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

Scuole di Dottorato in Matematica Pura e Computazionale

Seminario Dottorato 2020/21

Preface

This document offers an overview of the schedule of Seminario Dottorato 2020/21.

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 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.

As it happened for the last one, also this academic year has been strongly conditioned by the COVID-19 pandemic emergency, so that all sessions of the seminar – excepted the one in the Opening Day – have been held online. Anyway we feel that keeping up this activity has helped the PhD students to stay in contact among them and with the Department life.

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

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

Padova, June 20th, 2021

Corrado Marastoni, Tiziano Vargiolu

Abstracts (from Seminario Dottorato's webpage)

Wednesday 7 October 2020

Mathematical modeling in primary school: an example of research in Math Education

Simone PASSARELLA (Padova, Dip. Mat.)

Introducing the distributivity property of multiplication over addition is a well-known challenge in Mathematics Education, especially in primary school. In this seminar, after a brief introduction to the research field of Mathematics Education, the results of a cycle of design research in which 2nd-grade students are introduced to the key concept of distributivity of multiplication over addition are described. The focus will be in showing how the heuristics of didactical phenomenology, guided reinvention and emergent modelling may guide the design and implementation of a modelling activity to make distributivity property more accessible for primary school students.

Wednesday 18 November 2020

Homological algebra: deforming abelian groups using torsion pairs

Sergio PAVON (Padova, Dip. Mat.)

Given an abelian category, one can (often) construct its derived category. We illustrate this process for the familiar category Ab of abelian groups and group homomorphisms. We also give some motivation for this construction, as we briefly see that derived categories provides the "correct" environment for the many (co)homological theories found throughout mathematics. Having understood what the derived category of an abelian category is, we address the natural question of "derived equivalences": when do two abelian categories have the same derived category? This turns out to be an interesting but very difficult question. We then restrict ourselves to abelian categories obtained by "deforming" the category Ab of abelian groups: a major role in this construction is played by torsion pairs, whose definition we will illustrate with some examples. We will conclude with a nice theorem, which states that deforming Ab with any torsion pair enjoing a certain property ("hereditarity") always yields an abelian category derived equivalent to Ab. (This seminar is based on a joint work with J. Vitória.)

Wednesday 9 December 2020

Mathematical modeling in Finance

Andrea MAZZORAN (Padova, Dip. Mat.)

In mathematical finance, tools from probability, and more specifically from stochastic analysis, are applied to tackle problems arising in financial markets. Therefore, one needs to build suitable models to describe the phenomena that appear in the market. In the first part of the talk, I am going to present some basic probability preliminaries and some intuitive concepts in finance. Then, I will move on to consider the more famous models in financial mathematics, describing their main properties. To be accessible to a mathematical audience of non-experts, I will extensively use examples, graphics and intuitive definitions In the very last part of the talk, the more technical one, I will mention some result of my research, more precisely, I will introduce a joint calibration problem via stochastic-local volatility model along with an exercise based on real data.

Wednesday 16 December 2020

Weak KAM, Homogenization and Ergodic Control: an introduction

Hicham KOUHKOUH (Padova, Dip. Mat.)

This seminar is intended for non-specialists to whom I would first and succinctly introduce the general ideas behind the weak KAM theory in dynamical systems, the theory of homogenization in the analysis of PDEs and the theory of ergodic control. I will then and especially insist on the link between them, and which manifests itself in a particular PDE known as "ergodic stationary Hamilton-Jacobi equation". A qualitative study of these problems will also be discussed, and a link with an optimization problem in the space of measures will be mentioned. I will finally provide an example of Singular Perturbations in a control problem. For the sake of clarity, I will focus on the deterministic case, but much of these results are valid in a stochastic framework. A discussion will conclude the presentation.

Wednesday 3 February 2021

An introduction to the moduli space of smooth curves and its compactification

Angelina ZHENG (Padova, Dip. Mat.)

The moduli space of algebraic curves is a central object in algebraic geometry. The idea behind this space is that it answers to a classification problem, by allowing us to classify algebraic curves up to isomorphisms. Nonetheless, the geometry of this space is rather abstract and subtle, and only few general statements are known. The aim of this talk is to give an elementary introduction to the moduli space of npointed smooth curves of genus g and its compactification: we will define the main ingredients and present some basic properties. We will also discuss basic examples in order to get familiar with these spaces. One way to understand these spaces is by computing some topological invariants, such as their cohomology groups. In general, their full cohomology ring is still unknown, but we have a complete description for some values of g, which I will present in the final part.

Wednesday 17 February 2021

Limit theorems for Lévy flights on a 1D Lévy random medium

Elena MAGNANINI (Padova, Dip. Mat.)

The purpose of this introductory presentation is to derive limit theorems for a class of random walks on random point processes, which mathematically describe the motion of a particle on a random environment. In the physical literature, the continuous-time version of this process is called Lévy-Lorentz gas. This model embodies the features of many different natural systems that show the so-called superdiffusive behavior, from particles moving in a turbulent fluid to the motion of crawling amoeba and the foraging strategies of several animals. We will start by recalling some classic results and definitions regarding the distributional convergence of stochastic processes, then we will move to the presentation of our model, starting from the definition of a one-dimensional random environment. We will finally present the convergence results for the process Y which takes place on such environment and that under suitable hypotheses displays a super-diffusive behavior. In particular, we will determine the scale whereby Y converges to a non-null limit and the limit process to which it converges.

Wednesday 3 March 2021

A differential eigenproblem of electromagnetics

Michele ZACCARON (Padova, Dip. Mat.)

We start with a preliminary overview of Sobolev spaces, very useful in the modern approach to solve PDEs. We then introduce an eigenvalue problem arising from time-harmonic Maxwell's equations, deriving its so-called weak formulation and showing some first properties. The last part of the seminar will focus on the domain perturbation, inspecting the dependence of the eigenvalues upon variation of the region in which we set our problem, and presenting a result regarding shape optimization. The talk is of introductory type and its purpose is to let the audience have a look at some of the tools and concepts from the spectral theory of differential operators.

Wednesday 17 March 2021

Optimal stopping theory and American options

Francesco ROTONDI (Padova, Dip. Mat.)

The optimal stopping problem is a classical one within stochastic calculus theory. Formally, given

a gain process, the optimal stopping problem is about finding the stopping time that maximizes the expected gain. In discrete time, this problem is solved using dynamic programming techniques and setting up a backward recursion. This delivers both the optimal stopping rule and the optimal value process. This optimal value process coincides with the Snell Envelope of the gain process, namely, the smallest supermartingale dominating it. Optimal stopping theory closely relates to the American derivatives valuation problem. American derivatives are financial contracts characterized by a payoff process that depends on an underlying stochastic process (usually the price of a traded asset). The holder of an American derivative chooses when to cash in the payoff, trying to do so optimizing the gain. Therefore, the fair price of this derivate depends on its optimal stopping time. During my talk, which will be accessible also to non-specialists, I will first describe and then show how to solve the optimal stopping problem in a discrete time setting. Secondly, I will show how to apply these techniques for the valuation of American derivatives in the most standard setting. Finally, I will present a result on how to price these derivatives in a more sophisticated market model.

Wednesday 31 March 2021

On the simulation of planar homogeneous flows

Francesca TEDESCHI (Padova, Dip. Mat.)

This seminar presents a brief overview on Molecular Dynamics (NEMD) simulations. The basics of NEMD will be reviewed to make the presentation as accessible as possible. Molecular Dynamics is a deterministic tecnique, used to produce statistical data about the behavior of non-Newtonian fluids or other materials at the microscopic scale. Computational efficiency imposes to consider only a portion of fluid (represented by the simulation box), and to apply periodic boundary conditions, i.e., surrounding that box with periodic copies of its particles represented by a lattice deformed under the action of the imposed homogeneous flow. Requirements of compatibility and reproducibility, that guarantee the reliability of the simulation for an indefinite time, will be explained, together with the algorithms to correctly impose the periodic boundary conditions.

Wednesday 21 April 2021

Topics in Numerical Linear Algebra for High-Performance Computing

Monica DESSOLE (Padova, Dip. Mat.)

As computer architectures evolve, numerical algorithms have to cope with high resolution simulations and data integration that are now key to many research fields. Solution methods for real world problems require a constantly increasingly computational effort, demanding for more and more resources and tailored algorithms. In this talk we will introduce some recent high-performance algorithmic developments about solving dense linear systems, possibly numerical rank-deficient,

and we will present some parallel techniques for the solution of large-scale sparse systems of linear equations.

Wednesday 5 May 2021

An intuitive introduction to p-adic geometry

YUKIHIDE NAKADA (Padova, Dip. Mat.)

The field of p-adic numbers forms a bridge from number theory to analysis, mixing number-theoretic constructions with analytic ideas. And just as the complex numbers made possible complex analytic geometry, the p-adic numbers opened up a new, strange geometry over a number-theoretic field. The goal of this talk is to paint a broad picture of p-adic geometry accessible to mathematicians without any previous knowledge of number theory. We set the foundations with an intuitive introduction to the field of p-adic numbers and then give a brief introduction to socalled rigid geometry, the p-adic counterpart to complex analytic geometry.

Wednesday 19 May 2021

An introduction to singular control problems through an electricity market model

ALMENDRA AWERKIN VARGAS (Padova, Dip. Mat.)

We start motivating this seminar with the following question: at which electricity price it is optimal to increase the current installed power in order to obtain the maximum profit of selling the produced energy? We will see that the mathematics behind our model is called singular control theory and, always leaning in our electricity market example, we will briefly introduce the main concepts and results that define this specific branch of control theory. We conclude the talk answering our question considering the Italian market case.

Wednesday 9 June 2021

Synchronization and asymptotic dynamics of mechanical systems: an introduction

SARA GALASSO (Padova, Dip. Mat.)

Synchronization is a fascinating and eye-catching phenomenon, which spontaneously emerges from the collective behaviour of a huge variety of interacting systems. Examples permeate science, from fireflies to metronomes, from neurons to celestial bodies. In this seminar we shall focus on mechanical systems, having in mind, in particular, systems of coupled pendula. To construct a physical-mathematical model able to describe synchronicity patterns in their evolution, classical

tools from dynamical systems theory are essential. We will therefore recall the basic notions of invariant manifold and stability, as well as some results that will allow us to investigate, at an introductory level, the long-time asymptotic behaviour. Along with the theory, we will provide examples and examine simple models which should help us visualize some fundamental mechanisms underlying synchronization.

Mathematical modeling in primary school: an example of research in Math Education

Simone Passarella (*)

Abstract. Introducing the distributivity property of multiplication over addition is a well-known challenge in mathematics education, especially in primary school. In this contribution, after a brief introduction to the research field of mathematics education, the results of a cycle of design research in which 2nd-grade students are introduced to the key concept of distributivity of multiplication over addition are described. The focus will be in showing how the heuristics of didactical phenomenology, guided reinvention and emergent modelling may guide the design and implementation of a modelling activity to make distributivity property more accessible for primary school students.

European Didactics Traditions in Mathematics

Mathematics education, or better in the European context didactics of mathematics, refers to the discipline dealing with all aspects of teaching and learning mathematics. Several traditions concerning didactics of mathematics developed in Europe, both in the practice of learning and teaching at school and in research and development. Despite the variety of cultural, historical and political backgrounds, all these traditions share some common features, such as a strong connection with mathematics and mathematicians, the key role of theory, and the key role of design activities for learning and teaching (Blum, Artigue, Mariotti, Strasser and Van den Heuvel-Panhuizen, 2019).

Design activities in the didactics of mathematics can involve the design of tasks, lessons, teaching sequences, textbooks, curricula, assessments, and ICT-based material or programs for teacher education. It is through designed instructional material and processes, in which the intended *what* and *how* of teaching is operationalised, that learning environments for students can be created. In this direction, educational design forms a meeting point of theory and practice through which they influence each other reciprocally. In France, the design of mathematical tasks, situations and sequences of situations is essential to didactic research. This is clearly reflected in the methodology of didactical engineering within the

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: passarel@math.unipd.it. Seminar held on 7 October 2020.

theory of *didactical situations* (Brousseau, 1997) that emerged in the early 1980s. Designs are grounded in epistemological analyses, and situations are sought that capture the epistemological essence of the mathematics to be learned. In the last decade, the anthropological theory of the didactic (Chevallard and Sensevy, 2014) has developed its own design perspective that gives particular importance to identifying issues that question the world and have strong mathematical potential. In the Netherlands, a strong tradition in design can be found in the didactics of mathematics. At the end of the 1960s, the reform of mathematics education started with designing an alternative for the mechanistic mathematics education that then prevailed. Initial design activities were practice oriented, and the theory development that resulted in *Realistic Mathematics Education* (Freudenthal, 1991) grew from this practical work and later guided further design activities. Design implementation, including contexts, didactical models, longitudinal teaching-learning trajectories, textbook series, examination programs, mathematics events, and digital tools and environments, has been realised. In Italy, the role of design has also changed over time. The period from the mid-1960s to the mid-1980s was characterized by a deep epistemological concern and a strong pragmatic interest in improving classroom mathematics teaching. The focus was on the content and its well-crafted presentation in practice, based on conceptual analyses. The period from the mid-1980s to the present can be characterised by long and complex processes targeting the development of theoretical constructs based on teaching experiments, with the design of teaching and learning environments as both an objective and a means of the experimentation. Within the German didactic tradition, two periods can be distinguished. Before the 1970s and 1980s, design activities were mostly meant for developing learning and teaching environments for direct use in mathematics instruction. These design activities belonged to the long German tradition of *Stoffdidaktik*, which focused strongly on mathematical content and course development. In the 1970s, an empirical turn occurred, resulting in design activities done to study the effect of specified didactical variables through classroom experiments.

In the following part of this contribution, an example of research in mathematics education in which the role of design is central is being presented.

Mathematical Modelling in Primary School

This example is part of a PhD research project that aims at studying how mathematical modelling can be integrated in the regular school practice in the Italian context. specifically, it will be reported an example that focuses on the question of how to provide students with opportunities to be introduced to the concept of distributivity of multiplication over addition (DP) in a meaningful way. Indeed, students still encounter difficulties in understanding distributivity (Squire, Davies and Bryant, 2004), with consequently efforts in algebra and calculations. In this direction, it appears central to investigate how students can learn distributivity, what learning steps are needed to reinvent this concept and which learning activities may foster such learning steps. Our idea is that making students face with realistic and less stereotyped problems that take into consideration the experiential world of students, should support students' in making sense of mathematical concepts

(Chamberlin, Payne and Kettler, 2020), leading to a deeper understanding of the concept of distributivity. In this direction, mathematical modelling could play a central role in fostering a process of reinvention (Freudenthal, 1991) of the distributivity property offering students opportunities to attach meaning to this mathematical construct developed while solving real problems. Indeed, we believe that mathematical modelling could represent a valid educational strategy to improve students' understanding of DP, engaging them in meaningful real-life problem solving situations (Lesh and Lehrer, 2003). Not only a better understanding, but also a process of reinvention (Freudenthal, 1991) of DP is aimed to be achieved by students. Modelling could support such reinvention process of DP, making it more significant for students because rooting it in their personal experience. The aim of the research reported here is to design, implement, and evaluate a modelling activity in a reinvention process to introduce 2nd -grade students to the distributivity property in a meaningful way.

In the next two sections the theoretical perspectives followed in this study are presented.

Realistic Mathematics Education

Realistic Mathematics Education (RME) is a domain specific instruction theory for mathematics that offers a pedagogical and didactical philosophy on mathematical learning and teaching as well as on designing instructional materials for mathematics education. *Rich* and *realistic* situations are given a prominent position in the learning process and represent a starting point for the development of mathematical concepts and applications (Gravemeijer and Doorman, 1999). Realistic refers to problem situations that students can image and that are, at a certain stage, meaningful for them. Therefore, problems can come from the real world, but also from a fantasy world or from the formal world of mathematics, as long as the problems are experientially real in students' mind (Van den Heuvel-Panhuizen and Drijvers, 2014). However, this realistic connotation is not sufficient to have a valuable mathematical problem. The context, indeed, must also be *rich* (Freudenthal, 1991), referring to a situation that promotes a structuring process as a means of organizing phenomena, physical and mathematical, and even mathematics as a whole (Treffers, 1987). The core principles of RME have been synthetized in six educational principles (for a complete discussion see Van den Heuvel-Panhuizen and Drijvers, 2014): *activity principle*; *reality principle*; *level principle*; *intertwinement principle*; *interactivity principle*; *guidance principle*. Based on these principles RME also offers the following design heuristics: *guided reinvention*, *didactical phenomenology*, and *emergent models* (Gravemeijer, 1994).

Guided reinvention is based on the fact that the teaching of mathematics should be a human activity as opposed to a ready-made system (Freudenthal, 1973; 1991). When students progressively mathematize their own mathematical activity (Treffers, 1987) they can reinvent mathematics under the guidance of the teacher and the instructional design. This is the meaning of the first heuristic *guided reinvention*: students should experience the learning of mathematics as a process similar to the process by which mathematics was invented (Gravemeijer, 1994). Consequently, the role of the designer is fundamental in RME, since she/he has to foster this process of guided reinvention. In so doing, the designer can use different methods, such as using students' informal solution strategies as

a source supporting students' solutions in getting closer to the end goal (Bakker, 2004).

Mathematical concepts and tools serve to organize phenomena, both from daily life and from mathematics itself (Bakker, 2004). A phenomenology of a mathematical concept is an analysis of that concept in relation to the phenomena it organizes. *Didactical phenomenology* is the study of concepts in relation to phenomena with a didactical interest. In this perspective the challenge is to find phenomena that beg to be organized by the concepts that are to be taught (Freudenthal, 1983). In this research, the design of instructional materials will follow a didactical phenomenology approach, in which the goal is to find problem situations that could provide the basis for the development of the mathematical concept of DP.

Fundamental tools for bridging the gap between the informal, context-related mathematics and the more formal mathematics are models. Indeed, in RME *models of* a certain situation can become *models for* more formal reasoning (Gravemeijer 1994; 1999). The movement from situational to formal reasoning is described by the following four levels (Gravemeijer, Cobb, Bowers and Whitenack, 2000): *situational level*; *referential level*; *general level*; *formal level*. This shift is at the basis of the notion of emergent modelling, that will be treated in detail in the next section.

Mathematical modelling

The promotion of mathematical modelling is accepted as a central goal of mathematics education worldwide, especially if mathematics education aims to promote responsible citizenship (Kaiser, 2017). Modelling is a creative process of making sense of real-world phenomena through different phases of the modelling process, such as structuring real-world situations that have to be modeled; mathematising, i.e. translating a real situation into mathematical terms; working with the resulting model finding solutions to the mathematical problems; interpreting and validating results in relation to the real starting situation; monitoring the entire process of modelling.

Different perspectives concerning the way to integrate mathematical modelling into classroom teaching emerged, emphasizing either the solution of the original problem or the development of mathematical concepts or ideas. In this study with mathematical modelling we refer to *emergent modelling*, in which real-world examples and their relations to mathematics are central elements of the structure of the teaching and learning process (Freudenthal, 1991; Treffers, 1987).

Emergent modelling was initially developed by Gravemeijer (1999) with the meaning of supporting the emergence of formal mathematical ways of knowing. The underlying educational theory is RME, in which models have always employed to foster a process in which formal mathematics is re-invented by students themselves. Indeed, in this perspective modelling activities are used as a vehicle for the development, rather than applications, of mathematical concepts (Greer, Verschaffel and Mukhopadhyay, 2007). Students, starting from a real context, begin to model their informal mathematical strategies and arrive to reinvent mathematical concepts and applications they need, and that can subsequently be formalized in mathematical terms and generalized to other situations. As a consequence, emergent modelling can be seen as a long-term dynamic process from a *model of* stu-

dents' situated informal mathematical strategies to a *model for* more formal mathematical reasoning (Gravemeijer and Doorman, 1999), that favours understanding, reasoning and sense-making. This transition from *model of* to *model for* involves the constitution of a new mathematical reality (Streefland, 1985) that can be denoted as formal in relation to the original starting points of the students.

Design research phases

As stated in the previous section, the aim of this study consists in studying how mathematical modelling could support the introduction of DP at primary school. Consequently, *design research* is used as the research method of this study (Doorman, 2019; Bakker, 2018; Gravemeijer, 2004; Cobb et al., 2003; Edelson, 2002), since we needed to create an instructional environment with which it could be possible to study how and to what extent the suggested process could be fostered.

> (...) design research explicitly exploits the design process as an opportunity to advance the researchers' understanding of teaching, learning, and educational systems. Design research may still incorporate the same types of outcome-based evaluation that characterize traditional theory testing, however, it recognizes design as an important approach to research in its own right. (Edelson 2002, p. 107)

Design research is characterized by a cyclical process in which educational materials for learning environments are designed, implemented, and evaluated for following cycle(s) of (re)design and testing (Van Dijke, Dirjvers and Bakker, 2020; McKenney and Reeves, 2012; Gravemeijer, 2003). This example reports on a first cycle of design research, composed by three main phases (Figure 1): *design phase*; *teaching experiment*; *retrospective analysis* (Bakker, 2018; Gravemeijer, 2004).

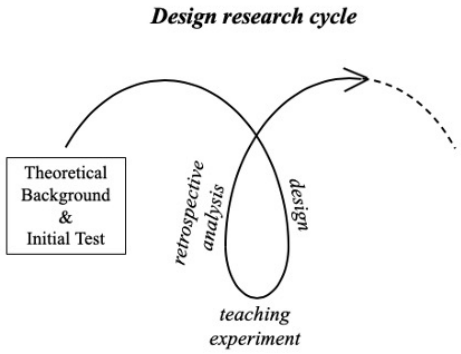


Figure 1. Design research cycle reported on this study.

The study was conducted in a 2nd-grade class (age 7) composed by nineteen students during two weeks of regular mathematics lessons. At the time of the activity, students involved in the study were working on multiplication in the set of natural numbers. In

particular, multiplication as iterated sum (Maffia and Mariotti 2018; Fischbein, Deri, Nello and Marino 1985) was introduced by the official mathematics teacher. Specifically, before the modelling activity students involved in the study did know the notion of multiplication as iterated sum and were able to perform multiplication between one-digit natural numbers and tenths.

In the next the phases of design research for this study are described in detail, focusing on: (i) the design of a Hypothetical Learning Trajectory; (ii) the presentation of the main results from the teaching experiment; (iii) the analysis of the results in relation to our initial hypothesis.

Design phase

The design phase of this research was explicated by the development of the three components of a Hypothetical Learning Trajectory (HLT): learning goal; hypothetical learning process; learning activities (Simon, 1995).

The learning goal is represented by the reinvention of DP.

The design heuristics of guided reinvention, didactical phenomenology and emergent modelling helped in designing a hypothetical learning process together with a set of instructional activities that fitted this learning route (Gravemeijer, 2004). Starting from the classroom initial level and following the heuristic of didactical phenomenology, we supposed that making students face with a problem situation in which the new mathematical concept of DP was needed, could stimulate them in reinventing that concept. In this direction, our hypothesis consisted in putting students face with a problem situation in which at a certain point the necessity of performing multiplications between a number with 2-digits and one with 1-digit emerged. Indeed, consider for example the multiplication 25×4, in which the first term has two digits and the second term has one digit. To perform this multiplication, 25 can be decomposed as $20 + 5$, and then multiply 20×4 and 5×4. In the first multiplication, 20×4, the two 2-digits number 20 is actually 2×10, and as we said before students were able to perform multiplications with tenths. The second multiplication, 5×4, is between only 1-digit numbers. In conclusion, $25 \times 4 = (20 + 5) \times 4 = (20 \times 4) + (5 \times 4)$, that is DP. Which problem situation could provide the development of the hypothesized reinvention process? The context for such problem situation should not only be realistic and rich, i.e. significant for students and mathematics, but also it must stimulate students in developing the DP from their informal solution strategies, in agreement with emergent modelling. In such a reinvention process, also the constraints given in the text of the problem should encourage (or not) the emergence of DP. In order to do that, a specific artifact (Bonotto, 2013) was produced and given to students together with the modelling task.

Since at the time of the activity the school in which the teaching experiment took place was under building renovation, we decided to choose as modelling task the following Tiling Problem:

The Tiling Problem

The school director decided to renovate the school. Students can design a floor tiling of their own classroom. The floor of your classroom was divided in six equal strips. Each group of students should tile a strip, using all the available types of floor tiles.

Figure 2. Modelling task.

Together with the task, to each student was given a booklet which included: the figure of the classroom divided in six stripes; the figure of each stripe that must be tiled by each single group; a brochure, that represents an artifact, including the shapes of the tiles that must be used (triangular, square, rectangular) with their relative costs; the task repeated in a clearer form.

The modelling activity was implemented in three subsequent phases. In the first phase, *getting started*, students were engaged in a series of activities in order to make them familiar with the problem situation they would have to face. The second phase of the modelling activity consisted in the *model construction*. In this phase each group had to solve *The Tiling Problem* and create a poster in which report the designed floor tiling and explain the strategy followed to calculate its total cost. In the final *presentation and discussion* phase, each group had to present to the classroom its project, explaining the steps followed to solve the task.

Teaching experiment: some results

Some results from the second phase of *model construction* are reported. In the *model construction phase* students, working in groups of three, had to create a model to solve the task explored in the previous lesson.

While solving the task, all the groups developed a similar strategy. The first step consisted in counting the number of all the tiles of the same type and multiply the number obtained with the relative cost. For example, one group counted 50 square tiles, 26 triangular tiles and 15 rectangular tiles. Then, the number of each type of tile was multiplied by its relative cost, highlighting the notion of multiplication as iterated sum, already known by students. In our example, students had to perform 50×6, 26×4, 15×10. While performing multiplications similar to the latter ones, students encountered the difficulty of multiplying a number with one digit with a number with two digits. Since several groups were not able to solve this problem, the teacher decided to reason about it in a whole class discussion. Some students suggested the strategy reported in the following dialogue (T=teacher; S1=first student; S2=second student) to calculate 6×57:

S1: *I write* $6 \times 57 = 57 \times 6$.
Then I divide 57 as 50 and 7?
T: *Divide?*
S1: *Write...?*
T: *Decompose.*
S1: *Yes, I decompose 57 as 50 plus 7!*
Then I calculate 50×6.
S2: *That is 300!*
S1: *Then* 6×7
S2: *42*
T: *Excellent, and with these number? (pointing at 300 and 42)*
S1: *I put them together!*
T: *How?*
S1: *I compose them?*
T: *What does it mean?*
S1: *I make the sum!*

The final step developed by students to solve *The Tiling Problem* was to sum the costs of each shape of tiles. In our example, students, having calculated $50 \times 6 = 300$, $26 \times 4 = 104$, $15 \times 10 = 150$, summed $300 + 104 + 150 = 554$, that represented the total cost in euros of their tiling design.

