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

Scuole di Dottorato in Matematica Pura e Computazionale

Seminario Dottorato 2018/19

Preface

This document offers a large overview of the eight months' schedule of Seminario Dottorato 2018/19. Our "Seminario Dottorato" (Graduate Seminar) is a double-aimed activity. At one hand, the speakers (usually Ph.D. students or post-docs, but sometimes also senior researchers) are invited to think how to communicate their researches to a public of mathematically well-educated but not specialist people, by preserving both understandability and the flavour of a research report. At the same time, people in the audience enjoy a rare opportunity to get an accessible but also precise idea of what's going on in some mathematical research area that they might not know very well.
Let us take this opportunity to warmly thank the speakers once again, in particular for their nice agreement to write down these notes to leave a concrete footstep of their participation. We are also grateful to the collegues who helped us, through their advices and suggestions, in building an interesting and culturally complete program.

Padova, June 20th, 2019

Corrado Marastoni, Tiziano Vargiolu

Abstracts (from Seminario Dottorato's web page)

Wednesday 3 October 2018

Exact and meta-heuristic approach for Vehicle Routing Problems

Nicola GASTALDON (Padova, Dip. Mat. and Trans-Cel s.n.c., Albignasego PD)

The Vehicle Routing Problem (VRP) includes a wide class of problems studied in Operations Research and relevant from both theoretical and practical perspectives. In its basic formulation, the problem is to find a set of routes for a given fleet of vehicles through a set of locations, so that each location is visited by exactly one vehicle and the total travel cost is minimized. Such problem is often enriched with many attributes rising from real-world applications, such as capacity constraints, pickup and delivery operations, time windows, etc. VRP belongs to the class of combinatorial optimization problems, and it is very hard to solve efficiently and researchers have developed many exact and (meta-)heuristic algorithms. The former takes advantage of the structure of the mathematical model to obtain a speedup through decomposition methods. The latter exploits heuristic techniques to obtain solutions that trade off quality and computational burden, such as evolutionary algorithms and neighborhood search routines.
In our research, we consider the VRP arising at Trans-Cel, a freight transportation company based in Padova. We devised a Tabu Search heuristic implementing different neighborhood search policies, and now embedded in the tool supporting the operation manager at Trans-Cel. The algorithm runs in an acceptable amount of time both in static and dynamic settings, and the quality of the solutions is assessed through comparison with results obtained by a Column Generation algorithm that solves a mathematical programming formulation of the problem. Current research aims at developing data-driven techniques that exploit the information available from the company's repositories to support stochastic transportation demand arising in real time.

Wednesday 21 November 2018

Congruent numbers, Heegner method and BSD conjecture

Yan HU (Padova, Dip. Mat.)

The "Congruent number problem" is an old unsolved major problem in number theory. In this seminar we provide a brief introduction to it. We will start from the original version of the problem, and lots of objects will be introduced during the talk. If time permits, some current progresses relateted to the BSD conjecture will also be described.

Wednesday 5 December 2018

Regular domain perturbation problems

Paolo LUZZINI (Padova, Dip. Mat.)

The study of the dependence of functionals related to partial differential equations and of quantities of physical relevance upon smooth domain perturbations is a classical topic and has been carried out by several authors.
In this talk we will give an introductory overview about regular domain perturbation problems. We will provide concrete examples, highlight the motivations and the possible applications, and present an outline of some new results obtained in collaboration with P. Musolino and R. Pukhtaievych.

Wednesday 19 December 2018

Real Options: An overview

Dimitrios ZORMPAS (Padova, Dip. Mat.)

Financial options are contracts that derive their value from the performance of an underlying asset. They give to their holder the right, but not the obligation, to buy/sell an asset at a predetermined price and time. Contracts similar to options have been used since ancient times. However, the most basic model for their pricing was proposed in the early 1970's leading to a Nobel prize in 1997.
In the late 1970's the term Real Options is coined by Stewart Myers. According to the real options approach an investment characterized by uncertainty and irreversibility is like a financial option on a real asset. For instance, a potential investor has the right but not the obligation to pay a given amount of money in order to make an investment and gain access to the corresponding profit flow. Using standard option pricing tools one can also study the option to leave a market, outsource production, mothball a production plant etc.
In this seminar, I will refer to the correspondence between financial and real options and then present the simplest model in the real options literature that has to do with a potential investor who is considering undertaking an uncertain and irreversible investment. Then I will present a number of applications of the real options approach from the broad literature of operations management and finally make a reference to applications of the real options approach in energy economics.

Wednesday 30 January 2019

Conservation law models for supply chains

Maria Teresa CHIRI (Padova, Dip. Mat.)

Many real situations are modelled by nonlinear hyperbolic first order partial differential equations (PDEs) in the form of conservation or balance laws. Beside the classical case of Euler equations of gas dynamics, such PDEs arise for instance in traffic flow, gas pipelines, telecommunication networks, blood flow in arteries.
In this talk, after a short review on the basic theory of scalar conservation laws, we introduce a new model for supply chains. Here, we are considering large volume production that allows a continuous description of the product flow in terms of conservation laws, accompanied by ordinary differential equations describing the processing capacities. A key feature of this model is the behaviour of solutions in presence of a discontinuous dynamics with respect to the unknown conserved quantity (number of parts being processed).
This is a joint work with Prof. Fabio Ancona from University of Padova.

Wednesday 13 February 2019

Mean field interacting particle systems and games

Guglielmo PELINO (Padova, Dip. Mat.)

Mean field theory studies the behaviour of stochastic systems with a large number of interacting microscopic units. Under the mean-field hypothesis, it is often possible to give a macroscopic easier description of the phenomena, which still allows to catch the main characteristics of the complex pre-limit model. The main purpose of the talk is to motivate a system of two coupled forward-backward partial differential equations, known as the mean field game system, which serves as a limit model for a particular class of stochastic differential games with N players. For reaching this goal, an introductive overview on macroscopic limits for mean field interacting particle systems and games under diffusive dynamics will be presented. In the last part of the talk I will briefly review my contributions in the context of finite state mean field games.

Wednesday 27 February 2019

On the Alexander polynomial of line arrangements in $\mathbb{P}^2$

Federico VENTURELLI (Padova, Dip. Mat.)

The Alexander polynomial was first introduced in the context of knot theory, and it was used to study the local topology of plane curve singularities; this notion was later extended to projective hypersurfaces (zero loci of a single polynomial equation in a projective space), which is the case that

will be discussed in this talk. The Alexander polynomial of a hypersurface V encodes information on the monodromy eigenspaces of $H^1(F, \mathbb{C})$, where F is the Milnor fibre of V; while these eigenspaces are well understood for smooth hypersurfaces, they are significantly harder to compute if the hypersurface is singular, even in the simplest cases i.e. hyperplane arrangements.
In my talk I will try to give a basic introduction to this problem, explaining how the combinatorics of a hyperplane arrangement can help in determining its Alexander polynomial and presenting some known results; throughout the exposition some detours will be made, in order to discuss explicit examples and to introduce (or clarify) concepts that could be unfamiliar to non-specialists.

Wednesday 13 March 2019

An introduction to stochastic control in discrete time with an application to the securitization of systematic life insurance risk

Maren Diane SCHMECK (Univ. Bielefeld, Germany)

The basic idea behind insurance is to diversify risks. If a systematic risk is involved, this idea does not work well any more. So the idea arose to transfer the insurance risk to financial markets. Even though not perfectly linked to the own portfolio, these securitisation products work similarly to a reinsurance contract. For an investor, the products give a possibility to diversify an investment portfolio. Also insurers may act as investor and in this way diversify their own risk to regions where they have not underwritten contracts.
The literature on securitisation products considers either the point of view of an investor, or the product is used to perform a Markovitz optimisation. From the point of view of an insurer, this only partially answers the question how to choose a securitisation portfolio. We will here use utility theory and stochastic control in discrete time to determine the optimal portfolio. In order to simplify the presentation we consider the case of a mortality catastrophe bond. Similar consideration would also apply for other securitisation products.
The first part of the presentation will give an introduction to the methodology that we use in our research: stochastic control in discrete time. That is, we will look at the dynamic programming principle, also called Bellman?s equation and some results about the optimal strategy.

Wednesday 27 March 2019

Covers and envelopes of modules

Giovanna Giulia LE GROS (Padova, Dip. Mat.)

Approximation theory of modules is the study of left or right approximations of modules, also known as covers or envelopes, with respect to certain classes of modules. For a class C of R-modules, the aim is to characterise the rings over which every module has a C-cover or a C-envelope and furthermore to characterise the class C itself. For example, if one considers the

class of injective modules, then it is well-known that every module has an injective envelope (or injective hull). Instead, Bass proved that projective covers rarely exist and characterised the rings over which every module admits a projective cover, which are known as perfect rings. Moreover, precovers and preenvelopes are strongly related to the notion of a cotorsion pair, which is a pair of Ext-orthogonal classes in the category of R-modules.
The aim of this talk is to give a basic introduction to the theory of covers and envelopes, and to describe them with respect to some well-known classes of R-modules, along with a review of concepts in homological algebra that will be useful in this exposition.

Wednesday 10 April 2019

Probability and Information in Finance

Claudio FONTANA (Padova, Dip. Mat.)

In mathematical finance, tools from stochastic analysis are applied to the study of investment and valuation problems arising in financial markets. In this talk, we introduce some basic and fundamental concepts and results, with a focus on no-arbitrage properties and optimal investment problems. After a general overview, we will discuss the role of information and explore the interplay between information, arbitrage, and optimal investment.

Wednesday 8 May 2019

An introduction to Riemann-Hilbert correspondence

Davide BARCO (Padova, Dip. Mat.)

The 21st Hilbert problem concerns the existence of a certain class of linear differential equations on the complex affine line with specified singular points and monodromic groups. Arising both as an answer and an extension to this issue, Riemann-Hilbert correspondence aims to establish a relation between systems of linear differential equations defined on a complex manifold and suitable algebraic objects encoding topological properties of the same systems. The goal was first achieved for systems with regular singularities, thanks to the works by Deligne, Kashiwara and Mebkhout. Moreover, Deligne and Malgrange established a generalized correspondence (called Riemann-Hilbert-Birkhoff correspondence) for systems with irregular singularities on complex curves, encoding and describing the Stokes phenomenon which arises in this case. In more recent years, the correspondence has been extended to take account of irregular points on complex manifolds of any dimension by D'Agnolo and Kashiwara.
In this talk we give a basic introduction on the subject by providing concepts and classical example from the theory.

Wednesday 22 May 2019

Including topographic effects in shallow water modeling

Elena BACHINI (Padova, Dip. Mat.)

Shallow water equations are typically used to model fluid flows that develop predominantly along the horizontal (longitudinal and lateral) direction. Indeed, the so-called Shallow Water (SW) hypothesis assumes negligible vertical velocity components. The typical derivation of the SW equations is based on the integration of the Navier-Stokes equations over the fluid depth in combination with an asymptotic analysis enforcing the SW assumptions. In the presence of a general terrain, such as a mountain landscape, the model must be adapted to geometrical characteristics, since the bottom surface can be arbitrarily non-flat, with non-negligible slopes and curvatures.
After an introduction on the standard SW model, we will present a new formulation of the two-dimensional SW equations in intrinsic coordinates adapted to general and complex terrains, with emphasis on the influence of the geometry of the bottom on the solution. The proposed model is then discretized with a first order upwind Godunov Finite Volume scheme. We will give an overview of the numerical method and then show some results. The results indicate that it is important to take into full consideration the bottom geometry and slope even for relatively mild and slowly varying curvatures.

Wednesday 12 June 2019

Serre's p-adic modular forms and p-adic interpolation of the Riemann zeta function

Giacomo GRAZIANI (Padova, Dip. Mat.)

The so-called zeta functions are among the most famous and discussed objects in mathematics, the simplest of which is the (in)famous Riemann zeta function. In order to work with them (and with the strictly related L-functions as well), mathematicians decided to isolate simpler pieces and hence ultimately to address the problem of their p-adic interpolation. In this seminar, after introducing the various objects involved, we will focus on easiest example of the Riemann zeta function and describe the surprising interpolation exploited by Serre using his notion of p-adic modular forms.

Exact and meta-heuristic approach for Vehicle Routing Problems

Nicola Gastaldon (*)

Abstract. The Vehicle Routing Problem (VRP) includes a wide class of problems studied in Operations Research and relevant from both theoretical and practical perspectives. In its basic formulation, the problem is to find a set of routes for a given fleet of vehicles through a set of locations, so that each location is visited by exactly one vehicle and the total travel cost is minimized. Such problem is often enriched with many attributes rising from real-world applications, such as capacity constraints, pickup and delivery operations, time windows, etc. VRP belongs to the class of combinatorial optimization problems, and it is very hard to solve efficiently and researchers have developed many exact and (meta-)heuristic algorithms. The former takes advantage of the structure of the mathematical model to obtain a speedup through decomposition methods. The latter exploits heuristic techniques to obtain solutions that trade off quality and computational burden, such as evolutionary algorithms and neighborhood search routines. In our research, we consider the VRP arising at Trans-Cel, a freight transportation company based in Padova. We devised a Tabu Search heuristic implementing different neighborhood search policies, and now embedded in the tool supporting the operation manager at Trans-Cel. The algorithm runs in an acceptable amount of time both in static and dynamic settings, and the quality of the solutions is assessed through comparison with results obtained by a Column Generation algorithm that solves a mathematical programming formulation of the problem. Current research aims at developing data-driven techniques that exploit the information available from the company's repositories to support stochastic transportation demand arising in real time.

1 Introduction

The Vehicle Routing Problem refers to a wide class of problems studied in Operations Research and nowadays still provides challenges in several real-world applications. In its basic formulation, the problem is to find a set of routes given a fleet of vehicles and a set of locations so that each location are visited by exactly one vehicle and the total travel cost is minimized. Such problem is often enriched with many attributes rising from the real-world cases, such as the capacity constraints, pickup and delivery operations, time windows, etc.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; and Trans-Cel s.n.c., Albignasego PD. E-mail: gastald88@gmail.com . Seminar held on October 3rd, 2018.

This problem belongs to the class of combinatorial optimization problems, and it is very hard to solve efficiently. In order to solve such problem in an acceptable computational time, researchers have developed many exact and (meta-)heuristic algorithms. The former takes advantage of the structure of the mathematical model to obtain a speedup through a decomposition into sub-problems. The latter exploits heuristic techniques to obtain solutions close to the optimum, such as evolutionary algorithms and neighborhood search routines.

2 The Vehicle Routing Problem (VRP)

The Vehicle Routing Problem (VRP) considers a fleet of vehicles and a set of locations and requires a solution to be a set of routes that minimizes the overall travel cost while visiting exactly once each location (see [13]). Such problem arises both in freight transportation companies and in the public transport context. In real world applications multiple attributes have to be considered, for instance there may be different type of operations (pickup or delivery operations) to be satisfied at each location, capacity constraints on vehicles, precedence policies on the route design, soft and hard time windows within which any operation can be performed etc. In particular there are cases when the decision maker must take into account simultaneously several attributes, and such setting is known as the Multi-Attribute Vehicle Routing Problem (MAVRP) [1].
In the Operations Reserch literature devoted to MAVRPs several exact approaches were studied, suitable for many variants of the VRP. In [2, 3] a set partition formulation is solved by cut-and-column generation algorithm. A real-world VRP involving several attributes (multiple capacities, hours-of-service regulations, open routes, split-delivery, client-vehicle compatibility constraints etc.) is solved in [5] through a column generation approach and a bounded bidirectional dynamic programming algorithm for the pricing problem.
In real-world applications often the operation manager needs to take decision in short time. On the other hand, exact approaches rarely meet the efficiency required, so many heuristics have been designed by the researchers. In [12] and [14] there is an overview of several flexible heuristics able to adapt to the VRP definitions arising in different settings, and handle a variety of objectives and side constraints. Meta-heuristic approaches, such as Tabu Search, Genetic Algorithms, Ant Colony Optimization etc., are very popular for solving MAVRPs [4]. The research presented in [15] hybridizes genetic algorithms and local search, and proposes a unified framework for solving a wide range of large-scale vehicle routing problems with time windows, route-duration constraints and further attributes related to client assignment.
We devised solving methods for the MAVRP on the base of Trans-Cel, a freight transportation company in Padova, dealing with long-medium hauls (north and middle Italy) where several attributes must be considered in the planning phase.
Trans-Cel offers a just-in-time service to its customers and adopts the *groupage* technique, meaning that different type of freight and/or packaging can be loaded in the same vehicle. The orders consist of multiple pickup and delivery operations, and all the pickup operations must be satisfied before any delivery operation. A priority is associated to orders, so that they can be regular, urgent or mandatory. A date for the operation to take place is

provided, as well as a time window that can be soft (i.e: delays w.r.t. the time window end are penalized in the objective function), or hard (i.e: delays w.r.t. the time window end are not feasible). Further data related to the order are the revenue obtained, the duration of the operations and possible loading facilities required (i.e: a tail lift).
Trans-Cel has a heterogeneous fleet, whose vehicles have their own volume and weight capacity. They also have a cost per distance unit and a fixed cost for deployment. A vehicle may or may not be equipped with particular loading facilities (i.e: tail lift), and it has a time window of availability during a day.
The routes are daily routes, they are open (they may or may not start and end at the depot) and they may start with some pending orders from the day before and finish with some pending orders for the day after. Drivers are subject to the European hours of service Regulation, which states that a break of 45 minutes is required after 4:45 hours of cumulated driving time and a night break of 9 hours is required after either 9 hours of cumulated driving time or 13 hours of cumulated working time (driving time + loading/unloading time).

3 Explicit and Implicit Formulation

The VRP can be modeled through a complete graph $G = (N, A)$, where N is the set of nodes and A the set of arcs. The node 0 corresponds to the depot, while the nodes $1, \dots, |N|$ represent customer locations. A distance parameter c_{ij} is associated with each arc $(i, j) \in A$ between node i and j. We set V as the set of available vehicles and we define a set of binary variables x_{ij}^k assuming value 1 if vehicle k traverses the arc (i, j), 0 otherwise. The VRP can be modeled by the following **Arc-Flow** formulation with 3 indices (see [13]):

$$\min \sum_{k \in V} \sum_{(i,j) \in A} c_{ij} x_{ij}^k \tag{1}$$

$$s.t. \sum_{k \in V} \sum_{i \in \delta^-(j)} x_{ij}^k = 1 \qquad \forall j \in N \setminus \{0\} \tag{2}$$

$$\sum_{k \in V} \sum_{j \in \delta^+(i)} x_{ij}^k = 1 \qquad \forall i \in N \setminus \{0\} \tag{3}$$

$$\sum_{k \in V} \sum_{j \in \delta^+(0)} x_{0j}^k \leq |V| \tag{4}$$

$$\sum_{j \in \delta^+(S)} x_{ij}^k \geq 1 \qquad \forall S \subseteq N, S \neq \emptyset, k \in V \tag{5}$$

$$x_{ij}^k \in \{0, 1\} \qquad \forall (i, j) \in A, k \in V \tag{6}$$

We observe that the objective function (1) is to minimize the overall cost of the solution. We have the flow conservation constraints (2) and (3) so that each node is visited by exactly one vehicle. Constraint (4) prevent the solution to use more vehicles than the size of the fleet. Constraints (5) are known as the *sub-tour elimination constraints*, that force

any tour in the solution to contain the depot node. This set of constraints is exponentially large, so solving methods must be devised to tackle such a complexity.

The model can be enriched with attributes featuring in variants of the classic VRP. For instance we can make the model take into account the capacity constraints on vehicles by defining the variables u_i^k representing the cumulated load on board of vehicle k after the operation at node i and new parameters: q_i which is the amount of load to pick up at node i, Q is the capacity of a vehicle.
We add the constraints

$$u_i^k - u_j^k + Qx_{ij}^k \leq Q - q_j \qquad \forall (i,j) \in A, k \in V$$

that substitutes sub-tour elimination constraints, and the capacity constraint

$$q_i \leq u_i^k \leq Q \qquad \forall i \in N, k \in V.$$

We observe that the model is composed of *difficult* (binding) constraints and *easy* (non-binding) constraints, for example:

$$q_i \leq u_i^k \leq Q \qquad \forall i \in N, \mathbf{k} \in \mathbf{V} \qquad \text{non binding}$$

$$\sum_{\mathbf{k} \in \mathbf{V}} \sum_{i \in \delta^-(j)} x_{ij}^k = 1 \; \forall j \in N \setminus \{0\} \qquad \text{binding}$$

In the constraint matrix, non-binding constraints are represented by diagonal blocks. The goal is to find a way to exploit such structure in order to devise a solving algorithm with good performance.

The VRP can be formulated as a **Set-Partitioning** problem ([13]):

$$\min \sum_{r \in \Omega} c_r x_r \tag{7}$$

$$s.t. \sum_{r \in \Omega} a_{ir} x_r = 1 \qquad \forall i \in N \setminus \{0\} \tag{8}$$

$$\sum_{r \in \Omega} x_r \leq |V| \tag{9}$$

$$x_r \in \{0,1\} \qquad \forall r \in \Omega \tag{10}$$

Where Ω is the set of *all* feasible routes, c_r is the cost of route $r \in \Omega$, a_{ir} is 1 if node i is visited by route r, 0 otherwise and x_r is 1 if route r is part of the solution, 0 otherwise. The objective function (7) represent the total cost of the solution, constraints (8) ensure that each node is covered by exactly one route and constraint (9) forces the number of routes to be less or equal to the number of vehicles available.
Let us observe that we are not able to directly solve such a Linear Problem as is, due to the (exponentially) large number of variables. We need to make use of suitable algorithms to deal with such a problem formulation.

4 Column Generation Algorithm

4.1 General

The Column Generation algorithm (CG) takes advantage of a decomposition of the original problem (called the **Master Problem (MP)**)into two sub-problems to obtain an efficient solving method:

- the *Reduced Master Problem (RMP)* is defined as the MP but on a restricted set of variables $\bar{\Omega} \subset \Omega$ in such a way that there exist at least a feasible solution;
- the *Slave Problem(SP)* is a problem whose solution is a variable in $\Omega \setminus \bar{\Omega}$.

We give the following remark on what can guarantee optimality of a solution for a Linear Programming problem:

Remark 1 Let x^* be a feasible basic solution to a linear problem $P = \min\{c^T x : Ax = b,\ x \geq 0\}$. If all variables out of base x_j^F have reduced cost $\bar{c}_j \geq 0$, then x^* is optimal.

Therefore if the SP builds a path r (variables in MP) with $\bar{c}_r < 0$, we can add it to RMP and solve it (e.g. with the simplex algorithm), else the solution to the current RMP is also optimal for the original problem defined on the entire Ω.
Moreover observe that since variables in the MP correspond to paths in the graph G of our VRP model, the *non-binding* constraints are handled only in the SP, whereas *binding* constraints are handled only in the RMP.

By the duality theory in linear programming, we know that given the dual solution u^* of a linear programming problem, the following relation holds:

$$\bar{c}_r = c_r - (u^*)^T A_r,$$

where c_r is the cost of path r in the objective function and A_r is the column related to path r in the matrix of constraints .
For each arc (i, j) we can define the reduced cost on the arc $\bar{c}_{ij} := c_{ij} - u_j$, so we can compute:

$$\bar{c}_r = c_r - (u^*)^T A_r = \sum_{(i,j)\in r} c_{ij} - \sum_{(i,j)\in r} u_j = \sum_{(i,j)\in r} (c_{ij} - u_j),$$

then

$$\bar{c}_r = \sum_{(i,j)\in r} \bar{c}_{ij}.$$

Observe that:

- finding reduced cost variables is equivalent to solving a *Shortest-Path Problem (SPP)*;
- solving a SPP on a graph $\bar{G}$ equal to G except that the costs on each arc (i, j) is set as $\bar{c}_{ij}$ provides a feasible route with the minimum reduced cost for the MP
- if the SPP solution has a non-negative value, we know that the optimal solution to RMP is also *optimal to MP*

- there exist efficient *dynamic programming* algorithms for the solution of SPP (e.g. Bellman-Ford algorithm)

We can sum up the overall **CG algorithm** by the iteration of the procedure below:

(a) We solve the RMP. We obtain x^* and u^*;

(b) We compute $\bar{c}_{ij}$ and solve SP;

(c) If $\bar{c}_r < 0$ we add r to RMP and restart, else x^* is optimal for MP .

Remark 2 The Column Generation algorithm provides a solution to the linear relaxation of the original problem (it is a bound to the optimum). A branching method need to be integrated in order to obtain the (integer) optimal solution (*Branch-and-Price*).

4.2 Trans-Cel Problem

In order to adapt the algorithm to Trans-Cel model, we need to take into account some observations. First, the objective function is to maximize profit ($\equiv$ minimize the sum of overall cost and missed revenues). This means that inconvenient orders can be rejected by the algorithm. there is a set of *Mandatory orders* O^M contained in the set of all orders O that cannot rejected by the algorithm. We define the following variables:

- y_r takes value 1 if route r is in the solution, 0 otherwise
- x_o takes value 1 if order o is rejected, 0 otherwise

We define $o(i) \in O$ as the order containing node $i \in N$, and q_o is the revenue of order o. We can state now the following formulation of the problem (see [11]):

$$\min \sum_{v \in V} \sum_{r \in \Omega^v} c_r y_r + \sum_{o \in O \setminus O^M} q_o x_o \tag{11}$$

$$\text{s.t.} \sum_{v \in V} \sum_{r \in \Omega^v} a_{ir} y_r = 1 \quad \forall\, i \in N : o(i) \in O^M \tag{12}$$

$$\sum_{v \in V} \sum_{r \in \Omega^v} a_{ir} y_r + x_{o(i)} = 1 \quad \forall\, i \in N : o(i) \in O \setminus O^M \tag{13}$$

$$\sum_{v \in V} \sum_{r \in \Omega^v} y_r \leq |V| \tag{14}$$

$$y_r \in \{0,1\} \quad \forall\, r \in \Omega^v, v \in V \tag{15}$$

$$x_o \in \{0,1\}, \quad \forall\, o \in O \setminus O^M \tag{16}$$

As observed before, we can use efficient dynamic programming algorithms to solve in pseudo-polynomial time the Shortest Path Problem. We mention two main algorithms, which are the **Bellman-Ford** algorithm, that have $O(|N||A|)$ complexity, and the **Dijkstra** algorithm, with $O(|N|\log(|N|))$ complexity has the requirement that each cost of the graph must be non-negative.

Definition 1 (Label) A *label* π_i for SPP from node s to node d is the cost of the shortest path from node s to node i.

Since we are dealing with a **Multi-Attribute VRP** we cannot use either of these algorithms, because *not all paths are feasible*. This type of problem is called **Resource-Constraint Elementary SPP (RCESPP)**, where *elementary* means that a path does not visit a node more than once and *resources* carry data useful to check feasibility at each node in a path (i.e. cumulated driving time). The idea is to generalize the concept behinds Bellman-Ford algorithm taking into account (not binding) *constraints on routes* (i.e. Capacity, Time Windows, ...) and evolution of *resources*. In this context, the definition of label is generalized by adding also the resources values to it:

$$\lambda_i = (\rho_i^1, \dots, \rho_i^r, \pi_i) \qquad \rho_i^m \text{ with } m = 1, \dots, r \text{ resources.}$$

The algorithm solving this problem is based on starting from an initial empty label at the source node and extend it through the reachable nodes in the graph till it is no more possible to build a path that may be an optimal solution to RCESPP.

Due to the large number of labels that may rise during the algorithm iterations, the time of convergence may increase dramatically. This is why there are two main sub-routine of the algorithm that must be implemented wisely:

- a **Dominance Rule** should be implemented as a criterion to discard labels that certainly will not take to an optimal path. We introduce the following notation: $(\rho_i^m)_{\lambda_i}$ represents the m-th resource of the label λ_i related to the node i. We say that for a node i a label λ_i *dominates* the label μ_i if for all $m = 1, \dots, r$ we have $(\rho_i^m)_{\lambda_i} > (\rho_i^m)_{\mu_i}$ and $(\pi_i)_{\mu_i} > (\pi_i)_{\lambda_i}$ ($\equiv$ *the path in* λ_i *is shorter and the set of all feasible extensions of* λ_i *contains the set of all feasible extensions of* μ_i).
- the **Extension of a Label** must be performed incrementally (this is the essence of dynamic programming).

Example 1 Let us consider a Maximum-Duration VRP, where we have a constraint on the maximum duration of a route.
$A = 10$ is the max duration of a route;
t_{ij} = travel time from node i to node j. c_{ij} = distance from node i to node j. $\lambda_i = (\tau_i, \pi_i)$, with τ_i as cumulated time elapsed from s to i.
The extension from i to j generates the label $\lambda_j = (\tau_j, \pi_j)$, by data of label λ_i:

$$\tau_j = \tau_i + t_{ij}$$

$$\pi_j = \pi_i + c_{ij}$$

The dominance rule imposes that both τ_j and π_j are less or equal to the dominated label.

$$(7, 2) \text{ dominates } (8, 5)$$

$$(9, 3) \text{ does not dominate } (8, 5)$$

If we discarded $(8, 5)$ keeping $(9, 3)$, we would miss all extensions of travel time 2 ($(9, 3)$ would break the time constraint), which may contain the optimal solution.

The steps of the algorithm can be summarized as below:

(a) **Initialize:** set initial label at source node s

(b) **Iterate:** for each node extend all labels related to it. Eliminate infeasible labels and dominated labels

(c) **Stop:** when all labels has been processed (no further extensions available)

(d) **return** label at d with smaller cost

In the case of Trans-Cel we are dealing with many attributes, so there are more complex domination rule and label extension function.
A label in such problem is defined as [11]:

$$L = (i, t^w, t^d, w, v, \mathcal{O}, \mathcal{U}, \pi_\rangle)$$

- i the last node of the path;
- t^w and t^d cumulative working and driving time;
- w and v volume and weight on board after visiting i;
- $\mathcal{O}$ is the set of "open orders";
- $\mathcal{U}$ is the set of *unreachable nodes*;
- π_i the path cost.

Further details on label extension and dominance rule can be found in [11].

5 Meta-Heuristic Approach

Exact Methods are often time-consuming when dealing with large-scale problems. Real world applications need fast answers to events. To this end, heuristic approaches have been deeply studied by researchers.
We can identify two main types of heuristics:

- *Constructive:* build a solution from scratch (e.g. Greedy)
- *Improving:* refine a given solution (e.g. Local Search)

The heuristics algorithm are designed in such a way that they provide a feasible solution of high quality in short time. Nevertheless they cannot guarantee the optimality.
Moreover many of these algorithms are set up in the solution space, which usually have a nasty topology (in combinatorial optimization the solution space is the discrete set of

all possible combinations of values assigned to decision variables), so analytical tools are not available (i.e: the gradient). For instance there is no definition for open or close sets, so a *neighborhood* of a solution cannot be defined through open sets but by a set of perturbations (*moves*).
The Local Search algorithm as is is based on the concept of exploring a neighborhood and updating the current solution as long as an improving one is found. As soon as no improving solutions are found during the exploration the algorithm stops. This means that such an algorithm risks to get stuck in local minima. Algorithms called **Meta-Heuristics** have been designed so that they can implement techniques to escape such local minima. Meta-Heuristic algorithms are mainly divided into three approaches:

- *Trajectory-Based* algorithms (i.e: Tabu Search, Simulated Annealing etc.);
- *Population-Based* algorithms (i.e: Ant Colony, Genetic Algorithm etc.);
- *Hybrid* algorithms of the two types.

We can say that Trajectory-Based algorithms rely on *exploration* of neighborhoods, whereas Population-Based algorithms exploit *sampling* and combining solutions at each iteration.
We now consider the **Tabu Search (TS)** algorithm, describing its main features.
The TS algorithm performs an exploration of one or more neighborhoods similarly to the Local Search, but as soon as no improving solution is found, it updates the current solution with the best one found even if worsening. This acceptance criterion takes place only for a predefined number of iterations. Moreover, in order to prevent cycling on solution already explored, moves are stored in a First-In-First-Out list so that an inverse move is not accepted during the exploration phase (is made *Tabu*).
We devised several tabu-search based heuristics fro the MAVRP arising in Trans-Cel, as described in [8], [6], [9], [11], [10] and [7]: in the following we summarize their main features.

We consider, among others, three different neighborhoods:

- **1-Order Relocation (1R)** : move one order from one route to another;
- **2-Order Swap (2S)** : swap two order in two different routes;
- **2-Order Relocation (2R)** : move two order from one route to another.

The TS performs an exploration of the neighborhoods above in a *Variable Neighborhood Descent* fashion. This means that we explore hierarchically each neighborhoods in such a way that when no improving solution is found in one neighborhood, we start exploring the next one. Whenever an improving solution is found, we start over the routine. If no improving solution is found in any of the neighborhoods, we update the current solution according to one of the following criteria:

(a) we choose the best solution found in all the neighborhoods (*TS DET*);

(b) we choose a random solution among the best 5 found in all the neighborhoods (*TS STOCH*);

The second criterion is useful for a diversification phase in the search procedure. Our full solving algorithm (that we denote by *RND*) runs a TS DET and 3 TS STOCH.

We also set up different enhancements for such algorithm in order to increase the efficiency:

- we **relaxed** some of the problem constraints on specific sub-routes of the solution to obtain a speed-up, then a **destroy-and-repair** phase is triggered to fix infeasibility. This relies on the fact that such infeasibility are rarely met;
- we set up a **granular** exploration of 2R. This means that we filter moves such that the pair of orders we want to relocate are "far" from each other given a metric in the order space;
- we designed a **parallel** exploration of each neighborhood through a decomposition of the moves. Each thread explores one of the sub-set of moves.

6 Computational Results

The algorithm was implemented in C++ and tests run on an Intel Core i5-5200 2.20 GHz CPU with 8 GB RAM. We made our experiments on real instances collected in Trans-Cel operations' office, and we partitioned the set of instances into different groups based on the number of nodes in the instance.

In Table 1, *CG bound* is the percentage of CG run that converged within a time limit of 1 hour, *CG opt* is the percentage of CG run that returned an integer optimal solution, *RND opt* is the percentage of instances where RND found the optimal solution, *RND gap* is the average, min and max values of the gap between RND sub-optimal solution and the CG bounds.

Group	CG bound (%)	CG opt (%)	RND opt (%)	RND gap (%)
0-40	100.0	77.8	66.7	0.8 (0.4 ; 1.1)
41-80	77.8	66.7	22.2	1.4 (0.4 ; 3.1)
81-90	90.0	60.0	0.0	0.8 (0.1 ; 1.4)
91-100	100.0	77.8	11.1	0.8 (0.2 ; 2.4)
101-116	66.7	33.3	0.0	0.6 (0.0 ; 0.9)
all	*88.4*	*65.1*	*20.9*	*0.9 (0.0 ; 3.1)*

Table 1. Results on real instances: optimality gap.

Notice that the optimality gap of RND is on average 0.6%, with worst case 3.1%.

In Table 2, *BI vs RND* is the gap between the initial solution found through a Best Insertion heuristic (BI) and RND, *DET vs RND* is the gap between TS DET and RND, and the remainder columns show the execution time in seconds for each mentioned routine.

Group	BI vs RND (%)	DET vs RND (%)	BI (s)	DET (s)	RND (s)
0-40	2.8 (0.0 ; 13.3)	0.1 (0.0 ; 0.8)	0.0 (0.0 ; 0.0)	0.3 (0.0 ; 2.4)	1.4 (0.0 ; 9.7)
41-80	5.6 (0.0 ; 14.6)	0.0 (0.0 ; 0.4)	0.1 (0.0 ; 0.3)	3.2 (0.0 ; 8.5)	12.9 (0.0 ; 32.3)
81-90	6.2 (0.2 ; 10.8)	0.7 (0.0 ; 3.3)	0.2 (0.1 ; 0.3)	5.4 (2.4 ; 12.0)	23.7 (14.1 ; 40.1)
91-100	8.8 (3.0 ; 14.1)	0.3 (0.0 ; 1.7)	0.4 (0.2 ; 0.8)	9.0 (3.8 ; 20.1)	40.3 (18.5 ; 79.9)
101-116	8.2 (3.5 ; 14.8)	0.2 (0.0 ; 1.0)	1.1 (0.3 ; 4.3)	19.1 (4.4 ; 56.7)	77.7 (22.9 ; 196.8)
all	*6.2 (0.0 ; 14.8)*	*0.3 (0.0 ; 3.3)*	*0.3 (0.0 ; 4.3)*	*6.5 (0.0 ; 56.7)*	*27.8 (0.0 ; 196.8)*

Table 2. Results on real instances: basic algorithms.

Observe that the RND algorithm obtains better results than the sole DET routine up to 3.3% improvement.

In the end, Tables 3 and 4 show the difference respectively in terms of objective gap and run time of RND, RND with granular filter (RND + F) and RND with granular filter and parallelized exploration (RND + F + 4P).

Group	RND	RND+F	RND+F+4P
0-40	-	0.0 (0.0;0.0)	0.0 (0.0 ; 0.4)
41-80	-	0.0 (0.0;0.4)	0.0 (-0.7 ; 0.4)
81-90	-	0.1 (-1.9;0.8)	0.2 (0.0 ; 0.8)
91-100	-	0.6 (-0.9;3.1)	0.8 (-0.9 ; 3.9)
101-116	-	0.0 (-0.2;0.2)	0.3 (-0.2 ; 1.0)
all	-	0.1 (-1.9;3.1)	0.3 (-0.9 ; 3.9)

Table 3. Results on real instances: speed-up effects. (gap towards RND).

Group	RND	RND+F	RND+F+4P
0-40	-	1.5 (0.0;9.8)	1.0 (0.0;6.1)
41-80	-	14.9 (0.0;43.5)	8.6 (0.0;19.6)
81-90	-	24.2 (12.4;44.5)	15.4 (6.5;29.7)
91-100	-	34.5 (16.7;67.9)	18.5 (8.1;33.3)
101-116	-	66.3 (23.3;96.0)	33.9 (11.8;48.6)
all	-	25.5 (0.0;96.0)	14.2 (0.0;48.6)

Table 4. Results on real instances: speed-up effects. (running times).

We observe that at the cost of negligible changes in the objective function we obtain a large increasing in efficiency thanks to filtering and parallel design.

References

[1] P. Arias, J. Caceres-Cruz, D. Guimarans, and A.A. Juan, *Rich vehicle routing problem: Survey.* ACM Computing Survey 2 (2014), :229–268.

[2] R. Baldacci, E. Bartolini, A. Mingozzi, and R. Roberti, *An exact solution framework for a broad class of vehicle routing problems.* Computational Management Science, 7/3 (2010), 229–268.

[3] R. Baldacci and A. Mingozzi, *A unified exact method for solving different classes of vehicle routing problems.* Mathematical Programming 120/2 (2008), 347–380.

[4] O. Bräysy, M. Gendreau, G. Hasle, A. Lokketangen, and J.Y. Potvin, *Metaheuristics for the vehicle routing problem and its extensions: a categorized bibliography.* In B. Golden, S. Ranghavan, and E. Wasil editors, "The vehicle routing problem: latest and new challanges", pp. 143–169. Springer, New York, USA, 2008.

[5] A. Ceselli, G. Righini, and M. Salani, *A column generation algorithm for a rich vehicle-routing problem.* Transportation Science 43/1 (2009), 56–69.

[6] L. De Giovanni, N. Gastaldon, I. Lauriola, and F. Sottovia, *A heuristic for multiattribute vehicle routing problems in express freight transportation.* In A. Sforza and C. Sterle, editors, "Optimization and Decision Science: Methodologies and Applications ODS 2017". Springer Proceedings in Mathematics & Statistics, volume 217, pp. 161–169. Springer International Publishing, Cham, 2017.

[7] L. De Giovanni, N. Gastaldon, M. Losego, and F. Sottovia, *Algorithms for a vehicle routing tool supporting express freight delivery in small trucking companies.* Accepted for publication on Transportation Research Proceedia, 2018.

[8] L. De Giovanni, N. Gastaldon, and F. Sottovia, *A rich vehicle routing problem in express freight transportation.* In "Book of Abstract", p. 33, 2016.

[9] L. De Giovanni, N. Gastaldon, and F. Sottovia, *A rich vehicle routing problem in express freight transportation.* Network Optimization Workshop - NOW 2017 Conference, p. 33, 2017.

[10] L. De Giovanni, N. Gastaldon, and F. Sottovia, *Express delivery in freight trans- portation: an application to the trucking industry.* In "Book of Extended Abstracts", pp. 570–573, 2018. Odysseus 2018 - Seventh International Workshop on Freight Transportation and Logistics.

[11] L. De Giovanni, N. Gastaldon, and F. Sottovia, *A two-level local search heuristic for pickup and delivery problems in express freight trucking.* Submitted to Networks, 2018.

[12] G. Laporte, S. Ropke, and T. Vidal, *Heuristics for the vehicle routing problem.* In P. Toth and D. Vigo editors, "Vehicle Routing: Problems, Methods, and Applications". MOS-SIAM Series on Optimization, pp. 87–116. 2014.

[13] P. Toth and D. Vigo, "An overview of vehicle routing problems". In P. Toth and D. Vigo editors, "The Vehicle Routing Problem". SIAM - Society for Industrial and Applied Mathematics, Philadelphia, USA, 2002.

[14] T. Vidal, T.G. Crainic, M. Gendreau, and C. Prins, *Heuristics for multi-attribute vehicle routing problems: A survey and synthesis.* European Journal of Operational Research 231/1 (2013), 1–21.

[15] T. Vidal, T.G. Crainic, M. Gendreau, and C. Prins, *A hybrid genetic algorithm with adaptive diversity management for a large class of vehicle routing problems with time-windows.* Computers & Operations Research 40/1 (2013), 475–489.

Congruent numbers, Heegner method and BSD conjecture

Yan Hu (*)

Abstract. The congruent number problem is an old unsolved major problem in number theory. In this note we will give a brief introduction to the topic. We will start from the original version of the problem, and lots of mathematical objects will be introduced. The progress of study of the congruent number problem and other related topics such as BSD conjecture will also be presented.

1 Introduction

Number Theory is one of the oldest branches in mathematics which is the study of properties of the positive integers. In this topic, a great many of its problems are simple to state yet very difficult to solve. The congruent number problem, the written history of which can be traced back at least a millennium, is the oldest unsolved major problem in number theory.

The original version of congruent number problem is written in an Arab manuscript [2] before 972 as follows:

Question 1.1 *(Congruent number problem(Original version) Given an integer n, find a (rational) square γ^2 such that $\gamma^2 \pm n$ are both (rational) square.*

We will talk about the progress of the congruent number problem and explain the connection among the congruent number problem with other open questions in number theory such as BSD conjecture.

2 Congruent numbers

We first want to explain what is a "congruent number".

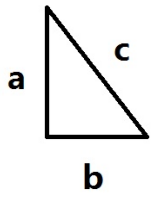
(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: sthuyan@gmail.com . Seminar held on November 21st, 2018.

Definition 2.1 A positive integer n is called a congruent number if there exist positive rational numbers a, b, c such that

$$a^2 + b^2 = c^2, \quad n = \frac{1}{2}ab.$$

We also give the equivalence form of congruent number as follows:

Definition 2.2 (Triangular version) A positive integer n is called a congruent number if it is the area of a right-angled triangle, all of whose sides have rational length.



,

Example 2.3 24 is a congruent number:

$$6^2 + 8^2 = 10^2,\ 24 = \frac{1}{2} \times 6 \times 8$$

So is 6:

$$3^2 + 4^2 = 5^2,\ 6 = \frac{1}{2} \times 3 \times 4$$

It is clear that we can assume n has no square factors.(Such n is called square-free integer)

Conjecture 2.4 *(Fibonacci) 1 is not a congruent number.*

It took 400 year until it was proved by Fermat using the method called infinite descent.

3 Congruent number problem

Now we can introduce the congruent number problem.

Question 3.1 *(Congruent number problem) Given an integer n, determining it is a congruent number or not.*

Some modern theory of number theory grew out of the study of this problem. For example, in the 17th century, Fermat gave a wonderful proof of the first special case of this problem. Fermat noted that his proof that 1 is not a congruent number also implies that there are no rational numbers x and y with $xy \neq 0$ such that $x^4 + y^4 = 1$. It also led Fermat to his so called Last Theorem (now solved by Andrew Wiles). Now let us talk a bit about Fermat's method.

3.1 Fermat's method

Theorem 3.2 (Fermat) *1,2,3 are non-congruent.*

Fermat's idea was based on the ancient Euclidean formula:

Lemma 3.3 (Euclidean formula(300 BC)) *Given (a,b,c) positive integers, pairwise co-prime and $a^2+b^2=c^2$. Then there is a pair of coprime positive integers (p,q) with $p+q$ odd, such that*

$$a=2pq,\ b=p^2-q^2, c=p^2+q^2.$$

Thus we have a congruent number generating formula:

$$n=pq(p+q)(p-q)/\square$$

Now we can give the sketch of the proof of Theorem 3.2.

