $t(\mathcal{M})_\infty$ for $t(\mathcal{M}) \otimes_{\mathbb{Q}} \mathbb{R}$. Then the period map is the isomorphism

$$\alpha_{\mathcal{M}}\colon V^+_{B,\infty} \xrightarrow{\sim} t(\mathcal{M})_\infty$$

for free F_∞-modules of rank 1.

We have an element $\omega_f \in V_{\mathrm{dR}}$, and also there is a map from V_{dR} to $t(\mathcal{M})_\infty$. Fix $\gamma \in V^+_B \backslash \{0\}$.

Definition 6.1 The period of f relative to γ is the determinant $\Omega^{(\gamma)}_\infty$ of $\alpha_{\mathcal{M}}$ computed with respect to the basis $\{\gamma\}$ of V^+_B and the image of ω_f in $t(\mathcal{M})_\infty$.

6.1.2 Motivic cohomology and the Gillet–Soulé height pairing for $\mathcal{M}$

Denote by $\mathrm{CH}^{k/2}_0(X/K)$ the abelian group of codimension $k/2$ homologically trivial cycles on X defined over K. Let us write $\mathrm{CH}^{k/2}_{\mathrm{arith}}(X/K)$ for the subgroup of $\Pi_B \Pi_\epsilon \cdot \mathrm{CH}^{k/2}_0(X/K)$ consisting of the classes of those cycles that admit an integral model having trivial intersection with all the cycles of dimension k supported on special fibers. Then we define the zeroth and first motivic cohomology gorup of $\mathcal{M}$ over K as

$$H^i_{\mathrm{mot}}(K,\mathcal{M})\colon = \begin{cases} 0 & \text{if } i = 0, \\ \mathrm{CH}^{k/2}_{\mathrm{arith}}(X/K)_F[\theta_f] & \text{if } i = 1. \end{cases}$$

We have the following conjecture about the motivic cohomology groups.

Conjecture 6.2 *$H^1_{\mathrm{mot}}(K,\mathcal{M})$ has finite dimension over F.*

If we assume that this conjecture is valid, then we have

Definition 6.3 The algebraic rank of $\mathcal{M}$ over K is $r_{\mathrm{alg}}(\mathcal{M}/K)\colon = \dim_F(H^1_{\mathrm{mot}}(K,\mathcal{M}))$.

We have the Gillet–Soulé height pairing

$$\langle \cdot, \cdot \rangle_{\mathrm{GS}}\colon \mathrm{CH}^{k/2}_{\mathrm{arith}}(X/K) \times \mathrm{CH}^{k/2}_{\mathrm{arith}}(X/K) \longrightarrow \mathbb{R}.$$

For each $\sigma \in \Sigma$, the GS pairing induce an F-bilinear pairing

$$\langle \cdot, \cdot \rangle_{\mathrm{GS},\sigma}\colon H^1_{\mathrm{mot}}(K,\mathcal{M}) \times H^1_{\mathrm{mot}}(K,\mathcal{M}) \longrightarrow \mathbb{R} = F \otimes_{F,\Sigma} \mathbb{R}.$$

Then if we consider the F_∞-module

$$H^1_{\mathrm{mot}}(K,\mathcal{M})_\infty\colon = \prod_{\sigma \in \Sigma} H^1_{\mathrm{mot}}(K,\mathcal{M}) \otimes_{F,\sigma} \mathbb{R},$$

which is free of rank $r_{\mathrm{alg}}(\mathcal{M}/K)$. We can define an F_∞-bilinear height pairing

$$\langle\cdot,\cdot\rangle_{\mathrm{GS},\infty}\colon H^1_{\mathrm{mot}}(K,\mathcal{M})_\infty \times H^1_{\mathrm{mot}}(K,\mathcal{M})_\infty \longrightarrow F_\infty$$

by the rule

$$((x_\sigma)_{\sigma\in\Sigma},(y_\sigma)_{\sigma\in\Sigma}) \longmapsto (\langle x_\sigma, y_\sigma\rangle_{\mathrm{GS},\sigma})_{\sigma\in\Sigma},$$

where $(x_\sigma)_{\sigma\in\Sigma}$, $(y_\sigma)_{\sigma\in\Sigma}$ belong to $H^1_{\mathrm{mot}}(K,\mathcal{M})\otimes_{F,\sigma}\mathbb{R}$.

Conjecture 6.4 *For the pairing $\langle\cdot,\cdot\rangle_{GS,\infty}$, it is non-degenerate.*

So when the pairing is non-degenerate, we can use it to define the regulator of $\mathcal{M}$. Set $r\colon = r_{\mathrm{alg}}(\mathcal{M}/K)$. If $r>0$, we fix a basis $\mathscr{B}=\{t_1,\cdots,t_r\}$ of $H^1_{\mathrm{mot}}(K,\mathcal{M})$ over F. We also can see that $\mathscr{B}$ is also a basis of $H^1_{\mathrm{mot}}(K,\mathcal{M})_\infty$ over F_∞.

Definition 6.5

1. If $r>0$, then the Gillet–Soulé $\mathscr{B}$-regulator of $\mathcal{M}$ over K is

$$\mathrm{Reg}_{\mathscr{B}}(\mathcal{M}/K)\colon = \det(\langle t_i,t_j\rangle_{\mathrm{GS},\infty})_{1\le i\le j\le r}\in F_\infty.$$

2. If $r=0$, then $\mathrm{Reg}(\mathcal{M}/K)\colon = 1$.

6.1.3 L-functions of $\mathcal{M}$

Since the étale realization V_p of $\mathcal{M}$ can be viewed as a p-adic representation of G_K, hence we define $L(\mathcal{M},s)$ the L-function of $\mathcal{M}$ to be

$$L(\mathcal{M},s)\colon = L(V_p,s).$$

We introduce the archimedean factor

$$L_\infty(\mathcal{M},s)\colon = (\frac{\sqrt{N}}{2\pi})^{s+k/2}\cdot\frac{\Gamma(s+k/2)}{(i\sqrt{N})^{k/2}}$$

Then we have

Definition 6.6 The completed L-function of $\mathcal{M}$ over $\mathbb{Q}$ is

$$\Lambda(\mathcal{M},s)\colon = L_\infty(\mathcal{M},s)\cdot L(\mathcal{M},s).$$

Definition 6.7

1. The analytic rank $r_{\mathrm{an}}(\mathcal{M}/K)$ of $\mathcal{M}$ over K is the order of vanishing of $L(\mathcal{M}/K,s)$ at $s=0$.
2. The $\Lambda^*(\mathcal{M},0)$ of $\Lambda(\mathcal{M},s)$ at $s=0$ is the leading term of the Taylor expansion of $\Lambda(\mathcal{M},s)$ at $s=0$.

6.1.4 The fundamental line of $\mathcal{M}$

Definition 6.8 The fundamental line of $\mathcal{M}$ is

$$\Delta(\mathcal{M})\colon = \mathrm{Det}_F^{-1}(H^1_{\mathrm{mot}}(\mathbb{Q},\mathcal{M}))\cdot \mathrm{Det}_F(H^1_{\mathrm{mot}}(\mathbb{Q},\mathcal{M})^*)\cdot \mathrm{Det}_F(t(\mathcal{M}))\cdot \mathrm{Det}_F^{-1}(V_B^+).$$

We define the $\mathbb{R}$-vector space

$$\Delta(\mathcal{M})_\infty\colon = \Delta(\mathcal{M})\otimes_{\mathbb{Q}}\mathbb{R},$$

then $\Delta(\mathcal{M})_\infty \cong \Delta(\mathcal{M})\otimes_F F_\infty$ is a free F_∞-module of rank 1.

If the Conjecture 6.4 holds true, we have isomorphism of F_∞-modules

$$H^1_{\mathrm{mot}}(\mathbb{Q},\mathcal{M})_\infty \cong H^1_{\mathrm{mot}}(\mathbb{Q},\mathcal{M})^*_\infty.$$

We have an isomorphism

$$\theta_\infty\colon \Delta(\mathcal{M})_\infty \xrightarrow{\sim} (F_\infty, 0)$$

of F_∞-modules.

We have the following conjecture.

Conjecture 6.9 *There exists $\zeta_f \in \Delta(\mathcal{M})$ such that the equality*

$$\theta_\infty(\zeta_f) = L^*(\mathcal{M},0)^{-1}$$

holds in $F_\infty^\times$.

The element ζ_f is called a zeta element and $\{\zeta_f\}$ is a basis of $\Delta(\mathcal{M})$ over F.

6.1.5 Shafarevich–Tate groups of $\mathcal{M}$

We choose a $\mathcal{O}_p$-lattice T_p in V_p, and we take $A_p\colon = V_p/T_p$. Since we can view V_p is a p-adic representation of G_K, so we can define the Bloch–Kato Selmer groups $H^1_f(K, A_p)$ as before. For a p-primary abelian group G, we denote by G_{div} the maximal p-divisible subgroup of G.

Definition 6.10

1. The Shafarevich–Tate group of $\mathcal{M}$ over K at p is

$$Ш_p(K,\mathcal{M})\colon = H^1_f(K,A_p)/H^1_f(K,A_p)_{\mathrm{div}}.$$

2. The Shafarevich–Tate group of $\mathcal{M}$ over K is

$$Ш(K,\mathcal{M})\colon = \oplus_p Ш_p(K,\mathcal{M}),$$

where p varies over all prime numbers.

6.1.6 Tamagawa ideals of $\mathcal{M}$

For each prime $\mathfrak{p}$ of F above p, if T is a finite $\mathcal{O}_\mathfrak{p}$-module and $T = \oplus_{\mathfrak{p}|p} T_\mathfrak{p}$ is its splitting as a product of $\mathcal{O}_\mathfrak{p}$-modules, then

$$\mathcal{I}_{\mathcal{O}_p}(T) = \prod_{\mathfrak{p}|p} \mathcal{I}_{\mathcal{O}_\mathfrak{p}}(T_\mathfrak{p}) \subset F_p$$

for every finite $\mathcal{O}_p$-modules T.

Now we fix a finite prime $\ell \neq p$, and any $G_{\mathbb{Q}_\ell}$-module M, as before we have $H^1_{\mathrm{ur}}(\mathbb{Q}_\ell, M)$. By definition, $H^1_f(\mathbb{Q}_\ell, V_p) = H^1_{\mathrm{ur}}(\mathbb{Q}_\ell, V_p)$ and there is an exact sequence

$$0 \longrightarrow H^0_f(\mathbb{Q}_\ell, V_p) \longrightarrow V_p^{I_\ell} \overset{\mathrm{Frob}_\ell - 1}{\longrightarrow} V_p^{I_\ell} \longrightarrow H^1_f(\mathbb{Q}_\ell, V_p) \longrightarrow 0.$$

Then we have two isomorphisms

$$\vartheta_\ell\colon \mathrm{Det}_{F_p}(H^0_f(\mathbb{Q}_\ell, V_p)) \cdot \mathrm{Det}^{-1}_{F_p}(H^1_f(\mathbb{Q}_\ell, V_p)) \overset{\sim}{\longrightarrow} \mathrm{Det}^{-1}_{F_p}(H^1_f(\mathbb{Q}_\ell, V_p)) \overset{\sim}{\longrightarrow} (F_p, 0).$$

Definition 6.11 The p-part of the Tamagawa ideal of $\mathcal{M}$ at ℓ is

$$\mathrm{Tam}^{(p)}_\ell(\mathcal{M})\colon = \vartheta_\ell(\mathrm{Det}^{-1}_{\mathcal{O}_p}(H^1_f(\mathbb{Q}_\ell, T_p))).$$

We consider the case $\ell = p$. Then there is an exact sequence of F_p-modules

$$0 \longrightarrow H^0_f(\mathbb{Q}_p, V_p) \longrightarrow D_{\mathrm{cris}}(V_p) \overset{(\varphi-1,\mathrm{pr})}{\longrightarrow} D_{\mathrm{cris}}(V_p) \oplus t(\mathcal{M})_p \longrightarrow H^1_f(\mathbb{Q}_p, V_p) \longrightarrow 0.$$

From $H^0_f(\mathbb{Q}_p, V_p) = 0$ and taking determinants, we obtain an isomorphism

$$\vartheta_p\colon \mathrm{Det}^{-1}_{F_p}(H^1_f(\mathbb{Q}_p, V_p)) \overset{\sim}{\longrightarrow} \mathrm{Det}^{-1}_{F_p}(t_p(V_p)).$$

Define the $\mathcal{O}_p$-submodule

$$\Lambda_p\colon = \vartheta_p(\mathrm{Det}^{-1}_{\mathcal{O}_p}(H^1_f(\mathbb{Q}_p, T_p)))$$

of $\mathrm{Det}^{-1}_{F_p}(t_p(V_p))$. Since $T_p(V_p)$ is a free F_p-module of rank 1 and fix an F_p-generator ω of $t(\mathcal{M})_p$, then ω is a generator of the free F_p-module $\mathrm{Det}_{F_p}(t(\mathcal{M})_p)$ of rank 1. We also know $\mathrm{Det}^{-1}_{F_p}(t(\mathcal{M})_p)$ is free of rank 1 over F_p, and let ω^{-1} denote the generator of $\mathrm{Det}^{-1}_{F_p}(t(\mathcal{M})_p)$. Now Λ_p is an $\mathcal{O}_p$-submodule of free F_p-module $\mathrm{Det}^{-1}_{F_p}(t(\mathcal{M})_p) = F_p \cdot \omega^{-1}$, so there exists an $\mathcal{O}_p$-ideal $\mathrm{Tam}_{p,\omega}(A_p)$ such that

$$\Lambda_p = \mathrm{Tam}_{p,\omega}(A_p) \cdot \omega^{-1}.$$

We define $T_B^+\colon = T_B^{\phi_\infty = 1}$ and $T_p^+\colon = H^0(\mathbb{R}, T_p)$. There are comparison isomorphisms

$$\mathrm{Comp}_{B,\text{ét}}\colon T_B^+ \otimes_\mathbb{Q} \mathbb{Q}_p \overset{\sim}{\longrightarrow} T_p^+ \otimes_{\mathbb{Z}_p} \mathbb{Q}_p, \qquad \mathrm{Comp}_{B,\mathrm{dR}}\colon V_B^+ \otimes_\mathbb{Q} \mathbb{Q}_p \overset{\sim}{\longrightarrow} T_p^+ \otimes_{\mathbb{Z}_p} \mathbb{Q}_p.$$

There is an induced isomorphism of $\mathcal{O}_p$-modules

$$\mathrm{Comp}_{B,\text{ét}}\colon T_B^+ \otimes_{\mathbb{Q}} \mathbb{Q}_p \xrightarrow{\sim} T_p^+.$$

Choose $\delta_f \in T_B^+ \backslash \{0\}$ and set $\omega_{\delta_f}\colon = \mathrm{Comp}_{B,\mathrm{dR}}(\delta_f) \in t(V_p)$.

Definition 6.12 The p-part of the Tamagawa ideal of $\mathcal{M}$ at p is

$$\mathrm{Tam}_p^{(p)}(\mathcal{M})\colon = \mathrm{Tam}_{p,\omega_{\delta_f}}(A_p).$$

In the archimedean case, we have the following definition of the Tamagawa ideal of $\mathcal{M}$.

Definition 6.13 The p-part of the Tamagawa ideal of $\mathcal{M}$ at ∞ is

$$\mathrm{Tam}_\infty^{(p)}(\mathcal{M})\colon = \mathrm{Det}_{\mathcal{O}_p}^{-1}(H_f^1(\mathbb{R}, T_p)).$$

6.1.7 p-torsion of $\mathcal{M}$

Let $H^1(\mathbb{Q}, T_p)_{\mathrm{Tors}}$ be the torsion submodule of $H^1(\mathbb{Q}, T_p)$.

Definition 6.14 The p-torsion part of $\mathcal{M}$ is

$$\mathrm{Tors}_p(\mathcal{M})\colon = \mathcal{I}^{-1}(H^0(G_s, A_p)^\vee) \cdot \mathcal{I}^{-1}(H^1(\mathbb{Q}, T_p)_{\mathrm{Tors}}).$$

6.1.8 Some linear algebra of lattices

Let $\mathscr{B} = \{t_1, \cdots, t_r\}$ be a basis of $H^1_{\mathrm{mot}}(\mathbb{Q}, \mathcal{M})$ as an F-vector space. This is also a basis of $H^1_{\mathrm{mot}}(\mathbb{Q}, \mathcal{M})_p$ over F_p.

We assume the p-part of Conjecture over $\mathbb{Q}$, so $\widetilde{\mathscr{B}}\colon = \{x_1, \cdots, x_r\}$ is a basis of $H_f^1(\mathbb{Q}, V_p)$ as an F_p-module.

6.1.9 Explicit formula

We choose $\gamma_f \in T_B^+ \backslash \{0\}$ and let $\Omega_\infty\colon = \Omega_\infty^{(\gamma_f)} \in F_\infty^+$ be the period. We consider the $\mathcal{O}_p$-submodule Λ_{γ_f} of T_p^+ generated by $\mathrm{Comp}_{B,\text{ét}}(\gamma_f)$ and set $\mathcal{I}_p(\gamma_f)\colon \mathcal{I}(T_p^+/\Lambda_{\gamma_f})$. Let us define the period

$$\Omega_{\mathcal{M}}\colon = \frac{\Omega_\infty}{(2\pi i)^{k/2}} \in (F \otimes_{\mathbb{Q}} \mathbb{C})^\times.$$

We have the following theorem.

Theorem 6.15 *Assume that*

1. *the conjecture 6.4 holds true.*

2. *we have the isomorphism* $H^1_{\mathrm{mot}}(K,\mathcal{M})_p \cong H^1_f(K,V_p)$ *of* F_p*-modules induced by the* p*-adic regulator.*

3. *the conjecture 6.9 is valid.*

4. $Ш_p(\mathbb{Q},\mathcal{M})$ *is finite.*

The p*-part of Bloch–Kato conjecture is equivalent to the equality*

$$\left(\frac{\Lambda^*(\mathcal{M},0)}{\Omega_{\mathcal{M}}\cdot \mathrm{Reg}_{\mathscr{B}}(\mathcal{M})}\right)=\frac{\mathcal{I}(Ш_p(\mathbb{Q},\mathcal{M}))\cdot \mathcal{I}_p(\gamma_f)\cdot \prod_{v\in S}\mathrm{Tam}_v^{(p)}(\mathcal{M})}{(\det(A_{\widetilde{\mathscr{B}}}))^2\cdot \mathrm{Tors}_p(\mathcal{M})}$$

of fractional $\mathcal{O}_p$*-ideals.*

Example 6.16

1. Let $E/\mathbb{Q}$ be an elliptic curve of rank zero with finite Tate–Shafarevich group. Let $M = H^1(E)(1)$. Then $(M_B^+)_\mathbb{R} \longrightarrow (M_{\mathrm{dR}}/F^0M_{\mathrm{dR}})_\mathbb{R}$ is an isomorphism between one-dimensional vector spaces, and choosing suitable $\mathbb{Q}$-generators is given by $1 \longmapsto \Omega_E = \int_{E(\mathbb{R})}\omega$, where ω is a holomorphic differential. It follows that the Beilinson–Deligne conjecture asserts that $L(M,0)/\Omega_E \in \mathbb{Q}^\times$.

2. Let M be the Artin motive associated to the one-dimensional Galois representation $G_\mathbb{Q} \longrightarrow \{\pm 1\}$ which cuts out the extension $\mathbb{Q}(i)/\mathbb{Q}$. It corresponds to the unique non trivial Dirichlet character $\chi\colon (\mathbb{Z}/4\mathbb{Z})^\times \longrightarrow \{\pm 1\}$. One can see $L(\chi,1)$ is an non zero rational multiple of $c^+(M)=\pi$. Indeed

$$L(\chi,1) = 1 - 1/3 + 1/5 - 1/7 + \cdots = \pi/4.$$

References

[1] S. Bloch and K. Kato, *Tamagawa numbers of motives and L-functions.* The Grothendieck Festschrift, volume I (1990), 333–400.

[2] A.J. Scholl, *Motives for modular forms.* Inventiones mathematicae 100 (1990), 419–430.

A duality based DMK approach to the L^1-norm and Total Variation regularization in optimization problems

Nicola Segala [(*)]

1 Introduction: Tikhonov Regularization in Variational Problems

Consider an open Lipschitz domain $\Omega \in \mathbb{R}^n$ and a Banach space $\mathcal{H}(\Omega)$ defined on Ω. In the most general case, the Tikhonov Regularization for a Variational Problem is the following optimization problem:

$$\begin{aligned} &\min_{u\in\mathcal{H}(\Omega)} \int_\Omega c(u,\epsilon(u)) + \int_\Gamma N(u,\epsilon(u)) + \lambda \int_\Omega R(\epsilon(u)) \\ &\qquad s.t. \quad g(u) \geq 0 \quad in\ \Omega \\ &\qquad\qquad\ \ b(u) = 0 \quad in\ \Gamma = \partial\Omega \end{aligned} \tag{1}$$

Where c is a positive cost function defined on Ω, N is a positive cost function defined on $\Gamma = \partial\Omega$, g is a linear or nonlinear constraint function, b define some boundary conditions, ϵ is a linear operator between Banach spaces, R is a positive convex function (the regularizer function) and λ is a positive parameter which controls the amount of regularization desired. It is clear that one can define any discrete optimization problem by opportunely discretize a continuous variational problem.

For example, one can consider the following discrete kernel regression regularized problem, which is clearly derived from a continuous one where the integrals have become summations over a discrete samples set:

$$\begin{aligned} &\min_f \sum_{i=1}^n \frac{|\, y_i - f(x_i)\,|^2}{2} + \lambda \sum_{i=1}^n \frac{|\, f(x_i)\,|^2}{2} \\ &\qquad s.t. \quad f(x) = \sum_{i=1}^n c_i k(x,x_i) \\ &\qquad\qquad x_i \in \mathbb{R}^2,\ i = 1,...,n \\ &\qquad\qquad \lambda \geq 0 \end{aligned} \tag{2}$$

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: nsegala@math.unipd.it. Seminar held on 15 February 2023.

In this example the regularization term is nothing that the $2 - norm$ squared of the regression function f and is modulated by it's Tikhonov parameter λ.
In Figure 1 we can see the effect of the Tikhonov regularization for problem (2) where we can clearly see that the regularization avoids the overfitting.

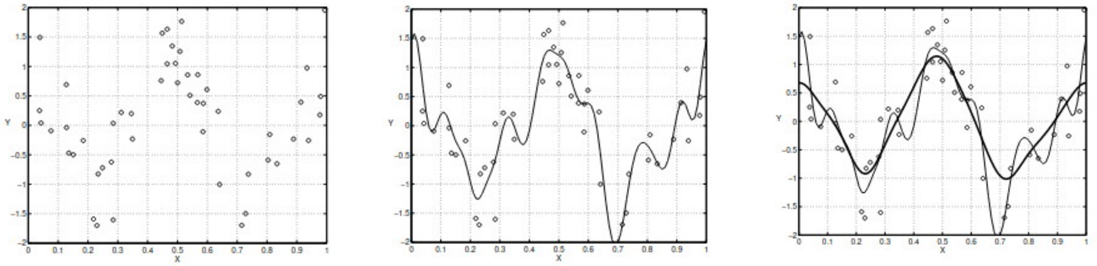


Figure 1: Samples (left), Overfitting (center), Regularized (right).

In this seminar we will focus our attention on a class of non common regularizers derived from the $1 - norm$, namely the $L_1 - norm$ of our design parameters which is sometimes called the *compressed sensing* regularization and the Total Variation (TV) regularizer.
These choices of regularization functionals have the nice property to improove the sparsity (*compressed sensing*) or to improove the local flatness (TV) of the optimal solutions. Since both the $L_1 - norm$ and the TV are non differentiable functionals , the use of such regularization strategies is hampered by the difficulty in finding efficient and robust numerical solution algorithms.

2 A duality based reformulation for the Total Variation Energy functional

Consider an open Lipschitz domain $\Omega \in \mathbb{R}^n$. Given a parameter $p \in \mathbb{R}$ with $1 < p \leq 2$ we define the p-Dirichlet energy on $W^{1,p}(\Omega)$ as:

$$E_p(\varphi) := \int_\Omega \frac{1}{p} |\nabla \varphi|^p \tag{3}$$

The dual definition of the p-Dirichlet energy reads as follows:

$$E^*_{p'}(\varphi) = \sup_\sigma - \int_\Omega \frac{|\sigma|^{p'}}{p'} + \int_\Omega \sigma \nabla \varphi$$

Where $p' = \frac{p}{p-1}$ is che conjugate exponent to p and $\sigma \in \left[W^{p'}(\mathrm{div}, \Omega)\right]^n := \{v \in [L^{p'}(\Omega)]^n | \operatorname{div} v \in L^{p'}(\Omega)$.
We recall here the Legendre transform for a sufficiently regular function $G(x)$, $x \in X$:

$$G^*(x^*) = \sup_{x \in X} \int_\Omega x^* x - G(x) \quad \forall x^* \in X^* \tag{4}$$

We call $x \in X$ the state variable and $x^* \in X^*$ the conjugate or adjoint variable. In particular if G is proper, l.s.c. and convex then is possible to show that:

$$G(x) = \sup_{x^* \in X^*} \int_\Omega xx^* - G^*(x^*) \tag{5}$$

Defining as state variable $x = |\sigma|^2$ and conjugate variable $x^* = \mu \geq 0$, applying twice the Legendre transform on $E^*_{p'}(\varphi)$ we obtain the following duality based reformulation:

$$\begin{aligned}\mathcal{L}^*_\beta(\varphi) := E^*_{p'}(\varphi) &= \sup_\sigma \left[-\sup_{\mu\geq 0} \int_\Omega \frac{\mu|\sigma|^2}{2} - \frac{1}{2}\int_\Omega \frac{\mu^{\frac{2-\beta}{\beta}}}{\frac{(2-\beta)}{\beta}} \right] + \int_\Omega \sigma \nabla \varphi \\ &= \sup_\sigma \inf_{\mu \geq 0} - \int_\Omega \frac{\mu|\sigma|^2}{2} + \frac{1}{2}\int_\Omega \frac{\mu^{\frac{2-\beta}{\beta}}}{\frac{(2-\beta)}{\beta}} + \int_\Omega \sigma \nabla \varphi\end{aligned} \tag{6}$$

Where $\beta := 2 - p$ and $0 \leq \beta < 1$.
As often happens in optimization theory, one can also consider the dual problem associated with $\mathcal{L}^*_\beta(\varphi)$. Since we are dealing with a suddle point problem, the standard theory of duality in these cases (see for example [2] and [3] for an exaustive treatment) tells us that the dual problem is achieve by exanching the "inf" with the "sup", therefore formally we define the following dual problem:

$$\mathcal{L}_\beta(\varphi) := \inf_{\mu\geq 0} \sup_\sigma - \int_\Omega \frac{\mu|\sigma|^2}{2} + \int_\Omega \sigma \nabla \varphi + \frac{1}{2}\int_\Omega \frac{\mu^{\frac{2-\beta}{\beta}}}{\frac{(2-\beta)}{\beta}} \tag{7}$$

We will see in the next section that the dual problem (7), once defined on opportune function spaces, has indeed some advantages and can be used to formulate a new definition for the total variation of a function in $BV(\Omega)$.

2.1 Equivalence between Total Variation Energy and $\mathcal{L}_\beta$ in the limit case $\beta = 1$

As in Section 2 we have seen the equivalence between the p-Dirichlet energy and our dual definition $\mathcal{L}^*_\beta(\varphi)$ in the case where $0 \leq \beta < 1$, one may wonder what happens in the limit case where $\beta = 1$, which implies $p = 1$. Consider an open Lipschitz domain $\Omega \in \mathbb{R}^n$, then for any function $\varphi \in W^{1,1}(\Omega)$ we define the Total Variation energy as:

$$TV(\varphi) := \int_\Omega |\nabla \varphi| \,. \tag{8}$$

By standard duality arguments, we can define the Total Variation energy for a function $\varphi \in BV(\Omega)$ as:

$$\begin{aligned} TV(\varphi) :=& \sup_{\substack{\sigma \in \mathcal{C}^\infty_c(\Omega)^n \\ |\sigma(x)| \leq 1 \; \forall x \in \Omega}} - \int_\Omega \varphi \operatorname{div} \sigma \\ =& \sup_{\substack{\sigma \in \mathcal{C}^\infty_c(\Omega)^n \\ |\sigma(x)| \leq 1 \; \forall x \in \Omega}} \int_\Omega \sigma \nabla \varphi \end{aligned} \tag{9}$$

Where one has to pay attention to the definition of $\nabla\varphi$ in (9). It is a known fact (see [4] Structure Theorem of BV functions) that if $\varphi \in BV(\Omega)$ then:

$$\begin{gathered}\nabla\varphi \in [\mathcal{M}(\Omega)]^n \\ \nabla\varphi = \nu\gamma \\ |\nabla\varphi| = \nu \in \mathcal{M}^+(\Omega) \\ \gamma : \Omega \mapsto \mathbb{R}^n \quad \nu - measurable \\ |\gamma(x)| \leq 1 \; \forall x \in \Omega\end{gathered} \tag{10}$$

Where $\mathcal{M}(\Omega)$ is the set of the Radon measures in Ω and $\mathcal{M}^+(\Omega)$ is the set of the positive Radon measures in Ω.

Motivated by these facts we introduce our candidate definition of the 1-Dirichlet Energy as:

$$\mathcal{L}_1^*(\varphi) := \sup_{\sigma\in[\mathcal{C}_c^\infty(\Omega)]^n}\left[\inf_{\mu\in\mathcal{M}^+(\Omega)} -\int_\Omega \frac{\mu|\sigma|^2}{2} + \int_\Omega \frac{\mu}{2}\right] + \int_\Omega \sigma\nabla\varphi \tag{11}$$

It is straightforward to see that:

$$\inf_{\mu\in\mathcal{M}^+(\Omega)} -\int_\Omega \frac{\mu|\sigma|^2}{2} + \int_\Omega \frac{\mu}{2} = \begin{cases} 0 & |\sigma(x)| \leq 1, \; \forall x \in \Omega \\ -\infty & |\sigma(x)| > 1, \; \forall x \in \Omega \end{cases} \tag{12}$$

Therefore, since we are looking for a supremum in σ in (11) we have that:

$$\mathcal{L}_1^*(\varphi) = TV(\varphi) \tag{13}$$

Consider now the candidate dual problem:

$$\mathcal{L}_1(\varphi) := \inf_{\mu\in\mathcal{M}^+(\Omega)} \sup_{\sigma\in[\mathcal{C}_c^\infty(\Omega)]^n} -\int_\Omega \frac{\mu|\sigma|^2}{2} + \int_\Omega \sigma\nabla\varphi + \int_\Omega \frac{\mu}{2} \tag{14}$$

The goal of this section is actually to prove that the dual problem defined in (14) is precisely the Total Variation of φ.

Consider now the following variational problem:

$$\begin{gathered} T_m(\varphi) = \inf_{\mu\in\mathcal{M}^+(\Omega)} \sup_{\sigma\in[\mathcal{C}_c^\infty(\Omega)]^n} -\int_\Omega \frac{1}{2}\mu|\sigma|^2 + \int_\Omega \sigma\nabla\varphi \\ s.t. \quad \int_\Omega \mu = m \end{gathered} \tag{15}$$

Then we can prove the following Theorem:

Theorem 2.1 *For any $\varphi \in BV(\Omega)$ we have that:*

$$T_m(\varphi) = \frac{(TV(\varphi))^2}{2m} \tag{16}$$

From Theorem 2.1 we have an immediate corollary:

Corollary 2.2 *For any $\varphi \in BV(\Omega)$ we have that:*

$$TV(\varphi) = 2T_{TV(\varphi)}(\varphi) \tag{17}$$

We are now ready to state the main Theorem of this section:

Theorem 2.3 *Given a function $\varphi \in BV(\Omega)$, we have that:*

$$TV(\varphi) = 2T_{TV(\varphi)}(\varphi) = \mathcal{L}_1(\varphi) \tag{18}$$

Moreover, the optimal measure μ^* for (14) satisfies $\mu^* = |\nabla \varphi|$.

Theorem 2.3 states the equivalence between $TV(\varphi)$ and $\mathcal{L}_1(\varphi)$, with the same arguments one can show the following:

Theorem 2.4 *Given a function $\varphi \in L^1(\Omega)$, we have that:*

$$||\varphi||_1 = \inf_{\nu \in \mathcal{M}^+(\Omega)} \sup_{\sigma \in \mathcal{C}_c^\infty(\Omega)} -\int_\Omega \frac{\nu|\sigma|^2}{2} + \int_\Omega \sigma\varphi + \int_\Omega \frac{\nu}{2} \tag{19}$$

The optimal measure ν^* for (19) satisfies $\nu^* = |\varphi|$.

Remark 2.5 For the readers that have familiarity with optimal transport theory, we point out that the results presented in Theorems 2.1 and 2.3 are an extension of the results presented in [1] by Bouchitté and Buttazzo for the L_1 Optimal Transport problem.

2.2 The regularized problem

In order to avoid ill-conditioning still maintaining sufficient accuracy, it is useful at least from a numerical point of view, to consider a regularized version for the functional defined in (14). In the framework of the Tikhonov regularization, consider the following regularized functional:

$$\mathcal{L}_{1,\delta}(\varphi) := \inf_{\mu \in \mathcal{M}^+(\Omega)} \sup_{\sigma \in [\mathcal{C}_c^\infty(\Omega)]^n} -\int_\Omega \frac{(\mu+\delta)|\sigma|^2}{2} + \int_\Omega \sigma \nabla \varphi + \int_\Omega \frac{\mu}{2} \tag{20}$$

where $\delta > 0$, $\delta << 1$ is a small Tikhonov parameter.
Since now $\mu + \delta > 0$ the supremum in (20) is in fact a maximum ant the optimal solution σ^* satisfies $\sigma^* = \frac{\nabla \varphi}{\mu+\delta}$.
The functional defined in (20) simplifies as follows:

$$\mathcal{L}_{1,\delta}(\varphi) = \inf_{\mu \in \mathcal{M}^+(\Omega)} \int_\Omega \frac{|\nabla \varphi|^2}{2(\mu+\delta)} + \int_\Omega \frac{\mu}{2} \tag{21}$$

The same goes for the L_1-norm in (19) and we define the regularized norm as:

$$||\varphi||_{1,\delta} = \inf_{\nu \in \mathcal{M}^+(\Omega)} \int_\Omega \frac{|\varphi|^2}{2(\nu+\delta)} + \int_\Omega \frac{\nu}{2} \tag{22}$$

3 A DMK (Dynamic Monge Kantorovich) approach to the numerical solution of 1-Harmonic functions with given profile boundary data

Consider now the problem of finding minimizers for the Total Variation energy. The Euler-Lagrange equation for the Total Variation energy involves the 1-Laplacian operator defined as:

$$\Delta_1(u) = -\operatorname{div}\left(\frac{\nabla u}{|\nabla u|}\right) \quad u \in W^{1,1}(\Omega) \tag{23}$$

There is a wide literature about the 1-Laplacian operator. Indeed, since the 1-Laplacian operator is not surjective, one is interested to compute 1-Harmonic functions on some open bounded Lipschitz domain $\Omega \in \mathbb{R}^n$ with non homogeneous Dirichlet boundary conditions on $\Gamma = \partial\Omega$. 1-Harmonic functions are related to the problem of finding surfaces with minimal mean-curvature and naturally, since the sparsity of the gradient is promoted by the L^1 norm, are flat solutions.
Consider for example a parametric surface on $\mathbb{R}^3$:

$$S(x,y) := \begin{pmatrix} x \\ y \\ z = u(x,y) \end{pmatrix} \tag{24}$$

where $u : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$. Consider now the level set:

$$\Gamma_c := \{\vec{x} \in \Omega | u(\vec{x}) = c\} \tag{25}$$

Then by classical differential geometry arguments, the unit normal to Γ_c is nothing that $\nu = \frac{\nabla u}{|\nabla u|}$ and the mean curvature is $H = \operatorname{div}(\nu) = \Delta_1(u)$, therefore 1-Harmonic functions are locally zero mean curvature hypersurfaces.
Note that, since the TV Energy is not differentiable, subgradients need to be used, making its numerical minimization highly nontrivial.
We can use the experience gained with the DMK scheme, see [5], to tackle this problem in a different way with the aim to develop alternative and more performing numerical minimization strategies. Consider an open bounded Lipschitz domain $\Omega \in \mathbb{R}^n$ and set $\Gamma = \partial\Omega$. Given a profile function $\gamma(x) : \Gamma \mapsto \mathbb{R}$ we look at the following variational problem:

$$\begin{gathered} \inf_{u \in W^{1,1}(\Omega)} \int_\Omega |\nabla u| \\ s.t.\ u(x) = \gamma(x) \quad x \in \Gamma \end{gathered} \tag{26}$$

Problem (26) can be simplified by introducing an appropriate lifting function $\bar{u}$ such that $\bar{u}(x) = \gamma(x)$ for all $x \in \Gamma$ and considering:

$$\inf_{u \in {W_0}^{1,1}(\Omega)} \int_\Omega |\nabla(u + \bar{u})| \tag{27}$$

By Theorem 2.3, we have the following Theorem.