As hypothesized in the HLT, while constructing their models students encountered the problem of calculating multiplications between a number with one digit and a number with two digits, that stimulated some students to reinvent the notion of DP (Figure 3).

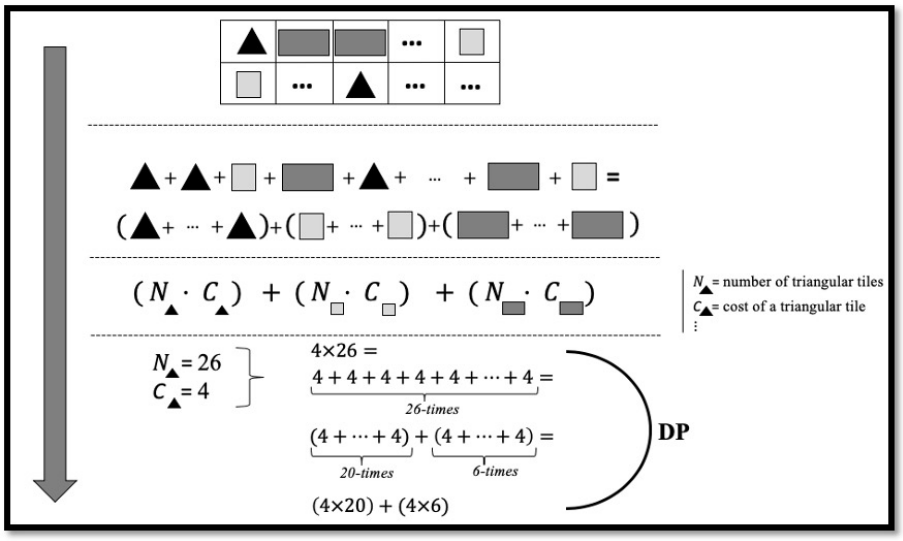


Figure 3. Reconstruction of students' model to solve the modelling task that let them to reinvent DP.

Retrospective analysis

The final phase of the design research cycle comprised the retrospective analysis, in which the HLT developed in the design phase was compared with students' actual learning occurred during the teaching experiment.

In agreement with the process of emergent modelling, the assignment given to students stimulated them to create and work with the new mathematical concept of DP. In the specific, the strategy developed by students to solve the task, that consisted in grouping the tiles with the same shape and then multiply by the associated costs, showed that they have been able to reinvent the mathematical concept of DP (Figure 3). This is evident also from the extract of the dialogue in which students explained their strategy to calculate 6×57. As a consequence, the hypothesis that multiplications between 1-digit and 2-digits numbers should permit students in reinventing the concept of DP was confirmed. Guided by the interaction with the teacher and peers, students have been able to reason and explain this property, that would be at the basis of their following strategies of calculus. In this way, properties of mathematical operations become meaningful for students, because no longer mechanical rules but rooted in their experience and directly constructed by students to solve a concrete problem from a meaningful context.

The reinvention process was possible not only thanks to the designed modelling sequence, but also to the use of a suitable artifact, represented by the brochure given to students. Indeed, having given students the shapes of the tiles to be used and the constraint to use all that shapes, guided them to face with the problem of performing multiplications between numbers with more than one digit, and consequently to the reformulation of DP, as hypothesised in the design phase.

Some considerations about possible modifications of the designed and implemented instructional sequence can be inferred, concerning two main facts: the need of students' self-evaluation and the role of the teacher. Students during the modelling activity worked always in groups. They could share their opinions, strategies, misunderstandings, but there was not enough space for individual reflection. Group work is important, but we believe that also individual moments in which students have time to reflect on the modelling activity, reconstruct the entire process, clarify doubts and express their ideas is fundamental in the learning process. Moreover, to achieve the described results, the role of the teacher was fundamental. The teacher, indeed, encouraged students to use their own methods; stimulated students to articulate and reflect on their personal beliefs, misconceptions and informal problem-solving and modelling strategies. However, the role of the teacher in the design of the learning trajectory was not clearly explicated. Instead, it should be more emphasized and clarified, since it plays a fundamental position in supporting students shift from concrete to symbolic representations of DP (Maffia and Mariotti, 2020). Learning should become a constructed understanding through a continuous interaction between teacher and students, that can be synthetized, using Freudenthal?s words, in teaching and learning as *guided reinvention*, reinforcing in this way mathematical reasoning and sense-making.

Conclusions

The present contribution followed two main directions: the first concerning a brief introduction to mathematics education, with particular attention to the different trends developed in the European contexts (Blum, Artigue, Mariotti, Strasser and Van den Heuvel-Panhuizen, 2019); the second concerning an example of research in mathematics education, that focuses on a modelling activity that aimed to make distributivity property accessible to primary school students (Passarella, 2021).

In the first part a description of the common features shared between the different European traditions in the didactics of mathematics were reported. Specifically, attention was given on the central role of design activities for learning and teaching.

In the second part an example of research in mathematics education was presented. This example followed the methodology of design research, and its aim consisted in designing, implementing, and evaluating a modelling activity in a reinvention process to introduce 2nd-grade students to the distributivity property in a meaningful way.

References

Bakker, A. (2004). "Design research in statistic education: on symbolizing and computer tools", CD-Bèta Press, Utrecht.

Bakker, A. (2018). "Design Research in Education. A practical Guide for Early Career Researchers", Routledge.

Blum, W., Artigue, M., Mariotti, A., Strasser, R., and Van den Heuvel-Panhuizen, M. (2019). *European Didactic Traditions in Mathematics: Introduction and Overview.* In W. Blum et al. (Eds), European Traditions in Didactics of Mathematics (pp. 1-10). New ICMI Studies no. 13. New York: Springer.

Bonotto, C. (2013). *Artifacts as sources for problem-posing activities.* Educational Studies in Mathematics, 83(1), 37–55.

Brousseau, G. (1997). "Theory of didactical situations in mathematics", Dordrecht: Kluwer.

Chamberlin, S., Payne, A. M., and Kettler, T. (2020). *Mathematical modelling: a positive learning approach to facilitate student sense making in mathematics.* International Journal of Mathematical Education in Science and Technology, DOI: 10.1080/0020739X.2020.1788185.

Chevallard, Y., and Sensevy, G. (2014). *Anthropological approaches in mathematics education, French perspectives.* In S. Lerman (Ed.), Encyclopedia of mathematics education (pp. 38–43), New York: Springer.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., and Schauble, L. (2003). *Design experiments in educational research.* Educational Researcher, 32(1), 9–13.

Doorman, M. (2019). *Design and research for developing local instruction theories.* Avances de Investigacion en Educacion Matematica, 15, 29–42.

Edelson, D.C. (2002). *Design Research: What We Learn When We Engage in Design.* Journal of the Learning Sciences, 11, 105–121.

Fischbein, E., Deri, M., Nello, M., and Marino, M. (1985). *The role of implicit models in solving verbal problems in multiplication and division.* Journal for Research in Mathematics Education, 16(1), 3–17.

Freudenthal, H. (1973). "Mathematics as an educational task", Dordrecht, the Netherlands: Reidel.

Freudenthal, H. (1983). "Didactical phenomenology of mathematical structures", Dordrecht, the Netherlands: Reidel.

Freudenthal, H. (1991). "Revisiting mathematics education", China lectures. Dordrecht: Kluwer.

Gravemeijer, K. (1994). "Developing realistic mathematics education", Utrecht: CD Bèta Press.

Gravemeijer, K. (1999). *How emergent models may foster the construction of formal mathematics.* Mathematical Thinking and Learning, 1(2), 155–177.

Gravemeijer, K. (2003). *Supporting Students' Development of Measuring Conceptions: Analyzing Students' Learning in Social Context.* Journal for Research in Mathematics Education, 12, 51–66.

Gravemeijer, K. (2004). *Learning Trajectories and Local Instruction Theories as Means of Support for Teachers in Reform Mathematics Education.* Mathematical Thinking and Learning, 6(2), 105–128.

Gravemeijer, K., Cobb, P., Bowers, J., and Whitenack, J. (2000). *Symbolizing, modeling, and instructional design.* In P. Cobb, E. Yackel and K. McClain (Eds.), Symbolizing and communicating in mathematics classrooms: Perspectives on discourse, tools, and instructional design (pp. 225–273). Mahwah, NJ: Lawrence Erlbaum Associates.

Gravemeijer, K. and Doorman, M. (1999). *Context Problems in Realistic Mathematics Education: A Calculus Course as an Example.* Educational Studies in Mathematics, 39(1-3), 111–129.

Greer, B., Verschaffel, L. and Mukhopadhyay (2007). *Modelling for life: mathematics and children's experience.* In W. Blum et al. (Eds), Modelling and applications in mathematics education (pp. 89-98). New ICMI Studies no. 10. New York: Springer.

Lesh, R., and Lehrer, R. (2003). *Models and Modeling Perspectives on the Development of Students and Teachers.* Mathematical Thinking and Learning, 5(2-3), 109–129.

Maffia, A., and Mariotti, A. (2018). *Intuitive and formal models of whole number multiplication: relations and emerging structures.* For the Learning of Mathematics, 38(3), 30–36.

McKenney, S., and Reeves, T. C. (2012). *Conducting educational design research.* London: Routledge.

Passarella, S. (2021). *Emergent modelling to introduce distributivity property of multiplication: a design research study in primary school.* International Journal of Mathematical Education in Science and Technology.

Simon, M.A. (1995). *Reconstructing mathematics pedagogy from a constructivist perspective.* Journal for Research in Mathematics Education, 26, 114–145.

Streefland, L. (1985). *Wiskunde als activiteit en de realiteit als bron (Mathematics as an activity and the reality as a source).* Tijdschrift voor Nederlands Wiskundeonderwijs (Nieuwe Wiskrant), 5 (1), 60–67.

Treffers, A. (1987). "Three dimensions. A model of goal and theory description in mathematics instruction-the Wiskobas project", Dordrecht: Reidel Publishing.

Van den Heuvel-Panhuizen, M., and Drijvers, P. (2014). *Realistic Mathematics Education.* In S. Lerman (Ed), Encyclopedia of Mathematics Education (pp. 521-525). Dordrecht, Heidelberg, New York, London: Springer.

Van Dijke-Droogers, M., Drijvers, P., and Bakker, A. (2020). *Repeated Sampling with a Black Box to Make Informal Statistical Inference Accessible.* Mathematical Thinking and Learning, 22(2), 116–138.

Homological algebra: deforming abelian groups using torsion pairs

Sergio Pavon (*)

Abstract. We gently introduce the concept of abelian categories, motivated by the example of the category of abelian groups, and we show that they provide all the ingredients needed for (co)homological theories. This direction leads us to the definition of the derived category of an abelian category, which is "the correct place to compute cohomologies". In the second part, we address a natural question: given two abelian categories, when do they have the same derived category (i.e., they are derived equivalent)? In particular we explore the case in which one of the abelian categories has been obtained from the other via a deformation procedure called "HRS-tilting". We give a direct criterion to check for a derived equivalence in some instances of this procedure, and a more explicit reformulation of it for categories of modules over a ring.

1 Abelian categories and their derived categories

1.1 Categories

Recall that a *category* is the datum of two main ingredients: a class of *objects*, and for every pair of objects, a set of *morphisms* from one to the other. The set of morphisms from an object A to an object B of a category $\mathcal{C}$ is denoted by $\mathsf{Hom}_{\mathcal{C}}(A, B)$; an element of this set is also denoted with an arrow $f\colon A \to B$. The definition of a category also requires a notion of *composition of morphisms* to be defined: given morphisms $f\colon A \to B$ and $g\colon B \to C$ there must be a morphism $g \circ f\colon A \to C$, satisfying some natural properties.

A motivating example could be the category Sets of sets, whose objects are sets and whose morphisms are functions; the composition of morphism is the usual composition of functions. Other examples, similar in spirit, are the categories Top of topological spaces with continuous functions, or the category Vec_k of k-vector spaces with k-linear applications, for a field k.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: sergio.pavon@math.unipd.it . Seminar held on 18 November 2020.

1.2 Abelian groups

We now introduce our main example, the *category of abelian groups*, denoted Ab. Its objects are the abelian groups, and its morphisms are the homomorphisms of groups.

Example 1 Some objects of Ab are

$$(\mathbb{Z},+),(\mathbb{Q},+),(\mathbb{R},+),(\mathbb{C}^*,\cdot),(\mathbb{S}^1 := \{z\in\mathbb{C}\colon |z|=1\},\cdot),\ \mathbb{Z}/n\mathbb{Z},\ \mathbb{Q}/\mathbb{Z},\ \mathbb{R}/\mathbb{Q}.$$

Some morphisms of Ab are

$$\begin{aligned}\mathrm{id}_A &\colon A\to A,\ a\mapsto a \quad \text{for every group } A\\ \pi_n &\colon \mathbb{Z}\to\mathbb{Z}/n\mathbb{Z},\ z\mapsto z \pmod n\\ \exp &\colon \mathbb{R}\to\mathbb{S}^1,\ t\mapsto e^{2\pi i t}.\end{aligned}$$

The category Ab enjoys some special properties. For example, notice that for every two objects A, B of Ab, the set $\mathsf{Hom}_{\mathsf{Ab}}(A,B)$ is not only a set, but an abelian group itself. Indeed, it has an addition operation defined by

$$\forall f,g\colon A\to B \qquad (f+g)\colon A\to B,\ (f+g)(a)=f(a)+_B g(a)$$

This makes $\mathsf{Hom}_{\mathsf{Ab}}(A,B)$ into an abelian group, with neutral element the *zero morphism* from A to B, namely $0\colon A\to B,\ a\mapsto 0_B$. Moreover, composition of morphism respects this operation, meaning that it is bilinear.

It is not common for a category to have Hom-*groups*, instead of just Hom-sets. Such a category is called *preadditive*. For a counterexample, notice that Sets is not preadditive, as there is not a good notion of "sum" of two functions between arbitrary sets. Conversely, Vec_k is another preadditive category.

We now give a list of some useful special properties enjoyed by Ab.

(a) as we said, it is preadditive

(b) it has a *zero object*, which in this case is the zero group

(c) it has the direct sum of any two objects; in this case, the usual direct sum of abelian groups

(d) all morphisms have a *kernel* and a *cokernel*; in this case,

$$\text{for } f\colon A\to B,\quad \ker f=\{a\in A\colon f(a)=0_B\} \quad\text{and}\quad \operatorname{coker} f = B/\operatorname{im} f$$

(e) **1st Isomorphism Theorem**: Any morphism $f\colon A\to B$ induces an isomorphism

$$\bar f\colon A/\ker f\to \operatorname{im} f.$$

(more abstractly, an isomorphism $\bar f\colon \operatorname{coker}\ker f\to\ker\operatorname{coker} f$).

1.3 Abelian categories

We then give the following definition.

Definition 2 A category $\mathcal{C}$ is called an *abelian category* if it satisfies the properties (1-5) above.

Example 3 Since abelian categories are modeled after Ab, clearly Ab is an abelian category. Other examples are

(a) The category ab of finitely generated abelian groups with group homomorphisms.

(b) Vec_k

(c) The category $\mathsf{Mod}(R)$ of (right) modules of any ring R, with R-linear maps.

(d) The category $\mathsf{Coh}(X)$ of coherent sheaves over a scheme X, with morphisms of sheaves.

In the second part we will meet some other abelian categories, possibly much different from these.

As counterexamples, the categories Sets and Top mentioned above are not abelian (not even preadditive). To produce some subtler counterxamples, a good direction could be to choose a subcategory of an abelian category (which will then be automatically preadditive) so that some properties of abelian categories fail to hold (e.g., some kernels don't exist, because we didn't include them).

1.4 Sequences

One of the reasons abelian categories are useful is that they provide all the tools needed to give the following definitions.

Definition 4 Let $\mathcal{A}$ be an abelian category, and let $a, b \in \mathbb{Z} \cup \{\pm\infty\}$, with $a < b$. A family of objects $\{A_i \in \mathcal{A} \colon a < i < b\}$ together with morphisms $\{\phi_i \colon A_i \to A_{i+1} \colon a < i < b-1\}$ is a *sequence in* $\mathcal{A}$ if $\phi_{i+1} \circ \phi_i = 0$ for every $a < i < b-2$.

A sequence in $\mathcal{A}$ is commonly depicted as a chain of arrows,

$$\cdots \xrightarrow{\phi_{i-2}} A_{i-1} \xrightarrow{\phi_{i-1}} A_i \xrightarrow{\phi_i} A_{i+1} \xrightarrow{\phi_{i+1}} \cdots$$

which may or may not go on indefinitely on either side, and such that the composition of consecutive arrow vanishes. Clearly in order for this definition to make sense there must be a notion of "zero morphism", hence we need a preadditive category.

Notice that to ask $\phi_{i+1} \circ \phi_i = 0$ is the same as requiring that $\operatorname{im} \phi_i \subseteq \ker \phi_{i+1}$. We may therefore give the following definition.

Definition 5 Let $\mathcal{A}$ be an abelian category and $A_\bullet = \{A_i, \phi_i \colon a < i < b\}$ be a sequence in $\mathcal{A}$. For $a+1 < i < b-1$, the *cohomology in degree i of $A_\bullet$* (or its *i-th cohomology*) is the quotient $H^i(A_\bullet) := \ker \phi_{i+1} / \operatorname{im} \phi_i$.

The reason for the name "cohomologies" will come in a few paragraphs. Clearly cohomologies only make sense in degrees where a morphism ends and another morphism starts, and that is why we exclude the possible extremal indices in the definition. Notice also that to define cohomologies we are using some more properties of abelian categories, namely the existence and properties of kernels and cokernels.

Example 6 Consider the chain of morphisms

$$A_\bullet : \qquad 0 \longrightarrow \mathbb{Z} \xrightarrow{\ i\ } \mathbb{Q} \xrightarrow{\exp_{|\mathbb{Q}}} \mathbb{S}^1 \longrightarrow 0$$

where i is the natural inclusion and $\exp_{|\mathbb{Q}}$ is the restriction of the exponential map introduced before. First of all, it is indeed a sequence, as the three possible composition of consecutive morphisms all vanish:

- $0 \longrightarrow \mathbb{Z} \longrightarrow \mathbb{Q}$ vanishes because its domain is the zero object;
- $\mathbb{Z} \longrightarrow \mathbb{Q} \longrightarrow \mathbb{S}^1$ vanishes because $\forall z \in \mathbb{Z} \quad \exp(z) = e^{2\pi i z} = 1$, which is the neutral element of $\mathbb{S}^1$;
- $\mathbb{Q} \longrightarrow \mathbb{S}^1 \longrightarrow 0$ vanishes because its codomain is the zero object.

We can therefore compute its cohomologies. If we number the degrees from left to right starting from -1, we have

- $H^0(A_\bullet) = \ker i / \operatorname{im} 0 = 0/0 = 0$;
- $H^1(A_\bullet) = \ker \exp / \operatorname{im} i = \mathbb{Z}/\mathbb{Z} = 0$;
- $H^2(A_\bullet) = \ker 0 / \operatorname{im} \exp = \mathbb{S}^1 / \operatorname{im} \exp \simeq (\mathbb{R}/\mathbb{Z})/(\mathbb{Q}/\mathbb{Z}) \simeq \mathbb{R}/\mathbb{Q}$.

1.5 Exact sequences

Definition 7 Let $\mathcal{A}$ be an abelian category and $A_\bullet$ a sequence in $\mathcal{A}$. If for some degree i we have $H^i(A_\bullet) = 0$, the sequence is said to be *exact in degree i*. A sequence which is exact in every degree is said to be *exact*.

Exact sequences prove to be a useful notational tool, to express relations between objects of the category. We give the following examples.

Example 8 Let $f\colon A \to B$ be a morphism in $\mathcal{A}$. Then the sequence

$$0 \longrightarrow A \xrightarrow{\ f\ } B$$

is exact if and only if f is a monomorphism (i.e. an injective homomorphism of groups, if we take $\mathcal{A} = \mathsf{Ab}$). Indeed, the only cohomology which is defined is in the middle degree: and there we have $H = \ker f / \operatorname{im} 0 = \ker f$, which is zero if and only if f is a monomorphism.

Example 9 Dually, let $g\colon B \to C$ be a morphism in $\mathcal{A}$. Then the sequence

$$B \xrightarrow{g} C \longrightarrow 0$$

is exact if and only if g is an epimorphism (i.e. a surjective homomorphism of groups, if $\mathcal{A} = \mathsf{Ab}$).

Example 10 Glueing the two examples above, we obtain a powerful notational tool: a sequence

$$0 \longrightarrow A \xrightarrow{f} B \xrightarrow{g} C \longrightarrow 0$$

is exact (and therefore called a *short exact sequence*) if and only f is a monomorphism, g is an epimorphism and $C \simeq B/\operatorname{im} f$. Since we can identify $\operatorname{im} f$ with A via the 1st Isomorphism Theorem, such an exact sequence tells us that B is an *extension* of C with A, which just means that $C = B/A$.

For example, some short exact sequences are

$$0 \longrightarrow \mathbb{Z} \xrightarrow{i} \mathbb{Q} \longrightarrow \mathbb{Q}/\mathbb{Z} \longrightarrow 0$$

$$0 \longrightarrow \mathbb{Z} \xrightarrow{\cdot 2} \mathbb{Z} \longrightarrow \mathbb{Z}/2\mathbb{Z} \longrightarrow 0$$

1.6 (Co)homological Theories and complexes

We now study another use of the sequences we have described so far, which is pervasive in modern mathematics.

When studying an object of some complicated type (e.g., a topological space, a smooth real manifold, an algebraic variety and so on) it can be useful to attach to it some algebraic data, in order to exploit the tools of (often linear) algebra. This is the idea behind all the (co)homological theories found in many fields of mathematics. The "algebraic data" mentioned above take the form of a sequence of objects in some abelian category, as we defined. Common choices for this abelian category are Vec_k and Ab.

Example 11 (De Rham Cohomology) Let M be a smooth real manifold. For every natural number $n \geq 0$, one can construct the real vector space $\Omega^n(M)$ of real n-forms on M; for example, $\Omega^0(M)$ is the space of smooth real-valued functions on M, and so on. For every $n \geq 0$, there is also the *exterior derivative morphism*, i.e. a $\mathbb{R}$-linear application $d\colon \Omega^n(M) \to \Omega^{n+1}(M)$, which sends an n-form to its differential. This morphisms have the property that the composition of two consecutive ones of them is the zero morphism, which is often written $d^2 = 0$. This allows us to associate to M a sequence in $\mathsf{Vec}_{\mathbb{R}}$

$$\Omega^\bullet(M): \qquad 0 \longrightarrow \Omega^0(M) \xrightarrow{d} \Omega^1(M) \xrightarrow{d} \Omega^2(M) \xrightarrow{d} \cdots$$

The cohomologies $H^i_{\mathrm{dR}}(M) := H^i(\Omega^\bullet(M))$ of this sequence, called the *de Rham cohomologies of* M, give some information on the original manifold M. For example, if we number

the degrees following the exponent of Ω, we have that

$$H^0_{\mathrm{dR}}(M) = H^0(\Omega^\bullet(M)) = \ker d = \{\text{locally constant functions}\}$$
$$H^1_{\mathrm{dR}}(M) = H^1(\Omega^\bullet(M)) = \ker d/\operatorname{im} d = \{\text{closed 1-forms}\}/\{\text{exact 1-forms}\}$$

So the dimension of $H^0_{\mathrm{dR}}(M)$ as a real vector space counts the number of connected components of M; while $H^1_{\mathrm{dR}}(M)$ measures how far is the fundamental theorem of calculus from holding on M.

To formalise the example above, we give the following definition.

Definition 12 Let $\mathcal{A}$ be an abelian category. A *(cochain) complex* of objects of $\mathcal{A}$ is just a (unbounded) sequence $A^\bullet$, in the sense above, regarded as a single object. The morphisms of $\mathcal{A}$ which appear in $A^\bullet$ are called the *differentials of* $A^\bullet$. Given two complexes $A^\bullet, B^\bullet$, we define a *morphism of complexes* between them as a family of morphisms of $\mathcal{A}$, $f = \{f_i \colon A_i \to B_i \colon i \in \mathbb{Z}\}$ such that the following diagram commutes:

$$\begin{array}{ccccccccccc}
A^\bullet & \cdots & \xrightarrow{d^A_{i-2}} & A_{i-1} & \xrightarrow{d^A_{i-1}} & A_i & \xrightarrow{d^A_i} & A_{i+1} & \xrightarrow{d^A_{i+1}} & \cdots \\
\downarrow f & & & \downarrow f_{i-1} & & \downarrow f_i & & \downarrow f_{i+1} & & \\
B^\bullet & \cdots & \xrightarrow{d^B_{i-2}} & B_{i-1} & \xrightarrow{d^B_{i-1}} & B_i & \xrightarrow{d^B_i} & B_{i+1} & \xrightarrow{d^B_{i+1}} & \cdots
\end{array}$$

We define the *category of complexes over* $\mathcal{A}$, denoted $\mathsf{C}(\mathcal{A})$, as the category with objects the complexes of objects of $\mathcal{A}$ and with morphisms the morphisms of complexes.

Example 13 As we saw, the de Rham cohomology theory associates to every smooth manifold M a complex $\Omega^\bullet(M)$ of real vector spaces. In addition, it also associates to every smooth function $\varphi \colon M \to N$ between smooth manifolds a morphism of complexes $\varphi^* \colon \Omega^\bullet(N) \to \Omega^\bullet(M)$, defined as follows: for an n-form ω on N, $(\varphi^*)^n(\omega) = \varphi^*\omega$ is the n-form on M obtained as the pullback of ω along φ. In fact, this assignment defines a *functor* from the category of smooth manifolds with smooth functions to the category $\mathsf{C}(\mathsf{Vec}_{\mathbb{R}})$. In essence, this is what all (co)homological theories are: functors between some category to a category of complexes. By the way, notice that in the case of de Rham cohomology, a smooth function $M \to N$ maps to a morphism of complexes $\Omega^\bullet(N) \to \Omega^\bullet(M)$, not the other way around. This means that the functor is *contravariant*, and this is the case for *cohomological* theories. If instead the functor is *covariant*, i.e. it doesn't exchange the direction of morphisms, then the theory is called *homological*.