Proof. [6]

1. Suppose 1 is congruent. Then is an integral right triangle with minimum area: $\square = pq(p+q)(p-q)$.
2. As all 4 factors are co-prime,
$$p=x^2, q=y^2, p+q=u^2, p-q=v^2.$$
3. Thus we have an equation with the solution as follows:
$$(u+v)^2+(u-v)^2=(2x)^2.$$
4. Then $(u+v, u-v, 2x)$ forms a right triangle and with a smaller area y^2. Contradiction!

Fermat called the method the infinite descent. □

The following example shows that it is also very difficult to compute a precise triangle when you already know the corresponding area.

Example 3.4 Zagier has computed a precise triangle with area 157:

$$157=\frac{1}{2}ab,\ a^2+b^2=c^2.$$

$a=\frac{411340519227716149383203}{2166655693714761309610}$,

$b=\frac{680329848782643505121754 0}{411340519227716149383203}$,

$c=\frac{224403517704336969924557513090674863160948472041}{8912332268928859588025535178967163570016480830}$.

Now we want to think about the congruent number problem in another way which is related to the arithmetic of elliptic curve.

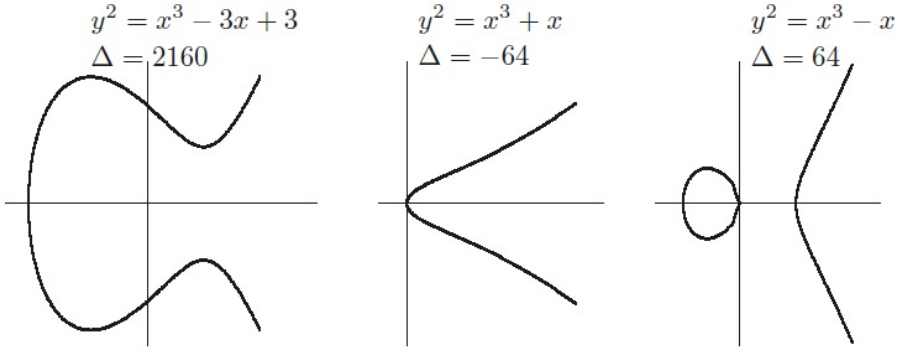
4 Congruent number problem revisited(Elliptic curve version)

An elliptic curve $E/\mathbb{Q}$ is given by:

$$E : y^2 = x^3 + ax + b,\ a, b \in \mathbb{Q};$$

where $\Delta := -16(4a^3 + 27b^2) \neq 0$.

The following figures show some types of elliptic curves:



Now we can give a new form of congruent number problem:

Question 4.1 *(Congruent number problem(Elliptic curve version)) For a positive integer n, find a rational point (x, y) with $y \neq 0$ on the elliptic curve:*

$$E_n :\ ny^2 = x^3 - x$$

The equivalence with the original version is given by:

$$x = \frac{p}{q} \Leftrightarrow (a, b, c) = (2pq, p^2 - q^2, p^2 + q^2).$$

Now our research objects have changed. We can study the congruent number problem by studying the elliptic curve over $\mathbb{Q}$.

5 Elliptic curve $E/\mathbb{Q}$

Recall that an elliptic curve $E/\mathbb{Q}$ is given by:

$$E : y^2 = x^3 + ax + b,\ a, b \in \mathbb{Q};$$

where $\Delta := -16(4a^3 + 27b^2) \neq 0$.

Write

$$E(\mathbb{Q}) = \{(x, y) \in \mathbb{Q}^2 : y^2 = x^3 + ax + b\} \cup \{\infty\}.$$

Basic problem: Given an elliptic E, find all the rational points on the curve.

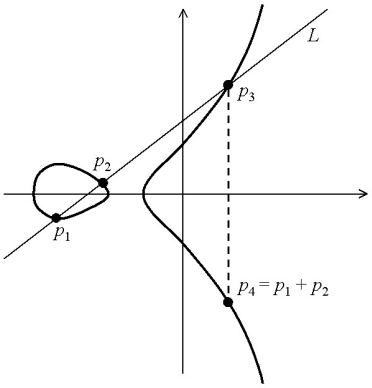
5.1 Addition law

Let $\mathcal{O} = \infty$ be the point "at infinity". We define an additional operation "+" on $E/\mathbb{Q}$ by the following law:

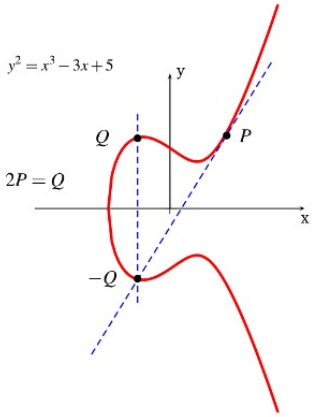
Proposition 5.1 *Let $P, Q \in E$, let L be the line through P and Q (if $P = Q$, let L be the tangent line to E at P), and let R be the third point of intersection of L with E. Let L' be the line through R and $\mathcal{O}$. Then L' intersects E at R, $\mathcal{O}$, and a third point. We denote that third point by $P + Q$.*

Proof. The proof and more details can be found in Chapter 3 of [4]. □

It is not so easy to understand what the addition law is only by the description above. Let us see the figures below which can help us understand it easily:



$P_1 + P_2 = P_4,\ P_3 = -P_4$



$Q = P + P = 2P$

The addition law on $E(\mathbb{Q})$ has the following properties:

Proposition 5.2
(a) $P + \mathcal{O} = \mathcal{O} + P = P$, *for all* $P \in E(\mathbb{Q})$.
(b) $P + (-P) = \mathcal{O}$, *for all* $P \in E(\mathbb{Q})$.
(c) $P + (Q + R) = (P + Q) + R$, *for all* $P, Q, R \in E(\mathbb{Q})$.
(d) $P + Q = Q + P$, *for all* $P, Q \in E(\mathbb{Q})$
In other words, under the addition $E(\mathbb{Q})$ is an abelian group with identity $\mathcal{O}$.

Proof. See the proof of Proposition 2.2 in [4]. □

Let us see a numerical example about the operation on the elliptic curve over $\mathbb{Q}$.

Example 5.3

$$E : y^2 = x^3 - 5x + 8.$$

The point $P = (1, 2)$ is on the curve $E(\mathbb{Q})$. Using the tangent line construction

$$2P = P + P = \left(-\frac{7}{4}, -\frac{27}{8}\right).$$

Let $Q = \left(-\frac{7}{4}, -\frac{27}{8}\right)$. Using the secant line construction, we find that

$$3P = P + Q = \left(\frac{553}{121}, -\frac{11950}{1331}\right).$$

5.2 Group structure of $E(\mathbb{Q})$

The understanding of the group structure of $E(\mathbb{Q})$ is a major question in modern number theory and arithmetic algebraic geometry. Thus we have the following theorem:

Theorem 5.6 (Mordell-Weil) *Let E be an elliptic curve over $\mathbb{Q}$. Then*

$$E(\mathbb{Q}) \simeq E(\mathbb{Q})_{tor} \oplus \mathbb{Z}^r$$

for some $r > 0$, where $E(\mathbb{Q})_{tor}$ is the finite torsion subgroup of $E(\mathbb{Q})$.

Proof. The proof and more details can be found in Chapter VIII of [4]. □

Remark 5.5

- The integer r is called the rank of $E(\mathbb{Q})$.
- The description of all possible $E(\mathbb{Q})_{tors}$ is clear:

Theorem 5.6 (Mazur) *There are exactly 15 possible finite groups for $E(\mathbb{Q})_{tors}$. In particular, $E(\mathbb{Q})_{tors}$ has order at most 16.*

Now we can study the congruent number problem in another point of view.

Question 5.7 *For a positive integer n, let E_n be the elliptic curve*

$$E_n : ny^2 = x^3 - x.$$

Then n is a congruent number if and only if $r = rank(E(\mathbb{Q})) > 0$. It means that there are infinitely many rational solutions (x, y) satisfying the equation of E_n.

Thus given an elliptic curve over $\mathbb{Q}$, determining the rank is one of the most important problems in the theory of elliptic curves.

5.3 L-series

In this section, we want to introduce that how we study the elliptic curve. In complex analysis, we know the famous Riemann zeta function:

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

where s is a complex variable.

Zeta function plays an very important role in number theory. Swinnerton-Dyer said: "the zeta function knows everything about number field, we just have to prevail on it to tell us". Thus we study the L-series which is the generalization of Riemann zeta function.

The L-series has very close connection to the elliptic curve. Before giving the definition of the L-series attached to the elliptic curve, we will first introduce some notations.

Let $E/\mathbb{Q}$ be an elliptic curve over $\mathbb{Q}$:

$$E/\mathbb{Q} : y^2 = x^3 + ax + b$$

$\Delta = -16(4a^3 + 27b^2) \neq 0$ discriminant of $E/\mathbb{Q}$.

$N_p = \sharp\{\texttt{solutions of } y^2 \equiv x^3 + ax + b \texttt{ mod } p\}$.

$a_p = p - N_p$.

For the a_p, we have the following theorem:

Theorem 5.8 (Hasse (1922)) *If $p \nmid \Delta$, then*

$$\mid a_p \mid \leq 2\sqrt{p}.$$

Remark 5.9

For each $x \pmod p$, there is a 50% chance that $y^2 \equiv x^3 + ax + b \pmod p$ has solution.

If $y^2 \equiv x^3 + ax + b \pmod p$ has solution, then we get two solutions (x, y) and $(x, -y)$.

Thus we might expect N_p is approximately

$$N_p \approx \frac{1}{2} \cdot 2 \cdot p = p.$$

Hence $\mid a_p \mid = \mid N_p - p \mid$ should be small compared with p.

Now we can define the L-series attached to $E/\mathbb{Q}$:

$$L(E,s) = \prod_{p \nmid 2\Delta} \left(1 - \frac{a_p}{p^s} + \frac{1}{p^{2s-1}}\right)^{-1}$$

s is a complex variable, $s \in \mathbb{C}$.

$L(E,s)$ is absolutely convergent for $Re(s) > \frac{3}{2}$. (Hasse)

$L(E,s)$ has holomorphic continuation to $\mathbb{C}$. (Wiles, et al.)

6 BSD Conjecture

We are now talking about a deep conjecture related to the congruent number problem. The conjecture originated from B. Birch and H.P.F. Swinnerton-Dyer [1] in the 60's.

Conjecture 6.1 (Birch and Swinnerton-Dyer) *The Taylor expansion of $L(E,s)$ at $s = 1$ has the form*

$$L(E,s) = c(s-1)^r + \textit{ higher order terms of } (s-1)$$

with $c \neq 0$ and $r =$ rank $E(\mathbb{Q})$. In particular $L(E,1) = 0$ if and only if $E(\mathbb{Q})$ is infinite.

Remark 6.1 For more details about the BSD conjecture, people can see [5].

6.1 Some results

BSD conjecture is still an open problem in the field of number theory and is widely recognized as one of the most challenging mathematical problems. The conjecture was chosen as one of the seven Millennium Prize Problems listed by the Clay Mathematics Institute, which has offered one million prize for the first correct proof. We only show two results here and then you will know how difficult to prove the conjecture.

Theorem 6.3 (Gross-Zagier, Kolyvagin) *The BSD conjecture is true if rank$(E(\mathbb{Q})) \leq 1$.*

Jerrold Tunnell made significant progress by connecting congruent numbers to elliptic curves. He showed that there is a formula for determining whether any positive number n is a congruent number or not, but the complete validity of his formula depends on the truth of BSD conjecture.

Theorem 6.4 (Tunnell (1983)) *Let n be an odd square-free positive integer. Consider the two conditions:*

(A) *n is a congruent number;*

(B) *the number of triples of integers (x,y,z) satisfying $2x^2+y^2+8z^2 = n$ is equal to twice the number of triples satisfying $2x^2 + y^2 + 32z^2 = n$.*

Then

- (A) *implies* (B)
- *If the BSD conjecture is true, then* (B) *implies* (A).

6.2 Application to congruent number problem

The L-series has a functional equation $s \Leftrightarrow 2-s$ with sign

$$\epsilon(n) = \begin{cases} 1 & n \equiv 1,2,3 \ mod\ 8 \\ -1 & n \equiv 5,6,7 \ mod\ 8 \end{cases}$$

This gives a partition $\mathbb{N} = S \sqcup T$ according to $\epsilon = \pm 1$, i.e, $\epsilon(n) = 1, n \in S$, otherwise $n \in T$.

Thus we have the following conjectures:

Conjecture 6.5 *100% of $n \in S$ are non-congruent numbers.*

Conjecture 6.6 *100% of $n \in T$ are congruent numbers.*

7 Heegner method

The first person to prove the existence of infinitely many square-free congruent numbers was Heegner. Heegner published his paper in 1952 as a 59 years old nonprofessional mathematican. He showed that every prime number p of the form $p = 8n + 5$ is a congruent number. In the same paper, Heegner solved Gauss's class number one problem. The importance of Heegner's paper was realized only in the late 1960s, after the discovery of the conjecture of Birch and Swinnerton-Dyer. We now introduce his method.

Definition 7.1 Heegner number is a square-free positive integer d such that the imaginary quadratic field $\mathbb{Q}(\sqrt{-d})$ has class number 1.

Example 7.2 Only 9 Heegner numbers: 1, 2, 3, 7, 11, 19, 43, 67, 163

Heegner's main idea of constructing solution to $E : y^2 = x^3 - x$ is by using modular functions:

$$f : \mathcal{H} := \{z \in \mathbb{C}, Re(z) > 0\} \to E(\mathbb{C}).$$

Example 7.3 $e^{\pi\sqrt{163}}$ is almost an integer.

$e^{\pi\sqrt{163}} = 262537412640768743.99999999999925...$

Consider the modular function:

$$j(\tau) = \frac{1}{q} + 744 + 196884q + 21493760q^2 + \cdots$$

where $q = e^{2\pi i\tau}$.

8 Conjecture

Following the BSD conjecture, we have the following conjecture concerning the distribution of congruent numbers:

Conjecture 8.1 *If $n \equiv 5, 6, 7$ mod 8, then n is congruent.*

Conjecture 8.2 *If If $n \equiv 1, 2, 3$ mod 8, then n has probability 0 to be congruent:*

$$\lim_{X \to \infty} \frac{\sharp\{n \leq X : n \equiv 1, 2, 3 \text{ mod } 8 \text{ and congruent}\}}{X} = 0.$$

9 Concluding remarks

In this note, we only give a brief introduction to the congruent number problem and also give some very basic knowledge about number theory and arithmetic algebraic geometry. Of course, many topics and applications have been omitted. For more details, readers can refer to the book [3].

References

[1] Birch, B.J. and Swinnerton-Dyer, H.P.F., *Notes on elliptic curves. II.* J. Reine Angew. Math. 218 (1965), 79–108.

[2] Dickson, L.E., "History of the Theory of Numbers II". Carnegie Intitute of Washington, 1920.

[3] Koblitz, N.I., "Introduction to elliptic curves and modular forms". Vol- ume 97. Springer Science & Business Media, 2012.

[4] Silverman, J., "The arithmetic of elliptic curves". Volume 106. Springer Science & Business Media, 2009.

[5] Zagier, D., *L-series of elliptic curves, the Birch?Swinnerton-Dyer conjecture, and the class number problem of Gauss.* Notices Amer. Math. Soc 31/7 (1984), 739–743.

[6] Zhang, S.-W., *Congruent numbers and Heegner points.* In "Colloquium De Giorgi 2010?2012", pp. 61?68. Springer, 2013.

Regular domain perturbation problems

Paolo Luzzini (*)

Abstract. The study of the dependence of functionals related to partial differential equations and of quantities of physical relevance upon smooth domain perturbations is a classical topic and has been carried out by several authors. In this note we give an introductory overview about regular domain perturbation problems. We provide concrete examples, highlight the motivations and the possible applications, and present an outline of some new results on the longitudinal fluid flow along a periodic array of cylinders.

1 An informal introduction

We start this note with an informal introduction to regular perturbation problems of domains. By means of concrete examples we try to give an idea of what this type of perturbation problems are and, moreover, we explain the motivations that lead to study such problems. With this aim, in this section we avoid to be too rigorous, preferring a colloquial exposition also understandable by the non-specialists.

Let Ω be a subset of $\mathbb{R}^n$ which represents a physical object, and let $J(\Omega)$ be a quantity (a functional) which depends on the shape of the object represented by Ω. A very simple example is the case in which the quantity $J(\Omega)$ is the volume of the object Ω, that is

$$J(\Omega) \equiv \mathrm{Vol}(\Omega).$$

Else, the set Ω could represent the shape of a drum, and $J(\Omega)$ could be its fundamental tone. As a last example, the set Ω could play the role of an airplane's wing, and $J(\Omega)$ could be the aerodynamic resistance on the wing. If one starts to smoothly deform the shape of Ω (see Figure 1, then possibly the quantity $J(\Omega)$ will be affected. Back to the previous examples, this deformation maybe produces a change in the sound of our drum, or maybe the aerodynamic of our wing changes. The question that we are interest in answering here is the following:

what can be said on the regularity of the map $\Omega \mapsto J(\Omega)$?

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: pluzzini@math.unipd.it . Seminar held on December 5th, 2018.

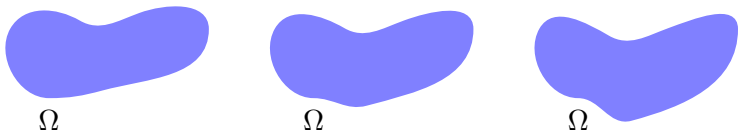


Figure 1

The previous question immediately raises an issue: at this point we do not have a structure on the class of admissible shapes which gives sense to say that the map $\Omega \mapsto J(\Omega)$ has some regularity. In the next sections we will see how to mathematically formalize a structure on the class of admissible shapes. However, here we leave this issue at an intuitive level: one has to think that the admissible perturbations of the shape of Ω are the ones which do not produce holes or cracks or other types of singularities, and there is some structure that measures such perturbations.

We now present with more details three examples. The first two examples come from mathematical problems, although they have a precise applied interpretation. The last example comes instead from an applied problem.

Example 1.1 Let u_Ω be the solution of the following Dirichlet problem for the Laplace equation

$$\begin{cases} \Delta u = 0 & \text{in } \Omega, \\ u = g & \text{on } \partial\Omega. \end{cases} \tag{1}$$

Let $x \in \Omega$. Let $J_1(\Omega)$ be the solution of the problem computed at x, *i.e.*

$$J_1(\Omega) \equiv u_\Omega(x).$$

Moreover, let $J_2(\Omega)$ be the energy integral of the solution of the problem, *i.e.*

$$J_2(\Omega) \equiv \int_\Omega |Du_\Omega(y)|^2 \, dy.$$

How do the functionals J_1, J_2 depend on the variation of the shape of Ω? We answer to this question in Subsection 2.1.

Example 1.2 Let $\lambda_1(\Omega)$ be the first eigenvalue of the Dirichlet Laplacian with homogeneous boundary conditions. We set

$$J(\Omega) \equiv \lambda_1(\Omega).$$

How does the functional J depend on the variation of the shape of Ω? We answer to this question in Subsection 2.2. We note that this problem is related to the example of the drum we made at the beginning of this section. Indeed, the pure tones of a drum are strictly related to the eigenvalues of the Dirichlet Laplacian (see *e.g.* Kac [9]).

Example 1.3 Let the domain Ω represent the shape of an airplane's wing, and let $J(\Omega)$ be the drag (aerodynamic resistance parallel to the fluid flow) or the lift (aerodynamic resistance perpendicular to the fluid flow) on the wing (see Figure 2). How do small changes and perturbations of the wing's shape affect the drag or the lift on the wing?

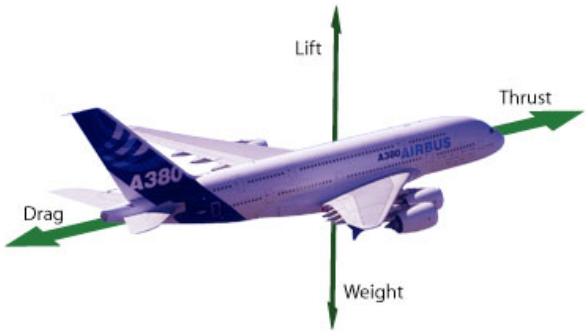


Figure 2

The reasons that lead to study these types of problems are several. Here we only mention that these problems arise in **shape optimization**. The aim of shape optimization is to find (if it exists) a set Ω^* which minimizes (or maximizes) a given functional J over a class of admissible sets $\mathcal{O}$:

$$\Omega^* \in \mathcal{O}, \quad J(\Omega^*) = \min_{\Omega \in \mathcal{O}} J(\Omega).$$

If one knows that the functional J is somehow "regular" (for example if it enjoys some differentiability properties), then one can apply differential calculus in order to find critical shapes as a first step in order to find optimal shapes. One can think about the finite dimensional case: the first step in order to find the points of minimum (and maximum) of a function of several real variables is to find the points where the gradient vanishes. Otherwise, for a constraint optimization problem one can use the Lagrange multipliers method. A generalizations of these techniques can be used also in the infinite dimensional case of shape optimization.

If we go back to Example 1.3 regarding the airplane's wing, a shape optimization problem could be to find the shape Ω of the cross-section of the wing which maximizes the lift $J(\Omega)$ under some constraints, *e.g.* fixed volume and fixed drag (see Figure 3). If one knows that the lift $J(\Omega)$ is differentiable with respect to the deformation of the shape of Ω, then by using tools from differential calculus one could in principle obtain some information on the critical shapes and accordingly on the optimal shapes.

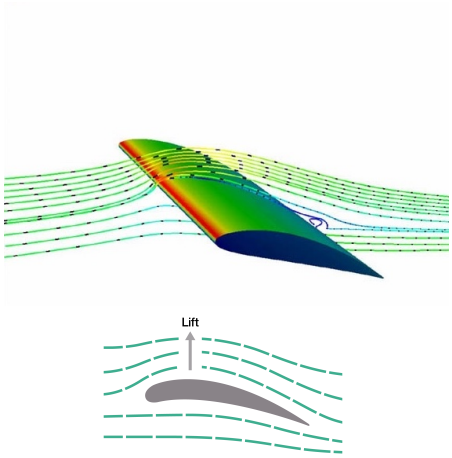


Figure 3

There exists a vaste literature regarding shape analysis and optimization of functionals related to partial differential equations or quantities of physical relevance, both from the theoretical and numerical point of view. Without being exhaustive we mention Sokołowski and Zolésio [19], Henrot and Pierre [8], Novotny and Sokołowski [17], Mohammadi and Pironneau [15], Laporte and Le Tallec [11].

2 Some known results

In this section we present some known results on regular perturbations of the problems introduced in Examples 1.1 and 1.2. Here we start to be less colloquial and more rigorous. Accordingly, first of all we have to formalize the shape perturbations of sets.

2.1 The Dirichlet problem for the Laplace equation

We fix Ω to be a bounded open subset of $\mathbb{R}^n$ such that

- Ω is connected;
- $\mathbb{R}^n \setminus \overline{\Omega}$ is connected;
- Ω is of class $C^{1,\alpha}$ for some $\alpha \in]0,1[$.

Then we consider the following class of diffeomorphisms on $\partial\Omega$.

$$\mathcal{A}_{\partial\Omega} \equiv \{\psi \in C^1(\partial\Omega, \mathbb{R}^n) : \psi, d\psi(y) \text{ are injective } \forall y \in \partial\Omega\}$$

If $\phi \in \mathcal{A}_{\partial\Omega}$, the Jordan-Leray separation theorem ensures that $\mathbb{R}^n \setminus \phi(\partial\Omega)$ has exactly two open connected components, and we denote by $\mathbb{I}[\phi]$ the bounded open connected component of $\mathbb{R}^n \setminus \phi(\partial\Omega)$ (see Figure 4).

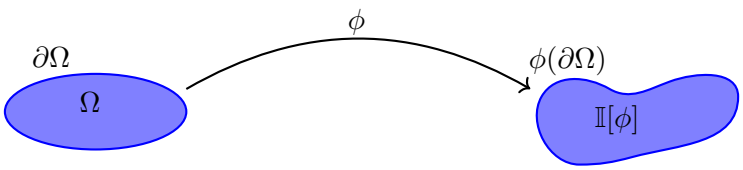


Figure 4

The ϕ-dependent set $\mathbb{I}[\phi]$ plays the role of the perturbed set, and the perturbation is made by perturbing the diffeomorphism $\phi \in \mathcal{A}_{\partial\Omega} \cap C^{1,\alpha}(\partial\Omega, \mathbb{R}^n)$, which is an open subset of the Banach space $C^{1,\alpha}(\partial\Omega, \mathbb{R}^n)$.

Next, we fix $g \in C^{1,\alpha}(\partial\Omega)$ and we consider the following Dirichlet problem for the Laplace equation in the ϕ-dependent set $\mathbb{I}[\phi]$.

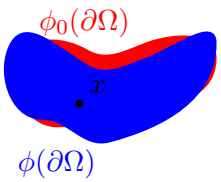
$$\begin{cases} \Delta u = 0 & \text{in } \mathbb{I}[\phi], \\ u = g \circ \phi^{-1} & \text{on } \phi(\partial\Omega). \end{cases}$$

It is well known that this problem admits a unique solution $u[\phi] \in C^{1,\alpha}(\overline{\mathbb{I}[\phi]})$. We are interested in understanding the dependence of the solution $u[\phi]$ and of its energy integral $\int_\Omega |Du[\phi]|^2 dy$ upon the diffeomorphisms ϕ, *i.e* upon the shape of the set. This problem has been considered in Lanza de Cristoforis [12], where is proved that these quantities depend analytically upon ϕ. Here, and throughout the present note, 'analytic' means always 'real analytic'. For the definition and properties of analytic operators in Banach spaces, we refer for example to Deimling [5, §15].

Theorem 2.1 (Lanza de Cristoforis '07) *Let $\phi_0 \in \mathcal{A}_{\partial\Omega} \cap C^{1,\alpha}(\partial\Omega, \mathbb{R}^n)$. Let $x \in \mathbb{I}[\phi_0]$. Then there exists an open neighborhood $\mathcal{U}$ of ϕ_0 in $\mathcal{A}_{\partial\Omega} \cap C^{1,\alpha}(\partial\Omega, \mathbb{R}^n)$ such that*

(i) *$x \in \mathbb{I}[\phi]$ for all $\phi \in \mathcal{U}$.*

(ii) *The map from $\mathcal{U}$ to $\mathbb{R}$ which takes ϕ to $u[\phi](x)$ is real analytic.*



Theorem 2.2 (Lanza de Cristoforis '07) *The map from $\mathcal{A}_{\partial\Omega} \cap C^{1,\alpha}(\partial\Omega, \mathbb{R}^n)$ to $\mathbb{R}$ which takes ϕ to $\int_{\mathbb{I}[\phi]} |Du[\phi]|^2 dy$ is real analytic.*

2.2 Eigenvalues of the Dirichlet Laplacian

We fix Ω to be an open subset of $\mathbb{R}^n$ such that

- Ω is connected;
- Ω is of finite measure.

Then we consider the following class of bi-Lipschitz homeomorphisms on Ω.

$$\mathbf{\Phi}_\Omega \equiv \left\{ \psi \in (\mathrm{Lip}(\Omega))^n : \inf_{\substack{x,y\in\Omega \\ x\neq y}} \left\{ \frac{|\psi(x)-\psi(y)|}{|x-y|} \right\} > 0 \right\}$$

If $\phi \in \mathbf{\Phi}_\Omega$, the ϕ-dependent set $\phi(\Omega)$ plays the role of the perturbed set, and the perturbation is made by perturbing the homeomorphism $\phi \in \mathbf{\Phi}_\Omega$ (see Figure 5).

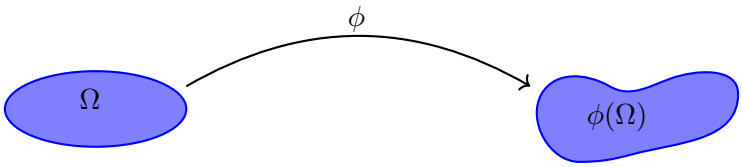


Figure 5

Next, we consider the eigenvalue problem for the Dirichlet Laplacian in the ϕ-dependent set $\phi(\Omega)$:

$$\begin{cases} -\Delta u = \lambda u & \text{in } \phi(\Omega), \\ u = 0 & \text{in } \partial\phi(\Omega). \end{cases}$$

More precisely, we consider its weak formulation. That is the problem

$$\int_{\phi(\Omega)} Dv(Dw)^t \, dx = \lambda \int_{\phi(\Omega)} vw \, dx \qquad \forall w \in W_0^{1,2}(\phi(\Omega))$$

in the unknown $v \in W_0^{1,2}(\phi(\Omega))$ (the Dirichlet eigenfunctions) and $\lambda \in \mathbb{R}$ (the Dirichlet eigenvalues). Such a problem is well known to have a sequence of eigenvalues

$$0 < \lambda_1[\phi] < \lambda_2[\phi] \leq \lambda_3[\phi] \leq \dots$$

which we write as many times as their multiplicity. Accordingly, we are interested in understanding the dependence of the eigenvalues $\lambda_k[\phi]$ ($k \in \mathbb{N}\setminus\{0\}$) upon the homeomorphism ϕ, *i.e.* upon the shape. Before stating the results, we must introduce a subclass of $\mathbf{\Phi}_\Omega$. Let $F \subseteq \mathbb{N}\setminus\{0\}$ be of finite cardinality. Let $\mathbf{\Phi}_\Omega^F$ be the subset of $\mathbf{\Phi}_\Omega$ of those transformations for which the eigenvalues with index in F may coincide but must not be equal to any of the remaining eigenvalues. That is

$$\mathbf{\Phi}_\Omega^F \equiv \Big\{ \phi \in \mathbf{\Phi}_\Omega : \lambda_k[\phi] \notin \{\lambda_m[\phi] : m \in F\} \quad \forall k \in \mathbb{N}\setminus(F\cup\{0\}) \Big\}.$$

The following result of Prodi [18] shows that simple eigenvalues depend analytically on the shape of the set.

Theorem 2.3 (Prodi '94) *Let $F = \{j\}$. Then the map from $\mathbf{\Phi}^F_\Omega$ to $\mathbb{R}$ which takes ϕ to $\lambda_j[\phi]$ is analytic.*

The case of multiple eigenvalues is more involved since the multiplicity of the eigenvalues may change under small shape perturbations. However, it turns out that the symmetric functions of the eigenvalues with index in F depend analytically on ϕ. In fact we have the following result of Lanza de Cristoforis and Lamberti [10].

Theorem 2.4 (Lamberti and Lanza de Cristoforis '04) *The map from $\mathbf{\Phi}^F_\Omega$ to $\mathbb{R}$ which takes ϕ to*

$$\Lambda_{F,s}[\phi] \equiv \sum_{\substack{j_1,\dots,j_s \in F \\ j_1<\dots<j_s}} \lambda_{j_1}[\phi]\cdots\lambda_{j_s}[\phi]$$

is analytic, for all $s = 1,\dots,|F|$.

The previous result is a generalization of the result of Prodi. Indeed, if $|F| = 1$ Theorem 2.4 immediately implies Theorem 2.3. Moreover, as a further corollary of Theorem 2.4 we have that a multiple eigenvalue depends analytically on those shape transformations which do not change the multiplicity of the eigenvalue. That is, if we set

$$\mathbf{\Theta}^F_\Omega \equiv \Big\{\phi \in \mathbf{\Phi}^F_\Omega : \lambda_m[\phi] \text{ have a common value } \lambda_F[\phi] \quad \forall m \in F\Big\},$$

then we have the following.

Corollary 2.5 *The map from $\mathbf{\Theta}^F_\Omega$ to $\mathbb{R}$ which takes ϕ to $\lambda_F[\phi]$ is real analytic.*

More recently, similar results have been proved for the eigenvalues of other operators. We mention without being exhaustive Grinfeld [6] for the C^2 regularity for the biharmonic operator, Buoso and Lamberti [3] for the analyticity for the Reissner-Mindlin system, and Buoso [2] for the analyticity for elliptic systems.

We conclude this section by explaining a possible application in the framework of shape optimization of the previous regularity results for eigenvalues. The celebrated Rayleigh-Krahn-Faber Theorem states that the ball minimizes the first eigenvalue of the Dirichlet Laplacian among all the domains in $\mathbb{R}^n$ with a prescribed finite measure. That is, if $\mathbb{B}_n$ is the n-dimensional ball with a fixed radius then $\lambda_1(\mathbb{B}_n) \le \lambda_1(D)$ for all $D \subseteq \mathbb{R}^n$ such that $|D| = |\mathbb{B}_n|$ (see Figure 6).

Figure 6

Do there exist other sets which minimize the first eigenvalues? The answer turns out to be negative. This can be proved exploiting the regularity of the first eigenvalues upon shape

deformations. Indeed, one can apply the Lagrange multipliers method and easily deduce that if a domain D minimizes the first eigenvalue of the Dirichlet Laplacian among the sets with a prescribed finite volume, then D must be a ball.

3 The longitudinal flow along a periodic array of cylinders

In this section we present some new results regarding the regular perturbation of a problem which comes from the study of the properties of porous materials. More precisely, we study the regularity of the longitudinal fluid flow along a periodic array of cylinders upon perturbations of the shape of the cross-section of the cylinders and of the periodicity structure.

First of all, we introduce the problem in an informal way. We consider an infinite periodic array of parallel cylinders of any shape and a Newtonian fluid which is flowing at low Reynolds numbers along the cylinders (see Figure 7).

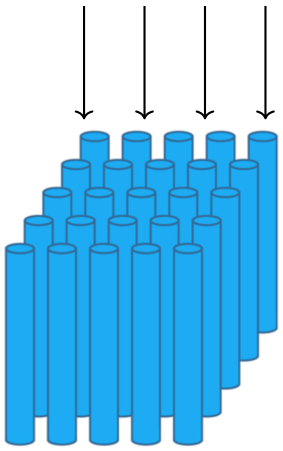


Figure 7

The equations that describe the fluid motion in the case of a Newtonian fluid are the Stokes flow equations:

$$\begin{cases} \mu \Delta \mathbf{u} - Dp = 0, \\ \operatorname{div} \mathbf{u} = 0. \end{cases}$$

where

- $\mathbf{u} = (u_1, u_2, u_3)$ is the velocity field of the fluid;
- Dp is pressure gradient;
- μ is the viscosity of the fluid.

We assume the so called no-slip condition, which says that the velocity field is zero at the solid boundary, *i.e.*

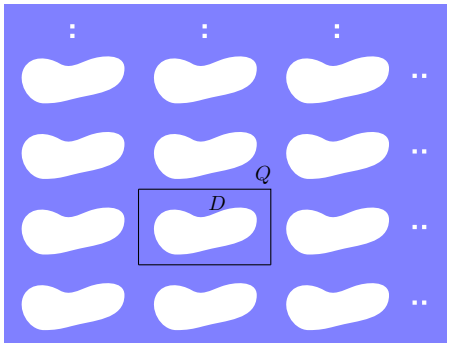
$$\mathbf{u} = 0 \text{ at the boundary of the cylinders.}$$

Moreover, we also assume that

- Dp is constant and parallel to the axis of the cylinders, and without loss of generality we assume $Dp = (0, 0, 1)$;
- $\mu = 1$.

Under such assumptions one readily verifies that the first two components of the velocity field are zero, and the third component u_3 satisfies a periodic Dirichlet problem for the Poisson equation with homogeneous boundary conditions in the cross-section of the cylinder's array (see *e.g.* Adler [1]). That is

$$\begin{cases} \Delta u_3 = 1, \text{ in } \overline{Q} \setminus \overline{D} \\ u_3 \text{ periodic w.r.t. the cell } Q, \\ u_3 = 0 \text{ on } \partial D. \end{cases}$$



Then we define the **average flow velocity over the cell** as

$$\Sigma \equiv \frac{1}{|Q|} \int_{Q \setminus D} u_3 \, dx.$$

We are interested in studying the dependence of the average flow velocity Σ upon the perturbation of the cylinder's cross-section shape and of the periodicity structure. This quantity has been studied by several authors. Here, for example, we mention Hasimoto [7], Mityushev and Adler [14] (see also references therein), and Musolino and Mityushev [16].

As before, first we have to formalize the shape perturbations of our setting. We fix $l \in]0, +\infty[$ and we define the periodicity Q_l cell and the periodicity matrix q_l as

$$Q_l \equiv]0, l[\times]0, 1/l[\,, \qquad q_l \equiv \begin{pmatrix} l & 0 \\ 0 & 1/l \end{pmatrix}.$$

We note that $|Q_l| = 1$ for all $l \in]0, +\infty[$. This choice helps making the computations simpler and the exposition clearer and it is of course physically meaningful. However, this restriction is not necessary and we could consider a more general periodic structure and a more general perturbation of the periodic structure. Next, we fix Ω as in Subsection 2.1 and we set

$$\mathcal{A}_{\partial\Omega}^{Q_1} \equiv \{\psi \in \mathcal{A}_{\partial\Omega} : \psi(\partial\Omega) \subseteq Q_1\}.$$

As in Subsection 2.1, if $\phi \in \mathcal{A}^{Q_1}_{\partial\Omega}$ we denote by $\mathbb{I}[\phi]$ the bounded open connected component of $\mathbb{R}^2 \setminus \phi(\partial\Omega)$. Next, if $l \in]0,+\infty[$ and $\phi \in \mathcal{A}^{Q_1}_{\partial\Omega}$ we set

$$\mathbb{S}[l,\phi]^- \equiv \mathbb{R}^2 \setminus \overline{\bigcup_{z\in\mathbb{Z}^2} (q_l z + q_l \mathbb{I}[\phi])},$$

(see Figure 8). The set $\mathbb{S}[l,\phi]^-$ plays the role of the cross-section of the cylinder's array. The perturbation of this set is performed by perturbing $(l,\phi) \in]0,+\infty[\times(\mathcal{A}^{Q_1}_{\partial\Omega} \cap C^{1,\alpha}(\partial\Omega,\mathbb{R}^2))$.

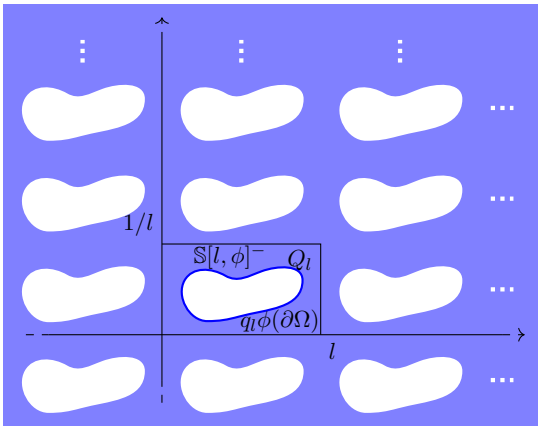


Figure 8

Now, we can rewrite the problem in the set $\mathbb{S}[l,\phi]^-$:

$$\begin{cases} \Delta u = 1 \text{ in } \mathbb{S}[l,\phi]^-, \\ u \text{ is } q_l\text{-periodic}, \\ u = 0 \text{ on } q_l\phi(\partial\Omega). \end{cases}$$

This problem is known to have a unique solution $u[l,\phi]$ in $C^{1,\alpha}(\overline{\mathbb{S}[l,\phi]^-})$. Then the average flow velocity over the cell can be written as

$$\Sigma[l,\phi] \equiv \int_{Q_l \setminus q_l\mathbb{I}[\phi]} u[l,\phi]\, dx.$$

We are interested in understanding the regularity of $\Sigma[\cdot,\cdot]$ upon (l,ϕ). Our main result, which solve this issue, is the following (see [13]).

Theorem 3.1 (L., Musolino, and Pukhtaievych '18) *The map*

$$\Sigma[\cdot\,,\cdot] \,:\quad]0,+\infty[\times\Big(\mathcal{A}^{Q_1}_{\partial\Omega} \cap C^{1,\alpha}(\partial\Omega,\mathbb{R}^2)\Big) \longrightarrow \mathbb{R}$$

which takes a pair (l,ϕ) *to* $\Sigma[l,\phi]$ *is real analytic.*

We conclude this note with some possible applications and future developments. As we have already said, this type of results are very useful in shape optimization: our result permits to apply differential calculus in order to find critical "rectangle-shape" pairs (l, ϕ) as a first step in order to maximize or minimize the average flow velocity $\Sigma[l, \phi]$ under some constraints. Furthermore, suppose that we have a one-parameter analytic family of pairs $(l_\delta, \phi_\delta)_{\delta \in]-\delta_0, \delta_0[}$ parametrized by a small parameter $\delta \in]-\delta_0, \delta_0[$. Then, by the analyticity of $\Sigma[\cdot, \cdot]$ we have

$$\Sigma[l_\delta, \phi_\delta] = \sum_{j=0}^{+\infty} c_j \delta^j, \tag{2}$$

for δ close to zero. For practical applications it is of interest to compute the coefficients $(c_j)_{j \in \mathbb{N}}$. Dalla Riva, Musolino, and Pukhtaievych [4] developed a completely constructive method to compute the coefficients for the effective conductivity of periodic two-phase dilute composite. The computation is based on the solutions of systems of integral equations. This type of approach can be exploited also in this case, in order to obtain an explicit expression for all the coefficients $\{c_j\}_{j \in \mathbb{N}}$ in the series (2). This is the object of future investigations and Theorem 3.1 provides the theoretical background for this aim.

References

[1] P.M. Adler, "Porous media: geometry and transports". Butterworth/Heinemann, 1992.

[2] D. Buoso, *Shape differentiability of the eigenvalues of elliptic systems.* Integral Methods in Science and Engineering: Theoretical and Computational Advances, Birkhäuser, 2015.

[3] D. Buoso and P.D. Lamberti, *Shape sensitivity analysis of the eigenvalues of the Reissner-Mindlin system.* SIAM J. Math. Anal. 47 (2015), 407–426.

[4] M. Dalla Riva, P. Musolino, and R. Pukhtaievych, *Series expansion for the effective conductivity of a periodic dilute composite with thermal resistance at the two-phase interface.* Asymptot. Anal. 111 (3-4) (2019), 217–250.

[5] K. Deimling, "Nonlinear Functional Analysis". Springer-Verlag, Berlin, 1985.

[6] P. Grinfeld, *Hadamard's formula inside and out.* J. Optim. Theory Appl. 146 (2010), 654–690.

[7] H. Hasimoto, *On the periodic fundamental solutions of the Stokes equations and their application to viscous flow past a cubic array of spheres.* J. Fluid Mech. 5 (1959), 317–328.

[8] A. Henrot and M. Pierre, "Variation et optimisation de formes. Une analyse géométrique". Mathématiques & Applications, Springer, Berlin, 2005.

[9] M. Kac, *Can One Hear the Shape of a Drum?.* American Mathematical Monthly 73, no. 4, part 2, 1966.

[10] P.D. Lamberti and M. Lanza de Cristoforis, *A real analyticity result for symmetric functions of the eigenvalues of a domain dependent Dirichlet problem for the Laplace operator.* J. Nonlinear Convex Anal. 5 (2004), no. 1, 19–42.

[11] E. Laporte and P. Le Tallec, "Numerical Methods in Sensitivity Analysis and Shape Optimization". Birkhäuser, 2003.

[12] M. Lanza de Cristoforis, *Perturbation problems in potential theory, a functional analytic approach.* J. Appl. Funct. Anal. 2 (2007), 197–222.

[13] P. Luzzini, P. Musolino, and R. Pukhtaievych, *Shape analysis of the longitudinal flow through a periodic array of cylinders.* J. Math. Anal. Appl. 477 (2019), no. 2, 1369–1395.

[14] V. Mityushev and P.M. Adler, *Longitudinal permeability of spatially periodic rectangular arrays of circular cylinders. II. An arbitrary distribution of cylinders inside the unit cell.* Z. Angew. Math. Phys. 53 (2002), 486–517.

[15] B. Mohammadi and O. Pironneau, "Applied Shape Optimization for Fluids". Oxford University Press, 2001.

[16] P. Musolino and V. Mityushev, *Asymptotic behavior of the longitudinal permeability of a periodic array of thin cylinders.* Electron. J. Differential Equations 290 (2015), 1–20.

[17] A.A. Novotny and J. Sokołowski, "Topological derivatives in shape optimization". Interaction of Mechanics and Mathematics. Springer, Heidelberg, 2013.

[18] G. Prodi, *Dipendenza dal dominio degli autovalori dell'operatore di Laplace.* Ist. Lombardo Accad. Sci. Lett. Rend. A. 128 (1994), 3–18.

[19] J. Sokołowski and J.P. Zolésio, "Introduction to Shape Optimization. Shape sensitivity analysis". Springer Series in Computational Mathematics, 16. Springer-Verlag, Berlin, 1992.