Theorem 3.1 *Problem (27) is equivalent to the following variational problem:*

$$\begin{aligned}\inf_{u\in {W_0}^{1,1}(\Omega)} TV(u+\bar{u}) &= \inf_{u\in {W_0}^{1,1}(\Omega)} \mathcal{L}_1(u+\bar{u}) = \\ = \inf_{\mu\in\mathcal{M}^+(\Omega)} \inf_{u\in {W_0}^{1,1}(\Omega)} &\sup_{\sigma\in[\mathcal{C}_c^\infty(\Omega)]^n} -\int_\Omega \frac{1}{2}\mu|\sigma|^2 + \int_\Omega \sigma\,\nabla(u+\bar{u}) + \int_\Omega \frac{1}{2}\mu\end{aligned} \tag{28}$$

Moreover the optimal solution triplet (μ^*, u^*, σ^*) *satisfies the following "Monge-Kantorovich" type equations:*

$$\begin{gathered}\sigma^* \in \partial||\nabla u^*||_1 \\ |\sigma^*(x)| \leq 1 \quad \forall x\in\Omega \\ |\sigma^*| = 1 \quad \textit{on } \mu^* > 0 \\ \mu^*\sigma^* - \nabla u^* = 0 \\ \operatorname{div}\sigma^* = 0 \\ u^*(x) = \gamma(x) \quad x\in\Gamma\end{gathered} \tag{29}$$

Using the KKT conditions for Problem (28) we can introduce a DMK like gradient flow (decreasing direction in μ) converging to the critical points:

$$\begin{cases}\mu(t)\sigma(t) - \nabla u(t) = 0 \\ \operatorname{div}\sigma(t) = 0 \\ u(t)(x) = \gamma(x) \quad x\in\Gamma \\ \dot{\mu}(t) = \mu(t)|\sigma(t)|^2 - \mu(t) \\ \mu(0) = \mu_0\end{cases} \tag{30}$$

Remark 3.2 (Some remarks on the numerical implementation: The regularized problem) Consider the gradient flow defined in (30). Since 1-Harmonic functions are flat solutions, our optimal density μ^* will inevitably goes to zero in some points, leading to very ill conditioned linear systems in the numerical discretization. Inspired by [5, 6] a very efficient choice is to consider the regularized variational problem defined in (21).
Rewriting Problem (28) using the approximated regularized functional $\mathcal{L}_{1,\delta}$ we have:

$$\begin{aligned}\inf_{u\in {W_0}^{1,1}(\Omega)} TV(u+\bar{u}) &\approx \inf_{u\in {W_0}^{1,1}(\Omega)} \mathcal{L}_{1,\delta}(u+\bar{u}) = \\ &= \inf_{\substack{\mu\in\mathcal{M}^+(\Omega) \\ u\in {W_0}^{1,1}(\Omega)}} \int_\Omega \frac{|\nabla(u+\bar{u})|^2}{2(\mu+\delta)} + \int_\Omega \frac{\mu}{2}\end{aligned} \tag{31}$$

And the corresponding DMK scheme rewrites as:

$$
\begin{cases}
-\operatorname{div}\left(\dfrac{\nabla u(t)}{(\mu(t)+\delta)}\right) = 0 \\
u(t)(x) = \gamma(x) \quad x \in \Gamma \\
\dot{\mu}(t) = \mu(t)\dfrac{|\nabla u(t)|^2}{(\mu(t)+\delta)^2} - \mu(t) \\
\mu(0) = \mu_0
\end{cases}
\tag{32}
$$

We conclude this section with some numerical solutions of Problem (26) using our regularized DMK scheme defined in (32). For the numerical test we consider a 2 dimensional square domain $\Omega = [0,1] \times [0,1]$ and two different profile functions $\gamma(x) : \mathbb{R}^2 \mapsto \mathbb{R}$:

- Test Case 1 (Cross Vault):

$$
\begin{cases}
\gamma(x,y) = \sqrt{0.25 - (x-0.5)^2} & y = 0 \text{ } or \text{ } y = 1 \\
\gamma(x,y) = \sqrt{0.25 - (y-0.5)^2} & x = 0 \text{ } or \text{ } x = 1
\end{cases}
\tag{33}
$$

- Test Case 2 (Tensile Structure):

$$
\begin{cases}
\gamma(x,y) = (x-0.5)\operatorname{sign}(x-0.5) & y = 0 \text{ } or \text{ } y = 1 \\
\gamma(x,y) = (y-0.5)\operatorname{sign}(y-0.5) & x = 0 \text{ } or \text{ } x = 1
\end{cases}
\tag{34}
$$

In Figure 2, we can see the comparison between the 1-harmonic solution for test case 1 (left) and the 1-harmonic solution for test case 2 (right) computed with our DMK solver (32).

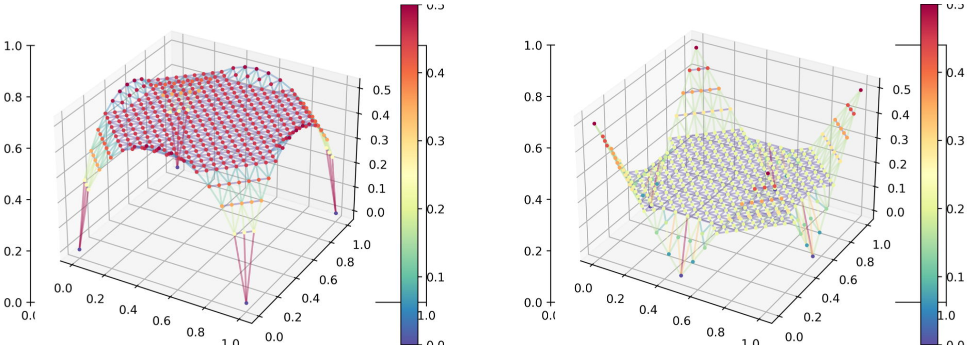


Figure 2: Test Case 1 (left) and Test Case 2 (right).

4 The Discrete 1-D signal TV Denoising

In this section we will see an application of our DMK scheme for the numerical solution of a classical denoising problem often denoted as the ROF (Rudin-Osher-Fatemi) problem.

Given a noisy sampled digital signal $b = [b_1, ..., b_n] \in \mathbb{R}^n$ and a parameter $\lambda > 0$, consider the following discrete optimization problem:

$$\min_{u=[u_1,..,u_n]\in\mathbb{R}^n} \frac{1}{2}\sum_{k=1}^{n} |u_k - b_k|^2 + \lambda \sum_{i=1}^{n-1} |u_{i+1} - u_i| \tag{35}$$

We can genuinely assign a differential structure to the collection of the n samples b by introducing the discrete counterparts of the differentials operator on graphs.
We define an undirected graph as a collection $\mathcal{G} = (E, V)$, where E is the set of $m = |E|$ edges, V the set of $n = |V|$ nodes. Each edge $e_i \in E$ is characterized by the pair $e_i = v_j v_k$, with $v_j, v_k \in V$. On a graph, we can define functions on nodes and functions on edges. We denote as $\mathcal{H}(V)$ and $\mathcal{H}(E)$ the Banach spaces of real-valued functions on V and E, respectively. We now introduce the graph gradient operator $\nabla : \mathcal{H}(V) \longrightarrow \mathcal{H}(E)$ as the $m \times n$ matrix whose (i, j)-element is

$$(\nabla)_{ij} = \begin{cases} -1 & e_i = v_j v_k,\ k \in \{1, \dots, n\} \\ 1 & e_i = v_k v_j,\ k \in \{1, \dots, n\} \\ 0 & \text{otherwise.} \end{cases} \tag{36}$$

Although the matrix ∇ relies on a edge orientations, we point out that such orientation is arbitrary and does not affect the construction of the operator so that $\mathcal{G}$ can be still considered as undirected.

Next, we define the graph divergence operator div : $\mathcal{H}(E) \longrightarrow \mathcal{H}(V)$, which can be expressed as

$$\mathrm{div} = -\nabla^T, \tag{37}$$

i.e., as the negative transposed $n \times m$ matrix of ∇.
In analogy to the continuous case, we define the weighted graph laplacian operator for a function $f \in \mathcal{H}(V)$ as:

$$(\Delta_c f)(u) := -div(c \odot \nabla f) = \nabla^T (Diag(c) \nabla f) \tag{38}$$

where $c \in \mathcal{H}(E)$ is a weight functions on the edges set, $\odot$ is the Hadamard product:

$$\begin{aligned} u \odot v : \mathbb{R}^n \times \mathbb{R}^n &\to \mathbb{R}^n \\ (u \odot v)_i &\mapsto u_i \cdot v_i \qquad i = 1, \dots, n \end{aligned}$$

and

$$\Delta_c = \nabla^T Diag(c) \nabla \tag{39}$$

is the c-weighted graph laplacian matrix.
Consider now the $1D$-graph of n time samples $G = (V, E)$, $V = \{t_1, .., t_n\}$, $E = \{(t_i, t_{i+1}) \mid i = 1, .., n-1\}$, then the $1D$-ROF problem (35) rewrites as:

$$\min_{u\in\mathcal{H}(V)} \frac{1}{2}\sum_{v\in V} |u(v) - b(v)|^2 + \lambda \sum_{e\in E} |\, \nabla u(e)\, | \tag{40}$$

As we have done for 1-Harmonic functions, we can use our regularized functional $\mathcal{L}_{1,\delta}$ to numerically solve Problem (40). Consider therefore the following approximated optimization problem:

$$\min_{u\in\mathcal{H}(V)} \min_{\substack{\mu\,\in\mathcal{H}(E)\\ \mu\geq 0}} \frac{1}{2}\sum_{v\in V}|u(v)-b(v)|^2+\lambda\sum_{e\in E}\frac{|\,\nabla u(e)\,|^2}{2(\mu+\delta)}+\lambda\sum_{e\in E}\frac{\mu}{2} \tag{41}$$

and the resulting DMK scheme:

$$\begin{cases} \Big(\lambda\,\Delta_{\frac{1}{\mu(t)+\delta}}+\mathbb{1}\Big)u(t)=b & \forall v\in V \\ \dot{\mu_{uv}}(t)=\mu_{uv}(t)\dfrac{|\,\nabla\,u(t)_{uv}|^2}{(\mu_{uv}(t)+\delta)^2}-\mu_{uv}(t) & \forall uv\in E \\ \mu_{uv}(0)=\mu_0 & \forall uv\in E \end{cases} \tag{42}$$

We conclude this section with an application of our proposed DMK scheme (42) for the TV denoise of a 128 samples very noisy digital signal (the noise is the 30 % of the original clean signal).
In Figure 3 , we can see the comparison between the original clean signal (green), the noisy added signal (grey) and the denoised signal (red) computed with our DMK solver (42).

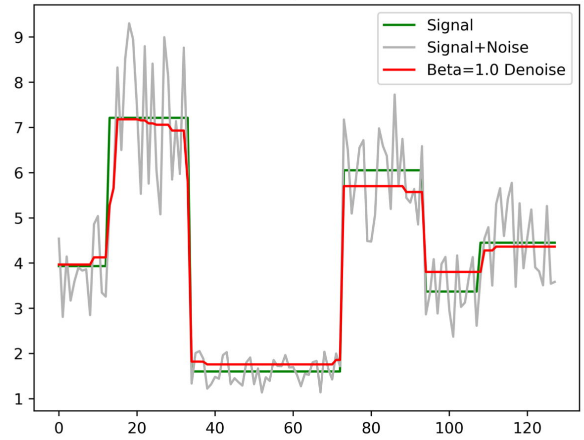


Figure 3: Original clean signal (green), noisy added signal (grey), denoised signal (red).

5 Compressed Modes for the Graph Laplacian

Nowadays, in the era of big data, clustering and reducing order models techniques play a very important role in data analysis.
Spectral clustering and PCA are widely used techniques in data analysis and data mining, taking advantages from very efficient numerical algorithms directly inherited from the huge weaponry of numerical linear algebra.

Despite that, when one has to deal with big data, such techniques suffer from the lack of sparsity of their final outputs.
To overcome this problem, the compressed modes CM technique (often referred in literature as Sparse PCA) was early introduced by Osher et al. in [7] for the laplacian matrix as an l_1 matrix norm Tikhonov regularization problem, in order to compute a sparse orthonormal approximated basis for the partial laplacian eigenvalue problem.
In this last section we will see an application of our DMK technique for the L_1-norm in order to provide an alternative new algorithm to the standard Bregman-Osher split iteration technique to solve the CM for the graph laplacian problem.
Let $\mathcal{G} = (E, V)$ be an undirected graph, where E is the set of $m = |E|$ edges, V the set of $n = |V|$ nodes.
Set a number k of compressed modes and denote as $U \in \mathbb{R}^{n \times k}$, $U \cdot, j = u_j$, $j = 1, .., k$ the orthonormal basis of the k approximated compressed eigenvectors, then the CM for the graph laplacian problem is the following Tikhonov regularized optimization problem:

$$\min_{\substack{U \in \mathbb{R}^{n\times k} \\ U^T U = \mathbb{1}}} \quad \frac{1}{2} U : \Delta U + \lambda ||U||_{1,1} \tag{43}$$

Where $A : B = \operatorname{tr}(A^T B)$ is the matrix scalar product, $||U||_{1,1} := \sum_{i,j} |U_{ij}|$ is the $l_{1,1}$ matrix norm and $\Delta = \nabla^T \nabla$ is the graph laplacian matrix defined in (39) with all weights equal to 1 for simplicity, without loosing generality.
Since the $l_{1,1}$ norm of a matrix is the sum of the L_1-norm of it's columns, we can tackle problem (43) with our approximated regularized functional defined in (22). Our new optimization problem becomes:

$$\min_{\substack{U \in \mathbb{R}^{n\times k} \\ U^T U = \mathbb{1}}} \min_{\substack{\nu \in \mathbb{R}^{n\times k} \\ \nu \geq 0}} \quad \frac{1}{2} U : \Delta U + \frac{\lambda}{2} U : (\nu_\delta)^{-*} \odot U + \frac{\lambda}{2} \nu : \mathbb{e}^{n\times k} \tag{44}$$

where $\nu \in \mathbb{R}^{n\times k}$, $\nu_{ij} \geq 0$, $\mathbb{e}^{n\times k} \in \mathbb{R}^{n\times k}$, $\mathbb{e}^{n\times k}_{ij} = 1$, $\nu_\delta := \nu + \delta \mathbb{e}^{n\times k}$ and $(A \odot B)_{ij} := A_{ij} B_{ij}$, $A^{-*}_{ij} = \frac{1}{A_{ij}}$.

5.1 Optimization on the Stiefel Manifold: A Feasible Update Preserving Scheme

Problem (44) is an optimization problem on the Stiefel manifold of order k defined as:

$$S^{n\times k} := \{U \in \mathbb{R}^{n\times k} \,|\, U^T U = \mathbb{1}\}, n \geq k \tag{45}$$

There is a wide literature about optimization on the Stiefel manifold, we will apply here a very beautiful technique presented in [9].
Let $f : S^{n\times k} \to \mathbb{R}$ a differentiable function and consider the following optimization problem:

$$\min_{U \in S^{n\times k}} f(U) \tag{46}$$

The Stiefel manifold inherits the Frobenius norm induced by the matrix scalar product therefore it is natural to speak about the projection into it's tangent space.
The most classical algorithm for optimization on the Stiefel manifold relies on the projected gradient descent, namely we look at the following ODE:

$$\dot{U} = -Prj_{T_U S^{n\times k}} \left(\nabla_U f(U)\right) \tag{47}$$

where $Prj_{T_U S^{n\times k}}$ denotes the projection into the tangent space of the Stiefel manifold at the point U and $\nabla_U f(U)$ is the gradient of the function f given by the Riesz representation theorem with respect to the matrix scalar product, e.g. $Df(U)\cdot\Psi = \nabla_U f(U) : \Psi$, $\forall\Psi \in T_U S^{n\times k}$.
The ODE in (47), once opportunely discretized, define a local descent direction iteration scheme for the objective function f, but inevitably at every iterations of the algorithm, the candidate minimizer for f ends outside the manifold.
As a consequence, considering for example the Esplicit Euler discretization for (47), the proper projected gradient descent iteration reads as follows:

$$U_{s+1} = Prj_{S^{n\times k}} \Big(U_s - dt\, Prj_{T_{U_s} S^{n\times k}} \left(\nabla_U f(U_s)\right)\Big) \tag{48}$$

where $Prj_{S^{n\times k}}$ is the minimum Frobenius norm projection into the the Steifel manifold of order k that can be easily computed via SVD as $\forall A \in \mathbb{R}^{n\times k}$:

$$Prj_{S^{n\times k}}(A) = UV^T \tag{49}$$

where $A = U\Lambda V^T$ is the SVD factorization.
The scheme defined in (48) is one of the most classical example of an update non-preserving scheme. On the other hand, one might be interested in an update preserving scheme, e.g. an optimization scheme that remains inside the manifold at every iteration provided an initial point that lies inside the manifold.
Update preserving schemes are a case of direct interest for optimization on the Stiefeld manifold and are performed mainly via Cayley transform.
Differently from the standard projected gradient descent scheme (48), update preserving schemes moves "zig-zagging" along a geodesic.

In Figure 4 we can see the difference between the projected gradient descent (blue arrow) and the update preseving scheme (red path) in a standard situation.
Following the work in [9], the starting point for an update preserving scheme on the Stiefel manifold is the projected gradient descent.
Using Lagrange multipliers or classical differential geometry arguments (see [9] and [8] for an exhaustive treatment) is easy to see that:

$$Prj_{T_U S^{n\times k}} \left(\nabla_U f(U)\right) = H(U)\, U, \quad \forall U \in S^{n\times k} \tag{50}$$

where:

$$H(U) = \nabla_U f(U)\, U^T - U\ \nabla_U f(U)^T$$

and clearly $H(U) = -H(U)^T$.

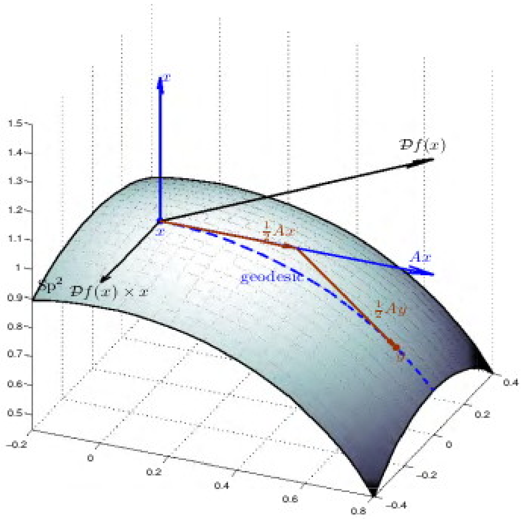


Figure 4: Projected Gradient (blue arrow) vs Update Preserving (red path).

As a consequence, the ODE in (47) rewrites as:

$$\dot{U} = -H(U)U \tag{51}$$

Observe now that if we use the Crank-Nicolson like scheme (or the mid point rule) to discretize (51) we have:

$$U_{s+1} = U_s - dt\, H\left(\frac{U_{s+1} + U_s}{2}\right)\left(\frac{U_{s+1} + U_s}{2}\right) \tag{52}$$

Rearranging (52) we obtain the following updating scheme:

$$U_{s+1} = \left(\mathbb{1} + \frac{dt}{2}\, H\left(\frac{U_{s+1} + U_s}{2}\right)\right)^{-1}\left(\mathbb{1} - \frac{dt}{2}\, H\left(\frac{U_{s+1} + U_s}{2}\right)\right) U_s \tag{53}$$

Now, since $H\left(\frac{U_{s+1}+U_s}{2}\right) = -H\left(\frac{U_{s+1}+U_s}{2}\right)^T$, by the Cayley Transform Theorem we have that if $U_s^T U_s = \mathbb{1}$ then $U_{s+1}^T U_{s+1} = \mathbb{1}$.
The scheme in (53) can be further simplified as:

$$U_{s+1} = \left(\mathbb{1} + \frac{dt}{2}\, H\left(U_s\right)\right)^{-1}\left(\mathbb{1} - \frac{dt}{2}\, H\left(U_s\right)\right) U_s \tag{54}$$

and also in this case we have an update preserving scheme.
In [9] it is shown that the scheme defined in (54) is a descent direction scheme providing an opportune line search and time stepping policy. We will use this technique in our numerical solver.

5.2 Numerical Solution

Consider our regularized optimization problem (44) and call:

$$f(U,\nu) = \frac{1}{2}U : \Delta U + \frac{\lambda}{2}U : (\nu_\delta)^{-*} \odot U + \frac{\lambda}{2}\nu : \mathbb{e}^{n\times k} \tag{55}$$

with the same notations as in (44). In order to use the minimization scheme defined in (54) we need to compute the gradient of (55) with respect to U in the frobenius norm and the projection of the gradient into the Stiefel tangent space. It easy to see that:

$$\begin{gathered} \nabla_U f(U,\nu) = \Delta U + \lambda(\nu_\delta)^{-*} \odot U \\ H(U) = \nabla_U f(U,\nu)U^T - U\,\nabla_U f(U,\nu)^T \end{gathered} \tag{56}$$

Therefore our discretized DMK scheme for problem (44) becomes:

$$\begin{cases} \left(\mathbb{1} + \dfrac{dt}{2} H\left(U_s\right)\right) U_{s+1} = \left(\mathbb{1} - \dfrac{dt}{2} H\left(U_s\right)\right) U_s \\ \nu_{s+1} = \nu_s + dt\,\left((\nu_\delta)_s^{-*} \odot U_s \odot U_s - \nu_s\right) \\ U_0^T U_0 = \mathbb{1} \\ \nu_0 > 0 \end{cases} \tag{57}$$

For our numerical example we consider the cyclic $1D$-graph with n=128 nodes $G = (V,E)$, $V = \{v_1,..,v_n\}$ and $E = \begin{Bmatrix} (v_i,v_{i+1}) & i = 1,..,n-1 \\ (v_n,v_1) & i = n \end{Bmatrix}$.

In Figure 5 we can see the comparison between some of the original cyclic $1D$-graph Laplacian matrix eigenvectors (the classical periodic sinusoidal functions) and some of the compressed modes computed with our DMK scheme for $k = 20$.

In Figure 6 we can see the comparison between the original k smallest laplacian eigenvalues and the k compressed eigenvalues i.e. the k smallest eigenvalues of the matrix $U^T\Delta U$ for $k = 20$.

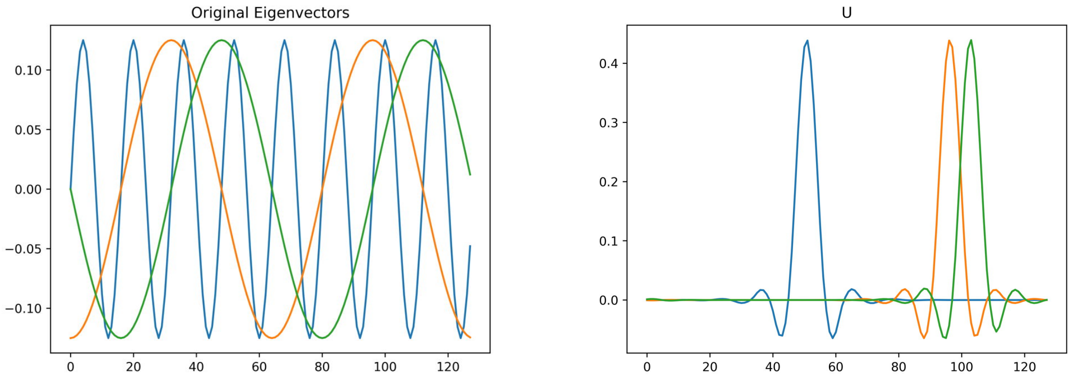


Figure 5: Original Eigenvectors (left), Compressed Eigenvectors (right).

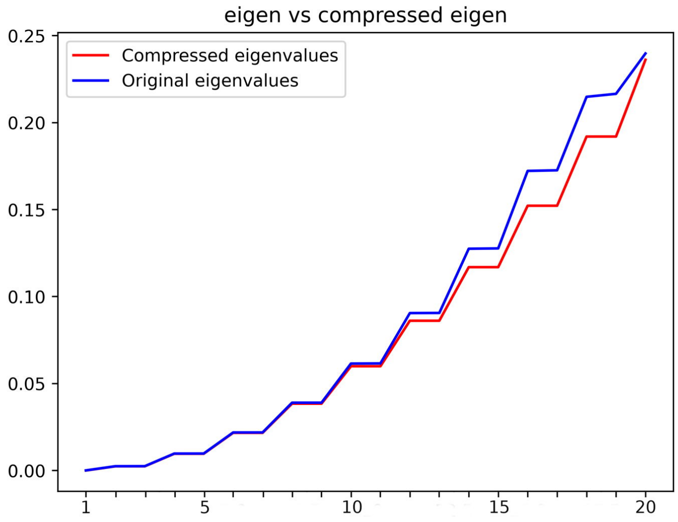


Figure 6: Original eigenvalues vs Compressed eigenvalues.

References

[1] Guy Bouchitté, Giuseppe Buttazzo, and Pierre Seppecher, *Mathématiques/mathematics shape optimization solutions via Monge-Kantorovich equation.* Comptes Rendus de l'Académie des Sciences - Series I-Mathematics, 324/10 (1997), 1185–1191.

[2] Michel Fortin, Daniele Boffi and Franco Brezzi, "Mixed Finite Element Methods and Applications". Springer, 1991.

[3] Ivar Ekeland and Roger Témam, "Convex Analysis and Variational Problems". Society for Industrial and Applied Mathematics, 1999.

[4] Lawrence C. Evans and Ronald E. Gariepy, "Measure theory and fine properties of functions". Textbooks in Mathematics. CRC Press, New York, 2015.

[5] Enrico Facca, Federico Piazzon, and Mario Putti, *Computing the l1 optimal transport density: a fem approach.* In Preparation, 2022.

[6] Enrico Facca, Federico Piazzon, and Mario Putti., *l1 transport energy.* Applied Mathematics & Optimization, 86/2 (2022), 1–40.

[7] Vidvuds Ozoliņš, Rongjie Lai, Russel Caflisch, and Stanley Osher, *Compressed modes for variational problems in mathematics and physics.* Proceedings of the National Academy of Sciences, 110/46 (oct 2013), 18368–18373.

[8] Hemant D. Tagare, *Notes on optimization on Stiefel manifolds.* 2011.

[9] Wen Zaiwen and Yin Wotao, *A feasible method for optimization with orthogonality constraints.* Mathematical Programming (dec 2013).

Micro-Macro limit: from the Follow-the-Leader model to the Lighthill-Whitham-Richards model

Mohamed Bentaibi [(*)]

1 Introduction

The Follow-the-Leader model (FtL) is a dynamical system describing the motion of N cars on a road lane, in which each car travels with a velocity that depends on its relative distance with respect to the one immediately in front. The Lighthill-Whitham-Richards (LWR) model is a hyperbolic conservation law where the solution is a macroscopic density that typically represents the dynamics of the average spatial concentration of vehicles.

With the FtL model we build a microscopic density which approximates the macroscopic one. Our main goal is to prove that the microscopic density converges to the macroscopic one. This occurs under suitable hypotheses on the dynamics, as well as strong convergence requests on the initial data. Additional stability results of the FtL model are also presented.

Let us first briefly describe the FtL model. We consider $N+1$ cars on a one-dimensional road lane. Let $\{x_j^N(0)\}_{j=0}^N$ denote the initial positions of the cars evolving in time according to the FtL dynamics. We have a trajectory $\{x_j^N(t)\}_{j=0}^N$, where each $x_j^N(t)$ travels with a velocity depending on the distance with respect to the car immediately in front of it $x_{j+1}^N(t)$. The leader $x_N^N(t)$ has no cars in front and thus it travels with the maximum velocity $v_{\max}$. A discrete function $\rho^{E,N}$, that we denote by *Eulerian microscopic density*, composed of $N \in \mathbb{N}$ regions each of mass $1/N$ is then defined by $\{x_j^N\}_{j=0}^N$ as

$$\rho^{E,N}(t,x) \coloneqq \sum_{j=0}^{N-1} \frac{1/N}{x_{j+1}^N(t) - x_j^N(t)} \chi_{[x_j^N(t), x_{j+1}^N(t))}(x) \qquad x \in \mathbb{R},\ t \geq 0. \tag{1.1}$$

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: `bentaibi@math.unipd.it`. Seminar held on 1 March 2023.

We now briefly describe the classical LWR model

$$\begin{cases} \rho_t + (f(\rho))_x = 0, \quad t > 0 \quad x \in \mathbb{R} \\ \rho(0,x) = \bar{\rho}(x) \quad x \in \mathbb{R} \end{cases} \tag{1.2}$$

with $\bar{\rho}$ a given initial data with compact support. The variable ρ describes a macroscopic density of cars, and the flux $f(\rho)$ at a point $x \in \mathbb{R}$ represents the number of cars passing through the given point $x \in \mathbb{R}$ per unit of time. We consider the following flux from now on:

$$f(\rho) \coloneqq \rho v(\rho).$$

We also assume that the maximal admissible density is $\rho_{max} \coloneqq 1$ and that $\|\bar{\rho}\|_{L^1(\mathbb{R})} = 1$.

From now on, we assume that the velocity function $v(\rho)$ satisfies the following assumptions:

$$v \in \mathrm{Lip}([0, \rho_{\mathrm{max}}]), \qquad v(\rho_{\mathrm{max}}) = 0, \qquad v \text{ decreasing.} \tag{V1}$$

We also use the notation $v_{max} := v(0)$ from now on. In some instances, we also make use of an additional assumption on the velocity:

$$\text{the map } [0, +\infty) \ni \rho \mapsto \rho v'(\rho) \in [0, +\infty) \text{ is non-increasing.} \tag{V2}$$

Informally, the main question is the following: does the Eulerian microscopic density built from the FtL converge to the “right" macroscopic density solution to the LWR?

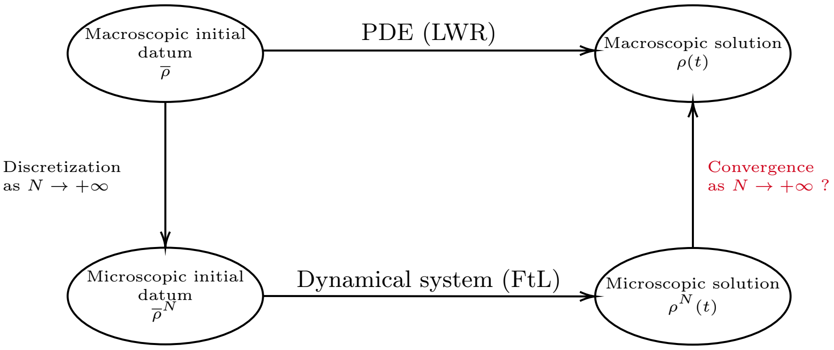

Macroscopic initial datum $\bar{\rho}$
PDE (LWR)
Macroscopic solution $\rho(t)$
Discretization as $N \to +\infty$
Convergence as $N \to +\infty$?
Microscopic initial datum $\bar{\rho}^N$
Dynamical system (FtL)
Microscopic solution $\rho^N(t)$

Figure 1: Problem statement.

These notes are organized as follows. In Section 2 we present the Follow-the-Leader dynamics. In Section 3 we present the Lighthill-Whitham-Richards model. In Section 4 we state the micro-to-macro result as well as a stability result from [1].

2 The Follow-the-Leader model

We start by considering $N+1$ vehicles with initial positions $\bar{x}_0^N < \dots < \bar{x}_N^N$ satisfying $\bar{x}_{i+1}^N - \bar{x}_i^N \geq l$ where

$$l := \frac{1}{N}$$

is the length of vehicles. This standard condition ensures non overlapping of cars.

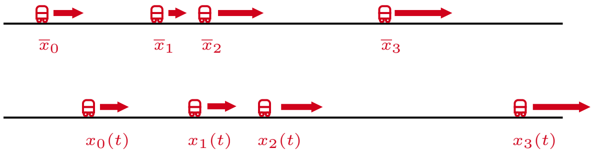


Figure 2: Illustration of cars moving according to the FtL for $N = 4$.

We now define the FtL dynamics:

Definition 2.1 The FtL model is

$$\begin{cases} \dot{x}_N^N = v_{max} \\ \dot{x}_j^N = v\left(\frac{l}{x_{j+1}^N - x_j^N}\right) \qquad j = 0, \dots, N-1 \\ x_i^N(0) = \bar{x}_j^N. \end{cases} \tag{2.1}$$

The FtL model describes the evolution of each car x_j^N that adapts its speed with respect to the distance with the car immediately in front x_{j+1}^N. We can build the Eulerian discrete density, which can be understood as a discrete approximation of the solution of the LWR model (1.2). Although it has been already presented in the introduction, we give the definition here.

Definition 2.2 Given $\{x_j^N(t)\}_{j=0}^N$ solution of (2.1), define the Eulerian discrete density as

$$\rho^{E,N}(t,x) := \sum_{j=0}^{N-1} \frac{1/N}{x_{j+1}^N(t) - x_j^N(t)} \chi_{[x_j^N(t), x_{j+1}^N(t))}(x) \qquad x \in \mathbb{R},\, t \geq 0. \tag{2.2}$$

Remark 2.1 (Discrete Minimum/Maximum Principle) The solution of the FtL model (2.1) and the corresponding Eulerian discrete density (2.2) satisfy a discrete minimum/maximum principle, corresponding to the well-known maximum principle for (1.2), see for example [2, Theorem 6.2.4]. Indeed, the following estimates hold:

$$\begin{aligned} &\min_{j=0,\dots,N-1} |x_{j+1}(t) - x_j(t)| \geq \min_{j=0,\dots,N-1} |\bar{x}_{j+1} - \bar{x}_j| \geq l; \\ &\rho^{E,N}(t,x) \leq \left\|\rho^{E,N}(0)\right\|_{L^\infty(\mathbb{R})} \leq 1 \qquad \forall x \in \mathbb{R},\, t \geq 0. \end{aligned} \tag{2.3}$$

A proof can be found in [3, Lemma 1].

3 The Lighthill-Whitham-Richards Model

Consider ρ as a macroscopic density of cars, i.e.

$$\rho(t,x) = \text{density of cars at a position } x \text{ of the road at time } t$$

and $f(\rho) \coloneqq \rho v(\rho)$ as the flux, i.e.

$$f(\rho(t,x)) = \frac{\text{total number of cars crossing point } x}{\text{unit time}}.$$

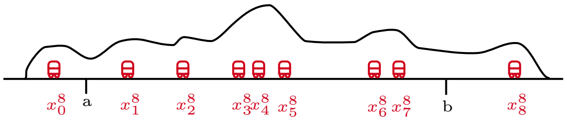


Figure 3: Illustration of density of cars.

By conservation of the number of cars in $[a,b]$,

$$\frac{d}{dt}\int_a^b \rho(t,x)dx = \text{flux of cars entering at a} - \text{flux of cars exiting at b} = -\int_a^b (\rho v(\rho))_x dx.$$

This is valid for every interval $[a,b]$, and we thus get the classical Lighthill-Whitham-Richards (LWR) model (1.2) with $\bar{\rho}$ a given initial data. Assume $\|\bar{\rho}\|_{L^1(\Omega)} = 1$ with maximal admissible density $\rho_{max} \coloneqq 1$. Assume that the velocity satisfies the same assumptions as in (2.1).

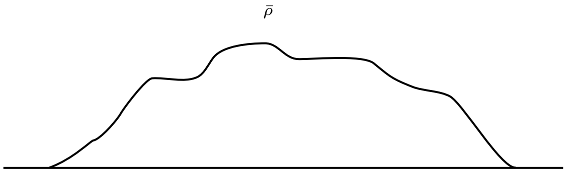


Figure 4: Initial data with compact support.

Even for smooth initial data, the solution of the Cauchy problem (1.2) may develop discontinuities in finite time due to the nonlinearlity of the flux, as see in Figure 5.

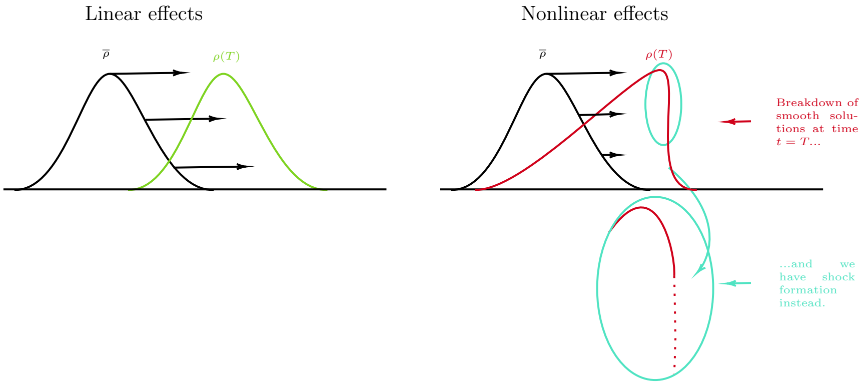

Linear effects
$\bar{\rho}$
$\rho(T)$
Nonlinear effects
$\bar{\rho}$
$\rho(T)$
Breakdown of smooth solutions at time $t = T$...
...and we have shock formation instead.

Figure 5: Linear vs. nonlinear effects.

Due to the nonlinear effects, we need to find a weaker notion of solution.

Definition 3.1 A function $\rho \in L^\infty((0,+\infty) \times \mathbb{R})$ is a weak (distributional) solution to (1.2) if it holds

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} [\rho(t,x)\varphi_t(t,x) + f(\rho(t,x))\varphi_x(t,x)]\, \mathrm{d}t\ \mathrm{d}x + \int_{\mathbb{R}} \bar{\rho}(x)\varphi(0,x)\mathrm{d}x = 0$$

for all $\varphi \in C_c^\infty([0,+\infty) \times \mathbb{R})$.

However, with weak solutions we lose uniqueness, and we need some mechanism to retrieve the solution that “physically makes sense". This is done by using the concept of entropy.