1.7 Quasi-isomorphisms and the derived category of $\mathcal{A}$

From the point of view of (co)homological theories, the important part of a complex are its cohomologies. Luckily, a morphism of complexes $f \colon A^\bullet \to B^\bullet$ always induces morphisms between the cohomologies, $H^i(f) \colon H^i(A^\bullet) \to H^i(B^\bullet)$, for all $i \in \mathbb{Z}$ in a natural way. (Incidentally, this means that $H^i \colon \mathsf{C}(\mathcal{A}) \to \mathcal{A}$ is a functor, for all $i \in \mathbb{Z}$; make sure also not to confuse $H^i(f)$ with the vertical morphism $f_i \colon A_i \to B_i$). Therefore, morphisms of

complexes can be used to compare the cohomologies of two complexes. In particular, it can happen that a particular morphism of complexes witnesses the fact that its domain and codomain complexes have "the same cohomologies":

Definition 14 Let $\mathcal{A}$ be an abelian category, $A^\bullet, B^\bullet \in \mathsf{C}(\mathcal{A})$ complexes and $f\colon A^\bullet \to B^\bullet$ a morphism of complexes. If for every $i \in \mathbb{Z}$ the induced morphism $H^i(f)\colon H^i(A^\bullet) \to H^i(B^\bullet)$ is an isomorphism, then f is called a *quasi-isomorphism*.

We immediately clear the field from a possible doubt.

Remark 15 Quasi-isomorphisms are not isomorphisms! A counterexample is the following, which obviously is not an isomorphism:

$$\begin{array}{ccccccccccccc} Z^\bullet & & \cdots & \longrightarrow & 0 & \longrightarrow & \mathbb{Z} & \xrightarrow{\mathrm{id}_\mathbb{Z}} & \mathbb{Z} & \longrightarrow & 0 & \longrightarrow & \cdots \\ \downarrow{\scriptstyle 0} & & & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \\ 0 & & \cdots & \longrightarrow & 0 & \longrightarrow & 0 & \longrightarrow & 0 & \longrightarrow & 0 & \longrightarrow & \cdots \end{array}$$

All the cohomologies of the complex $Z^\bullet$ are zero, and so are clearly the cohomologies of the zero complex. The induced maps on the level of the cohomologies are therefore $H^i(f) = \mathrm{id}_0\colon 0 \to 0$, isomorphisms for every $i \in \mathbb{Z}$, so f is a quasi-isomorphism.

This phenomenon is a bit disappointing, as it means that two complexes may have "the same cohomologies" while being essentially different from each other; and this doesn't play well with our idea of complexes as mere "containers" for their cohomologies. The solution to this problem is the following.

Construction 16 The reason why a quasi-isomorphism may not be an isomorphism is that it may lack an inverse morphism of complexes. This is morally the same reason why $\mathbb{Z}$ is not a field, as some elements lack a multiplicative inverse. The solution in that case is to add formal elements "$1/z$", for $z \in \mathbb{Z}^*$, with the property that they are inverses of the corresponding z. We follow the same path, and for every quasi-isomorphism of complexes $s\colon A^\bullet \to B^\bullet$ we introduce a formal new morphism $s^{-1}\colon B^\bullet \to A^\bullet$ such that $s^{-1} \circ s = \mathrm{id}_{A^\bullet}$ and $s \circ s^{-1} = \mathrm{id}_{B^\bullet}$. "Formal" means that there is no hope that s^{-1} will have a nice description as a diagram with vertical arrows, as usual morphisms of complexes do. This precedure, which is called *localisation*, can be made precise, with a bit of care (see e.g. [6]).

Definition 17 Let $\mathcal{A}$ be an abelian category. Under some assumptions on $\mathcal{A}$ (eg. if $\mathcal{A}$ has enough injectives or projectives) there exists a category $\mathsf{D}(\mathcal{A})$, called the *derived category of* $\mathcal{A}$, with objects the complexes of objects of $\mathcal{A}$ and as morphisms the usual morphisms of complexes, the formal inverses of quasi-isomorphisms and all compositions of these.

The hypothesis of this definition hold for a very general family of abelian categories. In particular, this is the case for our familiar categories Ab and Vec_k.

The derived category $\mathsf{D}(\mathcal{A})$ is very useful from many points of view. For one, it is a more natural place than $\mathsf{C}(\mathcal{A})$ to work with cohomologies of objects: indeed, here complexes "with the same cohomologies" are "the same". Even more, $\mathsf{D}(\mathcal{A})$ is not an abelian category

itself, but it still has some additional structure, namely it is a *triangulated* category. Inside it, cohomologies assume an even deeper meaning through the notion of *t-structures*, which we will not explore.

The derived category of an abelian category is also interesting for another reason. Indeed, some properties of an abelian category $\mathcal{A}$ can be read off its derived category $\mathsf{D}(\mathcal{A})$, making them *derived invariants*. As a consequence, if two abelian category have "the same" derived category (meaning that they are *derived equivalent*), then they must share these invariants. These invariants are of great interest both in examples coming from algebra (e.g. from the representation theory of algebras) and in those coming from algebraic geometry. Moreover, sometimes these two fields come in contact when an abelian category of algebraic origin is derived equivalent to one of geometric origin (as it is the case for $\mathsf{Rep}_k(\ \bullet \rightrightarrows \bullet\ )$ and $\mathsf{coh}(\mathbb{P}^1(k))$, for an algebraically closed field k).

2 Torsion pairs and HRS-tilting

In this second section, we are going to present a way to construct a new abelian category by "deforming" a given one. Later we will apply this procedure starting with Ab, and address the question of whether the category so obtain is derived equivalent to Ab.

2.1 Torsion pairs in abelian categories

The main tool in this "deformation" procedure are *torsion pairs*. We start with the first historical example, as a motivation.

Example 18 In Ab, let A be an abelian group. Recall that an element $a \in A$ is a *torsion element* if there exists a positive integer $n \geq 1$ such that $na = a + \cdots + a = 0$, and a *torsion-free element* otherwise. The group A is called *torsion* if all its elements are torsion elements; it is called *torsion-free* if its only torsion element is 0. For example, $\mathbb{Z}/n\mathbb{Z}$ is a torsion group for every $n \geq 1$; while $\mathbb{Z}$ is a torsion-free group. Notice that not every group is either torsion or torsion-free: an example is the circle $\mathbb{S}^1$. If we define $\mathcal{T}_{\mathsf{can}} := \{\text{torsion groups}\}$ and $\mathcal{F}_{\mathsf{can}} := \{\text{torsion-free groups}\}$, we can verify the following two properties:

- $\mathsf{Hom}_{\mathsf{Ab}}(T, F) = 0$ for every $T \in \mathcal{T}_{\mathsf{can}}, F \in \mathcal{F}_{\mathsf{can}}$;
- For every abelian group A there is a short exact sequence

$$0 \to T \to A \to F \to 0 \qquad \text{with } T \in \mathcal{T}_{\mathsf{can}}, F \in \mathcal{F}_{\mathsf{can}}.$$

The first property is easy to prove: let $f\colon T \to F$, and $t \in T$. Then there exists $n \geq 1$ such that $nt = 0$: and we have $nf(t) = f(nt) = f(0) = 0$, so $f(t)$ is a torsion element of F. By hypothesis, it is 0, so f is the zero morphism.

The proof of the second property is sketched with an example. Let $A = \mathbb{S}^1 \simeq \mathbb{R}/\mathbb{Z}$, and set T to be the set of torsion elements of A. It is easy to see that it is a subgroup, and in this case it is $T \simeq \mathbb{Q}/\mathbb{Z}$ (complex numbers with argument a rational fraction of a full

2π). The quotient A/T, which in this case is $(\mathbb{R}/\mathbb{Z})/(\mathbb{Q}/\mathbb{Z}) \simeq \mathbb{R}/\mathbb{Q}$, is always a torsion-free group, which plays then the role of F.

This example motivates the following definition.

Definition 19 Let $\mathcal{A}$ be an abelian category. A pair of classes $(\mathcal{T}, \mathcal{F})$ of objects of $\mathcal{A}$ is called a *torsion pair* in $\mathcal{A}$, if

- $\mathsf{Hom}_{\mathcal{A}}(T, F) = 0$ for every $T \in \mathcal{T}, F \in \mathcal{F}$;
- For every object A of $\mathcal{A}$ there is a short exact sequence

$$0 \to T \to A \to F \to 0 \qquad \text{with } T \in \mathcal{T}, F \in \mathcal{F}.$$

$\mathcal{T}$ is called the *torsion class*, and its objects the *torsion objects* (with respect to this torsion pair). Similarly, the class $\mathcal{F}$ and its objects are called *torsion-free*.

A torsion pair is a sort of "orthogonal decomposition" of the category $\mathcal{A}$.

Example 20 Some examples of torsion pairs in Ab:

- $(\mathcal{T}_{\mathsf{can}}, \mathcal{F}_{\mathsf{can}})$, the *canonical torsion pair* introduced above;
- $(\mathsf{Ab}, 0)$ and $(0, \mathsf{Ab})$, the *trivial torsion pairs*;
- $(\mathcal{D}, \mathcal{R})$, where $\mathcal{D} := \{\text{divisible groups}\}$ and $\mathcal{R} := \{\text{reduced groups}\}$. A group D is called *divisible* if for every $x \in D$ and $n \geq 1$ there exists $y \in D$ such that $x = ny$ (and so x can be "divided by n"); for example, $\mathbb{Q}$ is divisible. On the other hand, a group R is *reduced* if the only element which is divisible by every positive integer is 0: i.e. if $\bigcap_{n\geq 1} nR = 0$. For example, $\mathbb{Z}$ is reduced.

Notice that the first and the last of the examples above differ substantially: for the canonical torsion pair, "torsion" is a property defined by elements, while for $(\mathcal{D}, \mathcal{R})$ this is not the case. This can be made more precise by noticing that while for $(\mathcal{T}_{\mathsf{can}}, \mathcal{F}_{\mathsf{can}})$ *subgroups of torsion groups are torsion groups*, for $(\mathcal{D}, \mathcal{R})$ this is not the case, as shown by the counterexample $\mathcal{R} \ni \mathbb{Z} \leq \mathbb{Q} \in \mathcal{D}$. Therefore, we define:

Definition 21 A torsion pair $(\mathcal{T}, \mathcal{F})$ in an abelian category $\mathcal{A}$ is called *hereditary* if $\mathcal{T}$ is closed under subobjects (i.e., subobjects of torsion objects are torsion).

2.2 HRS-tilting

Notice that the definition of a torsion pair is asymmetrical; namely, if $(\mathcal{T}, \mathcal{F})$ is a torsion pair in $\mathcal{A}$, usually $(\mathcal{F}, \mathcal{T})$ is not a torsion pair. A procedure exists to "swap" the torsion and torsion-free classes: but in doing so, the abelian category $\mathcal{A}$ gets deformed into a new abelian category, called its *HRS-tilt* with respect to the torsion pair.

Theorem 22 [2, Proposition 2.1] *Let $\mathcal{A}$ be an abelian category admitting a derived category, and let $\mathbf{t} = (\mathcal{T}, \mathcal{F})$ be a torsion pair in $\mathcal{A}$. Then there exists an abelian category $\mathcal{H}_{\mathbf{t}}$ with*

- *as objects, the complexes of objects of $\mathcal{A}$ of the form*

$$A^\bullet : \qquad \cdots \to 0 \to 0 \to A^{-1} \to A^0 \to 0 \to 0 \to \cdots$$

  such that $H^{-1}(A^\bullet) \in \mathcal{F}$ and $H^0(A^\bullet) \in \mathcal{T}$;

- *as morphisms between two such objects $A^\bullet, B^\bullet$, the morphisms between them as objects of the derived category $\mathsf{D}(\mathcal{A})$.*

In fact, this new category $\mathcal{H}_\mathbf{t}$ is abelian for a non-trivial reason: it is the heart of a t-structure in $\mathsf{D}(\mathcal{A})$, as proven in the referenced proposition. But we will not go further in this direction.

This construction of $\mathcal{H}_\mathbf{t}$ fulfills its promise to swap $\mathcal{T}$ and $\mathcal{F}$. Indeed, notice that complexes of the form

$$\begin{aligned} F[1] = &\qquad \cdots \to 0 \to 0 \to F \to 0 \to 0 \to 0 \to \cdots, \text{ for } F \in \mathcal{F} \\ T[0] = &\qquad \cdots \to 0 \to 0 \to 0 \to T \to 0 \to 0 \to \cdots, \text{ for } T \in \mathcal{T} \end{aligned}$$

have the required form, so they are objects of $\mathcal{H}_\mathbf{t}$. The same reason mentioned above, involving t-structures, assures that $(\mathcal{F}[1], \mathcal{T}[0])$ is a torsion pair in $\mathcal{H}_\mathbf{t}$.

2.3 HRS-tilting and derived equivalences

The natural question arises of whether this abelian category $\mathcal{H}_\mathbf{t}$ is derived equivalent to $\mathcal{A}$. This was addressed in a paper by Chen, Han and Zhou [1], were they gave a necessary and sufficient criterion. Its application, however, in general is not so straightforward, for some technical reasons. In [4], we specialised it in to ways:

- instead of an arbitrary abelian category, we start with a Grothendieck category, i.e. a cocomplete abelian category in which filtered colimits are exacts, and which has a generator. This class of abelian categories includes many of the examples coming from algebra and geometry, namely all categories of modules over a ring (e.g. Ab) and all categories of coherent sheaves.

- instead of arbitrary torsion pairs, we focus on hereditary ones.

The result is the following.

Proposition 23 *Let $\mathcal{G}$ be a Grothendieck category, and $\mathbf{t} = (\mathcal{T}, \mathcal{F})$ be a hereditary torsion pair in $\mathcal{G}$. Then the HRS-tilt $\mathcal{H}_\mathbf{t}$ is derived equivalent to $\mathcal{G}$ if and only if there exists an exact sequence*

$$F \to G \to T \to 0$$

where G denotes a generator of $\mathcal{G}$, $T \in \mathcal{T}$ and $F \in \mathcal{F}$.

This criterion is quite direct, since it only amounts to finding a morphism from a torsion-free object F to G, such that its cokernel is torsion. However, we were able to also identify which object $F \in \mathcal{F}$ and which morphism $F \to G$ should appear in the sequence

above, if it exists. Therefore the criterion boils down to checking whether the cokernel of this morphism is torsion or not.

Theorem 24 [4, Theorem 5.6] *Let $\mathcal{G}, G, \mathbf{t}$ as before. Then $\mathcal{H}_\mathbf{t}$ is derived equivalent to $\mathcal{G}$ if and only if the object $G/\mathsf{tr}_{fG}(G)$ belongs to $\mathcal{T}$, where*

- *fG is the torsion-free part of G;*
- *$\mathsf{tr}_{fG}(G)$ is the* trace *of fG in G: namely the image of the canonical morphism*

$$\pi\colon fG^{(\mathsf{Hom}(fG,G))} \to G, \qquad (x_\varphi)_{\varphi\colon fG\to G} \mapsto \sum_{\varphi\colon fG\to G} \varphi(x_\varphi)$$

Sketch of proof. The idea of this proof is to show that π has "the smallest possible cokernel" among all morphisms $F \to G$.

In detail, one sees that the doman of π is torsion-free, and so π is a suitable morphism to play the role of $F \to G$ in the Proposition above. Therefore, if its cokernel $G/\mathsf{tr}_{fG}(G)$ is torsion, the Proposition guarantees a derived equivalence. Conversely, one can prove that whenever a sequence as in the Proposition exists, i.e. a morphism $F \to G$ with torsion cokernel T, then there exists an epimorphism $T \to G/\mathsf{tr}_{fG}(G)$; and so the latter object is torsion as well. □

This result can be translated in more concrete terms when $\mathcal{G}$ is the category $\mathsf{Mod}(R)$ of (say, right) modules over a ring R (which is a Grothendieck category). In this case, the ring R itself is a generator, so we may take $G = R$. As a consequence, also $\mathsf{tr}_{fR}(R)$ has a nice description:

Lemma 25 *Let R be a ring, $(\mathcal{T}, \mathcal{F})$ a torsion pair in $\mathsf{Mod}(R)$, and let tR be the torsion part (the torsion* ideal*) of R. Then $\mathsf{tr}_{fR}(R) = \mathsf{Ann}({}_R tR)$, where*

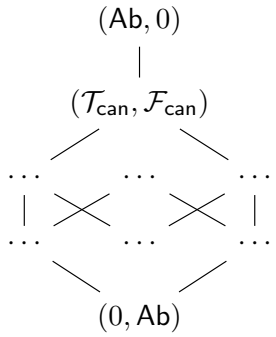
$$\mathsf{Ann}({}_R tR) := \{r \in R\colon rx = 0 \text{ for all } x \in tR\}.$$

In light of this fact, we conclude with a corollary of the Theorem above.

Proposition 26 [4, Corollary 5.11] *Let R be a commutative noetherian ring (for example, $R = \mathbb{Z}$, in which case $\mathsf{Mod}(R) = \mathsf{Ab}$). Then, for every hereditary torsion pair $\mathbf{t}$ in $\mathsf{Mod}(R)$, the HRS-tilt $\mathcal{H}_\mathbf{t}$ is derived equivalent to $\mathsf{Mod}(R)$.*

Proof. Use the notation of the Lemma above. Since R is commutative, for every $r \in R$ and $x \in t(R)$ we have $rx = xr$, in the definition of $\mathsf{Ann}({}_R tR)$. Moreover, since R is noetherian, tR is finitely generated, say $tR = (x_1, \dots, x_n)$. Then we can construct a morphism $R \to (tR)^n$ by $r \mapsto (x_1 r, \dots, x_n r)$, and it is clear that the kernel of this morphism is precisely $\{r \in R\colon xr = 0 \text{ for all } x \in tR\} = \mathsf{Ann}({}_R tR)$. Therefore, it factors through a monomorphism $R/\mathsf{Ann}({}_R tR) \hookrightarrow (tR)^n$. Now, the latter is a torsion module and the torsion pair is hereditary, so we conclude that $R/\mathsf{Ann}({}_R tR)$ is torsion as well. □

To see this Proposition in action, we may choose $R = \mathbb{Z}$, so that $\mathsf{Mod}(R) = \mathsf{Ab}$. In this category, hereditary torsion pairs, when ordered by inclusion of their torsion classes, form a complete lattice, which is isomorphic to the lattice of certain sets of primes (*specialisation closed* sets). This is a partial picture of it:



All of them produce a different abelian category via HRS-tilting; and still, all of these categories are derived equivalent to Ab.

2.4 Further readings

A standard introduction to category theory, from one of the two founders of the field, is [3], which is rich in examples from various parts of mathematics. More specifically on homological algebra, [6] explains in detail the constructions of the first part of this report and various (co)homological theories. The theory of hereditary torsion pairs, and in particular over a commutative noetherian ring, is exposed in [5] (among many other interesting things).

References

[1] X.-W. Chen, Z. Han and Y. Zhou, "Derived equivalences via HRS-tilts". Adv. Math. 354, 2019.

[2] D. Happel, I. Reiten, S. Smalø, "Tilting in abelian categories and quasitilted algebras". Mem. Amer. Math. Soc. 120, 1996.

[3] S. Mac Lane, "Categories for the Working Mathematician". Graduate Texts in Mathematics. Springer, New York NY, 1998.

[4] S. Pavon and J. Vitória, *Hearts for commutative noetherian rings: torsion pairs and derived equivalences*. arXiv:2009.08763.

[5] B. Stenström, "Rings of quotients". Die Grundlehren der Mathematischen Wissenschaften, Band 217. An introduction to methods of ring theory. Springer-Verlag, New York-Heidelberg, 1975.

[6] C. Weibel, "An introduction to homological algebra". Cambridge Studies in Advanced Mathematics, 38. Cambridge University Press, Cambridge, 1994.

Mathematical modeling in Finance

Andrea Mazzoran (*)

1 Introduction

The main topic of these notes consists in studying theoretical pricing models for those financial assets which are known as financial derivatives. Before we give the formal definition of the concept of a financial derivative we will, however, by means of a concrete example, introduce the single most important example: the European call option.

1.1 Problem Formulation

Let us thus consider a Swedish company, that we call *SWE*, which today (denoted by $t = 0$) has signed a contract with an American counterpart called *UNITED*. The contract stipulates that *UNITED* will deliver 1000 computer games to *SWE* exactly six months from now (denoted by $t = T$). Furthermore it is stipulated that *SWE* will pay 1000 US dollars per game to *UNITED* at the time of delivery (i.e. at time $t = T$). For the sake of the argument we assume that the present spot currency rate between the Swedish krona (SEK) and the US dollar is 8.00 SEK/$. One of the problems with this contract from the point of view of *SWE* is that it involves a considerable currency risk. Since *SWE* does not know the currency rate prevailing six months from now, this means that it does not know how many SEK it will have to pay at $t = T$. If the currency rate at $t = T$ is still 8.00 SEK/$ it will have to pay $8,000,000$ SEK, but if the rate rises to, say, 8.50 it will face a cost of $8,500,000$ SEK. For this reason, *SWE* has to face the problem of how to guard itself against this currency risk, and now we are going to list a number of natural possible strategies.

- The most naive strategy for *SWE* is perhaps to buy $\$1,000,000$ today at the price of $8,000,000$ SEK, and then keeping this money (in a Eurodollar account) for six months. The advantage of this procedure is of course that the currency risk is completely eliminated, but there are also some drawbacks. First of all the strategy above has the consequence of tying up a substantial amount of money for a long

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: mazzoran@math.unipd.it . Seminar held on 9 December 2020.

period of time, but an even more serious objection could be that SWE does not have access to $8,000,000$ SEK today.

- A more sophisticated arrangement, which does not require any outlays at all today, is that SWE goes to the forward market and buys a forward contract for $\$1,000,000$ with delivery six months from now. Such a contract may, for example, be negotiated with a commercial bank, and in the contract two things will be stipulated.
  - The bank will, at $t = T$, deliver $\$1,000,000$ to SWE.
  - SWE will, at $t = T$, pay for this delivery at the rate of K SEK/\$.

  The exchange rate K, which is called the forward price, (or forward exchange rate) at $t = 0$, for delivery at $t = T$, is determined at $t = 0$. By the definition of a forward contract, the cost of entering the contract equals zero, and the forward rate K is thus determined by supply and demand on the forward market. Observe, however, that even if the price of entering the forward contract (at $t = 0$) is zero, the contract may very well fetch a nonzero price during the interval $[0, T]$. Let us now assume that the forward rate today for delivery in six months equals 8.10 SEK/\$. If SWE enters the forward contract this simply means that there are no outlays today, and that in six months it will get $\$1,000,000$ at the predetermined total price of $8,100,000$ SEK. Since the forward rate is determined today, SWE has again completely eliminated the currency risk. However, the forward contract also has some drawbacks, which are related to the fact that a forward contract is a binding contract. To see this let us look at two scenarios.
  - Suppose that the spot currency rate at $t = T$ turns out to be 8.20. Then SWE can congratulate itself, because it can now buy dollars at the rate 8.10 despite the fact that the market rate is 8.20. In terms of the million dollars at stake SWE has thereby made an indirect profit of $8,200,000 - 8,100,000 = 100,000$ SEK, as one can check from Figure 1.1.
  - Suppose on the other hand that the spot exchange rate at $t = T$ turns out to be 7.90. Because of the forward contract this means that SWE is forced to buy dollars at the rate of 8.10 despite the fact that the market rate is 7.90, which implies an indirect loss of $8,100,000 - 7,900,000 = 200,000$ SEK, as one can check from Figure 1.2.

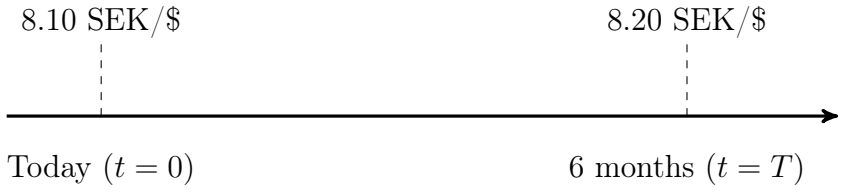


Figure 1.1. Indirect profit of $8,200,000 - 8,100,000 = 100,000$ SEK.

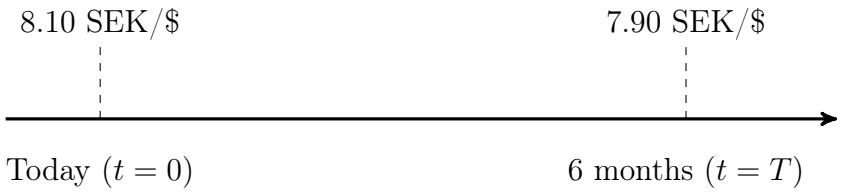


Figure 1.2. Indirect loss of $8,100,000 - 7,900,000 = 200,000$ SEK.

- What *SWE* would like to have of course is a contract which guards it against a high spot rate at $t = T$, while still allowing it to take advantage of a low spot rate at $t = T$. Such contracts do in fact exist, and they are called **European call options**. We will now go on to give a formal definition of such an option.

Definition 1.1.1 A European call option on the amount of X US dollars, with strike price (exercise price) K SEK/$ and exercise date T is a contract written at $t = 0$ with the following properties.

- The holder of the contract has, exactly at the time $t = T$, the right to buy X US dollars at the price K SEK/$.
- The holder of the option has no obligation to buy the dollars.

Concerning the nomenclature, the contract is called an option precisely because it gives the holder the option (as opposed to the obligation) of buying some underlying asset (in this case US dollars). A call option gives the holder the right to buy, whereas a put option gives the holder the right to sell the underlying object at a prespecified price. The prefix **European** means that the option can only be exercised at exactly the date of expiration. There also exist **American** options, which give the holder the right to exercise the option at any time before the date of expiration. Options of the type above (and with many variations) 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 object there are typically a large number of options with different dates of expiration and different strike prices.