Real Options: an overview

Dimitrios Zormpas (*)

Abstract. Financial options are contracts that derive their value from the performance of an underlying asset. They give to their holder the right, but not the obligation, to buy/sell an asset at a predetermined price and time. Contracts similar to options have been used since ancient times. However, the most basic model for their pricing was proposed in the early 1970's leading to a Nobel prize in 1997. In the late 1970's the term "real options" is coined by Stewart Myers. According to the real options approach an investment characterized by uncertainty and irreversibility is like a financial option on a real asset. For instance, a potential investor has the right but not the obligation to pay a given amount of money in order to make an investment and gain access to the corresponding profit flow. Using standard option pricing tools one can also study the option to leave a market, outsource production, mothball a production plant etc. In this seminar, I refer to the correspondence between financial and real options and then I present the simplest model in the real options literature. Finally, I make a reference to an application of the real options approach in energy economics.

Keywords: Investment analysis, Real options

JEL classification: C61, D92, G30

1 Introduction

Financial derivatives are contracts between two parties. For instance, a European call option on the amount of X units of a certain asset with strike price K and exercise date $T\,(>0)$ is a contract written at time $t=0$ with the following properties: The holder of the contract has the right, but not the obligation, to buy X units of the asset at time T from the issuer of the contract paying a price K. Similarly, a put option gives the right to the holder of the option to sell X units of a certain asset. On the contrary, an American call/put option allows the exercise of the option at any time before T.

According to Björk (2009), "Options of the type above are traded on options markets all over the world, and the underlying objects can be anything from foreign currencies to stocks, oranges, timber or pig stomachs. For a given underlying asset there are typically a large number of options with different expiration dates and different strike prices." The

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: zormpas@math.unipd.it . Seminar held on December 19th, 2018.

main question that the relevant literature addresses is "what is a fair price for an option-like contract", or, in other words, how much money should the writer of a financial option ask for writing such an option? The Black-Scholes-Merton model that was published in the early 1970's is the first model answering this question leading eventually to a Nobel prize in Economics in 1997. Since then the model has been refined and today more sophisticated versions of it are widely used worldwide.

In the late 1970's the term "real options" is coined by Stewart Myers who argues that an investment is like a financial option (e.g. an American call option) on a real asset. In fact, a potential investor (holder of the option) has the right but not the obligation to pay a given amount of money (the strike price) in order to make an investment (infinite or quasi-infinite horizon) and gain access to the corresponding profit flow (the underlying asset). Apart from the option to invest, real options are also the option to abandon a current business activity, the option to mothball a production plant reactivating it latter if the conditions improve etc. For an analytical overview of the real options approach see Dixit and Pindyck (1994).

2 The option to invest

In this section I discuss the simplest real option, namely, the option to invest.

Suppose that a potential investor is contemplating investing in a project of known cost $I > 0$ and that the project's value is fluctuating over time according to the following geometric Brownian motion:

$$\frac{dV_t}{V_t} = \alpha dt + \sigma dz_t \tag{1}$$

where $dz_t \sim N(0, dt)$. The question that the potential investor needs to answer is: "How big should V_t be before spending I?".

The value of the option to invest is:

$$F(V_t) = \max\left\{V_t - I, \frac{1}{1+rdt} E_t\left[F\left(V_t + dV_t\right)\right]\right\} \tag{2}$$

where $V_t - I$, is the termination value, i.e., the value of the option to invest at the time of the investment and $\frac{1}{1+rdt} E_t\left[F\left(V_t + dV_t\right)\right]$, is the continuation value, i.e., the value of postponing the investment for the next period. If the optimal investment threshold is V_T and $V_0 \geq V_T$ then the problem reduces to a mere maximization of the net-present value of the investment. On the contrary, in the more interesting case where $V_0 < V_T$, i.e., when the optimal investment threshold lies somewhere in the future, Eq. (2) reduces to $F(V_t)\, rdt = E_t\left[F\left(dV_t\right)\right]$ which gives:

$$\frac{1}{2}\sigma^2 V^2 F_{VV} + \alpha V F_V - rV = 0 \tag{3}$$

Eq. (3) needs to be solved subject to the following constraints:

$$F(0) = 0 \tag{4}$$

$$F(V_T) = V_T - I \tag{5}$$

$$F'(V_T) = 1 \tag{6}$$

The condition $F(0) = 0$ suggests that if the value of the project goes to zero, also the value of the option to invest in such a project goes to zero. Eq. (5) verifies that as soon as the option holder exercises the option s/he will receive exactly the termination value. Last, Eq. (6) is a standard smooth pasting condition. In order to solve this second-order ordinary differential equation of Eq. (3) we guess a solution of type $F(V_t) = AV^{\beta}$. From this we get:

$$\frac{1}{2}\sigma^2\beta\left(\beta-1\right)+\alpha\beta-r=0 \tag{7}$$

This has two roots:

$$\beta_{1,2} = \frac{1}{2} - \frac{\alpha}{\sigma^2} \pm \sqrt{\left(\frac{\alpha}{\sigma^2} - \frac{1}{2}\right)^2 + \frac{2r}{\sigma^2}} \text{ with } \beta_1 > 1, \beta_2 < 0 \tag{8}$$

Summing up, the general solution of $\frac{1}{2}\sigma^2 V^2 F_{VV} + \alpha V F_V - rV = 0$ is

$$F(V_t) = A_1 V^{\beta_1} + A_2 V^{\beta_2} \tag{9}$$

Thanks to $F(0) = 0$, Eq. (9) reduces to $F(V_t) = A_1 V^{\beta_1}$. From the value matching and the smooth pasting conditions we obtain:

$$A_1 = \frac{V_T - I}{V_T^{\beta_1}} > 0 \tag{10}$$

$$V_T = \frac{\beta_1}{\beta_1 - 1} I \tag{11}$$

Last, the value of the option for $V_0 < V_T$ is:

$$F(V_t) = (V_T - I)\left(\frac{V_0}{V_T}\right)^{\beta_1} \tag{12}$$

Note that $V_T = \frac{\beta_1}{\beta_1 - 1} I > I$. This means that when an investment is characterized by uncertainty and irreversibility the net-present value criterion which suggests that $V_T = I$ does not hold anymore. On the contrary, there is a wedge $\frac{\beta_1}{\beta_1 - 1} > 1$ that is capturing the value of the option to wait before investing.

Note that $\frac{\partial V_T}{\partial \sigma} > 0, \frac{\partial V_T}{\partial \alpha} > 0$ and $\frac{\partial V_T}{\partial r} < 0$. In words, as uncertainty soars up or the project becomes more promising, a higher threshold for investment is required. On the contrary, the more impatient the potential investor, the earlier the investment.

3 Investing in electricity production under a reliability options scheme

3.1 Capacity remuneration mechanisms and reliability options

The penetration of renewable energy sources in electricity systems worldwide contributes to the decarbonisation of electricity production and this is an important step towards a greener future. However, the phenomenon of decarbonisation sheds light on the issue of electric capacity security. Since energy production from photovoltaics or wind farms depends on weather patterns, the managers of electricity supply need to make sure that the supply of electric energy will meet the corresponding demand even when the weather patterns are unfavorable.

Capacity remuneration mechanisms are instruments used by energy-system operators worldwide explicitly to encourage investments in electricity production. Among them the reliability options scheme is gaining momentum. A reliability option is reminiscent of a financial option in the sense that it is a contract between a power plant (writer of the reliability option) and the energy-system operator (buyer of the reliability option) that gives the right to the latter to buy from the former energy at a predetermined price paying a premium in return. For more details see Andreis et al. (2018).

The question that I address below is: "How do reliability options affect investments in the energy sector?". In particular, what is the effect in the timing and the value of the option to invest?

3.2 Investing in electricity production under a reliability options scheme

A potential investor contemplates investing in a power plant. The price of electricity is assumed to fluctuate over time according to the following geometric Brownian motion:

$$\frac{dP_t}{P_t} = \alpha dt + \sigma dW_t \text{ with } P_0 = P \tag{13}$$

The unit production cost of electricity is assumed to be constant and equal to $c \geq 0$. The problem for the potential investor is: "when to invest if the sunk investment cost associated with the project is $I > 0$?".

Note that when there is not a reliability options scheme in place, the instantaneous profit is $\pi_t = P_t - c$ and the value of the option to invest for $P < P_T$ is $F(P_t) = \left(\frac{P_T}{r-\alpha} - \frac{c}{r} - I\right)\left(\frac{P}{P_T}\right)^{\beta_1}$ where $P_T = \frac{\beta_1}{\beta_1 - 1}(r-\alpha)\left(\frac{c}{r} + I\right)$ is the optimal investment threshold.

On the contrary, when a reliability options scheme is in place the instantaneous profit function is $\overline{\pi}_t = \min\{P_t, K\} - (c-m)$ where $K > 0$ is the strike price of the reliability options scheme and $m > 0$ is the premium that the power plant receives ex-ante. Alternatively, we can write:

$$\overline{\pi}_t = \begin{cases} P_t - n \text{ for } P_t \leq K \\ K - n \text{ for } P_t > K \end{cases} \tag{14}$$

where $n = c - m$.

The new profit function results in a new value function:

$$\overline{V}(P_t) = \begin{cases} AP_t^{\beta_1} + \frac{P_t}{r-\alpha} - \frac{n}{r} & \text{for } P_t \leq K \\ BP_t^{\beta_2} + \frac{K-n}{r} & \text{for } P_t > K \end{cases} \tag{15}$$

where $A = -\frac{r-\beta_2\alpha}{(r-\alpha)(\beta_1-\beta_2)r}K^{1-\beta_1} < 0$ and $B = -\frac{r-\beta_1\alpha}{(r-\alpha)(\beta_1-\beta_2)r}K^{1-\beta_2} < 0$.

The terms $AP_t^{\beta_1}$ and $BP_t^{\beta_2}$ capture the obligation of the power plant to respect the reliability options contract. More precisely, the former captures the obligation of the power plant to cash K as soon as $P_t > K$. On the contrary, the latter captures the obligation of the power plant to cash P_t whenever $P_t \leq K$.

As before, the value of the option to invest is:

$$\overline{F}(P_t) = \max\left\{\overline{V}(P_t) - I, \frac{1}{1+rdt}E_t\left[\overline{F}(P_t + dP_t)\right]\right\} \tag{16}$$

Note that two possible cases arise. On one hand, we might have a K that is larger than, or at most equal to, the optimal investment threshold or, on the other, K might be smaller than the optimal investment threshold.

One can show that, provided that $K \geq \frac{\beta_1}{\beta_1-1}(r-\alpha)\left(\frac{n}{r} + I\right)$, the optimal investment threshold is equal to

$$P_T^* = \frac{\beta_1}{\beta_1 - 1}(r-\alpha)\left(\frac{n}{r} + I\right) (\leq K) \tag{17}$$

and the value of the option to invest is equal to

$$\overline{F}(P_t) = \left(AP_T^{*\beta_1} + \frac{P_T^*}{r-\alpha} - \frac{n}{r} - I\right)\left(\frac{P}{P_T^*}\right)^{\beta_1} \tag{18}$$

Notably, $P_T^* < P_T$, i.e., the investment takes place earlier than in the case without a reliability options scheme in place. Unsurprisingly, $\partial P_T^*/\partial m < 0, \partial\overline{F}(P_t)/\partial m > 0$ and $\partial\overline{F}(P_t)/\partial K > 0$.

On the other hand, provided that $K \in \left(Ir + n, \frac{\beta_1}{\beta_1-1}(r-\alpha)\left(I + \frac{n}{r}\right)\right)$, the optimal investment threshold is equal to

$$P_T^{**} = \left[\frac{1}{B}\frac{\beta_1}{\beta_1 - \beta_2}\left(I - \frac{K-n}{r}\right)\right]^{\frac{1}{\beta_2}} \tag{19}$$

and the value of the option to invest is equal to:

$$\overline{F}(P_t) = \left(BP_T^{**\beta_2} + \frac{K-n}{r} - I\right)\left(\frac{P}{P_T^{**}}\right)^{\beta_1} \tag{20}$$

As expected: $\partial P_T^{**}/\partial m < 0$, $\partial P_T^{**}/\partial K < 0$, $\partial\overline{F}(P_t)/\partial K > 0$, $\partial\overline{F}(P_t)/\partial m > 0$. Interestingly, in this case the effect of the reliability options mechanism both on the timing and the value of the option to invest is ambiguous.

Summing up, here I present a simple extension of the standard real options model discussing investments in electricity production when a reliability options scheme is in place. A reliability options scheme sets a cap at the price of electricity but at the same time pays a premium to the power plants that write reliability options. I show how the combination of the price cap and the option premium that characterize the reliability option determine the timing and value of investments in the electricity sector.

4 Conclusion

The real options approach treats investments characterized by uncertainty and irreversibility. It builds on the idea that investments are like financial options in real assets in the sense that a potential investor has the right but not the obligation to pay the investment cost in order to gain access to the stochastic profit flow generated by the investment project under question. In this seminar I present the most simple model in the real options literature and then I refer to some results from my ongoing research work.

References

[1] Andreis, L. Flora, M., Fontini, F and Vargiolu, T., *Pricing Reliability Options under different electricity prices' regimes*. Mimeo (2018).

[2] Björk, T., "Arbitrage Theory in Continuous Time". Third Edition, Oxford University Press, 2009.

[3] Dixit, A., and Pindyck, R.S., "Investment under Uncertainty". Princeton University Press, Princeton, 1994.

Conservation laws with transition phase for supply chains

Maria Teresa Chiri (*)

Abstract. We present an overview of existing models for manufacturing process of industrial goods involving PDEs and in particular conservation laws. Then we introduce a new model for supply chains on a network based on conservation laws with discontinuous flux evolving on each arc (sub-chain) and on buffers of limited capacity in every junction (separating sub-chains). The dynamics of every arc is governed by a continuity equation describing the evolution of the density of objects processed by the supply chain. The flux is discontinuous at the maximal density since it admits different values according with the free or congested status of the supply chain.

1 Introduction

Conveyor belts are component used in automata distribution and warehousing, whose origin dates back to 1892, by Thomas Robinson. They were introduced for carrying coal, ores and other products, but not too late they found wide use in other different sectors. In fact nowadays conveyor systems have large application in industries for transportation of materials, goods and passengers, since they represent a quick and efficient technology which allows to move objects of different nature and have also some popular consumer application, as in supermarkets and airports. Hence conveyor belts constitute the pivot on which a more complex structure is based: we are talking about supply chain, a system of organizations, people, activities, information, and resources involved in moving/processing a product or service from supplier to customer.

In the last decade several mathematical models were developed in order to describe the flow of particles along a single conveyor belt and more generally along a chain or network of conveyor systems. The main distinction is between the microscopic (discrete) models which track each part in the material flow and macroscopic (continuous) models relying on conservation laws which determines the motion of the part density (DGHP10) The former models captures the most accurate dynamics but get computational extremely costly and

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: chiri@math.unipd.it . Seminar held on January 30th, 2019.

produce inefficient simulation times, while the latter are inspired to continuous traffic flow models and captures phenomena such as queuing and congestion.

Figure 1. Classic prototype conveyor belt.

After a brief overview of already existing continuous models for supply chains, based on scalar conservation laws, we will introduce a new model that presents more realistic features. It is based on scalar conservation laws with flux discontinuous in the conserved quantity. This choice of the flux was introduced for the first time in [AGH11] to study shutdown of the production line due to a failure, and the time evolution of the recovery of the production line once the failure has been repaired. If there is now a vast literature for the case of conservation laws with spatial discontinuous flux (see [BK08] for a brief introduction), the same cannot be said for the case that we consider. However there is no way to avoid this peculiarity of the flux since conveyors used in industrial settings include tripping mechanisms which allow for workers to immediately shut down the conveyor when a problem arises. Hence to get a realistic description we need to consider this kind of discontinuity. The first to consider scalar conservation laws with flux discontinuous in the conserved quantity was Gimse in [Gim93] in models for two phase flow in porous media where flow properties change abruptly at some saturation.
Dias and al. in [DFR05] analyze the limit case of a phase transition and studied the problem by resorting regularization of the flux function though some Friedrichs' mollifiers to fall back into classical theory (this is also the same approach used in [AGH11]).
In [BvcGMSG11] the authors find an extended framework for fluxes with jump discontinuity, introducing a concept of weak solution for the conservation law and establishing its existence for a class of fluxes that have at most countably number of monotone jumps.
A different point of view is given in [HJP13] where the authors introduce an explicit transition phase approach enlarging the set of variable for the equation and considering not only the density but also the phase which can be free or congested. The Hamilton Jacobi formulation associated to this kind of conservation law is studied in [AC19]. Although the HJ equation is also non-trivial to treat because of the discontinuity of the Hamiltonian in

the gradient, its solution reveals important information about the solution for the conservation law. Moreover under the hypotheses stated in the following it is possible to give an explicit formulation (Hopf Lax type) for the solution of the HJ.

2 Previous Models

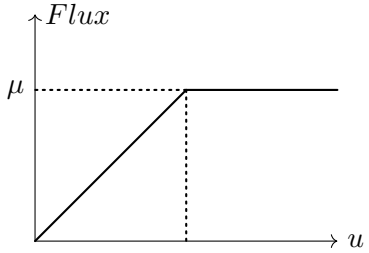
In this section we recall the main continuous models for supply chain introduced in the last years. Although they arise from such strictly applicative situations, these models offer important challenges also and above all from a theoretical point of view.

2.1 Armbruster- Degond-Ringhofer 2006

The first continuous model for conveyor belt was introduced in [ADR07], inspired by traffic flow models for which a large body theory had already been developed. It is based on conservation laws of the form

$$\partial_t u + \partial_x \min\{\mu, u\} = 0 \tag{2.1}$$

where the variable $x \in [a, b]$ represents the position along the single chain, $u : [0, +\infty) \times [a, b] \to [0, +\infty)$, function of time and position, stands for the product density, and μ is a bound on the rate of flux.



The number of parts processed is conserved but can be large, therefore a scalar conservation law is actually the most appropriate kind of equation to describe the physical behavior. Equation (2.1) is standard with a Lipschitsz continuous flux, hence it can be studied using the classical theory. However, although the flux is bounded, the density of parts can grow indefinitely which means that the chain has infinite capacity, this makes the model physically not very realistic. The defect was solved in the following model.

2.2 Armbruster - Gottlich - Herty 2011

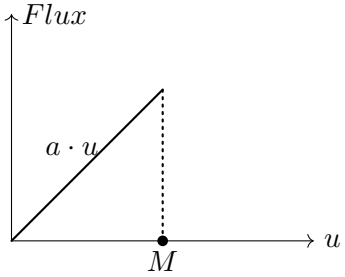
In [AGH11] the authors give a contribute to the body of continuous models by developing a model for supply chains or factories with finite work in progress. For evolution of parts they consider scalar conservation laws of the form

$$u_t + F_x(u)_x = 0 \tag{2.2}$$

with

$$F(u) = \begin{cases} a\,u & \text{if } u < M, \\ 0 & \text{if } u \geq M \end{cases}$$

where M is the maximum storage capacity in the processor.
The flux is discontinuous at maximum density and the most important consequence is the fact that the flux of the solution can not be uniquely determined through evaluation. Indeed assume that in a point a solution of (2.2) is exactly equal to M, then we are not able to say (in such a point) if the correspondent flux is 0 or $\lim_{s\to M^-} F(s)$.



Since at the beginning it was not clear how to deal directly with this kind of equation, the problem was studied by considering continuous approximation of the flux. A more direct approach was introduced in [HJP13].

2.3 Herty-Jorres-Piccoli 2013

In their work equation (2.2) is studied by modifying the dynamic in the following way: the flux F is replaced with a flux G with argument the density $u \in [0, M]$ and a second argument $\mathcal{S}$ attaining value in the finite set $\{\mathcal{F}, \mathcal{C}\}$ and representing the status of the belt. Here $\mathcal{F}$ is the free phase and $\mathcal{C}$ the congested. More explicitly $G : [0, M] \times \{\mathcal{F}, \mathcal{C}\} \to \mathbb{R}$ is given by

$$G(u, \mathcal{S}) = \begin{cases} F(u) & \text{if } 0 \leq u < M, \quad \mathcal{S} = \mathcal{F} \\ \lim_{u\to M^-} F(u) & \text{if } u = M, \quad \mathcal{S} = \mathcal{F} \\ F(u) & \text{if } 0 \leq u < M, \quad \mathcal{S} = \mathcal{C} \\ 0 & \text{if } u = M, \quad \mathcal{S} = \mathcal{C} \end{cases} \tag{2.3}$$

where the third case never occurs and is added just to have a well defined function G on the full domain $[0, M] \times \{\mathcal{F}, \mathcal{C}\}$.
The evolution of $(u(t, x), \mathcal{S}(t, x))$ corresponding to (2.2) is given by a conservation law paired to a state constraint :

$$\begin{cases} u_t + G(u, \mathcal{S})_x = 0 \\ \mathcal{S}(t, x) = \mathcal{C}(t, x) \implies u(t, x) = M \end{cases} . \tag{2.4}$$

The meaning of the state constraint is that the congested phase can appear only when $u(t,x) = M$, that is at maximal density.

Both (2.1) and (2.2) have been largely studied for evolution of parts not only on a single chain but also on network. A more recent model for supply chain is described in [AC19]. Here we give just a brief description of it.

3 A New Model

Consider a family of $n+m$ arcs joining at a node. We denote with indices $i \in \{1, ..., m\} = \mathcal{I}$ the incoming arcs, with $j \in \{1, ..., n\} = \mathcal{O}$ the outgoing arcs. On the $k-th$ arc the density of parts is described by the scalar conservation law

$$u_t + F_k(u)_x = 0$$

with $t > 0$, $x \in [-\infty, 0]$ for incoming and $x \in [0, \infty]$ for outgoing arcs. On the flux F_k we impose the following assumptions:

$$s \to F_k(s) \text{ smooth on } [0, M_k), \quad \partial_s^2 F_k \leq 0, \quad F(0) = 0 \text{ and } F_k(M_k) \in [0, N_k]. \tag{3.1}$$

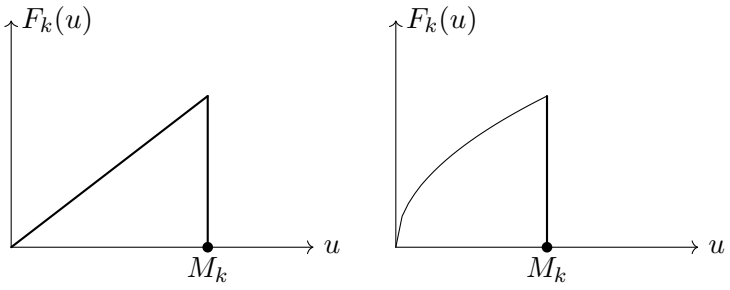


Figure 2. Two examples of fluxes which satisfy (3.1).

Given an initial data on each arc

$$u_k(0, x) = u_{k,0}(x) \quad k = 1, ..., n + m \tag{3.2}$$

we need to add a suitable set of boundary conditions in order to determine a unique solution. These boundary conditions provide additional constraints on the traces of the good densities

$$\bar{u}_k(t) = \begin{cases} \lim_{x \to 0^-} u_k(t, x) & k \in \mathcal{I} \\ \lim_{x \to 0^+} u_k(t, x) & k \in \mathcal{O} \end{cases}$$

near the junction.
Since we want to consider a realistic model, we assume that in the junction there is a buffer of limited capacity. The state of the buffer is represented by a function $q : [0, \infty) \to [0, M^b]$ which is the amount of goods from the incoming arcs waiting to enter in one of the outgoing

arcs with equal probability.
We require that the incoming fluxes are given by

$$\tilde{F}_i(t) = \min\{F_i(u_i(0,t)), \frac{M^b - q(t)}{|\mathcal{I}|}\} \quad i \in \mathcal{I}, \tag{3.3}$$

and the outgoing fluxes by

$$\tilde{F}_j(t) = \begin{cases} \frac{q(t)}{|\mathcal{O}|} & \text{if } (u_j(0,t), F_j(u_j(0,t))) \neq (M_j, 0) \\ 0 \ \text{else} \end{cases} \quad j \in \mathcal{O} \tag{3.4}$$

Conservation of the total number of parts implies that

$$\dot{q}(t) = \sum_{i \in \mathcal{I}} \tilde{F}_i(t) - \sum_{j \in \mathcal{O}} \tilde{F}_j(t). \tag{3.5}$$

Therefore the final model consists of a system of Conservation laws coupled with an Ordinary differential equation

$$\begin{cases} \partial_t u_k(t,x) + \partial_x F(u_k(t,x)) = 0 \quad k \in \mathcal{I} \cup \mathcal{O} \\ \dot{q}(t) = \sum_{i \in \mathcal{I}} \tilde{F}_i(t) - \sum_{j \in \mathcal{O}} \tilde{F}_j(t) \\ u_k(t,0) = u_{k,0}(x) \\ q(0) = c \quad c \geq 0 \end{cases} \tag{JP}$$

4 The Hamilton-Jacobi reformulation

As mentioned in section 2.2, a solution for equation (2.2) is defined not only by the conserved quantity, but also by the correspondent flux. For this reason we need to adapt the classical definition of solution to the case of scalar conservation laws with flux discontinuous in the conserved quantity.

Definition 4.1 Let $(u_0, f_0) \in L^\infty(\mathbb{R}) \times L^\infty(\mathbb{R})$ and $F \in L^\infty_{loc}(\mathbb{R})$ be regulated. We say that a couple (u, f) is a weak solution to the Cauchy problem if

$$\begin{gathered} u \in L^\infty([0,T]; L^\infty(\mathbb{R})), \quad f \in L^\infty([0,T]; L^\infty(\mathbb{R})), \\ f(x,t) \in \mathrm{co}F(u(x,t)) \text{ for a.a } (t,x) \in \mathbb{R}^+ \times \mathbb{R}, \\ \text{with } \mathrm{co}F(u) = [\min\{F(u^-), F(u^+)\}, \max\{F(u^-), F(u^+)\}] \end{gathered}$$

and the identity

$$-\int_0^\infty \int_{-\infty}^\infty u\phi_t dx dt - \int_0^\infty \int_{-\infty}^\infty f\phi_x dx dt = \int_{-\infty}^\infty u_0(x)\phi(0,x) dx$$

holds for all $\phi \in \mathcal{D}(\mathbb{R}^2)$.

Now consider the single Cauchy problem

$$\begin{cases} u_t + F(u)_x = 0 \quad t > 0, x \in \mathbb{R} \\ (u,f)(0,x) = (u_0, f_0)(x) \end{cases} \tag{CP1}$$

where F satisfies (3.1), $u_0 \in L^\infty(\mathbb{R})$ and $f_0(x) = F(u_0(x))$ if $u_0(x) < M$, $f_0(x) \in [0, N]$ if $u_0(x) = M$. The correspondent Hamilton-Jacobi reformulation is given by the following Cauchy problem

$$\text{(CP2)} \qquad \begin{cases} \omega_t + \hat{F}(\omega_x) = 0 \\ \omega(0, x) = \omega_0(x) \end{cases}$$

where $\omega_0(x) = \int_0^x u_0(s)ds$, $\hat{F}(s) = F(s)$ for $0 \le s < M$ and $\hat{F}(M) = \lim_{x \to +\infty} f_0(x)$.

Theorem 4.1 *The Cauchy problem* (CP2) *admits a unique viscosity solution in the sense of Ishii's definition [Ish85]. Moreover if ω is the solution of* (CP2)*, the couple $(\omega_x, -\omega_t)$ is a weak entropy solution of* (CP1)

The previous result is a strong and the key tool to prove the following.

Theorem 4.2 *Let the flux functions $F_k : [0, M_k] \to [0, N_k]$, with $k \in \mathcal{I} \cup \mathcal{O}$, be such that*

$$s \to F_k(s) \text{ smooth on } [0, M), \quad \partial_s^2 F_k \le 0, \quad F(0) = F(M) = 0,$$

$u_0^k \in L^\infty(\mathbb{R})$ and $\|u_0^k\| \le M_k$. The Cauchy Problem JP on the single junction has a unique admissible solution globally defined for all the time.

The solution is obtained as fixed point of a contractive map

$$q \to (\omega_i)_{i \in \mathcal{I}} \to G \to (\omega_j)_{j \in \mathcal{O}} \to \Big(G + \sum \omega_j(0, t) \Big) = \Lambda(q) \tag{4.1}$$

where

- $(\omega_i)_{i \in \mathcal{I}}$ is the vector of solutions of the boundary value problems for the HJ with Hamiltonian F_i
- G is the amount of parts that reach at time t the junction
- $(\omega_j)_{j \in \mathcal{O}}$ is the vector of solutions of the boundary value problems for the HJ with Hamiltonian F_j
- $G + \sum \omega_j(0, t)$ is the amount of parts inside the buffer waiting to enter the arcs j.

References

[AC19] Fabio Ancona and Maria Teresa Chiri, *Conservation laws with transition phase for conveyor belts.* In preparation (2019).

[ADR07] D. Armbruster, P. Degond, and C. Ringhofer, *Kinetic and fluid models for supply chains supporting policy attributes.* Bull. Inst. Math. Acad. Sin. (N.S.) 2 (2007), no. 2, 433–460. MR 2325723.

[AGH11] Dieter Armbruster, Simone Göttlich, and Michael Herty, *A scalar conservation law with discontinuous flux for supply chains with finite buffers.* SIAM J. Appl. Math. 71 (2011), no. 4, 1070–1087. MR 2823493.

[BK08] Raimund Bürger and Kenneth H. Karlsen, *Conservation laws with discontinuous flux: a short introduction.* J. Engrg. Math. 60 (2008), no. 3-4, 241–247. MR 2396483.

[BvcGMSG11] Miroslav Bulíček, Piotr Gwiazda, Josef Málek, and Agnieszka Świerczewska-Gwiazda, *On scalar hyperbolic conservation laws with a discontinuous flux.* Math. Models Methods Appl. Sci. 21 (2011), no. 1, 89–113. MR 2771334.

[DFR05] João-Paulo Dias, Mário Figueira, and José-Francisco Rodrigues, *Solutions to a scalar discontinuous conservation law in a limit case of phase transitions.* J. Math. Fluid Mech. 7 (2005), no. 2, 153–163. MR 2177124.

[DGHP10] Ciro D?Apice, Simone Göttlich, Michael Herty, and Benedetto Piccoli, "Modeling, simulation, and optimization of supply chains". Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2010, A continuous approach. MR 2665143.

[Gim93] Tore Gimse, *Conservation laws with discontinuous flux functions.* SIAM J. Math. Anal. 24 (1993), no. 2, 279–289. MR 1205526.

[HJP13] M. Herty, Ch. Jrres, and B. Piccoli, *Existence of solution to supply chain models based on partial differential equation with discontinuous flux function.* J. Math. Anal. Appl. 401 (2013), no. 2, 510–517. MR 3018001.

[Ish85] Hitoshi Ishii, *Hamilton-Jacobi equations with discontinuous Hamiltonians on arbitrary open sets.* Bull. Fac. Sci. Engrg. Chuo Univ. 28 (1985), 33–77. MR 845397.

Mean field interacting particle systems and games

Guglielmo Pelino (*)

Abstract. Mean field theory studies the behaviour of stochastic systems with a large number of interacting microscopic units. Under the mean-field hypothesis, it is often possible to give a macroscopic easier description of the phenomena, which still allows to catch the main characteristics of the complex pre-limit model. The main purpose of the talk is to motivate a system of two coupled forward-backward partial differential equations, known as the mean field game system, which serves as a limit model for a particular class of stochastic differential games with N players. For reaching this goal, an introductive overview on macroscopic limits for mean field interacting particle systems and games under diffusive dynamics will be presented. In the last part of the talk I will briefly review my contributions in the context of finite state mean field games.

1 Introduction

Originally formulated for applications to physics (more in particular statistical mechanics), mean field models have been since then employed in a wide range of different disciplines such as biology, sociology, economics, finance and computer science. In general, mean field theory deals with large systems of small interacting stochastic units. The basic purpose is to give a macroscopic description (usually deterministic) by studying the limit when the number of units diverges. The study of the resulting limit model allows then to retrieve informations on the pre-limit one. Thanks to the mean-field hypothesis - according to which the effect of all the other individuals on any given individual is approximated by a single averaged effect - it is often possible to perform the above-mentioned macroscopic limit procedure.

The main difference between mean field models for interacting particle systems and games is that in the first ones units follow prescribed laws of motion, thus they have zero-intelligence and we refer to them as particles. In games instead, the microscopic dynamics are controlled, and the interaction is given through some individual optimization procedure (minimization of a cost/ maximization of a reward functional), which puts the units in

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: pelino@math.unipd.it . Seminar held on February 13th, 2019.

competition. Mean field interacting particle systems will serve us as a benchmark for introducing the main techniques for the macroscopic limit procedure.

As stated in the abstract, the purpose of these notes is to justify the following system of two coupled forward-backward PDEs, known as the mean-field game system:

$$\begin{cases} -\partial_t u - \partial_{xx} u + \frac{|\partial_x u|^2}{2} = F(x, m(t)) \quad \text{in } [0,T] \times \mathbb{R}, \\ \partial_t m - \partial_{xx} m - \partial_x (m \partial_x u) = 0 \quad \text{in } [0,T] \times \mathbb{R}, \\ u(T,x) = G(x, m(T)), \ m(0,\cdot) = m_{(0)} \quad \text{in } \mathbb{R}. \end{cases} \tag{1}$$

Note that for simplicity we restricted ourselves to the one-dimensional case, but all the arguments can be extended easily to the multidimensional case, with analogous assumptions on the dynamics.

2 Preliminary notions

We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the underlying probability space in which all the processes we consider are living. The triple is made of a *sample space*, Ω, which models all the possible realizations $\omega \in \Omega$ of the randomness, $\mathcal{F}$, a σ-algebra of subsets of Ω, the collection of *events*, and $\mathbb{P} : \mathcal{F} \to [0,1]$, a probability measure which measures the likelihood of any possible event of $\mathcal{F}$.

A real-valued *random variable* X is a function $X : \Omega \to \mathbb{R}$, which is measurable with respect to the σ-algebra $\mathcal{F}$, by endowing $\mathbb{R}$ with its Borel σ-algebra $\mathcal{B}(\mathbb{R})$, i.e., for any $B \in \mathcal{B}(\mathbb{R})$ we have $X^{-1}(B) \in \mathcal{F}$.

The probability measure $\mathbb{P}$ induces a probability measure on the state space of X, $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, which we refer to as the *law* (or distribution) of the random variable. It is defined by

$$\text{Law}(X)(B) := \mathbb{P}(X \in B) = \mathbb{P}(\omega \in \Omega \ : \ X(\omega) \in B),$$

for any $B \in \mathcal{B}(\mathbb{R})$.

The definition we are aiming to is that of a *stochastic process*. Informally speaking, a stochastic process is a collection of random variables $X(t)$, indexed by some parameter t, which typically represents time. We consider here only continuous time stochastic processes in a finite interval of time, thus $t \in [0,T]$, with $T < \infty$. In order to properly define a stochastic process, we need the concept of filtration. A filtration is a collection of increasing σ-algebras $(\mathcal{F}_t)_{t\in[0,T]}$, such that $\mathcal{F}_t \subseteq \mathcal{F}$ for all $t \in [0,T]$, which carries the information available up to time t. We call $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$ a *filtered probability space*. We are now ready to give the following

Definition 1 (Stochastic process) A continuous-time **stochastic process** with values in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a family of $\mathbb{R}$-valued random variables $X = (X(t))_{t\in[0,T]}$ that are measurable with respect to $\mathcal{F}$. We say that X is **adapted** to the filtration $(\mathcal{F}_t)_{t\in[0,T]}$ if $X(t)$ is measurable with respect to $\mathcal{F}_t$.

The probability measure in the definition of a stochastic process induces a time-varying flow of probability measures on the state space of X. The definition is analogous to the

random variable case: for any $B \in \mathcal{B}(\mathbb{R})$, we set

$$m(t)(B) = \text{Law}(X(t))(B) := \mathbb{P}(X(t) \in B) = \mathbb{P}(\omega \in \Omega \ : \ X(t)(\omega) \in B).$$

We refer to $m(t)$ as the law or distribution of the process X at time t.

The processes we consider enjoy some additional properties: in particular, they are **Markov processes**. A stochastic process X is Markov if, for all $0 \leq s < t$,

$$\mathbb{P}(X(t) \in \cdot | \mathcal{F}_s) = \mathbb{P}(X(t) \in \cdot | X_s).$$

In words, the future of the process depends only on the current state and not on the whole previous history. Even though we do not give any proof, both PDEs which emerge in the mean field game system are derived by making an extensive use of the above property and its consequences.

3 Mean field interacting particle systems

In this section we address the problem of taking the macroscopic limit of a mean field interacting particle system, when at the microscopic level the particles evolve according to a system of interacting diffusion processes.

3.1 Single particle dynamics

A single particle evolves according to a **diffusion process**. A diffusion process is a particular Markov process with continuous paths. For our purposes, it can be defined as a solution to a stochastic differential equation (SDE) of the form

$$\begin{cases} dX(t) = b(t, X(t))dt + \sigma dW(t), \\ X(0) = x_0, \end{cases} \tag{2}$$

where $X(t) : \Omega \to \mathbb{R}$ is the state of the particle, $b : [0,T] \times \mathbb{R} \to \mathbb{R}$ is called *drift* function, $\sigma > 0$ and $(W(t))_{t \in [0,T]}$ is a standard one-dimensional Brownian motion. This equation can be thought of as a random perturbation of a deterministic ODE with vector field b, the strength of the perturbation being regulated by the parameter σ. In (2), the process $(W(t))_{t \in [0,T]}$ is defined as

Definition 2 Given a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in [0,T]}, \mathbb{P})$, we say that $(W(t))_{t \in [0,T]}$ is an $\mathbb{R}$-valued Brownian motion if it is an adapted stochastic process with the following properties:

- $W(0) = 0$ almost surely;
- $W(t) - W(s) \sim \mathcal{N}(0, t-s)$ for all $s < t$;
- $W(t_1) - W(s_1)$ and $W(t_2) - W(s_2)$ are independent for all $s_1 < t_1 \leq s_2 < t_2$;
- almost surely W has continuous paths.

The last property means that there exists an $A \in \mathcal{F}$ with $P(A) = 1$ such that for all $\omega \in \mathcal{F}$, we have that $t \to W(t)(\omega)$ is a continuous function of time. We work under regularity assumptions that guarantee that Equation (2) is well-posed. In our case, globally Lipschitz assumptions and sublinear growth for the drift function are sufficient for obtaining the well-posedeness of (2), analogously to the deterministic case.

3.2 N-particle dynamics

We now consider the N-particle version of Equation (2). Let X^i denote the stochastic process describing the state of the i-th particle in the system, where $i = 1, \dots, N$. The interacting particles system takes the form

$$\begin{cases} dX^i(t) = b(t, X^i(t), m^N(t))dt + \sigma dW^i(t), \\ X^i(0) \sim m_0. \end{cases} \tag{3}$$

In the above equation, the W^i's are N independent Brownian motions. We choose independent and identically distributed initial states $X^i(0)$'s according to a common law m_0.The drift function has an additional argument w.r.t the single particle case. Indeed, we have $b : [0, T] \times \mathbb{R} \times \mathbb{P}(\mathbb{R}) \to \mathbb{R}$, where we endowed $\mathbb{P}(\mathbb{R})$ with the Wasserstein metric. The term $m^N(t)$ is defined as the *empirical measure* of the N particles at time t,

$$m^N(t) := \frac{1}{N} \sum_{j=1}^{N} \delta_{X^j(t)}.$$

Thus, $m^N(t)$ is a random probability measure (the $X^j(t)$'s are random points!), which gives weight $\frac{1}{N}$ to the state of each particle in the system at time t. This form of interaction is called of mean-field type. Indeed, any particle in the system interacts with each other particle only through this macroscopic quantity m^N. This particularly symmetric and weak interaction is what gives us hopes to obtain a macroscopic limit for the model. In this case, the macroscopic limit would be a model for describing the evolution of the *density* of the particles, rather than describing each single microscopic unit in the system. Formally, the limit distribution of the particles is the limit of the sequence of the empirical measures m^N, when $N \to \infty$.

3.3 Macroscopic limit: McKean-Vlasov diffusions

The macroscopic limit of system (3) consists in proving the convergence of the sequence of the empirical measures $m^N(t)$ to some limit distribution $m(t)$, i.e. giving a law of large numbers. Recall that, if we have a sequence of random variables $(\xi_i)_{i=1,\dots,\infty}$ which are i.i.d with mean μ, then, the law of large numbers states that

$$\frac{1}{N} \sum_{i=1}^{N} \xi_i \xrightarrow{N \to \infty} \mu.$$

Thus, randomness is removed in the limit.

When looking for a macroscopic description of System (3), we are aiming to prove the same kind of averaging property

$$m^N(t) \rightharpoonup m(t), \tag{4}$$

for $N \to \infty$. Here, the convergence must be intended in the weak sense of stochastic processes. Specifically, if we think of $(m^N(t))_{t\in[0,T]}$ as a stochastic process with values in $\mathcal{P}(\mathbb{R})$, then we have that, for every continuous bounded function $f \in C_b([0,T];\mathbb{R})$, and for every $t \in [0,T]$,

$$\int_{\mathbb{R}} f(t,x)dm^N(t) \xrightarrow{N\to\infty} \int_{\mathbb{R}} f(t,x)dm(t),$$

almost surely in Ω. It can be shown that, for mean-field interacting particle systems, proving the law of large numbers (4) is equivalent to proving the *asymptotic indendepence* between the particles. This last property is known as *propagation of chaos*:

Definition 3 (Propagation of chaos)) Assume $(X^i(0))_{i=1,\dots,N} \sim m_0$ i.i.d. We say that System (3) propagates chaos if, for any $k \geq 1$, any k-uplet $(i_1,\dots,i_k)$ and any $0 < t \leq T$

$$\text{Law}(X^{i_1}(t),\dots,X^{i_k}(t)) \xrightarrow{N\to\infty} \text{Law}(X^{i_1}(t)) \otimes \cdots \otimes \text{Law}(X^{i_k}(t)).$$

The macroscopic limit of System (3) is thus described by an infinite number of particles $(X^i(t))_{i=1,\dots,\infty}$, which are all independent and identically distributed according to the limit distribution $m(t)$ in (4). Since they are all i.i.d, we can choose to describe one single *reference* particle, whose dynamics is thus

$$\begin{cases} dX(t) = b(t,X(t),m(t))dt + \sigma dW(t), \\ X(0) \sim m_0, \end{cases} \tag{5}$$

with $m(t) = \text{Law}(X(t))$. Equation (5) is a non standard SDE. In the literature it is referred to as McKean-Vlasov diffusion. It is non-linear in the sense that the solution is dependent on the law of the process itself. Another consequence of the law of large numbers (4) for System (3) is that, when the limit distribution $m(t)$ is regular enough to admit a density $m(t,x)$ (i.e. $m(t) = m(t,x)dx$), the density function $m(t,x)$ satisfies the Fokker-Planck PDE:

$$\begin{cases} \partial_t m(t,x) = -\partial_x \left[b(t,x,m(t))m(t,x)\right] + \frac{\sigma^2}{2}\partial_{xx} m(t,x), & \text{in } [0,T]\times\mathbb{R} \\ m(0) = m_0, & \text{in } \mathbb{R}. \end{cases} \tag{6}$$

The above PDE describes the evolution in time of the density of the infinite particles in the limit. It is precisely the macroscopic "easier" description of the phenomena which we referred to in the Introduction. This equation is one of the two equations appearing in the MFG system (1).

4 Mean field games

In this section we mimic the structure of the previous one, by adding in the story the control component. We introduce first a linear quadratic optimal control problem, our 1-player game, then we consider the N-player version of it. Under the mean-field assumptions on the cost functionals, we then address the macroscopic limit, describing the limit configuration of Nash equilibria through the mean field game system (1).

4.1 Stochastic control theory: "1-player" games

We modify the SDE (2) according to

$$dX^\alpha(t) = \alpha(t, X^\alpha(t))dt + \sigma dW(t), \tag{7}$$

where now $\alpha : [0,T] \times \mathbb{R} \to \mathbb{R}$ is the so-called *feedback control function*. It is chosen to minimize a cost functional of the form

$$J(\alpha) = \mathbb{E}\left[\int_0^T \left(\frac{|\alpha\left(s, X^\alpha(s)\right)|^2}{2} + F(X^\alpha(s))\right) ds + G(X^\alpha(T))\right]. \tag{8}$$

For every possible choice of control α, we have a different SDE (7), whose state is denoted by $X^\alpha(t)$ to explicitly highlight the dependence of the solution on the control function. We only allow controls which are such that (7) is a well-posed equation, and such that the cost is finite. The cost functional is made of three terms:

- $\frac{|\alpha(s)|^2}{2}$, which penalizes high velocities;
- $F : \mathbb{R} \to \mathbb{R}$ a running cost;
- $G : \mathbb{R} \to \mathbb{R}$ a terminal cost, paid when exiting the game at time T.