Theorem 3.1 (Kružkov's entropy solution) *Assume that the flux $f(\rho)$ is locally Lipschitz. For any given initial data $\bar{\rho} \in L^\infty$ with compact support, there exists a unique entropy solution $\rho \in L^\infty([0,+\infty)] \times \mathbb{R})$, i.e. it satisfies the entropy inequality*

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} [|\rho(t,x) - k|\varphi_t(t,x) + \operatorname{sign}(\rho(t,x) - k)[f(\rho(t,x)) - f(k)]\varphi_x(t,x)]dtdx$$
$$+ \int_{\mathbb{R}} |\bar{\rho}(x) - k|\varphi(0,x)dx \geq 0$$

for all $\varphi \in C_c^\infty([0,+\infty) \times \mathbb{R})$ with φ non-negative and for all constants $k \in \mathbb{R}$.

4 Microscopic to Macroscopic

Now that we have a microscopic density $\rho^{E,N}$ and a macroscopic density ρ, we would like to answer the following question: which properties of the initial data and/or of the convergence of the discretized system ensure convergence of the microscopic density $\rho^{E,N}$ to the macroscopic one ρ?

More precisely: let an initial configuration $\{x_j^N(0)\}_{j=0}^N$ be given, subject to the FtL dynamics. Build the corresponding microscopic density $\rho^{E,N}$ and consider the discrete approximation sequence $\left\{\rho^{E,N}\right\}_{N\in\mathbb{N}}$. Does this sequence converge to the solution ρ of the Cauchy problem (1.2)? In what topology and how arbitrary can the initial positioning be? Before introducing the result, we first present a fundamental condition on the support at initial time of $\rho^{E,N}$.

Definition 4.1 (Uniformly bounded initial support condition) We say that $\{x_j^N(t)\}_{j=0}^N$ satisfies the condition of the uniformly bounded initial support (4.1) if there exists a constant $K > 0$ independent of N, such that for all $N \in \mathbb{N}$ it holds

$$x_N^N(0) - x_0^N(0) < K. \tag{4.1}$$

Theorem 4.1 *Let $\bar{\rho} \in L^\infty(\mathbb{R})$ with compact support. Let v satisfy (V1). Let $\{x_j^N(t)\}_{j=0}^N$ follow the FtL dynamics, indexed by $N \in \mathbb{N}$, that satisfy the condition of the uniformly bounded initial support (4.1). Consider the corresponding Eulerian density $\rho^{E,N} \in L^\infty([0,+\infty)\times\mathbb{R})$ defined by (2.2). Let*

$$\rho^{E,N}(0) \rightharpoonup \bar{\rho}. \tag{4.2}$$

We have that

(1) *if $\bar{\rho}$ is of bounded variation and there exists $K > 0$ independent of N such that* $\mathrm{Tot.Var.}(\rho^{E,N}(0);\Omega) < K$, *i.e.*

$$\frac{1}{N}\left(\frac{1}{x_1^N(0)-x_0^N(0)} + \frac{1}{x_N^N(0)-x_{N-1}^N(0)} + \sum_{j=0}^{N-2}\left|\frac{1}{x_{j+2}^N(0)-x_{j+1}^N(0)} - \frac{1}{x_{j+1}^N(0)-x_j^N(0)}\right|\right) < K,$$

for all $N \in \mathbb{N}$, or

(2) *if assumption (V2) holds,*

then the sequence $\left\{\rho^{E,N}\right\}_{N\in\mathbb{N}}$ converges strongly to the weak entropy solution ρ of the Cauchy problem (1.2) in $L^1_{\mathrm{loc}}([0,+\infty)\times\mathbb{R})$ as $N\to\infty$.

Before presenting a stability result, we briefly introduce the Wasserstein distance. Given a certain transportation cost, how can we transport μ to ν in an optimal way, i.e. minimizing the total transportation cost?

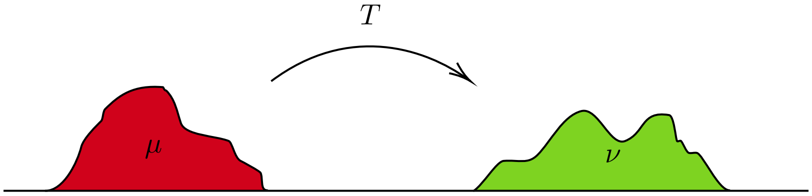


Figure 6: Transporting μ to ν.

Consider the cost $c(x,y) = |x-y|$. The Wasserstein distance W_1 is defined as

$$W_1(\mu,\nu) := \text{the minimal transport cost from } \mu \text{ to } \nu \text{ for the cost } c.$$

A fundamental question is the following: do we know how this map lookds like? In 1D definitely yes! There exists a unique nondecreasing T_{mon} such that

$$W_1(\mu,\nu) = \int_{\mathbb{R}} |x - T_{\text{mon}}(x)|dx \qquad T_{\text{mon}\,\#}\mu = \nu.$$

We are now ready to present a stability result. It is a microscopic stability result with respect to two different initial discretization schemes, uniformly in time.

Theorem 4.2 *Let $\Omega \subset \mathbb{R}$ be bounded. Assume that v satifies (V1). Let $\{x_j^N(t)\}_{j=0}^N$,$\{\tilde{x}_j^N(t)\}_{j=0}^N$ be solutions of the FtL, indexed by $N \in \mathbb{N}$ that satisfy the condition of the uniformly bounded initial support (4.1). Consider the corresponding discrete densities $\rho^{E,N}$, $\tilde{\rho}^{E,N} \in L^\infty([0,+\infty) \times \mathbb{R})$ defined by (1.1). If it holds*

$$\rho^{E,N}(0), \tilde{\rho}^{E,N}(0) \rightharpoonup \bar{\rho}$$

and

$$\lim_{N\to+\infty} \sum_{j=0}^{N-1} |x_{j+1}(0) - x_j(0) - (\tilde{x}_{j+1}(0) - \tilde{x}_j(0))| = 0,$$

then for all $T > 0$ it holds

$$\lim_{N\to+\infty} \sup_{t\in[0,T]} W_1(\rho^{E,N}(t), \tilde{\rho}^{E,N}(t)) = 0.$$

As a consequence,

(1) *if* $\text{Tot.Var.}\left(\rho^{E,N}(0);\mathbb{R}\right), \text{Tot.Var.}\left(\tilde{\rho}^{E,N}(0);\mathbb{R}\right) \leq K$ *for some $K > 0$ independent of N, then for all $T > 0$ it holds*

$$\lim_{N\to+\infty} \sup_{t\in[0,T]} \left\|\rho^{E,N}(t) - \tilde{\rho}^{E,N}(t)\right\|_{L^1(\Omega)} = 0,$$

(2) *if assumption (V2) holds, then for all $\delta, T > 0$ it holds*

$$\lim_{N\to+\infty} \sup_{t\in[\delta,T]} \left\|\rho^{E,N}(t) - \tilde{\rho}^{E,N}(t)\right\|_{L^1(\Omega)} = 0.$$

References

[1] Fabio Ancona, Mohamed Bentaibi, and Francesco Rossi, *On the stability of the many particle limit of the follow-the-leader model and convergence to the nonlinear scalar conservation law.* In preparation.

[2] Constantine M. Dafermos, "Hyperbolic conservation laws in continuum physics". Grundlehren vol. 325, Springer, 2000.

[3] Marco Di Francesco and Massimiliano D. Rosini, *Rigorous derivation of nonlinear scalar conservation laws from follow-the-leader type models via many particle limit.* Archive for Rational Mechanics and Analysis 217/3 (2015), 831–871.

On the volume of (half-)tubular neighborhoods of surfaces in sub-Riemannian geometry

Tania Bossio [(*)]

Abstract. In 1840 Steiner proved that the volume of the tubular neighborhood of a convex set in $\mathbb{R}^n$ is a polynomial of degree n in the size of the tube. The coefficients of such a polynomial carry information about the curvature of the set. In this talk we present Steiner-like formulas in the framework of sub-Riemannian geometry. In particular, we introduce the three-dimensional sub-Riemannian contact manifolds, which the first Heisenberg group is a special case of. Then, we show the asymptotic expansion of the volume of the half-tubular neighborhood of a surface and provide a geometric interpretation of the coefficients in terms of sub-Riemannian curvature objects.

1 Introduction

Sub-Riemannian geometry can be seen as a generalization of Riemannian geometry. Namely, sub-Riemannian manifolds are differential manifolds in which a particle moves with constraints on its velocity.

In other words, we refer to metric spaces in which the distance between points is measured considering the infimum over the "Riemannian" length of curves connecting the points and with the property that the velocity lies in an assigned subspace of preferable directions in the tangent plane at every point.

In this note we want to introduce the reader to the sub-Riemannian world considering the problem of finding a generalization of the Steiner formula. More precisely, we are interested in computing the volume of a half-tubular neighborhood of a surface embedded in a the three-dimensional contact manifold. This setting is a particular case of general sub-Riemannian manifolds, which the Heisenberg group is the most studied case of.

Moreover, we give a geometrical interpretation of the coefficients of the asymptotic expansion of the volume of the half-tube, with respect to the radius, in terms of curvature objects. The statements we obtain in [4] are a generalization of previously known results in the Heisenberg group (cf. [2, 6]).

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: `tania.bossio@math.unipd.it`. Seminar held on 15 March 2023.

The note is organized as follows. We start recalling the classical Weyl and Steiner formulas in the Euclidean setting. Then, we describe the geometry of a three-dimensional contact sub-Riemannian manifold, exploiting in detail the Heisenberg case. Finally, we present the new results contained in [4].

2 The euclidean setting

Let S be a compact hypersurface in $\mathbb{R}^n$. The *tubular neighborhood* of radius ε around S is defined as the set of points

$$T_\varepsilon = \{x \in \mathbb{R}^n \mid |x-y| < \varepsilon \quad \forall y \in S\}.$$

Weyl derived the following expression for the (Lebesgue) volume of T_ε

$$\mathrm{Vol}^{\mathbb{R}^n}(T_\varepsilon) = 2 \sum_{1\le 2k+1 \le n} C_{2k+1}(S)\varepsilon^{2k+1}.$$

The formula is a polynomial of degree n in the radius ε and the coefficients are integrals of certain curvature functions. In particular,

$$C_1(S) = \int_S dA, \qquad C_3(S) = \int_S K dA,$$

where K is the scalar curvature of S (the product of the principal curvatures of S), while dA denotes the induced surface measure. Moreover, the coefficients do not depend on how S is (isometrically) embedded in $\mathbb{R}^n$. The last fact is also the reason why in the formula there appear only coefficients of odd degree. Indeed, when the hypersurface is oriented, then the tube T_ε is split in in two *half-tubes*, i.e., $T_\varepsilon = S_\varepsilon^+ \cup S_\varepsilon^-$. When computing the volume of these subsets we obtain that

$$\mathrm{Vol}^{\mathbb{R}^n}(S_\varepsilon^\pm) = \sum_{1\le 2k+1\le n} C_{2k+1}(S)\varepsilon^{2k+1} \pm \sum_{1\le 2k\le n} C_{2k}(S)\varepsilon^{2k}.$$

This formula is due to Steiner and the coefficients of even degree carry information about the embedding of S in $\mathbb{R}^n$.

When specializing the formula in the case of a surface in $\mathbb{R}^3$ we obtain that

$$\mathrm{Vol}^{\mathbb{R}^3}(S_\varepsilon^+) = \varepsilon \mathrm{Area}(S) - \varepsilon^2 \int_S H\, dA + \frac{\varepsilon^3}{3}\int_S K\, dA, \tag{1}$$

where H is the mean curvature of S, computed with respect to the orientation of the surface given by the inward pointing normal of the surface S with respect to S_ε^+, and K is the Gaussian curvature of S. We refer the reader to the monograph of A. Gray [5] for an exhaustive overview of this subject.

3 Three-dimensional sub-Riemannian contact manifolds

In this section we present the three-dimensional sub-Riemannian contact manifolds, that are a particular case of the general sub-Riemannian manifolds. They are three-dimensional metric spaces in which a particle is constraint to move in a subspace generated by two fixed independent directions at each point. The Heisenberg group is the most known example of such type of manifolds. We refer the interested reader to [1] for a general overview of sub-Riemannian geometry and in particular to [1, §17.2] for the specific case of three-dimensional sub-Riemannian contact manifolds.

Definition 1 A three-dimensional sub-Riemannian contact manifold is a triple $(M, \mathcal{D}, g)$ where

- M is a three-dimensional smooth manifold,
- $\mathcal{D}$ is a rank two distribution in TM, locally defined as the kernel of a smooth one-form ω such that $\omega \wedge \mathrm{d}\omega \neq 0$,
- g is a smooth metric defined on $\mathcal{D}$ such that $\mathrm{vol}_g = \mathrm{d}\omega|_{\mathcal{D}}$.

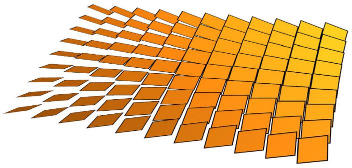


Figure 1: Courtesy of E. Le Donne, *Lecture notes on sub-Riemannian Geometry*, 2010.

The Most important consequence of the last definition is that the distribution satisfies the *Hörmander condition*. More precisely, the Lie algebra generated by $\mathcal{D}$ fills the whole tangent space ($\mathcal{L}\mathrm{ie}(\mathcal{D}) = TM$). In the case of a contact sub-Riemannian distribution the Lie bracket generating (Hörmander) condition specifies as follows: for any pair of independent vector fields $X, Y \in \mathcal{D}$, the Lie bracket between them $[X, Y]$ is never tangent to $\mathcal{D}$. Since to fill the whole tangent space to M at each point it suffices a first order Lie bracket relation, we say that the contact structure is of step 2.

Definition 2 Among all the transversal vector fields to $\mathcal{D}$ there is the *Reeb vector field* Z defined as the unique vector field such that

$$\omega(Z) = 1, \qquad \mathrm{d}\omega(Z, \cdot) = 0.$$

If we consider an orthonormal frame X, Y on $\mathcal{D}$ such that $\mathrm{d}\omega(X, Y) = -1$, we obtain the following bracket relations:

$$[X,Y] = c^X_{YX}X + c^Y_{YX}Y + Z \tag{0.1}$$

$$[X,Z] = c^X_{XZ}X + c^Y_{XZ}Y \tag{0.2}$$

$$[Y,Z] = c^X_{YZ}X + c^Y_{YZ}Y \tag{0.3}$$

for c^i_{jk} smooth functions on M.

Remark 3 Notice that the Reeb vector field Z appears only in the bracket relation between horizontal vector fields in $\mathcal{D}$. This fact is due to the general Cartan "magic" formula. In the case of a smooth 1-form ω and two vector fields V, W on a manifold, the Cartan formula explicates as follows:

$$\mathrm{d}\omega(V,W) + \omega[V,W] = V(\omega(W)) - W(\omega(V)).$$

In our case, the conclusion holds from the properties of the contact form ω.

3.1 The Three-dimensional Heisenberg group

The most important example of three-dimensional sub-Riemannian contact manifold is the three-dimensional *Heisenberg group* $\mathbb{H}$, that is defined as $\mathbb{R}^3$ equipped with the contact form

$$\omega = dz + \frac{1}{2}(ydx - xdy). \tag{2}$$

The horizontal distribution $\mathcal{D}$ is generated by the following orthonormal frame

$$X = \partial_x - \frac{y}{2}\partial_z, \qquad Y = \partial_y + \frac{x}{2}\partial_z,$$

where (x,y,z) are the standard coordinates in $\mathbb{R}^3$.

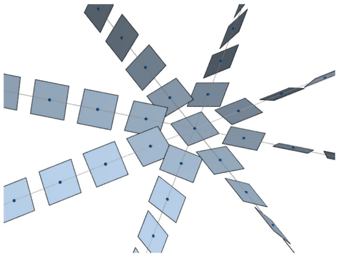


Figure 2: Courtesy of P. Pansu, *Géométrie du groupe d'Heisenberg*, 2018.

The Reeb vector field is obtained considering the only non trivial bracket relation

$$Z = [Y,X] = \partial_z.$$

3.2 The sub-Riemannian distance function

Definition 4 The sub-Riemannian distance between two fixed points $x, y \in M$ is defined as

$$d_{SR}(x,y) = \inf \left\{ \int_0^T g(\dot\gamma, \dot\gamma)^{\frac{1}{2}}\, dt \mid \gamma \text{ admissible for } (x,y) \right\} \tag{3}$$

where *admissible* for (x,y) means that $\gamma : [0,T] \to M$ is an absolute continuous curve joining the points and such that it is an horizontal curve, i.e., such that

$$\dot\gamma(t) \in \mathcal{D} \text{ for almost every } t \in [0,T].$$

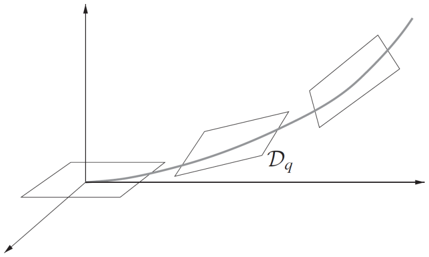


Figure 3: Courtesy of [1].

Notice that the length of an horizontal curve is invariant by reparametrization.

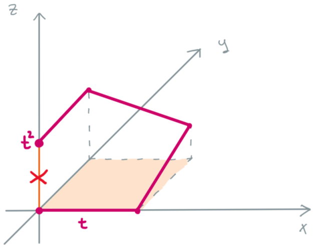
Since the horizontal distribution at every point is a subspace of the tangent plane, the set of the admissible curves is generally smaller that in the Riemannian setting, where the constrain on the velocity is trivial, i.e., $distr = TM$. At this point, the natural question is whenever d_{SR} is a finite distance. Are we able to connect every pair of points in the manifold trough an horizontal admissible curve?

3.3 Reachable points in the Heisenberg group

Let us consider again the Heisenberg group $\mathbb{H}$ as defined in Subsection 3.1. We showed that the Reeb vector field $Z = \partial_z$ is never belonging to the horizontal distribution. One can ask if it is possible to connect the origin to a point on the z-axis. In order to answer the question we present the following construction.

For a fixed $t > 0$ in $\mathbb{R}$, we define the curve $\gamma : [0, 4t] \to \mathbb{H}$ as

$$\gamma(s) = \begin{cases} (s, 0, 0) & s \in [0,t] \\ (t, s-t, \frac{t}{2}(s-t)) & s \in [t, 2t] \\ (t-(s-2t), t, \frac{t^2}{2} + \frac{t}{2}(s-2t)) & s \in [2t, 3t] \\ (0, t-(s-3t), t^2) & s \in [3t, 4t] \end{cases}. \tag{4}$$



The curve γ connects the origin to the point $(0, 0, t^2)$ and it is horizontal because the velocity belongs to $\mathcal{D}$ at any time. More precisely,

$$\dot{\gamma}(s) = \begin{cases} X_{\gamma(s)} & s \in [0, t] \\ Y_{\gamma(s)} & s \in [t, 2t] \\ -X_{\gamma(s)} & s \in [2t, 3t] \\ -Y_{\gamma(s)} & s \in [3t, 4t] \end{cases}.$$

Notice that the defined path measures how much the two vector fields X, Y do not commute. Moreover, $\gamma(4t) = \exp(t^2 \partial_z)$, where exp is the usual fexponential map. As $t \to 0$ we have that $[X, Y] = Z$ is the second order approximation of $\dot{\gamma}$.

3.4 The Rashevskii-Chow Theorem

The Hörmander condition is the key hypothesis that makes (M, d_{SR}) a proper metric space.

Theorem 5 *Let $(M, \mathcal{D}, g)$ be a three-dimensional sub-Riemannian contact manifold, then*

- *(M, d_{SR}) is a metric space;*
- *The metric topology is equivalent to the ambient topology.*

In particular, $d_{SR} : M \times M \to \mathbb{R}$ is continuous.

Actually, this result holds for every sub-Riemannian manifold. We refer the reader to Theorem 3.31 in [1] for more details.

Remark 6 The sub-Riemannian distance from a point in a contact sub-Riemannian manifold is Holder continuous of exponent $\frac{1}{2}$ (cf. [1]). The following figure shows the sub-Riemannian ball centered at the origin in the Heisenberg group. It looks like an apple with singularities at the poles.

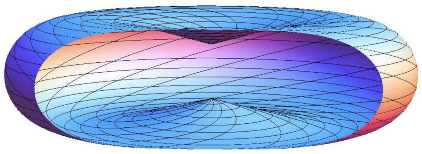


Figure 4: Courtesy of [1].

3.5 Length minimizers in the Heisenberg group

The curve in (4) it is not optimal in the sense that it does not realize the distance between its end-points. We recall that in the Heisenberg group a curve is admissible if its velocity lies in the kernel of the contact 1-form ω in (2). In other words, a curve $t \mapsto (x(t), y(t), z(t))$ is an admissible path if and only if $\dot{z}(t) = \frac{1}{2}(y(t)\dot{x}(t) - x(t)\dot{y}(t))$ or equivalently

$$z(t) = z(0) + \frac{1}{2}\int_0^T [y(t)\dot{x}(t) - x(t)\dot{y}(t)]\, dt.$$

If the starting point of the path is the origin, then $z(t)$ is the signed area of the domain bounded by the curve and the segment connecting $(0,0)$ with $(x(T), y(T))$. In this geometry, the length of an admissible tangent vector $(\dot{x}, \dot{y}, \dot{z}) \in \mathcal{D}$ is defined to be $\sqrt{\dot{x}^2 + \dot{y}^2}$, i.e., the length of the projection of the vector on the xy-plane. By construction, the sub-Riemannian length of an admissible curve in $\mathbb{H}$ is equal to the Euclidean length of its projection on the xy-plane.

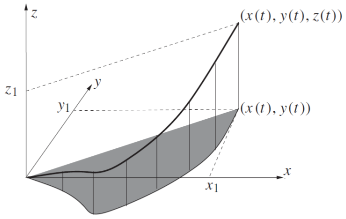


Figure 5: Courtesy of [1].

Therefore, in the Heisenberg group, to compute the shortest paths connecting the origin to a fixed point (x, y, z), we are reduced to solve the classical Dido isoperimetric problem. Namely, to find a shortest planar curve among those connecting $(0,0)$ to (x, y) and such that the signed area of the domain bounded by the curve and the segment joining the last points is equal to z. Solutions of the Dido problem are arcs of circles, and their lifts to $\mathbb{R}^3$ are spirals, where $z(t)$ is the area of the piece of disk cut by the chord connecting $(0,0)$ with $(x(t), y(t))$. A piece of such a spiral is a shortest admissible path between its endpoints while the planar projection of this piece is an arc of a circle. The spiral ceases to be a shortest path when its planar projection starts to run around the circle for a second time, i.e., when the spiral starts its second turn.

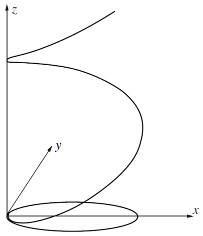


Figure 6: Courtesy of [1].

3.6 Length minimizers and the Hamiltonian description

In a three-dimensional contact sub-Riemannian manifold all length-minimizers arise as projections of a solution of an Hamiltonian equation (cf. [1, Prop. 4.8]). Namely, the Hamiltonian $H : T^*M \to \mathbb{R}$ related to the sub-Riemannian length minimization problem in a three-dimensional contact manifold is

$$H(\lambda) = H(p, x) = \frac{1}{2}\left[\langle p, X\rangle_x^2 + \langle p, Y\rangle_x^2\right],$$

where $p \in T_x^*M$ and $\langle\, ,\rangle$ denotes the duality pairing. Considering the canonical symplectic form $\sigma \in \Lambda^2(T^*M)$, the induced Hamiltonian vector field $\vec{H}$ on T^*M is defined by $\mathrm{d}H(\cdot) = \sigma(\cdot, \vec{H})$. Then, the Hamilton equation is

$$\dot{\lambda} = \vec{H}(\lambda). \tag{5}$$

Given a initial covector $\lambda_0 \in T^*M$, the unique solution $\lambda(t) = e^{t\vec{H}}(\lambda_0)$ to (5) is called normal extremal. Moreover, $\gamma(t) = \pi(\lambda(t))$, where $\pi : T^*M \to M$ is the canonical projection on M, is a locally length minimizing curve parametrized with constant speed $\sqrt{2H(\lambda_0)}$. To obtain arclength parametrized length minimizers one has to consider solutions of (5) with $\lambda_0 \in H^{-1}(1/2)$.

4 Half-tubes

The aim of our work is to generalize the Steiner formula (1) in the setting of the sub-Riemannian geometry. In particular we consider surfaces embedded in three-dimensional sub-Riemannian contact manifolds and we generalize the results previous known in the Heiseberg group (cf. [2, 6]).

Definition 7 Let S be a smooth surface bounding a smooth closed region Ω in a three-dimensional sub-Riemannian contact manifold $(M, \mathcal{D}, g)$. The half-tubular neighborhood of radius ε of the surface S is the set

$$S_\varepsilon = \{x \in M \setminus \Omega \mid 0 < \delta(x) < \varepsilon\},$$

where $\delta : M \to \mathbb{R}$ is the sub-Riemannian distance from the surface defined as

$$\delta(x) = \inf_{y \in S} d_{SR}(x, y),$$

with d_{SR} is defined in (3).

Our goal is to compute the volume ν of the set S_ε, where ν is the natural volume associated to the three-form $\omega \wedge \mathrm{d}\omega$ induced by the contact 1-form ω. In order to do that we first have to consider the regularity properties of δ. In fact, δ is not smooth everywhere. Specifically, it is smooth in neighborhoods of the surface that do not contain the so called characteristic points of the surface.

Definition 8 A point $x \in S$ is a characteristic point if the distribution and the tangent plane to S conincide at the point, i.e., if $\mathcal{D}_x = T_xS$.

The set of the characteristic points is closed and of zero measure in S. Before stating the result for the regularity of δ, we recall that the horizontal gradient of a differentiable function $f : M \to \mathbb{R}$ is the unique vector field $\nabla_H f \in \mathcal{D}$ such that $\mathrm{d}f|_{\mathcal{D}}(\cdot) = g(\nabla_H \delta, \cdot)$. Considering the orthonormal frame X, Y for $\mathcal{D}$ we have the following expression

$$\nabla_H f = (Xf)X + (Yf)Y.$$

We are ready to state the regularity result for the distance function from the surface δ. The version presented here refers to [7, Thm. 3.7].

Theorem 9 *Let S be a smooth compact surface bounding a closed domain Ω in a three-dimensional contact sub-Riemannian manifold M. Moreover, suppose that S does not contain characteristic points. Then,*

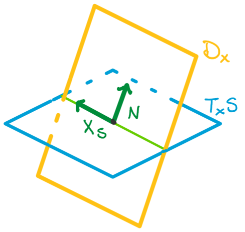
- $\delta : M \to \mathbb{R}$ *is* 1*-Lipschitz with respect to* d_{SR}*;*
- *There exists* $\varepsilon > 0$ *such that* $\delta : S_\varepsilon \to \mathbb{R}$ *is smooth;*
- *There exists a smooth diffeomorphism* $G : (0, \varepsilon) \times S \to S_\varepsilon$ *such that for all* $(t, p) \in (0, \varepsilon) \times S$
$$\delta\left(G(t,p)\right) = t \qquad \textit{and} \qquad dG(\partial_t) = \nabla_H \delta.$$

Moreover, on S_ε holds that $g(\nabla_H \delta, \nabla_H \delta)^2 = (X\delta)^2 + (Y\delta)^2 = 1$.

Hence, if the surface S does not contain characteristic points, then there exists a $\varepsilon > 0$ such that S_ε is a smooth domain in M. We proceed in describing the geometry of the surface out of the characteristic points, namely when the tangent plane to the surface is transversal to the distribution.

Definition 10 Let $x \in S$ be a non characteristic point. There are uniquely defined (up to a sign) the following unitary vector fields in $\mathcal{D}$:

- The *characteristic vector field* X_S is the vector field that generates $D_x \cap T_x S$;
- The *horizontal normal* N is the vector field orthogonal to X_S.



Let $f : M \to \mathbb{R}$ a locally defining function for S. Namely, S is locally defined as the zero level set of f with $\mathrm{d}f|_S \neq 0$. Then, we have the following expression for X_S and N

with respect to the frame X, Y in $\mathcal{D}$:

$$X_S = \frac{Yf}{\|\nabla_H f\|}X - \frac{Xf}{\|\nabla_H f\|}Y, \tag{0.4}$$

$$N = \frac{Xf}{\|\nabla_H f\|}X + \frac{Yf}{\|\nabla_H f\|}Y. \tag{0.5}$$

In particular, N is parallel to $\nabla_H f$.

In addition, outside of the characteristic points, we define the following sub-Riemannian curvature object relative to the surface.

Definition 11 The *sub-Riemannian mean curvature* of a surface S that does not contain characteristic points is the smooth function $\mathcal{H}$ defined in a neighborhood of S as

$$\mathcal{H} = -\mathrm{div}\left(\nabla_H \delta\right). \tag{6}$$

We obtain the following expression in terms of a local orthonormal frame X, Y in $\mathcal{D}$

$$\mathcal{H} = -XX\delta - YY\delta + c_{XY}^{Y}\left(X\delta\right) - c_{XY}^{X}\left(Y\delta\right).$$

We stress that (6) defines $\mathcal{H}$ not only on S but in a neighborhood of the surface. In particular, the derivative $N\mathcal{H}$ has a meaning.

The last ingredient we need for stating our main result is a smooth area form on the surface, the *induced sub-Riemannian area* defined on S as

$$dA = \iota_N(\omega \wedge \mathrm{d}\omega),$$

i.e. the 2-dimensional volume form on S obtained by restricting to N the volume associated to the contact structure induced by the form ω.

Considering the diffeomeorphism G in Theorem 9 and exploiting a sub-Riemannian version of the coarea formula, we obtain that

$$\nu(S_\varepsilon) = \int_0^\varepsilon \int_{G(t,S)} dA^t dt = \int_0^\varepsilon \int_S |\mathrm{d}_p G(t,p)| dA(p) dt,$$

where dA^t is the induced sub-Riemannian area form on the surface $G(t, S)$. This formula, which involves only sub-Riemannian quantities, can be obtained by its classical Riemannian counterpart. For the details we refer the interested reader to Proposition 41 and its proof in [4].

Computing the Taylor expansion with respect to t in 0 of $|\mathrm{d}_p G(t,p)|$, we obtain the following main result in [4] for the formula for the volume of the sub-Riemannian half-tubular neighborhood of a surface.

Theorem 12 *Let $(M, \mathcal{D}, g)$ be a contact three-dimensional sub-Riemannian manifold. Let S be a compact smooth surface that does not contain characteristic points and bounding a* *closed region Ω. The volume of the half-tubular neighborhood S_ε, is smooth with respect to ε and satisfies for $\varepsilon \to 0$*

$$\nu\left(S_{\varepsilon}\right)=\varepsilon \int_{S} dA-\frac{\varepsilon^{2}}{2} \int_{S} \mathcal{H} dA+\frac{\varepsilon^{3}}{6} \int_{S}\left(\mathcal{H}^{2}-N \mathcal{H}\right) dA+o(\varepsilon^{3}). \tag{7}$$

Here dA is the sub-Riemannian area measure on S, $\mathcal{H}$ is the mean sub-Riemannian curvature of S and N is the sub-Riemannian normal to the surface.

Remark 13 In contrast to the Euclidean case presented in (1), the formula (7) in Theorem 12 is not a polynomial in ε. In the Euclidean case the polynomial expansion is related to a specific choice of the volume, the Lebesgue one. In the sub-Riemannian case, even for the choice of the natural volume this is not true. Indeed, as proved in [2], in the first Heisenberg group $\mathbb{H}$ the volume of S_ε is analytic in ε.

Finally, to compute the coefficients of expansion (7) one a priori needs the knowledge of the explicit expression of δ, the sub-Riemannian distance, which in general is not possible. We provide a formula which permits to compute those coefficients only in terms of a function f locally defining S.

Before stating the formula, we introduce the Tanno connection, which is a canonical connection on contact manifold. We extensively exploited this technical tool for our computations

Definition 14 The Tanno connection ∇ is a linear metric connection (i.e., $\nabla g = 0$) on TM such that $\nabla Z = 0$. Furthermore, denoting by Tor the Torsion of ∇, we require that

(i) $\mathrm{Tor}(X,Y) = \langle X, JY\rangle Z = -\mathrm{d}\omega(X,Y)Z$ for all $X, Y \in \mathcal{D}$;

(ii) $\mathrm{Tor}(Z, JX) = -J\mathrm{Tor}(Z,X)$ for any vector field X on M.

Proposition 15 *Under the assumptions of Theorem 12, let us suppose that S is locally defined as the zero level set of a smooth function $f : M \to \mathbb{R}$ such that $\nabla_H f|_U \neq 0$ and $\langle \nabla_H f, \nabla_H \delta\rangle|_U > 0$. Then, the following formula is equivalent to (7):*

$$\nu(S_\varepsilon)=\varepsilon \int_S dA-\frac{\varepsilon^2}{2}\int_S \operatorname{div}\left(\frac{\nabla_H f}{\|\nabla_H f\|}\right) dA+\frac{\varepsilon^3}{6}\int_S a_3\, dA,$$

where

$$a_3=\int_S\left[2X_S\left(\frac{-Zf}{\|\nabla_H f\|}\right)-\left(\frac{Zf}{\|\nabla_H f\|}\right)^2-\kappa+\langle \mathrm{Tor}\left(Z,X_S\right),N\rangle\right] dA$$

with $\kappa = \langle R(X,Y)Y,X\rangle$ where R and Tor *are the curvature and the torsion operators associated with the Tanno connection of Definition 14.*

Remark 16 The second order coefficients appearing in (1) and (7) are both integral of a mean curvature[1]. One may then wonder if the same analogy holds for the third order

[1] The extra factor 2 in the sub-Riemannian formula is due to the fact that in the Euclidean case one defines the mean curvature as one half of the sum of the two principal curvatures.

coefficients. Namely, if the third coefficient of expansion (7) is the integral of a suitably defined sub-Riemannian Gaussian curvature of the surface as in (1).

A strategy to define sub-Riemannian objects is the Riemannian approximation. More precisely, we consider the Riemannian manifold (M, g^ε), where g^ε is the Riemannian metric obtained as the extension of the sub-Riemannian one g imposing that the Reeb vector field Z is orthogonal to $\mathcal{D}$ and with norm $1/\varepsilon$. The metric spaces (M, g^ε) converge to (M, g) in the Gromov-Hausdorff sense as $\varepsilon \to 0$. Therefore one can try to define the sub-Riemannian objects taking the limit of the corresponding Riemannian ones defined in these approximating manifolds.

The sub-Riemannian mean curvature $\mathcal{H}$ in (6) can be equivalently defined as the limit of the Riemannian mean curvatures H^ε of S with respect to the Riemannian extension g^ε converging to the sub-Riemannian metric g. The same does not hold for the Gaussian curvature.

In [3] the authors define the sub-Riemannian Gaussian curvature $\mathcal{K}_S$ of a surface S in the Heisenberg group $\mathbb{H}$ through Riemannian approximations. The expression of the limit, written in our notation is

$$\mathcal{K}_S = X_S(Z\delta) - (Z\delta)^2,$$

It can be checked that the integral of this quantity does not correspond to the third coefficient, that thanks to Proposition 15 specified in the Heisenberg case, rewrites as follows

$$a_3 = \int_S 2X_S\,(Z\delta) - (Z\delta)^2\; dA.$$

being $\kappa = 0$ and Tor the null operator in the Heisenberg group.

References

[1] A. Agrachev, D. Barilari, and U. Boscain, "A Comprehensive Introduction to Sub-Riemannian Geometry". Cambridge University Press, 2019.

[2] Z.M. Balogh, F. Ferrari, B. Franchi, E. Vecchi, and K. Wildrick, *Steiner's formula in the Heisenberg group.* Nonlinear Anal. 126 (2015), no. 1-2, 1–38.

[3] Z.M. Balogh, J.T. Tyson, and E. Vecchi, *Intrinsic curvature of curves and surfaces and a Gauss-Bonnet theorem in the Heisenberg group.* Math. Z. 287 (2017), no. 1-2, 1–38.

[4] D. Barilari and T. Bossio, *Steiner and tube formulae in 3D contact sub-Riemannian geometry.* Communications in Contemporary Mathematics, 2023.

[5] A. Gray, "*Tubes*, Second". Progress in Mathematics, vol. 221, Birkhäuser Verlag, Basel, 2004. With a preface by Vicente Miquel.

[6] M. Ritoré, *Tubular neighborhoods in the sub-Riemannian Heisenberg groups.* Adv. Calc. Var. 14 (2021), no. 1, 1–36.

[7] T. Rossi, *The relative heat content for submanifolds in sub-riemannian geometry.* ArXiv, 2022.

Ergodic Mean-Field Games with Riesz-type aggregation

Chiara Bernardini [*]

1 Introduction

The theory of Mean Field Games (MFG in short) has been introduced around 2006 in a series of seminal papers by Lasry and Lions [22–24], in order to model Nash equilibria of differential games with infinitely many interacting agents. Independently at about the same time, Caines, Huang and Malhamé [18–20] developed the analogous concept of "Nash centainly equivalence principle". Since then, the study of MFG rapidly grew, also encouraged by its powerful applications in a wide range of disciplines: equations of this kind arise in Economics, Finance, models of social systems and crowd motions. For a complete presentation of the theory and its applications, we refer the reader to P.-L. Lions series of lectures at Collège de France [29], the lectures by Gueant, Lasry and Lions [16] and also [1] among many others.

This Section is a basic introduction to Mean Field Games from a PDE viewpoint, providing the main fact we need to contextualize our problems.