We now see that *SWE* can insure itself against the currency risk very elegantly by buying a European call option, expiring six months from now, on a million dollars with a strike price of, for example, 8.00 SEK/$. If the spot exchange rate at T exceeds the strike price, say that it is 8.20, then *SWE* exercises the option and buys at 8.00 SEK/$. Should the spot exchange rate at T fall below the strike price, it simply abstains from exercising the option. Note, however, that in contrast to a forward contract, which by definition has the price zero at the time at which it is entered, an option will always have a nonnegative price, which is determined on the existing options market. This means that *SWE* will have the rather delicate problem of determining exactly which option they wish to buy, since a

higher strike price (for a call option) will reduce the price of the option. One of the main problems is to see what can be said from a theoretical point of view about the market price of an option like the one above. In this context it is worth noting that the European call has some properties which turn out to be fundamental.

- Since the value of the option (at $t = T$) depends on the future level of the spot exchange rate, the holding of an option is equivalent to a **future stochastic claim**.
- The option is a **derivative asset** in the sense that it is defined in terms of some **underlying** financial asset.

Since the value of the option is contingent on the evolution of the exchange rate, the option is often called a **contingent claim**. We can now formulate the two main problems that arise in the financial mathematical environment.

Main Problems: Take a fixed derivative as given.

- What is a "fair" price for the contract?
- Suppose that we have sold a derivative, such as a call option. Then we have exposed ourselves to a certain amount of financial risk at the date of expiration. How do we protect ("hedge") ourselves against this risk?

Let us look more closely at the pricing question above. There exist two natural and mutually contradictory answers.

Answer 1: "Using standard principles of operations research, a reasonable price for the derivative is obtained by computing the expected value of the discounted future stochastic payoff."

Answer 2: "Using standard economic reasoning, the price of a contingent claim, like the price of any other commodity, will be determined by market forces. In particular it will be determined by the supply and demand curves for the market for derivatives. Supply and demand will in their turn be influenced by such factors as aggregate risk aversion, liquidity preferences and so on, so it is impossible to say anything concrete about the theoretical price of a derivative."

The reason that there is such a thing as a theory for derivatives lies in the following fact.

Main Result: Both answers above are incorrect! It is possible (given, of course, some assumptions) to talk about the "correct" price of a derivative, and this price is not computed by the method given in *Answer 1* above. We can state the basic philosophy here. The main ideas are as follows.

Main Ideas

- A financial derivative is defined in terms of some underlying asset which already exists on the market.

- The derivative cannot therefore be priced arbitrarily in relation to the underlying prices if we want to avoid mispricing between the derivative and the underlying price.
- We thus want to price the derivative in a way that is consistent with the underlying prices given by the market.
- We are not trying to compute the price of the derivative in some "absolute" sense. The idea instead is to determine the price of the derivative in terms of the market prices of the underlying assets.

2 Some tools from Stochastic Analysis

2.1 Stochastic Integrals

In order to study asset pricing on financial markets in continuous time, we need to model asset prices as continuous time stochastic processes, and the most complete and elegant theory is obtained if we use diffusion processes and stochastic differential equations as our building blocks. What, then, is a diffusion? Loosely speaking we say that a stochastic process X is a diffusion if its local dynamics can be approximated by a stochastic difference equation of the following type.

$$X(t+\Delta t) - X(t) = \mu(t, X(t))\Delta t + \sigma(t, X(t))Z(t) \tag{2.1}$$

Here $Z(t)$ is a normally distributed disturbance term which is independent of everything which has happened up to time t, while μ and σ are given deterministic functions. The intuitive content of (2.1) is that, over the time interval $[t, t+\Delta t]$, the X-process is driven by two separate terms:

- a locally deterministic velocity $\mu(t, X(t))$;
- a Gaussian disturbance term, which is amplified by the factor $\sigma(t, X(t))$.

The function μ is called the (local) **drift** term of the process, whereas σ is called the **diffusion** term. In order to model the Gaussian disturbance terms we need the concept of a Wiener process.

Definition 2.1.1 A stochastic process W is called a Wiener process if the following conditions hold:

- $W_0 = 0$ almost surely;
- W_t has independent increments, i.e. $\forall k \geq 2$ and $0 \leq t_0 < t_1 < \cdots < t_k < \infty$ the random variables $\{W_{t_i} - W_{t_{i-1}}\}_{1 \leq i \leq k}$ are independent;
- W_t has stationary Gaussian centered increments, i.e. $\forall t > s \geq 0$ the random variables $W_t - W_s \sim \mathcal{N}(0, t-s)$;
- W_t has continuous trajectories, i.e. the map $t \mapsto W_t$ is almost surely continuous.

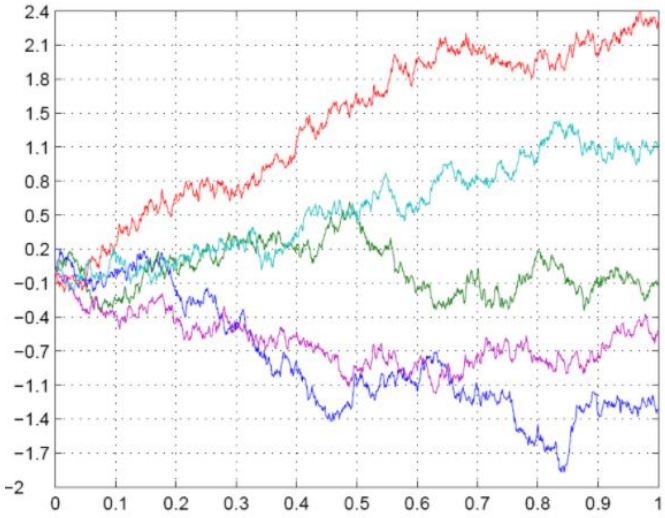


Figure 2.1 Some trajectories of a Wiener process.

We may now use a Wiener process in order to write (2.1) as

$$X(t+\Delta t) - X(t) = \mu(t, X(t))\Delta t + \sigma(t, X(t))\Delta W(t) \tag{2.2}$$

where $\Delta W(t)$ is defined by

$$\Delta W(t) = W(t+\Delta t) - W(t). \tag{2.3}$$

Let us now try to make (2.2) a bit more precise. It is then tempting to divide the equation by Δt and let Δt tend to zero. Formally we would obtain

$$\begin{aligned} &\dot{X}(t) = \mu(t, X(t)) + \sigma(t, X(t))v(t), \\ &X(0) = a \end{aligned} \tag{2.4}$$

where we have added an initial condition and where

$$v(t) = \frac{dW}{dt} \tag{2.5}$$

is the formal time derivative of the Wiener process W. If v were an ordinary (and well defined) process we would now in principle be able to solve (2.4) as a standard ordinary differential equation (ODE) for each v-trajectory. However, it can be shown that with probability 1 a Wiener trajectory is nowhere differentiable, see for example Karatzas and Shreve (2014), so the process v cannot even be defined. Thus this is a dead end.

Another possibility of making Equation (2.2) more precise is to let Δt tend to zero without first dividing the equation by Δt. Formally we will then obtain the expression

$$\begin{cases} dX(t) = \mu(t, X(t))dt + \sigma(t, X(t))dW(t), \\ X(0) = a \end{cases} \tag{2.6}$$

and it is now natural to interpret (2.6) as a shorthand version of the following integral equation

$$X(t) = a + \int_0^t \mu(s, X(s))ds + \int_0^t \sigma(s, X(s))dW(s) \tag{2.7}$$

In Equation (2.7) we may interpret the ds-integral as an ordinary Riemann integral. The natural interpretation of the dW-integral is to view it as a Riemann–Stieltjes integral for each W-trajectory, but unfortunately this is not possible since one can show that the W-trajectories are of locally unbounded variation. Thus the stochastic dW-integral cannot be defined in a naive way. As long as we insist on giving a precise meaning to Equation (2.2) for each W-trajectory separately, we thus seem to be in a hopeless situation. If, however, we relax our demand that the dW-integral in Equation (2.7) should be defined trajectory-wise we can still proceed. It is in fact possible to give a global (L^2-) definition of integrals of the form

$$\int_0^t g(s)dW(s) \tag{2.8}$$

for a large class of processes g. This new integral concept, the so called Itô integral, will then give rise to a very powerful type of stochastic differential calculus, the Itô calculus. The main steps to define this new integral concept are the following:

- define integrals of the type $\int_0^t g(s)dW(s)$;
- develop the corresponding differential calculus;
- analyze stochastic differential equations of the type (2.7) using the stochastic calculus above.

In order to proper define the stochastic integral and to construct it in details, we refer the reader, for example, to Karatzas and Shreve (2014) and Lamberton and Lapeyre (2007).

2.2 Stochastic Differential Equations

Let $M(n, d)$ denote the class of $n \times d$ matrices, and consider as given the following objects:

- a d-dimensional (column-vector) Wiener process W;
- a (column-vector valued) function $\mu : \mathbb{R}_+ \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$;
- a function $\sigma : \mathbb{R}_+ \times \mathbb{R}^n \longrightarrow M(n, d)$;
- a real (column) vector $x_0 \in \mathbb{R}^n$.

We now want to investigate whether there exists a stochastic process X which satisfies the **stochastic differential equation** (SDE)

$$\begin{aligned} dX_t &= \mu\left(t, X_t\right) dt + \sigma\left(t, X_t\right) dW_t \\ X_0 &= x_0. \end{aligned} \tag{2.9}$$

To be more precise we want to find a process X satisfying the integral Equation

$$X_t = x_0 + \int_0^t \mu\,(s, X_s)\,ds + \int_0^t \sigma\,(s, X_s)\,dW_s, \text{ for all } t \geq 0. \tag{2.10}$$

The standard method for proving the existence of a solution to the SDE above is to construct an iteration scheme of Cauchy–Picard type. The idea is to define a sequence of processes $X_0, X_1, X_2, \ldots$ according to the recursive definition

$$\begin{aligned} X_t^0 &\equiv x_0 \\ X_t^{n+1} &= x_0 + \int_0^t \mu\,(s, X_s^n)\,ds + \int_0^t \sigma\,(s, X_s^n)\,dW_s \end{aligned} \tag{2.11}$$

Having done this, one expects that the sequence $\{X^n\}_{n=1}^{\infty}$ will converge to some limiting process X, and that this X is a solution to the SDE. This construction can in fact be carried out, but here we only give the result, the interested reader may refer to Karatzas and Shreve (2014) for the proof.

Proposition 2.2.1 *Suppose that there exists a constant K such that the following conditions are satisfied for all x, y and t:*

$$\begin{aligned} \|\mu(t,x) - \mu(t,y)\| &\leq K\|x - y\| \\ \|\sigma(t,x) - \sigma(t,y)\| &\leq K\|x - y\| \\ \|\mu(t,x)\| + \|\sigma(t,x)\| &\leq K(1 + \|x\|). \end{aligned} \tag{2.12}$$

Then there exists a unique solution to the SDE (2.9). The solution has the following properties:

(a) X is $\mathcal{F}_t^W$-adapted;

(b) X has continuous trajectories;

(c) X is a Markov process;

(d) there exists a constant C such that

$$\mathbb{E}\left[\|X_t\|^2\right] \leq Ce^{Ct}\left(1 + \|x_0\|^2\right). \tag{2.13}$$

The fact that the solution X is $\mathcal{F}_t^W$-adapted means that for each fixed t the process value X_t is a functional of the Wiener trajectory on the interval $[0, t]$, and in this way an SDE induces a transformation of the space $C[0, \infty)$ into itself, where a Wiener trajectory $W(\omega)$ is mapped to the corresponding solution trajectory $X(\omega)$. Generically this transformation, which takes a Wiener trajectory into the corresponding X-trajectory, is enormously complicated and it is extremely rare that one can "solve" an SDE in some "explicit" manner. There are, however, a few nontrivial interesting cases where it is possible to solve an SDE, and the most important example for us is the equation below, describing the so-called Geometric Brownian motion (GBM).

2.2.1 Geometric Brownian Motion

Geometric Brownian motion will be one of our fundamental building blocks for the modeling of asset prices, and it also turns up naturally in many other places. Its SDE reads as

$$\begin{aligned} dX_t &= \alpha X_t dt + \sigma X_t dW_t \\ X_0 &= x_0. \end{aligned} \tag{2.14}$$

Thus we see that GBM can be viewed as a linear ODE, with a stochastic coefficient driven by white noise. See Figure 2.2, for a numerical simulation of GBM with $\alpha = 1$, $\sigma = 0.2$ and $X(0) = 1$. The smooth line is the graph of the expected value function $\mathbb{E}[X_t] = 1 \cdot e^{\alpha t}$. For small values of σ, the trajectory will (at least initially) stay fairly close to the expected value function, whereas a large value of σ will give rise to large random deviations. This can clearly be seen when we compare the simulated trajectory in Figure 2.2 to the three simulated trajectories in Figure 2.3 where we have $\sigma = 0.4$.

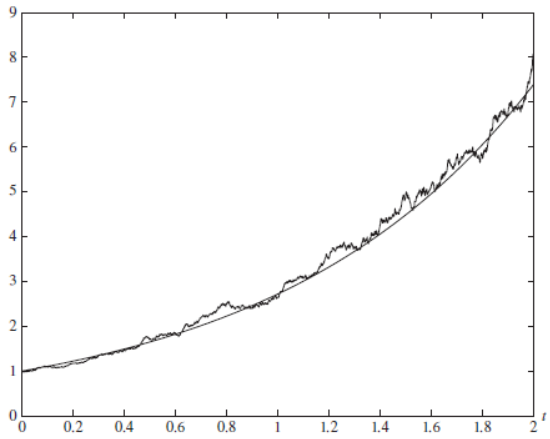


Figure 2.2 Geometric Brownian motion: $\alpha = 1$, $\sigma = 0.2$.

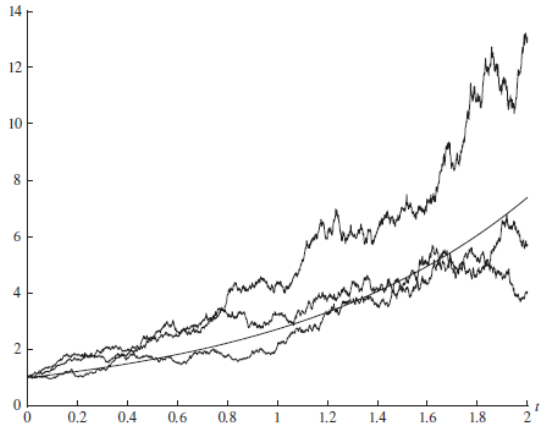


Figure 2.3 Geometric Brownian motion: $\alpha = 1$, $\sigma = 0.4$.

Inspired by the fact that the solution to the corresponding deterministic linear equation is an exponential function of time we are led to investigate the process Z, defined by $Z_t = \ln X_t$, where we assume that X is a solution and that X is strictly positive (see below). The Itô formula gives us

$$\begin{aligned} dZ &= \frac{1}{X}dX + \frac{1}{2}\left\{-\frac{1}{X^2}\right\}[dX]^2 \\ &= \frac{1}{X}\{\alpha X dt + \sigma X dW\} + \frac{1}{2}\left\{-\frac{1}{X^2}\right\}\sigma^2 X^2 dt \\ &= \{\alpha dt + \sigma dW\} - \frac{1}{2}\sigma^2 dt. \end{aligned} \tag{2.15}$$

Thus we have the equation

$$\begin{aligned} dZ_t &= \left(\alpha - \frac{1}{2}\sigma^2\right)dt + \sigma dW_t \\ Z_0 &= \ln x_0 \end{aligned} \tag{2.16}$$

This equation, however, is extremely simple: since the right-hand side does not contain Z it can be integrated directly to

$$Z_t = \ln x_0 + \left(\alpha - \frac{1}{2}\sigma^2\right)t + \sigma W_t \tag{2.17}$$

which means that X is given by

$$X_t = x_0 \cdot \exp\left\{\left(\alpha - \frac{1}{2}\sigma^2\right)t + \sigma W_t\right\} \tag{2.18}$$

Strictly speaking there is a logical flaw in the reasoning above. In order for Z to be well defined we have to assume that there actually exists a solution X to Equation (2.14) and we also have to assume that the solution is positive. As for the existence, this is covered by Proposition 2.2.1, but the positivity seems to present a bigger problem. We may actually avoid both these problems by regarding the calculations above as purely heuristic. Instead we define the process X by the formula (2.18). Then it is an easy exercise to show that X thus defined actually satisfies the SDE (2.14). Thus we really have proved the first part of the following result, which will be used repeatedly in the sequel. The result about the expected value follows easily.

Proposition 2.2.2 *The solution to the equation*

$$\begin{aligned} dX_t &= \alpha X_t dt + \sigma X_t dW_t \\ X_0 &= x_0 \end{aligned} \tag{2.19}$$

is given by

$$X(t) = x_0 \cdot \exp\left\{\left(\alpha - \frac{1}{2}\sigma^2\right)t + \sigma W(t)\right\}. \tag{2.20}$$

The expected value is given by

$$E[X_t] = x_0 e^{\alpha t}. \tag{2.21}$$

3 A general Overview on SLV models

One of the central problem in modern mathematical finance is derivative pricing. A derivative is a financial contract which value depends on an underlying asset which can be an equity stock, an interest rate or any different financial asset. The difficult concerning derivative pricing is to define a fair price. For this purpose a mathematical theory is needed. The well known Black-Scholes model was first introduced in 1973 and nowadays it represents an universal accepted framework for derivative pricing. In this introduction we briefly recall the main results of the standard theory, following Bj?ork (2009).

The original Black-Scholes model assumes the existence of a risk free asset B_t and of an underlying asset S_t, following respectively a deterministic and a Geometric Brownian motion dynamics:

$$dB_t = rB_t dt \tag{3.1}$$

$$dS_t = \mu S_t dt + \sigma S_t dW_t, \tag{3.2}$$

where the deterministic constant μ, σ and r represent respectively the local mean rate of return of the asset, the volatility of the asset and the short rate interest. W_t is a standard Wiener process. The model is widely employed as a useful approximation to reality, however its assumptions are not all empirically valid, see for example Bj?ork (2009).

Therefore, during the last two decades several models have been introduced with the aim of developing and generalizing the Black-Scholes framework for equity derivatives pricing. In particular, two main strands of research have been widely developed and used: Local Volatility (LV) models and Stochastic Volatility (SV) models.
Local Volatility models were introduced for the first time by Dupire et al. (1994) and Derman and Kani (1994) and they assume that the diffusion coefficient of the underlying asset is no longer a constant value but instead a deterministic function of time and of the underlying asset itself, namely

$$dS_t = (r - d)S_t dt + \eta_S (S_t, t) S_t dW_t,$$

where r and d are the risk-free interest rate and the dividend rate, respectively, and the function η_S is the so-called *local volatility function*.
In the Stochastic Volatility models instead, like in the highly celebrated Heston (1993), the volatility itself is considered to be a stochastic process v_t with its own dynamics. Thus, this is a two-factor model, driven by two correlated Wiener processes W_t and Z_t with correlation ρ, namely

$$\begin{aligned} dS_t &= (r - d)S_t dt + f(v_t) S_t dW_t \\ dv_t &= a(v_t, t)\, dt + c(v_t, t)\, dZ_t \\ dW_t dZ_t &= \rho dt, \end{aligned}$$

where r and d are the risk-free interest rate and the dividend rate, respectively, and f, a, c are functions of v_t.

Both of these models present many advantages, but also some drawbacks. The main advantage of local volatility models is their capability of a theoretically perfect fit of the market quoted plain vanilla options. If a good calibration of a local volatility model is

performed, the model can reproduce the market prices. Unfortunately, one of its main drawbacks is that its implied volatility dynamics is inconsistent with the observed one. On the other hand, a stochastic volatility model is able to reproduce a consistent dynamics, but not to fit exactly the market prices.
One of the possible answers to overcome such limitations is to look for a model with a stochastic volatility dynamics which can well replicate the market prices at the same time. The key ingredient to reach this is to merge the LV and the SV models characteristics in a suitable and consistent manner. More precisely, the main idea is to model the diffusion coefficient of the underlying asset process S_t as the product between a stochastic component $f(v_t)$ and a deterministic function $\ell(S_t, t)$. Thus, this generalized stochastic-local volatility (SLV) model, that dates back to Lipton (2002), is described by the following dynamics

$$\begin{aligned} dS_t &= (r-d)S_t dt + \ell\left(S_t, t\right) f\left(v_t\right) S_t dW_t \\ dv_t &= a\left(v_t, t\right) dt + c\left(v_t, t\right) dZ_t \\ dW_t dZ_t &= \rho dt, \end{aligned}$$

where r and d are the risk-free interest rate and the dividend rate, respectively, f,a,c are functions of v_t and ρ is the correlation between the Wiener processes W_t and Z_t. The function ℓ_S is the so-called *leverage function*.
The pivotal element of SLV models is the function ℓ_S that can be computed via an application of the Gyöngy lemma, see for example Gyöngy (1986).

References

Björk, T. (2009). "Arbitrage theory in continuous time", Oxford University Press.

Derman, E. and Kani, I. (1994). *Riding on a smile*. Risk, 7(2):32–39.

Dupire, B. et al. (1994). *Pricing with a smile*. Risk, 7(1):18–20.

Gyöngy, I. (1986). *Mimicking the one-dimensional marginal distributions of processes having an Itô differential*. Probability Theory and Related Fields, 71(4):501–516.

Heston, S. L. (1993). *A closed-form solution for options with stochastic volatility with applications to bond and currency options*. The Review of Financial Studies, 6(2):327–343.

Karatzas, I. and Shreve, S. (2014). "Brownian motion and stochastic calculus, volume 113", Springer.

Lamberton, D. and Lapeyre, B. (2007). "Introduction to stochastic calculus applied to finance", CRC Press.

Lipton, A. (2002). "The vol smile problem", Risk Magazine, 15:61–65.

Weak KAM, Homogenization and Ergodic Control: an introduction

Hicham Kouhkouh (*)

Abstract. This seminar is intended for non-specialists to whom I would first and succinctly introduce the general ideas behind the weak KAM theory in dynamical systems, the theory of homogenization in the analysis of PDEs and the theory of ergodic control. I will then and especially insist on the link between them, and which manifests itself in a particular PDE known as "ergodic stationary Hamilton-Jacobi equation". A qualitative study of these problems will also be discussed, and a link with an optimization problem in the space of measures will be mentioned. For the sake of clarity, I will focus on the deterministic case, but much of these results are valid in a stochastic framework.

1 Introduction

The aim of these notes is to summarize the seminar that the author gave as part of the PhD requirements in the department of mathematics of Padova University. The goal was to present this field of research to other graduate students who are not necessary familiar with and provide a qualitative insight of the different material presented.

The partial differential equations techniques related to the material presented in the sequel rely on viscosity methods. We refer to the book of Bardi & Capuzzo-Dolcetta [4] and to the user's guide [6] by Crandall, Ishii & Lions. Moreover, the present manuscript does not include sufficient references related to these topics. We refer for instance to the monograph [1] and the references therein. Finally, we refer to Fathi's book [11] (available online) for the weak KAM theory.

All the results are well known and are not due to the author. In particular, most of the material in §2 and §3 is taken from Evans [9], while §4 is taken from Barles [5]. See also the references therein for further details. And other references are quoted whenever it is needed. Mistakes or typos, if they exist, are of the responsibility of the author (H.K.).

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: kouhkouh@math.unipd.it . Seminar held on 16 December 2020.

2 On Weak KAM

Let $\mathbb{T}^n = [0,1]^n$ be the n-dimensional unit cube in $\mathbb{R}^n$ with opposite faces identified.

Definition 2.1 We call a *Lagrangian* $L(v,x)$ a function $L : \mathbb{R}^n \times \mathbb{T}^n \to \mathbb{R}$, where $x = (x_1, \dots, x_n) \in \mathbb{T}^n$ is the position variable and $v = (v_1, \dots, v_n) \in \mathbb{R}^n$ is the velocity.

If instead of $\mathbb{T}^n$ we are on a general manifold M, the *Lagrangian* L will be defined on its tangent bundle $TM \coloneqq \{(v,x) \,|\, x \in M \text{ and } v \in T_xM\}$. Indeed, $T(\mathbb{T}^n) = \mathbb{R}^n \times \mathbb{T}^n$.

We will hereafter make the following assumptions on the *Lagrangian* function L

(i) (uniform convexity) there exist constants $0 < \gamma \leq \Gamma$ such that

$$\gamma|\xi|^2 \leq \sum_{i,j=1}^{n} L_{v_i,v_j}(v,x)\xi_i\xi_j \leq \Gamma|\xi|^2$$

for all ξ, v, x; and

(ii) (periodicity) the mapping

$$x \mapsto L(v,x) \text{ is } \mathbb{T}^n\text{-periodic}$$

for all x.

Definition 2.2 Given a curve $\mathbf{x} : [0,T] \to \mathbb{R}^n$, we define its *action* to be

$$A_T[\mathbf{x}(\cdot)] \coloneqq \int_0^T L(\dot{\mathbf{x}}(t), \mathbf{x}(t))\, \mathrm{d}t$$

where $\dot{} = \frac{d}{dt}\cdot$ is the time derivative.

Definition 2.3 If $\mathbf{x} : [0,T] \times \mathbb{R}^n$ satisfies

$$A_T[\mathbf{x}(\cdot)] \leq A_T[\mathbf{y}(\cdot)]$$

for all curves $\mathbf{y}(\cdot)$ satisfying $\mathbf{y}(0) = \mathbf{x}(0)$ and $\mathbf{y}(T) = \mathbf{x}(T)$, we call $\mathbf{x}(\cdot)$ a minimizer of (or a *minimizing curve* for) the action $A_T[\cdot]$ on the time interval $[0,T]$.

The next theorem is a well known result in Calculus of Variations.

Theorem 2.4 (Euler-Lagrange equation) *Suppose that $x_0, x_T \in \mathbb{R}^n$ are given, and define the admissible class of curves*

$$\mathcal{A} \coloneqq \{\mathbf{y} \in C^2([O,T]; \mathbb{R}^n \,|\, \mathbf{y}(0) = x_0,\, \mathbf{y}(T) = x_T\}.$$

Suppose $\mathbf{x}(\cdot) \in \mathcal{A}$ and

$$A_T[\mathbf{x}(\cdot)] = \min_{\mathbf{y} \in \mathcal{A}} A_T[\mathbf{y}(\cdot)].$$

Then the curve $\mathbf{x}(\cdot)$ *solves the Euler-Lagrange equations*

$$-\frac{d}{dt}(D_v L(\dot{\mathbf{x}}, \mathbf{x})) + D_x L(\dot{\mathbf{x}}, \mathbf{x}) = 0 \quad (0 \leq t \leq T). \tag{2.1}$$

Example 2.5 The Lagrangian $L = \frac{|v|^2}{2} - W(x)$ is the difference between the kinetic energy $\frac{|v|^2}{2}$ and the potential energy $W(x)$. In this case the Euler-Lagrange equations (2.1) read

$$\ddot{\mathbf{x}} = -DW(\mathbf{x}).$$

Theorem 2.4 motivates the following definition.