The optimal control problem just formulated is of linear-quadratic type, since the control function appears linearly in the dynamics (7), and quadratically in the cost functional (8). When F and G are quadratic in X themselves, a typical condition for the *admissible* controls is given by

$$\mathbb{E}\left[\int_0^T \frac{|\alpha\left(s, X^\alpha(s)\right)|^2}{2} ds\right] < \infty,$$

together with α globally Lipschitz in the x-variable and of sublinear growth. In this context, we know that there exists a unique solution to the control problem. It can be found by defining the function

$$v(t,x) := \min_\alpha \mathbb{E}^{t,x}\left[\int_t^T \left(\frac{|\alpha(s, X^\alpha(s))|^2}{2} + F(X^\alpha(s))\right) ds + G(X^\alpha(T))\right], \tag{9}$$

where the expectation is conditioned on dynamics $(X^\alpha(s))_{t \le s \le T}$ starting at time t in $X^\alpha(t) = x$. We call v the *value function* of the optimal control problem. A basic principle in optimal control problems allows us to deduce an equation for the value function. Such

principle is known as *dynamic programming*, according to which we can recursively reconstruct the value function by going backward in time. In our context, if we fix any $h > 0$, it establishes that

$$v(t,x) = \min_{\alpha} \mathbb{E}\left[v(t+h, X^{\alpha}(t+h)) + \int_t^{t+h} \left(\frac{|\alpha(s)|^2}{2} + F(X^{\alpha}(s)) \right) ds \right].$$

If we now differentiate with respect to time the above expression, and apply It's formula to $\frac{d}{dt}\mathbb{E}\left[v(t, X^{\alpha}(t))\right]$, we get

$$\begin{cases} -\partial_t v - \frac{\sigma^2}{2}\partial_{xx} v + \frac{|\partial_x v|^2}{2} = F(x), & \text{in } [0,T] \times \mathbb{R} \\ v(T,x) = G(x), & \text{in } \mathbb{R}. \end{cases} \tag{10}$$

In order to close the argument, one is left to show what in literature is known as *Verification Theorem*: it establishes that, if one is able to solve the Hamilton-Jacobi-Bellman equation (10), then the solution provides the value function of the optimal control problem, and the optimal feedback function is given by

$$\alpha^*(t,x) = -\partial_x v(t,x).$$

Finally, we remark that the Hamilton-Jacobi-Bellman equation is the first of the two equations appearing in the mfg system (1), except for the coupling (the term on the right-hand side of the equality) which is different from Equation (10). More on this will come shortly.

4.2 N-player games

We now consider a game of the same type as before, but for a system of N controlled diffusions. The state of the i-th player is denoted by $X^i(t)$, for $i = 1, \dots, N$, and the dynamics are given by the system

$$\begin{cases} dX^i(t) = \alpha^i(t, \boldsymbol{X}(t))dt + \sigma dW^i(t), \\ X^i(0) \sim m_0, \end{cases} \tag{11}$$

where $\boldsymbol{X}(t) := (X^1(t), \dots, X^N(t))$ is the vector representing the state of the whole population, and $\alpha^i : [0,T] \times \mathbb{R}^N \to \mathbb{R}$ is the feedback control function for the i-th player. Observe that α^i now depends on the whole state of the population at time t. We associate to each player a cost functional J^i, of the form

$$J^i(\alpha^1, \dots, \alpha^N) := \mathbb{E}\left[\int_0^T \left(\frac{\left|\alpha^i(s)\right|^2}{2} + F\left(X^i(s), m^{N,i}(s)\right) \right) ds + G\left(X^i(T), m^{N,i}(T)\right) \right], \tag{12}$$

where

$$m^{N,i}(t) := \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} \delta_{X^j(t)}$$

is the empirical measure of the $N-1$ players except for the i-th. In this way, the interaction among players is of mean-field type. Observe that the cost functionals J^i's now depend on the strategies chosen by every player. In this context, it is not straightforward to even give a definition of solution for the N-player game. For our purposes, we are interested in one particular concept of equilibrium for the game, the so-called Nash equilibrium.

Definition 4 A strategy vector $\boldsymbol{\alpha}^* := (\alpha^{1,*}, \dots, \alpha^{N,*})$ is said to be a Nash equilibrium for the N-player game if, for each $i = 1, \dots, N$

$$J^i(\boldsymbol{\alpha}^*) = \min_{\beta} J^i([\beta; \boldsymbol{\alpha}^{*,-i}]),$$

with

$$[\beta; \boldsymbol{\alpha}^{*,-i}]_j := \begin{cases} \alpha^{*,j}, & j \neq i \\ \beta, & j = i. \end{cases}$$

In words, $\alpha^{*,i}$ is the optimal control for the i-th player, when the other $N-1$ players play the strategies $\boldsymbol{\alpha}^{*,-i}$.

With this concept of solution in mind, one can mimic the strategy shown in the previous section for the 1-player case for constructing an equilibrium in the N-player version of the model:

- define $v^{i,N}(t, \boldsymbol{x}) := \min_\beta J^i([\beta; \boldsymbol{\alpha}^{*,-i}])$ the value function of the i-th player, computed minimizing the choices for the i-th player's control, freezing the others' controls in the Nash equilibrium, and conditioning to have initial states at time t given by $\boldsymbol{X}(t) = \boldsymbol{x}$;
- by dynamic programming, the functions $v^{i,N}$'s solve a system of N coupled HJB equations;
- $\alpha^{i,*}(t, \boldsymbol{x}) = -\partial_{x_i} v^{N,i}(t, \boldsymbol{x})$ provides the Nash equilibrium by a Verification Theorem.

The system of N coupled HJB equations for the value functions, which we call Nash system, is given by

(13)
$$\begin{cases} -\partial_t v^{N,i}(t, \boldsymbol{x}) - \sum_{j=1}^N \partial_{x_j x_j} v^{N,i}(t, \boldsymbol{x}) + \frac{\left|\partial_{x_i} v^{N,i}(t, \boldsymbol{x})\right|^2}{2} + \sum_{j \neq i} \partial_{x_j} v^{N,j} \partial_{x_j} v^{N,i} = F(x_i, m^{N,i}), \\ v^{N,i}(T, \boldsymbol{x}) = G(x_i, m^{N,i}). \end{cases}$$

In general, the above system becomes highly not tractable for N big, even from a numerical point of view. However, the mean-field interaction, expressed in the cost functions F and G, generates a first *symmetrization* in the value functions, which can be proved to be of the form:

$$v^{N,i}(t, \boldsymbol{x}) = v^N(t, x_i, m^{N,i}). \tag{14}$$

Accordingly, the Nash equilibrium satisfies the same symmetric properties

$$\alpha^{i,*}(t, \boldsymbol{x}) = \alpha^{*,N}(t, x_i, m^{N,i}) = -\partial_{x_i} v^N(t, x_i, m^{N,i}).$$

4.3 Macroscopic limit: the mean field game system

The symmetric properties (14) of the value functions hint that we have a chance to observe a mean-field limit for the system (13). If we look at the optimal dynamics of the N players when each of them is in the Nash equilibrium, we have the following system of SDEs

$$dX^i(t) = \alpha^{*,N}(t, X^i(t), m^{N,i}(t))dt + \sigma dW^i(t). \tag{15}$$

If we compare it with the mean-field system of interacting particles given in (3) they look quite similar at first glance. The additional difficulty in obtaining a macroscopic limit for (15) is that the drift function - $\alpha^{*,N}$ - now depends explicitly on N, while in (3) it was fixed to a function b. Thus, the symmetries expressed in (14) may not suffice to apply standard results from the theory of propagation of chaos. The additional challenge here is to prove that the sequence of the value functions v^N admits some limit. A breakthrough was achieved with [2], where the authors obtained a rigorous convergence of the sequence v^N to a limit function U, which solves an infinite-dimensional PDE on the space of probability measures. Without giving any other detail, we here assume we can prove a law of large numbers/propagation of chaos for the empirical measures of the N players, with the heuristic motivation that when N grows one player influences the others less and less. Thus,

$$m^{N,i}(t) \rightharpoonup m(t) \in \mathcal{P}(\mathbb{R}),$$

for $N \to \infty$. In the limit configuration, we end up with an infinite number of players $(X^i(t))_{i=1,\dots,\infty}$, which are all independent and identically distributed. The *reference* player in this case optimizes its strategy by considering the distribution $m(t)$ of the other infinite players to be *fixed*. Thus, given $m(t)$, the reference player chooses α^*, the optimal control, which must coincide with the limit of the Nash equilibrium sequence $\alpha^{*,N}$. At the same time though, the rest of the population is also in the Nash equilibrium: we must then have that the distribution of the reference player in the optimal dynamics coincides with that of the rest of the population. In formulae, we have the mean-field condition on the optimal dynamics

$$\text{Law}(X^{\alpha^*}(t)) = m(t).$$

The above argument can be summarized by finding a fixed point of a map

$$m(t) \to u(t) \to \alpha^* \to X^{\alpha^*}(t) \to \tilde{m}(t),$$

where m is the distribution of the other players, u is the value function of the reference player and α^* its optimal control, and $\tilde{m}$ is the distribution of the optimal dynamics of the reference player. A mean-field game equilibrium is such that we have $m(t) = \tilde{m}(t)$ for every t. This fixed point map can be reformulated via the system of two PDEs (1) which we introduced in the beginning, which we here restate for clarity:

$$\begin{cases} -\partial_t u - \partial_{xx} u + \frac{|\partial_x u|^2}{2} = F(x, m(t)) \quad \text{in } [0,T] \times \mathbb{R}, \\ \partial_t m - \partial_{xx} m - \partial_x(m \partial_x u) = 0 \quad \text{in } [0,T] \times \mathbb{R}, \\ u(T,x) = G(x, m(T)),\ m(0,\cdot) = m_{(0)} \quad \text{in } \mathbb{R}. \end{cases} \tag{16}$$

The first of the two equations governs the optimal choice of the reference player *given* the distribution m of the other infinite players, by describing the backward in time evolution of its value function u. The second one is instead explaining how the distribution of players evolves with time. At the same time such distribution coincides with the distribution of the reference player itself (look at the u appearing in the equation for m).

5 Finite state mean field games and perspectives

When the processes describing the dynamics of the players take value in a finite space $\Sigma := \{1, \dots, d\}$ we talk about finite state mean field games. The processes describing the evolution of the players are *continuous-time Markov chains*, and the controls are the transition rates among the possible states. Specifically, we have that $\alpha^i_y(t, x, \boldsymbol{x}^{N,i})$ represents the rate at which player i decides to go from state x to state y, when $x \neq y$, $\boldsymbol{x}^{N,i}$ being the states of the other $N-1$ players at time t. In formulae

$$\mathbb{P}\left[X^i(t+h) = y | X^i(t) = x, \boldsymbol{X}^{N,i}(t) = \boldsymbol{x}^{N,i}\right] = \alpha^i_y(t, x, \boldsymbol{x}^{N,i})h + o(h).$$

For these kinds of models, results were obtained in two opposite scenarios:

(a) *Uniqueness* scenario: when the mean field game system has a unique, regular solution;

(b) *Non-uniqueness* scenario: when the mean field game system possesses multiple solutions.

In both cases, the focus was on stuyding the convergence of the N-player Nash equilibrium and dynamics to the limiting mean field game configuration(s).

The uniqueness scenario is analyzed in [3], which we refer to for details. Here we make only a list of the most important results proved:

- the rigorous convergence of the (unique) Nash equilibrium to the limiting (unique) solution to the MFG;
- a Law of Large Numbers for the empirical measures of the players;
- refined asymptotics for the latter: a Central Limit Theorem and a Large Deviation Principle.

For ensuring uniqueness of the limit mfg system, we employed the so-called *monotonicity* conditions on the costs: in the context of finite state space mean field games these read as, for any $m, m' \in P(\Sigma)$,

$$\sum_{x \in \Sigma} (F(x, m) - F(x, m'))(m(x) - m'(x)) \geq 0,$$

and the same for the cost function G. It is not hard to see that this condition implies - at least for local mean field games - that it is less costly for players to occupy states of

the space with a low density of players: essentially, players prefer to spread rather than to aggregate.

The non-uniqueness scenario (2) is instead treated in [4], where we restricted to a binary state space $\{-1, 1\}$, in order to have explicit computations for the solutions of the mean field game. Moreover, we considered $F \equiv 0$ and an anti-monotonic final cost G, meaning that here players tend to favor the aggregation with the state of the majority, but at the same time they are in a non-cooperative setting, thus making highly non-trivial to guess what happens at the limit level. Still, at the level of the N-player game we have existence and uniqueness of the Nash equilibrium. In [4] we proved:

- existence of multiple solutions to the MFG system;
- convergence of the (unique) Nash equilibrium to *one* solution of the MFG (the entropy solution to a conservation law, the master equation);
- the other solutions are *almost* Nash equilibria (ε-Nash, with $\varepsilon \xrightarrow{N\to\infty} 0$);
- Law of Large Numbers for initial values of the empirical measure outside an "*indecision*" point.

Finally, a perspective work aims at weakening the mean-field assumption. As stated in the introduction, the interaction graph in the mean-field case is a complete graph. Can we erase some connections and still hope to retrieve the same mean-field limit in the context of games? In particular, the goal of this future work is to consider the case of Erdos-Renyi graphs, where the edges between the nodes are all i.i.d and each of them has a probability p_N to exist and $1 - p_N$ to not be present. For N-player games on this class of graphs, we are interested in constructing approximate Nash equilibria by using the mean field limit configuration which one would find if the graph was complete under mean-field interaction, where the error in the approximation tends to 0 with N going to infinity. In particular, we want to allow p_N to go to 0 with N. Presumably, in order to find the same mean field limit one should have that p_N goes to zero at most with a speed such that $Np_N \to \infty$ for $N \to \infty$. Indeed, Np_N is the average degree in the Erdos-Renyi graph, and thus if the average degree tends to infinity we are likely to retrieve a mean-field effect on the game. The mean-field interacting particle system counterpart of this result is analyzed in [6], where they prove that the condition on the diverging average degree is indeed necessary for getting a mean-field limit.

References

[1] P. Cardaliaguet, "Notes on mean field games". Technical report, Université de Paris - Dauphine, September 2013.

[2] P. Cardaliaguet, F. Delarue, J.-M. Lasry, and P.-L. Lions, *The master equation and the convergence problem in mean field games.* `arXiv:1509.02505 [math.AP]`, September 2015.

[3] A. Cecchin and G. Pelino, *Convergence, fluctuations and large deviations for finite state mean field games via the master equation.* Stochastic Processes and their Applications, pre-published online, `doi.org/10.1016/j.spa.2018.12.002`, December 2018.

[4] A. Cecchin, P. Dai Pra, M. Fischer, and G. Pelino, *On the convergence problem in Mean Field Games: a two state model without uniqueness.* SIAM Journal on Control and Optimization (2019), to appear.

[5] J.-M. Lasry and P.-L. Lions, *Mean field games.* Japan. J. Math. 2/1 (2007), 229–260.

[6] R.I Oliveira, G. Reis, *Interacting diffusions on random graphs with diverging degrees: hydrodynamics and large deviations.* `https://arxiv.org/abs/1807.06898`, 2018.

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

On the Alexander polynomial of line arrangements in $\mathbb{P}^2$

Federico Venturelli (*)

Abstract. In [M] Milnor proved that to any homogeneous polynomial f in $n+1$ indeterminates one can associate the smooth locally trivial fibration $f : \mathbb{C}^{n+1} \setminus f^{-1}(0) \to \mathbb{C}^*$; the $\mathbb{C}$-linear automorphism induced by the geometric monodromy of the generic fibre F on the cohomology group $H^i(F, \mathbb{C})$ is called i-th algebraic monodromy of f, and its characteristic polynomial is called i-th Alexander polynomial of the projective hypersurface $V := V(f) \subset \mathbb{P}^n$. When V is smooth these Alexander polynomials are known by the work of Brieskorn [B]; however, much less is known already when V has only isolated singularities. In these notes we will present some results concerning the simplest case one can focus on in the latter situation: the computation of the first Alexander polynomial when $f \in \mathbb{C}[x, y, z]$ factors into linear homogeneous polynomials, i.e. when $V \subset \mathbb{P}^2$ consists of a collection of lines.

Organization of the text

In the first part of these notes we review some notions that are necessary in order to state the definition of the (first) Alexander polynomial: in particular, we show how the homotopy lifting property of fibrations allows us to define the geometric monodromy of F, and we recall the definition of (de Rham) cohomology groups with complex coefficients for complex manifolds. We then introduce the intersection lattice of a hyperplane arrangement and state a long-standing conjecture on the Alexander polynomial of non-central line arrangements in $\mathbb{P}^2$.

The second section is devoted to the presentation of a formula, due to Libgober (see [L1]), that allows for the explicit computation of the Alexander polynomial of a plane projective curve. Its introduction is justified by the fact that computing this polynomial using the definition is in most cases unfeasible, and it will enable us to state (and understand) a classical result of Zariski on the Alexander polynomial of a very particular plane curve. By using this formula, the interested reader will be able to verify the following (perhaps surprising) fact: the vast majority of line arrangements one can draw on paper have trivial

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: `venturel@math.unipd.it`. Seminar held on February 27th, 2019.

Alexander polynomial, where by 'trivial' we mean 'of the form $(t-1)^{r-1}$ where r is the number of lines'.

If one compares this formula with the expression of the Alexander polynomial of a non-central line arrangement suggested by the conjecture, the latter may seem quite mysterious; however, once we introduce the notion of k-net on a line arrangement we will see that some of this mystery might be explained. The highlight of the last part of this section is a recent result by Papadima and Suciu on some classes of arrangements that admit 3-nets (see [PS]), that seems to suggest that highly symmetric line arrangements are the only ones having a non-trivial Alexander polynomial.

In the last section we will first recall some notions from complex geometry (basic Hodge theory and the definition of deformation) and then proceed to illustrate an original result: namely, we will present a very specific class of line arrangements whose Alexander polynomial is trivial. The link between these two topics is provided by a result of Libgober (see [L2]) that relates the Alexander polynomial of a plane curve C to the irregularity of a surface S associated to C.

1 Explaining the title

In the same way vector bundles $\pi : E \to B$ encode the information of a family of vector spaces (the fibres of E_b of π) parametrised by a base space B, fibrations $\pi : X \to Y$ describe families of topological spaces parametrised by a base topological space Y; however, fibrations enjoy an additional property, called homotopy lifting property (HLP in the following). The exact statement of the HLP is rather involved, so we will only explain why it is important to us and what it allows us to do. But first, we define properly what a fibration is:

Definition 1.1 A fibration is a continuous surjective map $\pi : X \to Y$ between topological spaces that satisfies the HLP; it is called *locally trivial* if for any open subset U of Y the preimage $\pi^{-1}(U)$ is homeomorphic to $U \times Z$ for some topological space Z.

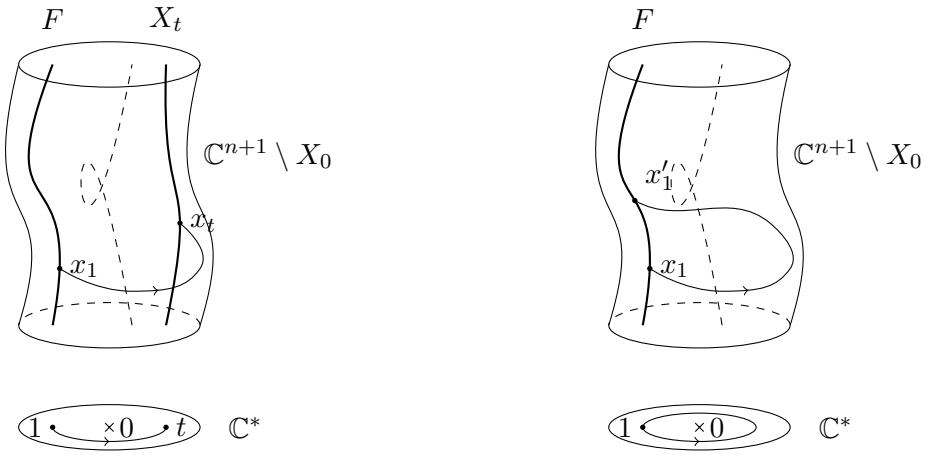
Assume now $f \in \mathbb{C}[x_0, \dots, x_n]$ is a homogeneous polynomial. For any $t \in \mathbb{C}$ we can define X_t as the zero-locus in $\mathbb{C}^{n+1}$ of the polynomial $f - t$ (sets like these are called *affine hypersurfaces*). If we want to think of all the X_t's as a continuous family depending on the parameter t, we can consider the polynomial map associated to f, i.e. $f : \mathbb{C}^{n+1} \to \mathbb{C}$ s.t. $\underline{x} \mapsto f(\underline{x})$; clearly $X_t = f^{-1}(t)$. An important theorem by Milnor (see [M]) tells us the following:

Theorem 1.1 *The restricted map $f : \mathbb{C}^{n+1} \setminus X_0 \to \mathbb{C}^*$ is a smooth locally trivial fibration.*

By *smooth* we mean that all the fibres, i.e. all the X_t's with $t \neq 0$, are smooth affine hypersurfaces (in particular we can think of them as complex manifolds) and that they are all diffeomorphic as differentiable manifolds. The removal of the fibre X_0 is clearly necessary: using the Jacobian Criterion, one can verify that the affine variety $V(f) \subset \mathbb{C}^{n+1}$ associated to a homogeneous polynomial is always singular at the origin. It is customary to denote the fibre X_1 by F, and to call it the *Milnor fibre* of f.

Now, why is the HLP so important? Fix a point in $\mathbb{C}^*$, for example (and for simplicity) 1, and consider any path $\alpha : [0,1] \to \mathbb{C}^*$ going from 1 to t: if we fix an element x_1 in the fibre $X_1 = F$, we can 'lift' the path α to a path α' in $\mathbb{C}^{n+1} \setminus X_0$ going from x_1 to some element $x_t \in X_t$. The HLP guarantees that this lifted path α' is unique up to homotopy: if β is a path in $\mathbb{C}^*$ different from α going from 1 to t that can be 'continuously deformed onto α', then the lifted path β' in $\mathbb{C}^{n+1} \setminus X_0$ is a path from $x_1 \in F$ to $x_t \in X_t$ that can be continuously deformed onto α'.

If the path α we choose is actually a loop ($\alpha(0) = \alpha(1) = 1$) then x_t is again an element of F (denote it by x_1'), but we will not always have $x_1 = x_1'$. The situation is the following: there is an action, called *monodromy action*, of the fundamental group $\pi_1(\mathbb{C}^*, 1)$ on the Milnor fibre F, or, equivalently, a representation $\rho : \pi_1(\mathbb{C}^*, 1) \to Aut(F)$. In particular, to a generator α of $\pi_1(\mathbb{C}^*, 1)$ we can associate an automorphism $h : F \to F$ of the Milnor fibre, called the *geometric monodromy* of F (or of f).



The lifting of a path and of a loop, respectively. The dashed line represents the removed singular fibre X_0.

We now need to recall the notion of (de Rham) cohomology groups of a complex manifold M. These are $\mathbb{C}$-vector spaces $H^i(M, \mathbb{C})$ defined as the quotient of the set of the closed complex-valued differentiable i-forms on M by the set of the exact complex-valued differentiable i-forms on M. For $i = 1$, the case we are mainly interested in, we get 1-forms i.e. expressions like this

$$\omega = \sum_{i=1}^{n} f_i(\underline{z}) dz_i + \sum_{i=1}^{n} g_i(\underline{z}) d\overline{z}_i \tag{1.1}$$

where $n := dim(M)$, f_i and g_i are complex-valued differentiable functions, the z_i's are holomorphic coordinates for M and dz_i,$d\overline{z}_i$ satisfy the formal properties of the differential. While their definition seems to be related to calculus, these cohomology groups do reflect geometric properties of M, as the following example shows.

Example 1 $H^1(\mathbb{C}^*, \mathbb{C})$ is the 1-dimensional $\mathbb{C}$-vector space generated by the class of the 1-form $\omega := \frac{dz}{z}$; on the other hand $H^1(\mathbb{C}, \mathbb{C}) = 0$ since all 1-forms on $\mathbb{C}$ are exact. In this case, the first cohomology group 'detects' the fact that $\mathbb{C}$ is simply connected while $\mathbb{C}^*$ is not.

It is a standard fact that an automorphism of a complex manifold M induces an automorphism of the spaces $H^i(M, \mathbb{C})$ for all i.

Now, to sum things up: if f is a homogeneous polynomial in $n+1$ indeterminates we can associate to it:

- Its zero-locus $V := V(f)$ in the projective space $\mathbb{P}^n$ (these sets are called *projective hypersurfaces*).
- The Milnor fibre F arising from the fibration $f : \mathbb{C}^{n+1} \setminus X_0 \to \mathbb{C}^*$.
- The geometric monodromy $h : F \to F$ and the induced automorphism $T : H^1(F, \mathbb{C}) \to H^1(F, \mathbb{C})$, called (first) *algebraic monodromy* of F.

We have explained all the notions we need in order to define the (first) Alexander polynomial:

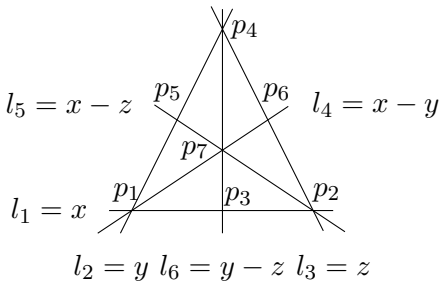
Definition 1.2 The (first) *Alexander polynomial* of the projective hypersurface V (of the polynomial f) is the characteristic polynomial Δ_V of T. We will refer to it simply as to the Alexander polynomial.

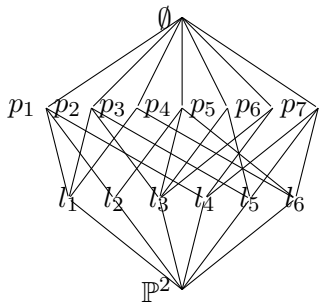
One can prove that the geometric monodromy is given by $h(x_0, \dots, x_n) = (\eta_d x_0, \dots, \eta_d x_n)$ where d is the degree of f and η_d is a primitive d-th root of unity. In particular h is unipotent (after applying h for d times we get the identity), so T is unipotent too; as a consequence we get that T is diagonalizable and Δ_V is a product of cyclotomic polynomials Φ_k with k dividing d.

We can now turn our focus to *hyperplane arrangements*: they are defined as finite collections $\mathcal{A} = \{H_1, \dots, H_r\}$ of codimension 1 linear subspaces of some $\mathbb{K}^n$, where $\mathbb{K}$ is any field. The number of questions that have been raised about such objects is huge, and so is the literature (see [D] for a recent introductory textbook on the topic); in order to study our problem, we will specialize the definition by requiring that the H_i are codimension 1 vector subspaces of $\mathbb{C}^3$ (so $\underline{0} \in H_i$ for all i). Setting $n = 3$ is actually not a critical change, but requiring that the H_i be vector spaces is: it allows us to associate to the arrangement $\mathcal{A}$ consisting of affine planes in $\mathbb{C}^3$ its projectivization, which is an arrangement of lines in $\mathbb{P}^2$. The latter objects are the ones we care about, and we will indicate them again by $\mathcal{A}$; they can be described as the zero-locus in $\mathbb{P}^2$ of a homogeneous polynomial f that factors into linear homogeneous polynomials; in particular they are projective hypersurfaces, so we can speak of their Alexander polynomial $\Delta_{\mathcal{A}}$.

To a line arrangement $\mathcal{A}$ we can associate the intersection lattice $L(\mathcal{A})$, which keeps track of the incidence relations between the various lines of $\mathcal{A}$. In the figure below we present the line arrangement that we will use as main example throughout these notes,

and the associated intersection lattice: this arrangement is usually referred to as the A_3 arrangement, and it can be realized by the polynomial $f = xyz(x-y)(x-z)(y-z)$:





There is a long-standing conjecture on the Alexander polynomial of line arrangements:

Conjecture 1.2 *The Alexander polynomial of a non-central line arrangement $\mathcal{A} \subset \mathbb{P}^2$ consisting of d lines has the form*

$$\Delta_{\mathcal{A}}(t) = (t-1)^{d-1}(t^2+t+1)^a[(t+1)(t^2+1)]^b = \Phi_1(t)^{d-1}\Phi_3(t)^a[\Phi_2(t)\Phi_4(t)]^b \tag{1.2}$$

and the exponents a, b are determined by $L(\mathcal{A})$.

By 'non-central' we mean that not all lines of $\mathcal{A}$ pass through the same point.

Remark 2 We have seen that for arrangements $\mathcal{A}$ of d lines $\Delta_{\mathcal{A}}$ is a product of factors $\Phi_k^{\alpha_k}$ where Φ_k is the k-th cyclotomic polynomial, $\alpha_k \geq 0$ and k divides d; as a consequence, one would expect the 'complexity' (number of non-trivial cyclotomic factors) of $\Delta_{\mathcal{A}}$ to increase with the number of divisors of d. Formula (1.2) suggests instead that the only divisors of k that matter are 1, 2, 3 and 4: but why should line arrangements of, say, $35 = 5 \cdot 7$ lines, which are arguably more complicated than those with, say, $6 = 2 \cdot 3$ lines, have an Alexander polynomial that is simpler than the one of the latter arrangements? A possible explanation is suggested in the next Section.

2 A formula for $\Delta_{\mathcal{A}}$ and the importance of symmetry

Since the Alexander polynomial is the characteristic polynomial of an automorphism of a vector space, we have a 'standard' way to compute it: find a basis $\mathcal{B}$ for $H^1(F, \mathbb{C})$, write the matrix $M_{\mathcal{B}}$ of T associated to $\mathcal{B}$ and compute the determinant of $M_{\mathcal{B}} - t \cdot Id$; however, finding a basis for $H^1(F, \mathbb{C})$ is in general a difficult task, unless F is quite simple.

Example 3 Pick $f = xy \subset \mathbb{C}[x, y, z]$ and let $\mathcal{A} = V(f) \subset \mathbb{P}^2$ (line arrangement consisting of two incident lines), then F is given by $V(xy-1) \subset \mathbb{C}^3$; we can compute $H^1(F, \mathbb{C})$ using the holomorphic de Rham complex of F (for those who know: because affine varieties are Stein spaces): if we call $S := \mathbb{C}[[x, \frac{1}{x}, z]]$ it reads

$$0 \to S \xrightarrow{d^0} Sdx \oplus Sdz \xrightarrow{d^1} Sdx \wedge dz \xrightarrow{d^2} 0$$

$$f \longrightarrow (\tfrac{df}{dx}dx, \tfrac{df}{dz}dz)$$

$$(gdx, hdz) \longrightarrow (\tfrac{dh}{dx} - \tfrac{dg}{dz})dx \wedge dz$$

We have $H^1(F, \mathbb{C}) = Ker(d_1)/Im(d_0) = \mathbb{C}\cdot[\frac{dx}{x}]$. Since the degree of f is 2 the geometric monodromy h is described by $h(x, \frac{1}{x}, z) = (-x, -\frac{1}{x}, -z)$, so we get $T([\frac{dx}{x}]) = [\frac{d(-x)}{(-x)}] = [\frac{dx}{x}]$ i.e. $M_{\mathcal{B}} = Id$ and $\Delta_{\mathcal{A}}(t) = t - 1$.

An alternative way to compute the Alexander polynomial is provided by a formula by Libgober, which we are going to describe and explain now. This is where the exposition will get (even) sketchier, since in order to avoid using the language and formalism of sheaves (and schemes) we will need to give some ad hoc definitions; the reader who wishes to know how things work should consult [L1] and [L2].

Let $C \subset \mathbb{P}^2$ be any curve, and let $\Sigma = \{p_1, \dots, p_r\}$ be the set of its singular points. To any p_i in Σ and to any germ of holomorphic function ϕ around p_i one can associate a non-negative rational number k_ϕ which, when positive, is called *constant of quasi-adjunction* of p_i relative to the function germ ϕ. For line arrangements any singular point p is an ordinary multiple point i.e. it has a local equation of the type $x^m - y^m = 0$ for some $m \in \mathbb{N}$; the constants of quasi-adjunction of p are summarized in the following table (holomorphic functions are analytic, so we can restrict to computing k_ϕ when ϕ is a monomial function germ):

Functions ϕ around p	Corresponding k_ϕ
Constants	$\frac{m-2}{m}$
x, y	$\frac{m-3}{m}$
x^2, xy, y^2	$\frac{m-4}{m}$
$\dots$	$\dots$
Monomials of degree $m-3$	$\frac{1}{m}$

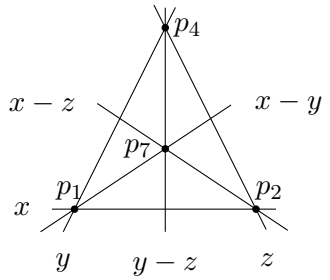
We can see immediately that for a line arrangement $\mathcal{A}$ the singular points that admit constants of quasi-adjunction are those whose order is 3 or bigger; we denote this set of points by $\Sigma' := \{p_1, \dots, p_s\}$. We are now ready to give two definitions:

(a) Let $k \in \mathbb{R}$ and $i = 1, \dots, s$. We denote by V_k^i the vector space of holomorphic function germs around p_i whose constant of quasi-adjunction is bigger than or equal to k (and set $V_k^i := \{0\}$ if there are none); we denote by Z_k the set of the finitely many couples (p_i, V_k^i), and define its *length* $l(Z_k)$ as the sum of the dimensions of the V_k^i.

(b) Let $k \in \mathbb{R}$ and $i = 1, \dots, s$. Denote by $J_i \subset \mathbb{C}[x, y, z]$ the ideal of the point $p_i \in \Sigma'$ (the set of all polynomials in $\mathbb{C}[x, y, z]$ vanishing on p_i). We denote by e_k^i the maximum degree of monomial function germs in V_k^i plus one if $V_k^i \neq \{0\}$, and set

$e_k^i := 0$ if $V_k^i = 0$; we also set $I_k := \cap_{i=1,\dots,s} J_i^{e_k^i}$. The *Hilbert function* of I_k in degree t is $h_{I_k}(t) := dim_{\mathbb{C}}(\mathbb{C}[x,y,z]/I_k)_t$.

This is probably rather confusing, so let us try and clarify things with an example. If we consider the A_3 arrangement we have $\Sigma' = \{p_1, p_2, p_4, p_7\}$ and $J_1 = (x, y)$, $J_2 = (x, z)$, $J_4 = (y, z)$, $J_7 = (x - y, x - z)$.



Since all points of Σ' have order 3 they admit only one constant of quasi-adjunction, namely $\frac{1}{3}$ (which is relative to constant function germs); this means we only have two possibilities:

- If $k \leq \frac{1}{3}$ we get $V_k^i = \mathbb{C}$ and $e_k^i = 1$ for all i, so $l(Z_k) = 4$ and $I_k = J_1 \cap J_2 \cap J_4 \cap J_7$.
- If $k > \frac{1}{3}$ we get $V_k^i = 0$ and $e_k^i = 0$ for all i, so $l(Z_k) = 0$ and $I_k = \mathbb{C}[x, y, z]$.

We can now describe Libgober's formula

Theorem 2.1 *Let $C \subset \mathbb{P}^2$ be a curve of degree d and let $k_1, \dots, k_n$ be all the constants of quasi-adjunction of the singular points of C. The Alexander polynomial of C is*

$$\Delta_C(t) = (t-1)^{r-1} \prod_{j=1,\dots,n}^{dk_j \in \mathbb{N}} [(t - e^{2\pi i k_j})(t - e^{-2\pi i k_j})]^{s(k_j)} \tag{2.1}$$

where r is the number of irreducible components of C, $N_d(k_j) := d - 3 - dk_j$ and the number $s(k_j)$ is given by the difference $l(Z_{k_j}) - h_{I_{k_j}}(N_d(k_j))$.

The number $s(k_j)$ is a generalisation of the so-called *defect of a linear system*. Precisely, for any finite set of points $\Sigma \subset \mathbb{P}^2$ and any $d \in \mathbb{N}$, we can define $S_d(\Sigma) := \{g \in \mathbb{C}[x, y, z]_d \text{ s.t. } g \text{ vanishes on } \Sigma\}$ (the linear system of homogeneous polynomials of degree d vanishing on Σ) and $def(S_d(\Sigma)) := |\Sigma| - codim_{\mathbb{C}}(S_d(\Sigma), \mathbb{C}[x, y, z]_d)$ (its defect); the latter value measures, in some sense, the dependence of the points in Σ with respect to curves of degree d.

Example 4 Consider $d = 1$ and $|\Sigma| = 3$: if the points of Σ are collinear (i.e. 'dependent with respect to lines') then $def(S_1(\Sigma)) = 1$; otherwise $def(S_1(\Sigma)) = 0$. Consider instead $d = 2$ and $|\Sigma| = 6$. The space of conics in $\mathbb{P}^2$ has dimension 5 (can you see why?), so it

is not always true that we can find a conic passing through all the points of Σ; if this is possible, we get $def(S_2(\Sigma)) = 1$, otherwise we get $def(S_2(\Sigma)) = 0$.

The number $s(k_j)$ generalises the defect by replacing $|\Sigma|$ (cardinality of a set of points) with $l(Z_{k_j})$ ('cardinality of a set of points with multiplicities') and $codim_{\mathbb{C}}(S_d(\Sigma), \mathbb{C}[x,y,z]_d)$ with $h_{I_{k_j}}(N_d(k_j))$.

Remark 5 The previous discussion on the number $s(k_j)$ as a generalisation of the defect shows that the Alexander polynomial of a plane curve C does not depend only on the number and type of singular points, but also on their relative position; this was known since the classical example by Zariski (see [Z]) of an irreducible ($r = 1$) curve of degree 6 having 6 cusps as singularities (i.e. points with local equation $y^2 - x^3 = 0$): if these cusps lie on a conic then $\Delta_C(t) = \Phi_6(t) = t^2 - t + 1$, otherwise $\Delta_C(t) = 1$ (compare with previous example).

By using formula (2.1) for computations, one quickly realizes that line arrangements with non-trivial Alexander polynomial are pretty rare; one of them is A_3 (what is its Alexander polynomial?), which enjoys particular symmetry properties: we can in fact partition its set of lines into three classes $\mathcal{A}_1 := \{x, y-z\}$, $\mathcal{A}_2 := \{z, x-y\}$ and $\mathcal{A}_3 := \{y, x-z\}$ of the same cardinality in such a way that lines from different $\mathcal{A}_i$ meet only at triple points and the number of lines in $\mathcal{A}_i$ passing through a triple point is constant in i. We can summarise the situation by saying that A_3 admits a 3-net with base locus $\{p_1, p_2, p_4, p_7\}$.

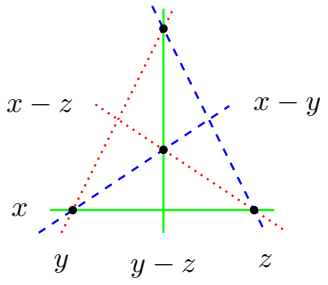
Definition 2.1 A k-net on a line arrangement $\mathcal{A}$ is a pair $(\mathcal{N}, \mathcal{X})$ where $\mathcal{N}$ is a partition of $\mathcal{A}$ into $k \geq 3$ classes $\mathcal{A}_1, \ldots, \mathcal{A}_k$ and $\mathcal{X}$ is a set of multiple points of order at least 3 (called *base locus*) such that:

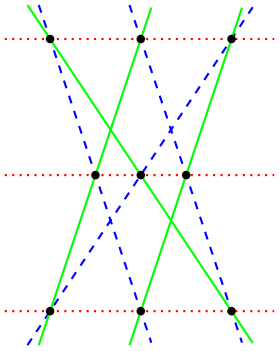
(a) The $\mathcal{A}_i$ contain the same number of lines.

(b) For any $l \in \mathcal{A}_i$ and $l' \in \mathcal{A}_j$ with $i \neq j$, the point $l \cap l'$ belongs to $\mathcal{X}$.

(c) For any $p \in \mathcal{X}$, there is exactly one line of $\mathcal{A}_i$ passing through p for each i.

Below we show the 3-nets on the A_3 arrangement and on the Pappus arrangement; the latter has Alexander polynomial $(t-1)^8(t^2+t+1)$.



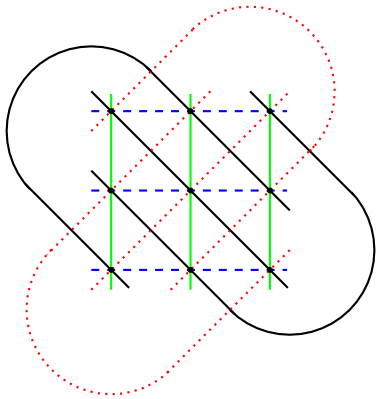


The following recent result (proved in [PS]) shows that the existence of a 3-net is in some cases a necessary and sufficient condition for the non-triviality of $\Delta_{\mathcal{A}}$:

Theorem 2.2 *Let $\mathcal{A}$ be a line arrangement consisting of d lines with multiple points of order 2 and 3 only, then:*

- $\Delta_{\mathcal{A}}(t) = (t-1)^{d-1}(t^2+t+1)^a$ *with* $0 \leq a \leq 2$.
- $a \neq 0$ *if and only if $\mathcal{A}$ admits a 3-net.*
- *a depends only on $L(\mathcal{A})$.*

This Theorem links the presence of a non-trivial factor Φ_3^a in $\Delta_{\mathcal{A}}$ to the existence of a 3-net on $\mathcal{A}$ with base locus given by multiple points of order three, under an admittedly very restrictive hypothesis on the multiplicities of the singular points. This provides a way to convince ourselves that the form of $\Delta_{\mathcal{A}}$ suggested by Conjecture 1.2 is not as strange as we initially thought: Yuzvinsky proved in fact (see [Y]) that on a line arrangement $\mathcal{A}$ there can be a k-net with base locus of cardinality at least two only if $k \leq 3, 4$ (the condition on the cardinality of the base locus is necessary: a central arrangement of k lines admits a k-net for any k). By analogy with the previous Theorem one could hypothesise that the presence of a non-trivial factor $[\Phi_2\Phi_4]^b$ is linked to the existence on a 4-net on $\mathcal{A}$. However, while arrangements admitting 3-nets are not too rare, only *one* non-central arrangement admitting a 4-net is known up to now: the so-called Hesse arrangement, a representation of which is shown in the picture below:



Its Alexander polynomial is $(t-1)^{11}[(t+1)(t^2+1)]^2$, so it is indeed non-trivial; another interesting feature of this arrangement is that if we remove all the lines of a single class, we get the Pappus arrangement.

3 Some complex geometry and the T_{2k} arrangement class

As anticipated, we begin this section by recalling some notions in complex geometry; the interested reader should refer to [V] for the details, especially for the Hodge theory part.