In the Mean Field Games theory, players are assumed to be indistinguishable and "rational", that is each of them optimizes his/her behavior by taking into account the behavior of the other players, in this sense each individual strategy is influenced by some averages of quantities depending on the states of the other agents. Moreover, agents are infinitesimal, namely they are small compared to the collection of all other controllers and hence individually have a negligible influence on the game. Each agent chooses his optimal strategy in view of global information that are available to him and that result from the action of all other players, which is described through the distribution law of the dynamical states. In this setting, the central concept is the notion of Nash equilibria, which describes how agents play in an optimal way by taking into account the others' strategies. In particular, there is a Nash equilibrium when no controller has interest to deviate unilaterally from the planned control. The key idea underlying the theory comes from Statistical Mechanics, and consists in a mean-field approach to describe equilibria in a system of many interacting

[*] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: `chiara.bernardini@math.unipd.it`. Seminar held on 29 March 2023.

identical particles.

Let us briefly describe the derivation of the MFG system in the simplest case where the state space is $\mathbb{R}^N$. Assume to have a differential game with infinitely many players, each agent controls his/her own dynamics, which is described by the following stochastic differential equation (SDE)

$$X_t = x + \int_0^t b(X_s, \alpha_s, m(s))ds + \sqrt{2\varepsilon} B_t$$

where B_t is a N-dimensional standard Brownian motion, α is the control and m is the distribution of the other players. The cost each agent want to minimize is

$$J(t,x,\alpha) = \mathbb{E}\left[\int_t^T L(X_s, \alpha_s, m(s))ds + G(x_T, m(T))\right].$$

We introduce the value function:

$$u(t,x) = \inf_{\alpha \in \mathcal{A}} J(t,x,\alpha)$$

where $\mathcal{A}$ is the set of admissible controls. Then at least formally, the value function u solves the Hamilton-Jacobi equation

$$\begin{cases} -\partial_t u - \varepsilon \Delta u + H(x, Du, m) = 0 \quad \text{in } (0,T) \times \mathbb{R}^d \\ u(x,T) = G(x, m(T)) \end{cases}$$

where

$$H(x,p,m) = \sup_\alpha [-b(x,\alpha,m) \cdot p - L(x,\alpha,m)].$$

By standard argument in control theory, if $\alpha^*(t,x)$ is defined as the maximum point in the definition of H when $p = Du(t,x)$, one can verify that α^* is the optimal feedback for the problem, and the drift is of the form $b(x, \alpha^*(t,x), m(t)) = -H_p(x, Du(t,x), m(t))$. If all agents argue in this way and if their associated noises are independent, the law of large numbers implies that their distribution evolves with a velocity which is due, on the one hand, to the diffusion, and, on the other hand, on the drift term. This leads to a Kolmogorov-Fokker-Planck equation. So, from a PDE view point, Nash equilibria of the differential game solve the following system, where the unknowns are the functions u and m

$$\text{(1)} \qquad \begin{cases} -\partial_t u - \varepsilon \Delta u + H(x, Du, m) = 0 \quad \text{in } (0,T) \times \mathbb{R}^d \\ \partial_t u - \varepsilon \Delta m - \operatorname{div}(m\, H_p(x, Du, m)) = 0 \quad \text{in } (0,T) \times \mathbb{R}^d \\ u(x,T) = G(x, m(T)) \end{cases} .$$

In our setting we assume that the dynamics of each player is described by the following controlled stochastic differential equation

$$dX_t = -v_t\, dt + \sqrt{2}\, dB_t$$

where v_t is the controlled velocity and B_t is a standard N dimensional Brownian motion. Moreover, the cost is of long time average type, namely

$$\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\int_0^T\left[\frac{|v_t|^{\gamma'}}{\gamma'}+V(X_t)-K_\alpha * m(X_t)\right]dt$$

where $\gamma'=\frac{\gamma}{\gamma-1}$ is the conjugate exponent of γ and $m(x)$ is the density of population at $x\in\mathbb{R}^N$. We will assume that the potential V is a locally Hölder continuous coercive function, that is there exist b and C_V positive constants such that

$$C_V^{-1}(\max\{|x|-C_V,0\})^b\le V(x)\le C_V(1+|x|)^b,\quad \forall x\in\mathbb{R}^N. \tag{2}$$

The assumption of V to be non-negative is not restrictive, we can assume more generally that V is bounded from below and shift appropriately λ.
The coupling in the system is given through the interaction term $-K_\alpha * m$, where K_α is the Riesz potential of order $\alpha\in(0,N)$ defined as

$$K_\alpha(x)=\frac{1}{|x|^{N-\alpha}}.$$

So, given $M>0$, we consider elliptic systems of the form

$$\begin{cases}-\Delta u+\frac{1}{\gamma}|\nabla u|^\gamma+\lambda=V(x)-\int_{\mathbb{R}^N}\frac{m(y)}{|x-y|^{N-\alpha}}dy\\ -\Delta m-\operatorname{div}(m\nabla u(x)\,|\nabla u(x)|^{\gamma-2})=0\\ \int_{\mathbb{R}^N}m=M,\quad m\ge 0\end{cases}\qquad\text{in }\mathbb{R}^N \tag{3}$$

where $\gamma>1$ and $\alpha\in(0,N)$ are fixed. Note that the unknowns in the system (3) are the functions u,m and the constant $\lambda\in\mathbb{R}$ which can be interpreted as a Lagrange multiplier, related to the mass constraint $\int_{\mathbb{R}^N}m=M$. In the ergodic setting, the distribution law of each player moving with optimal speed converges as $t\to+\infty$ to an invariant measure μ (independent of the initial position) and μ coincides, in a equilibrium regime for the game, with the density of the population m. From a PDE viewpoint, equilibria of the differential game are encoded by solutions of the system (3), where the Hamilton-Jacobi-Bellman equation takes into account the value of the game λ and the optimal speed $-\nabla u|\nabla u|^{\gamma-2}$ of the optimal control problem of a typical agent, and the Kolmogorov-Fokker-Planck equation gives the density of the overall population m.

In the case when $\gamma=\gamma'=2$, as pointed out in [22], using the Hopf-Cole transformation $v(x):=e^{-u(x)/2}$ we can reduce the MFG system (3) to a single PDE. In particular we observe that with the previous change of variable, setting $m(x)=v^2(x)$, the MFG system (3) is equivalent to the normalized Choquard equation

$$\begin{cases}-2\Delta v+(V(x)-\lambda)v=(K_\alpha * v^2)v\\ \int_{\mathbb{R}^N}v^2(x)dx=M,\qquad v>0\end{cases}\qquad\text{in }\mathbb{R}^N, \tag{4}$$

with associated energy

$$\mathcal{E}(v) = \int_{\mathbb{R}^N} 2|\nabla v|^2 + V(x)v^2 dx - \frac{1}{2}\int_{\mathbb{R}^N}\int_{\mathbb{R}^N} \frac{v^2(x)\, v^2(y)}{|x-y|^{N-\alpha}} dx\, dy.$$

Choquard-type equations have been intensively studied during the last decades and have appeared in the context of various mean-field type physical models (refer to [25–27, 31, 32] and references therein for a complete overview). Indeed their solutions are steady states of a generalized nonlinear Schrödinger equation, with attractive interaction potential given in terms of a Riesz interaction kernel, which is therefore weaker and with longer range than the usual power-type potential in nonlinear Schrödinger equation. Equation (4) was first studied by E. Lieb [25], who proved existence and uniqueness (up to translations) of solutions when $N = 3$ and $\alpha = 2$ by using symmetric decreasing rearrangement inequalities. Then, P.-L. Lions [26] proved that there exists a minimum of the energy associated to (4) when we restrict the infimum to functions with spherical symmetry, refer to [28, §3] and also [31, 32] for further results.

Going back to our Mean-Field Game system, the two distinctive features of our model are the following: the state space is the whole Euclidean space $\mathbb{R}^N$, and the coupling is aggregative and defined in terms of a Riesz-type interaction kernel. Usually, Mean-Field Game systems are considered in bounded domains, with Neumann or periodic boundary conditions, in order to avoid non-compactness issues. We recall some works in the non compact setting: in particular [3] in the linear-quadratic framework, [33] in the time-dependent case, [14] for regularity results and finally [7], where a system analogous to (3) has been considered, with power-type nonlinearity. In the unbounded setting, the dissipation induced by the Brownian motion has to be compensated by the optimal velocity, which is a priori unknown and depends by the distribution m itself and on the coercive potential V. The coercive potential V describes spatial preferences of agents and hence discourages them to be far away from the origin. Moreover, due to the presence of the Riesz-type interaction potential $-K_\alpha * m$ which represents the coupling between the individual and the overall population, every player of the game is attracted toward regions where the population is highly distributed. Most of the MFG literature focuses on the study of systems with competition, namely when the coupling descourages aggregation: this assumption is essential if one seeks for uniqueness of equilibria, and it is in general crucial in many existence and regularity arguments (see [15]). Focusing MFG systems, namely models with coupling which encourages aggregation, have been studied for instance in [7, 8, 10, 11, 13] in the stationary setting.

2 Existence and nonexistence of solutions

In this chapter we provide existence and nonexistence results of classical solutions solving the MFG system (3), where by *classical solution* we will mean a triple $(u, m, \lambda) \in C^2(\mathbb{R}^N) \times W^{1,p}(\mathbb{R}^N) \times \mathbb{R}$ for every $p \in (1, +\infty)$. Our focus will be to obtain classical solutions which satisfy some integrability conditions and boundary conditions at ∞ which will be meaningful from the point of view of the game. In particular, we will require some integrability properties of the optimal speed with respect to m, namely

$$m|\nabla u|^\gamma \in L^1(\mathbb{R}^N) \qquad Vm \in L^1(\mathbb{R}^N) \qquad \text{and} \qquad |\nabla m||\nabla u| \in L^1(\mathbb{R}^N). \tag{5}$$

Indeed, if one looks at the Kolmogorov equation, such integrability properties are important to ensure some minimal regularity of m and uniqueness of the invariant distribution itself (see [17, 30]). Regularity and boundedness of m is quite crucial in our setting: indeed, due to the aggregating forces, m has an intrinsic tendency to concentrate and hence to develop singularities. Moreover, the Lagrange multiplier λ will be uniquely defined as the generalized principal eigenvalue (see for details [4, 9, 21]): if $m \in L^1(\mathbb{R}^N)$ is fixed and such that $K_\alpha * m \in C^{0,\theta}(\mathbb{R}^N)$ for some $\theta \in (0,1)$, we define λ as

$$\lambda := \sup\left\{ c \in \mathbb{R} \,\middle|\, \exists v \in C^2(\mathbb{R}^N) \text{ solving } \Delta v + \frac{1}{\gamma}|\nabla v|^\gamma + c = V(x) - K_\alpha * m \right\}.$$

Once we know this value exists, it is possible to show that there exists $u \in C^2(\mathbb{R}^N)$ solving the HJB equation with such value λ, and that such solution u is coercive i.e.

$$u(x) \to +\infty \qquad \text{as } |x| \to +\infty \tag{6}$$

and moreover its gradient has polynomial growth (see [4, 6, 9, 21]). Note that (6) is a quite natural "boundary" condition for ergodic HJB equations on the whole space: indeed the optimal speed would give rise to an ergodic process, so in particular, at least heuristically $-\nabla u \cdot x < 0$ for $|x| \to +\infty$, (refer to [17] and references therein, for more information about ergodic problems on the whole space and their characterization in terms of Lyapunov functions). Existence results for such classical solutions will depend on the interplay between the dissipation (i.e. by the diffusive term in the system) and the aggregating forces (described in terms of the Riesz potential K_α and the coercive potential V). So, we get that the MFG system (3) shows three different regimes which correspond to $\alpha \in (0, N-2\gamma')$, $\alpha \in (N-2\gamma', N-\gamma')$ and $\alpha \in (N-\gamma', N)$. We will refer to $\alpha = N-2\gamma'$ as the *Hardy-Littlewood-Sobolev-critical exponent* and to $\alpha = N-\gamma'$ as the *mass-critical* (or L^2-critical) *exponent*, in analogy with the regimes appearing in the study of the Choquard equation (4) when $\gamma' = 2$. Obviously if $\gamma' \geq N$, there exists just one regime, which will be the mass-subcritical regime $\alpha \in [0, N)$, whereas if $\frac{N}{2} \leq \gamma' < N$ there will be just 2 regimes.

First of all we observe that for classical solutions to (3) with $V \equiv 0$ and which satisfy (5), a Pohozaev type identity holds:

$$(2-N)\int_{\mathbb{R}^N} \nabla u \cdot \nabla m \, dx + \left(1 - \frac{N}{\gamma}\right)\int_{\mathbb{R}^N} m|\nabla u|^\gamma dx = \lambda N M + \frac{\alpha+N}{2}\int_{\mathbb{R}^{2N}} \frac{m(x)m(y)}{|x-y|^{N-\alpha}} dxdy. \tag{7}$$

Also in presence of the potential a similar identity holds, under the additional integrability condition that $m\nabla V \cdot x \in L^1(\mathbb{R}^N)$. For MFG in the periodic setting with polynomial interaction potential an analogous Pohozaev identity has been proved in [10]. For the case of the Choquard equation we refer to [31] and references therein.

In the *Hardy-Littlewood-Sobolev-supercritical regime* $0 < \alpha < N - 2\gamma'$, the Pohozaev identity, together with the fact that $\lambda \leq 0$ (see [6, Lemma 2.11]), implies that solutions to the MFG system (3) do not exist. More precisely, we obtain the following nonexistence result.

Theorem 2.1 *Assume that $\alpha \in (0, N - 2\gamma')$ and $V \equiv 0$. Then, the MFG system (3) has no classical solutions $(u, m, \lambda) \in C^2(\mathbb{R}^N) \times W^{1,\frac{2N}{N+\alpha}}(\mathbb{R}^N) \times \mathbb{R}$ which satisfy (5) and (6).*

On the other hand in the *Hardy-Littlewood-Sobolev subcritical regime* $N - 2\gamma' < \alpha < N$ we obtain existence of classical solutions to the MFG system (3) by means of a Schauder fixed point argument (refer to [2] and see also [10]). More in detail, we consider a regularized version of problem (3), obtained by convolving the Riesz-interaction term with a sequence of standard symmetric mollifiers. Taking advantage of the fixed-point structure associated to the MFG system and exploiting the Schauder Fixed Point Theorem, we show that solutions to the "regularized" version of the MFG system do exist. Then we provide a priori uniform estimates on the solutions to the regularized problem, which allow us to pass to the limit and obtain a classical solution of the MFG system (3).

Theorem 2.2 *Assume that the potential V is locally Hölder continuous and satisfies (2). We have the following results:*

(i) *if $N - \gamma' < \alpha < N$ then, for every $M > 0$ the MFG system (3) admits a classical solution (u, m, λ);*

(ii) *if $N-2\gamma' < \alpha \leq N-\gamma'$ then, there exists a positive real value $M_0 = M_0(N, \alpha, \gamma, C_V, b)$ such that if $M \in (0, M_0)$ the MFG system (3) admits a classical solution (u, m, λ).*

Moreover in both cases there exists a constant $C > 0$ such that

$$|\nabla u(x)| \leq C(1 + |x|)^{\frac{b}{\gamma}} \qquad u(x) \geq C|x|^{\frac{b}{\gamma}+1} - C^{-1},$$

where $C = C(C_V, b, \gamma, N, \lambda, \alpha)$, $\sqrt{m} \in W^{1,2}(\mathbb{R}^N)$ and it holds

$$m|\nabla u|^\gamma \in L^1(\mathbb{R}^N), \qquad mV \in L^1(\mathbb{R}^N), \qquad |\nabla u|\,|\nabla m| \in L^1(\mathbb{R}^N).$$

Note that in the mass-subcritical case, solutions to the MFG exist for every mass M, whereas in the mass-supercritical case and mass-critical case (namely for $\alpha \in (N-2\gamma', N-\gamma']$) we provide existence just for sufficiently small masses, below some threshold M_0, due to the fact that in this case the interaction attractive potential is stronger than the diffusive part.

The Hardy-Littlewood-Sobolev critical exponent is not covered by our analysis. Indeed it is possible to prove existence of solutions to the regularized problem also in this case, for sufficiently small masses. Nevertheless in order to pass to the limit in the regularization, we need to obtain a priori L^∞ bounds on solutions m_k to the regularized problem, starting from uniform bounds in $L^{\frac{2N}{N+\alpha}} \cap L^1$. This is not possible at the critical level $\alpha = N - 2\gamma'$, due to critical rescaling properties of the Sobolev critical exponent: a priori uniform L^∞

bounds on m_k only holds in the range when we have a uniform bound in L^q, for $q > \frac{N}{\gamma'+\alpha}$ and $\frac{2N}{N+\alpha} > \frac{N}{\gamma'+\alpha}$ only in the Hardy-Littlewood-Sobolev subcritical regime. One way to circumvent this difficulty would be to obtain at the critical level $\alpha = N - 2\gamma'$, by using regularity estimates on the viscous Hamilton-Jacobi equation and on the Fokker Planck equation and a smallness condition on $\|m\|_{\frac{N}{N-\gamma'}}$, a priori uniform bounds on m in L^q for some $q > \frac{N}{N-\gamma'}$. This kind of result has been obtained recently in [12] for MFG in bounded domains with Neumann boundary conditions, and with a nonlinear Schrödinger type potential. This problem is related to the maximal regularity of solutions to viscous Hamilton-Jacobi equation $-\Delta u + |\nabla u|^\gamma = f(x)$. When $m \in L^{\frac{N}{N-\gamma'}}$, then by Hardy-Littlewood-Sobolev inequality $K_\alpha * m \in L^{\frac{N}{\gamma'}}$, which is a critical threshold in this setting.

2.1 The vanishing viscosity limit

Then, we consider an ergodic MFG system defined in the whole space $\mathbb{R}^N$ with an external confining potential V and Brownian noise which depends on $\varepsilon > 0$. More in detail, we take into account systems of the form

$$(8) \qquad \begin{cases} -\varepsilon\Delta u + \frac{1}{\gamma}|\nabla u|^\gamma + \lambda = V(x) - K_\alpha * m(x) \\ -\varepsilon\Delta m - \operatorname{div}(m\nabla u|\nabla u|^{\gamma-2}) = 0 \\ \int_{\mathbb{R}^N} m = M, \quad m \geq 0 \end{cases} \quad \text{in } \mathbb{R}^N.$$

Studying the asymptotic behavior of rescaled solutions to the MFG system (8) in the vanishing viscosity limit, we are able to prove existence of classical solutions to the potential-free MFG system. As a matter of fact, letting $\varepsilon \to 0$, the dynamic of each player is no subject anymore to the dissipation effect induced by the Brownian motion, so we expect aggregation of players. In particular, in the vanishing viscosity limit the mass m tends to concentrate, the introduction of the coercive potential V, which represents spatial preferences of agents, rules out this possibility and leads to concentration of mass around minima of the potential V.

Focusing on the *mass-subcritical regime* $N - \gamma' < \alpha < N$, where $\gamma' = \frac{\gamma}{\gamma-1}$ is the conjugate exponent of γ, we provide existence of classical solutions to the MFG system

$$(9) \qquad \begin{cases} -\Delta u + \frac{1}{\gamma}|\nabla u|^\gamma + \lambda = -K_\alpha * m(x) \\ -\Delta m - \operatorname{div}(m\nabla u|\nabla u|^{\gamma-2}) = 0 \\ \int_{\mathbb{R}^N} m = M, \quad m \geq 0 \end{cases} \quad \text{in } \mathbb{R}^N$$

where $\gamma > 1$ is fixed and $K_\alpha : \mathbb{R}^N \to \mathbb{R}$ is the Riesz potential of order $\alpha \in (0, N)$. Notice that by *classical solution* we mean a triple $(u, m, \lambda) \in C^2(\mathbb{R}^N) \times W^{1,p}(\mathbb{R}^N) \times \mathbb{R}$ for every $p \in (1, +\infty)$, solving the system. More precisely, we obtain the following existence result.

Theorem 2.3 *Let $N - \gamma' < \alpha < N$. Then, for every $M > 0$ there exists $(\bar{u}, \bar{m}, \bar{\lambda})$ classical solution to the MFG system (9). Moreover, there exist C_1, C_2, C_3 and C_4 positive constants such that*

$$\bar{u}(x) \geq C_1|x| - C_1^{-1}, \qquad |\nabla \bar{u}| \leq C_2$$

and

$$0 < \bar{m}(x) \leq C_3 e^{-C_4|x|}.$$

Remark 1 Solutions to the MFG system (9) are invariant by translation, namely if $(\bar{u}(x), \bar{m}(x), \bar{\lambda})$ is a classical solution to (9) then for every $x_0 \in \mathbb{R}^N$ and $c \in \mathbb{R}$, also $\big(\bar{u}(x+x_0)+c, \bar{m}(x+x_0), \bar{\lambda}\big)$ is a classical solution to (9). Therefore, the constants C_1 and C_4 appearing in the previous theorem, depend on the choice of the solution.

Theorem 2.3 partially completes the study of existence of solutions to the potential-free MFG system (9) started in [6]. In particular, in [6, Theorem 1.1], using a Pohozaev-type identity, one proves that if $0 < \alpha < N - 2\gamma'$, "regular" solutions to the MFG system (9) (namely satisfying some quite natural integrability conditions and boundary conditions at infinity) do not exist. It remains still open the problem of existence of solutions to (9) when $\alpha \in [N-2\gamma', N-\gamma']$.

On the other hand, the main result in [6] deals with the study of MFG systems with Riesz-type coupling and external confining potential V. More in detail, exploiting a Schauder fixed point argument, one proves that if $\alpha \in (N-\gamma', N)$ the MFG system admits a classical solution for every total mass $M > 0$, while if $\alpha \in (N-2\gamma', N-\gamma']$ a solution does exist at least for sufficiently small masses M, below some threshold value M_0 (see [6, Theorem 1.2] for more details). This different behavior is due to the fact that when $N-2\gamma' < \alpha \leq N-\gamma'$ the interaction attractive potential is stronger than the diffusive part, so if the total mass M is too large, the mass m tends to concentrate and hence to develop singularities. Notice that, using a fixed point approach, the presence of the coercive potential V adds compactness to the problem and proves to be essential to conclude, so the existence result in [6] does not cover the case when the potential V is identically 0. In order to deal with the potential-free system we take advantage a variational argument. This approach allows us to obtain some uniform (namely not depending on the viscosity parameter) estimates on the solutions, which will be crucial in the vanishing viscosity setting and which can not be obtained by means of a fixed point technique.

Finally, a similar MFG system but with local decreasing coupling defined in terms of a power-type function, has been studied in [7]. We point out that, in our setting the nonlocal attractive coupling models a long-range attractive force between players, moreover, in order to deal with the Riesz-term we need different techniques compared to the ones used in [7].

Let us summarize the main tools to prove our results. As J.-M. Lasry and P.-L. Lions first pointed out in [24], taking into account the variational nature of the MFG system, solutions to (8) are related to critical points of the following energy functional

$$\mathcal{E}(m,w) := \begin{cases} \int\limits_{\mathbb{R}^N} mL\left(-\frac{w}{m}\right) + V(x)\, m\, dx - \frac{1}{2} \int\limits_{\mathbb{R}^N} \int\limits_{\mathbb{R}^N} \frac{m(x)\, m(y)}{|x-y|^{N-\alpha}} dx\, dy & \text{if } (m,w) \in \mathcal{K}_{\varepsilon,M}, \\ +\infty & \text{otherwise} \end{cases} \tag{10}$$

where

$$L\left(-\frac{w}{m}\right) := \begin{cases} \frac{1}{\gamma'}\left|\frac{w}{m}\right|^{\gamma'} & \text{if } m > 0 \\ 0 & \text{if } m = 0,\ w = 0 \\ +\infty & \text{otherwise} \end{cases}$$

and the constraint set is defined as

$$\begin{aligned} \mathcal{K}_{\varepsilon,M} := \Big\{ (m,w) \in & (L^1(\mathbb{R}^N) \cap L^q(\mathbb{R}^N)) \times L^1(\mathbb{R}^N) \quad \text{s.t.} \quad \int_{\mathbb{R}^N} m\, dx = M, \quad m \geq 0 \text{ a.e.} \\ & \varepsilon \int_{\mathbb{R}^N} m(-\Delta\varphi)\, dx = \int_{\mathbb{R}^N} w \cdot \nabla\varphi\, dx \quad \forall \varphi \in C_0^\infty(\mathbb{R}^N) \Big\} \end{aligned} \tag{11}$$

with

$$q := \begin{cases} \frac{N}{N-\gamma'+1} & \text{if } \gamma' < N \\ \gamma' & \text{if } \gamma' \geq N \end{cases} . \tag{12}$$

Using some regularity results for the Kolmogorov equation, the Hardy-Littlewood-Sobolev inequality and the fact that V is non-negative, we prove that the energy $\mathcal{E}$ is bounded from below. By classical direct methods and compactness arguments, we obtain minimizers $(m_\varepsilon, w_\varepsilon)$ of $\mathcal{E}$. Finally, passing to another functional with linearized Riesz-term and using convex duality arguments, we are able to construct the associated solutions $(u_\varepsilon, m_\varepsilon, \lambda_\varepsilon)$ of the MFG system (8). Then, in order to investigate the behavior of the system in the vanishing viscosity limit, we define a suitable rescaling of u_ε, m_ε and λ_ε. We also translate the reference system by y_ε, where y_ε is a point of minimum for the value function u_ε, in this way around y_ε the mass remains positive and we can rule out vanishing of the total mass in the limit. We obtain a triple $(\bar{m}_\varepsilon, \bar{u}_\varepsilon, \tilde{\lambda}_\varepsilon)$ which solves the following MFG system

$$\begin{cases} -\Delta\bar{u}_\varepsilon + \frac{1}{\gamma}|\nabla\bar{u}_\varepsilon|^\gamma + \tilde{\lambda}_\varepsilon = \varepsilon^{\frac{(N-\alpha)\gamma'}{\gamma'-N+\alpha}} V\left(\varepsilon^{\frac{\gamma'}{\gamma'-N+\alpha}}(y+y_\varepsilon)\right) - K_\alpha * \bar{m}_\varepsilon(y) \\ -\Delta\bar{m}_\varepsilon - \operatorname{div}(\bar{m}_\varepsilon \nabla\bar{u}_\varepsilon |\nabla\bar{u}_\varepsilon|^{\gamma-2}) = 0 \\ \int_{\mathbb{R}^N} \bar{m}_\varepsilon = M \end{cases} .$$

Exploiting a concentration-compactness argument (refer to the seminal work of P.-L. Lions [28]) as done in [7], we are able to prove that there is no loss of mass when passing to the limit as $\varepsilon \to 0$. We show that in the vanishing viscosity limit, the rescaled solutions converge (up to sub-sequences) to $(\bar{u}, \bar{m}, \bar{\lambda})$ *classical* solution to the MFG system (9). Moreover, solutions to (9) are related to minimum points of the following energy

$$\mathcal{E}_0(m,w) := \int_{\mathbb{R}^N} \frac{m}{\gamma'}\left|\frac{w}{m}\right|^{\gamma'} dx - \frac{1}{2}\int_{\mathbb{R}^N}\int_{\mathbb{R}^N} \frac{m(x)m(y)}{|x-y|^{N-\alpha}} dx\, dy$$

over the constraint set

$$\mathcal{B} := \Big\{ (m,w) \in \mathcal{K}_{1,M} \;\Big|\; m(1+|x|^b) \in L^1(\mathbb{R}^N) \Big\}.$$

The following theorem states existence of solutions to (8) and concentration of mass.

Theorem 2.4 *Let $N - \gamma' < \alpha < N$. Assume that the potential V is locally Hölder continuous and satisfies (2). Then, for every $\varepsilon, M > 0$ there exists $(u_\varepsilon, m_\varepsilon, \lambda_\varepsilon)$ classical solution to (8), such that $(m_\varepsilon, -m_\varepsilon \nabla u_\varepsilon |\nabla u_\varepsilon|^{\gamma-2})$ is a minimum of the energy $\mathcal{E}$.*

Moreover, there exists a sequence $\varepsilon \to 0$ and a sequence of points $x_\varepsilon = \varepsilon^{\frac{\gamma'}{\gamma'-N+\alpha}} y_\varepsilon$ around which there is concentration of mass, namely for every $\eta > 0$ there exist $R, \varepsilon_0 > 0$ such that

$$\int_{B(x_\varepsilon, \varepsilon^{\frac{\gamma'}{\gamma'-N+\alpha}} R)} m_\varepsilon(x)\, dx \geq M - \eta$$

for all $\varepsilon < \varepsilon_0$ and

$$x_\varepsilon \to \bar{x}, \quad \text{as } \varepsilon \to 0$$

where $\bar{x}$ is a minimum point of the potential V and $V(\bar{x}) = 0$.

Remark 2 We assumed that the Hamiltonian H has the form $H(p) = \frac{1}{\gamma}|p|^\gamma$ for $\gamma > 1$ fixed, but actually the previous results hold also for more general assumptions on the Hamiltonian H, namely assuming that the Hamiltonian $H : \mathbb{R}^N \to \mathbb{R}$ is strictly convex, $H \in C^2(\mathbb{R}^N \setminus \{0\})$ and there exist C_H, $K > 0$ and $\gamma > 1$, such that $\forall p \in \mathbb{R}^N$ the following conditions hold

$$C_H|p|^\gamma - K \leq H(p) \leq C_H|p|^\gamma$$

$$\nabla H(p) \cdot p - H(p) \geq K^{-1}|p|^\gamma - K$$

$$|\nabla H(p)| \leq K|p|^{\gamma-1}.$$

References

[1] Achdou Y., Cardaliaguet P., Delarue F., Porretta A., Santambrogio F., “Mean field games”. Lecture Notes in Mathematics, Vol. 2281. CIME Foundation Subseries. Springer, vii+307 pp. ISBN: 978-3-030-59837-2; 978-3-030-59836-549-06 (35-06 49N80 91-06).

[2] Bardi M., Feleqi E., *Nonlinear elliptic systems and mean-field games.* NoDEA Nonlinear Differential Equations Appl. 23 (2016), no. 4, Art. 44, 32 pp.

[3] Bardi M., Priuli F.S., *Linear-quadratic N-person and mean-field games with ergodic cost.* SIAM J. Control Optim. 52:5 (2014), 3022–3052.

[4] Barles G., Meireles J., *On unbounded solutions of ergodic problems in $\mathbb{R}^m$ for viscous Hamilton-Jacobi equations.* Comm. Partial Differential Equations, 41:12, 1985-2003 (2016). doi:10.1080/03605302.2016.1244208.

[5] Bernardini C., *Mass concentration for Ergodic Choquard Mean-Field Games.* Submitted (2022). Preprint ArXiv: 2212.00132.

[6] Bernardini C., Cesaroni A., *Ergodic Mean-Field Games with aggregation of Choquard-type.* J. Differential Equations 364, 296-335 (2023). doi:10.1016/j.jde.2023.03.045.

[7] Cesaroni A., Cirant M., *Concentration of ground states in stationary Mean-Field Games systems.* Anal. PDE 12 (2019), no. 3, 737–787. doi:10.2140/apde.2019.12.737.

[8] Cesaroni A., Cirant M., *ntroduction to variational methods for viscous ergodic mean-field games with local coupling.* Contemporary research in elliptic PDEs and related topics, 221–246, Springer INdAM Ser., 33, Springer, Cham, 2019.

[9] Cirant M., *On the solvability of some ergodic control problems in* $\mathbb{R}^d$. SIAM J. Control Optim., 52:6 (2014), 4001–4026. doi:10.1137/140953903.

[10] Cirant M., *Stationary focusing Mean Field Games.* Comm. in Partial Differential Equations, (2016) 41, no 8, 1324–1346. doi:10.1080/03605302.2016.1192647.

[11] Cirant M., *On the existence of oscillating solutions in non-monotone mean-field games.* J. Differential Equations 266 (2019), no. 12, 8067–8093.

[12] Cirant M., Cosenza A., Verzini G., *Ergodic Mean Field Games: existence of local minimizers up to the Sobolev critical case.* (2023). Preprint Arxiv: Preprint ArXiv: 2301.11692.

[13] Gomes D.A., Nurbekyan L., Prazeres M., *One-dimensional stationary mean-field games with local coupling.* Dyn. Games Appl. 8 (2018), no. 2, 315–351.

[14] Gomes D.A., Pimentel E., *Local regularity for mean-field games in the whole space.* Minimax Theory Appl. 1:1 (2016), 65–82.

[15] Gomes D.A., Pimentel E.A., Voskanyan V., "Regularity theory for mean-field game systems". Springer Briefs in Mathematics. Springer, 2016.

[16] Guéant O., Lasry J.-M., Lions P.-L., *Mean field games and applications.* In: Carmona, R.A., et al. (eds.) Paris-Princeton Lectures on Mathematical Finance 2010. Lecture Notes in Mathematics, vol. 2003, pp. 205–266. Springer, Berlin (2011).

[17] Hasminskii R.Z., *Stochastic Stability of Differential Equations.* Sijthoff & Noordhoff, Alphen aan den Rijn, Netherlands (1980).

[18] Huang M., Malhamé R.P., Caines P.E., *Large population stochastic dynamic games: closed loop Mckean-Vlasov systems and the Nash certainty equivalence principle.* Commun. Inf. Syst. 6, 221–252 (2006).

[19] Huang M., Malhamé R.P., Caines P.E., *An invariance principle in large population stochastic dynamic games.* J. Syst. Sci. Complex. 20, 162–172 (2007).

[20] Huang M., Caines P.E., Malhamé R.P., *Large-population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized* ε*-Nash equilibria.* IEEE Trans. Automat. Control 52: 1560–1571 (2007).

[21] Ichihara N., *The generalized principal eigenvalue for Hamilton-Jacobi-Bellman equations of ergodic type.* Ann. Inst. H. Poincaré Anal. Non Linéaire 32:3 (2015), 623–650.

[22] Lasry J.-M., Lions P.-L., *Jeux à champ moyen. I. Le cas stationnaire.* C.R. Math. Acad. Sci. Paris, 343:9 (2006), 619–625. doi: 10.1016/j.crma.2006.09.019.

[23] Lasry J.-M., Lions P.-L., *Jeux à champ moyen. II - horizon fini et contrôle optimal.* Comptes Rendus Mathématique 343, 679–684 (2006).

[24] Lasry J.-M., Lions P.-L., *Mean field games.* Jpn. J. Math., 2:1 (2007) 229–260. doi:10.1007/s11537-007-0657-8.

[25] Lieb E.H., *Existence and uniqueness of the minimizing solution of Choquard's nonlinear equation.* Studies in Appl. Math. 57 (1976/77), no. 2, 93–105.

[26] Lions P.L., *The Choquard equation and related questions.* Nonlinear Anal. 4 (1980), no. 6, 1063–1072.

[27] Lions P.L., *Compactness and topological methods for some nonlinear variational problems of mathematical physics.* Nonlinear problems: present and future (Los Alamos, N.M., 1981), North-Holland Math. Stud., vol. 61, North-Holland, Amsterdam-New York, 1982, pp. 17–34.

[28] Lions P.L., *The concentration-compactness principle in the calculus of variations. I. The locally compact case.* Ann. Inst. H. Poincaré Anal. Non Linéaire, 1(2):109–145 (1984).

[29] Lions P.L., Cours au Collège de France.

[30] Metafune G., Pallara D., Rhandi A., *Global properties of invariant measures.* J. Funct. Anal., 223(2):396–424 (2005). doi:10.1016/j.jfa.2005.02.001.

[31] Moroz V., Van Schaftingen J., *A guide to the Choquard equation.* J. Fixed Point Theory Appl. 19 (2019), 773–813. doi:10.1007/s11784-016-0373-1.

[32] Moroz V., Van Schaftingen J., *Groundstates of nonlinear Choquard equations: existence, qualitative properties and decay asymptotics.* J. Funct. Anal. 265 (2013), no. 2, 153–184; doi: 10.1016/j.jfa.2013.04.007.

[33] Porretta A., *On the weak theory for mean field games systems.* Boll. Unione Mat. Ital. 10:3 (2017), 411–439.

Reciprocity laws

Eduardo Rocha Walchek [(*)]

Abstract. In this seminar, we give an introduction to reciprocity laws in number theory, since its origins in solving quadratic equations over finite fields to how it evolved – like ever more complex variations on the original theme – to the almost unrecognizable, yet still somehow related, modern explicit reciprocity laws. Our focus will be on the succession of the many results tied by this same name, introducing the relevant concepts and ideas along the way.

1 Introduction

> Zi Gong asked: "Is there any word that could
> guide a person throughout life?"
> The Master replied: "That would be
> *reciprocity*: never impose on others
> what you would not choose for yourself."
>
> Confucius, Analects XV.24

Reciprocity is one of many examples in mathematics where a succession of results evolving from an old concept remain tied to the same name even if the newer iterations don't bear, at first glance, any resemblance to the original idea. In what follows, we will see how a trick to cut down computations needed to figure out whether a quadratic equation has solutions modulo a prime transformed, through ever more complicated variations of this theme, to a study of abelian extensions, and later becoming an important addition to the toolbox of modularity. Although one could list dozens of so called reciprocity laws, we shall focus only in a few major milestones.

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: `eduardo.rochawalchek@phd.unipd.it`. Seminar held on 19 April 2023.

2 The early history of reciprocity (early 1800s)

2.1 Quadratic reciprocity

Let $p > 2$ be a prime[1]. After linear equations modulo p, the natural next step is to try to solve quadratic equations modulo p. This is quite the jump from the linear case, as here solutions are not guaranteed to exist. So one asks about conditions for the solvability of

$$ax^2 + bx + c \equiv 0 \mod p,$$

with $p \nmid a$. Since a and 2 are both invertible modulo p, one can complete squares to reduce to the equivalent problem of solving

$$(2ax + b)^2 \equiv b^2 - 4ac \mod p \iff x^2 \equiv d \mod p.$$

In other words, we want to find conditions for d to be a *quadratic residue* modulo p.