Definition 2.6 We define the *minimal action* to go in time T from an intial position x_0 to a final position x_T in $\mathbb{R}^n$ by the function

$$h_T(x_0, x_T) = A_T[\mathbf{x}(\cdot)] = \min_{\mathbf{y} \in \mathcal{A}} A_T[\mathbf{y}(\cdot)]$$

The minimal action plays an important role in the weak KAM theory, but not only. Indeed it allows us to define the so-called *Lax-Oleinik* semigroup which is behind many concepts in analysis, calculus of variations and dynamical systems.

Definition 2.7 Let $u \in C(\mathbb{T}^n)$ and set

$$T_t^- u(x) := \inf \left\{ u(\mathbf{x}(0)) + \int_0^t L(\dot{\mathbf{x}}(s), \mathbf{x}(s)) \mathrm{d}s \,\middle|\, \mathbf{x}(t) = x \right\}. \tag{2.2}$$

where the infimum is taken over all admissible curves defiend on $[0, t]$ and satisfying $\mathbf{x}(t) = x$. We call the family of nonlinear operators $\{T_t^-\}_{t \geq 0}$ the *Lax-Oleinik* semigroup.

Remark 2.8 Using the minimal action, the function $x \mapsto T_t^- u(x)$ can be expressed as

$$T_t^- u(x) = \inf_{y \in \mathbb{T}^n} \{u(y) + h_t(y, x)\}.$$

We state in the following some properties of the Lax-Oleinik semigroup when acting on the space of continuous functions on $\mathbb{T}^n$ endowed with the max norm $\| \cdot \|$.

Proposition 2.9 *Recall the family of nonlinear operators* $\{T_t^-\}_{t \geq 0}$ *defined above. Then the following hold*

(a) $T_t^- \circ T_s^- = T_{t+s}^-$ *(semigroup property).*

(b) $f \leq g$ *implies* $T_t^- f \leq T_t^- g$.

(c) $T_t^-(f+c) = T_t^- f + c$.

(d) $\|T_t^- f - T_t^- g\| \leq \|f - g\|$ *(nonexpansiveness).*

(e) *for all* $f \in C(\mathbb{T}^n)$, $\lim_{t\to 0} T_t^- f = f$ *uniformly.*

(f) *for all* f, $t \mapsto T_t^- f$ *is uniformly continuous.*

We are now ready to state an important result in weak KAM theory due to A. Fathi [10].

Theorem 2.10 (Weak KAM Theorem) *There exists a function* $u \in C(\mathbb{T}^n)$ *and a unique constant* $c \in \mathbb{R}$ *such that*

$$T_t^- u_- + c\,t = u_- \quad \text{for all } t \geq 0.$$

The proof relies on the following abstract theorem about nonlinear mappings on a Banach space X.

Theorem 2.11 (Common fixed points) *Suppose* $\{\phi_t\}_{t\geq 0}$ *is a semigroup of nonexpansive mappings of* X *into itself. Assume also for all* $t > 0$ *that* $\phi_t(X)$ *is precompact in* X *and for all* $x \in X$ *that* $t \mapsto \phi_t(x)$ *is continuous.*
Then there exists a point x^* *such that*

$$\phi_t(x^*) = x^* \quad \text{for all times } t \geq 0.$$

Sketch of proof. (Weak KAM Theorem) For $u, \hat{u} \in C(\mathbb{T}^n)$, we write

$$u \sim \hat{u} \quad \text{if} \quad u - \hat{u} \equiv \text{constant},$$

and define the equivalence class

$$[u] := \{\hat{u} \,|\, u \sim \hat{u}\}.$$

Set

$$X := \{[u] \,|\, u \in C(\mathbb{T}^n)\}$$

with the norm

$$\|[u]\| := \min_{a\in\mathbb{R}} \|u + a\mathbb{1}\|,$$

where $\mathbb{1}$ denotes the constant function identically equal to 1.

We have $T_t : X \to X$, according to property (3) of Proposition 2.9. Hence Theorem 2.11 insures the existence of a common fixed point

$$T_t^-[u]^* = [u]^* \quad (t \geq 0).$$

Selecting any representative $u_- \in [u]^*$, we see that

$$T_t^- u_- = u_- + c(t)$$

for some $c(t)$. The semigroup property implies $c(t+s) = c(t) + c(s)$, and consequently $c(t) = ct$ for some constant c. □

The weak KAM theorem stated above brings to light two components: a continuous function u_- and a unique constant c. In what follows we will see how to characterize these two elements, first through a partial differential equation, and then via an optimization problem on the space of measures.

Characterization of u.

Recalling the definition (2.2) of the Lax-Oleinik semigroup T_t^-, we have for any function $u(\cdot)$ and for all $\mathbf{x}(\cdot)$ such that $\mathbf{x}(t) = x$

$$T_t^- u(x) \leq u(\mathbf{x}(0)) + \int_0^t L(\dot{\mathbf{x}}(s), \mathbf{x}(s))\, \mathrm{d}s. \tag{2.3}$$

In particular, when plugging the function u_- and using the weak KAM theorem, one gets

$$T_t^- u_-(x) + ct = u_-(x) \leq u_-(\mathbf{x}(0)) + \int_0^t L(\dot{\mathbf{x}}(s), \mathbf{x}(s))\, \mathrm{d}s + ct.$$

Recalling $\mathbf{x}(t) = x$, one has

$$\frac{u_-(\mathbf{x}(t)) - u_-(\mathbf{x}(0))}{t} \leq \frac{1}{t} \int_0^t L(\dot{\mathbf{x}}(s), \mathbf{x}(s))\, \mathrm{d}s + c.$$

Assume now the gradient of u_- exists almost everywhere (for example when u_- is Lipschitz continuous) and the admissible curve $\mathbf{x}$ is Lipschitz continuous. Then letting $t \to 0$ and setting $\dot{\mathbf{x}}(0) = v \in \mathbb{R}^n$, one finally gets

$$v \cdot Du_-(x) \leq L(v, x) + c.$$

The latter holds for any $v \in \mathbb{R}^n$, that is

$$H(Du_-(x), x) \leq c,$$

where H is the Hamiltonian and is defined by $H(p, x) := \max_{v \in \mathbb{R}^n} \{v \cdot p - L(v, x)\}$, i.e. the Legendre-Fenchel conjugate of the Lagrangian L. When L is defined on the tangent bundle TM of a general manifold M, the Hamiltonian is defined on its cotangent bundle T^*M and p is referred to as the *costate* variable, or in mechanics as the *momentum*.

Finally, the previous computations show us that provided we have an equality in (2.3) (for example when using a minimizing curve), the function u_- solves the partial differential equation

$$H(Du_-(x), x) = c \tag{2.4}$$

almost everywhere in $\mathbb{R}^n$. We usually refer to such equation as an *ergodic Hamilton-Jacobi* equation. It appears that the right notion of (weak) solution to such an equation is u_- being a viscosity solution. The proof that this is indeed the case follows [8]. This is summarized in the following theorem.

Theorem 2.12 *The function u_- in the weak KAM theorem is a viscosity solution to (2.4).*

Characterization of c.

Before we characterize the unique constant c in the weak KAM theorem, we need some definitions.

Recall the Euler-Lagrange equation (2.1) and consider the corresponding initial-value problem

$$\begin{cases} -\dfrac{d}{dt}(D_v L(\dot{\mathbf{x}}, \mathbf{x})) + D_x L(\dot{\mathbf{x}}, \mathbf{x}) = 0 \\ \mathbf{x}(0) = x,\ \dot{\mathbf{x}}(0) = v. \end{cases}$$

We define the *flow map* $\{\phi_t\}_{t\in\mathbb{R}}$ on $T(\mathbb{T}^n)$ by the formula

$$\phi_t(v, x) \coloneqq (\mathbf{v}(t), \mathbf{x}(t)),$$

where $\mathbf{v}(t) \coloneqq \dot{\mathbf{x}}(t)$.

Definition 2.13 A probability measure μ on the tangent bundle $T(\mathbb{T}^n)$ is *flow invariant* if

$$\int_{T(\mathbb{T}^n)} \Phi(\phi_t(v, x)) d\mu = \int_{T(\mathbb{T}^n)} \Phi(v, x) d\mu$$

for each bounded continuous function Φ.

Flow invariance is somtimes also denoted by $\phi_t \sharp \mu = \mu$ for all t, where $\phi_t \sharp \mu$ is the push-forward of μ defined by

$$\int_{T(\mathbb{T}^n)} \Phi(\phi_t(v, x)) \mathrm{d}\mu = \int_{T(\mathbb{T}^n)} \Phi(v, x) \mathrm{d}(\phi_t \sharp \mu), \quad \forall\, \Phi \text{ bounded continuous.}$$

We will finally denote by π the projection of $T(\mathbb{T}^n)$ onto $\mathbb{T}^n$, that is for all $(v, x) \in T(\mathbb{T}^n)$

$$\pi(v, x) \coloneqq x.$$

In particular, one has $(\dot{\mathbf{x}}(t), \mathbf{x}(t)) = \phi_t(v, x)$ and $\mathbf{x}(t) = \pi(\phi_t(v, x))$.

Let us now rewrite (2.3) using the weak KAM theorem and the above definitions

$$u_-(\pi(\phi_t(v, x))) \leq u_-(\pi(v, x)) + \int_0^t L(\phi_s(v, x))\, \mathrm{d}s + ct$$

We can now integrate with respect to a flow invariant measure μ

$$\int_{T(\mathbb{T}^n)} u_-(\pi(\phi_t(v, x)))\, \mathrm{d}\mu \leq \int_{T(\mathbb{T}^n)} u_-(\pi(v, x))\, \mathrm{d}\mu + \int_{T(\mathbb{T}^n)} \int_0^t L(\phi_s(v, x))\, \mathrm{d}s \mathrm{d}\mu + ct.$$

Using Definition 2.13, one gets

$$0 \leq \int_0^t \int_{T(\mathbb{T}^n)} L(v,x)\,\mathrm{d}\mu\mathrm{d}s + ct$$

which finally yields after simplifying by t

$$-c \leq \int_{T(\mathbb{T}^n)} L(v,x)\,\mathrm{d}\mu, \quad \text{for all } \mu \text{ flow invariant.}$$

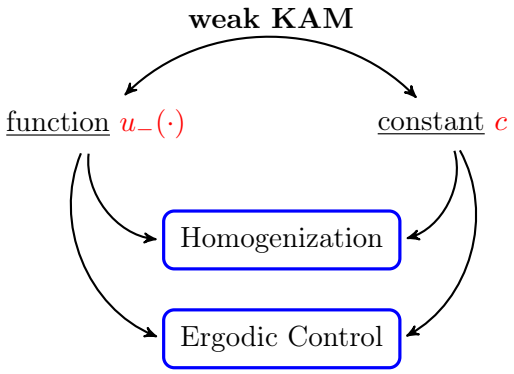
This motivates the following theorem.

Theorem 2.14 *The constant c from the weak KAM theorem is given by*

$$c = -\inf\left\{ \int_{T(\mathbb{T}^n)} L(v,x)\,d\mu \,\middle|\, \forall\, \mu \textit{ flow invariant, probability measure} \right\}.$$

This is an optimization problem over the set of probability measures satisfying the flow invariane constraint. Such a problem is known as *Mather variational problem*. And one of the goals of weak KAM theory is the study of the support of the measures solving this problem, also known as *Mather sets*, in terms of the underlying Hamiltonian dynamics.

To sum up.
The weak KAM theorem as we stated above puts in light two components: a function u_- and a constant c. We have seen that one can characterize each of them by means either of a partial differential equation or by an optimization problem. We will see in what follows the link of these two elements with other theories that although they have been developed independently, they are still intimately related to the latter.



3 On Homogenization

Homogenization is related to "averaging": we want to provide a *macroscopic* description to phenomena with *microscopic* behavior, through a smart averaging procedure. In other words, we want to extract "averaged" information from disordered/heterogenous media, or also to describe the behavior of a multiscale phenomenon. Such averaging is obtained with an aymptotic analysis. Indeed, denoting by $\varepsilon \ll 1$ the ratio between the microscopic and macroscopic levels, our phenomenon is then described by $\mathbf{x}$ (for the macroscopic level) and by $\frac{\mathbf{x}}{\varepsilon}$ (for the microscopic level). Homoegenization is then concerned with the study of what happens when $\varepsilon \to 0$.

We will consider in what follows an instance of first order partial differential equations and see how homogenization translates in this situation. The results in this section are due to Lions, Papanicolaou & Varadhan [12].

Theorem 3.1 (Homogenization) *Suppose g is bounded and uniformly continuous, and u^ε is the unique bounded, uniformly continuous viscosity solution of the initial-value problem*

$$\begin{cases} \partial_t u^\varepsilon + H\left(Du^\varepsilon, \dfrac{x}{\varepsilon}\right) = 0, & (t>0) \\ u^\varepsilon = g, & (t=0). \end{cases}$$

Then $u^\varepsilon \to u$ locally uniformly, where u solves the homogenized equation

$$\begin{cases} \partial_t u^\varepsilon + \overline{H}(Du) = 0, & (t>0) \\ u = g, & (t=0). \end{cases}$$

Before we sketch the proof, we need the following result which concerns $\overline{H}(\cdot)$ called *the effective Hamiltonian*, and the problem (3.1) is called *the Cell problem.*

Theorem 3.2 (Effective Hamiltonian) *For each vector $P \in \mathbb{R}^n$, there exists a unique real number $c(P)$ for which we can find a viscosity solution of*

$$\begin{cases} H(P + D_x v, x) = c(P) \\ v \ \text{ is } \mathbb{T}^n - \text{periodic.} \end{cases} \tag{3.1}$$

Remark 3.3 Notice that the problem (3.1) writes as (2.4). It suffices indeed to write for a fixed $P \in \mathbb{R}^n$, $\omega(x) := P\cdot x + v(x)$. Then $D\omega = P + D_x v$ and (3.1) becomes $H(D\omega, x) = c$.

Sketch of proof (Effective Hamiltonian). Step 1. (Existence) We approximate the problem (3.1) by a sequence of other problems

$$\varepsilon v^\varepsilon + H(P + D_x v^\varepsilon, x) = 0$$

for which it is easy to show existence of a unique solution v^ε by the usual viscosity methods. Now since H is periodic in x, uniqueness implies v^ε is also periodic.
We assume moreover we have the following uniform estimates

$$\max |D_x v^\varepsilon|, |\varepsilon v^\varepsilon| \le C.$$

Hence we may extract subsequences for which

$$v^{\varepsilon_j} \to v \ \text{ uniformly, } \varepsilon_j v^{\varepsilon_j} \to -c(P).$$

It is then straightforward to confirm that v is a viscosity solution of $H(P + D_x v, x) = c(P)$.

Step 2. (uniqueness of $c(P)$) Suppose also

$$H(P + D_x\hat{v}, x) = \hat{c}(P).$$

We may assume that $\hat{c}(P) > c(P)$ and that $\hat{v} < v$, upon adding a constant to v, if necessary. Then

$$\delta\hat{v} + H(P + D_x\hat{v}, x) > \delta v + H(P + D_x v, x)$$

in viscosity sense, if $\delta > 0$ small. The viscosity solution comparison principle then implies the contradiction $\hat{v} \geq v$. □

Sketch of proof (Homogenization). Let ϕ be a smooth function and suppose that $u - \phi$ has a strict maximum at the point (x_0, t_0). Define the perturbed test function

$$\phi^\varepsilon(x,t) \coloneqq \phi(x,t) + \varepsilon v\left(\frac{x}{\varepsilon}\right),$$

where v is a periodic viscosity solution of

$$H(P + Dv, x) = \overline{H}(P)$$

for $P = D\phi(x_0, t_0)$.

Assume for the rest of the discussion that v is smooth. Then ϕ^ε is smooth and $u^\varepsilon - \phi^\varepsilon$ attains a max at a point $(x_\varepsilon, t_\varepsilon)$ near (x_0, t_0). Consequently

$$\partial_t\phi^\varepsilon + H\left(D\phi^\varepsilon, \frac{x_\varepsilon}{\varepsilon}\right) \leq 0.$$

And then

$$\partial_t\phi + H\left(D\phi(x_\varepsilon, t_\varepsilon) + Dv\left(\frac{x_\varepsilon}{\varepsilon}\right), \frac{x_\varepsilon}{\varepsilon}\right) \approx \partial_t\phi + \overline{H}(D\phi(x_0, t_0)) \leq 0.$$

The reverse inequality similarly holds if $u - \phi$ has a strict minimum at the point (x_0, t_0). We refer to [7] for the case when v is not smooth. □

4 On Ergodic Control

Before we move to ergodic control, let us first recall the minimal action in Definition 2.6

$$h_T(x_0, x_T) = \min_{\mathbf{x}(\cdot)\in\mathcal{A}} \left\{ \int_0^T L(\dot{\mathbf{x}}(s), \mathbf{x}(s))\,\mathrm{d}s \;\middle|\; \mathbf{x}(0) = x_0, \mathbf{x}(T) = x_T \right\}.$$

This is a *calculus of variations* problem where we aim at finding a curve $s \mapsto \mathbf{x}(s)$ which minimizes during the time interval $[0, T]$ some running cost L (the Lagrangian) that depends on its velocity $\dot{\mathbf{x}}$ and position $\mathbf{x}$. In particular we are supposing we do have access to the velocity of each curve, i.e. we can observe (or measure) the velocity. Optimal control theory deals with the case where we do not have a direct access to the full vector of velocity, but we have rather the possibility to choose some parameter which intervenes

in its velocity. This is formulated by a controlled (or parametrized) ordinary differential equation, and the problem writes for example as

$$\begin{aligned} u_T(x_0, x_T) = \min_{\alpha(\cdot)} \int_0^T L(\alpha(s), \mathbf{x}(s))\, \mathrm{d}s \\ \text{subject to: } \dot{\mathbf{x}}(s) = F(\mathbf{x}(s), \alpha(s)), \quad s \in (0,T) \\ \mathbf{x}(0) = x_0,\ \mathbf{x}(T) = x_T. \end{aligned}$$

where $\alpha(\cdot) : [0,T] \to A$ is the control and A is a subset of $\mathbb{R}^n$ for example.

Remark 4.1 In the particular case of $F(\alpha, x) = \alpha$, we have $u_T(x_0, x_T) = h_T(x_0, x_T)$.

Ergodic control problems deal with the case where we are interested in the behavior of the function $u_T(\cdot,\cdot)$ as $T \to +\infty$, and also aim at characterizing the limit of $\frac{1}{T}u_T$ as $T \to +\infty$ in addition to the corresponding long time behavior of the optimal trajectories. In the sequel, we will drop the dependency of $u_T(\cdot)$ on the final position x_T, and consider it as a function of the initial position x_0 and of the time horizon T. We will then write $u(x,T)$.

Before we go any further, let us state a result on the characterization of the value function $u(\cdot,\cdot)$ as the unique solution to a first order Hamilton-Jacobi equation. We denote by $H(x,p)$ the Hamiltonian and is defined as

$$H(x,p) = \sup_{a \in A}\{-F(x,a)\cdot p - L(a,x)\}.$$

We assume in addition that H satisfies the following

(a) H is $\mathbb{Z}^n$-periodic in x, i.e. $H(x+z,p) = H(x,p)$ for all $z \in \mathbb{Z}^n$, for all $x, p \in \mathbb{R}^n$;

(b) H is coercive, i.e. $H(x,p) \to +\infty$ when $|p| \to +\infty$, uniformly in x.

Using Dynamic Programming principle, one can prove that the value function u_T satisfies in the viscosity sense the Hamilton-Jacobi equation

$$\partial_t u(x,t) + H(x, Du) = 0, \qquad \text{in } \mathbb{R}^n \times (0, +\infty) \tag{4.1}$$

complemented with the initial condition $u(x,0) = 0$.

Theorem 4.2 *Under the standing assumption, there exists a unique solution u to the Hamilton-Jacobi equation (4.1) that is periodic in x and Lipschitz contiuous in x and t on $\mathbb{R}^n \times (0, +\infty)$.*

Proof. See for instance [5, Theorem 10.1]. □

The first question one deals with in ergodic control is the limit (whether it exists or not) of the time average of the value function $\frac{1}{t}u(x,t)$ as $t \to +\infty$. This is the object of the next theorem.

Theorem 4.3 *Under the standing assumptions, there exists a constant $c \in \mathbb{R}$ such that*

$$\frac{u(x,t)}{t} \to c, \quad \text{as } t \to +\infty \quad \text{uniformly w.r.t. } x \in \mathbb{R}^n.$$

Proof. See for instance [5, Theorem 10.2]. □

In the light of the previous sections, we are tempted by asking whether one can characterize the constant c in the latter theorem. Moreover, a natural question that also arises is whether $u(x,t) - ct$ defines a certain particular function $v(x)$. In this case, one can assert that the value function $u(x,t)$ admits an asymptotic behavior of the form $u(x,t) \approx ct + v(x)$ as t goes to infinity.

Indeed, if one plugs $ct+v(x)$ in the Hamilton-Jacobi equation (4.1), then $(c, v(\cdot))$ should satisfy

$$c + H(x, Dv) = 0. \tag{4.2}$$

But this is nothing but a Cell problem (for $P = 0$) as in (3.1) or again an ergodic Hamilton-Jacobi equation as in (2.4). Remember that for such equation, both the constant c and the function $v(\cdot)$ are unknown. We then have the result due to Lions, Papanicolaou & Varadhan [12] and that we have already mentioned in Theorem 3.2.

Theorem 4.4 *Under the standing assumptions, there exists a unique constant $c \in \mathbb{R}$ such that the equation (4.2) has a periodic and Lipschitz continuous solutions.*

Under further assumptions, one can show that $u(x,t) - ct$ does indeed converge to $v(x)$. See for instance [5, Theorem 10.5].

Other important questions concern the behavior of the optimal trajectories as the time horizon goes to infinity. The work of Arisawa [2, 3] answered some of these questions.

5 Conclusion

An important problem in weak KAM is the search and study of subsets invariant by the flow of the dynamics. It appears that these invariant subsets are related to the support of the invariant measure that minimizes the action and which is in turn related to the regularity of the solution to the HJ equation. This connexion is very deep, and establishes a link between several fields of research.

Other important questions concern the constant c , its uniqueness and its physical (dynamical) interpretation. In fact, it is somehow the level-0 of the energy of the system. Note in addition that a direct link with quantum mechanics does exist which leaves open many other questions. An extension to a stochastic framework is also possible.

Finally, we saw the problem of calculus of variations (no agent) and a control problem (one agent), but we can do the same for the problem of games (two agents), and mean field games (infinitely many agents).

A Appendix: Viscosity solutions

We recall from [4] some basic theory of continuous viscosity solutions of the Hamilton-Jacobi equation (HJ) of the general form

$$F(x, u(x), Du(x)) = 0, \quad x \in \Omega$$

where Ω is an open domain of $\mathbb{R}^n$ and the Hamiltonian $F = F(x, r, p)$ is a continuous rela valued function on $\Omega \times \mathbb{R} \times \mathbb{R}^n$.

Definition A.1 A function $u \in C(\Omega)$ is a viscosity subsolution of (HJ) if, for any $\varphi \in C^1(\Omega)$

$$F(x_0, u(x_0), D\varphi(x_0)) \leq 0$$

at any local maximum point $x_0 \in \Omega$ of $u-\varphi$. Similarly, $u \in C(\Omega)$ is a viscosity supersolution of (HJ) if, for any $\varphi \in C^1(\Omega)$,

$$F(x_1, u(x_1), D\varphi(x_1)) \geq 0$$

at any local minimum point $x_1 \in \Omega$ of $u - \varphi$. Finally, u is a viscosity solution of (HJ) if it is simultaneously a viscosity sub- and supersolution.

Geometrically, this means that the vailidity of the subsolution (resp. supersolution) condition for u is tested on smooth functions "touching from above (resp. below)" the graph of u at x_0 (resp. x_1).

The following proposition explains the local character of the notion of viscosity solution and its consistency with the classical pointwise definition.

Proposition A.2

(a) *if $u \in C(\Omega)$ is a viscosity solution of (HJ) in Ω, then u is a viscosity solution of (HJ) in Ω', for any open set $\Omega' \subset \Omega$;*

(b) *if $u \in C(\Omega)$ is a classical solution of (HJ), that is, u is differentiable at any $x \in \Omega$ and*

$$F(x, u(x), Du(x)) = 0, \quad \forall\, x \in \Omega$$

then u is a viscosity solution of (HJ);

(c) *if $u \in C^1(\Omega)$ is a viscosity solution of (HJ), then u is a classical solution of (HJ).*

Proof. See in [4] proof of Proposition 1.3. □

Lemma A.3 *Let $u \in C(\Omega)$. Then,*

(a) *$p \in D^+u(x)$ if and only if there exists $\varphi \in C^1(\Omega)$ such that $D\varphi(x) = p$ and $u - \varphi$ has a local maximum at x;*

(b) *$p \in D^-u(x)$ if and only if there exists $\varphi \in C^1(\Omega)$ such that $D\varphi(x) = p$ and $u - \varphi$ has a local minimum at x.*

Proof. See in [4] proof of Lemma 1.7 □

As a direct consequence of the previous Lemma, the following new definition of viscosity solution turns out to be equivalent to the initial one: a function $u \in C(\Omega)$ is a viscosity subsolution of (HJ) in Ω if

$$F(x, u(x), p) \leq 0, \quad \forall\, x \in \Omega,\ \forall\, p \in D^+u(x)$$

a viscosity supersolution of (HJ) in Ω if

$$F(x, u(x), p) \geq 0, \quad \forall\, x \in \Omega,\ \forall\, p \in D^-u(x)$$

The following proposition is a consistency result that improves Proposition A.2.