In Section 1 we explained what a differentiable complex-valued 1-form on a complex manifold M of dimension n is: if we look back at the expression (1.1), we can see that we can write it as the sum of two 1-forms α and β where α only involves holomorphic differentials dz_i and β only involves antiholomorphic differentials $d\overline{z}_i$; any 1-form is thus the sum of a 1-form of *type* $(1,0)$ and a 1-form of type $(0,1)$. In general, a k-form on M of type (p,q) for $p+q=k$ (also called simply a (p,q)-form) is given by an expression like

$$\omega = \sum_{|I|=p,|J|=q} f_{I,J}(\underline{z}) dz_I \wedge d\overline{z}_J \quad \text{for } I, J \subset \{1,\dots,n\}$$

where $dz_I := dz_{i_1} \wedge \cdots \wedge dz_{i_p}$ and $d\overline{z}_J := d\overline{z}_{j_1} \wedge \cdots \wedge d\overline{z}_{j_q}$. If we differentiate a (p,q)-form ω, taking advantage of the properties of the differential we see that we obtain the sum of a $(p+1,q)$-form and a $(p,q+1)$-form. If we denote by $\Omega_M^{p,q}$ the space of (p,q)-forms on M, we can define the following operators:

$$\begin{aligned} &\partial : \Omega_M^{p,q} \to \Omega_M^{p+1,q} \text{ s.t. } \omega \mapsto (p+1,q)\text{-part of } d\omega && \textit{Holomorphic differential} \\ &\overline{\partial} : \Omega_M^{p,q} \to \Omega_M^{p,q+1} \text{ s.t. } \omega \mapsto (p,q+1)\text{-part of } d\omega && \textit{Dolbeault operator} \end{aligned}$$

Both ∂ and $\overline{\partial}$ give 0 if applied consecutively, so for any $p = 0,\dots,n$ we can consider the following complex

$$0 \to \Omega_M^{p,0} \xrightarrow{\overline{\partial}} \dots \xrightarrow{\overline{\partial}} \Omega_M^{p,n} \xrightarrow{\overline{\partial}} 0$$

and we can define

$$H^{p,q}(M) := \frac{Ker(\overline{\partial} : \Omega_M^{p,q} \to \Omega_M^{p,q+1})}{Im(\overline{\partial} : \Omega_M^{p,q-1} \to \Omega_M^{p,q})} \qquad h^{p,q} := dim_{\mathbb{C}} H^{p,q}(M)$$

The vector spaces $H^{p,q}(M)$ contain (p,q)-forms which are $\overline{\partial}$-closed but not $\overline{\partial}$-exact, and their dimensions $h^{p,q}$ are called *Hodge numbers* of M. When M falls into a particular class of complex manifolds, namely compact Kähler manifolds, for any $i = 0,\dots,n$ we have the following equalities:

$$\begin{aligned} H^i(M,\mathbb{C}) &= \bigoplus_{p+q=i} H^{p,q}(M) && \text{Hodge decomposition} \\ H^{p,q}(M) &= \overline{H^{q,p}(M)} && \text{Hodge symmetry} \end{aligned}$$

These two facts are very important for us, because we will be dealing with projective manifolds, which are in particular compact Kähler manifolds. The information on the cohomology of a projective manifold M can thus be resumed in the so called *Hodge diamond*; if $n = 2$, it reads

$$\begin{array}{ccccc} & & H^{2,2}(M) & & \\ & H^{2,1}(M) & & H^{1,2}(M) & \\ H^{2,0}(M) & & H^{1,1}(M) & & H^{0,2}(M) \\ & H^{1,0}(M) & & H^{0,1}(M) & \\ & & H^{0,0}(M) & & \end{array}$$

but using Hodge symmetry and Serre duality (for the latter see for example [H]) we can see that the only essential part of it is the lower leftmost triangle

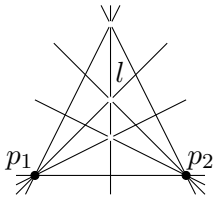
$$\begin{array}{ccc} H^{2,0}(M) & & H^{1,1}(M) \\ & H^{1,0}(M) & \\ & & H^{0,0}(M) \end{array}$$

One of the many reasons why the Hodge numbers are important is that they are 'constant for families of manifolds': what this means is that if $\pi : X \to B$ is a *deformation*, i.e. a morphism between complex manifolds satisfying some technical conditions and parametrising a family of manifolds as the fibres $X_b := \pi^{-1}(b)$, then $h^{p,q}(X_b) = h^{p,q}(X_{b'})$ for any p, q and $b, b' \in B$. We can relax the definition of deformation by requiring X and S to be simply projective varieties (zero-loci of homogeneous polynomials in some projective space); in this case the fibres X_b can have singularities, and we say that π is an *equisingular* deformation if all the X_b have the same singularities.

Now we go back to line arrangements. Consider 4 points p_i in $\mathbb{P}^2$ such that no three of them lie on the same line, and consider arrangements $\mathcal{A}$ consisting of

- $k-1$ lines passing through p_1 but noth through p_2.
- $k-1$ lines passing through p_2 but noth through p_1.
- A line passing through p_1 and p_2.
- A line l passing through p_3 and p_4.

with $k \geq 4$; we call any arrangement like this of *type* T_{2k}. Note that any arrangement of this type can have from zero to $k-1$ triple points lying on the line l (the next figure shows an arrangement of type T_8).



Our aim is to prove that all arrangements of type T_{2k} have trivial Alexander polynomial. Key to our argument is the following result by Libgober, a proof of which can be found also in [D]:

Theorem 3.1 *Let $C := V(f) \subset \mathbb{P}^2$ be a curve of degree d, $F \subset \mathbb{C}^3$ be the corresponding Milnor fibre, $S_C := V(y^d - f) \subset \mathbb{P}^3$ a surface and $\tilde{S_C} \subset \mathbb{P}^3$ its resolution of singularities. If we write the non-trivial part of the Alexander polynomial of C as $\prod_{1<k|d} \Phi_k^{\alpha_k}$, we have*

$$2h^{0,1}(\tilde{S_C}) = \sum_{1<k|d} deg(\Phi_k)\alpha_k$$

We can now sketch the proof of the aforementioned result on arrangements of type T_{2k}:

(a) First, note that an equisingular deformation of *any* arrangement $\mathcal{A}$ gives rise to an equisingular deformation of the surface $S_{\mathcal{A}}$, and from the latter we obtain a deformation of the smooth surface $\tilde{S_{\mathcal{A}}}$; thus, we get a family of complex projective manifolds of dimension 2, one of whose members is $\tilde{S_{\mathcal{A}}}$.

(b) Any two arrangements $\mathcal{A}$ and $\mathcal{A}'$ of type T_{2k} having the same number of triple points can be equisingularly deformed one onto the other; this means that arrangements of type T_{2k} can be divided into k 'equisingular deformation-equivalence' classes, depending on the number of triple points they have. For each of these classes we take a representative $\mathcal{A}_i$

(c) Up to an automorphism of $\mathbb{P}^2$, we can assume that $p_1 = (0 : 0 : 1)$ and $p_2 = (0 : 1 : 0)$; this makes it possible to write an explicit equation for each $\mathcal{A}_i$, and to compute their Alexander polynomial (which turns out to be trivial) using Libgober's formula (2.1).

(d) Using the previous Theorem, we can conclude that any $\mathcal{A}$ in the same class as $\mathcal{A}_i$ has trivial Alexander polynomial, because $h^{0,1}(\tilde{S_{\mathcal{A}}}) = h^{0,1}(\tilde{S_{\mathcal{A}_i}}) = 0$. We can conclude that all the arrangements of type T_{2k} have trivial Alexander polynomial.

Note that arrangements of type T_{2k} do not admit 3-nets nor 4-nets, so this result is another hint towards the fact that nets are necessary in order to have non-trivial Alexander polynomial.

References

[B] Brieskorn E., *Die Monodromie der Isolierten Singularitaten von Hyperflachen.* Manuscr. Math. 2 (1970), 103–161.

[D] Dimca A., "Hyperplane arrangements, an introduction". Universitext, Springer, New York, 2017.

[H] Hartshorne R., "Algebraic geometry". Springer, New York, 1977.

[L1] Libgober A., *Alexander invariants of plane algebraic curves.* In "Singularities, Part 2 (Arcata, Calif., 1981)", volume 40 of Proc. Symp. Pure Math, 135–143. Amer. Math. Soc., Providence, RI, 1983.

[L2] Libgober A., *Position of singularities of hypersurfaces and the topology of their complements.* J. Math-Sci. 82 (1996), 3194–3210. Algebraic Geometry, 5.

[M] Milnor J., "Singular points of complex hypersurfaces". Annals of Math. Studies, volume 61. Princeton Univ. Press, Princeton, NJ, 1968.

[PS] Papadima S., Suciu A., *The Milnor fibration of a hyperplane arrangement - From modular resonance to algebraic monodromy.* Proc. of the London Math. Soc. 114, no. 6 (2017), 961–1004.

[V] Voisin C., "Complex algebraic geometry and Hodge theory". Cambridge University Press, Cambridge, 2002.

[Y] Yuzvinski S., *Realization of finite abelian groups by nets in* $\mathbb{P}^2$. Compos. Math. 140, no. 6 (2004), 1614–1624.

[Z] Zariski O., *On the irregularity of cyclic multiple planes.* Ann. Math. 32 (1931), 485–511.

An introduction to stochastic control in discrete time with an application to life insurance

Maren Diane Schmeck (*)

Abstract. We review the basic theory of stochastic control in discrete time in intuitive, informal way. The dynamic programming principle we then apply to a problem in life insurance: We consider an insurance company, which wants to hedge against systematic mortality risk by trading into an adequate securitziation product, maximizing utility of terminal wealth. In this setting, we show how to calculate the optimal strategies.

1 Introduction

The first aim of this note is to give an intuitive introduction to dynamic programming in discrete time. Here, one aims at finding a optimal dynamic strategy $U_0,\ U_1,\ U_2, \ldots,\ U_T$ over a time horizont, giving an optimal strategy at each point of time. The dynamic programming principle then states that one can reformulate the problem, solving it backwards in time, in each step optimizing only over a decision at a single point of time. First one solves the problem for the last time period. The result is then used to optimize over the period before the last period, and so on. Here, we follow closely [2].
In this framework, we discuss a problem from life insurance that is taken from [1]. An insurance company is exposed to systematic mortality risk, which he/she wants to transfer to the financial market by trading into a mortality securitization product. The Insurer then wants to determine the trading strategy that maximises expected utility from terminal wealth. Here, we focus on the case of catastrophe mortality risk, that can effect all individuals of the homogeneous cohort of the insurer simultaneously. The second aim of this note is then to illustrate how to calculate the strategies in this framework.

(*)Center for Mathematical Economics (IMW), Bielefeld University, Universitätsstrasse, 33615 Bielefeld, Germany. E-mail: maren.schmeck@uni-bielefeld.de . Seminar held on March 13th, 2019.

2 Stochastic Control in Discrete Time

This section follows the textbook of Schmidli [2] in an informal way. Consider an iid series of random variables $\{Y_n : n \in \mathbb{N}^*\}$, modeling the stochastic changes over time. Consider further the natural filtration $\{\mathcal{F}_n\} = \{\mathcal{F}_n^Y\}$. Now, a decision is made at each point of time $n \in \mathbb{N}$, modeled through the random variable U_n. We assume that the decision U_n is adapted, that is only based on the information available at time n. Let U_n be in some space $\mathcal{U}$. Furthermore, let $\mathscr{U}$ denote the set of *admissible* strategies, i.e. the adapted strategies $U = \{U_n\}$ that are allowed.
Let's construct the controlled stochastic process. Let $X_0 = x$ be the initial value. The process at time $n+1$ is

$$X_{n+1} = f(X_n, U_n, Y_{n+1}) \,,$$

where f is a measurable function. Here, the next state of the process X only depends on the present state and the present decision.
Let $r(X_n, U_n)$ be the reward at time $n \in \mathbb{N}$. The value connected to some strategy U is then

$$V_T^U(x) = \mathbb{E}\left[\sum_{n=0}^{T} r(X_n, U_n) e^{-\delta n}\right] \,.$$

The parameter $\delta \geq 0$ is a discounting parameter. The goal is to maximize $V_T^U(x)$. Define therefore the *value function*

$$V_T(x) = \sup_{U \in \mathscr{U}} V_T^U(x) \,.$$

It is not feasible to find $V(x)$ by calculating the value function V_T^U for each possible strategy U. But it turns out that the *dynamic programming principle* helps to characterize the value function. It says that, if one knows the value function for $T-1$ steps, it is optimal to maximize only over the first step. If we are at time 1 then only $T-1$ time units are left and we denote the remaining value with $V_{T-1}(x)$ and $V_{T-1}^U(x)$.

Let U be an arbitrary strategy. Then $X_1 = f(x, U_0, Y_1)$ and

$$\begin{aligned} V_T^U(x) &= r(x, U_0) + \mathbb{E}\left[\sum_{n=1}^{T} r(X_n, U_n) e^{-\delta n}\right] \\ &= r(x, U_0) + e^{-\delta}\, \mathbb{E}\left[\sum_{n=0}^{T-1} r(X_{n+1}, U_{n+1}) e^{-\delta n}\right] \end{aligned}$$

Define $\widetilde{X}_n = X_{n+1}$, $\widetilde{U}_n = U_{n+1}$ and $\widetilde{Y}_n = Y_{n+1}$. Then $\widetilde{X}_{n+1} = f(\widetilde{X}_n, \widetilde{U}_n, \widetilde{Y}_{n+1})$.

$$\begin{aligned} V_T^U(x) &= r(x, U_0) + e^{-\delta}\, \mathbb{E}\left[\mathbb{E}\left[\sum_{n=0}^{T-1} r(X_{n+1}, U_{n+1}) e^{-\delta n} \middle| X_1,\, U_0\right]\right] \\ &= r(x, U_0) + e^{-\delta}\, \mathbb{E}\left[V_{T-1}^{\widetilde{U}}(X_1)\right] \end{aligned}$$

Thus, by conditioning on time 1 we have found a relationship between $V_T^U(x)$ and $V_{T-1}^{\widetilde{U}}(X_1)$ for some strategies U and $\widetilde{U}$ as described above. Now, taking the supremum over the strategies $U \in \mathscr{U}$, one can show that the problem reduces to finding the optimal strategy for the fist point of time, if one knows the value function V_{T-1}. Thus, the decisions that have to be taken at the following times enter indirectly V_{T-1}, and one has to approach the problem backwards in time.

Theorem 1 (Bellman's equation) *Suppose that $V_T(x)$ is finite. The function $V_T(x)$ fulfils the dynamic programming principle*

$$V_T(x) = \sup_{u \in \mathcal{U}} \left\{ r(x,u) + e^{-\delta}\, \mathbb{E}\left[V_{T-1}(f(x,u,Y))\right] \right\} , \tag{1}$$

where Y is a generic random variable with the same distribution as Y_n.

For the proof of the Bellmann's equation, see Schmidli [2]. Thus, the Bellmann principle tell us that one can solve the problem backwards, finding the optimal strategies and value function $V_1(x)$ if one stands at time $T-1$ and only 1 step is left. This can then be used to find the optimal strategy if we are two time steps before final time T

$$V_2(x) = \sup_{u \in \mathcal{U}} \left\{ r(x,u) + e^{-\delta}\, \mathbb{E}\left[V_1(f(x,u,Y))\right] \right\} ,$$

and so on. That is, we aproach the problem recursively with a series of optimization problems. Each of these problems optimizes only over one decision $u \in \mathcal{U}$, and not over a series of decisions as in the initial formulation.

3 Example from 'Mortality Options: the Point of View of an Insurer' [1]

This example is taken from Schmeck and Schmidli [1]. There, we consider a life insurer, who faces systematic mortality risk. Let's consider the special case of catastrophe mortality risk: for example due to an epidimia or extreme weather conditions. The Insurer wants to transfer this risk to the financial market by trading into a securitization product (SP). Both insurance contract and SP are assumed to make payments at terminal time n only.

We consider two indices: I models the cohort of the insurer, L the reference portfolio. We further assume that the interest rate r is constant and that the price of a unit of the index is given by the conditional expectation of the discounted payoff. Conditional on the number of survivors in the reference portfolio at time t, L_n is binomially distributed with parameters L_t and ${}_{n-t}p_{x+t}(t)$. Thus the price of a unit of the index at time $t \leq n$ is ${}_{n-t}p_{x+t}(t)(1+r)^{-(n-t)}L_t$.
The insurer starts with a wealth $W_0 = w$. At time k, the insurer buys θ_k units of the

mortality bond. Thus the wealth at time $k+1$ is then given by

$$\begin{aligned} W_{k+1} &= \Big(W_k - \theta_k \frac{{}_{n-k}p_{x+k}(k)L_k}{(1+r)^{n-k}}\Big)(1+r) + \theta_k \frac{{}_{n-k-1}p_{x+k+1}(k+1)L_{k+1}}{(1+r)^{n-k-1}} \\ &= W_k(1+r) + \theta_k \frac{{}_{n-k-1}p_{x+k+1}(k+1)L_{k+1} - {}_{n-k}p_{x+k}(k)L_k}{(1+r)^{n-k-1}} \,. \end{aligned}$$

Denote by

$$V_n(w,x,I_0,L_0) = \sup_{\theta_k} \mathbb{E}[u(W_n - f_I(I_n))]$$

the maximal expected utility for a time horizon of length n. The variable x denotes the underlying cohort.

Recursively, we get

$$V_{n+1}(w,x,I_0,L_0) = \sup_{\theta} \mathbb{E}\Big[V_n\Big(W_1, x+1, I_1, L_1\Big)\Big] \,.$$

This is the Bellman equation connected to our problem.

At time $s < t$ the best prior estimate of the probability of death for a member of the cohort in the period $(t-1,t]$ is ${}_1q_{x+t-1}(0)$. At time t, the realised death probability is ${}_1q_{x+t-1}(t) = {}_1q_{x+t-1}(0)Z_t$, where $\{Z_t\}$ are iid positive variables with expected value 1. We assume in addition that ${}_1q_{x+t-1}(t) \le 1$.
Suppose the value function $V_n(w,x,I,L)$ is known. We consider now a securitisation product with payoff in $n+1$. The value of the wealth process at time 1 reads

$$W_1 = W_0(1+r) + \theta_0 \frac{{}_np_{x+1}(1)L_1 - {}_{n+1}p_x(0)L_0}{(1+r)^n} \,.$$

Consider the index L. Because we assume a large portfolio, we can model the index as $L_{n+1} = L_n\, {}_1p_{x+n}(n+1) = L_n(1-(1-{}_1p_{x+n}(0))Z_{n+1})$. This yields in particular

$$L_1 = {}_1p_x(1)L_0 = (1-(1-{}_1p_x(0))Z_1)L_0 \,.$$

Thus

$$\begin{aligned} {}_np_{x+1}(1)L_1 - {}_{n+1}p_x(0)L_0 &= {}_np_{x+1}(0)(1-(1-{}_1p_x(0))Z_1)L_0 - {}_{n+1}p_x(0)L_0 \\ &= (1-Z_1)({}_np_{x+1}(0) - {}_{n+1}p_x(0))L_0 \,, \end{aligned}$$

where we used that ${}_np_{x+1}(1) = {}_np_{x+1}(0)$ since the left hand side is the best estimate for the future. Because ${}_np_{x+1}$ here models a survival probability when the cohort is $x+1$ years old (that is from time 1 on), we get the first information on the realised mortality in period 2. Note that ${}_np_{x+1}(0) > {}_{n+1}p_x(0)$. We assume that $I_{n+1} = I_n(1-(1-{}_1p_{x+n}(0))\tilde{Z}_{n+1})$, also assuming a large portfolio. Then the process I behaves in the same way as L but with different random variables.

Consider the exponential utility function $u(x) = -\mathrm{e}^{-\alpha x}$ for some $\alpha > 0$. We claim that $V_n(w,x,I,L) = -\exp\{-\alpha_n w + f_n(x,I,L)\}$ for some function f_n and $\alpha_n = \alpha(1+r)^n$. We find by induction

$$\begin{aligned} & V_{n+1}(w,x,I,L)\mathrm{e}^{\alpha_n w(1+r)} \\ &= \sup_\theta -\mathbb{E}\Big[\exp\Big\{-\alpha_n\theta\frac{(1-Z_1)({}_np_{x+1}(0) - {}_{n+1}p_x(0))L}{(1+r)^n} + f_n(x+1,I_1,L_1)\Big\}\Big] \\ &= -\inf_\theta \mathbb{E}\Big[\exp\Big\{-\alpha_n\theta\frac{(1-Z_1)({}_np_{x+1}(0) - {}_{n+1}p_x(0))L}{(1+r)^n} + f_n(x+1,I_1,L_1)\Big\}\Big]\,. \end{aligned}$$

We have $\alpha_{n+1} = \alpha_n(1+r) = \alpha(1+r)^{n+1}$ and, with $I_1 = (1-(1-{}_1p_x(0))\widetilde{Z}_1)I$ and $L_1 = (1-(1-{}_1p_x(0))Z_1)L$,

$$\begin{aligned} & f_{n+1}(x,I,L) \\ &= \inf_\theta \log\mathbb{E}\Big[\exp\Big\{-\alpha_n\theta\frac{(1-Z_1)({}_np_{x+1}(0) - {}_{n+1}p_x(0))L}{(1+r)^n} + f_n(x+1,I_1,L_1)\Big\}\Big] \\ &= \inf_\theta \log\mathbb{E}\Big[\exp\Big\{-\alpha\theta(1-Z_1)({}_np_{x+1}(0) - {}_{n+1}p_x(0))L + f_n(x+1,I_1,L_1)\Big\}\Big]\,. \end{aligned} \tag{2}$$

Since $V_0(x) = -\mathrm{e}^{-\alpha(w-I_0)}$, the form is proven. Thus, the initial wealth is not relevant for the optimal strategy in this modelling framework, which is to be expected for the exponential utility function. As a specific model, assume that

$$\widetilde{Z}_1 = \gamma U_1 + (1-\gamma)Z_1\,, \tag{3}$$

where $Z_1,\ U_1 \sim \mathcal{U}[\frac{1}{2},\frac{3}{2}]$ are independent and $\gamma \in (0,1)$. As we assume a linear dependence between $\widetilde{Z}$ and Z, its dependence structure is determined by its correlation coefficient being equal to $1-\gamma$.

3.1 One time step before maturity of the contracts

After stating the relevant modelling assumptions, we use the relation (2) to calculate the optimal strategy and optimal function $f_1(x,I,L)$. That is, the insurer is one time step before maturity of the insurance contract. We get $\alpha_0 = \alpha$ and $f_0(I) = \alpha I$. Noting ${}_0p_{x+1}(0) = 1$, we obtain

$$\begin{aligned} & f_1(x,I,L) \\ &= \inf_\theta \log\mathbb{E}\big[\exp\{-\alpha\theta(1-Z)({}_0p_{x+1}(0) - {}_1p_x(0))L + f_0(1-(1-{}_1p_x(0))\widetilde{Z})I)\}\big] \\ &= \inf_\theta \log\mathbb{E}\big[\exp\{-\alpha\theta(1-Z)(1-{}_1p_x(0))L + \alpha(1-(1-{}_1p_x(0))\widetilde{Z})I\}\big] \\ &= \alpha\,{}_1p_x(0)I + \inf_\theta \log\mathbb{E}\big[\exp\{\alpha(1-{}_1p_x(0))[\theta L - (1-\gamma)I](Z-1)\}\big]\big\} \\ &\qquad + \log\mathbb{E}\big[\exp\{\alpha(1-{}_1p_x(0))\gamma I(1-U)\}\big] \\ &= \alpha\,{}_1p_x(0)I + \log\frac{\sinh(\frac{1}{2}\alpha(1-{}_1p_x(0))\gamma I)}{\frac{1}{2}\alpha(1-{}_1p_x(0))\gamma I} \\ &\qquad + \inf_\theta \log\frac{\sinh(\frac{1}{2}\alpha(1-{}_1p_x(0))[\theta L - (1-\gamma)I])}{\frac{1}{2}\alpha(1-{}_1p_x(0))[\theta L - (1-\gamma)I]}\,. \end{aligned}$$

Now, $g(x) = \frac{1}{x}\sinh(x)$ takes its unique minimum in $x = 0$ with $g(0) = 1$. Thus,

$$\theta^* := (1-\gamma)\frac{I}{L}\,, \tag{4}$$

and

$$f_1(x, I, L) = \alpha\, {}_1p_x(0)I + \log \frac{\sinh(\frac{1}{2}\alpha(1 - {}_1p_x(0))\gamma I)}{\frac{1}{2}\alpha(1 - {}_1p_x(0))\gamma I}\,.$$

Note that f_1 does not depend on L. Because part of the mortality in the reference portfolio changes as the mortality in the own portfolio, the strategy is to hedge this dependent part of possible survivors with the securitisation product. Therefore, the form of our optimal portfolio is due to the simple model we use. Note that, if we replace L by ζL, we can get the same value by choosing θ/ζ. Therefore, a θ proportional to $1/L$ is to be expected.

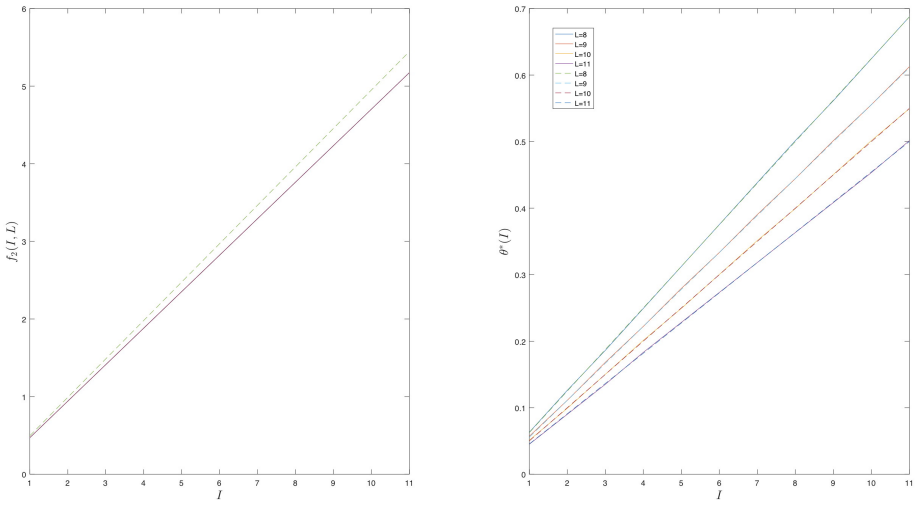


Figure 1. Left: The optimal functions $f_2(I, L)$ (solid line) and $f_1(I, L)$ (dashed). Right: The optimal strategies θ_2^* (solid line) and θ_1^* (dashed) if there are two and one time step left before maturity of the products.

3.2 Two time steps before maturity of the contracts

Taking into account that f_1 does not depend on L we have that

$$\begin{aligned} f_2(x, I, L) &= \inf_\theta \log \mathbb{E}\left[\exp\{-\alpha\theta(1-Z)({}_1p_{x+1}(0) - {}_2p_x(0))L + f_1(x+1, I_1, 0)\}\right] \\ &= \inf_\theta \log \mathbb{E}\Big[\exp\{-\alpha\theta(1-Z)({}_1p_{x+1}(0) - {}_2p_x(0))L \\ &\qquad +\alpha\, {}_1p_{x+1}(0)I_1\}\frac{\sinh(\frac{1}{2}\alpha(1 - {}_1p_{x+1}(0))\gamma I_1)}{\frac{1}{2}\alpha(1 - {}_1p_{x+1}(0))\gamma I_1}\Big]\,. \end{aligned}$$

where we have used the relation (2) and $I_1 = (1 - (1 - {}_1p_x(0))\widetilde{Z})I$, where again $\widetilde{Z} = (1-\gamma)Z + \gamma U$ and $Z,\ U \sim \mathcal{U}[\frac{1}{2}, \frac{3}{2}]$. Note that the expectation on the right hand side does not have an explicit analytical representation. Nevertheless, we would like to illustrate its behaviour numerically for the parameters $\gamma = 0.5$, ${}_1p_x(0) = 0.99$, ${}_1p_{x+1}(0) = 0.95$, ${}_2p_x(0) = {}_1p_x(0)\,{}_1p_{x+1}(0) = 0.9405$ and $\alpha = 0.5$.

The first panel in Figure 1 shows the resulting optimal function $f_2(I, L)$. As in the first time step, it is independent of L. For comparison we have added $f_1(I, L)$ (dashed line). Note that we have added the subscripts to the θ if we want to emphasize that θ_2 is a position to hold if two time steps are left, and θ_1 is the position to hold if only one time step is left. The last panel shows θ_2^* as a function of I, note the rather linear shape. Again for comparison, we have added the optimal strategy θ_1^* (dashed line) if only one unit of time is left before the insurance product is paid out. The lines seem to be identical.

References

[1] Schmeck, M. D., and Schmidli, S., *Mortality Options: The Point of View of an Insurer*. Available at `https://pub.uni-bielefeld.de/record/2935798` (2019).

[2] Schmidli, H., "Stochastic Control in Insurance". Springer-Verlag, London, 2008.

Covers and envelopes of modules

Giovanna Giulia Le Gros [(*)]

Abstract. Approximation theory of modules is the study of left or right approximations of modules, also known as covers or envelopes, with respect to certain classes of modules. For a class C of R-modules, the aim is to characterise the rings over which every module has a C-cover or a C-envelope and furthermore to characterise the class C itself. For example, if one considers the class of injective modules, then it is well-known that every module has an injective envelope (or injective hull). Instead, Bass proved that projective covers rarely exist and characterised the rings over which every module admits a projective cover, which are known as perfect rings. Moreover, precovers and preenvelopes are strongly related to the notion of a cotorsion pair, which is a pair of Ext-orthogonal classes in the category of R-modules.
The aim of this note is to give a basic introduction to the theory of covers and envelopes, and to describe them with respect to some well-known classes of R-modules, along with a review of concepts in homological algebra that will be useful in this exposition.

1 Preliminaries

In these notes, the ring R will always denote an associative ring with unit.

Definition 1.1 A *right R-module* is an abelian group $(M, +)$ with a right action of R on M, $\cdot : M \times R \to M$, such that the following conditions hold.

(a) $(m + n) \cdot r = m \cdot r + n \cdot r$ for every $m, n \in M$ and $r \in R$.

(b) $m \cdot 1_R = m$ for every $m \in M$.

(c) $m \cdot (r + s) = m \cdot r + m \cdot s$ for every $m \in M$ and $r, s \in R$.

(d) $m \cdot (rs) = (m \cdot r) \cdot s$.

For convenience the right action $m \cdot r$ will simply be denoted mr.

Let M, N be right R-modules. A *homomorphism of right R-modules* is a map $f : M \to N$ such that f preserves the abelian group operation and is right R-linear. That is,

[(*)]Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: legros@math.unipd.it . Seminar held on March 27th, 2019.

(a) $f(m+n) = f(m) + f(n)$,

(b) $f(m \cdot r) = f(m) \cdot r$ for every $m \in M$ and $r \in R$.

Let Mod-R denote the collection (or more precisely, the category) of all right R-modules. The left R-modules are defined analogously with the action of R on the left, however, when R is commutative, the right R-modules and left R-modules coincide. In these notes we refer always to right R-modules, although everything can be written in terms of left R-modules.

Example 1.2 Let K be a field. Then Mod-K is exactly the category of vector spaces over K. This category is particularly well-behaved as every K-module is isomorphic to a direct sum of copies of K.

Example 1.3 An abelian group G can be considered a $\mathbb{Z}$-module, and every homomorphism of abelian groups a $\mathbb{Z}$-module homomorphism. Therefore, one can think of Mod-$\mathbb{Z}$ as the category of abelian groups, where the action of $\mathbb{Z}$ is $x \cdot n = \underbrace{x + \cdots + x}_{n \text{ times}}$ for $x \in G$ and $n \in \mathbb{Z}$.

We will say a diagram (as below) of R-module homomorphisms *commutes* if the homomorphism is the same no matter what path of homomorphisms you choose. That is, the following diagram commutes exactly when $g' \circ f = f' \circ g$.

$$\begin{array}{ccc} A & \xrightarrow{f} & B \\ {\scriptstyle g}\downarrow & & \downarrow{\scriptstyle g'} \\ C & \xrightarrow{f'} & D \end{array}$$

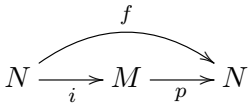
The arrow $\hookrightarrow$ will always denote an injective homomorphism and the arrow $\twoheadrightarrow$ will always denote a surjective homomorphism.

Definition 1.4 The *direct sum* of R-modules M and N is a module denoted $M \oplus N$ where every element can be written as a unique sum $m + n$ where $m \in M$ and $n \in N$.
For an infinite set of modules $\{M_i\}_{i \in I}$ indexed by I, the direct sum $\bigoplus_{i \in I} M_i$ is the R-module where every element can be written as a unique finite sum of elements in the M_i.
If a module L can be written as $M \oplus N$ for submodules $M, N \subseteq L$, then M and N are referred to as *direct summands* of L, or one writes $M, N \underset{\oplus}{\leq} L$.

For an index set I, the module $M^{(I)}$ will denote the direct sum $\bigoplus_{i \in I} M_i$ where $M_i = M$ for each i.

Some useful properties of modules are preserved under direct sums and direct summands, therefore it will be useful to know when a module decomposes as a direct sum of two of its submodules. In fact, for $N, N' \leq M$, $M = N \oplus N'$ if and only if $N + N' = M$ and $N \cap N' = 0$.

Similarly, suppose $f : N \to N$ is an isomorphism that factors through M. Then the following diagram commutes (that is if $p \circ i = f$), and $i(N)$ is a direct summand of M.



In fact, $M = i(N) \oplus \mathrm{Ker}p$.
Direct sums of copies of the ring are called the *free modules*, that is, the R-modules which have a basis, or a linearly independent generating set. We recall the following fact which will be useful, as well as some useful classes of modules.

Remark 1.5 For every module $M \in$ Mod-R, there exists a surjective map $R^{(\alpha)} \twoheadrightarrow M$ for some cardinal α.

An R-module D is *divisible* in Mod-R if for every non-zero divisor $r \in R$, $D = Dr$. For example, $\mathbb{Q}$ is divisible in Mod-$\mathbb{Z}$ as $\mathbb{Q}n = \mathbb{Q}$ for every $n \in \mathbb{Z}$. Also, every K-vector space is divisible since K itself is divisible in Mod-K.

A right R-module M is *torsion* if for every element $m \in M$ there exists a non-zero divisor $r \in R$ such that $mr = 0$. For any module M, there exists a maximal unique torsion submodule denoted $t(M)$.
A right R-module M is *torsion-free* if its unique torsion submodule is 0, or equivalently if for every element $m \in M$ and every non-zero divisor r in R, $m \cdot r = 0$ implies that $m = 0$. In the case of abelian groups, these two notions coincide with the usual notion of torsion and torsion-free abelian groups. For example, $\mathbb{Z}$ and $\mathbb{Q}$ are torsion-free but $\mathbb{Z}/2\mathbb{Z}$, $\mathbb{Q}/\mathbb{Z}$, $\mathbb{Z}/3\mathbb{Z} \oplus \mathbb{Z}$ are not. The modules $\mathbb{Q}/\mathbb{Z}$ and $\mathbb{Z}/2\mathbb{Z}$ are torsion, and the maximal torsion submodule of $\mathbb{Z}/3\mathbb{Z} \oplus \mathbb{Z}$ is $\mathbb{Z}/3\mathbb{Z}$.
A submodule of a torsion module is always torsion, and similarly a submodule of a torsion-free module is always torsion-free. The only module that is both torsion and torsion-free is 0.

2 Short exact sequences

The next two sections will introduce briefly some constructions in Mod-R. More information about the content of the following two sections can be found any standard text on homological algebra, for example [4].
For an R-homomorphism f, one can define the following useful R-modules.

Definition 2.1 The *kernel* of a homomorphism $f : M \to N$ is the R-module

$$\mathrm{Ker}f = \{m \in M : f(m) = 0\} \leq M.$$

The *image* of a homomorphism $f : M \to N$ is the R-module

$$\mathrm{Im}f = \{n \in N : n = f(m) \text{ for some } m \in M\} = f(M) \leq N.$$

The *cokernel* of a homomorphism $f : M \to N$ is the R-module

$$\mathrm{Coker} f \cong N/\mathrm{Im} f.$$

A homomorphism is injective if and only if its kernel is 0. Dually, a homomorphism is surjective if and only if its image is the whole codomain ($\mathrm{Im} f = N$) or equivalently $\mathrm{Coker} f = 0$.

The sequence

$$A \xrightarrow{f} B \xrightarrow{g} C$$

is called *exact at* B if $\mathrm{Ker} g = \mathrm{Im} f$. That is if both $g \circ f = 0$ and if $g(b) = 0$ for $b \in B$, there is an $a \in A$ such that $f(a) = b$.

Examples 2.2 The following are examples of sequences which are exact at the centre module. The first two examples are in Mod-$\mathbb{Z}$, while the rest are in Mod-R for some ring R. We let 2 denote multiplication by 2 and nat the natural homomorphism from a module to its quotient by a submodule (that is, it sends each element to the equivalence class containing it).

- $$\mathbb{Z} \xrightarrow{2} \mathbb{Z} \xrightarrow{\mathrm{nat}} \mathbb{Z}/2\mathbb{Z}$$
- $$\mathbb{Z}/4\mathbb{Z} \xrightarrow{2} \mathbb{Z}/4\mathbb{Z} \xrightarrow{2} \mathbb{Z}/4\mathbb{Z}$$
- For I a right ideal of R.
$$I \hookrightarrow R \xrightarrow{\mathrm{nat}} R/I$$
- Moreover, if I a right ideal of R as above and $R^{(\alpha)} \twoheadrightarrow I$ is a surjective map for some cardinal α, the following row is exact where ϕ is the composition of the surjective map $R^{(\alpha)} \twoheadrightarrow I$ and the inclusion map $I \hookrightarrow R$.

$$\begin{array}{ccccc} R^{(\alpha)} & \xrightarrow{\quad\phi\quad} & R & \xrightarrow{\mathrm{nat}} & R/I \\ & \searrow \quad \cup \quad \nearrow & & & \\ & I & & & \end{array}$$

For any $f : A \to B$, the following are exact.

- $$\mathrm{Ker} f \to A \xrightarrow{f} B$$

  It follows from the above sequence that $0 \to A \xrightarrow{f} B$ exact if and only if f is injective (that is, $\mathrm{Ker} f = 0$).

- $$A \xrightarrow{f} B \to \mathrm{Coker} f \cong B/f(A)$$

  It follows from the above sequence that $B \xrightarrow{g} C \to 0$ is exact if and only if f is surjective (that is, $\mathrm{Im} g = C$).

For reasons now clear from the last two examples, $0 \to A \to B$ and $B \to C \to 0$ will be used to denote injective and surjective homomorphisms, respectively.

Definition 2.3 The sequence

$$0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0$$

is called a *short exact sequence* if f is injective, g is surjective, and $\mathrm{Im} f = \mathrm{Ker} g$. That is, the sequence is exact at each module A, B, and C.

The following lemma characterises a type of short exact sequence which simplifies in a nice way, called a *split* short exact sequence.

Lemma 2.4 *A short exact sequence* $0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0$ *splits if any of the following hold.*

(a) *There exists* $f' : B \to A$ *such that* $f' \circ f = \mathrm{id}_A$.

(b) *There exists* $g' : C \to B$ *such that* $g \circ g' = \mathrm{id}_C$.

(c) *There exists an isomorphism* h *such that the following diagram commutes.*

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A & \xrightarrow{f} & B & \xrightarrow{g} & C & \longrightarrow & 0 \\ & & \| & & \cong\downarrow h & & \| & & \\ 0 & \longrightarrow & A & \xrightarrow{i_A} & A \oplus C & \xrightarrow{p_C} & C & \longrightarrow & 0 \end{array}$$

2.1 Projective and injective modules

Short exact sequences are strongly related to two classes of modules: the projective modules, and the injective modules, which are defined dually to each other.

A module P is *projective* if for every map $P \xrightarrow{h} C$ and surjection $B \xrightarrow{g} C$, there exists a map $h' : P \to B$ such that $g \circ h' = h$.

$$\begin{array}{ccccc} & & P & & \\ & \overset{\exists h'}{\swarrow} & \downarrow h & & \\ B & \xrightarrow[g]{} & C & \longrightarrow & 0 \end{array}$$

We will denote by $\mathcal{P}$ the class of projective modules in Mod-R.

Example 2.5 In Mod-$\mathbb{Z}$, the projective modules are of the form $\mathbb{Z}^{(\alpha)}$ for all cardinals α. That is, they are exactly the free abelian groups. More generally, the projective modules in Mod-R are exactly the directs summands of free modules (direct sums of copies of R), so P is projective if and only if $P \underset{\oplus}{\leq} R^{(\alpha)}$ for some cardinal α.
It follows that arbitrary direct sums of projective modules are projective.

We now define the dual of the projective modules.
A module E is *injective* if for every map $A \xrightarrow{k} E$ and injective map $A \xrightarrow{f} B$, there exists a map $k' : B \to E$ such that $k' \circ f = k$.

$$\begin{array}{ccccc} 0 & \longrightarrow & A & \xrightarrow{f} & B \\ & & {\scriptstyle k}\downarrow & \swarrow_{\exists k'} & \\ & & E & & \end{array}$$

We will denote by $\mathcal{I}$ the class of injective modules in Mod-R. We will give some examples of injective modules. In general, they are less easily characterised than the projective modules.

Example 2.6 In general, every injective module is divisible. In Mod-$\mathbb{Z}$, the converse also holds, that is an abelian group is injective if and only if it is divisible. For example, $\mathbb{Q}$ and $\mathbb{Q}/\mathbb{Z}$ are injective modules, as it is clear that $\mathbb{Q} = n\mathbb{Q}$ for every $n \in \mathbb{Z}$, and moreover $n(\mathbb{Q}/M) = \mathbb{Q}/M$ for any submodule M of $\mathbb{Q}$.

The injective modules have the following nice property.

Lemma 2.7 *Every R-module can be embedded in an injective R-module.*

Example 2.8 We will give an example in the case of abelian groups. Take $M \in$ Mod-$\mathbb{Z}$. We know there exists a short exact sequence of the following form.

$$0 \to S \to \bigoplus \mathbb{Z} \to M \to 0$$

Then $M \cong (\bigoplus \mathbb{Z})/S \subseteq (\bigoplus \mathbb{Q})/S$ and $\mathbb{Q}/S$ is divisible, so is injective in Mod-$\mathbb{Z}$.

The following two lemmas will make precise the relationship between short exact sequences and the injective and projective modules.

Lemma 2.9 *The module P is projective if and only if every short exact sequence of the form $0 \to A \xrightarrow{f} B \xrightarrow{g} P \to 0$ splits.*

Proof of forward direction. Suppose the module P is projective. Then for every surjection $g : B \to P$, the identity map on P factors through g as follows.

$$\begin{array}{ccccccccc} & & & & & & P & & \\ & & & & & \swarrow_{\exists g'} & \downarrow{\scriptstyle \mathrm{id}_P} & & \\ 0 & \longrightarrow & A & \xrightarrow{f} & B & \xrightarrow[g]{} & P & \longrightarrow & 0 \end{array}$$

Therefore the sequence splits by Lemma 2.4. □

Lemma 2.10 *The module E is injective if and only if every short exact sequence of the form $0 \to E \xrightarrow{f} B \xrightarrow{g} C \to 0$ splits.*

Proof of forward direction. Suppose the module E is injective. Then for every injection $f : E \to B$, the identity map on E factors through f as follows.

$$\begin{array}{ccccccccc} 0 & \longrightarrow & E & \xrightarrow{f} & B & \xrightarrow[g]{} & C & \longrightarrow & 0 \\ & & {\scriptstyle \mathrm{id}_E}\downarrow & \swarrow & {\scriptstyle \exists f'} & & & & \\ & & E & & & & & & \end{array}$$

Therefore the sequence splits by Lemma 2.4. □

3 Extensions

Fix $A, C \in$ Mod-R. The short exact sequences of the form $0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0$ are called *extensions of C by A*.
Two extensions are said to be *equivalent* if there exists an isomorphism h such that the following diagram commutes.

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A & \xrightarrow{f} & B & \xrightarrow{g} & C & \longrightarrow & 0 \\ & & \| & & {\scriptstyle \cong}\downarrow{\scriptstyle h} & & \| & & \\ 0 & \longrightarrow & A & \xrightarrow{f'} & B' & \xrightarrow{g'} & C & \longrightarrow & 0 \end{array}$$

It can be shown that these extensions form an abelian group denoted $\mathrm{Ext}^1_R(C, A)$. The identity element of this abelian group corresponds to the split extension, that is the extension that has the property of Lemma 2.4. It follows that $\mathrm{Ext}^1_R(C, A) = 0$ if and only if every extension of C by A splits.
This construction gives us another equivalent characterisation of projective modules and injective modules.

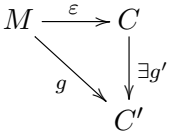
Lemma 3.1 *A module P is projective if and only if $\mathrm{Ext}^1_R(P, M) = 0$ for every $M \in$ Mod-R. Dually, a module E is injective if and only if $\mathrm{Ext}^1_R(M, E) = 0$ for every $M \in$ Mod-R.*

4 Envelopes and preenvelopes

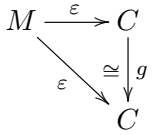
The remaining sections follow [3] or [2].

Let $\mathcal{C}$ be a class in Mod-R closed under isomorphisms (that is, if $C \in \mathcal{C}$ and $C \cong D$ then also $D \in \mathcal{C}$) and M a right R-module.
A *$\mathcal{C}$-preenvelope* of M is a homomorphism $\varepsilon : M \to C$ where $C \in \mathcal{C}$ with the property

that for every homomorphism $g : M \to C'$ with $C' \in \mathcal{C}$, there exists $g' : C \to C'$ such that $g'\varepsilon = g$.



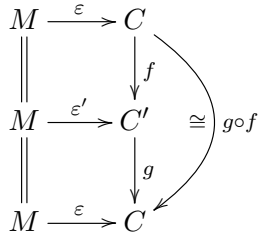
A $\mathcal{C}$-*envelope* of M is a $\mathcal{C}$-preenvelope with the property that for every $g : C \to C$ such that $g\varepsilon = \varepsilon$, g is an isomorphism.



The existence of a $\mathcal{C}$-envelope or a $\mathcal{C}$-preenvelope of a module depends often on the class $\mathcal{C}$ and the module M. If M does have a $\mathcal{C}$-envelope, we can describe the relationship between the $\mathcal{C}$-preenvelopes and $\mathcal{C}$-envelopes of a module M.

Lemma 4.1 *Suppose M has a $\mathcal{C}$-envelope $\varepsilon : M \to C$ and a $\mathcal{C}$-preenvelope of M $\varepsilon' : M \to C'$. Then $C' = D \oplus D'$ where $D \cong C$ and* $\mathrm{Im}\varepsilon \subseteq D$. *Additionally, $\varepsilon : M \to D$ is a $\mathcal{C}$-envelope of M.*
Moreover, it follows that $\varepsilon' : M \to C'$ is a $\mathcal{C}$-envelope of M if and only if M is not contained in a proper direct summand of C'.