A way to identify quadratic residues was first posed by Legendre (1798), through the Legendre symbol[2]:

Definition 1 (Legendre symbol) Let $p > 2$ be a prime and a an integer such that $\gcd(a, p) = 1$. We define

$$\left(\frac{a}{p}\right) = \begin{cases} 1, & \text{if } a \text{ is a quadratic residue mod } p, \\ -1, & \text{otherwise.} \end{cases}$$

In particular, the Legendre symbol is multiplicative, that is,

$$\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right)$$

(see [IR90, Proposition 5.1.2] for more properties of the Legendre symbol).

Let us consider specificaly equations of the form

$$x^2 \equiv q \mod p,$$

where p and q are two distinct odd primes. Let $f_q(x) \doteq x^2 - q \in \mathbb{Z}[x]$. Since $p \neq q$, then f_q is either irreducible or splits into two different linear factors[3]. The Legendre symbol detects the (ir)reducibility of f_q:

$$f_q \text{ mod } p \in \mathbb{F}_p[x] \text{ splits into linear factors} \iff \left(\frac{q}{p}\right) = 1.$$

We can see this equation from two *reciprocal* perspectives:

[1] As usual, $p = 2$ breaks everything and should be treated separately, so we will take the liberty to leave out any exceptional cases whenever it is convenient in favor of conveying the general ideas.

[2] The historical context of Legendre's study of quadratic residues was in the study of primes of the form $x^2 + ny^2$, a topic also studied by Fermat and Euler. See [Cox13] for more information.

[3] If $p = q$, f_q is the two linear factors are equal, however we will ignore all the ramified cases.

- Fixing p and varying q, the focus is on the base field, $\mathbb{F}_p$. To determine which q are quadratic residues, since there are finitely many classes modulo p, finitely many symbols must be computed, which can be done by a laborious but feasible method known as (one of the many) Gauss Lemma, *cf.* [IR90, p. 52, Lemma].
- Fixing q and varying p, the focus is now on the polynomial $x^2 - q \in \mathbb{Z}[x]$. To determine over which fields $\mathbb{F}_p$ this polynomial has roots, one would need to compute infinitely many symbols: every time we change p, the algorithm must be redone from scratch.

It would be great if, in the second scenario, one could "invert the symbol", so the fixed part is the modulus and, like in the first scenario, only finitely many symbols must be computed and somehow "invert" them back, bypassing the need of infinite computation. This procedure is the Quadratic Reciprocity Law, the origin of reciprocity:

Theorem 2 (Quadratic reciprocity law) *Let $p, q > 2$ be primes. Then*

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2}\cdot\frac{q-1}{2}}$$

[IR90, Chap. 5, §2, Theorem 1].

In practice, one computes $\left(\frac{a}{q}\right)$ for all $1 \leq a \leq q-1$ and, for each p, we have that $\left(\frac{p}{q}\right) = \left(\frac{a}{q}\right)$ for some $a \equiv p \mod q$. Theorem 2 then gives us $\left(\frac{q}{p}\right)$ for all p.

The quadratic reciprocity law is due to Gauss (1801)[(i)]. Gauss also noticed that the "natural habitat" of quadratic reciprocity is $\mathbb{Z}[i]$, as a consequence of biquadratic reciprocity over $\mathbb{Z}[i]$ (with a biquadratic residue symbol taking values in quartic roots of unity), which he stated, but never published a proof of it (1828[(ii)], 1830[(iii)]). The first published proof is due to Eisenstein (1844)[(iv)]. For the biquadratic reciprocity law, which we will not consider here, we refer the reader to [IR90, Chap. 9, §7–10].

2.2 The reciprocity problem

In hopes of generalizing the quadratic reciprocity law, the leitmotif to guide all the next "classical" reciprocity laws is:

Problem 3 (Reciprocity) *Given $f \in \mathbb{Z}[x]$, how does f split modulo a prime p? In the case where $f(x) = x^n - a$, how to compute the associated residue symbol? What is the best ring over which this question can be considered?*

[(i)] C. F. Gauss, *Disquisitiones Arithemeticae*, 1801, §151, where Gauss refers to it as a "fundamental theorem".

[(ii)] C. F. Gauss, *Theoria residuorum biquadraticorum, Commentatio prima*, 1828.

[(iii)] C. F. Gauss, *Theoria residuorum biquadraticorum, Commentatio secunda*, 1832.

[(iv)] F. G. Eisenstein, *Lois de réciprocité*, J. Reine Angew. Math. 28, 53–67, 1844.

2.3 Cubic reciprocity

Eisenstein (1844, *ibid.*) also proved the cubic reciprocity, which happens naturally in $\mathbb{Z}[\zeta_3]$, where ζ_3 is a primitive cubic root of the unity. Let μ_3 be the set of cubic roots of unity.

In $\mathbb{Z}[\zeta_3]$, a rational prime p can be

- Ramified, only occours for $p = 3 = \zeta_3^2 \cdot (1+\zeta_3)^2$ (we will ignore this case);
- Split into a product of two (conjugate) primes, if $p \equiv 1 \mod 3$;
- Inert, remaining a prime, if $p \equiv 2 \mod 3$.

Similarly to how the Legendre symbol take values in square roots of unity, the cubic power symbol should take values in cubic roots of unity. Let π be a prime with $\gcd(\pi, 3) = 1$ (or $N\pi \doteq \pi\bar{\pi} \neq 3$) and $a \in \mathbb{Z}[\zeta_3]$ such that $\pi \nmid a$. By Fermat's Little Theorem (a.k.a. the fact that $(\mathbb{Z}[\zeta_3]/\pi\mathbb{Z}[\zeta_3])^\times$ is a group of order $N\pi - 1$),

$$a^{N\pi-1} \equiv 1 \mod \pi \implies a^{(N\pi-1)/3} \equiv \zeta_3^k \mod \pi.$$

Definition 4 (Cubic residue symbol) Let π be a prime such that $N\pi \neq 3$ and $\pi \nmid a$. Define

$$\left(\frac{a}{\pi}\right)_3 \doteq \zeta_3^k,$$

where k is in the unique residue class modulo 3 defined as above.

Once again, the symbol is multiplicative and detects if a is a cubic residue or not:

$$\left(\frac{a}{\pi}\right)_3 = 1 \iff x^3 \equiv a \mod \pi \text{ has a solution}$$

[IR90, Proposition 9.3.3].

The units in $\mathbb{Z}[\zeta_3]$ are six: ± 1, $\pm\zeta_3$ and $\pm\zeta_3^2$. Since the cubic residue symbol can yield different values for associated primes, we would like to choose one among the associates of a prime to state a reciprocity law free of ambiguities. Direct calculation shows that:

Proposition 5 *Let π be a prime with $N\pi \neq 3$. Then there exists only one associate π' of π such that $\pi \equiv 2 \mod 3$.*

We call *primary* a prime π which is its only associate satisfying the condition above.

Theorem 6 (Cubic reciprocity law) *If π_1 and π_2 are two primary primes,*

$$\left(\frac{\pi_1}{\pi_2}\right)_3 = \left(\frac{\pi_2}{\pi_1}\right)_3,$$

[IR90, Chap. 9, §3, Theorem 1].

2.4 Eisenstein cyclotomic reciprocity

Cubic reciprocity is a special case of cyclotomic reciprocity, also due to Eisenstein (1850)[(v)]. We first define the mth power symbol in general, and then specialize the discussion to $m = \ell$ a prime.

Let $m \geq 2$ be an integer and ζ_m a primitive mth root of unity. Let $\mathfrak{p} \subseteq \mathbb{Z}[\zeta_m]$ be a prime ideal that is prime to m, that is, $m \notin \mathfrak{p}$ (this way all primes in $\mathbb{Z}[\zeta_m]$ above ramified rational primes are excluded). Again Fermat's Little Theorem implies

$$\begin{aligned} a^{N\mathfrak{p}-1} \equiv 1 \mod \mathfrak{p} &\implies \prod_{i=0}^{m-1} (a^{(N\mathfrak{p}-1)/m} - \zeta_m^i) \equiv 0 \mod \mathfrak{p} \\ &\implies a^{(N\mathfrak{p}-1)/m} \equiv \zeta_m^k \mod \mathfrak{p}. \end{aligned}$$

Definition 7 (Power residue symbol) Let $\mathfrak{p}$ be a prime such that $m \notin \mathfrak{p}$ and $a \in \mathbb{Z}[\zeta_m] \setminus \mathfrak{p}$. Define

$$\left(\frac{a}{\mathfrak{p}}\right)_m \doteq \zeta_m^k,$$

where k is in the unique residue class modulo m defined as above.

Being a Dedekind domain, any principal ideal $(b) = \prod_i \mathfrak{p}_i^{e_i}$, with $b \in \mathbb{Z}[\zeta_m]$ prime to m factors as a product of primes prime to m, and we can define

$$\left(\frac{a}{b}\right)_m \doteq \left(\frac{a}{(b)}\right)_m \doteq \prod_i \left(\frac{a}{\mathfrak{p}_i}\right)_m^{e_i}.$$

More details can be found at [IR90, Chap. 14, §2].

From now to the end of the section, take $m = \ell$ to be a rational prime. Call *primary* a non-unity a that is prime to ℓ and congruent to a rational integer mod $(1 - \zeta_\ell)^2$. Again, this is a uniqueness condition to avoid ambiguity coming from the many associates of a prime [IR90, p. 206, Lemma].

Theorem 8 (Eisenstein cyclotomic reciprocity) *If a is prime to ℓ and b is primary and prime to a,*

$$\left(\frac{a}{b}\right)_\ell = \left(\frac{b}{a}\right)_\ell$$

[IR90, Chap. 13, §2, Theorem 1].

Thus one can describe the behavior of the splitting of the cyclotomic polynomial:

Corollary 9 *The cyclotomic polynomial*

$$\Phi_\ell(x) = \frac{x^\ell - 1}{x - 1} = x^{\ell-1} + \cdots + x + 1$$

[(v)] F. G. Eisenstein, *Beweis der allgemeinsten Reciprocitäsgesetze zwischen reellen und komplexen Zahlen*, Verhandlungen der Königlich Preußische Akademie der Wissenschaften zu Berlin, 189–198, 1850.

splits into distinct linear factors modulo p if and only if $p \equiv 1 \mod \ell$.[(4)]

3 Reciprocity from power equations to field extensions (late 1800s)

3.1 Kummer cyclotomic reciprocity

At around the same time Eisenstein was proving the reciprocity laws stated above, the well known Fermat's Last Theorem, on the non-existence of non-trivial integral solutions to the equation $x^n + y^n = z^n$ for $n \geq 3$, at the time a 200 years old open problem, had already seen many wrong or unfruitful attempts of a proof, but one particular wrong assumption in a proof by Lamé (1847)[(vi)] was a stumble into something way greater than if first seemed: the concept of ideals, and with it the branch of number theory we nowadays call algebraic, was about to see the light of the day.

Lamé's attempted proof relied heavily on the uniqueness of factorization in $\mathbb{Z}[\zeta_\ell]$, which is not true in general. Had it been true, one could write from a non-trivial solution two different factorizations of the same thing:

$$z^\ell = \prod_{i=1}^{\ell}(x + \zeta_\ell^i y) = x^\ell + y^\ell.$$

This inspired Kummer to fix Lamé's the proof in the case of so called *regular primes*, considering instead factorizations of ideals.

The ring of integers $\mathcal{O}_K$ of a number field[(5)] K is an example of *Dedekind domain*, where non-zero prime ideals are maximal, and fractional ideals[(6)] factor uniquely as a product of finitely many maximal ideals with integral coefficients [Neu99, p. 22, Corollary 3.9]. Therefore, the set of fractional ideals with the usual product is a (free) abelian group $J_{\mathbb{Q}(\zeta_\ell)}$. The subset $P_{\mathbb{Q}(\zeta_\ell)}$ of principal ideals is a subgroup, and the quotient $C\ell_{\mathbb{Q}(\zeta_\ell)} = J_{\mathbb{Q}(\zeta_\ell)}/P_{\mathbb{Q}(\zeta_\ell)}$ is a finite group, called the *class group* of $\mathbb{Q}(\zeta_\ell)$. Its cardinality $h_{\mathbb{Q}(\zeta_\ell)}$ is the *class number* of $\mathbb{Q}(\zeta_\ell)$, *cf.* [Neu99, p. 36, Theorem 6.3].

Definition 10 A prime ℓ is said to be regular if ℓ does not divide the class number of $\mathbb{Q}(\zeta_\ell)$.

In the case that ℓ is a regular prime, Kummer proved the following statement, which is the key for proving Fermat's Last Theorem in the case $n = \ell$ ([Rib79, p. 86, §3A]):

Lemma 11 *If α is a primary unit in $\mathbb{Z}[\zeta_\ell]$, then $\alpha = \beta^\ell$, where ℓ is another unit* [Rib79, p. 86, Lemma 3.4].

[(4)] This statement also holds for the mth cyclotomic polynomial when m is not a prime. See [Wym72, p. 575] for details.

[(5)] By a *number field* we mean a finite extension of $\mathbb{Q}$.

[(6)] A *fractional ideal* $\mathfrak{f}$ of a Dedekind domain $\mathcal{O}$ is $d^{-1} \cdot \mathfrak{a}$, where $d \in \mathcal{O} \setminus \{0\}$ and $\mathfrak{a}$ is an ideal of $\mathcal{O}$.

[(vi)] G. Lamé, *Démonstration générale du théorème de Fermat sur l'impossibilité en nombres entiers de l'equation $x^n + y^n = z^n$*, C. R. Acad. Sci. Paris, 24, 310–314, 1847.

Back to reciprocity, suppose ℓ is a regular prime and let h be the class number of $\mathbb{Q}(\zeta_\ell)$. If $\mathfrak{p}$ is a prime ideal in $\mathbb{Z}[\zeta_\ell]$ not divisible by ℓ, $\mathfrak{p}^h$ is trivial in the class group of $\mathbb{Q}(\zeta_\ell)$ and therefore is principal, let us say $\mathfrak{p}^h = (\alpha)$, for some $\alpha \in \mathbb{Z}[\zeta_\ell]$, that we may again assume to be primary. Since $\ell \nmid h$, there exists $h' \in \mathbb{Z}_{>0}$ such that $h' \equiv h^{-1} \mod \ell$. Informally, this is a way to invert h in $\mathfrak{p}^h = (\alpha)$, which motivates the following definition:

Definition 12 (Kummer symbol) Let $\mathfrak{p}$ be as above and $\mathfrak{q}$ be another prime ideal prime to both $\mathfrak{p}$ and ℓ. We define

$$\left(\frac{\mathfrak{p}}{\mathfrak{q}}\right)_\ell \doteq \left(\frac{\alpha^{h'}}{\mathfrak{q}}\right)_\ell .$$

The above definition independs of the choice of α and h', by the previous lemma. These symbols satisfying the following reciprocity law, extending Eisenstein's:

Theorem 13 (Kummer cyclotomic reciprocity, 1858[(vii)]) *If $\mathfrak{p} \neq \mathfrak{q}$ are prime ideals that are prime to ℓ,*

$$\left(\frac{\mathfrak{p}}{\mathfrak{q}}\right)_\ell = \left(\frac{\mathfrak{q}}{\mathfrak{p}}\right)_\ell .$$

3.2 Brief account on Kummer field theory

With Kummer's reciprocity law, the focus of reciprocity starts to shift from the polynomial itself to its splitting field, that is, the field generated by all of its roots.

Definition 14 (Kummer extension) A Kummer extension is an abelian Galois extension L/K of (Galois group of) exponent n, where K is a field containing all nth roots of unity (we always assume $n \nmid \mathrm{char}(K)$ unless otherwise stated).

The main example (and the motivation to consider such extensions) is the splitting field of $f(x) = x^n - a \in K[x]$.

Let L/K be a Galois extension. There is an exact sequence

$$0 \longrightarrow \mu_n \longrightarrow L^\times \xrightarrow{\cdot^n} (L^\times)^n \longrightarrow 0.$$

The long sequence of Galois cohomology gives ([Sil09, p. 419, Proposition B.2.3])

$$0 \longrightarrow K^\times \cap (L^\times)^n/(K^\times)^n \longrightarrow \mathrm{H}^1(\mathrm{Gal}(L/K), \mu_n) \longrightarrow \mathrm{H}^1(\mathrm{Gal}(L/K), L^\times).$$

Since the Galois action over μ_n is transitive, the third term of the above sequence is $\mathrm{Hom}(\mathrm{Gal}(L/K), \mu_n)$, and the fourth term is zero, by Hilbert's theorem 90 ([MilFT, p. 71]). Therefore, we have an isomorphism

$$\frac{K^\times \cap (L^\times)^n}{(K^\times)^n} \cong \mathrm{Hom}(\mathrm{Gal}(L/K), \mu_n)\colon a = \alpha^n \text{ (for } \alpha \in L) \mapsto \left(\sigma \mapsto \frac{\sigma(\alpha)}{\alpha}\right),$$

[(vii)] E.E. Kummer, *Über die allgemeinen Reziprozitätsgesetze der Potenzreste*, Ber. K. Akad. Wiss. Berlin, 158–171, 1858.

see [MilFT, p. 75].

Theorem 15 (Main theorem of Kummer theory) *Let K be a field containing all nth roots of unity ($n \nmid \mathrm{char}(K)$). There's a 1-to-1 correspondence between*

- *Abelian extensions L/K of exponent n;*
- *Subgroups Δ of $K^\times$ containing $(K^\times)^n$.*

An extension L corresponds to $\Delta_L = K^\times \cap (L^\times)^n$, and a subgroup Δ corresponds to $L_\Delta = K(\Delta^{1/n})$ [MilFT, p. 75, Theorem 5.30].

In sum, $K^\times/(K^\times)^n$ tells us about exponent n abelian extensions of K. The following is the "converse" of the motivating example:

Corollary 16 *Every n-cyclic extension of a field K containing all nth roots of unity is the splitting field of $x^n - a$ for some $a \in K$.*

4 Reciprocity in the genesis of Class Field Theory (early 1900s)

4.1 Remarks on global and local fields

We make this quick intermission to recall the concepts of global and local fields.

A *place* in a normed field K is a class of topologicaly equivalent norms. A norm $|\cdot|$ is said to be archimedean or not depending on whether they *don't* or do satisfy the strong triangle inequality

$$|x+y| \leq \max\{|x|, |y|\}.$$

Very informaly, a *global field* K is a jack of all trades, but master of none: it has many places, but none of them are complete. Completing a global field K respect to a place v gives a *local field* K_v, the best extension of K where one can consider v, but the tradeoff is that any other place $w \neq v$ won't extend in a meaningful way to K_v.

For us, the relevant example of a global field is a number field $K/\mathbb{Q}$.[7] Ostrowski's theorem gives a complete list of all places existing in K:

Theorem 17 (Ostrowski) *Let K be a number field. There are bijections between:*

- *Prime ideals of $\mathcal{O}_K$ and non-archimedean places of K: the prime 0 is related to the trivial norm and a non-zero, thus maximal, prime ideal $\mathfrak{p}$ of $\mathcal{O}_K$ corresponds to the $\mathfrak{p}$-adic norm*

  $$|x|_{\mathfrak{p}} = (N\mathfrak{p})^{-v_{\mathfrak{p}}(x)},$$

  where $v_{\mathfrak{p}}(x)$ is the exponent of $\mathfrak{p}$ in the factorization of the fractional ideal (x);
- *Embeddings of K into $\mathbb{R}$ or $\mathbb{C}$ and archimedean places: each embedding σ gives the norm $|\sigma(\cdot)|_\infty$, where $|\cdot|_\infty$ is the usual norm in $\mathbb{R}$ or $\mathbb{C}$ (complex embeddings come in conjugated pairs, both giving the same place)* [Con].

[7] A *global field* is always a number field or a function field of an algebraic curve over a finite field.

It is common to say that archimedean places in a number field come from primes at infinity, as the non-archimedean places come from finite primes. More precisely, for $K = \mathbb{Q}$, we have for each prime p a p-adic non-archimedean norm given by

$$|\pm p^a p_1^{a_1} \cdots p_n^{a_n}|_p \doteq p^{-a}$$

and the usual norm $|\cdot|_\infty$, the only archimedean place[(8)]. The completion of $\mathbb{Q}$ at $|\cdot|_p$ gives $\mathbb{Q}_p$, the field of p-adic numbers, whose ring of integers is $\mathbb{Z}_p$, the ring of p-adic integers, which is also the p-adic closure of $\mathbb{Z}$ (for more information on p-adic numbers, see [Gou97]). The completion of $\mathbb{Q}$ at $|\cdot|_\infty$ is $\mathbb{Q}_\infty = \mathbb{R}$.

Although $\mathbb{R}$ and $\mathbb{C}$ are completions of number fields at places, they are often not considered local fields, as usually this label is reserved to non-archimedean local places, which, differently from the archimedean local fields, are fraction fields of *discrete valuation rings* (DVR), a well-desired property in this context.

Proposition 18 (Discrete valuation rings) *Let R be a ring and $K = \mathrm{Frac}(R)$. The following are equivalent:*

- *There exists a group homomorphism $v\colon K^\times \to \mathbb{Z}$ such that*

  $$v(x+y) \geq \min\{v(x), v(y)\},$$

  that is, a discrete valuation, and $R = \{x \in K^\times;\ v(x) \geq 0\} \cup \{0\}$;

- *R is a local principal ideals domain;*

- *There exists an irreducible $\pi \in R$ such that any $z \in K$ can be written uniquely as $z = u\pi^n$, with $n \in \mathbb{Z}$ and $u \in R^\times$, and $z \in R \iff n \geq 0$.*

A ring satisfying the above equivalent properties is said to be a discrete valuation ring [AM69, p. 94, Proposition 9.2].

Remark 19 If v is a discrete valuation in a field K, $|\cdot|_v = c^{-v(\cdot)}$ is a non-archimedean norm, for any $c > 1$. This, in fact, gives a 1-to-1 correspondence between finite places of a global field and discrete valuations on it, and both concepts are often used interchangeably.

The irreducible at the last item in the list above is called a *local parameter* of K_v and it generates the only maximal ideal $\mathfrak{m}_v$ of $\mathcal{O}_{K,v}$. The quotient $\mathcal{O}_{K,v}/\mathfrak{m}_v$, called the *residue field* of K_v, is a finite field of characteristic p when K is a number field and v is a place over the rational prime p. One key property of local fields is in how they relate to their residue fields: simple roots of integral polynomials modulo $\mathfrak{m}_v$ lift to integral solutions in K_v:

Proposition 20 (Hensel's Lemma) *Let K be a number field, v be a finite place in K, and $f \in \mathcal{O}_{K,v}[T]$ a polynomial. Let $\overline{f} \in k_v[T]$ be the reduction of f modulo $\mathfrak{m}_v$. If there is $\overline{a} \in k_v$ such that $\overline{f}(\overline{a}) = \overline{0}$ and $\overline{f}'(\overline{a}) \neq \overline{0}$, then there is a unique $\alpha \in \mathcal{O}_v$ such that $f(\alpha) = 0$ and $\overline{\alpha} = \overline{a}$* [EP05, p. 22, Corollary 1.3.2].

[(8)] This references the fact that all non-trivial norms over $\mathbb{R}$ are equivalent.

Another curious property is that the local field is a ring of meromorphic series on the local parameter (compare with the last item in Proposition 18):

Proposition 21 *Let K be a local field with a (non-archimedean) place v, a local parameter π and R a set of representatives in $\mathcal{O}_{K,v}$ of the residue classes of the quotient $\mathcal{O}_{K,v} \twoheadrightarrow \mathcal{O}_{K,v}/(\pi)$. Then every element $x \in K^\times$ can be uniquely written as a convergent series (in the norm induced by the valuation v)*

$$x = \sum_{i=v(x)}^{\infty} r_i \pi^i,$$

with coefficients $r_i \in R$ [EP05, p. 23, Proposition 1.3.5].

4.2 Hilbert reciprocity law

Quadratic residues are just one particular case of the study of quadratic forms over finite fields (or the reduction modulo primes of a quadratic form on a global or local field). Another particular case would be the study of *norm residues*, the "two-dimensional" version of the problem. Let K be a local field.

Definition 22 (Hilbert symbol) For $a, b \in K$, define

$$(a,b) \doteq \begin{cases} 1, & \text{if } ax^2 + by^2 = z^2 \text{ has non-zero solutions,} \\ -1, & \text{otherwise.} \end{cases}$$

An equivalent way to write the Hilbert symbol is in terms of subgroups $K^\times/(K^\times)^2$, which ties back to the main theorem of Kummer theory:

Proposition 23 $(a,b) = 1 \iff a \in N_b \doteq \{z^2 - by^2 \neq 0;\ y, z \in K\}$*, a subgroup of $K^\times$ containing $(K^\times)^2$.*

Let K be a number field, v be a place and denote by $(\cdot,\cdot)_v$ the Hilbert symbol in K_v. The following lemma shows that, for a pair $a, b \in K$, for most places v the associated symbol $(a,b)_v$ is 1:

Lemma 24 *For $a, b, c \in \mathcal{O}_{K,\mathfrak{p}}^\times$, $ax^2 + by^2 = c$ has a solution in $\mathcal{O}_{K,\mathfrak{p}}$.*

Proof. The idea is to find a solution modulo $\mathfrak{p}$ and use Hensel's Lemma (Proposition 20) to lift it to a solution in $\mathcal{O}_{K,\mathfrak{p}}$.

The residue field $\mathcal{O}_{K,\mathfrak{p}}/\mathfrak{p}\mathcal{O}_{K,\mathfrak{p}}$ is a finite extension of $\mathbb{F}_p$, thus is of the form $\mathbb{F}_q$ with $q = p^r$, where there are $(q+1)/2$ non-zero residue classes[(9)]. Thus both the sets of classes of the form $ax^2 \bmod \mathfrak{p}\mathcal{O}_{K,\mathfrak{p}}$ and the set of classes of the form $c - by^2 \bmod \mathfrak{p}\mathcal{O}_{K,\mathfrak{p}}$ are of

[(9)] In odd characteristic, $\mathbb{F}_q^\times \to \mathbb{F}_q^\times : a \mapsto a^2$ is a morphism of degree 2 kernel, giving $(q-1)/2$ non-zero residue classes, plus 0.

cardinality $(q+1)/2$, totalizing $q+1$ classes. By the pigeonhole principle, since $\#\mathbb{F}_q = q$, there is a class in both sets, giving

$$a\bar{x}_0^2 + b\bar{y}_0^2 \equiv c \mod \mathfrak{p}\mathcal{O}_{K,\mathfrak{p}}.$$

Suppose without loss of generality that $\bar{x}_0 \neq 0$. Then lift $\bar{y}_0$ to any $y_0 \in \mathcal{O}_{K,\mathfrak{p}}$ and use Hensel's Lemma (Proposition 20) to lift $\bar{x}_0$ to $x_0 \in \mathcal{O}_{K,\mathfrak{p}}$ such that

$$ax_0^2 + by_0^2 = c.$$

□

Corollary 25 *Given $a, b \in K$, for all but finitely many places, $(a,b)_v = 1$.*

Proof. Apply the above lemma for a place not over 2 and not dividing a or b. □

The remaining places balance themselves out, as shown by Hilbert (1897)[(viii)]:

Theorem 26 (Hilbert reciprocity law) *For $a, b \in K$, $\prod_v (a,b)_v = 1$, where the product ranges over all (finite and infinite) places of K.*

Let us consider the case $K = \mathbb{Q}$ in particular. Take $p \neq q$ to be two odd rational primes. From Lemma 24, $(p,q)_\ell = 1$ for all $\ell \notin \{p,q\}$.

Lemma 27 *For $a \in \mathbb{Z}_p^\times$, $(a,p)_p = \left(\dfrac{a}{p}\right)$, the Legendre symbol.*

Proof. The equation $ax^2 + py^2 = z^2$ has a non-zero solution if and only if $a \equiv (z/x)^2 \mod p$ is a quadratic residue, so the lemma follows from the definitions. □

Therefore, $(p,q)_q = \left(\dfrac{p}{q}\right)$ and $(p,q)_p = (q,p)_p = \left(\dfrac{q}{p}\right)$. Direct calculation gives

Lemma 28 *For $a, b \in \mathbb{Q}_2$, $(a,b)_2 = (-1)^\lambda$, where*

$$\lambda = \left(\tfrac{2^{-v_2(a)}a-1}{2}\right)\left(\tfrac{2^{-v_2(b)}b-1}{2}\right) + v_2(a)\left(\tfrac{(2^{-v_2(b)}b)^2-1}{8}\right) + v_2(b)\left(\tfrac{(2^{-v_2(a)}a)^2-1}{8}\right)$$

and v_2 is the 2-adic valuation.

In particular, $(p,q)_2 = (-1)^\lambda$, with $\lambda = \left(\frac{p-1}{2}\right)\left(\frac{q-1}{2}\right)$. Lastly, $(a,b)_\infty = 1 \iff a > 0$ or $b > 0$.

Corollary 29 *Hilbert's reciprocity law over $\mathbb{Q}$ implies the quadratic reciprocity law.*

[(viii)] D. Hilbert, *Die Theorie der algebraischen Zahlkörper*, Jahresbericht der Deutschen Mathematiker-Vereinigung, 4: 175–546, 1897.

4.3 Artin reciprocity law

Three years later, Hilbert publishes his famous list of problems. Among them, the ninth one reads:

Problem 30 (Hilbert's 9th problem) *Find the most general reciprocity law for the norm residues of kth order in a general algebraic number field, where k is a power of a prime.*

The abelian case was solved by Artin (1927[ix], 1930[x]), and is in the genesis of Class Field Theory, the study of abelian field extensions. Below we state the local version of this result (for the global version see [MilCFT, p. 158, Theorem 3.5]):

Theorem 31 (Artin's local reciprocity law) *Let K be a (non-archimedean) local field. There's an isomorphism*

$$\varphi_K \colon K^\times \xrightarrow{\sim} \operatorname{Gal}(K^{\mathrm{ab}}/K),$$

where K^{ab} is the maximal abelian extension of K, that is, the union of all finite abelian extensions of K. If L/K is one such extension, φ_K restricts to an isomorphism

$$\varphi_{L/K} \colon K^\times/N_{L/K}(L^\times) \xrightarrow{\sim} \operatorname{Gal}(L/K),$$

where $N_{L/K}(a) = \prod_{\sigma\in\operatorname{Gal}(L/K)} \sigma(a)$ [MilCFT, p. 20, Theorem 1.1].

4.4 Generalized Hilbert reciprocity law

Let K be a local field containing all nth roots of unity (with $\operatorname{char}(K) \nmid n$) and $L = K(\sqrt[n]{K^\times})$ be the maximal abelian extension of K with exponent n.

On one hand, Artin reciprocity gives us a canonical isomorphism

$$K^\times/(K^\times)^n = K^\times/N_{L/K}(L^\times) \cong \operatorname{Gal}(L/K).$$

On the other hand, we saw that Kummer theory implies

$$K^\times/(K^\times)^n \cong \operatorname{Hom}(\operatorname{Gal}(L/K), \mu_n).$$

Therefore, we have a non-degenerate bilinear pairing

$$(\cdot,\cdot)\colon K^\times/(K^\times)^n \times K^\times/(K^\times)^n \to \mu_n,\ (\sigma,\phi) = \phi(\sigma),$$

which coincides with the Hilbert symbol for $n = 2$: in fact, $ax^2 + by^2 = z^2$ has a non-zero solution if and only if b is a norm of some element in $K(\sqrt{a})$.[10]

[10] If $ax^2 + by^2 = z^2$ has a non-zero solution, assume, switching a and b if necessary, that $y \neq 0$. Then $b = \left(\frac{z}{y}\right)^2 - a\left(\frac{x}{y}\right)^2 = N\left(\frac{z+\sqrt{a}x}{y}\right)$. Conversely, $b = N(\alpha + \beta\sqrt{a}) = \alpha^2 - a\beta^2 \implies a\beta^2 + b\cdot 1^2 = \alpha^2$.

[ix] E. Artin, *Beweis des allgemeinen Reziprozitäsgesetzes*, Abh. Math. Semin. Univ. Hamb. 5: 353–363, 1927.

[x] E. Artin, *Idealklassen in Oberkörpern und allgemeines Reziprozitätsgesetzes*, Abh. Math. Semin. Univ. Hamb. 7: 46–51, 1930.

The generalized Hilbert symbol relates to the nth power residue symbol the same way the previous Hilbert symbol relates to the Legendre symbol. The following is the corresponding statement of Lemma 27 in this scenario:

Lemma 32 *For $a \in \mathcal{O}_{K,\mathfrak{p}}^\times$ and π a local parameter of $\mathcal{O}_{K,\mathfrak{p}}$, $(a,\pi)_\mathfrak{p} = \left(\dfrac{a}{\mathfrak{p}}\right)_n$.*

Lemmas 24 and 32, together with the Hilbert reciprocity law, imply:

Corollary 33 (Generalized Hilbert reciprocity law) *For $a, b \in K$,*

$$\left(\frac{a}{b}\right)_n = \left(\frac{b}{a}\right)_n \prod_{\mathfrak{p}|n,\infty} (a,b)_\mathfrak{p}.$$

5 Explicit reciprocity laws

In parallel to the formulas relating the symbols among themselves, the natural but more difficult problem of describing the symbols themselves *explicitely* has also been studied. We saw above how some symbols can be easily computed, but some, for example the generalized Hilbert symbol for $\mathfrak{p} \mid n$, can be trickier to compute.

The earliest notable result in this vein dates back to Kummer[(xi)]:

Theorem 34 (Kummer) *Let $K = \mathbb{Q}_\ell(\zeta_\ell)$, for $\ell \neq 2$, and π be a local parameter of $\mathcal{O}_K$. For $a, b \in 1 + (\pi)$,*[(11)] *we have that*

$$(a,b) = \zeta_\ell^{\mathrm{res}(\log(a(\pi))\cdot \mathrm{dlog}(b(\pi))\cdot \pi^{-\ell})},$$

where a and b are seen as power series on π (cf. Proposition 21*),* log *is the ℓ-adic logarithm* (defined analogously as in [Gou97, p. 112, Definition 4.5.2])*,* dlog *is the logarithmic derivative* $\mathrm{dlog}(f) = f'/f$ *and* res *denotes the formal residue, that is, the coefficient of index -1. See e.g.* [Vos00], [Li].

After Kummer, Artin–Hasse (1928)[(xii)], later extended by Iwasawa (1968)[(xiii)], proved what is known as the *classical explicit reciprocity law.*

Remark 35 Vostokov ([Vos00]) points out two different lines of explicit reciprocity laws: the one with an approach more similar to the original one by Kummer, in which Shafarevich

[(11)] These are the so called *principal units.* In fact, $\mathcal{O}_K^\times = (1 + (\pi)) \times (\mathcal{O}_K/(\pi))^\times$, and seeing the latter factor as the ℓth roots of unity, we have that every unit in $\mathcal{O}_K^\times$ is the product of one a principal unit and an ℓth root of unity.

[(xi)] E.E. Kummer, *Über die allgemeinen Reziprozitätsgesetze der Potenzreste*, Ber. K. Akad. Wiss. Berlin, 158–171, 1858.

[(xii)] E. Artin, H. Hasse, *Die beiden Ergänzungssatz zum Reziprozitätsgesetz der ℓ^n-ten Potenzreste im Körper der ℓ^n-ten Einheitswurzeln*, Abh. Math. Sem. Univ. Hamb. 6: 146–162, 1928.

[(xiii)] K. Iwasawa *On explicit formulas for the norm residue symbols*, J. Math. Soc. Japan 20, 151–165, 1968.

(1950)[xiv], later extended by Brückner (1979)[xv] and Vostokov (1980)[xvi], have obtained similar explicit formulas for the Hilbert symbol for finite extensions of $\mathbb{Q}_\ell(\zeta_\ell)$; and the one starting from the approach of Artin–Hasse–Iwasawa, the line which modern reciprocity laws usually follow.

5.1 Classical explicit reciprocity law

Consider the tower of fields $K_n = \mathbb{Q}_\ell(\zeta_{\ell^n})$ and a norm-compatible set of roots of unity $(\zeta_{\ell^n})_n$. By mapping $\zeta_{\ell^n} \mapsto 1$, we can write the generalized Hilbert symbol as

$$(\cdot,\cdot)_n \colon K_n^\times \times K_n^\times \to \mathbb{Z}/\ell^n\mathbb{Z}.$$

Fixing the first term and taking inverse limit in the second (respect to the norm for the fields), we get a pairing

$$(\cdot,\cdot)\colon K_m^\times \times \varprojlim_n K_n^\times \longrightarrow \varprojlim_n \mathbb{Z}/\ell^n\mathbb{Z} = \mathbb{Z}_\ell,$$

which we can rewrite as a map

$$\varprojlim_n K_n^\times \longrightarrow \mathrm{Hom}_{\mathrm{cont}}(K_m^\times, \mathbb{Z}_\ell)\colon \alpha \mapsto (-,\alpha),$$

where K_m has the ℓ-adic norm topology and $\mathbb{Z}_\ell$ is discrete. The ℓ-adic exponential map $\exp\colon K_m \to K_m^\times \otimes_{\mathbb{Z}_\ell} \mathbb{Q}_\ell$ (*cf.* [Gou97, Definition 4.5.6]) induces a pullback

$$\exp^*\colon \mathrm{Hom}_{\mathrm{cont}}(K_m^\times, \mathbb{Z}_\ell) \longrightarrow \mathrm{Hom}_{\mathrm{cont}}(K_m, \mathbb{Q}_\ell)\colon \phi \mapsto \phi \circ \exp.$$

Consider the composition map

$$\lambda_m\colon \varprojlim_n K_n^\times \xrightarrow{(\cdot,\cdot)} \mathrm{Hom}_{\mathrm{cont}}(K_m^\times, \mathbb{Z}_\ell) \xrightarrow{\exp^*} \mathrm{Hom}_{\mathrm{cont}}(K_m, \mathbb{Q}_\ell) \xrightarrow{\sim} K_m,$$

where the last isomorphism is

$$K_m \longrightarrow \mathrm{Hom}_{\mathrm{cont}}(K_m, \mathbb{Q}_\ell)\colon x \mapsto \mathrm{Tr}_{K_m/\mathbb{Q}_\ell}(x \cdot -).$$

Let $R = \mathbb{Z}_\ell[\![t-1]\!]$ be the ring of formal power series in $t-1$ with coefficients in $\mathbb{Z}_\ell$ and the norm map $N\colon R^\times \to R^\times\colon f(t) \mapsto \displaystyle\prod_{i=1}^{\ell} f(\zeta_\ell^i t^{1/\ell})$ induced by $t \mapsto t^\ell$.