Proposition A.4

(a) *If $u \in C(\Omega)$ is a viscosity solution of (HJ), then*

$$F(x, u(x), Du(x)) = 0$$

at any point $x \in \Omega$ where u is differentiable;

(b) *if u is locally Lipschitz continuous and it is a viscosity solution of (HJ), then*

$$F(x, u(x), Du(x)) = 0, \quad \textit{almost everywhere in } \Omega.$$

Proof. See in [4] proof of Proposition 1.9. □

References

[1] O. Alvarez and M. Bardi, "Ergodicity, stabilization, and singular perturbations for Bellman-Isaacs equations". American Mathematical Soc., 2010.

[2] M. Arisawa, *Ergodic problem for the Hamilton-Jacobi-Bellman equation. i. Existence of the ergodic attractor.* In Annales de l?Institut Henri Poincare (C) Non Linear Analysis, vol. 14, Elsevier (1997), 415–438.

[3] M. Arisawa, *Ergodic problem for the Hamilton-Jacobi-Bellman equation. ii..* In Annales de l?Institut Henri Poincare (C) Non Linear Analysis, vol. 15, Elsevier (1998), 1–24.

[4] M. Bardi and I. Capuzzo-Dolcetta, "Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations". Springer Science & Business Media, 2008.

[5] G. Barles, "First-order Hamilton-Jacobi equations and applications". CIME Course, (2011).

[6] M.G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations.* Bulletin of the American Mathematical Society, 27 (1992), 1–67.

[7] L.C. Evans, *Periodic homogenisation of certain fully nonlinear partial differential equations.* Proceedings of the Royal Society of Edinburgh Section A: Mathematics, 120 (1992), 245–265.

[8] L.C. Evans, "Partial differential equations". Vol. 19 of Grad. Stud. Math., AMS, Providence, 2002.

[9] L.C. Evans, *Weak kam theory and partial differential equations.* In Calculus of variations and nonlinear partial differential equations, Springer (2008), 123–154.

[10] A. Fathi, *Théoreme KAM faible et théorie de Mather sur les systemes lagrangiens.* Comptes Rendus de l'Académie des Sciences, Series I-Mathematics, 324 (1997), 1043–1046.

[11] A. Fathi, "Weak KAM theorem in lagrangian dynamics". 2003.

[12] P.-L. Lions, G. Papanicolaou, and S.S. Varadhan, *Homogenization of Hamilton-Jacobi equations.* Unpublished work (1986).

An introduction to the moduli space of smooth curves and its compactification

Angelina Zheng (*)

Abstract. The moduli space of algebraic curves is a central object in algebraic geometry. The idea behind this space is that it answers to a classification problem, by allowing us to classify algebraic curves up to isomorphisms. Nonetheless, the geometry of this space is rather abstract and subtle, and only few general statements are known. The aim of this talk is to give an elementary introduction to the moduli space of n-pointed smooth curves of genus g and its compactification: we will define the main ingredients and present some basic properties. We will also discuss basic examples in order to get familiar with these spaces. One way to understand these spaces is by computing some topological invariants, such as their cohomology groups. In general, their full cohomology ring is still unknown, but we have a complete description for some values of g, which I will present in the final part.

1 Introduction

We set our ground field to be the field of complex numbers $\mathbf{C}$. The main goal of this seminar is to give a basic introduction to the moduli space of n-pointed smooth curves of genus g and its compactification. First of all, we will give the definition of a moduli space. The idea behind these spaces is that they allow us to classify objects up to some equivalence ralation. We will start with a very basic examples, coming from linear algebra.

Example 1 We would like to classify finite-dimensional vector spaces over a field k up to ismorphisms. We know that two vector spaces V, W are isomorphic if and only if $\dim_k V = \dim_k W$. Thus we can write

$$\{\text{finite-dimensional vector spaces over } k\}/\text{iso} \cong \mathbf{Z}_{\geq 0}$$
$$[V] \mapsto \dim_k V.$$

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: zheng@math.unipd.it . Seminar held on 3 February 2021.

So we can identify all finite-dimensional vector spaces with the set of positive integers. Note however that $\mathbf{Z}_{\geq 0}$ is not a vector space. Indeed, we would like our "classifying space" to have a similar structure of the objects that we are classifying. Consider then the following example.

Example 2 We would like to classify all lines in $\mathbf{C}^{n+1}$, up to the equivalence relation "being parallel", which is the same as classifying all distinct lines through the origin. Each line is then determined by a non-zero vector $v \in \mathbf{C}^{n+1}\backslash\{0\}$ and two lines are equal if and only they are spanned by two vectors $v, w \in \mathbf{C}^{n+1}$ which are linearly dependent, i.e if there exists $\lambda \in \mathbf{C}^*$ s.t. $v = \lambda w$. Thus,

$$\{\text{lines in}\mathbf{C}^{n+1}\}/\sim \quad \cong \quad \mathbf{C}^{n+1}\backslash\{0\}/\mathbf{C}^* \quad =: \quad \mathbf{P}^n$$

Here, both lines and the projective space have the structure of algebraic varieties and this is what we are actually looking for in a moduli space.
We will indeed consider the category of algebraic varieties over $\mathbf{C}$: $Var_{\mathbf{C}}$. (Note however that the following can be adapted to any category). First of all we need to define a *moduli problem* (in the category of varieties).

Definition 1 A moduli problem is the data of

- A class of objects, hence varieties, $\mathcal{A}$, plus a notion of what it means to have a family $\mathcal{F}(\mathbb{B})$ of objects in $\mathcal{A}$ over a variety $\mathbb{B}$;
- for any variety $\mathbb{B}$, an equivalence relation on $\mathcal{F}(\mathbb{B})$, and consequently on the objects in $\mathcal{A}$.

If $\mathcal{M}$ is a variety, such that:

(a) $\mathcal{M} = \mathcal{A}/\sim$;

(b) for any family $f : \mathfrak{X} \to \mathbb{B}$, the map

$$\begin{aligned} \phi_f : \mathbb{B} &\to \mathcal{M} \\ b &\mapsto [\mathfrak{X}_b] \end{aligned}$$

is a morphism of varieties;

then $\mathcal{M}$ is a **coarse moduli space**.
If there also exists a *universal family* $\mathcal{U} \xrightarrow{g} \mathcal{M}$ such that any family $f : \mathfrak{X} \to \mathbb{B}$ makes the following diagram commutative:

$$\begin{array}{ccc} \mathfrak{X} \cong \mathbb{B} \times_{\mathcal{M}} \mathcal{U} & \longrightarrow & \mathcal{U} \\ \downarrow f & & \downarrow g \\ \mathbb{B} & \xrightarrow{\phi_f} & \mathcal{M} \end{array}$$

then $\mathcal{M}$ is a **fine moduli space**.
Since we are only interested in giving an idea of the spaces defined above, we won't enter in details and we won't discuss families but we will only think of moduli spaces as set of isomorphism classes.
Consequently, we won't even distinguish coarse and fine moduli spaces, whose difference is mostly based on the presence of non-trivial automorphisms for the ojects in $\mathcal{A}$, which do not allow the existence of a universal family.

2 $\mathcal{M}_{g,n}$

We would like to classify

$$\mathcal{A} = \{\text{smooth irreducible projective } n\text{-pointed curves of genus } g\}/\sim;$$

Even though we won't discuss families, just for completeness, here $\mathcal{F}(\mathbb{B}) = \{\text{proper flat morphisms } f : \mathcal{C} \to \mathbb{B};\ \mathcal{C}_b \in \mathcal{A}\}$; where the isomorphism can be extended naturally to families.
In the following we recall the definition of the main ingredients.

Definition 2

- A *curve* C is a variety of dimension 1.
- an n-pointed curve is curve C together with n distinct points $p_1, \dots, p_n \in C$;
- C is *projective* if it is the vanishing locus of a finite set of homogeneous polynomials $f_1, \dots, f_r \in \mathbf{C}[x_0, \dots, x_N]$;
- C is *irreducible* if it is not the union of two proper closed subsets (in the Zariski topology);
- C is *smooth* if $\operatorname{rank}\left(\frac{\partial f_i}{\partial x_j}(p)\right) \geq N - 1$ for any $p \in C$.

Definition 3 Let C be a smooth projective curve. Its *geometric genus* is defined to be

$$p_g(C) = \dim H^0(C, \Omega^1_C),$$

where Ω^1_C is the cotangent line bundle of C.
Its *arithmetic genus* is defined to be

$$p_a(C) = 1 - \chi(C, \mathcal{O}_C),$$

where $\mathcal{O}_C$ is the structure sheaf of C.

Note that for smooth projective irreducible curves, $p_g(C) = p_a(C)$.
All this terminology might sound very complicated, however, there is another approach for studying algebraic curves, which is via Riemann surfaces. Indeed, since we are working over the complex number, if we consider C with the complex topology instead, we get a

compact connected manifold of complex dimension 1, i.e. a compact connected oriented real surface withoud boundary, called compact *Riemann surfaces*.
From the classification of closed connected surfaces, each Riemann surface is homeomorphic to a sphere or a connected sum of g tori, $g \geq 1$.

Definition 4 The number g is called the *topological genus* of the surface.

It turns out that the arithmetic (or geometric) genus of an algebraic curve is equal to the topological genus of the underlying surface.

Definition 5 An n-pointed curve is a curve C together with n distinct points $p_1, \dots, p_n \in C$.

From now on we will consider only curves having g, n such that $2g - 2 + n > 0$. The reason behind this assumption will be discussed later in the next section.

Theorem 2.1 (Deligne-Mumford) *There exist coarse moduli spaces $\mathcal{M}_g$ and $\mathcal{M}_{g,n}$, which are irreducible quasi-projective algebraic varieties of dimension $3g-3$ and $3g-3+n$, respectively.*

We end this section by discussing some examples.

Example 3 Le $g = 0$. From the assumption $2g - 2 + n > 0$ we require $n \geq 3$.

$\mathcal{M}_{0,3}$: We recall first the following fact

Proposition 2.2 *A projective smooth curve of genus 0 must be isomorphic to $\mathbf{P}^1$.*

As a consequence, any 3-pointed curve of genus 0 is isomorphic to $(\mathbf{P^1}; p_1, p_2, p_3)$. Moreover, $Aut(\mathbf{P}^1) = PGL(2) = GL(2)/\mathbf{C}^*$ and its action on $\mathbf{P}^1$is 3-transitive, meaning that for any three distinct points p_1, p_2, p_3 there is a unique automorphism sending them to $0, 1, \infty$. Thus $(\mathbf{P^1}; p_1, p_2, p_3) \cong (\mathbf{P^1}, 0, 1, \infty)$ and $\mathcal{M}_{0,3} = pt$.

$\mathcal{M}_{0,4}$: From what we saw above, any curve $(C; p_1, p_2, p_3, p_4)$ can be uniquely identifed with $(\mathbf{P}^1; 0, 1, \infty; t)$, for some $t \neq 0, 1, \infty$.
Thus, $\mathcal{M}_{0,4} = \mathbf{P}^1 \backslash \{0, 1, \infty\}$.

$\mathcal{M}_{0,n}$: We can generalize the result obtained for $n = 4$: each curve $(C; p_1, \dots, p_n)$ can be uniquely identifed with $(\mathbf{P}^1; 0, 1, \infty, t_1, \dots, t_{n-3})$ with $t_i \neq 0, 1, \infty$ and $t_i \neq t_j$. Thus $\mathcal{M}_{0,n} = \{(t_1, \dots, t_{n-3}) \in (\mathbf{P}^1 \backslash \{0, 1, \infty\})^{n-3} | \, t_i \neq t_j\}$.

Example 4 Let $g = 1$. From the assumption $2g - 2 + n > 0$ we require $n \geq 1$.
A smooth curve of genus 1 with a marked point is an *elliptic curve*, which is isomorphic to the quotient of $\mathbf{C}$ by a rank 2 lattice $\Lambda : z_1\mathbf{Z} + z_2\mathbf{Z}$ with z_1, z_2 linearly independent. Two lattices Λ, Λ' are isomorphic if $\Lambda = c\Lambda'$ for some $c \in \mathbf{C}^*$. Hence

$$\mathcal{M}_{1,1} = \{\text{rank 2 lattices}\}/\mathbf{C}^*.$$

However we can say more, for instance, if we rescale and rotate the lattice, then $\Lambda = \mathbf{Z}+\tau\mathbf{Z}$ with $\tau \in \mathfrak{h}$, the upper-half plane. Moreover, choosing another basis of the lattice Λ will give us another point $\tau' \in \mathfrak{h}$ such that $\Lambda = \mathbf{Z} + \tau\mathbf{Z} \cong \Lambda' = \mathbf{Z} + \tau'\mathbf{Z}$. The group of base change is

$$SL(2,\mathbf{Z}) = \{\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathbf{Z}_{2\times 2} |\, ad - bc = 1\}$$

which acts on $\mathfrak{h}$ by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \tau = \frac{a\tau + b}{c\tau + d}.$$

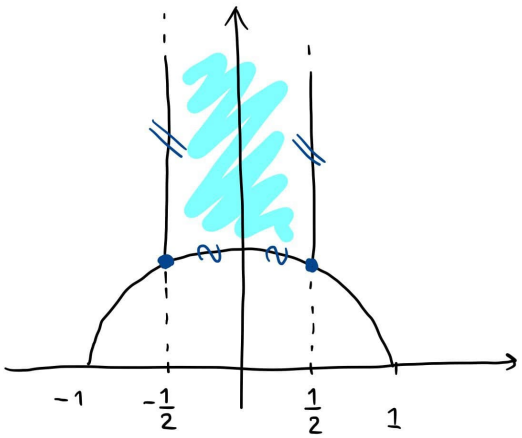
Hence,

$$\mathcal{M}_{1,1} = \mathfrak{h}/SL(2,\mathbf{Z}).$$

We can also represent this set of point with a *fundalmental domain* in $\mathfrak{h}$ by considering the action of $SL(2,\mathbf{Z})$. The group is generated by two transformations: $T : \tau \mapsto \tau + 1$, $S : \tau \mapsto -\frac{1}{\tau}$, which are respectively the translation to the right by 1, and the inversion with respect to the unit circle followed by the reflection with respect to the imaginary axis. This allows us to identify all points in $\mathfrak{h}$ that differ from each other by a translation by 1 and the points inside the unit circle with those outside. To be more precise, we can identify $\mathcal{M}_{1,1}$ with

$$\mathcal{F} := \{\tau \in \mathfrak{h} | \frac{1}{2} \leq \mathrm{Re}\,\tau \leq \frac{1}{2};\, |\tau| \geq 1\},$$

where two distinct points τ_1, τ_2 are equivalent only if $\mathrm{Re}(\tau_1) = \pm\frac{1}{2}$ and $\tau_2 = \tau_1 \pm 1$ or if τ_1 is on the unit circle and $\tau_2 = -\frac{1}{\tau_1}$.



Remark 1 The *j-invariant* $j : \mathfrak{h} \to \mathbf{C}$ is invariant under the action of $SL(2,\mathbf{Z})$ and restricted to the domain defined above gives us a bijection. Thus we may identify $\mathcal{M}_{1,1}$ with $\mathbf{C}$.

Note: $\mathcal{M}_{g,n}$ is not compact!

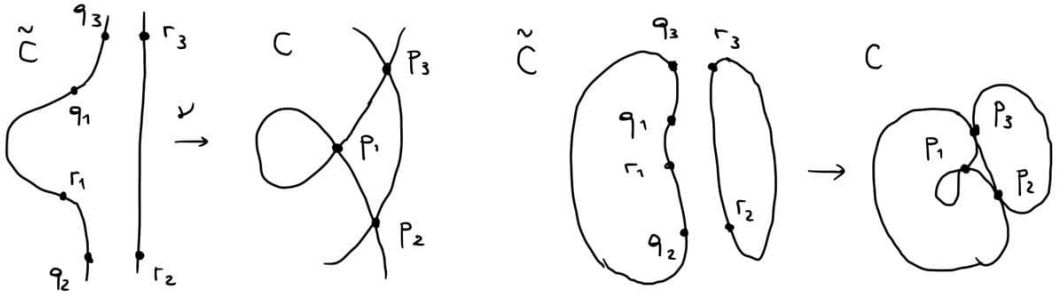
3 $\overline{\mathcal{M}}_{g,n}$

In this section we want to compactify the coarse moduli space of (n- pointed) smooth curves of genus g.
The first guess is to consider singular curves, namely *nodal* curves.

Definition 6 A *nodal* curve is a complete connected curve such that every singular point is a *node*, i.e. a point with a neighborhood which is analytically isomorphic to a neighborhood of 0 in $\{(x, y) \in \mathbf{C}^2 | \, xy = 0\} \subset \mathbf{C}^2$.

Definition 7 An *n-pointed nodal* curve is a nodal curve C, together with n distinct points $p_1, \ldots, p_n \in C$.

For any nodal curve C there exists a *normalization* $\nu : \tilde{C} \to C$ that behaves as follows.



The normalization separetes all the two branches of each node and if the curve is reducible, then its normalization is the disjoint union of the normalization of each component. In particular $\{q_i, r_i\} = \nu^{-1}(p_i)$ for each node p_i, with $q_i \neq r_i$.
Let $\tilde{C}$ be the normalization of a nodal curve C, with irreducible components $\{\tilde{C}_1, \ldots, \tilde{C}_r\}$, The geometric genus of a nodal curve is the genus of its normalization $\tilde{C}$:

$$p_g(C) = p_g(\tilde{C}) = \sum_{i=1}^{r} p_g(\tilde{C}_i),$$

and its arithmetic genus is

$$p_a(C) = p_g(\tilde{C}) - r + \#\{\text{nodes}\} + 1.$$

In the following we will consider the arithmetic genus. The reason behind this choice is because the arithmetic genus is constant on families of curves while the geometric genus might change.

However, it turns out that there are too many nodal curves having same genus to obtain a compact moduli space: in general a sequence of nodal curves would never converge to a smooth curve. In fact, we need to resctict to *stable* curves.

Definition 8 A *stable* curve is a nodal curve having only finitely many automorphisms.

There is a very explicit criterion for a nodal curve to be stable.

Proposition 3.1 *Let $(C; p_1, \dots, p_n)$ be a nodal curve with n marked points. It is stable if and only if, for any irreducible component of its normalization $\tilde{C}_i \subset \tilde{C}$ satifies*

$$2g(\tilde{C}_i) - 2 + \#Q \cap \tilde{C}_i + \#N \cap \tilde{C}_i > 0,$$

where Q is the set of pre-images of the nodes of C and N is the set of marked points.

Remark 2 An n-pointed smooth curve of genus g has finitely many automorphisms if and only if $2g - 2 + n > 0$.

Hence, the reason behind the assumption $2g-2+n > 0$ we left unanswered is to ensure that the smooth curves that we considered were automatically stable, so that $\emptyset \neq \mathcal{M}_{g,n} \subset \overline{\mathcal{M}}_{g,n}$.

Theorem 3.2 (Deligne-Mumford-Knudsen) *There exist coarse moduli spaces $\overline{\mathcal{M}}_g$ and $\overline{\mathcal{M}}_{g,n}$ of stable curves and n-pointed stable curves of genus g, which are irreducible projective algebraic varieties dimension $3g-3$ and $3g-3+n$, respectively.*

According to the *Stable reduction theorem* and the *valuative criterion of properness* $\overline{\mathcal{M}}_g$ and $\overline{\mathcal{M}}_{g,n}$ are actually complete and give a compactification for $\mathcal{M}_g$ and $\mathcal{M}_{g,n}$, respectively.

We will discuss some example also for these new moduli spaces, but before doing so, we present a nice characterization of nodal curves by some combinatorial objects: weighted graphs. Let us recall their definition

Definition 9 A *graph* Γ consists of

- A finite set of vertices $V(\Gamma)$;
- a finite set of edges $E(\Gamma)$;
- a finite set of legs $L(\Gamma)$.

A *weighted* graph is a graph Γ together with a function $w : V(\Gamma) \to \mathbf{Z}_{\geq 0}$.
We say that two weighted graphs (Γ, w) and (Γ', w') are isomorphic if there is a bijectition $f : V(\Gamma) \to V(\Gamma')$ such that $w'(f(v)) = w(v)$ for any vertex $v \in V(\Gamma)$ and f also preserves edges and legs attached to each vertex.

The *genus* of a weighted graph (Γ, w) is

$$g(\Gamma) = \sum_{v \in V(\Gamma)} w(v) + 1 - \#V(\Gamma) + \#E(\Gamma).$$

A *stable graph* Γ is a weighted graph for which

$$2w(v) - 2 + \#\{\text{incident legs or edges to } v\} > 0$$

for any $v \in V(\Gamma)$. So in particular,

- $w(v) = 0$ and $\#\{\text{incident legs or edges to } v\} \geq 3$, or
- $w(v) = 1$ and $\#\{\text{incident legs or edges to } v\} \geq 1$, or
- $w(v) > 2$.

Given a nodal curve C we can define its *dual graph* Γ_C as follows:

- Vertices correspond to the components $\{\tilde{C}_1, \dots \tilde{C}_r\}$;
- edges are given by the points of $Q \cap \tilde{C}_i$;
- legs are given by the points of $N \cap \tilde{C}_i$;
- the weight function assigns to each vertex the genus of the corresponding component.

With this definition, one can check that on dual graphs the notions of genus and stability coincide exactly with the ones on the associated nodal curve.

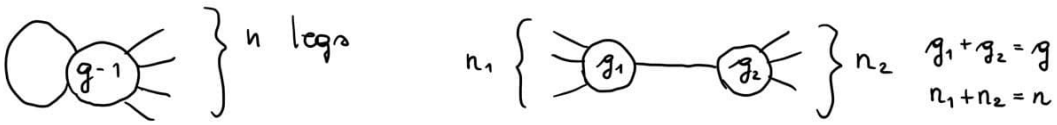
Thanks to this representation via graphs we also get a nice description of the boundary $\partial\mathcal{M}_{g,n} := \overline{\mathcal{M}}_{g,n} \backslash \mathcal{M}_{g,n}$:

$$\partial\mathcal{M}_{g,n} = \bigcup_{\Gamma : \#E(\Gamma)=1} \mathcal{M}_\Gamma,$$

with Γ a stable graph of genus g and n legs and

$$\mathcal{M}_\Gamma := \{(C; p_1, \dots, p_n) | \Gamma_C \cong \Gamma\}.$$

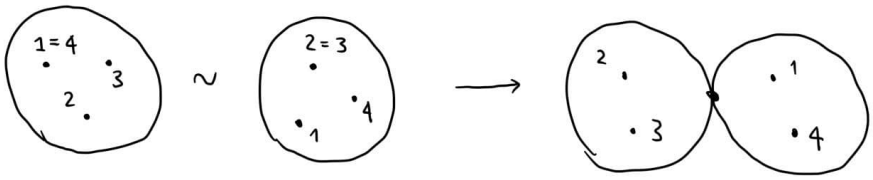
Hence we get two types of graphs:



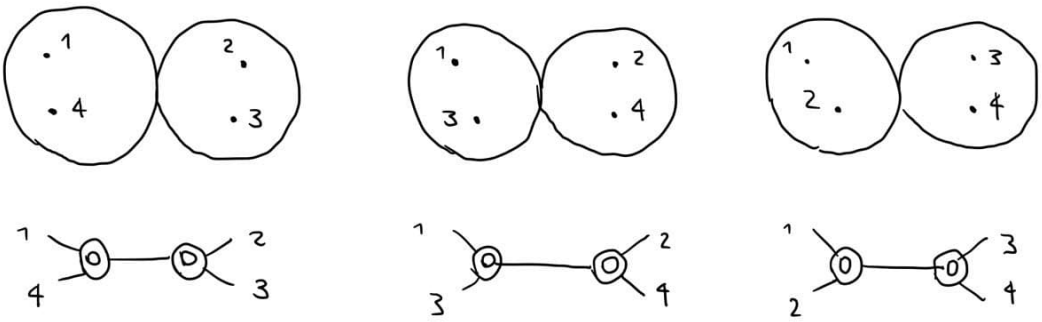
Finally, we go back to examples 3 and 4 and see the corresponbing compactification.

Example 5 We saw that $\mathcal{M}_{0,3}$ is a point, hence it is already compact. Indeed the unique stable curve of genus 0 with three marked points is smooth.

$\mathcal{M}_{0,4} = \mathbf{P}^1 \backslash \{0, 1, \infty\}$ so we expect that its compactification is obtained by adding three points. To get the compactification, we need to allow t to take values $0, 1, \infty$.
Let $(C; p_1, p_2, p_3, p_4) \cong (\mathbf{P}^1; 0, 1, \infty, t)$. When $t \to 0$, we get $p_1 = p_4$. However, changing its local coordinate $x \to x/t$ gives us $(C; p_1, p_2, p_3, p_4) \cong (\mathbf{P}^1; 0, 1/t, \infty, 1)$ and thus $p_2 = p_3$ for $t \to 0$. Since there is no reason to prefer one local coordinate to the other the right thing to do is to include both limit curves in the description of the limit:



The right-hand component corresponds to the initial local coordinate x, while the left-hand component corresponds to the local coordinate $\frac{x}{t}$.
Repeating for $t \to 1$, $t \to \infty$, gives us similar descriptions, hence we obtain $\overline{\mathcal{M}}_{0,4}$ from $\mathcal{M}_{0,4}$ by adding the following three points.



Example 6 Recall that $\mathcal{M}_{1,1}$ can be identified with the points in the complex line $\mathbf{C}$. Therefore $\overline{\mathcal{M}}_{1,1} \cong \mathbf{P}^1$. It is obtained by adding one point ∞, which corresponds to the only curve of genus 1 with a marked point having exactly one node.

4 Rational cohomology of some moduli spaces of curves

In this last section, we would like to present some result on the cohomology of moduli spaces. We will restrict to unmarked smooth curves of genus g. From what we discussed from the previous section, the geometry of these spaces is rather mysterious. One way to study their geometry is to compute some topological invariant, such as their (rational) cohomology groups, which are defined as the singular cohomology groups with rational coefficients. Indeed, we will consider only rational coefficients beacuse of the properties that hold in this case, such as Poincaré duality. The importance of this invariant lies on its ring structure and on the geometric information that we can deduce from its elements: given a closed algebraic subset $S \subset \mathcal{M}_g$ of complex codimension c, we can associate to S a cohomology class $[S] \in H^{2c}(\mathcal{M}_g; \mathbf{Q})$. Moreover, if $S' \subset \mathcal{M}_g$ is another algebraic set meeting S transversally, then the class $[S \cap S']$, associated to the intersection between S and S', is equal to the *cup product* $[S] \smile [S']$, which is the product that defines the ring

structure.