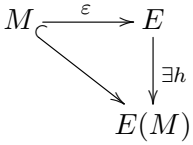
Proof. In the following diagram, the existence of g and f follow by the $\mathcal{C}$-preenvelope properties of ε' and ε respectively. That $g \circ f$ is an isomorphism follows from the $\mathcal{C}$-envelope property of ε.



So $C' = f(C) \oplus \mathrm{Ker}g$. As f is injective, $f(C) \cong C$, so we set $D = f(C)$ and $D' = \mathrm{Ker}g$ as in the statement of the lemma. Also $\mathrm{Im}\varepsilon \subseteq D$ follows by the commutativity of the top square. Using the $\mathcal{C}$-envelope property of ε again, it follows that $\varepsilon : M \to D$ is a $\mathcal{C}$-envelope of M.
If M is not contained in a proper direct summand of C', it forces that $\mathrm{Ker}g = 0$, so ε' is already a $\mathcal{C}$-envelope. For the converse, if ε' is a $\mathcal{C}$-envelope and is contained in a proper direct summand of D, it would contradict the envelope property of ε'. □

The above lemma also tells us that two $\mathcal{C}$-envelopes of M are isomorphic. For the next examples we will fix the class $\mathcal{C}$ to be the class of injective modules $\mathcal{I}$.

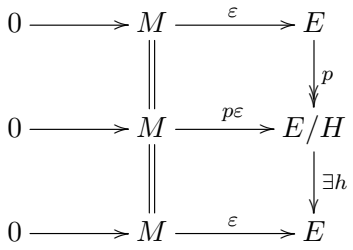
Example 4.2 Recall from Lemma 2.7 that every module can be embedded in an injective module. It follows that an $\mathcal{I}$-preenvelope of M must be an injective homomorphism.



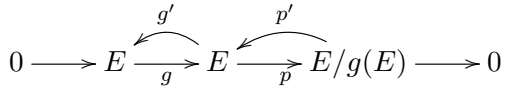
There is a characterisation of when an injective preenvelope is an injective envelope for any module M. First, recall that a submodule N of E is *essential* in E (or $N \leq_e E$) if for $H \leq E$, $H \cap N = 0$ implies that $H = 0$.

Lemma 4.3 *An $\mathcal{I}$-preenvelope $\varepsilon : M \to E$ is an $\mathcal{I}$-envelope if and only if M is essential in E.*

Proof. For the forward direction, suppose that $\varepsilon : M \to E$ is a $\mathcal{I}$-envelope. Take $H \leq E$ such that $H \cap M = 0$ and denote the projection map by $p : E \to E/H$. Then the composition $p\varepsilon : M \to E/H$ is injective. Therefore, there exists $h : E/H \to E$ since E is an injective module as in the following commuting diagram.



By the envelope property of ε, hp is an automorphism so p is an isomorphism and $H = 0$. For the converse, suppose that $\varepsilon : M \to E$ is a $\mathcal{I}$-preenvelope such that $M \leq_e E$. Take $g : E \to E$ such that $g\varepsilon = \varepsilon$. We will show that g is an isomorphism.
To see that g is injective, note $\mathrm{Ker}g \cap \varepsilon(M) = 0$ since if $\varepsilon(m) = g\varepsilon(m) = 0$, then $m = 0$ by the injectivity of ε. So $\mathrm{Ker}g = 0$.
Now we will show that g is surjective. Since we have just shown that g is injective, we have the following short exact sequence.



As E is injective, the sequence splits and $E = g(E) \oplus p'(E/g(E))$. However, $\varepsilon(M) \subseteq g(E)$ (since $\varepsilon(M) = g\varepsilon(M) \subseteq g(E)$) so $\varepsilon(M) \cap \mathrm{Ker}p = 0$ and $\mathrm{Ker}p = 0$ as $\varepsilon(M) \leq_e E$. □

Injective envelopes were shown to exist as stated in the following proposition.

Proposition 4.4 (Eckmann and Schopf (1953)) *Every R-module has an injective envelope.*

Example 4.5 In Mod-$\mathbb{Z}$, the following is an injective envelope.

$$0 \to \mathbb{Z} \to \mathbb{Q}$$

It is straightforward to see this as $\mathbb{Z}$ is essential in $\mathbb{Q}$. More generally, if R is an integral domain (that is, contains no zero divisors and is commutative), then the injective envelope of R is its field of fractions.

There are certain properties of a module that are preserved by its envelope, as we see in the following example.

Example 4.6 An $\mathcal{I}$-envelope of a torsion-free module M is itself torsion-free. To see this, take the following $\mathcal{I}$-envelope of M.

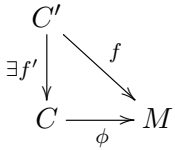
$$0 \to M \to E$$

If $t(E)$ is the torsion submodule of E, then $t(E) \cap M = 0$ as it is a submodule of both a torsion and a torsion-free module, so is 0. Thus $t(E) = 0$ since M is essential in E.

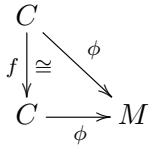
5 Covers and precovers

As before, let $\mathcal{C}$ be a class closed under isomorphisms in Mod-R and M a right R-module.

Definition 5.1 A *$\mathcal{C}$-precover* of M is a homomorphism $\phi : C \to M$ where $C \in \mathcal{C}$ with the property that for every homomorphism $f : C' \to M$ where $C' \in \mathcal{C}$, there exists $f' : C' \to C$ such that $\phi f' = f$.



A *$\mathcal{C}$-cover* of M is a $\mathcal{C}$-precover with the additional property that for every homomorphism $f : C \to C$ such that $\phi f = \phi$, f is an isomorphism.



As with envelopes and preenvelopes, the existence of a $\mathcal{C}$-cover or a $\mathcal{C}$-precover of a module depends often on the class $\mathcal{C}$ and the module M. Dually to the case of envelopes, if M does have a $\mathcal{C}$-cover, we can describe the relationship between the $\mathcal{C}$-precovers and $\mathcal{C}$-covers of a module M.

Lemma 5.2 *Suppose M has a $\mathcal{C}$-cover $\phi : C \to M$ and a $\mathcal{C}$-precover of M $\phi' : C' \to M$. Then $C' = D \oplus D'$ where $D \cong C$ and $D' \subseteq \mathrm{Ker}\phi'$. Additionally, $\phi' \restriction_D$ is a $\mathcal{C}$-cover of M. Moreover, it follows that $\phi' : C' \to M$ is a $\mathcal{C}$-cover of M if and only if C' contains no non-zero direct summands contained in* $\mathrm{Ker}\phi'$.

Proof. Dual to Lemma 4.1. □

Certain properties of the class $\mathcal{C}$ allow us to describe the $\mathcal{C}$-precovers and $\mathcal{C}$-covers. For example, if $R \in \mathcal{C}$ and $\phi : C \to M$ is a $\mathcal{C}$-precover, then ϕ must be surjective. This is because every element in the module M is in the image of a homomorphism from R. Therefore, one applies the precover property to get the following commutating diagram where $m \in M$ and $f_m : 1_R \mapsto m$.

$$\begin{array}{ccc} R & & \\ {\scriptstyle \exists g}\downarrow & \searrow^{f_m} & \\ C & \xrightarrow[\phi]{} & M \end{array}$$

Therefore, every element of M must be in the image of ϕ, so ϕ is surjective.

In the next examples, we will fix $\mathcal{C}$ to be the class of projective modules, $\mathcal{P}$.
First we note that $M \in$ Mod-R, and P projective, $\phi : P \to M$ is a $\mathcal{P}$-precover if and only if ϕ is surjective. This follows by the lifting property of the projective modules, and we conclude that every module has a $\mathcal{P}$-precover.

$$\begin{array}{ccccc} P' & & & & \\ {\scriptstyle \exists f'}\downarrow & \searrow^{f} & & & \\ P & \xrightarrow[\phi]{} & M & \longrightarrow & 0 \end{array}$$

Furthermore, analogously to the case of injective envelopes and essential submodules, we can describe $\mathcal{P}$-covers more explicitly. Recall that a module K is *superfluous* in P (or $K \ll P$) if for $H \leq P$, $H + K = P$ implies that $H = P$

Lemma 5.3 *A $\mathcal{P}$-precover $\phi : P \to M$ is a $\mathcal{P}$-cover if and only if* Kerϕ *is superfluous in P.*

Proof. Dual to Lemma 4.3. □

Unlike in the case of injective envelopes, projective covers do not necessarily exist for every module, though projective precovers do.

Example 5.4 In Mod-$\mathbb{Z}$, an abelian group M has a $\mathcal{P}$-cover if and only if M is projective. For example, consider the following $\mathcal{P}$-precover of $\mathbb{Z}/2\mathbb{Z}$.

$$0 \to \text{Ker}\phi \to \mathbb{Z} \xrightarrow{\phi} \mathbb{Z}/2\mathbb{Z} \to 0$$

Then ϕ is a $\mathcal{P}$-cover if and only if Kerϕ is superfluous in $\mathbb{Z}$. But the only superfluous submodule of $\mathbb{Z}$ is 0 (for any ideal $a\mathbb{Z} \neq 0$ of $\mathbb{Z}$, there exists a proper ideal $b\mathbb{Z}$ such that $a\mathbb{Z} + b\mathbb{Z} = \mathbb{Z}$). So ϕ is not a $\mathcal{P}$-cover.

For the next examples we will look at projective covers of cyclic modules. In particular, one sees that the existence of a projective cover is related to ring theoretic properties of

the ring. To this end, recall that the *Jacobson radical* of a ring R is the intersection of maximal right ideals of R, $J(R) := \bigcap_{\mathfrak{m} \leq_R \max} \mathfrak{m}$. The ideal $J(R)$ is superfluous in R. If not, there would exist a proper ideal H of R such that if $J(R) + H = R$. As H is proper, there is $\mathfrak{m} \supseteq H$, so $\mathfrak{m} = J(R) + \mathfrak{m} = R$, a contradiction.
Moreover, a right ideal I is superfluous in R if and only if $I \leq J(R)$. It follows that $J(R)$ is equal to the sum of the superfluous ideals of R, $J(R) = \sum_{K \ll R} K$.
We now can describe projective covers of cyclic modules.

Example 5.5 For I a right ideal of R, the following is a $\mathcal{P}$-cover if and only if $I \subseteq J(R)$.

$$0 \to I \to R \to R/I \to 0$$

Therefore, Example 5.4 follows from the fact that the Jacobson radical of $\mathbb{Z}$ is 0.

Example 5.6 For any right ideal I, if R/I has a $\mathcal{P}$-cover then either $I \subseteq J(R)$ or I contains a non-zero direct summand of R (that is, there is a decomposition $L_1 \oplus L_2 = R$ such that $L_1 \subseteq I$).
To see this, let $0 \to K \to P_{R/I} \to R/I \to 0$ be the $\mathcal{P}$-cover of R/I extracted from $R \overset{\text{nat}}{\to} R/I$ as in the following diagram from Lemma 5.2.

$$\begin{array}{ccccccccc}
0 & \longrightarrow & K \oplus L & \longrightarrow & P_{R/I} \oplus L & \longrightarrow & R/I & \longrightarrow & 0 \\
 & & \| & & \| & & \| & & \\
0 & \longrightarrow & I & \longrightarrow & R & \xrightarrow{\text{nat}} & R/I & \longrightarrow & 0
\end{array}$$

So L is a non-zero direct summand of R such that $L \leq I$.

5.1 Flat modules

For more information about flat modules and direct limits (which are not defined here), refer to any standard text on algebra, for example [2].

Definition 5.7 A *flat right R-module* is a module F such that for every injective map $f : A \to B$ of left R-modules, the map induced by the tensor product $F \otimes_R f : F \otimes_R A \to F \otimes_R B$ is also injective.

For abelian groups, the flat modules $\mathcal{F}$ are exactly the torsion-free groups. Over a general ring R all flat modules are torsion-free, but the converse doesn't necessarily hold. For every ring R, all projective modules (and therefore all free modules) are flat in Mod-R for any ring R. Moreover, the flat modules have the property that a direct limit of flat modules is still flat, as well as the following property.

Remark 5.8 (Govorov-Lazard Theorem) Every flat module can be written as a direct limit of finitely generated free modules.

Therefore, $\mathcal{P} \subseteq \varinjlim \mathcal{P} = \mathcal{F}$ always holds.

In 1981, Enochs conjectured that every R-module has a flat cover. This was proven to be true in 2001 as stated in the following theorem.

Theorem 5.9 (Bican, El Bashir, Enochs (2001)) *Every R-module has a flat cover.*

In some sense, this can be seen as the dual to the existence of injective envelopes, even though the projective modules are constructed dually to the injective modules. It is then natural to ask over what rings does every module have a projective cover. These rings were characterised by Bass in the following classical theorem, both in terms of homological properties of Mod-R and ring theoretic properties of R.

Theorem 5.10 (Theorem P, Bass (1960) [1]) *For a ring R, the following conditions are equivalent.*

(a) R is right perfect (that is, every right R-module has a projective cover).

(b) Every flat right R-module is projective.

(c) The projective right R-modules are closed under direct limits ($\varinjlim \mathcal{P} = \mathcal{P}$).

(d) Every decreasing chain of left principal ideals terminates. That is, for the sequence of principal ideals

$$Ra_1 \supseteq Ra_2 \supseteq \cdots \supseteq Ra_i \supseteq \cdots$$

where $a_i \in R$, there exists an $n \in \mathbb{N}$ such that $Ra_n = Ra_m$ for all $m \geq n$.

Examples of right perfect rings include left artinian rings. Also, a commutative domain is perfect if and only if it is a field.

6 Cotorsion pairs

Let $\mathcal{C}$ be a class of right R-modules closed under isomorphisms. The *right* Ext^1_R*-orthogonal class* is the class

$$\mathcal{C}^{\perp} = \{M \in \text{Mod-}R : \mathrm{Ext}^1_R(C, M) = 0 \text{ for all } C \in \mathcal{C}\}.$$

The *left* Ext^1_R*-orthogonal class* is the class

$${}^{\perp}\mathcal{C} = \{M \in \text{Mod-}R : \mathrm{Ext}^1_R(M, C) = 0 \text{ for all } C \in \mathcal{C}\}.$$

Definition 6.1 A *cotorsion pair* is a pair of classes $(\mathcal{A}, \mathcal{B})$ which are Ext-orthogonal to each other. More precisely, $\mathcal{A} = {}^{\perp}\mathcal{B}$ and $\mathcal{B} = \mathcal{A}^{\perp}$.

This means that for every $A \in \mathcal{A}$ and $B \in \mathcal{B}$, the following short exact sequence splits.

$$0 \to B \to M \to A \to 0$$

Note that for every cotorsion pair $(\mathcal{A}, \mathcal{B})$, $\mathcal{P} \subseteq \mathcal{A}$ and $\mathcal{I} \subseteq \mathcal{B}$.

Examples 6.2 The following are all cotorsion pairs. From the lemmas concerning the properties of projective and injective modules, it is clear that the first two pairs are cotorsion pairs.

- $(\text{Mod-}R, \mathcal{I})$ where $\mathcal{I}$ is the class of injective modules.
- $(\mathcal{P}, \text{Mod-}R)$ where $\mathcal{P}$ is the class of projective modules.
- For a class $\mathcal{C}$, $\left({}^{\perp}(\mathcal{C}^{\perp}), \mathcal{C}^{\perp}\right)$ is a cotorsion pair and is said to be *generated by* $\mathcal{C}$. Dually, $\left({}^{\perp}\mathcal{C}, ({}^{\perp}\mathcal{C})^{\perp}\right)$ is a cotorsion pair and is said to be *cogenerated* by $\mathcal{C}$.
- $(\mathcal{F}, \mathcal{C})$ where $\mathcal{F}$ denotes the class of flat modules and $\mathcal{C} := \mathcal{F}^{\perp}$.

The notion of Ext-orthogonal classes allow us to define a particular type of preenvelope and precover.

A *special $\mathcal{C}$-preenvelope* of M is a $\mathcal{C}$-preenvelope ε such that ε is injective and $\text{Coker}\varepsilon \in {}^{\perp}\mathcal{C}$.

$$0 \to M \xrightarrow{\varepsilon} C \to C/M \to 0$$

A *special $\mathcal{C}$-precover* of M is a $\mathcal{C}$-precover ϕ such that ϕ is surjective and $\text{Ker}\phi \in \mathcal{C}^{\perp}$.

$$0 \to \text{Ker}\phi \to C \xrightarrow{\phi} M \to 0$$

A *complete cotorsion pair* in Mod-R is a cotorsion pair $(\mathcal{A}, \mathcal{B})$ such that every module in Mod-R has a special $\mathcal{A}$-precover or equivalently every module has a special $\mathcal{B}$-preenvelope. It follows that every $\mathcal{P}$-precover and $\mathcal{I}$-preenvelope is special.
Any short exact sequence of the form $0 \to M \to B \to A \to 0$ is a special $\mathcal{B}$-preenvelope and dually $0 \to B \to A \to M \to 0$ is a special $\mathcal{A}$-precover.

Examples 6.3 The following are examples of complete cotorsion pairs.

- $(\mathcal{P}, \text{Mod-}R)$ and $(\text{Mod-}R, \mathcal{I})$.
- $(\mathcal{F}, \mathcal{C})$ where $\mathcal{F}$ is the class of flat modules.
- $\left({}^{\perp}(\mathcal{S}^{\perp}), \mathcal{S}^{\perp}\right)$ where $\mathcal{S}$ is a set in Mod-R.
- More generally, for $n \geq 0$, $(\mathcal{P}_n, \mathcal{P}_n^{\perp})$ where $\mathcal{P}_n = \{M \in \text{Mod-}R : \text{p.dim}M \leq n\}$. So $M \in \mathcal{P}_n$ if there exists an exact sequence of the following form with $P_i \in \mathcal{P}$.

$$0 \to P_n \to P_{n-1} \to \cdots \to P_1 \to P_0 \to M \to 0$$

- $({}^{\perp}\mathcal{I}_n, \mathcal{I}_n)$ where $\mathcal{I}_n = \{M \in \text{Mod-}R : \text{i.dim}M \leq n\}$ where i.dim is defined analogously to p.dim above.
- $(\mathcal{F}_n, \mathcal{F}_n^{\perp})$ where $\mathcal{F}_n = \{M \in \text{Mod-}R : \text{fl.dim}M \leq n\}$ where fl.dim is defined analogously to p.dim above.

Thinking of the projective modules as the left-hand class of a complete cotorsion pair, the following theorem generalises the implication (3) $\implies$ (1) of Theorem 5.10 of Bass. Instead of stating the result in its full generality, we will state it using cotorsion pairs.

Theorem 6.4 ([Enochs and Xu, [5]]) *Let $(\mathcal{A}, \mathcal{B})$ be a complete cotorsion pair such that $\mathcal{A}$ is closed under direct limits. Then every module admits an $\mathcal{A}$-cover.*

For example, suppose $\mathcal{P} = \mathcal{F}$. Then we already have that $(\mathcal{P}, \text{Mod-}R)$ is a complete cotorsion pair and furthermore by assumption $\mathcal{P} = \mathcal{F}$ is closed under direct limits, so by the above theorem every module has a $\mathcal{P}$-cover.
Additionally, once we know that the flat cotorsion pair $(\mathcal{F}, \mathcal{C})$ is in fact a cotorsion pair and is complete, then we know that every module admits a $\mathcal{F}$-cover. This was one of the methods used to show that every module has a flat cover.
Instead, the converse is still an open problem.

Conjecture 6.5 ("Enochs Conjecture") *Let $(\mathcal{A}, \mathcal{B})$ be a complete cotorsion pair. If $\mathcal{A}$ is covering then $\mathcal{A}$ is closed under direct limits.*

By Bass's Theorem 5.10, Enoch's conjecture holds for the projective modules $\mathcal{P}$. That is, if $\mathcal{P}$ is covering then $\mathcal{P}$ is closed under direct limits.

References

[1] Hyman Bass, *Finitistic dimension and a homological generalization of semiprimary rings.* Trans. Amer. Math. Soc. 95 (1960), 466–488.

[2] Edgar E. Enochs and Overtoun M. G. Jenda, "Relative homological algebra". Volume 30 of de Gruyter Expositions in Mathematics. Walter de Gruyter & Co., Berlin, 2000.

[3] Rüdiger Göbel and Jan Trlifaj, "Approximations and endomorphism algebras of modules". Volume 1, volume 41 of de Gruyter Expositions in Mathematics. Walter de Gruyter GmbH & Co. KG, Berlin, extended edition, 2012. Approximations.

[4] C.A. Weibel, "An Introduction to Homological Algebra". Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995.

[5] Jinzhong Xu, "Flat covers of modules". Volume 1634 of Lecture Notes in Mathematics. Springer-Verlag, Berlin, 1996.

Probability and Information in Finance

Claudio Fontana

Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy
E-mail: fontana@math.unipd.it

Abstract. In this work, we aim at answering the following question: How much is an investor willing to pay in order to learn some additional information that allows to achieve arbitrage? Whenever such a value exists, we call it the value of informational arbitrage. We provide a general answer to the above question by relying on an indifference valuation approach. To this effect, we establish some new results on models with additional information and study optimal investment-consumption problems in the presence of initial information and arbitrage. We characterize when the value of informational arbitrage is universal, in the sense that it does not depend on the preference structure. This short expository paper is based on [CCF18], to which we refer for full details and proofs.

1 Introduction

The notion of information plays a particularly important role in the analysis of investment decisions. In line with economic intuition, access to more precise sources of information gives an informational advantage when trading in financial markets. The problem of quantifying such an informational advantage (which we generically call *additional information*) represents a central question in finance, dating back to the early contributions [LV68, Mor74, Wil89]. In mathematical finance, this problem has been first addressed in the seminal works [PK96, EGK97] and then substantially studied in general settings (see, e.g., [AIS98, ABS03, ADI06, Hil05]). The starting point of the present work is represented by [ABS03]. In that paper, the authors adopt an indifference valuation approach and derive a monetary value for additional information in a general semimartingale model.

The availability of additional information increases the profitability of investment strategies and, in extreme cases, can lead to the appearance of arbitrage opportunities[(1)]. This happens for instance in cases of insider trading, where arbitrage profits can be achieved by exploiting some private information on the assets traded in the market. Such extreme cases are not allowed by the setting considered in [ABS03]. In this work,

[(1)] Let us recall that an arbitrage opportunity is an admissible trading strategy requiring zero net investment that generates a non-negative non-zero payoff at some future date.

we explicitly allow for the possibility that the additional information yields arbitrage opportunities. We shall introduce and study the *value of informational arbitrage*, namely a monetary value for some additional information that can be exploited to generate arbitrage. This leads to new results on (no-)arbitrage in initially enlarged filtrations and allows to consider several interesting situations motivated by insider trading phenomena.

2 A Motivating Example

In order to illustrate some of the main concepts and tools, we present a simple example for the value of informational arbitrage. Let $W = (W_t)_{t\in[0,1]}$ be a Brownian motion on a filtered probability space $(\Omega, \mathcal{A}, \mathbf{F}, \mathbb{P})$, where $\mathbf{F} = (\mathcal{F}_t)_{t\in[0,1]}$ is the $\mathbb{P}$-augmentation of the natural filtration of W and $T = 1$ represents a fixed investment horizon. We consider an elementary financial market where a single risky asset with discounted price $S = (S_t)_{t\in[0,1]}$ is traded, with

$$S_t = \exp(W_t - t/2), \qquad \text{for all } t \in [0,1]. \tag{2.1}$$

The tuple $(\Omega, \mathbf{F}, \mathbb{P}; S)$ represents an arbitrage-free and complete financial market. Indeed, S is a $\mathbb{P}$-martingale and, hence, absence of arbitrage holds in the strong form of NFLVR (see [DS94, Fon15]). Furthermore, W has the martingale representation property on $(\Omega, \mathbf{F}, \mathbb{P})$ and, therefore, every contingent claim can be attained by self-financing trading in the market (market completeness).

In this context, the *ordinary (publicly available) information* is represented by the filtration $\mathbf{F}$, corresponding to the information generated by the price process S itself. We assume that the *additional information* is generated by the observation at time $t = 0$ of the realization of the random variable $L = \mathbf{1}_{\{W_1\geq 0\}}$. Due to (2.1), this additional information is equivalent to the knowledge at $t = 0$ of whether the final price of the risky asset will be above or below the threshold $1/\sqrt{e}$. A similar information has been considered in [PK96, Example 4.6] and [AI05, Example 2.12].

To the random variable L, we associate the *initially enlarged filtration* $\mathbf{G} = (\mathcal{G}_t)_{t\in[0,1]}$, defined as

$$\mathcal{G}_t = \mathcal{F}_t \vee \sigma(L), \qquad \text{for all } t \in [0,1]. \tag{2.2}$$

It is easy to check that the financial market $(\Omega, \mathbf{G}, \mathbb{P}; S)$ allows for arbitrage opportunities, which can also be realized by means of suitably chosen buy-and-hold strategies. Intuitively, the initial information revealed by L contains an anticipation on the final value of the stock price and, therefore, a trader having access to the information flow $\mathbf{G}$ can exploit this informational advantage when trading in the market. In this situation, we say that the random variable L leads to *informational arbitrage*. For $\mathbf{H} \in \{\mathbf{F}, \mathbf{G}\}$ and $v > 0$, we denote by $\mathcal{A}^{\mathbf{H}}(v)$ the set of all $\mathbf{H}$-predictable processes $\theta = (\theta_t)_{t\in[0,1]}$ such that θ is S-integrable on $(\Omega, \mathbf{H}, \mathbb{P})$ and the value process $V^{v,\theta} := v + \int_0^{\cdot} \theta_u \, \mathrm{d}S_u$ is non-negative. This corresponds to self-financing trading starting from initial wealth v, under a solvability constraint.

Let us consider an agent with initial wealth $v > 0$ and preferences described by a strictly increasing and concave utility function $U : (0, +\infty) \to \mathbb{R}$. Suppose that at $t = 0$,

before the beginning of trading, the agent has the possibility of learning the content of the additional information L by paying a price $\pi > 0$. The agent has to choose between the following two options:

(i) Construct a self-financing portfolio with initial wealth v on the basis of the ordinary information $\mathbf{F}$, solving the problem

$$\sup_{\theta \in \mathcal{A}^{\mathbf{F}}(v)} \mathbb{E}\big[U\big(V_1^{v,\theta}\big)\big] =: u^{\mathbf{F}}(v). \tag{2.3}$$

(ii) Pay the price π in order to learn the information L and then construct a self-financing portfolio with the residual capital $v-\pi$ by relying on the information $\mathbf{G}$, solving the problem

$$\sup_{\theta \in \mathcal{A}^{\mathbf{G}}(v-\pi)} \mathbb{E}\big[U\big(V_1^{v-\pi,\theta}\big)\big] =: u^{\mathbf{G}}(v-\pi). \tag{2.4}$$

Note that, since S is a martingale on $(\Omega, \mathbf{F}, \mathbb{P})$, problem (2.3) is solved by the trivial strategy $\theta^{*,\mathbf{F}} \equiv 0$, so that $u^{\mathbf{F}}(v) = U(v)$. We denote by $\theta^{*,\mathbf{G}}$ the optimal strategy for problem (2.4).

In the spirit of [ABS03], we want to determine the maximal price $\pi(v)$ that an agent is ready to pay for learning the additional information L, determined as the solution $\pi = \pi(v) \in \mathbb{R}_+$ to the equation

$$u^{\mathbf{G}}(v-\pi) = u^{\mathbf{F}}(v). \tag{2.5}$$

We call the quantity $\pi(v)$ the *value of informational arbitrage*. Since the knowledge of L allows to exploit arbitrage opportunities, we have that $u^{\mathbf{G}}(v) > u^{\mathbf{F}}(v)$, for every $v > 0$. This implies that an investor will always be willing to pay some strictly positive price to learn the realization of L before starting to trade. In the context of the present example, we can state the following result.

Theorem 2.1 *Let $U : (0, +\infty) \to \mathbb{R}$ be any strictly increasing and concave utility function. Then, for every $v > 0$, it holds that*

$$\pi(v) = v/2.$$

Moreover, the optimal strategy for every informed agent with initial wealth v is explicitly given by

$$\theta_t^{*,\mathbf{G}} = \big(\mathbf{1}_{\{W_1 \geq 0\}} - \mathbf{1}_{\{W_1 < 0\}}\big) \frac{1}{\sqrt{2\pi(1-t)}} \exp\left(-\frac{W_t^2}{2(1-t)}\right) \frac{v}{S_t}, \qquad \text{for } t \in [0,1].$$

The most striking aspect of the above result is that there is no dependency on the utility function U. In this sense, the value $v/2$ represents a *universal* value of informational arbitrage. Similarly, the strategy $\theta^{*,\mathbf{G}}$ is optimal for every utility function U and has several interesting features:

(i) it represents an arbitrage opportunity for the informed agent, since $V_t^{0,\theta^*,\mathbf{G}} \geq -v/2$ for all $t \in [0,1]$ and $V_1^{0,\theta^*,\mathbf{G}} = v/2 > 0$ a.s. Moreover, it corresponds to the strategy realizing the optimal arbitrage in $(\Omega, \mathbf{G}, \mathbb{P}; S)$, in the sense of [CT15];

(ii) it generates the numéraire portfolio (see [KK07]) for the financial market $(\Omega, \mathbf{G}, \mathbb{P}; S)$;

(iii) the strategy is always long or short in the risky asset depending on the content of the additional information L revealed before the beginning of trading;

(iv) the strategy is a bet on the risky asset: the position on the risky asset increases if the asset price decreases and, vice versa, decreases if the asset price increases;

(v) it holds that $\lim_{t\to 1} \theta_t^{*,\mathbf{G}} = 0$, meaning that the position in the risky asset is fully liquidated at the end of the investment horizon.

Proof. Referring to [CCF18] for a complete proof of Theorem 2.1, let us briefly discuss the intuition behind Theorem 2.1. As mentioned above, it holds that $u^{\mathbf{F}}(v) = U(v)$. On the other hand, suppose that at $t = 0$ one pays the price $\pi(v) = v/2$ in order to learn the realization of $L = \mathbf{1}_{\{W_1 \geq 0\}}$ and starts trading according to the strategy $\theta^{*,\mathbf{G}}$. By applying Itô's formula, it can be verified that the associated wealth process is given by

$$V_t^{v/2,\theta^*,\mathbf{G}} := \frac{v}{2} + \int_0^t \theta_u^{*,\mathbf{G}} \,\mathrm{d}S_u = v\left(\Phi\left(\frac{-W_t}{\sqrt{1-t}}\right)\mathbf{1}_{\{W_1<0\}} + \Phi\left(\frac{W_t}{\sqrt{1-t}}\right)\mathbf{1}_{\{W_1\geq 0\}}\right),$$

for all $t \in [0,1]$, where Φ denotes the distribution function of a standard Normal random variable. In particular, it holds that

$$V_0^{v/2,\theta^*,\mathbf{G}} = v/2 \qquad \text{and} \qquad V_1^{v/2,\theta^*,\mathbf{G}} = v \qquad \text{a.s.}$$

This shows that, in the presence of the additional information L, an initial capital of $v/2$ is sufficient to reach a final wealth equal to v with probability one. This is possible since L generates informational arbitrage. In turn, this implies that

$$u^{\mathbf{G}}(v/2) \geq \mathbb{E}\left[U\left(\frac{v}{2} + \int_0^1 \theta_u^{*,\mathbf{G}} \,\mathrm{d}S_u\right)\right] = U(v).$$

Conversely, for every strategy $\theta \in \mathcal{A}^{\mathbf{G}}(v/2)$, Jensen's inequality implies that

$$\begin{aligned}
\mathbb{E}\left[U\left(\frac{v}{2} + \int_0^1 \theta_u \,\mathrm{d}S_u\right)\right] \;&\leq\; U\left(\mathbb{E}\left[\frac{v}{2} + \int_0^1 \theta_u \,\mathrm{d}S_u\right]\right) \\
&=\; U\left(2\,\mathbb{E}\left[\left(\frac{v}{2} + \int_0^1 \theta_u \,\mathrm{d}S_u\right)\Big| W_1 \geq 0\right] + 2\,\mathbb{E}\left[\left(\frac{v}{2} + \int_0^1 \theta_u \,\mathrm{d}S_u\right)\Big| W_1 < 0\right]\right) \\
&\leq\; U(v).
\end{aligned}$$

Taking the supremum over all $\theta \in \mathcal{A}^{\mathbf{G}}(v/2)$, we obtain the inequality $u^{\mathbf{G}}(v/2) \leq U(v)$. This proves the claim that $u^{\mathbf{G}}(v/2) = u^{\mathbf{F}}(v)$, so that $\pi(v) = v/2$. □

3 The General Semimartingale Framework

Motivated by the example considered in Section 2, we now discuss the value of informational arbitrage in a general semimartingale setting, with respect to a general additional information.

3.1 Setting

We consider a filtered probability space $(\Omega, \mathcal{A}, \mathbf{F} = (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$ supporting a d-dimensional non-negative semimartingale $S = (S_t)_{t\in[0,T]}$, representing the discounted prices of d risky assets, with investment horizon $T < +\infty$. As in Section 2, the filtration $\mathbf{F}$ represents the ordinary information flow. We suppose that there exists a unique Equivalent Local Martingale Measure (ELMM) $\mathbb{Q}$ for S on $(\Omega, \mathcal{F}_T)$, under which S is a $\mathbb{Q}$-local martingale. The existence of an ELMM implies that the financial market $(\Omega, \mathbf{F}, \mathbb{P}; S)$ is arbitrage-free (in the sense of NFLVR), while the uniqueness of $\mathbb{Q}$ implies that every $\mathbb{Q}$-local martingale can be represented as a stochastic integral of S. The latter property corresponds to the completeness of the financial market $(\Omega, \mathbf{F}, \mathbb{P}; S)$.

3.2 The additional information filtration

The additional information is represented by an $\mathcal{A}$-measurable random variable L, taking values in a Lusin space E and with unconditional law λ. The corresponding enlarged filtration $\mathbf{G} = (\mathcal{G}_t)_{t\in[0,T]}$ is defined as in (2.2), i.e., the smallest filtration containing $\mathbf{F}$ and such that L is $\mathcal{G}_0$-measurable. For each $t \in [0,T]$, we denote by ν_t a regular version of the $\mathcal{F}_t$-conditional law of L. We shall work under the following standing assumption, which corresponds to the well-known *density hypothesis* introduced in the seminal work [Jac85].

Assumption 3.1 For all $t \in [0,T]$, it holds that $\nu_t \ll \lambda$ in the a.s. sense.

Assumption 3.1 implies the existence of a regular family of densities $q : \Omega \times [0,T] \times E \to \mathbb{R}_+$ such that $\nu_t(\mathrm{d}x) = q_t^x \, \lambda(\mathrm{d}x)$ a.s. for all $t \in [0,T]$. In the present context, Assumption 3.1 has several fundamental consequences. First and foremost, as shown in [Jac85], Assumption 3.1 implies the validity of the so-called *H'-hypothesis* (i.e., every $\mathbf{F}$-semimartingale is also a $\mathbf{G}$-semimartingale). In a frictionless financial market, the failure of the semimartingale property is incompatible with the solution of portfolio optimization problems. In addition, Assumption 3.1 allows to prove a new martingale representation result in the initially enlarged filtration $\mathbf{G}$.

Proposition 3.2 *Let $M = (M_t)_{t\in[0,T]}$ be a local martingale on $(\Omega, \mathbf{G}, \mathbb{P})$. Then there exists a $\mathbf{G}$-predictable S-integrable process $K = (K_t)_{t\in[0,T]}$ such that*

$$M_t = \frac{Z_t}{q_t^L}\left(M_0 + \int_0^t K_u \, \mathrm{d}S_u\right) \qquad \text{a.s. for all } t \in [0,T],$$

where $Z = (Z_t)_{t\in[0,T]}$ denotes the density process of $\mathbb{Q}$ on $(\Omega, \mathbf{F}, \mathbb{P})$.

In the proof of Proposition 3.2, the results of [Fon18] play a central role. In the present setting, the relevance of this proposition consists in the fact that it allows to transfer the

martingale representation property from the original space $(\Omega, \mathbf{F}, \mathbb{Q})$ onto the enlarged space $(\Omega, \mathbf{G}, \mathbb{P})$.

The following result, which relies on [AFK16], will play a fundamental role in the analysis of the (no-)arbitrage properties of the financial market in the presence of additional information.

Theorem 3.3

(i) *NUPBR holds in* $\mathbf{G}$ *if and only if the set* $\{q^x = 0 < q^x_-\}$ *is evanescent for* λ*-a.e.* $x \in E$.

(ii) *NFLVR holds in* $\mathbf{G}$ *if and only if* $q^x_T > 0$ *a.s. for* λ*-a.e.* $x \in E$.

Moreover, if the condition appearing in part (i) *holds, then the financial market* $(\Omega, \mathbf{G}, \mathbb{P}; S)$ *admits a unique* equivalent local martingale deflator *(ELMD), given by the process* Z/q^L.[2]

Remark 3.4 The necessary and sufficient conditions appearing in Theorem 3.3 closely resemble analogous conditions for the absence of arbitrage under absolutely continuous changes of probabilities, as considered in [Fon14]. This is due to the deep relation existing between enlargement of filtrations and absolutely continuous changes of probabilities, see [Jac85, Son13, Yoe85].

3.3 Admissible portfolios and preferences

We fix a *stochastic clock* $\kappa = (\kappa_t)_{t\in[0,T]}$, which is a non-decreasing $\mathbf{F}$-adapted bounded process with $\kappa_0 = 0$ and such that $\mathbb{P}(\kappa_T > 0|\sigma(L)) > 0$ a.s. The stochastic clock κ represents the notion of time according to which consumption is assumed to occur. A *portfolio* is defined as a triplet $\Pi = (v, \theta, c)$, where $v > 0$ represents the initial capital, $\theta = (\theta_t)_{t\in[0,T]}$ is an S-integrable process representing the holdings in the d risky assets and $c = (c_t)_{t\in[0,T]}$ is a non-negative process representing the consumption rate (with respect to the clock κ). For an ordinary agent, the strategy θ and the consumption process c are required to be measurable with respect to the $\mathbf{F}$-predictable and $\mathbf{F}$-optional sigma-fields, respectively. On the other hand, an informed agent is allowed to construct portfolios by choosing $\mathbf{G}$-predictable strategies θ and $\mathbf{G}$-optional consumption processes c. The value process $V^{v,\theta,c} = (V^{v,\theta,c}_t)_{t\in[0,T]}$ of a portfolio $\Pi = (v, \theta, c)$ is defined as

$$V^{v,\theta,c}_t := v + \int_0^t \theta_u \, \mathrm{d}S_u - \int_0^t c_u \, \mathrm{d}\kappa_u, \qquad \text{for all } t \in [0,T].$$

Given initial wealth $v > 0$, a couple (θ, c) is said to be *admissible* if $V^{v,\theta,c}_t \geq 0$ a.s. for all $t \in [0,T]$, denoted by $(\theta, c) \in \mathcal{A}^{\mathbf{H}}(v)$, for $\mathbf{H} \in \{\mathbf{F}, \mathbf{G}\}$.

We assume that preferences are defined with respect to intermediate consumption and/or terminal wealth. To this effect, we introduce a stochastic utility field $U : \Omega \times$

[2] We recall that, in the financial market $(\Omega, \mathbf{G}, \mathbb{P}; S)$, an equivalent local martingale deflator is a strictly positive $\mathbf{G}$-local martingale $Z^{\mathbf{G}} = (Z^{\mathbf{G}}_t)_{t\in[0,T]}$ with $Z^{\mathbf{G}}_0 = 1$ such that ZS is a $\mathbf{G}$-local martingale. The existence of an ELMD is a weaker property than the existence of an ELMM (see, e.g., [Fon15]) and is equivalent to the validity of NUPBR (see [KK07]).

$[0,T] \times \mathbb{R}_+ \to \mathbb{R} \cup \{-\infty\}$, satisfying suitable technical requirements (see [CCF18] for full details). For $\mathbf{H} \in \{\mathbf{F}, \mathbf{G}\}$, the optimal investment-consumption problem of an agent having access to the information flow $\mathbf{H}$ and starting with initial wealth $v > 0$ is given by

$$u^{\mathbf{H}}(v) := \sup_{(\theta,c)\in\mathcal{A}^{\mathbf{H}}(v)} \mathbb{E}\left[\int_0^T U(\omega, u, c_u(\omega))\,\mathrm{d}\kappa_u(\omega)\right]. \tag{3.1}$$

In order to ensure that the investment-consumption problem (3.1) is well-posed in $\mathbf{G}$, we shall work under the *standing assumption* that the condition appearing in part (i) of Theorem 3.3 holds. Problem (3.1) can then be solved by means of martingale methods, taking into account the possible presence of arbitrage and non-trivial initial information. In the general incomplete market case, problem (3.1) can be solved by means of convex duality techniques, by relying on [CCFM17]. For logarithmic and power utility functions, problem (3.1) can be shown to admit explicit solutions, fully characterized in terms of the processes κ, Z and $q^x|_{x=L}$, thereby generalizing [ABS03, Corollary 4.7].

3.4 The value of informational arbitrage

As in Section 2, our main goal consists in studying the *value of informational arbitrage*, defined as the solution $\pi = \pi(v) \in \mathbb{R}_+$ to the equation

$$u^{\mathbf{F}}(v) = u^{\mathbf{G}}(v - \pi).$$

In this section, we present a short outline of the main results on the value of informational arbitrage obtained in [CCF18]. First, as long as the optimal investment-consumption problem (3.1) is well-posed in $\mathbf{G}$, the utility indifference price $\pi(v)$ always exists, is finite and unique, for every $v > 0$. Moreover, for every utility stochastic field, $\pi(v)$ is always strictly positive if the additional information L generates arbitrage opportunities in $\mathbf{G}$. Under suitable conditions, we also derive universal lower and upper bounds for $\pi(v)$.

In the case of logarithmic and power utility functions, the value $\pi(v)$ can be explicitly computed. For instance, if κ is deterministic and $U(\omega, t, y) = \log(y)$, for all $(\omega, t) \in \Omega \times [0,T]$, it holds that

$$\pi^{\log}(v) = v\left(1 - \exp\left(-\frac{\int_0^T \mathbb{E}[\log(q_u^L)]\,\mathrm{d}\kappa_u}{\kappa_T}\right)\right).$$

Several interesting properties of the value of informational arbitrage can be deduced, for instance:

(a) the value of informational arbitrage is increasing with respect to the investor's initial wealth;

(b) in the presence of intermediate consumption, the value of informational arbitrage is lower than in the case of utility from terminal wealth only (compare with [LPS10]);

(c) $\pi^{\log}(v)$ is related to the Shannon information between L and $\mathcal{F}_T$, which reduces to the entropy of L whenever L is a discrete $\mathcal{F}_T$-measurable random variable (see [ADI06]).

Finally, coming back to the example considered in Section 2, the most striking feature of the example is represented by the fact that the value of informational arbitrage is universal, in the sense that it does not depend on the utility function. This situation is fully characterized by the following result.

Theorem 3.5 *Suppose that $\kappa = \delta_T$ and $\mathbb{Q} = \mathbb{P}$. Then the following are equivalent:*

(i) *the random variable q_T^L is a.s. equal to a constant;*

(ii) *for every $v > 0$, there exists a value $\pi(v)$ such that, for every increasing concave utility function U, it holds that $u^{\mathbf{F}}(v) = u^{\mathbf{G}}(v - \pi(v))$.*

In this case, for every utility function U, the indifference value $\pi(v)$ is given by the universal value

$$\pi(v) = v\left(1 - \frac{1}{q_T^L}\right).$$

Furthermore, under the conditions of Theorem 3.5, it can be shown that the optimal wealth process for an informed agent with initial wealth v who acquires the additional information at the indifference price $\pi(v) = v(1 - 1/q_T^L)$ is given by vq_t^L/q_T^L a.s., for all $t \in [0, T]$. The proof is based on a stochastic dominance argument. It is interesting to remark that the result of Theorem 3.5 is non-trivial only if the random variable L generates arbitrage opportunities in $\mathbf{G}$.

In [CCF18] we illustrate the above results by means of several explicit examples, in the context of continuous as well as discontinuous price processes and multi-dimensional financial markets. In particular, we discuss a general class of models where the conditions of Theorem 3.5 are satisfied.

References

[ABS03] J. Amendinger, D. Becherer, and M. Schweizer, *A monetary value for initial information in portfolio optimization.* Finance Stoch., 7/1 (2003), 29–46.

[ADI06] S. Ankirchner, S. Dereich, and P. Imkeller, *The Shannon information of filtrations and the additional logarithmic utility of insiders.* Ann. Probab. 34/2 (2006), 743–778.