Theorem 36 (Classical explicit reciprocity law, Artin–Hasse, Iwasawa) *Let $(\theta_n(t))_{n\geq 1}$ be a norm-compatible sequence in $R^\times$. Then $(\theta_n(\zeta_{\ell^n}))_n$ is a norm-compatible sequence in $\varprojlim_n K_n^\times$ and we have the explicit formula*

$$\lambda_m(u) = \ell^{-m}\zeta_{\ell^m} \cdot \mathrm{dlog}\,\theta_m(\zeta_{\ell^m}),$$

see e.g. [Li].

[xiv] I. R. Shafarevich, *A general reciprocity law*, Mat. Sbornik, 26: 113–146, 1950.

[xv] H. Brückner, *Explizites Reziprozitätsgesetz und Anwendungen*, Vorlesungen aus dem Fachbereich Mathematik der Universität Essen, vol. 2, 1979.

[xvi] S. V. Vostokov, *An explicit form of the reciprocity law*, Izv. Akad. Nauk SSSR Ser. Mat., 42 (6): 1288–1321, 1980.

5.2 Another look at classical explicit reciprocity

Consider the map from before:

$$\lambda_m \colon \varprojlim_n K_n^\times \xrightarrow{(\cdot,\cdot)} \mathrm{Hom}_{\mathrm{cont}}(K_m^\times, \mathbb{Z}_\ell) \xrightarrow{\exp^*} \mathrm{Hom}_{\mathrm{cont}}(K_m, \mathbb{Q}_\ell) \xrightarrow{\sim} K_m.$$

Since $\mathrm{Gal}(\overline{K}_m/K_m)$ acts trivially over $\mu_{\ell^n} \cong \mathbb{Z}/\ell^n\mathbb{Z}$, we have that

$$\mathrm{Hom}_{\mathrm{cont}}(\mathrm{Gal}(\overline{K}_m/K_m), \mu_{\ell^n}) \cong H^1(K_m, \mu_{\ell^n}),$$

and as we have seen, $K_m^\times/(K_m^\times)^{\ell^n} \cong \mathrm{Gal}(\overline{K}_m/K_m)$. Since the Hilbert symbol maps into continuous homomorphisms $K_m^\times \to \mu_{\ell^n}$ killing $(K_m^\times)^{\ell^n}$, taking inverse limits the first map above becomes

$$\varprojlim_n K_n^\times \longrightarrow H^1(K_m, \mathrm{Ta}_\ell \mathbb{G}_m),$$

where $\mathrm{Ta}_\ell \mathbb{G}_m \doteq \varprojlim_n \mu_{\ell^n}$ is the Tate module of the multiplicative group $\mathbb{G}_m$. The ℓ-adic exponential map induces a map

$$\exp^* \colon H^1(K_m, \mathrm{Ta}_\ell \mathbb{G}_m) \longrightarrow \mathrm{Lie}(\mathbb{G}_m)(K_m) = K_m.$$

Combining all those remarks, we get a new version of the map

$$\lambda_m \colon \varprojlim_n K_n^\times \xrightarrow{(\cdot,\cdot)} H^1(K_m, \mathrm{Ta}_\ell \mathbb{G}_m) \xrightarrow{\exp^*} K_m.$$

We notice three things that frequently appear together in “modern” number theory:

(1) A tower of things, where we take some sort of compatible system of points (in this case, fields K_n and norm-compatible set of roots $(\zeta_{\ell^n})_n$);

(2) A representation (in this case, $\mathrm{Ta}_\ell \mathbb{G}_m$).

(3) A cohomology with the things from the tower in the “base slot” and the representation in the “coefficients slot” (in this case, Galois).

In a context where there is some version of these three concepts above, “modern” explicit reciprocity laws will follow a similar approach to explicitly compute the correspondent “symbols” that appear in the context.

5.3 An example of modern explicit reciprocity law

We now sketch an example of a modern explicit reciprocity law from [CH18] in a setting that generalizes the three items identified above:

(1) A tower of *modular curves* and a system of points corresponding to the moduli of *elliptic curves* compatible by cyclic isogenies;

(2) A representation, coming from a *modular form*;

(3) A cohomology, in this case, étale (a nicely behaving cohomology for ℓ-adic fields, if $\ell \nmid N$, the tame level of the modular curves).

Let us briefly describe each item. In what follows, K is an imaginary quadratic field (that is, $\mathbb{Q}(\sqrt{-D})$ for some $D > 0$), ℓ is a fixed prime and $N \geq 4$ is an integer prime to ℓ and such that any $p \mid N$ splits on K (the *Heegner hypothesis* on K).

5.3.1 Towers of modular curves

An *elliptic curve* over a field K is an irreducible genus 1 projective curve with a K-rational point. Besides the algebraic variety structure, they also have a abelian group structure, making them abelian varieties. A map $\varphi\colon E_1 \to E_2$ between two elliptic curves over K that respects both the algebraic variety over K and group structures is an *isogeny* over K. An isogeny invertible by another isogeny (defined over the same field) is an *isomorphism* over said field. Over $\mathbb{C}$, elliptic curves look like toruses of the form $\mathbb{C}/(\mathbb{Z}\omega_1 \oplus \mathbb{Z}\omega_2)$, with $\omega_1, \omega_2 \in \mathbb{C}$ linearly independent, and isogenies can be thought of in the level of the lattices defining each elliptic curve:

$$\exists \varphi\colon \mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z}) \to \mathbb{C}/(\mathbb{Z} + \tau'\mathbb{Z}) \iff \exists m(\mathbb{Z} + \tau\mathbb{Z}) \subseteq (\mathbb{Z} + \tau'\mathbb{Z}).$$

Refer to [Sil09] for general information on elliptic curves.

Modular curves are quotients of the upper halfplane $\mathcal{H} \doteq \{z \in \mathbb{C};\ \operatorname{Im}(z) > 0\}$ by the action of arithmetic subrgoups Γ of $\mathrm{SL}_2(\mathbb{Z})$, called *congruence subgroups*, on $\mathcal{H}$ via Möbius transformations:

$$\Gamma \times \mathcal{H} \to \mathcal{H}\colon \left(\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right), \tau\right) \mapsto \tfrac{a\tau + b}{c\tau + d}.$$

Depending on the group Γ, the correspondent modular curve $Y(\Gamma) = \Gamma\backslash\mathcal{H}$ classifies elliptic curves endowed with certain torsion structure modulo isomorphisms preserving said torsion structures, that is, they are *moduli spaces* for the moduli problem we just described. For example, the modular curve $Y_0(N\ell^n)$ associated to the group

$$\Gamma_0(N\ell^n) = \left\{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z});\ c \equiv 0 \mod N\ell^n\right\},$$

for any $n \in \mathbb{Z}_{\geq 0}$, classifies triples (E, C_N, C_{ℓ^n}) consisting of elliptic curve E and cyclic subrgoups C_N and C_{ℓ^n} of orders N and ℓ^n, respectively, modulo isomorphisms $E \xrightarrow{\sim} E'$ sending C_N to C'_N and C_{ℓ^n} to C'_{ℓ^n}. There are natural projections $Y_0(N\ell^n) \twoheadrightarrow Y_0(N\ell^m)$ for $n \geq m$, giving the tower of modular curves we wanted. See [DS05, §1.5] for further information on modular curves as moduli spaces of elliptic curves.

The hypothesis $N \geq 4$ grants the representability of this moduli problem, that is, the existence of a *universal elliptic curve* $\mathcal{E} \twoheadrightarrow Y_0(N\ell^n)$, the family whose sections are all moduli classes.

Consider the elliptic curves $E_n = \mathbb{C}/(\mathbb{Z}\ell^n\tau \oplus \mathbb{Z})$ for some *CM point* $\tau \in \mathcal{H} \cap K$ and the natural isogenies $\varphi_n\colon E \to E_n$, where $E \doteq E_0$. Then

$$\operatorname{graph}(\varphi_m) \subseteq E \times E_m \hookrightarrow E \times \mathcal{E} \implies \operatorname{graph}(\varphi_m)^r \hookrightarrow E^r \times \mathcal{E}^r \doteq W_r,$$

where W_r is a *Kuga–Sato variety* and $r \in \mathbb{Z}_{\geq 2}$. We then define a *generalized Heegner cycle* $\Delta \doteq \epsilon \operatorname{graph}(\varphi_m) \in \mathrm{CH}^{k-1}(W_k)$, living in the group of codimension $k-1$ cycles (formal sums of subvarieties), *cf.* [CH18, §4]. The numbers N and k may seem arbitrary here, but they actually relate to next ingredient: modular forms.

5.3.2 Representations coming from modular forms

A *modular (resp. cuspidal) form* of weight k in a modular curve $Y = \Gamma\backslash\mathcal{H}$ is an analytic function $f\colon \mathcal{H} \to \mathbb{C}$ that is well-behaved (resp. vanishing) at the cusps of Y satisfying

$$f\left(\tfrac{az+b}{cz+d}\right) = (cz+d)^k f(z) \text{ for all } \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma.$$

To a modular form f as above, one can associate an *ℓ-adic representation* V_f, a $\mathbb{Q}_\ell$-vector space with an action of $\mathrm{Gal}(\overline{\mathbb{Q}}_\ell/\mathbb{Q})$, defined as a quotient of the first étale cohomology group with base in Y.

The close relation between modular forms and elliptic curves is one of the highlights of 20th century number theory: the *modularity theorem* states that the representation coming from an elliptic curve is isomorphic to one coming from a weight 2 cuspidal modular form over $Y_0(N)$ for some N (Wiles (1995)[xvii], Taylor–Wiles (1995)[xviii], Breuil–Conrad–Diamond–Taylor (2001)[xix]), and its most celebrated consequence is Fermat's Last Theorem, already mentioned earlier. See [DS05] for more on modular forms.

5.3.3 Cohomology

Let f be a cusp form for $Y_0(N)$ of weight k. Recall the generalized Heegner cycle $\Delta \in \mathrm{CH}^{k-1}(W_r)$ from before. The *Abel–Jacobi map*

$$\mathrm{AJ}\colon \mathrm{CH}^{k-1}(W_r) \longrightarrow H^1_{\text{ét}}(Y_0(N\ell^n), V_f)$$

maps Δ into a *generalized Heegner class* $z_{f,n}$ ([CH18, §4.2]). This gives a corestriction-compatible system $(z_{f,n})_n$ in the tower of $\mathbb{Q}_\ell$-vector spaces $(H^1_{\text{ét}}(Y_0(N\ell^n), V_f))_n$, called an *Euler system* ([CH18, §4.3]). Taking projective limits, we define a *big Heegner class*

$$z_f \doteq \varprojlim_n z_{f,n} \in \varprojlim_n H^1_{\text{ét}}(Y_0(N\ell^n), V_f)$$

([CH18, §5.2]).

5.3.4 Piecing it all together

Recall the map λ_m from §5.2. In this context,

- The domain is $\varprojlim_n H^1_{\text{ét}}(Y_0(N\ell^n), V_f)$, where we have the big Heegner class z_f;
- The "symbol" is the *de Rham pairing* (after seeing the classes as differentials over a rigid analytic variety via the Eichler–Shimura isomorphism);
- The map $\exp^*$ is the *Bloch–Kato dual exponential map* (see [CH18, §4.5]);

[xvii] A. Wiles, *Modular elliptic curves and Fermat's Last Theorem.* Ann. of Math., v. 141, n. 3, p.443–551, 1995.

[xviii] R. Taylor, A. Wiles, *A Ring-theoretic properties of certain Hecke algebras*, Ann. of Math., v. 141, n. 3, p. 553–572, 1995.

[xix] C. Breuil, B. Conrad, F. Diamond, R. Taylor, *On the modularity of elliptic curves over $\mathbb{Q}$: wild 3-adic exercises*, J. Amer. Math. Soc., v. 14, n. 4, p. 843–939, 2001.

The map λ_m becomes a *Perrin-Riou map*

$$\mathcal{L}\colon \varprojlim_n H^1_{\text{ét}}(Y_0(N\ell^n), V_f) \longrightarrow \text{some Iwasawa algebra},$$

cf. [CH18, §5.1]. But the focus of modern explicit reciprocity laws is not so much in computing the symbol itself, but in its relation to L-functions attached to modular forms, where it ties to important open problems, such as the Bloch–Kato conjecture.

Let f be a cusp form over $Y_0(N)$ of weight k and χ a character over K (a 1-dimensional representation of $\mathrm{Gal}(\overline{K}/K)$). One can define a *Rankin L-function* $L(f, \chi, s)$, which satisfies a functional equation centered at the *special point* $k/2$, where it attains a *special value* (see [CH18, §5.3]).

Conjecture 37 (Bloch–Kato) *One has* $\mathrm{ord}_{s=k/2} L(f, \chi, s) = \dim \mathrm{Sel}_{K_f}(K, V_f)$, *where* $\mathrm{Sel}(K, V_f)$ *is a subspace of* $H^1(K, V_f)$ *called the Bloch–Kato Selmer group associated to* f *(see [CH18, §1]).*

In our ℓ-adic context, one instead interpolates the special values of the (complex) L-function associated to f obtaining a ℓ-adic L-function $L_\ell(f, \chi)$. We now vaguely state an explicit reciprocity law arising in this context (which we ironically won't be able to state explicitly, neither plan to do so):

Theorem 38 ([CH18, Theorem 5.7], vaguely stated) *Under some conditions,* z_f *generates* $\mathrm{Sel}(K, V_f)$ *and* $\mathcal{L}(z_f)$ *coincides with* $L_\ell(f, \chi)$ *up to a explicit scalar multiple.*

5.3.5 Summary

- Reciprocity started as a way to compute symbols detecting existence of nth roots in modular arithmetic;
- Focus shifted towards the study of (abelian) field extensions, culminating in Artin's description of the Galois group of the maximal abelian extension of a local field;
- In parallel, explicit calculations of the symbols were happening in towers of fields mapping into cohomology over representations;
- In modern number-theoretic contexts associated to elliptic curves/modular forms, explicit reciprocity laws relate the étale cohomology of Galois representation associated to modular forms over towers of modular curves to L-functions associated to said modular forms, à la Bloch–Kato.

5.3.6 Where to go from here?

In the setting we described above there is much room for generalizations, leading to many currently active research lines:

- Other fields: instead of a quadratic imaginary field, one can work over totally real fields (that is, finite extensions of $\mathbb{Q}$ that can be embedded in $\mathbb{R}$);

- Other towers of things: instead of modular curves classifying elliptic curves (dimension 1 abelian varieties), one might consider, for example, Shimura varieties classifying abelian surfaces (the 2-dimensional version of the problem);
- Do everything in families: ℓ-adically interpolate the representations to get big Heegner classes associated to *families* of modular forms, either or Hida or of Coleman type;
- Other approaches to the "tower of things": instead of taking the inverse limit over the *base* side of the cohomology, one can take the inverse limit over the *coefficient* side: coefficients will get more complicate, but the base becomes much simpler.

References

[AM69] M.F. Atiyah, I.G. Macdonald, "Introduction to Commutative Algebra". Addison-Wesley Pub. Co., 1969.

[CH18] F. Castella, M.-L. Hsieh, *Heegner cycles and p-adic L-functions.* Math. Ann. 370 (2018), 567–628.

[Con] K. Conrad, "Ostrowski for number fields". Notes available on `https://kconrad.math.uconn.edu/blurbs/gradnumthy/ostrowskinumbfield.pdf`.

[Cox13] D.A. Cox, "Primes of the form $x^2 + ny^2$: Fermat, class field theory and complex multiplication". 2nd ed., 2013.

[DS05] F. Diamond, J. Shurman, "A First Course in Modular Forms". Springer GTM 228, 2005.

[EP05] A.J. Engler, A. Prestel, "Valued fields". Springer, 2005.

[Gou97] Q. Gouvêa, "p-adic Numbers, An Introduction". 2nd ed., Springer, 1997.

[Gra06] D. Grant, *Geometric proofs of reciprocity laws.* J. reine angew. Math. 586 (2006), 91–124.

[IR90] K. Ireland, M. Rosen, "A Classical Approach to Modern Number Theory". 2nd ed., GTM 84, Springer, 1990.

[Lem09] F. Lemmermeyer, *Jacobi and Kummer's Ideal Numbers.* Abh. Math. Semin. Univ. Hamb. 79, No. 2, 165–187, 2009.

[Li] C. Li, "Kato's explicit reciprocity laws". notes available on `https://www.math.columbia.edu/~chaoli/doc/ExplicitReciprocity.html`.

[MilCFT] J. Milne, "Class Field Theory". Notes available on `https://www.jmilne.org/math/CourseNotes/cft.html`.

[MilFT] J. Milne, "Field and Galois Theory". Notes available on `https://www.jmilne.org/math/CourseNotes/ft.html`.

[Neu99] J. Neukirch, "Algebraic Number Theory". Springer, 1999.

[Rib79] P. Ribenboim, "13 Lectures on Fermat's Last Theorem". Springer, 1979.

[Sil09] J. Silverman, "Arithmetic of Elliptic Curves". 2nd ed., GTM 106, Springer, 2009.

[Vos00] S. Vostokov, *8. Explicit formulas for the Hilbert symbol.* Geometry & Topology Monographs vol. 3, part I, sec. 8 (2000), 81–89.

[Wym72] B.F. Wyman, *What is a Reciprocity Law?.* Amer. Math. Monthly 79:56 (1972), 571–586.

Optimal control with a stochastic switching time: introduction and solution approaches

Maddalena Muttoni [(*)]

Abstract. When planning an optimal policy, a farsighted decision-maker should account for the possibile occurrence of disruptive events over the course of the time horizon. For example, when planning the optimal emission abatement policy, account for a possible climate catastrophe; when planning industrial production, account for an unpredictable disruption that may affect the producer's profit.

In the optimal control framework, a stochastic switching time is a random instant, modeled as a positive random variable, which marks a regime shift – i.e., an abrupt and irreversible change in the system – which splits the planning horizon into two stages. The shift may affect the payoff and/or the state trajectory in several ways, all of which are included in the analysis of the most general scenario. In search for the optimal policy under this kind of uncertainty, two methods are featured in the literature: the "backward" approach and the "heterogeneous" one. The two approaches will be described, compared, and then applied to a marketing toy model.

1 Introduction

In the context of dynamical systems, we call *stochastic switching time* an event which

- occurs at a random time τ;
- changes abruptly the nature of the system;
- splits the time horizon into two stages: a Stage 1 before the occurrence of τ, and a Stage 2 afterwards.

In this work we consider a single switching time only.

This framework finds many applications in various areas, such as epidemiology, climate change, production, and marketing, to name a few. For example, when planning the optimal emission abatement policy, account for a possible climate catastrophe; when planning industrial production, account for an unpredictable disruption that may affect the producer's profit.

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: mmuttoni@math.unipd.it. Seminar held on 3 May 2023.

In the context of optimal control, we are interested in seeing how the optimal strategy adjusts to the regime shift, i.e., how it changes going from Stage 1 to Stage 2, upon the occurrence of τ.

We are also interested in comparing the results of planners with different degrees of information about the switching time. Of course, more information will lead to a higher expected payoff. By computing the difference in the planners' expected payoffs, we can evaluate the cost of information about the switching time.

In this regard, we can distinguish between a *myopic* and a *farsighted* planner:

Definition 1 A *myopic* planner does not take into account the possibility of a switching time: they solve a single-stage optimal control problem.

Definition 2 A *farsighted* planner, on the contrary, anticipates the occurrence of a switching time and has some information about the probability distribution of τ, as will be better specified later on.

Simple optimal control problem

An optimal control problem is a dynamic optimization problem where, at every time in a given programming interval $[0, T]$, the agent sets the value of the control variable u from a given control set U (a topological space in general), i.e., they choose their control function, or *strategy*,

$$u : [0, T] \to U.$$

The strategy enters the state dynamics, influencing the evolution of the state variable $x \in \mathbb{R}^n$, whose initial value is given. Assuming the existence and uniqueness of the solution of such dynamics, the resulting *trajectory* is the absolutely continuous function

$$x : [0, T] \to \mathbb{R}^n.$$

The pair (u, x) of control and state trajectory is called a process (see [5]).

The planner's objective is to maximize a certain payoff, which is the sum of an intertemporal term and a salvage value. The first is the integral of a profit flow over time, which depends on the strategy and the corresponding state trajectory; the latter is a lump sum which depends on the final state $x(T)$.

In absence of constraints on the final state $x(T)$, the planner solves the following problem.

$$\underset{u(t)\in U}{\text{maximize}} \left[\int_0^T g\big(t, x(t), u(t)\big)\, dt + S\big(x(T)\big) \right]$$

subject to:

$$\begin{cases} \dot{x}(t) = f\big(t, x(t), u(t)\big) & \text{for } t \in [0, T] \\ x(0) = x_0 \end{cases}$$

where:

- $g(t, x, u)$ is the running payoff;

- $S(x)$ is the salvage value function;
- $f(t,x,u)$ is the state dynamics.

There are two solution approaches to this problem: *Dynamic programming* and Pontryagin's *maximum principle.*

Dynamic programming relies on the general mathematical principle of embedding the original problem in a large class of problems, each starting at $t \in [0,T]$ from the initial state $\mathtt{x}$. The corresponding "value function" can be defined as follows:

$$V(t,\mathtt{x}) := \sup_{u(\theta)\in U} \left[\int_t^T g\big(\theta, x(\theta), u(\theta)\big)\, d\theta + S\big(x(T)\big) \right]$$

subject to:

$$\begin{cases} \dot{x}(\theta) = f\big(\theta, x(\theta), u(\theta)\big) & \text{for } \theta \in [t,T] \\ x(t) = \mathtt{x} \end{cases}$$

Of course, if an optimal control exists, that sup is actually a max, as it is attained by the optimal control.

If the value function is differentiable, it is the classical solution of the Hamilton-Jacobi-Bellman (HJB) PDE and terminal condition:

$$\begin{cases} -\partial_t V(t,x) = \max_{\mathtt{u}\in U} \big\{ g(t,x,\mathtt{u}) + \partial_x V(t,x) \cdot f(t,x,\mathtt{u}) \big\} \\ V(T,x) = S(x) \end{cases} \tag{1}$$

and the $\Phi(t,x)$ which maximizes the RHS is the optimal feedback (i.e., it depends on the current state of the system) strategy, as long as there exists a unique solution to the resulting state dynamics. More precisely,

Theorem 1 *Let W be a continuously differentiable solution of system (1) of HJB equation and terminal condition on $[0,T] \times \mathbb{R}^n$. Let $\Phi(t,x)$ maximize the RHS of the HJB equation. If there exists a unique state trajectory $x(t)$ resulting from the feedback strategy $\Phi(t,x)$, then $u(t) = \Phi\big(t, x(t)\big)$ is the optimal control and W is the value function.*

Pontryagin's Maximum Principle, instead, provides necessary conditions for the maxima of the problem. We call the variable $\lambda \in \mathbb{R}^n$ the adjoint, or co-state, variable. If (u,x) is an optimal process, then there exists an absolutely continuous co-state trajectory

$$\lambda : [0,T] \to \mathbb{R}^n$$

such that:

$$\begin{cases} -\dot{\lambda}(t) = \partial_x g\big(t, x(t), u(t)\big) + \lambda(t)\partial_x f\big(t, x(t), u(t)\big)^{(\mathbf{2})} \\ \lambda(T) = \partial_x S\big(x(T)\big) \end{cases}$$

and, for almost every $t \in [0,T]$,

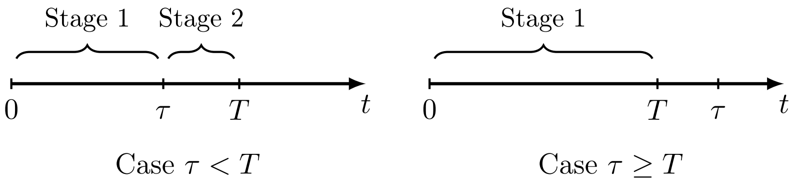
$$u(t) \in \arg\max_{\mathtt{u}\in U} H\big(t, x(t), \mathtt{u}, \lambda(t)\big).$$

[2] $[\lambda \partial_x f]_i = \lambda_1 \partial_{x_i} f_1 + \cdots + \lambda_n \partial_{x_i} f_n$

Two-stage optimal control with stochastic switching time

We are interested in integrating a stochastic switching time into the optimal control framework described in the previous paragraph.

We model the stochastic switching time τ as an absolutely continuous random variable taking values in $[0, +\infty)$. Due to the finiteness of the time horizon, τ could occur either during the programming interval $[0, T]$, splitting it into a Stage 1 and a Stage 2, or after T, leaving the whole interval in Stage 1.



We describe the probability distribution of τ through a quantity called the *hazard rate* of τ at time t:

$$\lim_{dt \to 0^+} \frac{\mathbb{P}\big(\tau \leq t + dt \,\big|\, \tau > t\big)}{dt} = \eta\big(t, x(t), u(t)\big) \tag{2}$$

where $\eta(t, x, u)$ is the hazard rate function: the hazard rate may be endogenous, i.e., dependent on the state of the system and possibly on the control variable, as well.

Remark 1 Since the hazard rate is defined only on the programming interval $[0, T]$, we can derive the distribution of τ only in $[0, T]$. However the distribution in the remaining interval $(T, +\infty)$ is not our concern, and suffice it to say that it can be arbitrarily prolonged.

The switch may have one or more simultaneous effects on the system, such as a change the state dynamics, in the running payoff, in the salvage value function, and in the control set; it may entail an endogenous lump sum $K\big(\tau, x(\tau), u(\tau)\big)$ at the switching time; it may induce a jump discontinuity in the state trajectory such that

$$x(\tau^+) = \varphi\big(\tau, x(\tau), u(\tau)\big).$$

See Fig. 1 for a schematic representation (and the respective notation) of the effects of the switch.

Observe that, if the state is assumed to be continuous in τ, then $\varphi(\tau, x, u) = x$.

	Stage 1	**Switch**	**Stage 2**
State dynamics	$f_1(t, x, u)$		$f_2(\tau, t, x, u)$
Running payoff	$g_1(t, x, u)$		$g_2(\tau, t, x, u)$
Salvage value	$S_1(x)$		$S_2(\tau, x)$
Control set	U_1		U_2
State jump		$\varphi(\tau, x, u)$	
Lump sum		$K(\tau, x, u)$	

Figure 1: Possible effects of τ.

Information structure

In this context, a *farsighted* planner is a planner who, in Stage 1, knows that τ could occur at any time, knows the hazard rate function $\eta(t, x, u)$ defined in (2), and the effects that the switch will have on the system.

Because the planner does not know when τ will occur, they need to plan a Stage 1 strategy for the whole time horizon.

$$u_1(t), \quad x_1(t) \quad \text{for } t \in [0, T]$$

If the switching time occurs during the programming interval, the planner realizes that τ has occurred and when. Therefore, in the Stage 2 interval $[\tau, T]$ they will be able to implement the optimal strategy for the specific occurrence of τ. The Stage 2 process turns out to be a family of processes, parameterized by the realization $s \in [0, T]$ of the switching time τ, where each of them is defined on the Stage 2 interval $[s, T]$:

$$u_2(s, t), \quad x_2(s, t) \quad \text{for } s \in [0, T], \; t \in [s, T]$$

Remark 2 In general, there could be different control and state variables in Stage 2 compared to Stage 1.

Due to the stochasticity of τ, the planner can only aim at maximizing the expectation of the total payoff, which takes different forms depending on the fact that τ occurs during the programming interval or afterwards. The most general formulation of the switching time optimal control problem is the following:

$$\begin{aligned} \underset{\substack{u_1(t) \in U_1 \\ u_2(s,t) \in U_2}}{\text{maximize}} \; \mathbb{E}\Bigg[\chi_{\tau < T} \Big\{ \int_0^\tau g_1\big(t, x_1(t), u_1(t)\big)\, dt + K\big(\tau, x_1(\tau), u_1(\tau)\big) \\ + \int_\tau^T g_2\big(\tau, t, x_2(\tau, t), u_2(\tau, t)\big)\, dt + S_2\big(\tau, x_2(\tau, T)\big) \Big\} \\ + \chi_{\tau \geq T} \Big\{ \int_0^T g_1\big(t, x_1(t), u_1(t)\big)\, dt + S_1\big(x_1(T)\big) \Big\} \Bigg] \end{aligned}$$

subject to:

$$\begin{cases} \dot{x}_1(t) = f_1\big(t, x_1(t), u_1(t)\big) & \text{for } t \in [0, T] \\ \qquad x_1(0) = x_0 & \\ \dot{x}_2(s, t) = f_2\big(s, t, x_2(s, t), u_2(s, t)\big) & \text{for } t \in [s, T] \\ \qquad x_2(s, s) = \varphi\big(s, x_1(s), u_1(s)\big) & \\ \text{Hazard rate of } \tau \text{ at time } t\text{:} \;\; \eta\big(t, x_1(t), u_1(t)\big) & \end{cases}$$

Please observe the abuse of notation in the formulas above and from now on: $\dot{x}_2(s, t) = \partial_t x_2(s, t)$.

It is worth noting that the initial condition of the Stage 2 problem is a function of the Stage 1 variables at the switch: $x_2(s, s) = \varphi\big(s, x_1(s), u_1(s)\big)$. This implies that the Stage 2 problem cannot be solved independently, *unless* one applies dynamic programming, where an optimal control problem is solved for every possible initial value.

Remark 3 In general, the Stage 2 data f_2, g_2, S_2 may also depend on the Stage 1 state variable at the switch, $x_1(s)$. With this purpose, let us introduce the auxiliary Stage 2 state variable $\tilde{x}_2$ such that $\tilde{x}_2(s,t) = x_1(s)$, i.e.,

$$\begin{cases} \dot{\tilde{x}}_2(s,t) = 0 & \text{for } t \in [s,T] \\ \tilde{x}_2(s,s) = x_1(s). \end{cases}$$

Adding $\tilde{x}_2$ to the Stage 2 state variables, and updating the Stage 2 dynamics and initial value with those of $\tilde{x}_2$, we can omit the dependence of f_2, g_2, S_2 on $x_1(s)$ without loss of generality.

Reformulation

We compute the expectation with the aid of the auxiliary Stage 1 state variable $z_1(t) := \mathbb{P}(\tau > t)$, which is the probability of still being in Stage 1 at time t. To view it as a state variable, we write its dynamics and initial value:

$$\begin{cases} \dot{z}_1(t) = -\eta\big(t, x(t), u(t)\big) z_1(t) \\ z_1(0) = 1 \end{cases}$$

where the dynamics is derived from the definition of hazard rate (2). Then, the probability density of τ at time t is:

$$f_\tau(t) = -\dot{z}_1(t) = \eta\big(t, x(t), u(t)\big) z_1(t).$$

Both the formulations above are used to compute the expectation. After basic integral manipulations, the resulting objective is:

$$\begin{aligned} \underset{\substack{u_1(t)\in U_1 \\ u_2(s,t)\in U_2}}{\text{maximize}} \Bigg[\int_0^T z_1(t) \Big\{ & g_1\big(t, x_1(t), u_1(t)\big) \\ & + \eta\big(t, x_1(t), u_1(t)\big) \Big[K\big(t, x_1(t), u_1(t)\big) \\ & \qquad + \int_t^T g_2\big(t, \theta, x_2(t,\theta), u_2(t,\theta)\big)\, d\theta + S_2\big(t, x_2(t,T)\big) \Big] \Big\}\, dt \\ & + z_1(T) S_1\big(x_1(T)\big) \Bigg] \end{aligned} \tag{3}$$

subject to:

$$\begin{cases} \dot{x}_1(t) = f_1\big(t, x_1(t), u_1(t)\big) & x_1(0) = x_0 \\ \dot{z}_1(t) = -\eta\big(t, x_1(t), u_1(t)\big) z_1(t) & z_1(0) = 1 \\ \dot{x}_2(s,t) = f_2\big(s, t, x_2(s,t), u_2(s,t)\big) & x_2(s,s) = \varphi\big(s, x_1(s), u_1(s)\big) \end{cases}$$

where we updated the Stage 1 dynamics with the new variable z_1.

2 Solution approaches

There are two possible ways of solving the two-stage optimal control problem defined above: the *backward* approach and the *heterogeneous* one.

The backward approach is based on dynamic programming: it involves solving the Stage 2 problem for every possible occurrence s of the switch and for every possible initial state, and then plugging the Stage 2 value function into the Stage 1 problem, which is solved as a simple optimal control problem, assuming optimal behavior in Stage 2. The two stages are solved separately (in reverse order) at the cost of computing the Stage 2 value function for every possible value of x_2 at the switch, instead of just the value that it will take from the condition $x_2(s,s) = \varphi\big(s, x_1(s), u_1(s)\big)$. For more details on this approach, see e.g. [3].

The heterogeneous approach is based on Pontryagin's Maximum Principle (see [3]): one derives necessary conditions for the solution by calculating the co-state functions for both stages, as solutions of the corresponding adjoint system, and letting the optimal strategies satisfy the resulting maximality conditions. The two stages are necessarily solved together, because x_1 and u_1 enter the initial conditions of Stage 2, and (as we will see) the Stage 2 co-states enter the Stage 1 adjoint equations.

In what follows, we will determine the optimal control of the switching time problem, applying the two approaches and comparing their results.

Backward approach

First, let us solve the Stage 2 problem with dynamic programming. Since, in general, the Stage 2 data may depend on the realization s of the switching time τ, the Stage 2 value function V_2 will depend on s as well:

$$V_2(s,t,\mathtt{x}) := \sup_{u(\theta)\in U_2} \left[\int_t^T g_2\big(s,\theta,x(\theta),u(\theta)\big)\, d\theta + S_2\big(s,x(T)\big) \right]$$

subject to:

$$\begin{cases} \dot{x}(\theta) = f_2\big(s,\theta,x(\theta),u(\theta)\big) & \text{for } \theta \in [t,T] \\ x(t) = \mathtt{x} \end{cases}$$

If $V_2(s,\cdot,\cdot)$ is differentiable, then it is the solution of the corresponding system of HJB equation and terminal condition (parametrized by s):

$$\text{(4)} \qquad \begin{cases} -\partial_t V_2(s,t,x) = \max\limits_{\mathtt{u}\in U_2} \big\{ g_2(s,t,x,\mathtt{u}) + \partial_x V_2(s,t,x)\cdot f_2(s,t,x,\mathtt{u}) \big\} \\ V_2(s,T,x) = S_2(s,x) \end{cases}$$

The optimal feedback strategy $\Phi_2(s,t,x)$ maximizes the RHS of the HJB equation in (4):

$$\Phi_2(s,t,x) \in \operatorname*{arg\,max}_{\mathtt{u}\in U_2} \big\{ g_2(s,t,x,\mathtt{u}) + \partial_x V_2(s,t,x)\cdot f_2(s,t,x,\mathtt{u}) \big\}.$$

Let (u_1, x_1) be a feasible Stage 1 process. Let the trajectory $t \mapsto x_2(s,t)$ satisfy the Cauchy problem:

$$\begin{cases} \dot{x}_2(s,t) = f_2\big(s,t,x_2(s,t),\Phi_2(s,t,x_2(s,t))\big) & \text{for } 0 \le s \le t \le T \\ x_2(s,s) = \varphi\big(s,x_1(s),u_1(s)\big), \end{cases}$$

then, the optimal control for Stage 2, given (u_1, x_1), is

$$u_2(s,t) = \Phi_2\big(s,t,x_2(s,t)\big).$$

By Bellman's Priciple of Optimality, given (u_1, x_1) and assuming (u_2, x_2) as above, we can write:

$$\int_t^T g_2\big(s,\theta,x_2(s,\theta),u_2(s,\theta)\big)\,d\theta + S_2\big(s,x_2(s,T)\big) = V_2\big(s,t,x_2(s,t)\big)$$

In particular, for $s = t$, we can substitute $x_2(t,t) = \varphi\big(t,x_1(t),u_1(t)\big)$, yielding:

$$V_2\big(t,t,x_2(t,t)\big) = V_2\big(t,t,\varphi(t,x_1(t),u_1(t))\big). \tag{5}$$

Assuming optimal behavior in Stage 2, in (3) we can substitute the Stage 2 payoff with (5), obtaining the following objective for Stage 1:

$$\begin{aligned} \underset{\substack{u_1(t)\in U_1\\ u_2(s,t)\in U_2}}{\text{maximize}} \Bigg[\int_0^T z_1(t) \Big\{ & g_1\big(t,x_1(t),u_1(t)\big) \\ & + \eta\big(t,x_1(t),u_1(t)\big)\Big[K\big(t,x_1(t),u_1(t)\big) + V_2\big(t,t,\varphi(t,x_1(t),u_1(t))\big)\Big]\Big\}\,dt \\ & + z_1(T)S_1\big(x_1(T)\big)\Bigg] \end{aligned}$$

subject to:

$$\begin{cases} \dot{x}_1(t) = f_1\big(t,x_1(t),u_1(t)\big) \\ \quad x_1(0) = x_0 \\ \dot{z}_1(t) = -\eta\big(t,x_1(t),u_1(t)\big)z_1(t) \\ \quad z_1(0) = 1 \end{cases}$$

This is a simple optimal control problem, which can be solved again with dynamic programming.