The full cohomology ring of $\mathcal{M}_g$ is known only for $g \leq 4$.
Let $\mathbf{Q}(-k)$ denote the Hodge structure of Tate of weight $2k$.

Theorem 4.1 *$\mathcal{M}_2$ has the rational cohomology of a point;*

Theorem 4.2 (Looijenga [1])

$$H^i(\mathcal{M}_3; \mathbf{Q}) = \begin{cases} \mathbf{Q}, & i = 0; \\ \mathbf{Q}(-1), & i = 2; \\ \mathbf{Q}(-6), & i = 6; \\ 0, & \textit{otherwise.} \end{cases}$$

Theorem 4.3 (Tommasi [3])

$$H^i(\mathcal{M}_4; \mathbf{Q}) = \begin{cases} \mathbf{Q}, & i = 0; \\ \mathbf{Q}(-1), & i = 2; \\ \mathbf{Q}(-2), & i = 4; \\ \mathbf{Q}(-3), & i = 5; \\ 0, & \textit{otherwise.} \end{cases}$$

What do we know for $g \leq 5$?
Let $\mathcal{T}_g \subset \mathcal{M}_g$ be the locus in $\mathcal{M}_g$ of trigonal curves, i.e. non-hyperelliptic curves admitting a degree 3 map to $\mathbf{P}^1$. . For $g \leq 4$, this space is not really interesting because it either coincides with $\mathcal{H}_g$ or its complement. Thus, since it is known that $\mathcal{H}_g$ has the cohomology of a point, $g = 5$ represents the first case in which the cohomology of $\mathcal{T}_g$ cannot be automatically determined from that of $\mathcal{H}_5$ and $\mathcal{M}_5$.

Theorem 4.4 (- [4])

$$H^i(\mathcal{T}_5; \mathbf{Q}) = \begin{cases} \mathbf{Q}, & i = 0; \\ \mathbf{Q}(-1), & i = 2; \\ \mathbf{Q}(-3), & i = 5; \\ \mathbf{Q}(-11), & i = 12; \\ 0, & \textit{otherwise.} \end{cases}$$

This result not only represents and advance in the computation of the cohomology of $\mathcal{M}_5$, but hopefully also for that of $\mathcal{T}_g$, for any $g \geq 5$.

References

[1] Eduard Looijenga, *Cohomology of $\mathcal{M}_3$ and $\mathcal{M}_3^1$*. Mapping class groups and moduli spaces of Riemann surfaces 150 (1993), 205–228.

[2] Johannes Schmitt, "The moduli space of curves". Lecture notes, 2020.

[3] Orsola Tommasi., *Rational cohomology of the moduli space of genus 4 curves*. Compositio Mathematica 141/2 (2005), 359–384.

[4] Angelina Zheng, *Rational cohomology of the moduli space of trigonal curves of genus 5*. arXiv preprint arXiv:2007.12150v2 (2020).

[5] Dimitri Zvonkine, "An introduction to moduli spaces of curves and their intersection theory". Lecture notes, 2014.

Limit theorems for Lévy flights on a 1D Lévy random medium

Elena Magnanini (*)

Abstract. We study a random walk on a point process given by an ordered array of points $(\omega_k,\, k \in \mathbb{Z})$ on the real line. The distances $\omega_{k+1} - \omega_k$ are i.i.d. random variables in the domain of attraction of a β-stable law, with $\beta \in (0,1) \cup (1,2)$. The random walk has i.i.d. jumps such that the transition probabilities between ω_k and ω_ℓ depend on $\ell - k$ and are given by the distribution of a $\mathbb{Z}$-valued random variable in the domain of attraction of an α-stable law, with $\alpha \in (0,1) \cup (1,2)$. Since the defining variables, for both the random walk and the point process, are heavy-tailed, we speak of a *Lévy flight on a Lévy random medium*. For all combinations of the parameters α and β, we prove the annealed functional limit theorem for the suitably rescaled process, relative to the optimal Skorokhod topology in each case. When the limit process is not càdlàg, we prove convergence of the finite-dimensional distributions. When the limit process is deterministic, we also prove a limit theorem for the fluctuations, again relative to the optimal Skorokhod topology.

Mathematics Subject Classification (2020): 60G50; 60G55; 60F17; 82C41; 60G51

Keywords: Random walk on point process; Lévy random medium; Lévy flights; Stable distributions; Anomalous diffusion; Stable processes.

1 Introduction

The expression 'Lévy random medium' indicates a stochastic point process, in some space, where the distances between nearby points have heavy-tailed distributions. Processes of this kind have been receiving a surge of attention, of late, both in the physical and mathematical literature; cf., respectively, [BFK, S, BCV, BDLV, ZDK] and [BCLL, BLP, MS]. They model a variety of situations that are of interest in the sciences. In particular, they are used as supports for various kinds of random walks, in order to study phenomena of anomalous transport and anomalous diffusion. An incomplete list of general or recent references on this topic includes [SZF, KRS, CGLS, ZDK, ACOR, ROAC].

The random medium that we consider in this talk is perhaps the most natural choice for a Lévy random medium in the real line: a sequence of random points $\omega = (\omega_k,\, k \in \mathbb{Z})$, where $\omega_0 = 0$ and the nearest-neighbor distances $\zeta_k = \omega_{k+1} - \omega_k$ are i.i.d. variables in

(*) Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: elena.magnanini@unipd.it. Seminar held on 17 February 2021.

the normal domain of attraction of a β-stable variable, with $\beta \in (0,1) \cup (1,2)$. Here β is the index of the stable distribution, not the skewness parameter, which equals 1 because $\zeta_k > 0$.

A random walk $Y = (Y_n,\, n \in \mathbb{N})$ takes place on ω according to the following rule. Independently of ω, there exists a random walk $S = (S_n,\, n \in \mathbb{N})$ on $\mathbb{Z}$ with $S_0 = 0$ and i.i.d. increments in the normal domain of attraction of an α-stable variable, with $\alpha \in (0,1) \cup (1,2)$. We define $Y_n := \omega_{S_n}$. This means that the process Y performs the same jumps as S, but on the marked points ω_k instead of $\mathbb{Z}$. For example, if a realization of S is $(0, 3, -1, \ldots)$, the process Y starts at the origin of $\mathbb{R}$, then jumps to the third marked point to the right of 0, then to the first marked point to the left of 0, and so on. In other words, S *drives* the dynamics of Y on the Lévy medium. For this reason we call it the *underlying random walk*.

Our process of interest is Y. We may describe it as a *Lévy flight on a one-dimensional Lévy random medium*. This phrase is borrowed from the physical literature, where the term 'Lévy flight' usually indicates a discrete-time random walk with long-tailed instantaneous jumps. This is in contrast to a 'Lévy walk', which in general designates a *persistent*, continuous- or discrete-time, random walk with long-tailed trajectory segments that are run at constant finite speed [ZDK]. A Lévy walk is often seen as an interpolation of a Lévy flight. In this seminar we give *annealed* limit theorems for Y in all cases $\alpha, \beta \in (0,1) \cup (1,2)$, identifying in each case both the scale n^γ whereby

$$\left(\frac{Y_{\lfloor nt \rfloor}}{n^\gamma} \,,\, t \in [0, +\infty) \right) \tag{1}$$

converges to a non-null limit, and the limit process. In all cases we prove the optimal, or at least morally optimal, functional limit theorem, meaning that we show distributional convergence of the process with respect to (w.r.t.) the strongest Skorokhod topology that applies there. When the limit process is not càdlàg (or càglàd) we show convergence of the finite-dimensional distributions. Finally, in the cases where the limit of (1) is deterministic, we prove a functional limit theorem for the corresponding fluctuations, again relative to the optimal topology.

2 Model

2.1 Setup

As mentioned in the introduction, the Lévy flight on random medium that we consider is a random walk performed over the points of a certain random point process. We proceed to define all the necessary constructions.

Random medium. Let $\zeta := (\zeta_i,\, i \in \mathbb{Z})$ be a sequence of i.i.d. positive random variables. We assume that the law of ζ_i belongs to the normal basin of attraction of a β-stable distribution, with $\beta \in (0,1) \cup (1,2)$. In the case $\beta \in (0,1)$, this means that, as $n \to +\infty$,

$$\frac{1}{n^{1/\beta}} \sum_{i=1}^{n} \zeta_i \xrightarrow{\mathrm{d}} Z_1^{(\beta)}, \tag{2}$$

where $Z_1^{(\beta)}$ is a stable variable of index β and skewness parameter 1 (because $\zeta_i > 0$). In the case $\beta \in (1,2)$ we have instead

$$\frac{1}{n^{1/\beta}} \sum_{i=1}^{n} (\zeta_i - \nu) \xrightarrow{\mathrm{d}} \tilde{Z}_1^{(\beta)}, \tag{3}$$

for a stable variable $\tilde{Z}_1^{(\beta)}$ of index β. In this case, necessarily, ν is the expectation of ζ_i and the skewness parameter is 0.

The random medium associated to $(\zeta_i, i \in \mathbb{Z})$ is defined to be:

$$\omega_0 = 0, \qquad \omega_k = \begin{cases} \sum_{i=1}^{k} \zeta_i & \text{if } k > 0, \\ 0 & \text{if } k = 0, \\ -\sum_{i=k}^{-1} \zeta_i & \text{if } k < 0. \end{cases} \tag{4}$$

This determines a point process $\omega := (\omega_k,\, k \in \mathbb{Z})$ on $\mathbb{R}$ that we call *Lévy random medium* to emphasize the fact that the distribution of ζ_i has a heavy tail. Each point ω_k will be called a *target*. In other words, the distances between neighboring targets are drawn according to independent random variables ζ_i.

Underlying random walk. We consider a $\mathbb{Z}$-valued random walk $S := (S_n,\, n \in \mathbb{N})$, with $S_0 = 0$ and i.i.d. increments $\xi_i := S_i - S_{i-1}$ that are independent of ζ (and thus of ω). In other words, S is given by

$$S_0 = 0, \qquad S_n = \sum_{i=1}^{n} \xi_i \quad \text{for } n \in \mathbb{Z}^+. \tag{5}$$

The law of ξ_i belongs to the normal basin of attraction of an α-stable distribution, with $\alpha \in (0,1) \cup (1,2)$. This means that convergences analogous to those given in (2) and (3) apply to the ξ_i, with limit random variables denoted by $W_1^{(\alpha)}$ and $\widetilde{W}_1^{(\alpha)}$, respectively. We will refer to S as the *underlying* random walk.

Random walk on the random medium. The *random walk on the random medium* $Y := (Y_n,\, n \in \mathbb{N})$ is defined to be:

$$Y_n := \omega_{S_n}, \quad n \in \mathbb{N}. \tag{6}$$

In other words, Y performs the same jumps as S, but on the points of ω; see Figure 1 for a hands-on explanation. In the following we will focus on the derivation of the asymptotic law of Y, under suitable scaling, with respect to the probability measure $\mathbb{P}$ governing the entire system (medium and dynamics). This is sometimes referred to as the *annealed* or *averaged* law of Y.

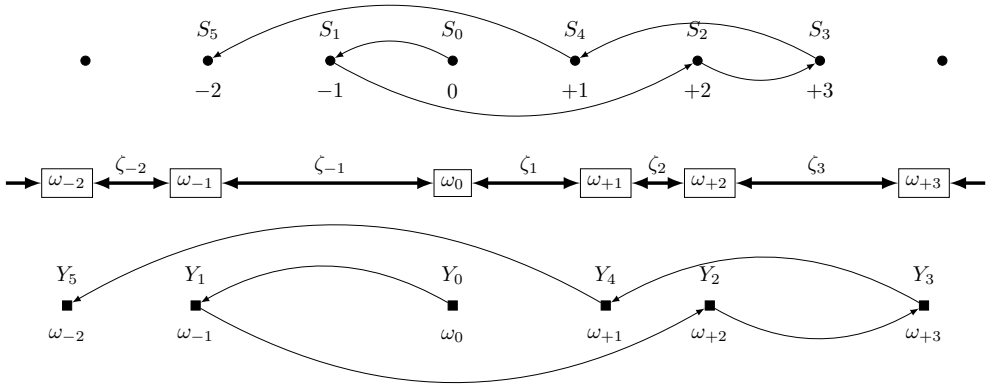


Figure 1 Top: A realization of the underlying random walk S on $\mathbb{Z}$. Middle: A realization of the random medium ω, with inter-distances given by ζ_i. Bottom: The corresponding process Y jumps between the targets ω according to the walk S.

Before recalling certain basic facts about the processes ω and S, and stating our main results on the process Y, let us fix the notation for spaces of càdlàg functions endowed with certain Skorokhod topologies.

3 Càdlàg functions and Skorokhod topologies

Given I, an interval or a half-line contained in $\mathbb{R}^+ := [0, +\infty)$, we denote by $\mathcal{D}(I) \equiv \mathcal{D}(I; \mathbb{R})$ the space of all càdlàg functions $f : I \longrightarrow \mathbb{R}$, where we recall that these are right-continuous functions with left limits at all points of their domain. If I is an interval or a half-line intersecting $(-\infty, 0)$, or $I = \mathbb{R}$, we consider a less customary function space: $\mathcal{D}(I)$ is the space of all functions $f : I \longrightarrow \mathbb{R}$ such that $s \mapsto f(s)$ is càdlàg for $s \geq 0$ and $s \mapsto f(-s)$ is càdlàg for $s \geq 0$ (in other words, the restriction of f to $I \cap (-\infty, 0]$ is càglàd). Notice that this implies that f is continuous at 0. We also use the abbreviations $\mathcal{D}^+ \equiv \mathcal{D}(\mathbb{R}^+)$ and $\mathcal{D} \equiv \mathcal{D}(\mathbb{R})$. Lastly, we denote by $\mathcal{D}_0$ and $\mathcal{D}_0^+$ the subspaces of nondecreasing functions of $\mathcal{D}$ and $\mathcal{D}^+$, respectively.

In this section we introduce two notions of distance/topology that turn out to be crucial in the following. A complete treatment of these topologies can be found, e.g., in [W2, Sections 3.3. & 11.5].

Definition 3.1 Let I be a bounded interval (which can be closed, open or half-open). For $f, g \in \mathcal{D}(I)$, denote

$$d_{J_1,I}(f,g) := \inf_{\lambda : I \to I} \max \left\{ \sup_{t \in I} |f \circ \lambda(t) - g(t)| \, , \, \sup_{t \in I} |\lambda(t) - t| \right\}, \tag{7}$$

where the infimum is taken over all increasing homeomorphisms $\lambda : I \longrightarrow I$. This defines a distance on $\mathcal{D}(I)$, which we refer to as the J_1 or $J_1(I)$ distance.

This metric induces a topology and a notion of limit in $\mathcal{D}(I)$ which can be reformulated as follows: given $(f_n)_{n\in\mathbb{N}}$ and f in $\mathcal{D}(I)$, the sequence f_n is said to converge to f in the J_1 topology, and we write

$$f_n \to f \qquad \text{in } (\mathcal{D}(I), J_1), \tag{8}$$

as $n \to \infty$, if there exists a sequence of increasing homeomorphisms $\lambda_n : I \longrightarrow I$ such that

$$\lim_{n\to\infty} \sup_{t\in I} |f_n \circ \lambda_n(t) - f(t)| = 0, \tag{9}$$

$$\lim_{n\to\infty} \sup_{t\in I} |\lambda_n(t) - t| = 0. \tag{10}$$

If we think of f_n as describing the spatial motion of some particle, the function $\lambda : I \longrightarrow I$ of (7) is sometimes called the *time change*. Requiring the time change to be a homeomorphism is occasionally too strong a condition. One has a weaker topology if they only require that λ be a (possibly discontinuous) bijection:

Definition 3.2 If I is a bounded interval and $f, g \in \mathcal{D}(I)$, the J_2 or $J_2(I)$ distance $d_{J_2,I}(f, g)$ is defined as in the r.h.s. of (7), but with the infimum taken over all bijections $\lambda : I \longrightarrow I$. The notions of J_2-convergence in all cases of I are derived as seen earlier for J_1.

5 Know results

We now recall some elementary functional limit theorems for suitable rescalings of the processes ω and S, cf. (4) and (5).

By definition, for all $k \in \mathbb{Z}$, ω_k is a sum of $|k|$ i.i.d. random variables ζ_i in the normal domain of attraction of a β-stable distribution. We first deal with the case $\beta \in (0, 1)$. For every $s \in \mathbb{R}$ we define

$$\hat{\omega}^{(n)}(s) := \begin{cases} \dfrac{\omega_{\lfloor ns \rfloor}}{n^{1/\beta}} & \text{if } s \geq 0, \\ \dfrac{\omega_{\lceil ns \rceil}}{n^{1/\beta}} & \text{if } s < 0. \end{cases} \tag{11}$$

Let $(Z_{\pm}^{(\beta)}(s),\ s \geq 0)$ be two i.i.d. càdlàg Lévy β-stable processes such that $Z_{\pm}^{(\beta)}(0) = 0$ and $Z_{\pm}^{(\beta)}(1)$ is distributed like $Z_1^{(\beta)}$, as introduced in (2) (these two conditions uniquely determine the common distribution of the processes). Set

$$Z^{(\beta)}(s) := \begin{cases} Z_+^{(\beta)}(s) & \text{if } s \geq 0, \\ -Z_-^{(\beta)}(-s) & \text{if } s < 0. \end{cases} \tag{12}$$

Then (see, e.g., [W2, Section 4.5.3]), as $n \to \infty$,

$$\hat{\omega}^{(n)} \xrightarrow{\mathrm{d}} Z^{(\beta)} \qquad \text{in } (\mathcal{D}, J_1). \tag{13}$$

When $\beta \in (1,2)$, the average distance $\nu := \mathbb{E}[\zeta_i]$ between successive targets is finite and positive by assumptions. So, at first order, a Strong Law of Large Numbers holds. More in detail, setting

$$\bar{\omega}^{(n)}(s) := \begin{cases} \dfrac{\omega_{\lfloor ns \rfloor}}{n} & \text{if } s \geq 0, \\ \dfrac{\omega_{\lceil ns \rceil}}{n} & \text{if } s < 0, \end{cases} \tag{14}$$

we have

$$\bar{\omega}^{(n)} \xrightarrow{\text{a.s.}} \nu \,\mathrm{id} \qquad \text{in } (\mathcal{D}, J_1). \tag{15}$$

as $n \to \infty$. Here and in the rest of the discussion id denotes the identity function, on whatever domain it is defined. Furthermore, a functional convergence similar to (13) holds for the fluctuations around this Law of Large Numbers. More explicitly, for $s \in \mathbb{R}$, define

$$\tilde{\omega}^{(n)}(s) := \begin{cases} \dfrac{\sum_{i=1}^{\lfloor ns \rfloor} (\zeta_i - \nu)}{n^{1/\beta}} & \text{if } s \geq 0, \\ \dfrac{-\sum_{i=\lceil ns \rceil}^{-1} (\zeta_i - \nu)}{n^{1/\beta}} & \text{if } s < 0. \end{cases} \tag{16}$$

Then, as $n \to \infty$,

$$\tilde{\omega}^{(n)} \xrightarrow{\text{d}} \tilde{Z}^{(\beta)} \qquad \text{in } (\mathcal{D}, J_1), \tag{17}$$

where the process $\tilde{Z}^{(\beta)}$ is defined similarly to $Z^{(\beta)}$, cf. (12), but with $\tilde{Z}^{(\beta)}_{\pm}(1)$ distributed like $\tilde{Z}^{(\beta)}_1$, introduced in (3).

Analogous limit theorems hold for the continuous-time rescaled versions of the underlying random walk S. By definition, S_n is a sum of n i.i.d. random variables ξ_i in the normal domain of attraction of an α-stable distribution. We distinguish two regimes, depending on the values of α and $\mu := \mathbb{E}[\xi_i]$ (when applicable).

The first regime corresponds to the cases $\alpha \in (0,1)$, or $\alpha \in (1,2)$ with $\mu = 0$. In these situations, the drift of the underlying random walk is either undefined or null. In either case, it does not affect the convergence of the process

$$\hat{S}^{(n)}(t) := \frac{S_{\lfloor nt \rfloor}}{n^{1/\alpha}}, \tag{18}$$

which we define for $t \geq 0$. In fact, let $W^{(\alpha)}$ denote a Lévy α-stable process with $W^{(\alpha)}(0) = 0$ and $W^{(\alpha)}(1)$ distributed like $W^{(\alpha)}_1$ (the latter variable has been defined after (5). Then, as $n \to \infty$,

$$\hat{S}^{(n)} \xrightarrow{\text{d}} W^{(\alpha)} \qquad \text{in } (\mathcal{D}^+, J_1). \tag{19}$$

When $\alpha \in (1,2)$ and $\mu \neq 0$, set, for $t \geq 0$,

$$\bar{S}^{(n)}(t) := \frac{S_{\lfloor nt \rfloor}}{n}. \tag{20}$$

By the functional version of the Strong Law of Large Numbers,

$$\bar{S}^{(n)} \xrightarrow{\text{a.s.}} \mu\,\mathrm{id} \qquad \text{in } (\mathcal{D}^+, J_1), \tag{21}$$

as $n \to \infty$. As for the fluctuations, defining

$$\tilde{S}^{(n)}(t) := \frac{\sum_{i=1}^{\lfloor nt \rfloor}(\xi_i - \mu)}{n^{1/\alpha}}, \tag{22}$$

we get

$$\tilde{S}^{(n)} \xrightarrow{\text{d}} \widetilde{W}^{(\alpha)} \qquad \text{in } (\mathcal{D}^+, J_1). \tag{23}$$

where $\widetilde{W}^{(\alpha)}$ is a Lévy α-stable process with $\widetilde{W}^{(\alpha)}(0) = 0$ and $\widetilde{W}^{(\alpha)}(1)$ distributed like $\widetilde{W}_1^{(\alpha)}$ (again defined after (5)).

5 Results

We now present our convergence results [SBBLM] for the Lévy flight Y which, as we shall see, strongly depend on the values of α and β. All theorems are stated using the notation established in the previous section.

We first analyze the case $\beta \in (0,1)$, corresponding to an infinite expected distance between the targets of the random medium

Theorem 5.1 *Let $\beta \in (0,1)$ and assume that either $\alpha \in (0,1)$ or $\alpha \in (1,2)$ with $\mu = 0$. For $t \in \mathbb{R}^+$ define*

$$\hat{Y}^{(n)}(t) := \hat{\omega}^{(n)} \circ \hat{S}^{(n)}(t) = \frac{Y_{\lfloor nt \rfloor}}{n^{1/\alpha\beta}}, \tag{24}$$

where $\hat{\omega}^{(n)}$ and $\hat{S}^{(n)}$ have been introduced, respectively, in (11) and (18). Then the finite-dimensional distributions of $\hat{Y}^{(n)}$ converge to those of $Z^{(\beta)} \circ W^{(\alpha)}$, i.e., for any $m \in \mathbb{Z}^+$ and $t_1, \dots, t_m \in \mathbb{R}^+$,

$$\left(\hat{Y}^{(n)}(t_1), \dots, \hat{Y}^{(n)}(t_m)\right) \xrightarrow{d} \left(Z^{(\beta)}(W^{(\alpha)}(t_1)), \dots, Z^{(\beta)}(W^{(\alpha)}(t_m))\right), \tag{25}$$

as $n \to \infty$.

Theorem 5.1 is rather weak, in that it only proves convergence of the finite-dimensional distributions of the process $\hat{Y}^{(n)}$ defined in (24). Observe, however, that the limit process $Z^{(\beta)} \circ W^{(\alpha)}$ has trajectories that are not càdlàg with positive probability (see for example the explanation around (2.9) of [BLP]). Therefore, a functional limit theorem w.r.t. a Skorokhod topology is not the natural result to expect. On the other hand, when $\alpha \in (1,2)$ and $\mu \neq 0$, the assertion can be strengthened as follows.

Theorem 5.2 *Let $\beta \in (0,1)$ and $\alpha \in (1,2)$ with $\mu \neq 0$. For $t \in \mathbb{R}^+$ define*

$$\hat{Y}^{(n)}(t) := \hat{\omega}^{(n)} \circ \bar{S}^{(n)}(t) = \frac{Y_{\lfloor nt \rfloor}}{n^{1/\beta}}, \tag{26}$$

cf. (11) and (20). Then, as $n \to \infty$,

$$\hat{Y}^{(n)} \xrightarrow{d} \mathrm{sgn}(\mu)\,|\mu|^{1/\beta}\, Z_+^{(\beta)} \qquad \text{in } (\mathcal{D}^+, J_2). \tag{27}$$

Remark 5.1 Since $Z_+^{(\beta)} \overset{\mathrm{d}}{=} Z_-^{(\beta)}$, one could put either process in the r.h.s. of (27), irrespectively of the sign of μ.

Remark 5.2 The convergence (27) fails in the topology J_1, or even M_1 [W2, Section 3.3]. The topology J_2 is thus the strongest among the classical Skorochod topologies with respect to which the convergence holds.

Next we consider the case $\beta \in (1,2)$, where the inter-distances of the random medium have finite mean.

Theorem 5.3 *Let* $\beta \in (1,2)$ *and recall the notation (14), (18) and (20).*

(a) *Assume* $\alpha \in (0,1)$, *or* $\alpha \in (1,2)$ *with* $\mu = 0$. *For* $t \in \mathbb{R}^+$ *set*

$$\hat{Y}^{(n)}(t) := \bar{\omega}^{(n)} \circ \hat{S}^{(n)}(t) = \frac{Y_{\lfloor nt \rfloor}}{n^{1/\alpha}}. \tag{28}$$

Then, as $n \to \infty$,

$$\hat{Y}^{(n)} \xrightarrow{d} \nu W^{(\alpha)} \qquad \text{in } (\mathcal{D}^+, J_1). \tag{29}$$

(b) *Assume* $\alpha \in (1,2)$ *and* $\mu \neq 0$. *Setting*

$$\bar{Y}^{(n)}(t) := \bar{\omega}^{(n)} \circ \bar{S}^{(n)}(t) = \frac{Y_{\lfloor nt \rfloor}}{n} \tag{30}$$

one has

$$\bar{Y}^{(n)} \xrightarrow{d} \nu\mu\,\mathrm{id} \qquad \text{in } (\mathcal{D}^+, J_1). \tag{31}$$

As stated in point 2 above, when $\alpha \in (1,2)$ and $\mu \neq 0$, the sequence of processes $\bar{Y}^{(n)}$ converges to a multiple of the identity function. The next theorem gives the explicit asymptotics of the fluctuations of $\bar{Y}^{(n)}$ around its deterministic limit.