[AFK16] B. Acciaio, C. Fontana, and C. Kardaras, *Arbitrage of the first kind and filtration enlargements in semimartingale financial models.* Stoch. Proc. Appl. 126/6 (2016), 1761–1784.

[AI05] S. Ankirchner and P. Imkeller, *Finite utility on financial markets with asymmetric information and structure properties of the price dynamics.* Stoch. Proc. Appl. 41/3 (2005), 479–503.

[AIS98] J. Amendinger, P. Imkeller, and M. Schweizer, *Additional logarithmic utility of an insider.* Stoch. Proc. Appl. 75/2 (1998), 263–286.

[CCF18] H.N. Chau, A. Cosso, and C. Fontana, *The value of informational arbitrage.* Preprint (available at `https://arxiv.org/abs/1804.00442`), 2018.

[CCFM17] H.N. Chau, A. Cosso, C. Fontana, and O. Mostovyi, *Optimal investment with intermediate consumption under no unbounded profits with bounded risk.* J. Appl. Prob. 54/3 (2017), 710–719.

[CT15] H.N. Chau and P. Tankov, *Market models with optimal arbitrage.* SIAM J. Financ. Math. 6 (2015), 66–85.

[DS94] F. Delbaen and W. Schachermayer, *A general version of the fundamental theorem of asset pricing.* Math. Ann. 300/1 (1994), 463–520.

[EGK97] R.J. Elliott, H. Geman, and B.M. Korkie, *Portfolio optimization and contingent claim pricing with differential information.* Stoch. Stoch. Rep. 60/3-4 (1997), 185–203.

[Fon14] C. Fontana, *No-arbitrage conditions and absolutely continuous changes of measure.* In C. Hillairet, M. Jeanblanc, and Y. Jiao, editors, "Arbitrage, Credit and Informational Risks", volume 5 of Peking University Series in Mathematics, pp. 3–18. World Scientific, Singapore, 2014.

[Fon15] C. Fontana, *Weak and strong no-arbitrage conditions for continuous financial markets.* Int. J. Theor. Appl. Finance 18/1 (2015), 1550005.

[Fon18] C. Fontana, *The strong predictable representation property in initially enlarged filtrations under the density hypothesis.* Stoch. Proc. Appl. 128/3 (2018), 1007–1033.

[Hil05] C. Hillairet, *Comparison of insiders' optimal strategies depending on the type of side-information.* Stoch. Proc. Appl. 115/10 (2005), 1603–1627.

[Jac85] J. Jacod, *Grossissement initial, hypothèse (H'), et théorème de Girsanov.* In T. Jeulin and M. Yor editors, "Grossissements de Filtrations: Exemples et Applications", volume 1118 of Lecture Notes in Mathematics, pp. 15–35. Springer, Berlin-Heidelberg, 1985.

[KK07] I. Karatzas and C. Kardaras, *The numéraire portfolio in semimartingale financial models.* Finance Stoch. 11/4 (2007), 447–493.

[LPS10] J. Liu, E. Peleg, and A. Subrahmanyam, *Information, expected utility, and portfolio choice.* J. Financ. Quant. Anal. 45/5 (2010), 1221–1251.

[LV68] I.H. La Valle, *On cash equivalents and information evaluation in decisions under uncertainty: Part I: Basic theory.* J. Am. Stat. Assoc. 63/321 (1968), 252–276.

[Mor74] J.R. Morris, *The logarithmic investor's decision to acquire costly information.* Manage. Sci. 21/4 (1974), 383–391.

[PK96] I. Pikovsky and I. Karatzas, *Anticipative portfolio optimization.* J. Appl. Prob. 28/4 (1996), 1095–1122.

[Son13] S. Song, *Local solution method for the problem of enlargement of filtration.* Preprint (available at `https://arxiv.org/abs/1302.2862`), 2013.

[Wil89] M. Willinger, *Risk aversion and the value of information.* J. Risk Insur. 56/2 (1989), 320–328.

[Yoe85] C. Yoeurp, *Théorème de Girsanov généralisé et grossissement d'une filtration.* In T. Jeulin and M. Yor editors, "Grossissements de Filtrations: Exemples et Applications", volume 1118 of Lecture Notes in Mathematics, pp. 172–196. Springer-Verlag, 1985.

An introduction to Riemann-Hilbert correspondence

Davide Barco (*)

Abstract. The 21st Hilbert problem concerns the existence of a certain class of linear differential equations on the complex affine line with specified singular points and monodromic groups. Arising both as an answer and an extension to this issue, Riemann-Hilbert correspondence aims to establish a relation between systems of linear differential equations defined on a complex manifold and suitable algebraic objects encoding topological properties of the same systems. The goal was first achieved for systems with regular singularities, thanks to the works by Deligne, Kashiwara and Mebkhout. Moreover, Deligne and Malgrange established a generalized correspondence (called Riemann-Hilbert-Birkhoff correspondence) for systems with irregular singularities on complex curves, encoding and describing the Stokes phenomenon which arises in this case. In more recent years, the correspondence has been extended to take account of irregular points on complex manifolds of any dimension by D'Agnolo and Kashiwara.
In this short note we give a basic introduction on the subject by providing concepts and classical example from the theory.

The original 21st Hilbert problem was stated in the following way:

"In the theory of linear differential equations with one independent variable z, I wish to indicate an important problem, one which very likely Riemann himself may have had in mind. This problem is as follows: To show that there always exists a linear differential equation of the Fuchsian class, with given singular points and monodromic group."

Arising both as an answer and an extension to this issue, Riemann-Hilbert correspondence aims to establish a relation between systems of linear differential equations defined on a complex manifold X and suitable algebraic objects encoding topological properties of the same systems.

We are then concerned with analytic linear differential equations, that is, equations of the form

$$a_n(z)\partial_z^n f + a_{n-1}\partial_z^{n-1} f + ... + a_0(z) f = 0 \tag{1}$$

where $a_i(z) \in \mathcal{O}_U \ \forall i = 0, ..., n$, $z \in U$, U open connected subset of $\mathbb{C}$ and ∂_z the derivation

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: davide.barco@math.unipd.it . Seminar held on May 8th, 2019.

associated with z, i.e. $\partial_z = \frac{d}{dz}$.

It is usual to classify the points in U with respect to a differential equation in the following way.

Definition Let $z_0 \in U$ and consider the differential equation (1). Then

- z_0 is said to be an **ordinary** point if $a_n(z_0) \neq 0$
- z_0 is said to be a **regular singular** point if $a_n(z_0) = 0$ and

$$\operatorname*{ord}_{z=z_0} a_n - n \leq \operatorname*{ord}_{z=z_0} a_j - j \quad \forall j = 0, ..., n-1$$

- z_0 is a **irregular singular** point otherwise.

Notice that, thanks to Cauchy theorem, we are always able to solve the differential equation on any simply connected open subset of U not containing singular points.

The above differential equation setting can be expressed in a more algebraic way by introducing the ring of differential operators and its module theory.

Definition The **ring of holomorphic differential operators** $\mathcal{D}_U$ is the subalgebra of $End_{\mathbb{C}}(\mathcal{O}_U)$ generated by $\mathcal{O}_U$ and ∂_z. More explicitly

$$\mathcal{D}_U = \bigoplus_{k \in \mathbb{N}} \mathcal{O}_U \partial_z^k$$

and its elements are of the form

$$P = a_n(z)\partial_z^n + a_{n-1}(z)\partial_z^{n-1} + ... + a_0(z)$$

where $a_i(z) \in \mathcal{O}_U$, $\forall i = 0, ..., n$

Let us see the relation between these two different descriptions.

Let $P \in \mathcal{D}_U$, we associate to this operator the left $\mathcal{D}_U$-module $\mathcal{M}_P := \frac{\mathcal{D}_U}{\mathcal{D}_U P}$, quotient of $\mathcal{D}_U$ by the ideal generated by P.

In this setting, if we consider the set $\mathcal{S}ol_U(\mathcal{M}_P) := Hom_{\mathcal{D}_U}(\mathcal{M}_P, \mathcal{O}_U)$, we get an isomorphism

$$Hom_{\mathcal{D}_U}(\mathcal{M}_P, \mathcal{O}_U) \simeq \{\psi \in Hom_{\mathcal{D}_U}(\mathcal{D}_U, \mathcal{O}_U) \mid \psi(P) = 0\}$$

We see then, by $Hom_{\mathcal{D}_U}(\mathcal{D}_U, \mathcal{O}_U) \simeq \mathcal{O}_U$ $(\psi \to \psi(1))$ that

$$Hom_{\mathcal{D}_U}(\mathcal{M}_P, \mathcal{O}_U) \simeq \{f \in \mathcal{O}_U \mid Pf = 0\}$$

In this way, we have shown how $\mathcal{M}_P$ encodes all information about the differential equation associated to P.

Let us describe some examples.

- Consider $P = \partial_z z \partial_z = z\partial_z^2 + \partial_z$: this equation has a regular singular point in 0.

  Notice moreover that, in each open ball not containing 0, a local basis for the solutions of the differential equation $Pf = 0$ is given by $\{1, \log(z)\}$, where $\log(z) = \log(|z|) + i \cdot \arg(z)$ is the principal value of the complex logarithm.

  The complete description of $\mathcal{S}ol_U(\mathcal{M}_P)$ for $U = B(z_0, r)$ is then

$$\mathcal{S}ol_U(\frac{\mathcal{D}_U}{\mathcal{D}_U(z\partial_z^2 + \partial_z)}) = \begin{cases} \mathbb{C} \cdot 1 + \mathbb{C} \cdot \log(z) & \text{if } z_0 \neq 0 \text{ and } 0 \notin U \\ \mathbb{C} & \text{if } z_0 = 0 \end{cases} \tag{2}$$

  Let γ be the unit circle in $\mathbb{C}$, $1 \in \gamma$ and $f(z) = \alpha \cdot 1 + \beta \cdot \log(z)$ a solution of $z\partial_z^2 + \partial_z$ in a small ball centered in z_0. If we perform analytic continuation of f along γ, we obtain a function $\tilde{f}(z)$ in a neighbourhood of z_0 defined as follows:

$$\tilde{f}(z) = \alpha \cdot 1 + \beta(\log(z) + 2\pi i) = (\alpha + 2\pi i\beta) \cdot 1 + \beta \cdot \log(z)$$

  This corresponds exactly to a change of basis from $\{1, \log(z)\}$ to $\{1, \log(z) + 2\pi i\}$, which can be described by

$$\begin{bmatrix} 1 & 2\pi i \\ 0 & 1 \end{bmatrix} \tag{3}$$

  We will call this matrix the **monodromy matrix** associated to $z\partial_z^2 + \partial_z$

- Consider as another example $P = z\partial_z - \lambda$: 0 is still regular singular.

  In this case, a local basis for the solution of $Pf = 0$ is given by $\{z^\lambda\}$, where $z^\lambda = e^{\lambda \log(z)}$.

  Notice that, if $\lambda \in \mathbb{N}$, then the solutions are monomials of degree λ: they are therefore defined in the whole complex line and 0 is then a 'false' singular point.

  If $\lambda \in \mathbb{C} \setminus \mathbb{N}$, the analytic continuation of $f(z) = \alpha z^\lambda$ along γ is then $\tilde{f}(z) = \alpha e^{\lambda(\log(z)+2\pi i)} = \alpha e^{2\pi i\lambda} z^\lambda$ and the corresponding monodromy matrix is $(e^{2\pi i\lambda})$.

  Again, notice that if $\lambda \in \mathbb{Z}_{<0}$, solutions are of the kind $\frac{\alpha}{z^\lambda}$: there is no monodromy involved and solution can be defined in a whole punctured neighbourhood of 0.

More in general, suppose $P \in \mathcal{D}_U$ with $\{z_1, ...z_m\}$ as regular singular points, z_0 ordinary point for P.

Recall that $\pi_1(U \setminus \{z_1, ..., z_m\}, z_0) \simeq \mathbb{Z}^m$, where a basis is provided by the classes $[\gamma_1]$ of the loops γ_i based in z_0 and encircling only the singular point z_i, $i = 1, ..., m$.

The **monodromy representation**

$$Mon_P : \pi_1(U \setminus \{z_1, ..., z_m\}, z_0) \to GL_n(\mathbb{C})$$

is then a morphism of groups associating to each $[\gamma_i]$ the monodromy matrix coming from the comparison between a local basis for the solutions at z_0 and its analytic continuation along a representative of $[\gamma_i]$.

From the examples above, it is clear that we can associate to each differential equation its monodromic representation: what Hilbert aimed at was then solving the inverse problem in dimension 1.

It is possible to pose such a question also in higher dimensions.

The main ingredient for the topological part is the equivalent in higher dimensions of the datum of $\mathrm{Sol}(\mathcal{M}_P)$ together with a monodromic representation: such objects are called **$\mathbb{C}$-constructible sheaves** and their bounded derived category is denoted by $D^b_{\mathbb{C}-c}(\mathbb{C}_X)$.

On the differential equation side, we can also extend the definition of the ring of holomorphic differential operators to a general analytic manifold X.

Then, analytic differential operators are replaced by **holonomic $\mathcal{D}$-modules**, while analytic differential operators with only regular singularities are replaced by **regular holonomic $\mathcal{D}$-modules**.

Their bounded derived categories are respectively denoted by $D^b_h(\mathcal{D}_X)$ and $D^b_{rh}(\mathcal{D}_X)$.

We are then able to state the Hilbert problem using these categories.

The following fundamental result, known as Riemann-Hilbert correspondence for regular holonomic $\mathcal{D}$-modules, constitutes then the positive answer to this classical question.

Proposition (Kashiwara [3]) *Let X be an analytic manifold. Then the functor*

$$\mathcal{S}ol_X := R\mathcal{H}om_{\mathcal{D}_X}(\cdot, \mathcal{O}_X) : D^b_{rh}(\mathcal{D}_X) \to D^b_{\mathbb{C}-c}(\mathbb{C}_X)$$

is an equivalence of categories. Moreover, by restricting to the hearts of suitable t-structures on both sides, this provides an equivalence of categories between the full subcategories $Mod_{rh}(\mathcal{D}_X)$ and $Perv(X)$ (perverse sheaves).

When trying to extend the above result to include differential equations with irregular singular points, several issues arise.

Consider for example $P = z^2\partial_z + 1$ in $\mathbb{C}$: 0 is the only singular point, it is irregular singular and a local basis for the solutions is given by $e^{\frac{1}{z}}$.

First, notice that all the solutions have an essential singularity at 0: this is a general feature which distinguishes regular and irregular singularities.

As such, solution displays different growth behaviour in the two cases.

- In the case of regular singularities, solutions are of moderate growth in all sectors S departing from 0:

$$\exists N \in \mathbb{N},\ C > 0 \ \text{ s.t. } |f(z)| < C|z|^{-N} \quad \forall z \in S$$

- In the case of irregular singularities, solutions are instead exponentially decreasing in certain sectors and exponentially increasing in others.

The second noticeable thing is that the local system arising as solution of the $\mathcal{D}$-module $\mathcal{M}_P$ is the same as the local system $\mathcal{S}ol(\mathcal{M}_{z\partial_z+1})$.

In other words, $\mathcal{S}ol$ is not able to distinguish between the two differential equations and the equivalence as written above can not be extended to take care of irregular singularities.

However, there is another feature displayed by such differential operators, the so-called *Stokes phenomenon*.

In order to give an idea of what the Stokes phenomenon is, we need a better understanding of the structure of the solutions of a differential equation.

In particular, we need to have information about formal solutions, i.e. solutions obtained by requiring that a series solves the equation without necessarily converging.

There is a classical result concerning the formal structure of solutions for a differential operator in dimension 1 having an irregular singular point.

Theorem (Hukuhara-Levelt-Turrittin; see [5]) *Suppose $0 \in U$ is irregular singular for the differential operator P. There exist a basis of formal solutions $\hat{F}$ for $Pf = 0$ given by*

$$\hat{f}_i(z) = e^{\phi_i(\frac{1}{z^d})} z^{\lambda_i} F_i(z^{\frac{1}{d}}),\ i = 1, ..., n$$

where $d \in \mathbb{N}_{>0}$, $\lambda_i \in \mathbb{C}$, $\phi_i \in t^{-1}\mathbb{C}[t^1]$ and F_i are formal power series.

The ϕ_i's are called **exponential factors**.

It is usual to associate to each couple of them two set of directions: they are the directions at which Stokes phenomenon appears.

Define the **Stokes line** relative to the pair of exponential factors (ϕ_i,ϕ_j) by the condition $Re(\phi_i - \phi_j) = 0$: along this line the magnitude of e^{ϕ_i} and e^{ϕ_j} is the same.

Define the **anti-Stokes line** relative to the pair of exponential factors (ϕ_i, ϕ_j) by the condition $Im(\phi_i - \phi_j) = 0$: when crossing this line, e^{ϕ_i} and e^{ϕ_j} go from a dominant (subdominant) to a subdominant (dominant) behaviour with respect to each other.

You can improve the preceding result. However, there is a price to pay: you have to choose a direction.

Lemma (Borel, Ritt) *Let $\theta \in S^1$. There exists a sector S centered at the direction θ and $F = \{f_i \in \mathcal{O}_S\}_{i=1,...,n}$ such that F is a basis for the solution of $Pf = 0$ in S and $f_i \sim \hat{f}_i$, that is, $\forall z \in S$ and $i = 1, ..., n$ the following holds*

$$\forall N \in \mathbb{N}\ \exists C > 0 \text{ s.t. } |f_i(z) - \hat{f}_i^N(z)| < C|e^{\phi_i(z)} z^{\lambda_i + \frac{N}{d}}| \tag{4}$$

where $\hat{f}_i^N$ denotes $\hat{f}_i$ in which the formal term F_i have been truncated at the N-th term.

We now have all necessary ingredients: let us describe the Stokes phenomenon with an example.

Consider the differential operator $\partial_z^2 - z$: the associated differential equation is called **Airy equation**, it has an irregular singular point at ∞.

Its exponential factors are $\phi_\pm = \pm\frac{2}{3}z^{\frac{3}{2}}$. Its solutions are two entire holomorphic functions called $Ai(z)$ and $Bi(z)$.

The restrictions $Ai(x)$ and $Bi(x)$ to the real line of such functions is of interest in physics.

In 1857 Stokes noticed that the asymptotic behaviours of $Ai(x)$ for $x \to \pm\infty$ were different.

$$Ai(x) \approx \frac{1}{\sqrt{\pi}}|x|^{-\frac{1}{4}} \cos(\frac{2}{3}|x|^{\frac{3}{2}} - \frac{\pi}{4}),\ x \to -\infty$$

$$Ai(x) \approx \frac{1}{2\sqrt{\pi}}|x|^{-\frac{1}{4}} e^{-\frac{2}{3}|x|^{\frac{3}{2}}},\ x \to \infty$$

The explanation for such a change in the asymptotics can be found in the complex line.

In the case of Airy equation, the Stokes lines for (ϕ_-, ϕ_+) are at directions $-\frac{\pi}{3}, \pi, \frac{5\pi}{3}$, while the anti-Stokes lines at directions $0, \frac{2\pi}{3}, \frac{4\pi}{3}$.

When looking at the complex picture and crossing the anti-Stokes direction $\frac{2\pi}{3}$, the function $Ai(z)$ has a change of asymptotic lift from

$$Ai(z) = \frac{1}{2\sqrt{\pi}} z^{-\frac{1}{4}} e^{-\frac{2}{3}z^{\frac{3}{2}}} (1 + ...)$$

to

$$Ai(z) = \frac{1}{2\sqrt{\pi}} z^{-\frac{1}{4}} e^{-\frac{2}{3}z^{\frac{3}{2}}} (1 + ...) + i\frac{1}{2\sqrt{\pi}} z^{-\frac{1}{4}} e^{+\frac{2}{3}z^{\frac{3}{2}}} (1 + ...)$$

This phenomenon, which explains the different asymptotic behaviour for $Ai(x)$, is called **Stokes phenomenon**.

Following the glimpse given from the Airy equation, we can describe the Stokes phenomenon in dimension 1 as follows.

We can cover a pointed neighborhood of the singularity with a family of sectors for which Borel-Ritt lemma holds.

Suppose S, S' in this family are such that $S \cap S' \neq \emptyset$ and , $F = \{f_i\}, F' = \{f_i'\}$ are asymptotic to the formal solutions $\hat{F}$ as in Hukuhara-Levelt-Turrittin theorem.

We can then compare F and F' in the common sector

$$f_i' = \sum_{j=1}^{n} a_{ij} f_j\ .$$

The a_{ij} are called Stokes multipliers, (a_{ij}) the Stokes matrix: this set of data completely encodes the Stokes phenomenon.

Deligne and Malgrange proved, moreover, that it is possible to recover the original differential equation from this information, hence establishing an extended version of the original Riemann-Hilbert correspondence called **Riemann-Hilbert-Birkhoff correspondence**.

Several efforts have been made in order to provide a description fitting for higher dimensions.

As explained above for the one dimensional case, one of the keypoints is to keep track of the growth of the solutions near the singularity.

In particular, this is related to the study of the behaviour of $Re(\phi_i)$ and their mutual interactions.

The idea developed by D'Agnolo and Kashiwara is to consider an extra real variable to take care of this issue.

They then introduced the category $E^b_{\mathbb{R}-c}(I\mathbb{C}_X)$ of **enhanced ind-sheaves**, algebraic objects defined on $X \times \mathbb{R}$ (no more on X as the classical local systems) which locally encode information about the growth of the exponential factors by considering $\{Re(\phi_i) \geq t\}$.

The theory they developed has brought to the following result, extending the one for regular holonomic $\mathcal{D}$-modules.

Proposition (D'Agnolo-Kashiwara [1]) *Let X be an analytic manifold. Then there exists a fully faithful functor*

$$\mathcal{S}ol^E_X : D^b_h(\mathcal{D}_X) \to E^b_{\mathbb{R}-c}(I\mathbb{C}_X).$$

Moreover, there is a way of reconstructing: there exist a functor $\mathcal{H}om(\cdot, \mathcal{O}^E_X)$ and an isomorphism

$$\mathcal{M} \to \mathcal{H}om(\mathcal{S}ol^E(\mathcal{M}), \mathcal{O}^E_X)$$

functorial in $\mathcal{M} \in D^b_h(\mathcal{D}_X)$.

References

[1] Andrea D'Agnolo, Masaki Kashiwara, *Riemann-Hilbert correspondence for holonomic $\mathcal{D}$-modules.* Publ. Math. Inst. Hautes Etudes Sci. 123/1 (2016), 69–197; `arXiv:1311.2374`.

[2] Ryoshi Hotta, Kiyoshi Takeuchi, Toshiyuki Tanisaki, "$\mathcal{D}$-Modules, Perverse Sheaves, and Representation Theory". Progress in Mathematics, Vol. 236, Birkhauser Basel, 2008.

[3] Masaki Kashiwara, *The Riemann-Hilbert Problem for Holonomic Systems.* Publ. RIMS, Kyoto Univ. 20 (1984), 319–365.

[4] Masaki Kashiwara, "$\mathcal{D}$-modules and microlocal calculus". Translation of Mathematical Monographs, Book 217, American Mathematical Society, 2002.

[5] A. H. M. Levelt, *Jordan decomposition for a class of singular differential operators.* Ark. Mat. 13 (1975), 1–27.

[6] Bernard Malgrange, "Equations differentielles àcoefficients polynomiaux". Progress in Mathematics, Vol. 96, Birkhauser Boston , Cambridge, MA, 1991.

[7] George Gabriel Stokes, *On the discontinuity of arbitrary constants which appear in divergent developments.* Trans. Cambridge Philos. Soc. 10 (1857), 105–128.

Including topographic effects in shallow water modeling

Elena Bachini (*)

Abstract. Shallow water models of geophysical flows must be adapted to geometrical characteristics in the presence of a general bottom topography with non-negligible slopes and curvatures, such as a mountain landscape. We study a shallow water model defined intrinsically on the bottom surface from a mathematical and numerical point of view.

1 Introduction

Shallow water equations are typically used to model fluid flows that develop predominantly along the horizontal (longitudinal and lateral) direction. Indeed, the so-called Shallow Water (SW) hypothesis assumes negligible vertical velocity components.

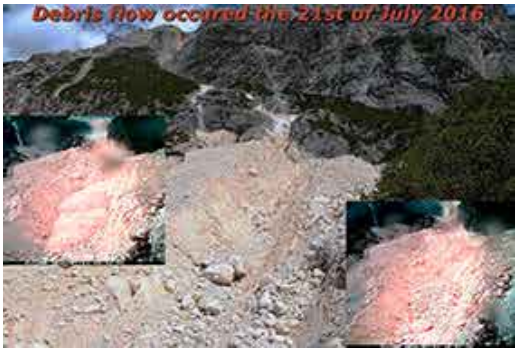


Figure 1. An example of landslide (2016).

Application of SW equations ranges from large-scale models in meteorology, oceanography [9], or tsunami modeling [12], but also to smaller scale models of river morphology [11], avalanches [8], debris flows and landslides [7, 10] and, at even smaller scales, lubrication theory. Most of these applications must consider a general bottom topography, commanded for example in the field of our interest by mountain landscapes (see fig. 1), introducing mathematical difficulties that have not yet a comprehensive solution. This is in contrast with the increasing necessity of reliable models, both in environmental and industrial applications.

Our particular interest is to take into account as much as possible the influence of the topography in the flow equations. The dynamics are now well understood in the flat case, but the situation is different for the non-flat case, especially working in several dimensions. In this report, we will present a new formulation of the two-dimensional SW equations in intrinsic coordinates adapted to general and complex terrains, with emphasis on the influence of the geometry of the bottom on the solution. The proposed model is then

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: elena.bachini@phd.unipd.it . Seminar held on May 22nd, 2019.

discretized with a first order upwind Godunov Finite Volume scheme. We will give an overview of the numerical method and then show some results. The results indicate that it is important to take into full consideration the bottom geometry and slope even for relatively mild and slowly varying curvatures.

2 Shallow Water Model

The typical derivation of the SW equations is based on the integration of the time-averaged Navier-Stokes equations over the fluid depth in combination with an asymptotic analysis enforcing the SW assumptions. Consider the classical incompressible Navier-Stokes equations on an open domain $\Omega \subset \mathbb{R}^3$ as:

$$\nabla \cdot \vec{u} = 0 \,, \tag{1a}$$

$$\frac{\partial \vec{u}}{\partial t} + \nabla \cdot (\vec{u} \otimes \vec{u}) = -\frac{1}{\rho}\nabla p + \frac{1}{\rho}\nabla \cdot \mathbb{T} + \vec{g} \,, \tag{1b}$$

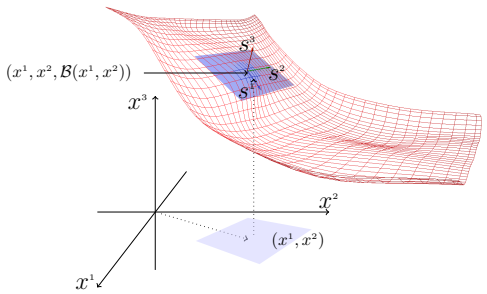
where $\vec{u} : \Omega \times [0,T] \to \mathbb{R}^3$ is the fluid velocity, ρ its density, assumed constant, $p : \Omega \times [0,T] \to \mathbb{R}$ is the fluid pressure, $\mathbb{T} : \Omega \to \mathbb{R}^{3\times 3}$ the stress tensor, and $\vec{g}$ the gravity acceleration. Note that we have used the product rule of differentiation and the incompressibility condition (1a) to write the convective term in conservative form. We assume that the domain boundary $\partial\Omega$ is smooth and formed by the union of the bottom surface ($\mathcal{S}_\mathcal{B}$), the free surface ($\mathcal{S}_\mathcal{F}$), and the lateral surface. Smoothness is detailed by the hypothesis that all these surfaces are regular and can be identified by the graph of some function. For almost flat bottom topographies, fluid depth is evaluated along the direction normal to the bottom surface. This approach is generally used to model large scale ocean dynamics or atmospheric flows, where the bottom boundary is the geo-sphere and the normal direction coincides with the direction of gravitational forces. In these cases, the SW approximation, essentially stating that the fluid vertical velocity is small compared to the horizontal components.

In the case of a generally curved bottom, our case of interest, the essence of the SW approximation requires that the integration path along which depth averaging is performed be at any point orthogonal to the fluid velocity. This path has been identified with the so called "cross-flow" curves [5, 6]. Integrating along small segments of this path would allow to define a specific discharge which is a constant of the motion in all the parallel cross-flow paths. Unfortunately, the knowledge of these cross-flow paths would require the knowledge of the velocity of the fluid and to know a priori which point of the free surface the path would intersect. Thus this would be an implicit definition of the problem, since the velocity is one of the unknowns, and would not give rise to a solvable system of equations. For this reason, Fent et al. [6] approximate the "cross-flow" path with the direction normal to the bottom starting from a NS system defined on a curvilinear coordinate reference frame defined on the bottom geometry. The system of SW Equations (SWE) resulting from depth integration turns out to be closely related to the model of Bouchut and Westdickenberg [4], that also presents a model of SWE on arbitrary topography, and shares similar approximations and limitations in terms of geometry of the bed topography.

Following Fent et al. [6] we define a local curvilinear reference system (LCS) positioned on the surface representing the topography of the bottom. All the developments, including depth integration, will be carried out with respect to this local reference system. We would like to describe the motion of a fluid particle using a coordinate system that satisfies the following two main conditions:

(a) the first two coordinates run along the bottom surface $\mathcal{S}_\mathcal{B}$, their tangent vectors belonging at each point $\mathbf{P} \in \mathcal{S}_\mathcal{B}$ to the tangent plane $T_\mathbf{p}\mathcal{S}_\mathcal{B}$;

(b) the third coordinate crosses the surface orthogonally so a vector tangent to $\mathcal{S}_\mathcal{B}$ is everywhere orthogonal to $\vec{\mathbf{N}}$, the surface normal vector.



Regarding the ensuing reference frame, the previous requests amount to asking that there exist three vector fields $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3$ in $\mathbb{R}^3$ such that

$$\mathbf{t}_1(\mathbf{P}), \mathbf{t}_2(\mathbf{P}) \in T_\mathbf{p}\mathcal{S}_\mathcal{B} \quad \forall \mathbf{P} \in \mathcal{S}_\mathcal{B} \ ,$$

are vector fields in the tangent plane of $\mathcal{S}_\mathcal{B}$ at point $\mathbf{P}$, and $\mathbf{t}_3(\mathbf{P})$ is orthogonal to the other two frame vectors and such that the right-hand rule is satisfied. Moreover, we ask that $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3$ commute in all $\mathbb{R}^3$ and, to ensure numerical stability, be pairwise orthogonal. Note that normalization of the other two basis vectors cannot be done, as this would amount to assume a zero curvature of $\mathcal{S}_\mathcal{B}$ at $\mathbf{P}$, loosing all the geometric information we would like to preserve in our LCS. The associated metric tensor, as a consequence of the orthogonality property, becomes the diagonal matrix given by:

$$\mathcal{G} := \begin{pmatrix} \|\mathbf{t}_1(\mathbf{P})\|^2 & 0 & 0 \\ 0 & \|\mathbf{t}_2(\mathbf{P})\|^2 & 0 \\ 0 & 0 & \|\mathbf{t}_3(\mathbf{P})\|^2 \end{pmatrix} = \begin{pmatrix} h_{(1)}^2 & 0 & 0 \\ 0 & h_{(2)}^2 & 0 \\ 0 & 0 & 1 \end{pmatrix} , \tag{2}$$

and the metric coefficients enter in the definition of the expressions of the differential operators that appear in the Navier-Stokes equations, i.e., we will have a new definition for the gradient of a scalar function, the divergence of a vector field, and the divergence of a tensor field.

Remark 1 We will use the standard notation with the subscript $\mathcal{G}$ to denote the intrinsic differential operators. Namely, $\nabla_\mathcal{G}$ and $\nabla_\mathcal{G}\cdot$ will be the gradient and divergence symbol, respectively.

The derivation of the SWE starts from the formulation of Navier-Stokes equations in the local coordinate system. Using new the definition of the differential operators, the

Navier-Stokes equations given in eq. (1) can be written in the LCS as:

$$\nabla_{\mathcal{G}} \cdot \vec{u} = 0 \,, \tag{3a}$$

$$\frac{\partial \vec{u}}{\partial t} + \nabla_{\mathcal{G}} \cdot (\vec{u} \otimes \vec{u}) = -\frac{1}{\rho}\nabla_{\mathcal{G}} p + \frac{1}{\rho}\nabla_{\mathcal{G}} \cdot \mathbb{T} + \vec{g} \,. \tag{3b}$$

We need to complete the system considering appropriate boundary conditions. Using the LCS, the bottom and free surfaces are given by:

$$\mathcal{S}_{\mathcal{B}} := \left\{ (s^1, s^2, s^3) \in \mathbb{R}^3 \text{ such that } s^3 = \mathcal{B}(s^1, s^2) \equiv 0 \right\} \,,$$
$$\mathcal{S}_{\mathcal{F}} := \left\{ (s^1, s^2, s^3, t) \in \mathbb{R}^3 \times [0, T] \text{ such that } s^3 = \mathcal{F}(s^1, s^2, t) \equiv \eta(s^1, s^2, t) \right\} \,,$$

where $\eta(s^1, s^2, t) = \mathcal{F}(s^1, s^2, t) - \mathcal{B}(s^1, s^2)$ denotes the fluid depth. We assume that the bottom is not eroding and thus maintains a fixed geometry, while the fluid surface is a function of time. The kinematic conditions postulate that the free surface moves with the fluid and that the bottom is impermeable. Moreover, assuming that the external actions on the fluid surface are negligible, the dynamic condition at the fluid-air interface translates into a zero-stress boundary equation (we do not present here the expressions for the boundary conditions, see [2] for the complete formulation).

Starting from the Navier-Stokes equations written in the local curvilinear coordinate system as given in eq. (3), we perform depth integration along the normal direction s^3 from $s^3 = \mathcal{B}(s^1, s^2) \equiv 0$ to $s^3 = \mathcal{F}(s^1, s^2, t) \equiv \eta(s^1, s^2, t)$. As example of computation, we report here the integration of the continuity equation (3a). Applying Leibniz rule and the kinematic boundary conditions, we obtain:

$$\begin{aligned}
\int_0^\eta \nabla_{\mathcal{G}} \cdot \vec{u} =& \int_0^\eta \frac{1}{h_{(1)} h_{(2)}} \left(\frac{\partial \left(h_{(1)} h_{(2)} u^1 \right)}{\partial s^1} + \frac{\partial \left(h_{(1)} h_{(2)} u^2 \right)}{\partial s^2} + \frac{\partial \left(h_{(1)} h_{(2)} u^3 \right)}{\partial s^3} \right) = \\
=& \frac{1}{h_{(1)} h_{(2)}} \frac{\partial}{\partial s^1} \int_0^\eta h_{(1)} h_{(2)} u^1 + \frac{1}{h_{(1)} h_{(2)}} \frac{\partial}{\partial s^2} \int_0^\eta h_{(1)} h_{(2)} u^2 \\
& + u^3 \Big|_{s^3=\eta} - \frac{u^1}{h_{(1)}} \frac{\partial \mathcal{F}}{\partial s^1} \Big|_{s^3=\eta} - \frac{u^2}{h_{(2)}} \frac{\partial \mathcal{F}}{\partial s^2} \Big|_{s^3=\eta} \\
& - u^3 \Big|_{s^3=0} + \frac{u^1}{h_{(1)}} \frac{\partial \mathcal{B}}{\partial s^1} \Big|_{s^3=0} + \frac{u^2}{h_{(2)}} \frac{\partial \mathcal{B}}{\partial s^2} \Big|_{s^3=0} = \\
=& \frac{\partial \eta}{\partial t} + \nabla_{\mathcal{G}} \cdot \int_0^\eta \underline{\vec{u}} \,,
\end{aligned}$$

where $\underline{\vec{u}} := [u^1, u^2]^{\mathrm{T}}$ reduces to the two component tangential velocity, and the curvilinear divergence operator $\nabla_{\mathcal{G}} \cdot$ is adapted to the two-dimensional setting. Recall that application of Leibniz rule requires enough regularity of both bottom and free surfaces as well as the velocity vector $\vec{u}$. Analogous computations are done for the momentum equations.

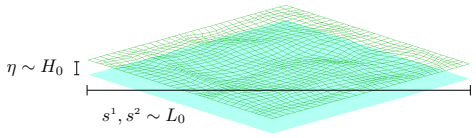


Figure 2. Example of a thin and wide layer of fluid over an horizontal bottom surface.

The classical SW hypothesis states that the characteristic depth of the fluid is smaller than the characteristic wavelength. In the present context, this statement is equivalent to our assumption of small normal velocity. To see this, assume a setting with a relatively thin and wide fluid moving on the terrain surface (fig. 2). Denote with L_0 the length scale in a direction tangential to the bottom and with H_0 the length scale of the fluid depth measured along the normal. The shallow water scaling assumes that $H_0/L_0 = \epsilon \ll 1$.

We would like to connect this idea with the order of approximation of the model in our curvilinear setting. Denote by V_0 the scale of the contravariant tangential velocity components u^1 and u^2, and by W_0 the scale of the contravariant normal component u^3. Formal application of the chain rule of differentiation to the continuity equation (3a) yields:

$$0 = \nabla_{\mathcal{G}} \cdot \vec{u} = \frac{1}{h_{(1)}h_{(2)}} \left(\frac{\partial}{\partial s^1}(h_{(1)}h_{(2)}u^1) + \frac{\partial}{\partial s^2}(h_{(1)}h_{(2)}u^2) + \frac{\partial}{\partial s^3}(h_{(1)}h_{(2)}u^3) \right) =$$
$$= \underbrace{\frac{\partial u^1}{\partial s^1}}_{\mathcal{O}\left(\frac{V_0}{L_0}\right)} + \underbrace{\frac{\partial u^2}{\partial s^2}}_{\mathcal{O}\left(\frac{V_0}{L_0}\right)} + \underbrace{\frac{\partial u^3}{\partial s^3}}_{\mathcal{O}\left(\frac{W_0}{H_0}\right)} + \underbrace{\frac{u^1}{h_{(1)}}}_{\mathcal{O}(V_0)} \frac{\partial h_{(1)}}{\partial s^1} + \underbrace{\frac{u^1}{h_{(2)}}}_{\mathcal{O}(V_0)} \frac{\partial h_{(2)}}{\partial s^1} + \underbrace{\frac{u^2}{h_{(1)}}}_{\mathcal{O}(V_0)} \frac{\partial h_{(1)}}{\partial s^2} + \underbrace{\frac{u^2}{h_{(2)}}}_{\mathcal{O}(V_0)} \frac{\partial h_{(2)}}{\partial s^2} .$$

With standard properties of the BigO notation, we obtain a relation between the normal velocity scale and the tangential one:

$$W_0 \sim \underbrace{\max\left\{\epsilon, H_0\frac{\partial h_{(1)}}{\partial s^1}, H_0\frac{\partial h_{(2)}}{\partial s^1}, H_0\frac{\partial h_{(1)}}{\partial s^2}, H_0\frac{\partial h_{(2)}}{\partial s^2}\right\}}_{\epsilon_{\mathcal{G}}} \times V_0 \,, \tag{4}$$

where $\epsilon_{\mathcal{G}}$ is the new "geometric" aspect ratio, that connects local curvatures information to the global length scale parameter ϵ. Hence, the SW approximation can be restated by the assumption $\epsilon_{\mathcal{G}} \ll 1$, which effectively adds a restriction on the shape of the bottom surface that ensures that the derivatives of the metric coefficients are of the order of $1/L_0$.

Using the formal expansions in powers of $\epsilon_{\mathcal{G}}$ in the normally integrated NS system we obtain our reduced formulation, which we name Bottom-Adapted Shallow Water (BASW) equations, as given in the next theorem. We use the following notation. The couple (s^1, s^2) indicates the curvilinear coordinate system associated with the LCS with the ensuing metric tensor $\mathcal{G}_{sw}$ given by the principal 2-minor of eq. (2). The vector $\vec{q} = \left[\eta U^1, \eta U^2\right]^{\mathrm{T}}$ denotes the depth-averaged velocity vector, while the tensor

$$\mathbf{T}_{sw} = \eta \begin{bmatrix} \mathbf{T}^{11} & \mathbf{T}^{12} \\ \mathbf{T}^{21} & \mathbf{T}^{22} \end{bmatrix}$$

is the principal 2-minors of $\mathbf{T}$. Vector $\mathbf{f}_{\mathcal{B}} = [\tau_b^1, \tau_b^2]^T$ is the vector field accounting for bed friction. Then, we can state the following theorem.

Theorem 2.1 *The bottom-adapted shallow water equations, written with respect to the LCS, are given by:*

$$\frac{\partial \eta}{\partial t} + \nabla_{\mathcal{G}} \cdot \vec{q} = 0 \,, \tag{5a}$$

$$\frac{\partial \vec{q}}{\partial t} + \nabla_{\mathcal{G}} \cdot \left(\frac{1}{\eta}(\vec{q} \otimes \vec{q}) + \left(\frac{g\eta^2}{2} \frac{\partial x^3}{\partial s^3} \right) \mathcal{G}_{sw}^{-1} \right) + \frac{g\eta^2}{2} \nabla_{\mathcal{G}} \left(\frac{\partial x^3}{\partial s^3} \right) + g\eta \nabla_{\mathcal{G}}(x^3) - \frac{1}{\rho} \left(\nabla_{\mathcal{G}} \cdot \mathbf{T}_{sw} + \mathbf{f}_{\mathcal{B}} \right) = 0 \,. \tag{5b}$$

They provide an approximation of order $\mathcal{O}(\epsilon_{\mathcal{G}}^2)$ of the Navier-Stokes equations, under the assumption of thin fluid layer, $\eta = \mathcal{O}(\epsilon_{\mathcal{G}})$.

As already mentioned in the introduction, the BASW model is similar to the model proposed by Bouchut and Westdickenberg [4] and the two model share fundamental mathematical properties:

Proposition 2.2 *The BASW system defined in eq. (5) is invariant under rotation, it admits a conserved energy in the absence of stresses, and is well-balanced.*

We refer to Bachini et al. [2] for the complete development of the computations and the proofs.

3 Numerical Solution

The intrinsic nature of the developed SWE allows the formulation of a Finite Volume (FV) discretization, adapted to our intrinsic setting. We assume that our final system (5) is defined on a compact subset of the bottom surface, $\Gamma \subset \mathcal{S}_{\mathcal{B}}$, and that a well-defined curvilinear boundary, denoted by $\partial\Gamma = \partial\overline{\Gamma}$, exists. System (5) can be written in divergence form as the balance law:

$$\frac{\partial \mathbf{U}}{\partial t} + \mathrm{div}_{\mathcal{G}} \underline{\underline{F}}(\mathbf{s}, \mathbf{U}) + \mathbf{S}(\mathbf{s}, \mathbf{U}) = 0 \,. \tag{6}$$

Here the conservative variable is given by $\mathbf{U} = \left[\eta, \eta U^1, \eta U^2\right]^{\mathrm{T}} = \left[\eta, q^1, q^2\right]^{\mathrm{T}}$, where $\eta : \Gamma \times [0,T] \to \mathbb{R}$, and $\mathbf{q} = \left[q^1, q^2\right]$, $\mathbf{q} : \Gamma \times [0,T] \to \mathbb{R}^2$. The flux function $\underline{\underline{F}}$ takes the form

$$\underline{\underline{F}}(\mathbf{s}, \mathbf{U}) = \begin{bmatrix} q^1 & q^2 \\ \dfrac{(q^1)^2}{\eta} + \dfrac{g\eta^2}{2h_{(1)}^2} \dfrac{\partial x^3}{\partial s^3} & \dfrac{q^1 q^2}{\eta} \\ \dfrac{q^1 q^2}{\eta} & \dfrac{(q^2)^2}{\eta} + \dfrac{g\eta^2}{2h_{(2)}^2} \dfrac{\partial x^3}{\partial s^3} \end{bmatrix} = \begin{bmatrix} \underline{F}^{\eta} \\ \\ \underline{\underline{F}}^{\mathbf{q}} \end{bmatrix} \,. \tag{7}$$

Note that the flux $\underline{\underline{F}}$ is a function of $\mathbf{s}$ because of the appearance of the components $h_{(\mathrm{i})}$ of the metric tensor $\mathcal{G}_{sw}$ and the presence of the bottom slope $\partial x^3/\partial s^3$. The symbol $\mathrm{div}_{\mathcal{G}}$ denotes the divergence operator applied to the flux function as divergence of a vector for the first row and divergence of a 2×2 tensor for the last two rows. We can define it as $\mathrm{div}_{\mathcal{G}} =$

$[\nabla_{\mathcal{G}}{}^{\eta}\cdot, \nabla_{\mathcal{G}}{}^{\mathbf{q}}\cdot]^{\mathrm{T}}$. The source function $\mathbf{S}$ comprises the metric tensor coefficients, the bottom slope and its derivatives, the two-dimensional averaged stress tensor $\mathbf{T}_{sw}$, the bottom friction parameter τ_b, and the conserved variable η. We summarize this dependency by explicitly writing it out in $\mathbf{S}(\mathbf{s}, \eta)$. We have then:

$$\mathbf{S}(\mathbf{s},\eta) = \begin{bmatrix} 0 \\ \dfrac{g\eta^2}{2h_{(1)}^2}\dfrac{\partial}{\partial s^1}\left(\dfrac{\partial x^3}{\partial s^3}\right) + \dfrac{g\eta}{h_{(1)}^2}\dfrac{\partial x^3}{\partial s^1} - \dfrac{1}{\rho}\left[\nabla_{\mathcal{G}}\cdot\mathbf{T}_{sw}\right]^{(1,\cdot)} - \dfrac{\tau_b^1}{\rho} \\ \dfrac{g\eta^2}{2h_{(2)}^2}\dfrac{\partial}{\partial s^2}\left(\dfrac{\partial x^3}{\partial s^3}\right) + \dfrac{g\eta}{h_{(2)}^2}\dfrac{\partial x^3}{\partial s^2} - \dfrac{1}{\rho}\left[\nabla_{\mathcal{G}}\cdot\mathbf{T}_{sw}\right]^{(2,\cdot)} - \dfrac{\tau_b^2}{\rho} \end{bmatrix} = \begin{bmatrix} S^{\eta} \\ \\ \mathbf{S}^{\mathbf{q}} \end{bmatrix} . \tag{8}$$

The regularity assumption on the bottom surface implies the uniform continuity of the flux and source functions with respect to $\mathbf{s}$.