The value function V associated to such a problem is a function of time and the state variables x and z:

$$\begin{aligned} V(t,\mathtt{x},\mathtt{z}) := \sup_{u(\theta)\in U_1} \Bigg[\int_t^T z(\theta) \Big\{ & g_1\big(\theta,x(\theta),u(\theta)\big) \\ & + \eta\big(\theta,x(\theta),u(\theta)\big)\Big[K\big(\theta,x(\theta),u(\theta)\big) + V_2\big(\theta,\theta,\varphi(\theta,x(\theta),u(\theta))\big)\Big]\Big\}\,d\theta \\ & + z(T)S_1\big(x(T)\big)\Bigg] \end{aligned}$$

subject to:

$$\begin{cases} \dot{x}(\theta) = f_1\big(\theta, x(\theta), u(\theta)\big) & \text{for } \theta \in [0,T] \\ \quad x(t) = \mathtt{x} \\ \dot{z}(\theta) = -\eta\big(\theta, x(\theta), u(\theta)\big) z(\theta) \\ \quad z(t) = \mathtt{z} \end{cases}$$

Due to the particular structure of the problem, the value function V can be decomposed as:

$$V(t,x,z) = zV^c(t,x)$$

We call $V^c(t,x)$ the "current" value function. It satisfies the system (6) of HJB equation and terminal condition:

$$(6) \qquad \begin{cases} -\partial_t V^c(t,x) = \max\limits_{\mathtt{u}\in U_1} \Big\{ g_1(t,x,\mathtt{u}) + \partial_x V^c(t,x) \cdot f_1(t,x,\mathtt{u}) \\ \qquad\qquad +\eta(t,x,\mathtt{u}) \Big[K(t,x,\mathtt{u}) + V_2\big(t,t,\varphi(t,x,\mathtt{u})\big) - V^c(t,x) \Big] \Big\} \\ V^c(T,x) = S_1(x) \end{cases}$$

The optimal feedback strategy $\Phi_1(t,x)$ maximizes the RHS of the HJB equation:

$$\begin{aligned} \Phi_1(t,x) \in \operatorname*{arg\,max}_{\mathtt{u}\in U_1} \Big\{ & g_1(t,x,\mathtt{u}) + \partial_x V^c(t,x) \cdot f_1(t,x,\mathtt{u}) \\ & + \eta(t,x,\mathtt{u}) \Big[K(t,x,\mathtt{u}) + V_2\big(t,t,\varphi(t,x,\mathtt{u})\big) - V^c(t,x) \Big] \Big\} \end{aligned}$$

Let the trajectory $x_1(t)$ satisfy the Cauchy problem:

$$\begin{cases} \dot{x}_1(t) = f_1\big(t, x(t), \Phi_1(t, x(t))\big) & \text{for } t \in [0,T] \\ x_1(0) = x_0 \end{cases}$$

Then, the optimal control for Stage 1 is

$$u_1(t) = \Phi_1\big(t, x_1(t)\big).$$

Heterogeneous approach

By [13], if the suitable regularity assumptions on the data hold, and if $(u_1, x_1, z_1, u_2, x_2, z_2)$ constitute an optimal 2-stage process, then there exist co-state functions $\lambda_x(t)$, $\lambda_z(t)$, and $\xi_x(s,t)$, $\xi_z(s,t)$ such that the following conditions hold.

1a. The Stage 1 strategy satisfies the following local maximality condition for almost every $t \in [0,T]$:

$$\begin{aligned} \Big\{ & z_1(t)\partial_u \big[g_1\big(t, x_1(t), u_1(t)\big) + \eta\big(t, x_1(t), u_1(t)\big) K\big(t, x_1(t), u_1(t)\big) \big] \\ & + \lambda_x(t)\partial_u f_1\big(t, x_1(t), u_1(t)\big) + \xi_x(t,t)\partial_u \varphi\big(t, x_1(t), u_1(t)\big) \\ & + \big[\xi_z(t,t) - \lambda_z(t)\big] \partial_u \eta\big(t, x_1(t), u_1(t)\big) z_1(t) \Big\} \cdot \big(\mathtt{u} - u_1(t)\big) \le 0 \quad \text{for all } \mathtt{u} \in U_1 \end{aligned}$$

It is convenient to define the “current” co-state functions

$$\lambda^c_x(t) := \lambda_x(t)/z_1(t), \qquad \xi^c_x(s,t) := \xi_x(s,t)/z_2(s,t)$$

1b. If φ and η do not depend on u, condition 1a. is substituted by the following maximality condition:

$$\begin{aligned} u_1(t) \in \operatorname*{arg\,max}_{\mathtt{u}\in U_1} \Big\{ & g_1\big(t, x_1(t), \mathtt{u}\big) + \eta\big(t, x_1(t)\big) K\big(t, x_1(t), \mathtt{u}\big) \\ & + \lambda^c_x(t) \cdot f_1\big(t, x_1(t), \mathtt{u}\big) \Big\} \end{aligned}$$

2. The Stage 2 strategy satisfies the usual maximality condition:

$$u_2(s,t) \in \operatorname*{arg\,max}_{\mathtt{u}\in U_2} \Big\{ g_2\big(s, t, x_2(s,t), \mathtt{u}\big) + \xi^c_x(s,t) \cdot f_2\big(s, t, x_2(s,t), \mathtt{u}\big) \Big\}$$

3. The co-state functions are solutions of the following adjoint system:

$$\begin{cases} -\dot\lambda^c_x(t) = \partial_x\big[g_1\big(t, x_1(t), u_1(t)\big) + \eta\big(t, x_1(t), u_1(t)\big) K\big(t, x_1(t), u_1(t)\big)\big] + \lambda^c_x(t)\partial_x f_1\big(t, x_1(t), u_1(t)\big) \\ \qquad\qquad + \big[\xi^c_x(t,t)\partial_x\varphi(t, x_1(t), u_1(t)) - \lambda^c_x(t)\big]\eta\big(t, x_1(t), u_1(t)\big) \\ \qquad\qquad + \big[\xi_z(t,t) - \lambda_z(t)\big]\partial_x\eta\big(t, x_1(t), u_1(t)\big) \\ \qquad \lambda^c_x(T) = \partial_x S_1\big(x_1(T)\big) \\ -\dot\lambda_z(t) = g_1\big(t, x_1(t), u_1(t)\big) + \eta\big(t, x_1(t), u_1(t)\big) K\big(t, x_1(t), u_1(t)\big) \\ \qquad\qquad + \big[\xi_z(t,t) - \lambda_z(t)\big]\eta\big(t, x_1(t), u_1(t)\big) \\ \qquad \lambda_z(T) = S_1\big(x_1(T)\big) \\ \\ -\dot\xi^c_x(s,t) = \partial_x g_2\big(s, t, x_2(s,t), u_2(s,t)\big) + \xi^c_x(s,t)\partial_x f_2\big(s, t, x_2(s,t), u_2(s,t)\big) \\ \qquad \xi^c_x(s,T) = \partial_x S_2\big(s, x_2(s,T)\big) \\ -\dot\xi_z(s,t) = g_2\big(s, t, x_2(s,t), u_2(s,t)\big) \\ \qquad \xi_z(s,T) = S_2\big(s, x_2(s,T)\big) \end{cases}$$

Approach comparison

After presenting the two approaches, let us compare them to see the respective pros and cons, both from the computational point of view and the information that can be deduced from their solutions.

The backward approach offers the advantage of deriving the optimal strategy in feedback form, which is very handy if the planner has access to the value of the state variable at every time. However, this comes at the cost of having to compute $V_2(s,t,x)$ for every t, x – not just $t = s$ and $x = \varphi\big(s, x_1(s), u_1(s)\big)$ – which is no easy task in general. Moreover, the computation of a value function, such as V^c and V_2, suffers from the curse of dimensionality (both analytically and numerically): the complexity of the problem increases with the dimension of x (i.e., the number of state variables).

The heterogeneous approach, on the contrary, allows to derive the strategy only in the open-loop form. On the other hand, it allows for a compact and unified representation of the model, the necessary conditions, and the dynamics, where the interaction between the two stages is made explicit. Also, the problem is numerically tractable in the general case and for any number of state variables.

3 Marketing model

Let us introduce a marketing problem where the two approaches presented above can be applied. We consider a firm that plans the advertising campaign to increase the demand for its product. We model the problem based on the well know “Advertising Goodwill model” introduced by Nerlove and Arrow in [10], although in finite time.

A firm controls advertising $a(t)$ in order to increase the demand for its product. The Goodwill $G(t)$ is a state variable that represents the past and present effects of advertising on the demand of a given product.

The following assumptions hold:

- the demand $D(t)$ is proportional to the Goodwill;
- advertising increases the Goodwill;
- in absence of advertising, the Goodwill tends to zero exponentially at rate δ.

The firm maximizes their utility, which is the sum of an inter-temporal term and a salvage value. The integral collects the stream of profit over time: it includes the net profit from the product's sales $(p-c)D(G)$ and the advertising cost $\kappa a^2/2$. The salvage value, proportional to the final Goodwill, captures the interest of the firm in sustaining the brand value.

The resulting problem is formulated below:

$$\text{(7)} \qquad \underset{a(t)\geq 0}{\text{maximize}} \left\{ \int_0^T \left[(p-c)D\big(G(t)\big) - \frac{\kappa}{2}a(t)^2\right] dt + \sigma G(T) \right\}$$

subject to:

$$\begin{cases} \dot{G}(t) = a(t) - \delta G(t) & \text{for } t \in [0,T] \\ G(0) = G_0 \end{cases}$$

Variables	**Parameters**
$a(t)$ = advertising (control)	p = unit selling price ($p > c$)
$G(t)$ = Goodwill (state)	c = unit production cost
$D(G) = \pi G$ = demand ($\pi > 0$)	κ = marginal advertising cost
	σ = weight of final Goodwill
	δ = Goodwill deprecation rate

Now, a myopic planner (who is not expecting any changes), computes the optimal advertising strategy by solving problem (7). From Pontryagin's Maximum Principle we get the following optimal control:

$$a(t) = \big[\lambda_G(t)/\kappa\big]^+$$

where $\lambda_G(t)$ is the co-state function that satisfies the following adjoint system:

$$\begin{cases} -\dot{\lambda}_G(t) = (p-c)\pi - \delta\lambda_G \\ \lambda_G(T) = \sigma. \end{cases}$$

Solving for $\lambda_G(t)$, we get

$$\lambda_G(t) = \psi + \chi e^{-\delta(T-t)} > 0$$

where

$$\psi = (p-c)\pi/\delta, \qquad \chi = \sigma - \psi,$$

so that

$$a(t) = \lambda_G(t)/\kappa.$$

Switching time adaptation

Let us introduce a stochastic switching time in the model described above. Suppose that the higher the demand, the greater the risk of an abrupt rise in the unit production cost c. Based on these assumptions, the effect of the switch is simply the following change in the running payoff function:

$$c = \begin{cases} c_1 & \text{for } t \le \tau \\ c_2 > c_1 & \text{for } t > \tau, \end{cases}$$

whereas the hazard rate is increasing in the demand. Specifically, we model a hazard rate that is affine in the demand:

$$\eta(D) = \alpha D + \beta,$$

Hence, the hazard rate of τ at time t is (with an abuse of notation with respect to the argument of the function η):

$$\eta\big(G(t)\big) = \alpha\pi G(t) + \beta. \tag{8}$$

All the other data (dynamics, salvage value, control set) do not change upon the switching time, there is no lump sum at τ, and the Goodwill is continuous at τ, i.e., $G_2(s,s) = G_1(s)$.

Solution: Backward approach

We begin by observing that the Stage 2 data g_2, f_2, S_2 do not depend on s. This implies that the Stage 2 value function V_2 will not depend on s either:

$$V_2(s,t,G) = V_2(t,G).$$

Suppose that V_2 is differentiable, then, it solves the following system:

$$\begin{cases} -\partial_t V_2(t,G) = \max\limits_{\mathtt{a}\geq 0} \Big\{ (p-c_2)\pi G - \dfrac{\kappa}{2}\mathtt{a}^2 + \partial_G V_2(t,G)(\mathtt{a} - \delta G) \Big\} \\ V_2(T,G) = \sigma G \end{cases} \tag{9}$$

The RHS is maximized by the following feedback strategy:

$$\Phi_2(t,G) = \big[\partial_G V_2(t,G)/\kappa\big]^+.$$

Let us suppose that $\partial_G V_2(t,G) \geq 0$, hence $\Phi_2(t,G) = \partial_G V_2(t,G)/\kappa$. After computing the maximum in the RHS by substituting $\Phi_2(t,G)$, we obtain the following PDE:

$$-\partial_t V_2(t,G) = (p-c_2)\pi G - \partial_G V_2(t,G)\delta G + \partial_G V_2(t,G)^2/(2\kappa).$$

Suppose that there exist functions $A(t)$ and $B(t)$ such that

$$V_2(t,G) = A(t)G + B(t),$$

then, system (9) translates to:

$$\begin{cases} -\dot{A}(t)G - \dot{B}(t) = (p-c_2)\pi G - A(t)\delta G + A(t)^2/(2\kappa) \\ A(T)G + B(T) = \sigma G \end{cases}$$

By separating the terms of degree 1 and 0 with respect to G, we obtain a system of backward ODEs in $A(t)$ and $B(t)$:

$$\begin{cases} -\dot{A}(t) = (p-c_2)\pi - \delta A(t) & A(T) = \sigma \\ -\dot{B}(t) = A(t)^2/(2\kappa) & B(T) = 0 \end{cases}$$

Let us solve for $A(t)$ first:

$$A(t) = \psi_2 + \chi_2 e^{-\delta(T-t)} > 0 \tag{10}$$

where

$$\psi_2 = (p-c_2)\pi/\delta, \qquad \chi_2 = \sigma - \psi_2.$$

As for $B(t)$, we limit ourselves to expressing it as an integral function:

$$B(t) = \int_t^T A(\theta)^2/(2\kappa)\, d\theta.$$

Now, $\partial_G V_2(t,G) = A(t)$ is positive, in accordance with our assumption, so that indeed

$$\Phi_2(t,G) = A(t)/\kappa.$$

This is a *degenerate* feedback strategy, i.e., it does not depend on G. As a consequence, it is not necessary to compute the resulting state trajectory to find the strategy as a function of time:

$$a_2(s,t) = A(t)/\kappa.$$

This strategy does not depend on the switching time either.

Now let us solve the Stage 1 problem. Recalling (6), the Stage 1 system of HJB equation and terminal condition is:

$$\begin{cases} -\partial_t V^c(t,G) = \max\limits_{\mathtt{u}\in U_1} \Big\{ (p-c_1)\pi G - \dfrac{\kappa}{2}\mathtt{a}^2 + \partial_G V^c(t,G)(\mathtt{a} - \delta G) \\ \qquad\qquad\qquad\qquad + \eta(G)\Big[V_2(t,G) - V^c(t,G)\Big]\Big\} \\ V^c(T,G) = \sigma G \end{cases} \tag{11}$$

where $V_2(t,G) = A(t)G + B(t)$.

The RHS of the HJB equation in (11) is maximized by

$$\Phi_1(t,G) = \big[\partial_G V^c(t,G)/\kappa\big]^+. \tag{12}$$

Suppose that $\partial_G V^c(t,G) \geq 0$. After computing the max in (11) by substituting (12), and recalling the formulation of the hazard rate function (8), the PDE becomes:

$$\begin{aligned} -\partial_t V^c(t,G) =& (p-c_1)\pi G + \partial_G V^c(t,G)^2/(2\kappa) - \partial_G V^c(t,G)\delta \\ &+ (\alpha\pi G + \beta)\Big[V_2(t,G) - V^c(t,G)\Big] \end{aligned}$$

The term $\alpha\pi G V^c(t,G)$ makes the PDE hard to solve analytically, so here we only present the analytical solution of the case $\alpha = 0$.

Constant hazard rate

With $\alpha = 0$, the hazard rate is $\eta(G) = \beta$ constant, meaning that τ is an exponential random variable, and as such its distribution is independent of the system. Such a distribution can represent an unpredictable, external event such as a disruption in the supply chain, or a pandemic or war outbreak.

The updated system, having substituted $\eta(G) = \beta$, reads:

$$\begin{cases} -\partial_t V^c(t,G) = (p-c_1)\pi G - \partial_G V^c(t,G)\delta G + \partial_G V^c(t,G)^2/(2\kappa) \\ \qquad\qquad\qquad\qquad + \beta\big[V_2(t,G) - V^c(t,G)\big] \\ V^c(T,G) = \sigma G \end{cases} \tag{13}$$

Suppose that there exist functions $C(t)$ and $D(t)$ such that:

$$V^c(t,G) = C(t)G + D(t).$$

then, system (13) translates to:

$$\begin{cases} -\dot{C}(t)G - \dot{D}(t) = (p-c_1)\pi G - C(t)\delta G + C(t)^2/(2\kappa) + \beta\big[A(t)G + B(t) - C(t)G - D(t)\big] \\ C(T)G + D(T) = \sigma G. \end{cases}$$

By separating the terms of degree 1 and 0 with respect to G, we obtain a system of backward ODEs in $C(t)$ and $D(t)$:

$$\begin{cases} -\dot{C}(t) = (p - c_1)\pi - \delta C(t) + \beta\big[A(t) - C(t)\big] & C(T) = \sigma \\ -\dot{D}(t) = C(t)^2/(2\kappa) + \beta\big[B(t) - D(t)\big] & D(T) = 0 \end{cases}$$

Let us solve for $C(t)$ first:

$$C(t) = \gamma + \chi_2 e^{-\delta(T-t)} + \mu e^{-(\delta+\beta)(T-t)} > 0 \tag{14}$$

where

$$\gamma = [(p - c_1)\pi + \beta\psi_2]/(\delta + \beta), \qquad \mu = \psi_2 - \gamma.$$

As for $D(t)$, we limit ourselves to expressing it as an integral function:

$$D(t) = \int_t^T e^{-\beta(\theta-t)}\big[\beta B(\theta) + C(\theta)^2/(2\kappa)\big]\, d\theta.$$

Solution: heterogeneous

We begin by observing that φ and η do not depend on the control, so the Stage 1 optimal strategy satisfies the maximality condition 1b. Coupling it with the maximality condition 2. for the Stage 2 optimal strategy, we get:

$$a_1(t) = \big[\lambda_G^c(t)/\kappa\big]^+, \qquad a_2(t) = \big[\xi_G^c(s,t)/\kappa\big]^+$$

The co-state functions satisfy the adjoint system 3., which for this problem translates to:

$$\begin{cases} -\dot{\lambda}_G^c(t) = (p - c_1)\pi - \lambda_G^c(t)\delta + \big[\xi_G^c(t,t) - \lambda_G^c(t)\big]\eta\big(G_1(t)\big) \\ \qquad\qquad\qquad\qquad\qquad\qquad + \big[\xi_z(t,t) - \lambda_z(t,t)\big]\eta'\big(G_1(t)\big) \\ \qquad \lambda_G^c(T) = \sigma \\ -\dot{\lambda}_z(t) = (p - c_1)\pi G_1(t) - \frac{\kappa}{2}a_1(t)^2 + \big[\xi_z(t,t) - \lambda_z(t)\big]\eta\big(G_1(t)\big) \\ \qquad \lambda_z(T) = \sigma G_1(T) \\ -\dot{\xi}_G^c(s,t) = (p - c_2)\pi - \xi_G^c(s,t)\delta \\ \qquad \xi_G^c(s,T) = \sigma \\ -\dot{\xi}_z(s,t) = (p - c_2)\pi G_2(s,t) - \frac{\kappa}{2}a_2(s,t)^2 \\ \qquad \xi_z(s,T) = \sigma G_2(s,T) \end{cases} \tag{15}$$

We solve for $\xi_G^c(s,t)$:

$$\xi_G^c(s,t) = A(t) > 0,$$

where $A(t)$ was defined in equation (10). Then,

$$a_2(s,t) = A(t)/\kappa,$$

as we obtained with the backward approach in the previous section.

Now we would like to solve for $\lambda^c_G(t)$, however, recalling the formulation of the hazard rate (8), if $\alpha \neq 0$, the FW-BW ODE system of state and co-states is coupled and nonlinear, hence it is difficult to solve analytically. For this reason, we only present the analytical solution of the case where $\alpha = 0$.

Constant hazard rate

As we discussed in the previous section, with $\alpha = 0$, the hazard rate is $\eta(G) = \beta$ constant, meaning that τ is an exponential random variable, and as such its distribution is independent of the system. Such a distribution can represent an unpredictable, external event such as a disruption in the supply chain, or a pandemic or war outbreak.

After substituting $\eta(G) = \beta$ in (15), the ODE for $\lambda^c_G(t)$ is autonomous:

$$\begin{cases} -\dot{\lambda}^c_G(t) = (p - c_1)\pi - \lambda^c_G(t)\delta + \left[\xi^c_G(t,t) - \lambda^c_G(t)\right]\beta \\ \lambda^c_G(T) = \sigma \end{cases}$$

Solving for $\lambda^c_G(t)$ we obtain:

$$\lambda^c_G(t) = C(t) > 0,$$

where $C(t)$ was defined in equation (14). Then,

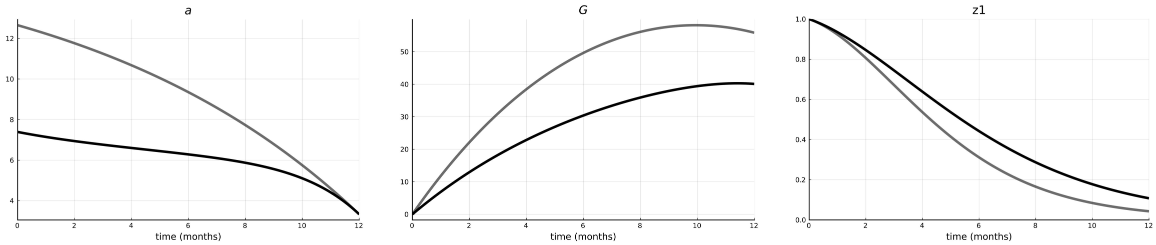
$$a_1(t) = C(t)/\kappa,$$

as we obtained with the backward approach in the previous section.

Numerical simulations

In this section we simulate the scenario where $\alpha > 0$, i.e., the hazard rate depends explicitly on the Goodwill, that we could not solve analytically. The code that was used to make the simulations is a gradient descent method which is based on the heterogeneous approach.

In Fig. 2 we present a comparison between a myopic planner (in gray), and the Stage 1 of a farsighted planner (in black).



Advertising Goodwill z_1

Figure 2: Myopic (gray) vs Farsighted (black).

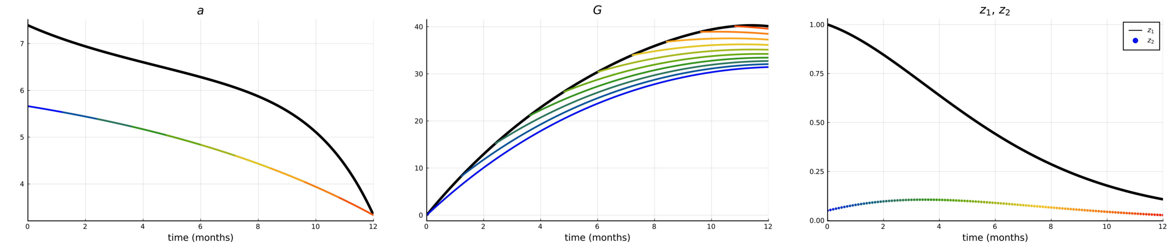
We observe that the farsighted planner implements a more cautious advertising strategy in anticipation of the switch. This happens because higher advertising would lead to a higher Goodwill and hence a higher risk of switching early. The rightmost plot is $z_1(t)$, i.e., the probability of still being in Stage 1 at time t: observe that it is higher in the farsighted case.

With the chosen set of parameters, the expected payoff of the farsighted planner is about 75, whereas for the myopic planner is about 45.

In Fig. 3 we represent the optimal behavior of the farsighted planner in Stage 1 (black) and in Stage 2 (color). Consider the Goodwill's plot. Each colored line originating from time s represents the Stage 2 behavior if the switching time occurs at time s; if $\tau = s$, we follow the Stage 1 line until time s, then we follow the corresponding Stage 2 line from s to T.

Now consider the advertising's plot. Observe that all the Stage 2 lines lie on the same line. This is because, as we found analytically, the Stage 2 strategy does not depend on s (in this linear-state model, not in general).

The optimal advertising is abruptly lowered upon the switch, due to the reduced profit margin $(p - c_2) < (p - c_1)$.



Advertising

Goodwill

z_1, z_2

Figure 3: Farsighted Stage 1 (black) vs Stage 2 (color).

4 Future developments

We aim to consider two additional planners, with intermediate degrees of information compared to the two considered here:

- a myopic planner who realizes when τ occurs and adjusts their strategy to the new regime;
- a farsighted planner who is obliged to persist with an initially declared strategy despite the switch.

Overall, we can represent the four planners in a table, as in Fig. 4.

	Myopic	**Farsighted**
Cannot update	1	3
Can update	2	4

Figure 4: Four scenarios.

In this work we considered scenarios 1 and 4, which are respectively the worst and the best case. It would be interesting to see how scenarios 2 and 3 affect the expected payoff.

References

[1] A. Buratto, M. Muttoni, S. Wrzaczek, and M. Freiberger, *Should the COVID-19 lockdown be relaxed or intensified in case a vaccine becomes available?.* PLOS ONE, 17 (2022), p. e0273557.

[2] H. Dawid, M.Y. Keoula, M. Kopel, and P.M. Kort, *Product innovation incentives by an incumbent firm: A dynamic analysis.* Journal of Economic Behavior & Organization, 117 (2015), pp. 411–438.

[3] E. Dockner, S. Jørgensen, N.V. Long, and G. Sorger, "Differential Games in Economics and Management Science". Cambridge University Press, Nov. 2000.

[4] M. Freiberger, *Two-stage-optimal-control.* (2023).

[5] D. Grass, J.P. Caulkins, G. Feichtinger, G. Tragler, D.A. Behrens, et al., "Optimal control of nonlinear processes". Berlino, Springer, 2008.

[6] D. Gromov and E. Gromova, *On a Class of Hybrid Differential Games.* Dynamic Games and Applications, 7 (2017), pp. 266–288.

[7] J.L. Haunschmied, R.M. Kovacevic, W. Semmler, and V.M. Veliov, eds., *Dynamic Economic Problems with Regime Switches.* Vol. 25 of Dynamic Modeling and Econometrics in Economics and Finance, Springer International Publishing, Cham, 2021.

[8] M. Kuhn and S. Wrzaczek, *Rationally Risking Addiction: A Two-Stage Approach.* In "Dynamic Economic Problems with Regime Switches", J.L. Haunschmied, R.M. Kovacevic, W. Semmler, and V.M. Veliov, eds., Dynamic Modeling and Econometrics in Economics and Finance, Springer International Publishing, Cham, 2021, pp. 85–110.

[9] N.V. Long, F. Prieur, M. Tidball, and K. Puzon, *Piecewise closedloop equilibria in differential games with regime switching strategies.* Journal of Economic Dynamics and Control, 76 (2017), pp. 264–284.

[10] M. Nerlove and K.J. Arrow, *Optimal advertising policy under dynamic conditions.* Economica (1962), pp. 129–142.

[11] S. Polasky, A. de Zeeuw, and F. Wagener, *Optimal management with potential regime shifts.* Journal of Environmental Economics and Management, 62 (2011), pp. 229–240.

[12] Y. Tsur and A. Zemel, *Coping with Multiple Catastrophic Threats.* Environmental and Resource Economics, 68 (2017), pp. 175–196.

[13] V.M. Veliov, *Optimal control of heterogeneous systems: Basic theory.* Journal of Mathematical Analysis and Applications, 346 (2008), pp. 227–242.

[14] S. Wrzaczek, M. Kuhn, and I. Frankovic, *Using Age Structure for a Multi-stage Optimal Control Model with Random Switching Time.* Journal of Optimization Theory and Applications, 184 (2020), pp. 1065–1082.

Cutting pattern optimization in sawmill

Enrico Vicario [(*)]

Abstract. Optimizing the cutting of wood from log to board is a step within the lumber production chain that has great potential for optimization. This processing in industry is carried out in lines with machines that are increasingly automated and have varying degrees of flexibility in cutting execution. In this work, a specific problem of generating optimal cutting patterns was modeled in order to address it as a Mixed Integer Linear Problem (MILP). The mathematical formulation has been tested on real instances and the result was compared, in terms of obtained value and computation time, with an ad-hoc greedy heuristics.

1 Introduction

Sawmill log cutting is a process that can be performed with various cutting technologies and different levels of automation. In general, the log cutting process involves removing the bark and then reducing the log into boards. This process can be carried out with milling machines and circular or band saws, with or without the use of laser technologies or other sensors to detect log dimensions and characteristics. The choice of technology depends on production requirements and the quality of the wood to be processed.

Some lines prefer cutting speed over flexibility; in these cases typically logs are presorted into homogeneous groups by diameter and the cutting pattern used remains the same for the entire production run. In some lines, although the main cutting pattern remains fixed, there remain degrees of freedom that can be optimized, such as the rotation and alignment of the log with respect to the cutting pattern, or the alignment and size of only the outermost boards [5].

On the other hand, other cutting lines, which are certainly more interesting from an optimization point of view, allow for variation in the size of each individual cut that is made on each log.

Consequently, the optimization process must be aligned with the actual possibilities of the line and the machines that will run it, which translates into constraints for optimization but also into different algorithmic choices. For example, in [1] a method based on two levels of dynamic programmig was used for cutting patterns consisting of side-by-side

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: `evicario@math.unipd.it`. With MICROTEC Srl,Via Miranese 56, 30171 Venezia, Italy; E-mail: `enrico.vicario@microtec.eu`. Seminar held on 31 May 2023.

vertical cuts, each split into further side-by-side cuts. In [4], in addition to the case of vertical side-by-side cuts (live sawing), the case of cutting patterns with a central block (cant sawing) and optimization with successive rotations (grade sawing) is also analyzed, applying algorithms based on dynamic programming in all cases. In [2], on the other hand, the problem as filling a circumference with rectangles is addressed by setting up a Mixed Integer Nonlinear Programming (MINLP) model; this approach generalizes the problem but untangles it from the actual cutting patterns of the line, so the obtained solution is unlikely to be cut in an automated line.

The cutting pattern describes the way a log is cut in terms of type and position of the cuts, and determines a log cutting solution. Generally, the cutting pattern is represented with a 2D visualization of the pattern of boards, represented by rectangles, that will then be cut from the log. This representation remains valid in some contexts but is not always sufficient to represent all the degrees of freedom that may be present in the cutting line. For example, some boards may be cut at inclinations that are independent of the main direction of the log, or even follow a curved profile if allowed by the cutting machine.

The main result of cutting a log are the boards, and the tasks of optimization is to maximize their overall value. The board's value depends on its size and quality. The quality is generally defined by a set of rules regarding defects or other properties of the board that must be satisfied. For example, the presence of knots or cracks can affect the quality of the board and thus its value.

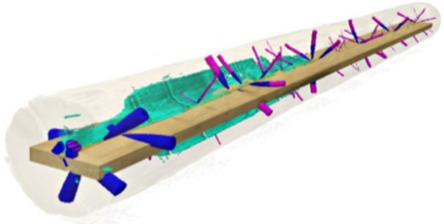


Figure 1: Virtual board on CT modeled log

The incorporation of new measurement technologies within the cutting lines, such as computed tomography scan, has made it possible to obtain accurate modeling of the defects within the log before cutting and thus have a positional value of the board (see Figure 1) to be used in the context of log cutting pattern optimization [6].

There are also other aspects that are generally complex to model and include in optimization and therefore happen to be often not properly considered, such as

- the value of byproducts generated during cutting such as sawdust and wood chips;
- the utilization time and machinery wear and tear that a given cutting pattern involves;
- the change in value and demand for given products during the current production;
- the later processing of boards for more accurate prediction of the value;

- the unavoidable and non-negligible uncertainty in the execution of the cutting pattern;
- the errors in the measurement on which optimization is based.

2 A cutting pattern optimization problem

The optimization problem presented here arises from an industrial application where it is necessary to find out the best cutting pattern based only on the log diameter and a set of products to be used to compose it. What makes this problem complicated is the large number of products available.

The shape of the log is then assumed as a cylinder: this assumption reduces the log optimization problem into a easier 2D optimization, since the rotation or inclination of the cutting pattern as well as different slopes of the side cuts cannot give a better solution. We also consider values such that the length of the products or their longitudinal position in the log have not influence in the research of the optimum.

Under these conditions, we can consider the products as rectangles and the log shape as a circumference of given diameter. In addition, the value of the boards is not dependent on their position inside the log and the boards are valid only if fully contained in the circle.

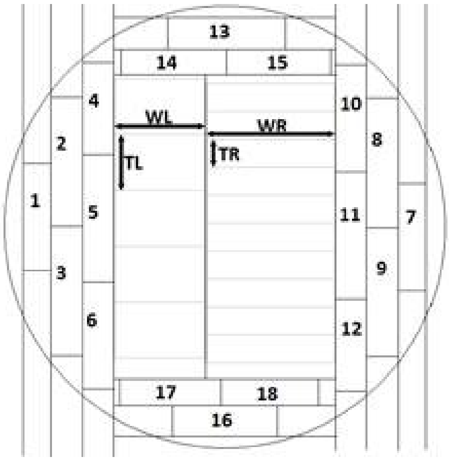


Figure 2: Construction diagram of the cutting pattern

It is a matter of constructing the best cutting pattern in two dimensions, in order to fill the area defined by a cylindrical trunk, maximizing the overall value of the inserted products. The products are supplied in two separate lists, one for the main products and one for the side products; this differentiation is due to the structure of the cutting pattern as shown in Figure 2. The feasible cutting pattern is structured as follows:

(a) One or two main product blocks. Each of these blocks consists of a sequence of boards of the same size stacked on top of each other. The two blocks may consist of boards of different sizes and therefore have different overall block heights;

(b) In the right and left of the two main blocks, there is a sequence of 1 to 3 side product cuts. These cuts can contain one or more products of the same thickness placed side by side (boards 1 to 12 in Figure 2);

(c) At the top and bottom of the cutting pattern, there are two further sequences of 1 or 2 cuts of side boards (13 to 18 in Figure 2).

The various cuts and boards are spaced by a fixed thickness (the thickness of the saw that will make the cut).

Additional specific limitations due to the mechanics of the line are ignored: any combination of boards that meets what is specified in the constraints described above is considered valid.

It is possible to make an indicative estimate of the number of cutting pattern combinations possible depending on the number of products available. Let S be the number of thicknesses, and K the approximate number of different product widths for the same thickness. Applying basic formulas of combinatorial calculus (combinations and dispositions), we obtain:

$$N_{comb} \approx (D_{s,3}C_{k,3}C_{k,2}C_{k,1})^2(D_{s,2}C_{k,2}C_{k,1})^2(SK)^2(SK)^2 \approx S^{14}K^{22}$$

A dedicated heuristic was developed at Microtec to address this specific problem, obtaining at first glance satisfactory results in reasonable time for industrial use. To validate the results obtained with this heuristic, it was chosen to formulate the problem as Mixed Integer Linear Programming (MILP). This made it possible to use standard solving techniques already implemented and optimized in commercial software. Moreover, this approach, through constraint relaxation, provided an upper bound for optimization.

3 MILP formulation

Let M and S two lists of products (main boards and side boards), where for each product they are defined its width w, thickness t and value v.

$$w_i^M, t_i^M, v_i^M, i = 1...N^M$$
$$w_i^S, t_i^S, v_i^S, i = 1...N^S$$

Let R_C also be the radius of the cylinder to be optimized. The cylinder is assumed centered in (0,0). Let s be the saw thickness assumed for simplicity fixed for all cuts.

The cutting pattern can be conveniently modeled by decomposing it in the 10 single lateral cuts (L, R, T, B) and in the 2 main cuts (C0, C1) as shown in Figure 3. Given the symmetry of the cutting pattern and the cylinder on which it will be applied, it is convenient to define the variables and constraints of each cut considering it in a basic direction and then return it to its actual position during the addition of further constraints that consider the cutting pattern in its entirety.

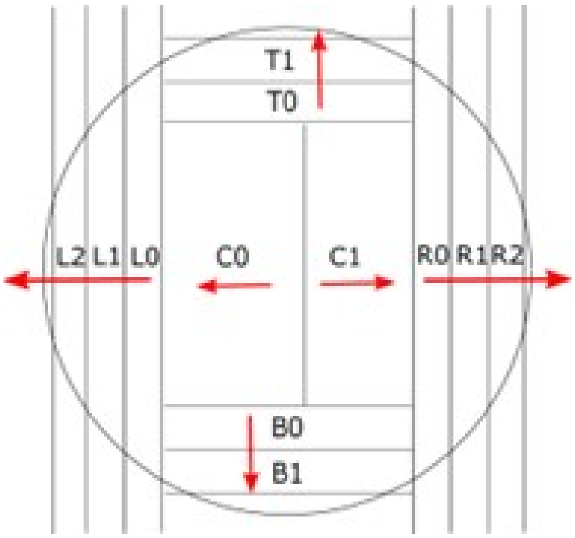


Figure 3: Cut types and order

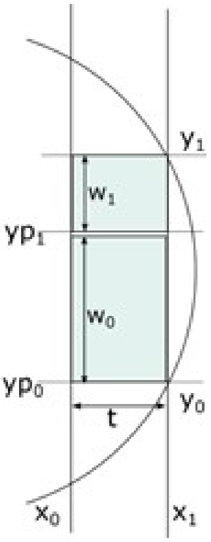


Figure 4: Side cut model

3.1 Template for a generic side cut

We define, for each side cut (Figure 4), the following variables:

- cut limits (considering only the products actually present)

$$x_0, x_1 \in [0, R_C],\ y_0, y_1 \in [-R_C, R_C];$$

- thickness of the cut (must coincide with the thickness of the products present)

$$t \in [0, R_C];$$

- denoting by N^P the maximum number of products allowed for the cut, position and width of products

$$y_i^P \in [-R_C, R_C],\ w_i^P \in [0, 2R_C],\ i \in 1...N^P;$$

- product presence:

  $P_{ij}^S \in [0, 1]$ *integer*, $i \in 1...N^P, j \in 1...N^S$ equal to 1 if at position i there is the product j,

  $P_i \in [0, 1]$ *integer*, $i \in 1...N^P$ equal to 1 if at position i there is a product.