Finite Volume scheme

The derivation of the scheme starts from the definition of the computational mesh. We assume that there exists a surface triangulation $\mathcal{T}(\Gamma)$ formed by the union of non-intersecting geodesic triangles with vertices on Γ (edges are geodesics). Obviously, we have that $\mathcal{T}(\Gamma) = \cup_{i=1}^{N_T} T_i = \overline{\Gamma}$ and $\sigma_{ij} = T_i \cap T_j$ is an internal geodesic edge. We will also use the approximate triangulation $\mathcal{T}_h(\Gamma)$ defined by the piecewise linear surface identified by the union of 2-simplices in $\mathbb{R}^3$ (flat three-dimensional triangles) with vertices coinciding with the vertices of $\mathcal{T}(\Gamma)$. We assume that this triangulation is closely inscribed in $\mathcal{T}(\Gamma)$ in the sense of Morvan [13] (the tangent spaces of $\mathcal{T}(\Gamma)$ and of $\mathcal{T}_h(\Gamma)$ are close in some sense). Quantities belonging to the approximated triangulation $\mathcal{T}_h$ will be identified with the subscript h. Thus the symbol $\sigma_{h,ij}$ will identify the common edge between triangles $T_{h,i}$ and $T_{h,j}$.

We start our work on $\mathcal{T}(\Gamma)$, where the divergence and integration by parts theorems are naturally defined. Following a standard development workflow for FV methods, we test eq. (6) with a piecewise constant (in space and time) function and apply the divergence theorem to obtain the following set of equations valid for all triangles $T_i \in \mathcal{T}(\Gamma)$ and for $t \in [t^k, t^{k+1}]$:

$$\mathbf{U}_i^{k+1} = \mathbf{U}_i^k - \frac{1}{|T_i|}\sum_{j=1}^{N_{\sigma(i)}} |\sigma_{ij}| \int_{t^k}^{t^{k+1}} \mathbf{F}_{ij}(\mathbf{U})\, dt - \int_{t^k}^{t^{k+1}} \mathbf{S}_i(\eta)\, dt \ ,$$

where we use the cell-averaged and edge-averaged quantities defined intrinsically in $\mathcal{T}(\Gamma)$ as:

$$\mathbf{U}_i = \frac{1}{|T_i|}\int_{T_i} \mathbf{U}\, d\mathbf{s} \ , \qquad \mathbf{F}_{ij} = \frac{1}{|\sigma_{ij}|}\int_{\sigma_{ij}} \left\langle \underline{\underline{F}}, \nu_{ij} \right\rangle_{\mathcal{G}} d\sigma \ , \qquad \mathbf{S}_i = \frac{1}{|T_i|}\int_{T_i} \mathbf{S}\, d\mathbf{s} \ . \tag{9}$$

We denote by $|T_i|, |\sigma_{ij}|$ the area of the cell T_i and the length of the curvilinear edge σ_{ij}, respectively, and ν_{ij} is the outer normal to the edge σ_{ij}. Note that the quantities $\mathbf{F}_{ij}, \mathbf{S}_i$ are depending only on the unknown $\mathbf{U}$ but not on the space variable $\mathbf{s}$, since

they are integrated in space. Moreover, it is important to underline that no numerical approximations are done up to this point. Now we need to devise numerically computable approximations of the above quantities. Thus, the following steps need to be appropriately defined: (i) time stepping; (ii) normal fluxes on edges; (iii) quadrature rules; (iv) Riemann problem.

For the time integration we use a first order explicit Euler time stepping scheme, while in space we consider a mid-point rule both on edges and cells. To maintain a well-balanced scheme we use an adaptation of the approach of Audusse et al. [1], Bouchut [3] and include the source terms in the flux. Then, the following FV equations are defined for each T_i:

$$\tilde{\mathbf{U}}_{h,i}^{k+1} = \mathbf{U}_{h,i}^{k} - \frac{\Delta t}{|T_{h,i}|} \sum_{j=1}^{N_{\sigma(i)}} |\sigma_{h,ij}| \left[\mathbf{F}_{h,ij}(\mathbf{U}_{h,i}^{k}, \mathbf{U}_{h,j}^{k}) + \mathbf{S}_{h,ij}(\mathbf{U}_{h,i}^{k}, \mathbf{U}_{h,j}^{k}) \right] , \tag{10}$$

where $\mathbf{F}_{h,ij}$ is the numerical approximation of the edge-averaged normal flux $\mathbf{F}_{ij}$ at σ_{ij}, and $\mathbf{S}_{h,ij}$ is calculated so that $\sum_{j=1}^{N_{\sigma(i)}} |\sigma_{h,ij}| \, \mathbf{S}_{h,ij}$ is a consistent quadrature rule for the last integral in eq. (9) and maintains the discrete version of the well-balance property. As standard in FV scheme, the flux on the edges is computed by solving appropriate Riemann problems, that in our case need to be adapted to the geometric framework.

To be able to perform all these computations we require the approximation of the relevant surface quantities. By assumption we have all the geometric information at the nodes as given data of the problem, and we decide to perform linear interpolation of these values to compute geometric quantities at the quadrature points. Again we refer to [2] for the full description of our approximations.

Simulations

A number of test cases performed over slowly varying bottom topography are used to show the effectiveness of the numerical approach and to verify the importance of considering the geometric features of the bed topography in the equations. We use a global parametrization $x^3 = \mathcal{B}(x^1, x^2)$ of the bottom surface $\mathcal{S}_{\mathcal{B}}$, with $\mathcal{B}$ a sufficiently smooth height function, whereby we start from a regular triangulation of a rectangular subset $\mathcal{A} \subset \mathbb{R}^2$ and move the nodes vertically on $\mathcal{S}_{\mathcal{B}}$. All the test cases simulate a gravity-driven fluid in a dam-break setting, without any stress tensor. The initial conditions are defined to initiate a dam-break phenomenon, with water depth in any case small enough to exclude the issue of the intersection of the local normals so that the coordinate transformation is always a diffeomorphism. Here we present the simulations of two test cases: a simple one-dimensional domain with simple one-dimensional curvature, and a centrally symmetric surface. Figure 3 shows the distribution in space of the metric coefficients.

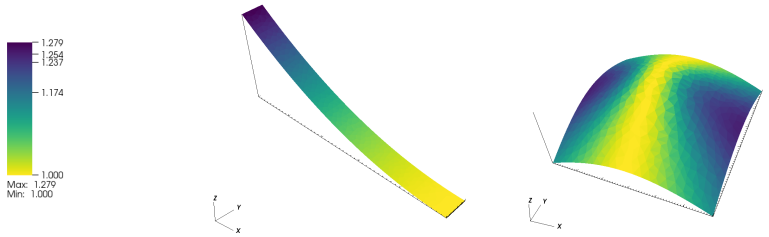


Figure 3. Spatial distribution of the metric coefficients $h_{(1)}$ for the parabola case (left) and the hyperboloid-central-bump (right).

The parabola test case considers a simple one-dimensional flow where the effects of curvature in the model can be verified. We simulate the breaking of a dam located at x^1 =2.0 m, with initially zero velocity everywhere and water depth of 0.5 m upstream and 0.2 m downstream the dam. No-flow boundary conditions are imposed everywhere except at the outlet boundary, where a free outflow is enforced. Figure 4 shows the calculated distribution of the water depth η at times t = 0.00 s, 0.50 s, 1.00 s and 1.50 s. The progress of the dam-break wave towards the outlet is characterized by a variable speed of propagation. The downwind shock initially smoothed by the numerical viscosity introduced by the 1st order solver is sharpened downstream by curvature effects, as the decreasing slope is decelerating the wave front. Also the upstream wave seems to sharpen, as evidenced by a shorter wave length at the end of the simulation.

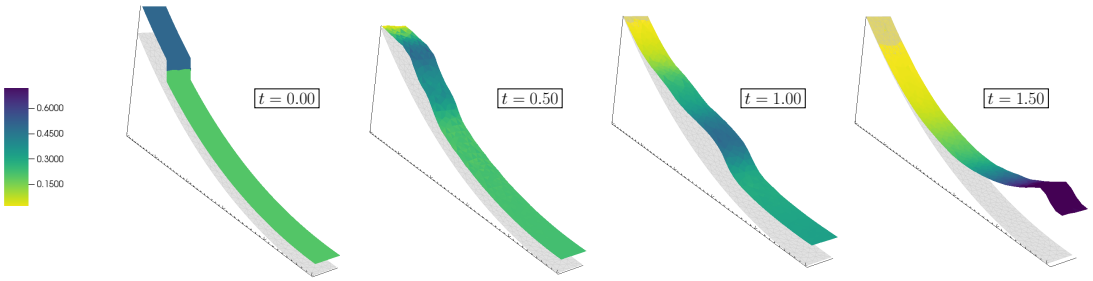


Figure 4. Parabola case: evolution of the gravity wave, shown both as color codes and depth elevation, the latter with a vertical magnification factor of 2.0.

The second test case presents an "almost" centrally symmetric domain, hyperboloid with a central bump (HCB for simplicity), and it is designed to verify the ability of the FV scheme to maintain symmetry on an unstructured grid. The initial conditions outline a central area of radius 0.5 m with upstream water depth of 2.0 m and downstream water depth of 1.0 m, and zero initial velocity.

Figure 5 shows the numerically evaluated evolution of the initial wave in terms of water depth η at times t = 0.0 s, 0.20 s, 0.40 s and 0.60 s. The initial wave moves downward

with radial velocity vectors towards the outlet. The dynamics of the flow is such that the downstream portion of the initial dam-break wave accelerates faster than the upstream region because of the larger bottom slope. Some oscillations are created by the Riemann solver at the tail of the downstream wave, but these remain bounded and do not interfere with the trailing wave. Nonetheless, the numerical results shows a rather symmetric wave pattern, demonstrating the robustness of the chosen numerical approach. This is further evidenced in fig. 6 (left panel), where the velocity vectors at $t = 0.20$ s are shown. The streamflows at the three different sections, located at a radial distance from the center of 1.0 m, 1.75 m and 2.5 m are shown in fig. 6 (right panel).

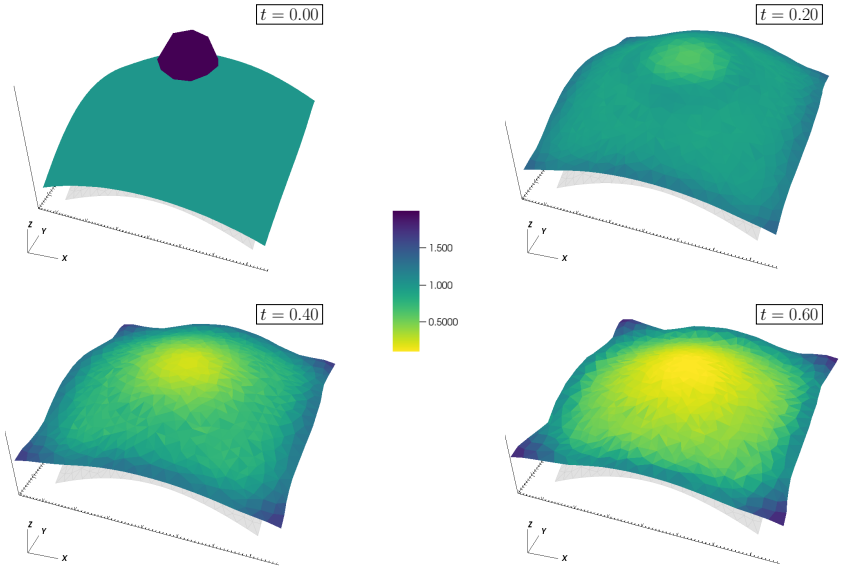


Figure 5. HCB: water depth at initial time ($t = 0.0$), and at $t = 0.20,\ 0.40,\ 0.60$.

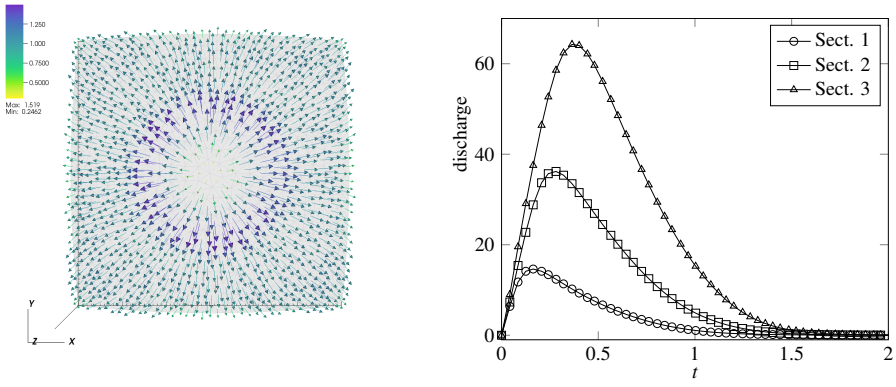


Figure 6. HCB: velocity vectors at $t = 0.20$ (left) and streamflows at the three preselected sections (right).

References

[1] E. Audusse, F. Bouchut, M. Bristeau, R. Klein, and B. Perthame, *A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows.* SIAM J. Sci. Comput. 25/6 (2004), 2050–2065. doi: 10.1137/S1064827503431090.

[2] E. Bachini, I. Fent, and M. Putti, *Geometrically intrinsic modeling of shallow water flows.* ESAIM- Math. Model. Num., submitted.

[3] F. Bouchut., "Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws and Well-Balanced Schemes for Sources". Volume 2/2004. Birkhauser Verlag, Basel, Switzerland, 01 2004. ISBN 3-7643-6665-6.

[4] F. Bouchut and M. Westdickenberg, *Gravity driven shallow water models for arbitrary topography.* Comm. Math. Sci. 2/3 (Sept. 2004), 359–389.

[5] M. Boutounet, L. Chupin, P. Noble, and J.P. Vila, *Shallow water viscous flows for arbitrary topography.* Commun. Math. Sci. 6/1 (2008), 29–55.

[6] I. Fent, M. Putti, C. Gregoretti, and S. Lanzoni, *Modeling shallow water flows on general terrains.* Adv. Water Resour. 121 (2018), 316–332.

[7] L. Francarollo and H. Capart, *Riemann wave description of erosional dam-break flows.* J. Fluid Mech. 461 (2002), 1–46.

[8] J.M.N.T. Gray, M. Wieland, and K. Hutter, *Gravity-driven free surface flow of granular avalanches over complex basal topography.* Phil. Trans. R. Soc. A, 455 (1985), 1841–1874, May 1999.

[9] R.L. Higdon, *Numerical modelling of ocean circulation.* Acta Num., 15:385, May 2006.

[10] R.M. Iverson and D.L. George, *A depth-averaged debris-flow model that includes the effects of evolving dilatancy. I.*. Physical basis. Proc. R. Soc. London 470(2170): 20130819–20130819, July 2014.

[11] S. Lanzoni, A. Siviglia, A. Frascati, and G. Seminara, *Long waves in erodible channels and morphodynamic influence.* Water Resour. Res., 42:W06D17, 2006.

[12] R. LeVeque, D. George, and M. Berger, *Tsunami modelling with adaptively refined finite volume methods.* Acta Num. 20 (2011), 211–289.

[13] J.-M. Morvan, "Generalized Curvatures". Volume 2 of Geometry and Computing. Springer Science & Business Media, Berlin, Heidelberg, May 2008.

Serre's p-adic modular forms and p-adic interpolation of the Riemann zeta function

Giacomo Graziani (*)

Abstract. The so-called zeta functions are among the most famous and discussed objects in mathematics, the simplest of which is the (in)famous Riemann zeta function. In order to work with them (and with the strictly related L-functions as well), mathematicians decided to isolate simpler pieces and hence ultimately to address the problem of their p-adic interpolation. In this seminar, after introducing the various objects involved, we will focus on easiest example of the Riemann zeta function and describe the surprising interpolation exploited by Serre using his notion of p-adic modular forms.

1 Introduction

In 1637 Fermat claimed he could prove that all the integral solutions of the equation

$$a^n + b^n = c^n, \quad n \geq 3$$

have $abc = 0$. As it is well known he didn't provide any proof of this fact, so the quest for proving (or disproving) this statement begun. A remarkable attempt was that of Lamé in 1847: one can suppose $n = p$ is a prime and, for a primitive p-th root of unity ξ, he used prime factorisation in the ring $\mathbb{Z}[\xi]$ to prove Fermat's claim. As Liouville pointed out, Lamé's proof had a flaw, namely the fact that $\mathbb{Z}[\xi]$ doesn't have prime factorisation in general (as $\mathbb{Z}$ does, for example), but Kummer's work, that set the fundations of both modern algebra and algebraic number theory, showed that something could be saved.

Definition 1.1 Let p be a prime number, we say that it is *regular* if all the ideals $I \subseteq \mathbb{Z}[\xi]$ such that I^p is principal are principal.

Kummer made Lamé's proof work for regular primes, but we know that not all primes are regular (the first irregular one is 37). Even more, we know that there are infinitely many irregular primes, while it is still an open problem to determine if the regular ones

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: giacomo.graziani@math.unipd.it . Seminar held on June 12th, 2019.

are finite or not. In any case Kummer also gave a criterion to estabilish the regularity of a prime in terms of Bernoulli numbers (Definition 3.1).

Theorem 1.2 *Let $p \geq 3$ be a prime, then p is regular if and only if it divides the denominator of B_k for some $k = 2, \dots, p-3$.*

Proof. [Was82, Theorem 5.34, p. 78]. □

In view of computations of Euler (Remark 3.4) we see that the theorem of Kummer reduces the study of the regularity of primes to the study of congruence properties of special values

$$\zeta(-n) = -\frac{B_{n+1}}{n+1} \quad \text{for } n \geq 0 \text{ odd}$$

of the Riemann zeta function (that are indeed rational!). As we will see, the introduction of the p-adic numbers will give a nice tool for dealing with congruences modulo powers of p and therefore it is natural to look for a p-adic interpolation of the Riemann zeta function based on these special values, which will be seen as elements of $\mathbb{Q}_p$. More precisely we would like to prove a statement of the kind

Claim *There exists an analytic function $\zeta^\times$ on $\mathbb{Z}_p$ such that $\zeta(-n) = \zeta^\times(-n)$.*

This task can be accomplished in various ways, notably following the measure-theoretic approach of Kubota and Leopoldt (see [CoSu06]). Here we will follow the (almost) elementary approach of J.-P. Serre, built on previous works of Klingen and Siegel, to the interpolation of the special values discussed above, which has the advantage of taking us through another fundamental theme of modern number theory: that of complex and p-adic modular forms.

2 Riemann zeta function

Let us start considering one of the most famous functions in mathematics, the *Riemann zeta function*

$$\zeta(s) = \sum_{n \geq 1} n^{-s}.$$

This is a holomorphic function $\{\Re(s) > 1\} \to \mathbb{C}$ which can be analytically continued to $\mathbb{C} \setminus \{1\}$, having a simple pole at $s = 1$ with residue 1, and it satisfies the *functional equation*

$$\zeta(s) = 2^s \pi^{s-1} \sin\left(\frac{1}{2}\pi s\right) \Gamma(1-s)\, \zeta(1-s).$$

Here $\Gamma(s)$ denotes the analytic continuation of the function

$$s \mapsto \int_0^\infty t^{s-1} e^{-t} dt$$

from $\{\Re(s) > 0\}$ to $\mathbb{C} \setminus \mathbb{Z}_{<0}$. In view of the factor $\sin\left(\frac{1}{2}\pi s\right)$ appearing in the relation above, we find the so-called *trivial zeros* of ζ, namely we have

$$\zeta(-2n) = 0 \quad \text{for } n \in \mathbb{N}.$$

It can be shown, and this is essentially the same thing as the Fundamental Theorem of Arithmetic, that ζ can be reprensented as an *Euler product*

$$\zeta(s) = \prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right)^{-1}.$$

Corollary 2.1 *There are infinitely many primes.*

Proof. The harmonic series is divergent hence the product must be infinite. □

Note 2.2 This proof, given by Euler, was acknowledged by Dirichlet as one of the inspirations behind his theorem on arithmetic progressions.

3 Bernoulli numbers

A discussion about Riemann ζ function can't be untied from that on Bernoulli numbers.

Definition 3.1 Define the Bernoulli numbers B_k as

$$\frac{X}{e^X - 1} = \sum_{k \geq 0} B_k \frac{X^k}{k!}.$$

Lemma 3.2 *The Bernoulli numbers are rational and, for $k \geq 3$ odd, $B_k = 0$.*

Proof. The expansion of

$$\frac{e^X - 1}{X} = \frac{1}{X} \sum_{k \geq 1} \frac{X^k}{k!} = \sum_{k \geq 1} \frac{X^{k-1}}{k!} \in \mathbb{Q}[[X]].$$

Since the constant term is not zero, we see that it is invertible in the ring $\mathbb{Q}[[X]]^\times$, hence its inverse has rational coefficients. To prove the last statement we just need to see that the function

$$\frac{X}{e^X - 1} + \frac{X}{2} = 1 + \sum_{k \geq 2} B_k \frac{X^k}{k!}$$

is even, which is a straightforward computation. □

Example 3.3 Here we list the first few Bernoulli numbers

$$B_0 = 1,\ B_1 = -\frac{1}{2}, B_2 = \frac{1}{6}, B_3 = 0, \ldots, B_{12} = -\frac{691}{2730}, \ldots$$

Remark 3.4 It was shown by Euler, with a direct computation, that for every integer $n \geq 1$ the equality

$$\zeta(2n) = (-1)^{n+1} \frac{(2\pi)^{2n}}{(2n)!} \cdot \frac{B_{2n}}{2}$$

holds. It follows that

$$\zeta(-n) = (-1)^n \frac{B_{n+1}}{n+1} \quad \text{for every } n \geq 1 \text{ odd.}$$

3.4 Complex modular forms

We let Γ denote the group $\mathrm{SL}_2(\mathbb{Z})$, that is, the group of 2×2 matrices with entries in $\mathbb{Z}$ and determinant 1 and we let $\mathfrak{h} = \{\tau \in \mathbb{C} \mid \Im(\tau) > 0\}$ be the complex upper half plane. For $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$ and $\tau \in \mathfrak{h}$ we set

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

the so-called *Moebius transformation*. One checks that

$$\Im\left(\frac{a\tau + b}{c\tau + d}\right) = \frac{\Im(\tau)}{|c\tau + d|^2} > 0$$

hence this action preserves $\mathfrak{h}$. Modular forms are holomorphic functions on $\mathfrak{h}$ which are particularly well behaved with respect to this action of Γ.

Definition 4.1 Let $k \in \mathbb{Z}$, a *modular form of weight k* (and level Γ) is a holomorphic function $f : \mathfrak{h} \to \mathbb{C}$ such that

- for every $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$ we have

$$f\left(\frac{a\tau + b}{c\tau + d}\right) = f(\tau)(c\tau + d)^k; \tag{1}$$

- f is "holomorphic at infinity".

Note 4.2 We explain the condition at infinity: note that $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in \Gamma$, therefore a holomorphic function f that satisfies Equation (1) has the property that for every $\tau \in \mathfrak{h}$

$$f(\tau + 1) = f(T \cdot \tau) = f(\tau),$$

which means that is periodic of period 1. From classical analysis one knows that there exist a_n's with

$$f(\tau) = \sum_{n\in\mathbb{Z}} a_n e^{2\pi i n\tau}$$

(Fourier expansion). The condition at infinity is that $a_n = 0$ if $n < 0$. One usually sets

$$q = q(\tau) = e^{2\pi i\tau}$$

and $f(q) = \sum_{n\geq 0} a_n q^n$ is called the *q-expansion* of f.

Theorem 4.3 (q-expansion principle) *Let f, g be two modular forms with the same weight k, then $f = g$ if and only if they have the same q-expansion.*

Definition 4.4 For $k \in \mathbb{Z}$ denote M_k the $\mathbb{C}$-vector space of modular forms of weight k.

Theorem 4.5 *We have*

1. $M_k = 0$ *for* $k < 0$*, when* k *is odd and for* $k = 2$*;*
2. $\dim_{\mathbb{C}} M_k = 1$ *for* $k = 0, 4, 6, 8, 10$*;*
3. $\dim_{\mathbb{C}} M_k = 1 + \dim_{\mathbb{C}} M_{k-6}$.

Proof. [Ser94, Théorème 4, p. 88]. We can give a quick proof of the first fact: note that $\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \in \Gamma$ and, when k is odd, Equation (1) reads $f(\tau) = -f(\tau)$ and this shows that $M_k = 0$ when k is odd. All the other statements require a more sophisticated approach involving the Theorem of Riemann-Roch applied to compactified Riemann surface $\left(\frac{\mathfrak{h}}{\Gamma}\right)^* \simeq \mathbb{P}^1$. □

4.1 Eisenstein series

Let $k > 2$ be an even integer, we define

$$G_k(z) = \sum_{(0,0)\neq(m,n)\in\mathbb{Z}^2} \frac{1}{(m+nz)^k},$$

this is a well-defined holomorphic function $\mathfrak{h} \to \mathbb{C}$ (it is important to have $k \neq 2$ here) and it is clear that it satisfies Equation (1) in weight k. The q-expansion of G_k is given by

$$G_k(q) = 2\zeta(k)\left(1 - \frac{2k}{B_k}\sum_{n\geq 1}\sigma_{k-1}(n)\, q^n\right)$$

where

$$\sigma_{k-1}(n) = \sum_{d|n} d^{k-1},$$

in particular it is a modular form of weight k.

Definition 4.6 Let $k \geq 4$ be an even natural number, set

$$E_k = \frac{1}{2\zeta(k)} \cdot \frac{B_k}{2k} \cdot G_k = -\frac{B_k}{2k} + \sum_{n \geq 1} \sigma_{k-1}(n)\, q^n,$$

called the *Eisenstein series of weight k.*

Theorem 4.7 *A basis for the $\mathbb{C}$-vector space M_k is given by the monomials $E_4^\alpha E_6^\beta$ with $4\alpha+6\beta = k$. In particular the graded $\mathbb{C}$-algebra $M = \sum M_k$ is isomorphic to the polynomial algebra $\mathbb{C}[E_4, E_6]$ where E_i is given weight i.*

Example 4.8 We have $E_8 = E_4^2$, $E_{10} = E_4 E_6$, $E_{12} = \frac{1}{691}\left(441E_4^3 + 250E_6^2\right)$.

4.2 Examples from the geometry of elliptic curves and *Monstrous moonshine*

A first example is the *modular discriminant* $\Delta = (60G_4)^3 - 27(140G_6)^2$: it is a modular form of weight 12 with no zeros on $\mathfrak{h}$. Its q-expansion is given by

$$\Delta(q) = q \prod_{n \geq 1} (1-q^n)^{24}.$$

Its importance lies in the fact that a Weierstrass equation

$$y^2 = 4x^3 - g_4 x - g_6,$$

defines an elliptic curve (i.e. it is smooth) if and only if $g_4^3 - 27g_6^2 \neq 0$. In view of the moduli theoretic approach to the theory of modular forms, this shows that Δ has no zero on $\mathfrak{h}$ and a zero at infinity.

Another example arising from the same context is the *j-invariant*

$$j = 1728\frac{(60G_4)^2}{\Delta}$$

which is not exactly a modular form, since it has a pole at infinity. It has the property that two Weierstrass equations as above define isomorphic elliptic curves if and only if j takes the same value.

Note The j-invariant is also famous in view of its q-expansion

$$j(q) = \frac{1}{q} + 744 + 196884q + 21493760q^2 + 864299970q^3 + \cdots \in \frac{1}{q} + \mathbb{Z}[[q]].$$

Let $\mathbb{M}$ be the Monster Group, then if r_n is the sequence of the (increasing) dimensions of its irreducible representations, we have

$$\begin{aligned} 1 &= r_1 \\ 196884 &= r_1 + r_2 \\ 21493760 &= r_1 + r_2 + r_3 \\ 8642999720 &= 2r_1 + 2r_2 + r_3 + r_4 \end{aligned}$$

and so on. This surprising connection (and subsequent extensions) are known as *Monstrous moonshine* and their explanation set connections among modular functions, string theory, quantum gravity and representation theory.

5 p-adic modular forms

Remark 5.1 In view of the so-called *q-expansion principle* we can identify a modular form with its q-expansion.

For an integer k we denote $\mathcal{M}_k \subseteq M_k$ the subset of weight k modular forms whose coefficients are rationals with p-adic norm ≤ 1.

Definition 5.2 Given $f = \sum_{n\geq 0} a_n q^n \in \mathbb{Q}_p\left[[q]\right]$, we define

$$|f|_p = \sup_{n\geq 0} |a_n|_p \,.$$

A *p-adic modular form* is an element $f \in \mathbb{Q}_p\left[[q]\right]$ for which there exists a sequence $f_i \in \mathcal{M}_{k_i}$ with $f_i \to f$ uniformly (i.e. $|f - f_i|_p \to 0$).

5.1 p-adic weights

An important issue is to determine what happens to the weights k_i when taking the limit in Definition 5.2.

Theorem 5.3 *Let f, g be in M_k and M_t respectively (i.e. with rational coefficients) and suppose that*

$$|f - g|_p \leq p^{-m} \, |f|_p$$

for some $m \geq 1$. Then if $f \neq 0$ we have

$$k \equiv t \pmod{(p-1)\, p^m}.$$

Proof. [Ser73, Théorème 1, p. 198]. □

Recall that the Chinese Reminder Theorem states that for every $m \geq 1$ we have an isomorphism

$$\frac{\mathbb{Z}}{(p-1)\, p^m} \simeq \frac{\mathbb{Z}}{p-1} \times \frac{\mathbb{Z}}{p^m}$$

and, in view of our description of elements of $\mathbb{Z}_p$ we have

Corollary 5.4 *Let f be a p-adic modular form and suppose that $f_i \in \mathcal{M}_{k_i}$ is a sequence converging to f. Then the sequence of weights $(k_i)_i$ defines an element of $\frac{\mathbb{Z}}{p-1} \times \mathbb{Z}_p$ which does not depend on the sequence $(f_i)_i$.*

Definition 5.5 Call $X = \frac{\mathbb{Z}}{p-1} \times \mathbb{Z}_p$ the *p-adic weight space*. It comes with a natural map $\epsilon : X \to \mathrm{End}^{\mathrm{c}}_{\mathbb{Z}}\left(\mathbb{Z}_p{}^{\times}\right)$ given by

$$v^{\kappa} := \epsilon\left(\kappa\right)\left(v\right) = v_1^u v_2^s$$

where $\kappa = (u, s) \in X$ and $v = v_1 v_2 \in \mathbb{Z}_p{}^{\times} \simeq \mu_{p-1} \times (1 + \mathfrak{m}_p)$. We say that κ is *even* if $u \in \frac{2\mathbb{Z}}{p-1}$, otherwise we say that it is *odd*.

5.2 Getting the job done

Fix an element $\kappa \in X$ and a sequence of positive even integers $(k_i)_i$ such that $|k_i|_\infty \to \infty$ for the usual absolute value and $k_i \to \kappa$ in X. For a natural number d we have p-adically

$$\lim_{i\to\infty} d^{k_i - 1} = \begin{cases} 0 & p \mid d \\ d^{\kappa - 1} & p \nmid d \end{cases}$$

and hence

$$\lim_{i\to\infty} \sigma_{k_i - 1}\left(n\right) = \sigma^{\times}{}_{\kappa-1}\left(n\right) := \sum_{d \mid n,\, p \nmid d} d^{\kappa - 1}.$$

It follows that

$$\lim_{i\to\infty} \left(E_{k_i} + \frac{B_{k_i}}{2k_i} \right) = \sum_{n \geq 1} \sigma_{\kappa-1}{}^{\times}\left(n\right) q^n.$$

Theorem 5.6 (Serre) *Let*

$$f^{(i)} = \sum_{n \geq 0} a_n^{(i)} q^n$$

be a sequence of p-adic modular forms of weights $\kappa^{(i)}$ and suppose that

1. *for every $n \geq 1$, the sequence $a_n^{(i)}$ has a p-adic limit a_n;*
2. *the sequence $\kappa^{(i)}$ has a limit κ in X.*

Then the sequence $a_0^{(i)}$ has a p-adic limit a_0 and moreover $f = \sum_{n\geq 0} a_n q^n$ is a p-adic modular form of weight κ with $f^{(i)} \to f$.

In view of Serre's Theorem we conclude that there exists the limit

$$a_0 = \lim_{i\to\infty} \left(-\frac{B_{k_i}}{2k_i} \right),$$

but in view of our discussion on Riemann ζ function this means that there exists

$$\zeta^{\times}\left(1 - \kappa\right) = \lim_{i\to\infty} \zeta\left(1 - k_i\right)$$

defined on the even elements of $X \setminus \{1\}$. Since $X = \frac{\mathbb{Z}}{p-1} \times \mathbb{Z}_p$, we will see $\zeta^{\times}$ as a function in two variables, we then have $p - 1$ functions

$$s \mapsto \zeta^{\times}\left(u, s\right)$$

such that

Theorem 5.7

1. $\zeta^{\times}(\bullet, u) \equiv 0$ *for* u *even, while for* u *odd it has a finite number of zeros on* $\mathbb{Z}_p$*;*
2. *For* $n \leq 0$ *an integer, let* $u = n \pmod{p-1}$*, then*

$$\zeta^{\times}(n, u) = \left(1 - p^{-n}\right)\zeta(n) = \prod_{\ell \neq p \text{ prime}} \left(1 - \ell^{-n}\right)^{-1}$$

is the prime-to-p part *of the Riemann* ζ *function.*

Proof.

1. The first statement follows from the fact the $B_k = 0$ for odd k, while the second follows from the analyticity property of $\zeta^{\times}$ (we won't talk about) on Weierstrass Preparation Theorem.
2. This is more complicated and follows after comparing $\zeta^{\times}$ with Kubota-Leopoldt p-adic ζ function, see [Ser73, Théorème 3, p. 206].

□

A p-adic numbers

Over the rational numbers $\mathbb{Q}$ we have the standard absolute value $|x|_{\infty} = \max\{x, -x\}$ and this leads to the well known field of real numbers. There are other norms on $\mathbb{Q}$: fix a prime number p and, for $\frac{a}{b} \in \mathbb{Q}\setminus\{0\}$ write

$$\frac{a}{b} = p^n \frac{x}{y}$$

where both x, y are coprime with p. Of course here n ranges in $\mathbb{Z}$. We set

$$\left|\frac{a}{b}\right|_p = p^{-n}.$$

Example A.1 We have

- $|m|_p \leq 1$ for every $m \in \mathbb{Z}$, called the *non-archimedean* condition
- $\left|\frac{7}{8}\right|_2 = 8$; $\left|\frac{140}{297}\right|_5 = \frac{1}{5}$.

This assigment satisfies

- $|x|_p = 0 \iff x = 0, \qquad |1|_p = 1$
- $|xy|_p = |x|_p |y|_p$

so no surprise so far. The interesting fact about $|\bullet|_p$ is that the the non-archimedean condition is equivalent to the following *ultrametric inequality*

$$|x+y|_p \leq \max\left\{|x|_p, |y|_p\right\}$$

which is much stronger than the usual condition $|x+y| \leq |x| + |y|$. We will list a bunch of consequences later.

Definition A.2 We define $\mathbb{Q}_p$ as the completion of $\mathbb{Q}$ with respect to the absolute value $|\bullet|_p$. We define

$$\mathbb{Z}_p = \left\{x \in \mathbb{Q}_p \mid |x|_p \leq 1\right\}$$
$$\mathfrak{m}_p = \left\{x \in \mathbb{Q}_p \mid |x|_p < 1\right\}.$$

Remark A.3 Using the ultrametric property one sees that $\mathbb{Z}_p$ is a ring (the point being that it is closed under addition), furthermore it is a local ring whose maximal ideal is $\mathfrak{m}_p$ which is principal generated by p and $\mathbb{Q}_p$ is its fraction field. This definition allows to describe $\mathbb{Z}_p$ explicitly: it is the completion of $\mathbb{Z}$ with respect to $|\bullet|_p$, hence an element $x \in \mathbb{Z}_p$ is an (equivalence class of) sequence $(a_n)_n \subseteq \mathbb{Z}$ with the property that for every $m \in \mathbb{N}$ there exists an integer N such that

$$a_n \equiv a_{n+1} \mod p^m \quad \text{for every } n \geq N,$$

that is which is definitively constant modulo p^m for every m. A consequence of this is that there are well defined maps $\mathbb{Z}_p \to \mathbb{Z}/p^m$ (since every element is eventually constant mod p^m) whose kernel is $\mathfrak{m}_p^m = p^m\mathbb{Z}_p$).

Example A.4 Another remarkable consequence of the ultrametric property is that, given a sequence $(a_n)_n \subseteq \mathbb{Q}_p$, then

$$\sum_{n\geq 0} a_n \text{ is convergent if and only if } \lim_{n\to 0} a_n = 0.$$

As an application we check that $1 + \mathfrak{m}_p$ (where 1 here denotes the constant sequence $(1, 1, \dots)$) is a multiplicative group: in fact given $x \in \mathfrak{m}_p$ we have formally

$$\frac{1}{1+x} = 1 - x + x^2 - \cdots = \sum_{n\geq 0} (-1)^n x^n$$

but this series is convergent by the criterion above, since $x \in \mathfrak{m}_p$ means that $|x|_p < 1$.

Note A.5 As a way to become more confident with p-adic numbers, we discuss here a topic that will play a major role: that of p-adic interpolation. In the form we are going

to need it, the problem asks if, given a function $f : \mathbb{Z} \to X$ where X is some space, there exists a continuous function $f : \mathbb{Z}_p \to X$ that restricts to f. Of course this amounts to check that f is continuous with respect to the p-adic topology. Let $x \in 1+\mathfrak{m}_p$, the simplest function, also in view of Example A.4, we would like to interpolate is

$$\begin{aligned} f_x : \mathbb{Z} &\to 1 + \mathfrak{m}_p \\ n &\mapsto x^n \end{aligned}$$

We compute

$$|x^p - 1|_p = |x-1|_p \left|x^{p-1} + \cdots + x + 1\right|_p \leq p^{-1} |x-1|_p$$

since $x \equiv 1 \pmod{p}$ hence $x^{p-1} + \cdots + x + 1 \equiv 0 \pmod{p}$. More generally for $n \in \mathbb{Z}$ write $n = tp^r$ with $p \nmid t$, then

$$|x^n - 1|_p \leq p^{-r} |x-1|_p = |n|_p |x-1|_p .$$

It follows that f_x is continuous with respect to the p-adic topology and hence the function

$$\begin{aligned} f_x : \mathbb{Z}_p &\to 1 + \mathfrak{m}_p \\ \alpha = [(a_n)_n] &\mapsto x^\alpha := \lim_{n\to\infty} x^{a_n} \end{aligned}$$

is well defined and does not depend on the particular sequence we pick to represent α. The group $1 + \mathfrak{m}_p$ is particularly important, in fact one can show that there exists a natural group homomorphism (called the Teichmüller character) $[\bullet] : {\mathbb{F}_p}^\times \to {\mathbb{Z}_p}^\times$ such that the composition

$${\mathbb{F}_p}^\times \xrightarrow{[\bullet]} {\mathbb{Z}_p}^\times \xrightarrow{(\mathrm{mod}\ p)} {\mathbb{F}_p}^\times$$

is the identity. This shows that there is a natural decomposition

$${\mathbb{Z}_p}^\times \simeq \mu_{p-1} \times (1 + \mathfrak{m}_p) .$$

This will play a role talking about p-adic weights.

Exercise A.6 Show that the map $f_x : \mathbb{Z}_p \to 1 + \mathfrak{m}_p$ in Note A.5 is injective for $x \neq 1$.

B This is very nice, but why bother?

I am writing this section because everybody, at some point, wonders why people care so much about Riemann hypothesis.

Conjecture (Riemann hypothesis, or RH) *The zeros of ζ in the strip $0 \leq \Re(s) \leq 1$ lie on the line $\Re(s) = \frac{1}{2}$.*

Let $x > 0$ be a real number, we define

$$\pi(x) = \#\{\text{primes} \leq x\},$$

so that $\pi(1) = 0$, $\pi(5) = \pi(5.57) = \pi(6) = 3$. A major progress in number theory is the following result proved independently by de la Vallée-Poussin and Hadamard in 1896

Theorem (Prime Number Theorem)

$$\lim_{x\to\infty} \frac{\pi(x)}{x/\ln(x)} = 1.$$

Which means that for x very large, $\pi(x)$ can be computed as $x/\ln(x)$ with good approximation.

The proof of the Prime Number Theorem has little to do with Riemann hypothesis, in fact de la Vallée-Poussin and Hadamard had to prove that ζ has no zeros on the line $\Re(s) = 1$, while an elementary proof given by Selberg and Erdos in 1949 made no use of ζ at all (even if de la Vallée-Poussin and Hadamard's estimations were more precise). In any case assuming it one can prove

Theorem *Assuming RH we have*

$$\pi(x) = \frac{x}{\ln(x)} + \mathcal{O}\left(x^{\frac{1}{2}}\ln(x)\right).$$

We give now a funny application of *RH*: let $n = p_1^{k_1}\dots p_t^{k_t}$ be the decomposition into prime factors of a natural number n, where the p_i's are distinct and $k_i \geq 1$. We define

$$\mu(n) = \begin{cases} (-1)^t & \text{if } t_i = 1 \text{ for every } i \\ 0 & \text{otherwise} \end{cases}$$

called the Möbius function, and for $x \in \mathbb{R}^{>0}$

$$M(x) = \sum_{n\in\mathbb{N},\, n\leq x} \mu(n).$$

Theorem B.1 *Without any condition, we have*[(3)]

$$M(x) = \Omega\left(x^{\frac{1}{2}}\right).$$

[(3)] There is some ambiguity in the meaning of $\Omega(\bullet)$. Following the Hardy–Littlewood definition, we say that (always understood for $x \to +\infty$)

$$f(x) = \Omega(g(x)) \text{ if } \limsup_{x\to\infty} \left|\frac{f(x)}{g(x)}\right| > 0.$$

Proof. The funny part of this Theorem is its proof (cfr. [TiHB86, Theorem 14.26 (B), p. 371]). First one shows it assuming RH is false, and then one shows it assuming RH is true. In the end this gives an unconditional proof. □

Another fact about $M(x)$ is the following

Theorem B.2 *RH is equivalent to*

$$M(x) = \mathcal{O}\left(x^{\frac{1}{2}+\epsilon}\right).$$

References

[Ser73] J.-P. Serre, *Formes modulaires et fonctions zéta p-adiques.* Modular functions of one variable III, Proc. Internat. Summer School, Univ. Antwerp (1972),191–268; Lecture Notes in Math. 350, Springer, Berlin (1973).

[Ser74] J.-P. Serre, *Fonctions Zéta p-adiques.* Mémoires de la S.M.F., tome 37 (1974), 157–160.

[Was82] L.C. Washington, "Introduction to Cyclotomic Fields". Springer GTM 83, 1982.

[TiHB86] E.C. Titchmarsh, R.D. Heat-Brown, "The Theory of the Riemann Zeta-function". Second Edition, Oxford Science Publications, 1986.

[Pat88] S.J. Patterson, "An introduction to the theory of the Riemann Zeta-Function". Cambridge 14, 1988.

[Ser94] J.-P. Serre, "Cours d'Arithmétique". 4e édition, Presses Universitaires de France, 1994.

[CoSu06] J. Coates, R. Sujatha, "Cyclotomic Fields and Zeta Values". Springer, 2006.