The following constraints are then inserted:

- thickness of the cut: $x_0 + t = x_1$;

- presence of the product: $\sum_{j=1}^{N^S} p_{ij}^S = p_i, i \in 1...N^P$ (since p_i is binary, this constraint also includes the fact that at most one product can be active);
- thickness of the cut equal to that of the present products:

$$\sum_{j=1}^{N^S} p_{ij}^S t_j^S \leq t + K(1 - p_i), \quad \sum_{j=1}^{N^S} p_{ij}^S t_j^S \leq t + K(1 - p_i), \;\; i \in 1...n^P$$

  (we consider the possibility that the cut is not present, in which case the constraint has no effect for K sufficiently large);

- product width (for convenience we also add a saw thickness if the product is present): $\sum_{j=1}^{N^S} p_{ij}^S (w_j^S + s) = w_i^P, \;\; i \in 1...n^P$;
- vertical position of products:
  $y_i^P = y_{i-1} + w_{i-1}^P, \;\; i \in 1...n^P$;
- compulsory presence of the cut (needed for the innermost side cut (L0, B0, R0, T0) to force the presence of at least one cut:
  $p_1 = 1$;
- continuous presence with respect to the index order:
  $p_i \leq p_{i-1} \; i \in 2...N^P$;
- vertical cut limits:
  $y_0 = y_o^P, y_1 = y_0 - s + \sum_{t=1}^{N^P} w_i^P$;
- constraints relative to the log:

  we must ensure that the products are inside the cylinder. It is sufficient to verify that the ends (x_1, y_0) and (x_1, y_1) are inside the circle of radius R_c. Linear constraints are then added on these variables to approximate a circle. The chosen approximation is such that it enlarges the original domain of the problem while keeping the maximum distance from the circle below 0.5mm; for the variable (x_1, y_0) positive slope constraints, (x_1, y_0) negative slope constraints.

  Let d be the maximum permissible distance between the circumference and the polygon defined by the linear constraints to be inserted (see Figure 5). It is easy to determine the angular step θ needed for a succession of these constraints:

$$\theta = 2 * \cos^{-1}(1 + \frac{d}{R_c}).$$

  For each of the two points, N constraints will be needed, with N defined by $N = \lceil \frac{\pi}{2\theta} \rceil$.

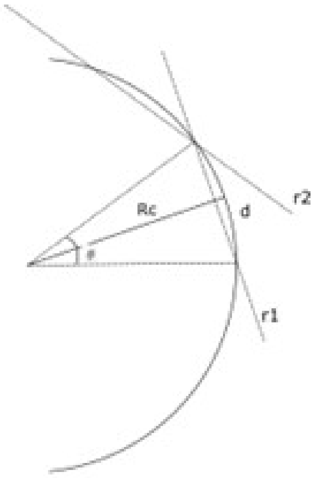


Figure 5: Shape constraints

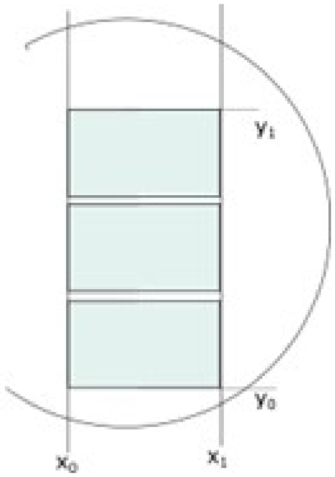


Figure 6: Main cut model

And then:

$$y_1 \leq -\frac{x_1}{\tan(\frac{\theta}{2}+k\theta)} + \frac{R_C}{\sin(\frac{\theta}{2}+k\theta)} + M(1-p_1), \quad k \in 1...N,$$

$$y_0 \geq \frac{x_1}{\tan(\frac{\theta}{2}+k\theta)} - \frac{R_C}{\sin(\frac{\theta}{2}+k\theta)} - M(1-p_1), \quad k \in 1...N.$$

In the event that the side cut may not appear, a term M is added to make the constraint redundant if no products are selected in the cut (M>0 large enough).

Contribution to objective function: each side cut inserted into the model contributes the following value to the objective function:

$$\sum_{i=1}^{N^P}\sum_{j=1}^{N^S} p_{ij}^S v_j^S.$$

3.2 Template for a main cut block

Let N^{MAX} be the maximum number of products that can be inserted in the cut (in standard conditions we can assume it to be 20). We define, for each of the central cuts, the following variables (see Figure 6):

- limits of the main cut:

$$x_0, x_1 \in [-R_C, R_C], y_0, y_1 \in [-R_C, R_C];$$

- selected product:

$$p_i \in [0,1], integer, i \in 1...N^M;$$

- quantity of products:

$$n_i^P \in [0, N^{MAX}], integer, i \in 1...N^M.$$

  Defining the quantity for each product, instead of the quantity of the selected product only, allows the following constraint of the height of cut to be expressed by maintaining linearity.

We add the following constraints:

- only one selected product:

$$\sum_{i=1}^{n^M} p_i = 1;$$

- width and height of the cut:

$$x_0 + \sum_{i=1}^{N^M} p_i w_i^M = x_1, \;\; y_0 - s + \sum_{i=1}^{N^M} n_i^P (t_i^M + s) = y_1;$$

- constraints relative to the log:

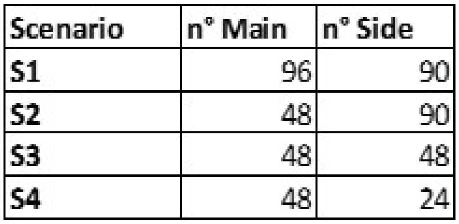
  they are expressed in an analogous way to the case of the side cut.

Contribution to objective function: each of the main cuts contributes the following value to the objective function:

$$\sum_{i=1}^{N^M} n_i^P v_i^M.$$

3.3 Global constraints for joining cuts

Additional constraints are added to create the sequences of cuts and join main and side cuts. The direction of the various cuts must be taken into account.



Scenario	n° Main	n° Side
S1	96	90
S2	48	90
S3	48	48
S4	48	24

Figure 7: Scenarios

4 Results

The model was implemented in C++ using the Gurobi API (version 9.11) and ran on a PC with an Intel(R) Core(TM) i7-7820HQ CPU, clock 2.90GHz, 4 cores, 8 logical processors.

Four scenarios were defined with a different amount of defined products (see Figure 7). The side products had thickness between 12mm and 40mm, width between 75mm and 175mm. The main products had thickness between 27mm and 57mm, width between 95mm and 200mm

Each of the four scenarios was tested with two different log diameters (400mm and 500mm). Figure 8 shows an example of the optimization result for the scenario S1 and a log diameter of 500mm.

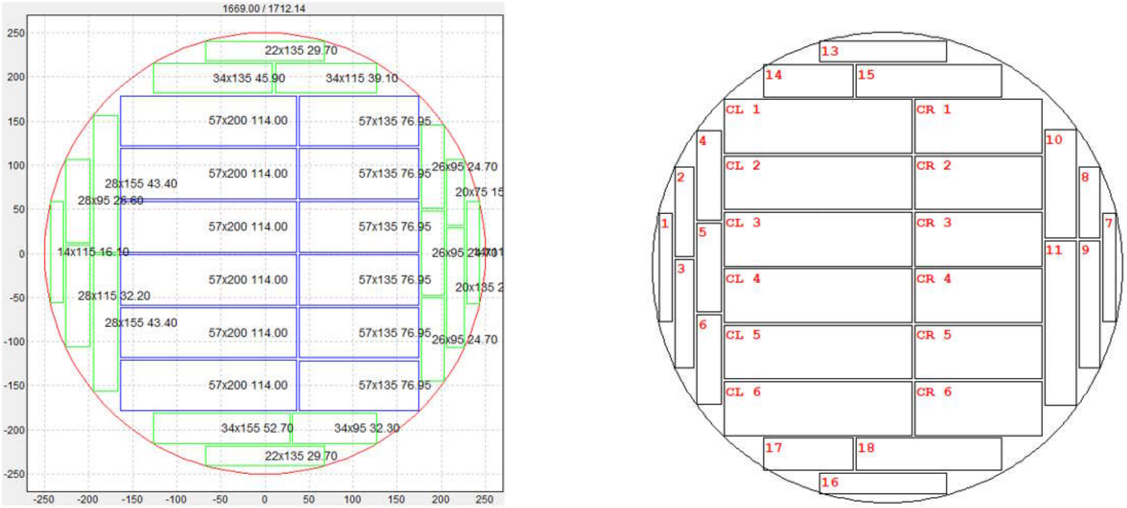
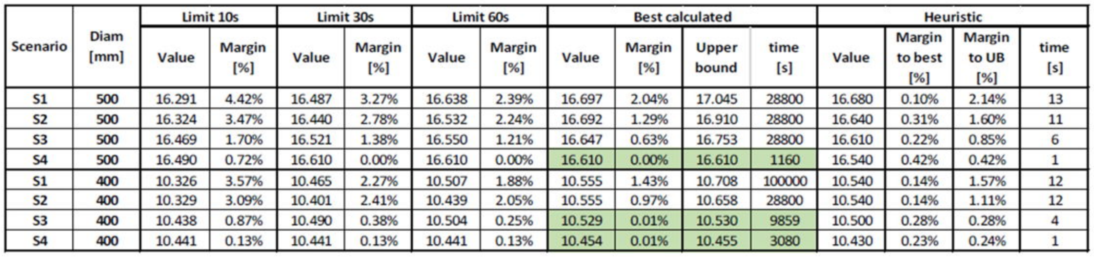


Figure 8: Solution obtained using MILP solver (left) or heuristic algorithm (right)



Scenario	Diam [mm]	Limit 10s		Limit 30s		Limit 60s		Best calculated				Heuristic			
		Value	Margin [%]	Value	Margin [%]	Value	Margin [%]	Value	Margin [%]	Upper bound	time [s]	Value	Margin to best [%]	Margin to UB [%]	time [s]
S1	500	16.291	4.42%	16.487	3.27%	16.638	2.39%	16.697	2.04%	17.045	28800	16.680	0.10%	2.14%	13
S2	500	16.324	3.47%	16.440	2.78%	16.532	2.24%	16.692	1.29%	16.910	28800	16.640	0.31%	1.60%	11
S3	500	16.469	1.70%	16.521	1.38%	16.550	1.21%	16.647	0.63%	16.753	28800	16.610	0.22%	0.85%	6
S4	500	16.490	0.72%	16.610	0.00%	16.610	0.00%	16.610	0.00%	16.610	1160	16.540	0.42%	0.42%	1
S1	400	10.326	3.57%	10.465	2.27%	10.507	1.88%	10.555	1.43%	10.708	100000	10.540	0.14%	1.57%	12
S2	400	10.329	3.09%	10.401	2.41%	10.439	2.05%	10.555	0.97%	10.658	28800	10.540	0.14%	1.11%	12
S3	400	10.438	0.87%	10.490	0.38%	10.504	0.25%	10.529	0.01%	10.530	9859	10.500	0.28%	0.28%	4
S4	400	10.441	0.13%	10.441	0.13%	10.441	0.13%	10.454	0.01%	10.455	3080	10.430	0.23%	0.24%	1

Figure 9: Computational results

The computational results obtained under the different conditions tested are reported in Figure 9. A time limitation of 28800 seconds (8 hours) was generally applied for all the test, except in one configuration where an higher time limit was tested (100000 seconds, about 28 hours). In the simplest 3 cases only, the solver reached an optimum (highlighted in green). In more complex scenarios, calculating the optimal solution can take several days; although it cannot be stated with certainty, it doesn't seem that the incumbent solution will have further significant improvements as time increases. The heuristic approach, in a really short time, always found a solution below 0.42% of difference respect to the best found by the MILP solver, and below 2.14% respect to the upper bound in the worst case.

The MILP approach considered here, although not directly usable in the specific application, certainly allowed us to verify the goodness of the heuristics developed. A similar approach can be easily extended to other types of optimization based on cutting patterns in which the shape of the log remains modelable with linear constraints (a convex shape) but finds its limitations in other applications in which it is necessary to consider the actual shape of the log and in which the value of the products is dependent on their position within the log itself.

References

[1] J. Geerts, *Mathematical solution for optimising the sawing pattern of a log given its dimensions and its defect core.* New Zealand J. Forestry Sci, vol.14, pp.124–134, Mar. 1984.

[2] I. Hinostroza, L. Pradenas, V. Parada, *Board cutting from logs: Optimal and heuristic approaches for the problem of packing rectangles in a circle.* Int. J. Prod. Econ., vol. 145, no. 2, pp. 541–546, Oct. 2013.

[3] P.P. Alvarez, J.R. Vera, *Application of robust optimization to the sawmill planning problem.* Ann. Oper. Res., vol. 219, no. 1, pp. 457–475, Aug. 2014.

[4] S.M. Bhandarkar, X. Luo, R.F. Daniels, E.W. Tollner, *Automated planning and optimization of lumber production using machine vision and computed tomography.* IEEE Trans. Autom. Sci. Eng., vol. 5, no. 4, pp. 677-695, Oct. 2008.

[5] C.G. Lundahl and A. Grönlund, *Increased yield in sawmills by applying alternate rotation and lateral positioning.* Forest Products J,vol. 60, no. 4, pp. 331-338, Jul 2010.

[6] A. Rais, E. Ursella, E. Vicario, F. Giudiceandrea, *The use of the first industrial X-ray CT scanner increases the lumber recovery value: case study on visually strength-graded Douglas-fir timber.* Annals of forest science, 74(2), 28, 2017.

Hopf algebras and Kaplansky's sixth conjecture

Elisabetta Masut [(*)]

These notes would be a gentle approach to some results about the non-existence of Hopf orders for a Hopf algebra.
In Section 1, we introduce the notion of a Hopf algebra and we present a key example of such an algebra. In Section 2, we make a brief excursus on representations theory for Hopf algebras. Section 3 aims at understanding how we can construct new Hopf algebras by means of deformations by twists. Finally, in Section 4, we define what a Hopf order is in order to present the mentioned results in Section 5.

Throughout these sections, K will be a field and unadorned tensor products are intended to be over K.

1 Hopf algebras

The goal of this section is to understand what a Hopf algebra is. Interested readers can find more details in [8].

We start giving an equivalent definition of an associative algebra, by means of commutative diagrams; this will allow us to dualize the definition.

Definition 1.1 A K-*algebra* is a K-vector space A, together with two K-linear maps, multiplication $m\colon A\otimes A\to A$ and unit $u\colon K\to A$, such that the following diagrams are commutative (the two lower maps in the second diagram are given by scalar multiplication):

a) associativity

$$\begin{tikzcd}
A\otimes A\otimes A \arrow[r, "m\otimes\mathrm{id}"] \arrow[d, "\mathrm{id}\otimes m"'] & A\otimes A \arrow[d, "m"] \\
A\otimes A \arrow[r, "m"] & A
\end{tikzcd}$$

b) unit

$$\begin{tikzcd}
& A\otimes A \arrow[dd, "m"] & \\
K\otimes A \arrow[ur, "u\otimes\mathrm{id}"] \arrow[dr] & & A\otimes K \arrow[ul, "\mathrm{id}\otimes u"'] \arrow[dl] \\
& A &
\end{tikzcd}$$

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: emasut@math.unipd.it. Seminar held on 14 June 2023.

Remark 1.2 We stress the fact that the above definition is equivalent to the usual notion of an associative unitary algebra.
Indeed, the commutativity of the first diagram means that for all $a, b, c \in A$

$$m \circ (m \otimes \mathrm{id})(a \otimes b \otimes c) = m \circ (\mathrm{id} \otimes m)(a \otimes b \otimes c),$$

that is $(ab)c = (ab)c$.
Moreover, the unit element 1_A of the algebra A is given by $u(1_K)$. Indeed, the commutativity of the unit diagram leads to the following equality

$$m \circ (u \otimes \mathrm{id})(1_K \otimes a) = m \circ (\mathrm{id} \otimes u)(1_K \otimes a) = 1_K a$$

for all $a \in A$. The above equation implies that

$$u(1_K)a = au(1_K) = a,$$

and so that $u(1_K)$ is the unit of the algebra A.

Definition 1.3 For any pair of K-spaces V and W, the *flip map* $\tau \colon V \otimes W \to W \otimes V$ is given by $\tau(v \otimes w) = w \otimes v$, for every $v \in V$ and $w \in W$.

An algebra A is commutative if and only if $m \circ \tau = m$ on $A \otimes A$.

This equivalent definition of algebra allows us to give the definition of a coalgebra, by simply reversing the direction of the arrows.

Definition 1.4 A *K-coalgebra* is a K-vector space C, together with two K-linear maps, the comultiplication $\Delta \colon C \to C \otimes C$ and the counit $\varepsilon \colon C \to K$, such that the following diagrams are commutative:

a) coassociativity

$$\begin{tikzcd}
C \arrow[r, "\Delta"] \arrow[d, "\Delta"'] & C \otimes C \arrow[d, "\Delta\otimes\mathrm{id}"] \\
C \otimes C \arrow[r, "\mathrm{id}\otimes\Delta"] & C \otimes C \otimes C
\end{tikzcd}$$

b) counit

$$\begin{tikzcd}
& C \arrow[dl, "1\otimes"'] \arrow[dr, "-\otimes 1"] \arrow[dd, "\Delta"] & \\
K \otimes C & & C \otimes K \\
& C \otimes C \arrow[ul, "\varepsilon\otimes\mathrm{id}"] \arrow[ur, "\mathrm{id}\otimes\varepsilon"'] &
\end{tikzcd}$$

We say that C is cocommutative if $\tau \circ \Delta = \Delta$.

Example Let G be a finite group. The main example we have considered in this seminar is the group algebra KG, that is the K-vector space with basis G. This vector space is

both an algebra and a coalgebra.
The algebra structure is naturally inherited from the group structure of G. Specifically, let $x, y \in KG$, that is $x = \sum_{g\in G} k_g g$ and $y = \sum_{h\in H} \tilde{k}_h h$, with $k_g, \tilde{k}_h \in K$. Then,

$$xy = \sum_{g,h\in G} k_g \tilde{k}_h gh.$$

Moreover, the unit element of KG is simply $1_{KG} = 1_K 1_G$.
For the coalgebra structure, we need to define a comultiplication and a counit map, which make the previous diagrams commutative. We define these maps on the basis elements and then we extend them by linearity. In particular,

$$\Delta(g) := g \otimes g$$

and

$$\varepsilon(g) := 1,$$

for all $g \in G$. It can be verified that these maps satisfy the coassociativity and the counit diagram.

In a coalgebra elements which have Δ and ϵ as the ones for the group elements are very important and they deserve a name.

Definition 1.5 Let C be a coalgebra, we say that $c \in C$ is a *group-like* element, if $\Delta(c) = c \otimes c$, and $\varepsilon(c) = 1$.

The next step is to combine the notion of algebra and coalgebra. For that, we need to observe that, given an algebra A (respectively coalgebra C), we can endow $A \otimes A$ (respectively $C \otimes C$) with an algebra (respectively coalgebra) structure.
Specifically, consider an algebra (A, m, u). Then, $A \otimes A$ is an algebra, with multiplication $m_{A\otimes A} = (m \otimes m) \circ (\mathrm{id} \otimes \tau \otimes \mathrm{id})$ and unit $u_{A\otimes A} = (u_A \otimes u_A)(\phi^{-1})$, where $\phi\colon K \otimes K \to K$ is the natural isomorphism.
Analogously, let (C, Δ, ε) be a coalgebra. Then, $C \otimes C$ is a coalgebra, with comultiplication $\Delta_{C\otimes C} = (\mathrm{id} \otimes \tau \otimes \mathrm{id})(\Delta \otimes \Delta)$ and counit $\varepsilon_{C\otimes C} = \phi \circ (\varepsilon \otimes \varepsilon)$, where ϕ is as above.
This allows us to give the following definition.

Definition 1.6 A K-space B endowed with an algebra structure (B, m, u) and a coalgebra structure (B, Δ, ε) is called a *bialgebra* if one of the following equivalent conditions holds:

(a) Δ and ε are algebra morphisms;

(b) m and u are coalgebra morphism.

Example Consider the group algebra KG. As we have just seen, the group algebra is both an algebra and a coalgebra. For being a bialgebra, we need that one of the conditions in Definition 1.6 holds, in particular that Δ and ε are algebra morphisms. We check it only for the comultiplication map, for the counit it is immediate.

The fact that $\Delta(1_{KG}) = 1_{KG\otimes KG} = 1_{KG} \otimes 1_{KG}$ arises from the definition of the unit of KG and of the comultiplication.
The last condition to verify is that $\Delta(gh) = \Delta(g) \cdot \Delta(h)$, for every $g, h \in G$. By definition of Δ in KG there holds

$$\Delta(gh) = gh \otimes gh.$$

Then, by definition of product in $KG \otimes KG$,

$$\Delta(g) \cdot \Delta(h) = (m \otimes m)(\mathrm{id} \otimes \tau \otimes \mathrm{id})(g \otimes g \cdot h \otimes h) = (m \otimes m)(g \otimes h \otimes g \otimes h) = gh \otimes gh.$$

This gives us the desired equality.

Given a bialgebra H, we need a "special" map $S\colon H \to H$ for having the Hopf algebra structure.
For constructing this map, we introduce the following definition.

Definition 1.7 Let C be a coalgebra and let A be an algebra. Then $\mathrm{Hom}_K(C, A)$ becomes an algebra under the *convolution product*:

$$(f * g)(c) = m \circ (f \otimes g)(\Delta(c))$$

for all $f, g \in \mathrm{Hom}_K(C, A)$ and $c \in C$. The unit element in $\mathrm{Hom}_K(C, A)$ is $u\varepsilon$.

We are now in a position to give the definition of a Hopf algebra.

Definition 1.8 Let $(H, m, u, \Delta, \varepsilon)$ be a bialgebra. Then H is a *Hopf algebra* if there exists an element $S \in \mathrm{Hom}_K(H, H)$ which is an inverse to id_H under convolution $*$. S is called an *antipode* for H.

In other words, S is such that for every $h \in H$:

$$m \circ (S \otimes \mathrm{id}_H)\Delta(h) = m \circ (\mathrm{id}_H \otimes S)\Delta(h) = \epsilon(h)1_H.$$

Example Consider again our key example, that is the group algebra KG. We show that it is a Hopf algebra.
We need to find the inverse under the convolution product of the identity map $\mathrm{id}\colon KG \to KG$. This means that we have to find a map $S\colon KG \to KG$ such that for every $g \in G$

$$(S * \mathrm{id})(g) = 1_{KG}.$$

Applying the definition of the convolution product we get

$$(S * \mathrm{id})(g) = S(g)g,$$

for all $g \in G$. Since $S(g)g$ has to be equal to 1_{KG}, we have that $S(g) = g^{-1}$ for all $g \in G$.

Example Another example we discussed during the seminar was the dual of a group algebra, when the group is finite (for the construction see [8, Example 1.3.6]). Indeed, the family of finite-dimensional Hopf algebras is self dual: the dual of a Hopf algebra is a Hopf algebra (for more details see [8, Section 1]).

2 Representation theory for Hopf algebras

In this section, we analyze an interesting property on representations of Hopf algebras.

The Hopf algebra structure allows us to define some operations on representations.
Firstly, recall that a Hopf algebra representation is a representation of its underlying algebra, that is an H-module.

Definition 2.1 A left H-module X consists of an abelian group $(X, +)$ and an operation $. : H \times X \to X$, such that for all $h, h' \in H$ and $x, x' \in X$ we have:

(1) $h.(x + x') = h.x + h.x'$

(2) $(h + h').x = h.x + h'.x$

(3) $(hh').x = h.(h'.x)$

(4) $1_H.x = x$.

Hence, for having a representation of H, we need an action of H on a vector space, which satisfies the above properties.

It is known that for representations of groups, we can define the trivial representation, the tensor product of representations and the dual representation. The group algebra representations inherit these properties.
In particular, let V and W be two representations of the group algebra KG. Then,

(1) the *trivial representation* is defined on K as

$$g.k = k$$

for all $g \in G$, $k \in K$;

(2) the *tensor product* of V and W is a representation via

$$g.(v \otimes w) = g.v \otimes g.w,$$

for all $g \in G$ and $v \in V, w \in W$;

(3) the *dual representation* of V has underlying vector space V^* and it is defined

$$(g.f)(v) = f(g^{-1}.v)$$

for all $f \in V^*, g \in G, v \in V$.

Hopf algebra representations mimic these operations, by means of the counit, comultiplication and antipode. Specifically, let X and Y be two representations of a Hopf algebra H, then

(1) the *trivial representation* on K is defined as

$$h.k = \varepsilon(h)k$$

for all $h \in H$, $k \in K$;

(2) the *tensor product* of X and Y is a representation via

$$h.(x \otimes y) = \Delta(h).(x \otimes y),$$

for all $h \in H$ and $x \in X, y \in Y$

(3) the *dual representation* of X, which is defined on X^* via

$$(h.\phi)(x) = \phi(S(h).x)$$

for all $\phi \in X^*, h \in H, x \in X$.

Remark 2.2 To conclude this brief excursus on representations, it is worth mentioning that the category of representations of a Hopf algebra forms a monoidal category ([5, Definition 2.1.1]). This is a consequence of the fact that we can define the tensor product of representations and the trivial representation, which are compatible thanks to the coassociativity.

3 Hopf algebras deformations

The Hopf algebras we need to understand, are deformations of group algebras. This section aims at explaining the technique we used to deform such algebras.

We start to explain what we mean by deforming an algebra, since it is a more familiar structure.

Let A be an algebra over a field K. We want to define a new operation on A for every $a, b \in A$:

$$a * b = \alpha(a, b)ab$$

where $\alpha \colon A \times A \to K$ is a bilinear map. In particular we require that this new operation is a multiplication and it means that:

(1) the operation $*$ should be associative i.e. $(a * b) * c = a * (b * c)$ for every $a, b, c \in A$;

(2) the identity element for $*$ should be 1_A.

How can we translate these conditions into some restrictions on $\alpha(-,-)$?

(1) Let us compute $(a * b) * c$ for every $a, b, c \in A$:

$$(a * b) * c = (\alpha(a, b)ab) * c = \alpha(a.b)(ab) * c = \alpha(a, b)\alpha(ab, c)(ab)c$$

and in the same way:

$$a * (b * c) = a * (\alpha(b, c)bc) = \alpha(b, c)a * (bc) = \alpha(b, c)\alpha(a, bc)a(bc).$$

Since the original multiplication was associative, these two expressions are equal if and only if:

$$\alpha(a, b)\alpha(ab, c) = \alpha(b, c)\alpha(a, bc) \tag{3.1}$$

for every $a, b, c \in A$.

(2) Let us compute $a * 1_A$ for every $a \in A$:

$$a * 1_A = \alpha(a, 1_A)a1_A = \alpha(a, 1_A)a$$

and $1_a * a$:

$$1_A * a = \alpha(1_A, a)1_Aa = \alpha(1_A, a)a.$$

In order to have $a = a * 1_A = 1_A * a$, the following must hold for every $a \in A$:

$$\alpha(a, 1_A) = \alpha(1_A, a) = 1. \tag{3.2}$$

Definition 3.1 A bilinear map $\alpha\colon A \times A \to K$ satisfying (3.1) and (3.2) is said to be a *2-cocycle.*

In this way we have altered the multiplication structure of an algebra A; can we do the same for a Hopf Algebra?
The problem of deforming the algebra structure of a Hopf Algebra is the fact that we would like the compatibility with the coalgebra structure to hold also for the deformed algebra. Furthermore, we have also to verify if the antipode is again an antipode or if there exists another one.

Hopf algebras have a double nature: they are both algebras and coalgebras. Hence, we can deform the coalgebra structure.
For this purpose, given an initial Hopf algebra H, we need an invertible element $J \in H \otimes_K H$, such that:

$$(1_H \otimes J)(\mathrm{Id} \otimes \Delta)(J) = (J \otimes 1_H)(\Delta \otimes \mathrm{Id})(J); \tag{3.3}$$

$$(\varepsilon \otimes \mathrm{Id})(J) = (\mathrm{Id} \otimes \varepsilon)(J) = 1_H. \tag{3.4}$$

The element J is called *Twist* or *dual cocycle.*

The *Drinfeld Twist* of H is the new Hopf Algebra denoted by H_J and defined as:

- H as underlying vector space;

- As algebra structure $H_J = H$ and the same counit;
- New comultiplication: $\Delta_J(h) = J\Delta(h)J^{-1}$, for all $h \in H$;
- New antipode: $S_J(h) = u_J S(h) u_J^{-1}$, for all $h \in H$;

where u_J is a precise invertible element of H, constructed by means of J.
The conditions on J ensure us that H_J is really a Hopf algebra.

In our research work, we deformed group algebras using Movshev strategy for constructing a twist. For completeness, we briefly show how this technique works (more details can be found in [7]).

Let G be a finite group and consider M an abelian subgroup of G of central type, that is $M \simeq E \times E$ for some group E. Consider $\widehat{M}$ the group of characters of M, i.e. $\widehat{M} = \{\chi\colon M \to K^\times;\ \chi \text{ is a group homomorphism}\}$.
Let $\phi \in \widehat{M}$ and consider the idempotent, primitive central element associated to ϕ:

$$e_\phi = \frac{1}{|M|}\sum_{m\in M}\phi(m)m^{-1}.$$

To proceed with the construction, we need another definition.

Definition 3.2 Let M be a group. A *normalized 2-cocycle for M is a map $\alpha\colon M\times M \to K$, such that for every $m', m'', m''' \in M$:*

$$\alpha(m', m'')\alpha(m'm'', m''') = \alpha(m'', m''')\alpha(m', m''m'''), \tag{3.5}$$

$$\alpha(m', 1_M) = \alpha(1_M, m') = 1. \tag{3.6}$$

A normalized 2-cocycle gives rise to a cocycle on the group algebra KM. It is enough to extend it by bilinearity.
The idempotent elements and a normalized 2-cocycle allow us to define a twist.

Proposition 3.3 *Let M be an abelian group. Consider $\sigma\colon \widehat{M}\times\widehat{M} \to K^\times$ a normalized 2-cocycle on $\widehat{M}$. Then, J, defined as follows:*

$$J = \sum_{\psi,\phi}\sigma(\psi,\phi)e_\phi \otimes e_\psi \tag{3.7}$$

is a twist for KM and for every group algebra KG, where G contains M.

4 Hopf orders

In this section, we present the definition of a Hopf order over $\mathbb{Z}$ of a $\mathbb{Q}$-Hopf algebra.

Definition 4.1 Let H be a finite-dimensional Hopf algebra over $\mathbb{Q}$. A *Hopf order* over $\mathbb{Z}$ is a $\mathbb{Z}$-subalgebra X such that

- X is a free $\mathbb{Z}$-module, i.e. it is a $\mathbb{Z}$-module that has a basis;
- $X \otimes_{\mathbb{Z}} \mathbb{Q} \simeq H$;
- $\Delta(X) \subseteq X \otimes_{\mathbb{Z}} X$;
- $\varepsilon(X) \subseteq \mathbb{Z}$;
- $S(X) \subseteq X$

Remark 4.2 For simplicity, we restricted the definition of a Hopf order on a familiar Dedekind domain, i.e. $\mathbb{Z}$. In general, the above definition is given substituting $\mathbb{Z}$ with a Dedekind domain R, $\mathbb{Q}$ with the field of quotients of R and X with a finitely generated and projective R-module.
More interested readers can refer to [6].

Example Let G be a finite group. Consider $\mathbb{Q}G$ with the structure of a Hopf algebra described in Section 1. Then, $\mathbb{Z}G$ is a Hopf order of $\mathbb{Q}G$.

Example This example aims to show that in a group algebra, Hopf orders are not unique. Consider $G \simeq C_2 \times C_2$, that is $G = \langle \sigma, \tau \,|\, \sigma^2 = \tau^2 = 1_G,\ \sigma\tau = \tau\sigma \rangle$. Let $\mathbb{Q}G$ be its associated group algebra. Then, by the previous example $\mathbb{Z}G$ is a Hopf order. Moreover, also the $\mathbb{Z}$-algebra generated by $\{\frac{1+\tau}{2}, \sigma\}$ is a Hopf order, strictly containing $\mathbb{Z}G$.

5 Non-existence of Hopf orders

An initial motivation for understanding whether a Hopf algebra admits a Hopf order is related to Kaplansky's sixth conjecture. This conjecture is a generalization of Frobenius theorem which is one of the fundamental theorems in Representation theory for finite groups. Its statement is the following.

Frobenius Theorem. *The dimension of any complex irreducible representation of G divides the order of G.*

Since a representation of a group corresponds to a representation of its group algebra, then this theorem is inherited by the complex group algebra and its representations. For this reason, Kaplansky wondered if the same theorem could hold for any complex finite-dimensional semisimple Hopf algebra.

Kaplansky's sixth conjecture. *Let H be a complex finite-dimensional semisimple Hopf algebra. The dimension of every irreducible representation of H divides the dimension of H.*

An important contribution to this topic is due to Larson who proved a weaker version of this conjecture. Indeed, he proved that if the Hopf algebra admits a Hopf order, then the conjecture is satisfied. For completeness, we state his result in a formal way.

Larson Theorem. *Let H be a split semisimple finite-dimensional Hopf algebra over a number field K, i.e. H is isomorphic to a direct sum of matrix algebras with coefficients in K. Suppose that H admits a* ***Hopf order*** *X over a* ***number ring*** *$R = \mathcal{O}_K$. Then, the dimension of every irreducible representation of H divides* $\dim(H)$.

At this point a natural question arises: does every complex finite-dimensional semisimple Hopf algebra admit a Hopf order?

In [2] and [3] Cuadra and Meir presented some families of Hopf algebras which satisfy the conjecture but do not admit a Hopf order. These families were obtained deforming some specific group algebras, by means of a twist constructed with Movshev's strategy. As we have seen in Section 3, such a deformation does not alter the algebra structure: hence, as algebras they are group algebras. This implies that the conjecture is satisfied.
These examples show that the admission of Hopf orders is not a necessary condition for satisfying the conjecture. Moreover, they reveal an arithmetic difference between semisimple Hopf algebras and group algebras: group algebras can be defined over the integers, while in general Hopf algebras don't.

At this point, it became interesting to understand if the existence of Hopf orders for a Hopf algebra is a rare phenomenon.
In [1] we proved the non-existence result for some deformations of KG, where G is a finite non-abelian simple group. For being precise, we give here the statement of this result.

Theorem 5.1 *Let K be a number field and G a finite non-abelian simple group. Then, there is a twist Ω for KG, arising from a 2-cocycle on an abelian subgroup of G, such that $(KG)_\Omega$ does not admit a Hopf order over $\mathcal{O}_K$.*

On the other side, Cuadra and Meir in [4] gave some conditions for which a twisted group algebra admits a Hopf order. For more interested readers, we state here their result.

Theorem 5.2 *Let K be a number field with ring of integers R. Let G be a finite group and M an abelian subgroup of G of central type. Consider the twist $J \in KM \otimes KM$ afforded by a non-degenerate 2-cocycle $\omega\colon \widehat{M} \times \widehat{M} \to K^\times$.
Fix a Lagrangian decomposition $\widehat{M} \simeq L \times \widehat{L}$. Suppose that L is contained in a normal abelian subgroup N of G. Then $(KG)_J$ admits a Hopf order over R.*

References

[1] G. Carnovale, J. Cuadra and E. Masut, *Non-existence of integral Hopf orders for twists of several simple groups of Lie type.* Accepted in Publ. Mat. ArXiv:2108.12324.

[2] J. Cuadra and E. Meir, *On the existence of orders in semisimple Hopf algebras.* Trans. Amer. Math. Soc. 368 (2016), no. 4, 2547–2562.

[3] _______, *Non-existence of Hopf orders for a twist of the alternating and symmetric groups.* J. London Math. Soc. (2) 100 (2019), no. 1, 137–158.

[4] _______, *Existence of integral Hopf orders in twists of group algebras.* ArXiv, 2022, `https://arxiv.org/pdf/2211.00097.pdf`. Preprint.

[5] P. Etingof, S. Gelaki, D. Nikshych, and V. Ostrik, "Tensor Categories". Mathematical Surveys and Monographs, 205. American Mathematical Society, 2015.

[6] R.G. Larson, *Orders in Hopf algebras.* J. Algebra 22 (1972), 201–210.

[7] M. Movshev, *Twisting in group algebras of finite groups.* Funct. Anal. Appl. 27 (1994), 240–244.

[8] S. Montgomery, *Hopf algebras and their action on rings.* CBMS Regional Conference Series in Mathematics 82. Amer. Math. Soc., Providence, RI, 1993.