Theorem 5.4 *Let* $\alpha, \beta \in (1,2)$ *with* $\mu \neq 0$, *and let* $\bar{Y}^{(n)}$ *be the process defined in (30).*

(a) *If* $\alpha < \beta$ *define*

$$\tilde{Y}^{(n)}(t) := \frac{n(\bar{Y}^{(n)}(t) - \nu\mu t)}{n^{1/\alpha}}. \tag{32}$$

Then, when $n \to \infty$,

$$\tilde{Y}^{(n)} \xrightarrow{d} \nu\widetilde{W}^{(\alpha)} \qquad \text{in } (\mathcal{D}^+, J_1), \tag{33}$$

where $\widetilde{W}^{(\alpha)}$ *has been defined after (23).*

(b) *If $\alpha > \beta$ define*

$$\tilde{Y}^{(n)}(t) := \frac{n(\bar{Y}^{(n)}(t) - \nu\mu t)}{n^{1/\beta}}. \tag{34}$$

Then, when $n \to \infty$,

$$\tilde{Y}^{(n)} \xrightarrow{d} \operatorname{sgn}(\mu)\,|\mu|^{1/\beta}\,\tilde{Z}_+^{(\beta)} \qquad \text{in } (\mathcal{D}^+, J_2), \tag{35}$$

where $\tilde{Z}_+^{(\beta)}$ has been defined after (17).

(c) *If $\alpha = \beta$ define*

$$\tilde{Y}^{(n)}(t) := \frac{n(\bar{Y}^{(n)}(t) - \nu\mu t)}{n^{1/\alpha}}. \tag{36}$$

Let $\tilde{Z}_+^{(\alpha)}$ and $\widetilde{W}^{(\alpha)}$ be two independent α-stable processes, as previously defined. As $n \to \infty$,

$$\tilde{Y}^{(n)} \xrightarrow{d} \operatorname{sgn}(\mu)\,|\mu|^{1/\beta}\,\tilde{Z}_+^{(\alpha)} + \nu\widetilde{W}^{(\alpha)} \qquad \text{in } (\mathcal{D}^+, J_2). \tag{37}$$

References

[ACOR] R. Artuso, G. Cristadoro, M. Onofri, M. Radice, *Non-homogeneous persistent random walks and Lévy-Lorentz gas.* J. Stat. Mech. Theory Exp. 2018, no. 8, 083209, 13 pp.

[BFK] E. Barkai, V. Fleurov, J. Klafter, *One-dimensional stochastic Lévy-Lorentz gas.* Phys. Rev. E 61 (2000), no. 2, 1164–116.

[BR] N. Berger, R. Rosenthal, *Random walks on discrete point processes.* Ann. Inst. Henri Poincaré Probab. Stat. 51 (2015), no. 2, 727–755.

[BCLL] A. Bianchi, G. Cristadoro, M. Lenci, M. Ligabò, *Random walks in a one-dimensional Lévy random environment.* J. Stat. Phys. 163 (2016), no. 1, 22–40.

[BLP] A. Bianchi, M. Lenci, F. Pène, *Continuous-time random walk between Lévy-spaced targets in the real line.* Stochastic Process. Appl. 130 (2020), no. 2, 708-732.

[SBBLM] S. Stivanello, G. Bet, A. Bianchi, M. Lenci, E. Magnanini, *Limit theorems for Lévy flights on a 1D Lévy random medium.* arXiv preprint arXiv:2007.03384.

[B] P. Billingsley, "Convergence of probability measures". John Wiley and Sons, Inc., New York-London-Sydney, 1968.

[BCV] R. Burioni, L. Caniparoli, A. Vezzani, *Lévy walks and scaling in quenched disordered media.* Phys. Rev. E 81 (2010), 060101(R), 4 pp.

[BDLV] R. Burioni, S. di Santo, S. Lepri, A. Vezzani, *Scattering lengths and universality in superdiffusive Lévy materials*. Phys. Rev. E 86 (2012), 031125, 7 pp.

[CF] P. Caputo, A. Faggionato, *Diffusivity in one-dimensional generalized Mott variable-range hopping models*. Ann. Appl. Probab. 19 (2009), no. 4, 1459–1494.

[CFP] P. Caputo, A. Faggionato, T. Prescott, *Invariance principle for Mott variable range hopping and other walks on point processes*. Ann. Inst. Henri Poincaré Probab. Stat. 49 (2013), no. 3, 654–697.

[CGLS] G. Cristadoro, T. Gilbert, M. Lenci, D. P. Sanders, *Transport properties of Lévy walks: an analysis in terms of multistate processes*. Europhys. Lett. 108 (2014), no. 5, 50002, 6 pp.

[EK] S. N. Ethier, T. G. Kurtz, "Markov processes. Characterization and convergence". John Wiley & Sons, Inc., New York, 1986.

[JS1987] J. Jacod, A. Shiryaev, "Limit theorems for stochastic processes". Springer-Verlag, Berlin, 1987.

[KRS] R. Klages, G. Radons, I. M. Sokolov (eds.), "Anomalous Transport: Foundations and Applications". Wiley-VCH, Berlin, 2008.

[K] N. Kubota, *Quenched invariance principle for simple random walk on discrete point processes*. Stochastic Process. Appl. 123 (2013), no. 10, 3737–3752.

[L] P. Levitz, *From Knudsen diffusion to Levy walks*. Europhys. Lett. 39 (1997), no. 6, 593–598.

[MS] M. Magdziarz, W. Szczotka, *Diffusion limit of Lévy-Lorentz gas is Brownian motion*. Commun. Nonlinear Sci. Numer. Simul. 60 (2018), 100–106.

[ROAC] M. Radice, M. Onofri, R. Artuso, G. Cristadoro, *Transport properties and ageing for the averaged Lévy-Lorentz gas*. J. Phys. A 53 (2020), no. 2, 025701, 16 pp.

[R] A. Rousselle, *Annealed invariance principle for random walks on random graphs generated by point processes in* $\mathbb{R}^d$. Markov Process. Related Fields 22 (2016), no. 4, 653–696.

[S] M. Schulz, *Lévy flights in a quenched jump length field: a real space renormalization group approach*. Phys. Lett. A 298 (2002), no. 2-3, 105–108.

[SZF] M. Shlesinger, G. Zaslavsky, U. Frisch (eds.), "Lévy Flights and Related Topics in Physics". Lecture Notes in Physics 450. Springer-Verlag, Berlin, 1995.

[W1] W. Whitt, *Some useful functions for functional limit theorems*. Math. Oper. Res. 5 (1980), no. 1, 67–85.

[W2] W. Whitt, "Stochastic-process limits. An introduction to stochastic-process limits and their application to queues". Springer-Verlag, New York, 2002.

[ZDK] V. Zaburdaev, S. Denisov, J. Klafter, *Lévy walks*. Rev. Mod. Phys. 87 (2015), 483–530.

[Z] L. Zhu, *Large deviations for one-dimensional random walks on discrete point processes*. Statist. Probab. Lett. 97 (2015), 69–75.

A differential eigenproblem of electromagnetics

Michele Zaccaron [(*)]

Abstract. We start with a preliminary overview of Sobolev spaces, very useful in the modern approach to solve PDEs. We then introduce an eigenvalue problem arising from time-harmonic Maxwell's equations, deriving its so-called weak formulation and showing some first properties. The last part of the seminar will focus on the domain perturbation, inspecting the dependence of the eigenvalues upon variation of the region in which we set our problem, and presenting a result regarding shape optimization. The talk is of introductory type and its purpose is to let the audience have a look at some of the tools and concepts from the spectral theory of differential operators.

1 Sobolev spaces

1.1 Lebesgue spaces

We start introducing the standard Lebesgue spaces of functions. Let Ω be an open set of $\mathbb{R}^n$. For $p \in [1, \infty)$ we define the space $L^p(\Omega)$ as the set of all Lebesgue-measurable functions such that the p-th power of their absolute value is Lebesgue integrable, that is

$$L^p(\Omega) := \{\, f : \Omega \to \mathbb{R} \text{ (or } \mathbb{C}) \text{ measurable} : \int_\Omega |\, f|^p dx < \infty \,\},$$

endowed with the norm $\|f\|_{L^p(\Omega)} = \|f\|_p := \left(\int_\Omega |f|^p dx\right)^{1/p}$. The space $L^p(\Omega)$ is a normed vector space, complete with respect to the metric obtained from its norm, thus a Banach space. In the case $p = 2$ we can define an inner product on $L^2(\Omega)$, namely

$$(f, g)_2 := \int_\Omega f g \, dx, \qquad f, g \in L^2(\Omega)$$

which makes it a Hilbert space. For $p = \infty$, with $L^\infty(\Omega)$ denotes the space of Lebesgue-measurable functions that have finite *essential supremum*, namely

$$L^\infty(\Omega) := \{\, f : \Omega \to \mathbb{R} \text{ (or } \mathbb{C}) \text{ measurable} : |\, f(x)| < \infty \text{ almost everywhere in } \Omega \,\}.$$

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: zaccaron@math.unipd.it . Seminar held on 3 March 2021.

The norm

$$\|f\|_\infty := \text{ess sup}_{x\in\Omega}\,|f(x)| = \inf\{\, C \geq 0 : |\, f(x)| \leq C \text{ for almost every } x \in \Omega\,\}$$

makes it a Banach space. We also introduce the space of locally integrable functions, that is

$$L^p_{loc}(\Omega) := \{\, f : \Omega \to \mathbb{R} \text{ (or } \mathbb{C}) \text{ measurable} : f \in L^p(K) \text{ for all } K \subset \Omega \text{ compact}\,\}\,.$$

Notice that $L^p(\Omega) \subset L^1_{loc}(\Omega)$ for all $p \in [1,\infty]$. In our setting, two functions equal almost everywhere (that is, they differ only on a set of Lebesgue measure zero) are equal. In the following we will sometimes recall the fact that we are working not with functions themselves, but with their equivalent classes w.r.t. the equivalence relation of being equal except a set of zero measure, writing "a.e.", which stands for "almost everywhere".

1.2 Weak derivation

If we consider a C^1 function $f : \mathbb{R} \to \mathbb{R}$, then integration by parts (Gauss theorem) tells us that

$$\int_{\mathbb{R}} f(x)\varphi'(x)\,dx = -\int_{\mathbb{R}} f'(x)\varphi(x)\,dx \quad \text{for all } \varphi \in C^\infty_c(\mathbb{R}).$$

In order to make sense of the left integral above, we just need to assume that $f \in L^1_{loc}(\mathbb{R})$. At this point we ask if there exist a function $g \in L^1_{loc}(\mathbb{R})$ such that

$$\int_{\mathbb{R}} f(x)\varphi'(x)\,dx = -\int_{\mathbb{R}} g(x)\varphi(x)\,dx \quad \text{for all } \varphi \in C^\infty_c(\mathbb{R}).$$

If the answer is yes, we then say that f admits a weak derivative (in $\mathbb{R}$) and the weak derivative of f is g, which we will denote with f'_w.

We will now generalize the concept. Let Ω be an open set of $\mathbb{R}^n$ and φ a smooth function on Ω. If $\alpha = (\alpha_1,\dots,\alpha_n) \in \mathbb{N}^n$ is a multi-index, then $|\alpha| = \alpha_1 + \ldots + \alpha_n$ and

$$D^\alpha\varphi = \frac{\partial^{\,|\alpha|}\varphi}{\partial x_1^{\alpha_1}\cdots\partial x_n^{\alpha_n}}.$$

If for $f \in L^1_{loc}(\Omega)$ there exist a $g \in L^1_{loc}(\Omega)$ such that

$$\int_\Omega f\, D^\alpha\varphi\,dx = (-1)^{\,|\alpha|}\int_\Omega g\,\varphi\,dx \quad \text{for all } \varphi \in C^\infty_c(\Omega),$$

then we call $g = D^\alpha_w f$ the *weak α-th derivative* of f. Here and in the following we will omit the subscript w and we will just write $D^\alpha f$ when referring to weak derivatives. Some properties of the weak derivative include:

(i) If the weak derivative of a function exists, it is uniquely determined a.e.

(ii) If $f \in C^k(\Omega)$ then the classical and the weak derivatives coincide.

(iii) If all first partial weak derivatives are 0 then the function is constant a.e.

(iv) $D^\alpha(D^\beta f) = D^\beta(D^\alpha f) = D^{\alpha+\beta} f$.

One can easily prove the uniqueness almost everywhere using the following important result.

Theorem 1 (Fundamental Lemma of Calculus of Variations) *Let $h \in L^1_{loc}(\Omega)$. If*

$$\int_\Omega h\,\varphi\,dx = 0 \quad \textit{for all } \varphi \in C_c^\infty(\Omega)$$

then $h = 0$ (a.e.).

For many other properties of weak derivatives and for a complete and detailed introduction to Sobolev spaces we refer to [2].

2 Sobolev spaces

Having introduced the notion of weak derivatives, we can introduce the so-called Sobolev spaces. For any $k \in \mathbb{N}$ and $1 \le p \le \infty$ set

$$W^{k,p}(\Omega) := \{\, f \in L^p(\Omega) : D^\alpha f \in L^p(\Omega) \text{ for all } \ |\,\alpha| \le k \,\}, \qquad W^{0,p}(\Omega) := L^p(\Omega).$$

Endowed with the norm $\|f\|_{W^{k,p}(\Omega)} = \left(\sum_{|\alpha|\le k} \|D^\alpha u\|_p^p\right)^{1/p}$ if $1 \le p < \infty$, and $\|u\|_{W^{k,\infty}(\Omega)} = \max_{|\alpha|\le k} \|D^\alpha u\|_\infty$ if $p = \infty$, it becomes a Banach space. Again, the case $p = 2$ has additional structure and properties due to its "duality". Indeed, endowing $H^k(\Omega) := W^{k,2}(\Omega)$ with the following inner product

$$(u,v)_{H^k(\Omega)} = \sum_{|\alpha|\le k} \int_\Omega D^\alpha u\, D^\alpha v\, dx$$

it becomes a Hilbert space. Observe that the Hilbert norm $\|u\| := (u,u)^{1/2}_{H^k(\Omega)}$ is equivalent to the standard one.

Remark 1 Why this notion of weak derivative? It was first introduced when dealing with PDEs: mathematicians noticed that the most important quantities of the PDE solutions were embodied in the integrability of the functions.

Motivated by the study of classical differential equations, arising especially from mathematical models describing physical phenomena, together with the introduction of the so-called weak (or variational) formulation, in the past century the approach to differential equations changed and it was observed that the classical spaces C^0, C^1, C^2 etc. were not exactly the right spaces in which set and solve PDE problems. The Sobolev spaces are their modern replacement. Indeed, quantities or properties of the underlying mathematical model are usually expressed in terms of integral norms, which correspond to "energies" relevant to the physical phenomena the PDE models.

Note that Sobolev spaces are not the only spaces one can use. There are many type of function spaces (Orlicz spaces, Hardy spaces, Morrey–Campanato spaces, Hölder spaces, Besov spaces etc.): one may choose the space that suits in the best way possible the differential equation and the goals of its study.

2 Maxwell equations

2.1 Introduction to the eigenproblem

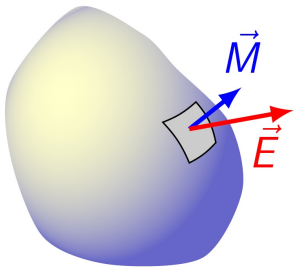
Recall the divergence and the curl (or rotor) of a vector field $F = (F_1, F_2, F_3)$ in $\mathbb{R}^3$:

$$\operatorname{div} F = \partial_1 F_1 + \partial_2 F_2 + \partial_3 F_3, \qquad \operatorname{curl} F = \begin{pmatrix} \partial_1 \\ \partial_2 \\ \partial_3 \end{pmatrix} \times \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix} = \begin{pmatrix} \partial_2 F_3 - \partial_3 F_2 \\ -\partial_1 F_3 + \partial_3 F_1 \\ \partial_1 F_2 - \partial_2 F_1 \end{pmatrix}$$

The *time-harmonic* Maxwell's equations in a medium filling a region Ω of $\mathbb{R}^3$ read:

$$\operatorname{curl} E - i\omega\mu M = 0, \quad \operatorname{curl} M + i\omega\varepsilon E = 0, \tag{1}$$

where E, M are respectively the electric field and magnetic field, ε and μ are the electric permeability and the magnetic permeability of the medium, and $\omega > 0$ is the angular frequency.



If the medium is isotropic and homogeneous, ε and μ are constants, and thus we can normalize them by setting $\varepsilon = \mu = 1$. We can pair this differential coupled system with the perfect conductor boundary conditions, namely

$$E \times \nu = 0, \quad M \cdot \nu = 0 \qquad \text{on } \partial\Omega.$$

Here ν denotes the outer unit normal to the boundary of Ω. We will assume Ω to be bounded and with Lipschitz boundary, hence ν is defined almost everywhere on $\partial\Omega$. Note that the solutions E, M are both divergence-free, because they are the curls. We refer to [12] for more details about time-harmonic Maxwell's equations.

Operating by curl in each of the equations (1) and setting $\lambda = \omega^2$, one obtains the following two boundary-value problems:

$$\begin{cases} \operatorname{curl}\operatorname{curl} E = \lambda E, & \text{in } \Omega, \\ \operatorname{div} E = 0, & \text{in } \Omega, \\ E \times \nu = 0, & \text{on } \partial\Omega, \\ \operatorname{curl} E \cdot \nu = 0, & \text{on } \partial\Omega, \end{cases} \tag{2}$$

and

$$\begin{cases} \operatorname{curl}\operatorname{curl} M = \lambda M, & \text{in } \Omega, \\ \operatorname{div} M = 0, & \text{in } \Omega, \\ M \cdot \nu = 0, & \text{on } \partial\Omega, \\ \operatorname{curl} M \times \nu = 0, & \text{on } \partial\Omega. \end{cases} \tag{3}$$

As one would expect, each of the two problems provide the same eigenvalues (cf., [19]). We will focus on the first one, meaning we will study the existence of a couple (E, λ) with $E : \Omega \to \mathbb{R}^3$ a vector field and $\lambda \in \mathbb{R}$ such that the following *electric problem* holds

$$(\mathcal{E}) \quad \begin{cases} \operatorname{curl}\operatorname{curl} E = \lambda E & \text{on } \Omega, \\ \operatorname{div} E = 0 & \text{on } \Omega, \\ E \times \nu = 0 & \text{on } \partial\,\Omega. \end{cases} \tag{4}$$

Note that a solution E to (4) automatically satisfies the boundary condition $\operatorname{curl} E \cdot \nu = 0$ on $\partial\Omega$ (see [14, Lemma 2.3]), thus we omitted it. Before proceeding any further, we recall a useful integration formula.

Lemma 1 (Green's formula) *Let Ω be a bounded Lipschitz set in $\mathbb{R}^3$. Then*

$$\int_\Omega \operatorname{curl} F \cdot G \, dx = \int_\Omega F \cdot G \, dx + \int_{\partial\Omega} (\nu \times F) \cdot G \, d\sigma$$

for any $F, G \in C^1(\overline{\Omega})^3$.

Multiplying both sides of the equation $\operatorname{curl}\operatorname{curl} E = \lambda E$ by a vector field F and integrating over Ω we get

$$\int_\Omega \operatorname{curl}\operatorname{curl} E \cdot F \, dx = \lambda \int_\Omega E \cdot F \, dx.$$

By Lemma 1 we have that

$$\int_\Omega \operatorname{curl}\operatorname{curl} E \cdot F \, dx = \int_\Omega \operatorname{curl} E \cdot \operatorname{curl} F \, dx + \int_{\partial\Omega} \operatorname{curl} E \cdot (F \times \nu) \, d\sigma.$$

Hence, supposing that also F satisfies the electric boundary condition, i.e. $F \times \nu = 0$ on $\partial\Omega$, we arrive at

$$\int_\Omega \operatorname{curl} E \cdot \operatorname{curl} F \, dx = \lambda \int_\Omega E \cdot F \, dx. \tag{5}$$

Setting $F = E$, we see that necessarily $\lambda \geq 0$.

2.2 The appropriate Hilbert space

Before stating the weak formulation of the electric problem $(\mathcal{E})$ we will introduce the appropriate underlying Hilbert space.

Definition 1 A vector field $u \in L^2(\Omega)^3$ possesses a weak curl in $L^2(\Omega)$ if there exist a vector field $w \in L^2(\Omega)^3$ such that

$$\int_\Omega u \cdot \operatorname{curl} \psi \, dx = \int_\Omega w \cdot \psi \, dx \qquad \text{for all vector fields } \psi \in C_c^\infty(\Omega)^3.$$

Similarly, it possesses a weak divergence in $L^2(\Omega)$ if there exists a scalar function $p \in L^2(\Omega)$ such that

$$\int_\Omega u \cdot \nabla\varphi \, dx = -\int_\Omega p\,\varphi \, dx \qquad \text{for all } \varphi \in C_c^\infty(\Omega).$$

Define the following Hilbert spaces:

$$H(\mathrm{curl},\Omega) = \{\, u \in L^2(\Omega)^3 : u \text{ has weak curl in } L^2(\Omega)^3 \,\}$$

equipped with the inner product

$$(u,v)_{H(\mathrm{curl},\Omega)} := \int_\Omega u \cdot v \, dx + \int_\Omega \mathrm{curl}\, u \cdot \mathrm{curl}\, v \, dx,$$

and

$$H(\mathrm{div},\Omega) = \{\, u \in L^2(\Omega)^3 : u \text{ has weak divergence in } L^2(\Omega)^3 \,\},$$

with inner product

$$(u,v)_{H(\mathrm{div},\Omega)} := \int_\Omega u \cdot v \, dx + \int_\Omega \mathrm{div}\, u \ \mathrm{div}\, v \, dx.$$

We consider the intersection $H(\mathrm{curl},\Omega) \cap H(\mathrm{div},\Omega)$, but observe that we still lack the electric boundary condition $u \times \nu = 0$ on $\partial\Omega$. The condition requires precisely that the tangential part of u must be null at the boundary. But we first need to clarify what do we mean when we say that the tangential component of the vector field u is zero on the boundary: in general u is defined only a.e., and thus computing its values on a set of Lebesgue measure zero such as $\partial\Omega$ does not make sense, or at least it is not clearly defined yet. In order to make this clear, we introduce the *tangential trace* operator γ_τ, from $C^\infty(\overline{\Omega})^3$ to $L^2(\partial\Omega)^3$, defined as

$$\gamma_\tau \psi = \psi \times \nu|_{\partial\Omega}, \qquad \psi \in C^\infty(\overline{\Omega})^3.$$

Then

Theorem 2 [7, Thm. 2.11] *The tangential trace mapping γ_τ can be extended by continuity to a linear and continuous map, still denoted by γ_τ, from $H(\mathrm{curl},\Omega)$ into $H^{-1/2}(\partial\Omega)^3$. Moreover, the following Green's formula holds: if $u \in H(\mathrm{curl},\Omega)$ then*

$$\int_\Omega \mathrm{curl}\, u \cdot \varphi \, dx - \int_\Omega u \cdot \mathrm{curl}\, \varphi \, dx = \int_{\partial\Omega} \gamma_\tau u \cdot \varphi \, d\sigma \tag{6}$$

for all $\varphi \in H^1(\Omega)^3$.

Observe that formula (6) is nothing but the generalization of the Green's formula introduced in Lemma 1.

We now define

$$H_0(\mathrm{curl},\Omega) := \mathrm{Ker}\, \gamma_\tau \subset H(\mathrm{curl},\Omega).$$

In particular, for any $u \in H_0(\mathrm{curl},\Omega)$ we have that

$$\int_\Omega \mathrm{curl}\, u \cdot \varphi \, dx = \int_\Omega u \cdot \mathrm{curl}\, \varphi \, dx$$

holds for all $\varphi \in H^1(\Omega)^3$. It turns out that $H_0(\mathrm{curl}, \Omega)$ is the closure in $H(\mathrm{curl}, \Omega)$ of smooth vector fields compactly supported in Ω, that is

$$H_0(\mathrm{curl}, \Omega) = \overline{C_c^\infty(\Omega)^3}^{H(\mathrm{curl},(\Omega))}.$$

The space $H_0(\mathrm{curl}, \Omega)$ incorporates the correct notion of electric boundary condition of having zero tangential trace.

Setting

$$X_N(\Omega) := H_0(\mathrm{curl}, \Omega) \cap H(\mathrm{div}, \Omega)$$

we equip it with the inner product

$$\int_\Omega u \cdot v \, dx + \int_\Omega \mathrm{curl}\, u \cdot \mathrm{curl}\, v \, dx + \int_\Omega \mathrm{div}\, u \ \mathrm{div}\, v \, dx.$$

Finally, we take the subspace of $X_N(\Omega)$ corresponding to divergence-free vector fields, namely

$$X_N(\mathrm{div}\, 0, \Omega) = \{\, u \in X_N(\Omega) : \mathrm{div}\, u = 0 \,\}.$$

The space $X_N(\mathrm{div}\, 0, \Omega)$ will be the appropriate Hilbert space in which we set the study of the electric eigenproblem $(\mathcal{E})$ and derive its weak formulation.

2.3 Weak formulation and some properties

To simplify the notation, we denote $H_\mathcal{E} := X_N(\mathrm{div}\, 0, \Omega)$. Duplicating the same argument that led us to (5), we derive the weak form of problem $(\mathcal{E})$:

$$\int_\Omega \mathrm{curl}\, E \cdot \mathrm{curl}\, F \, dx = \lambda \int_\Omega E \cdot F \, dx \quad \text{for all } F \in H_\mathcal{E} \tag{7}$$

in the unknowns $\lambda \geq 0$ the eigenvalue and $E \in H_\mathcal{E}$. From now on when we refer to the problem $(\mathcal{E})$ we will have in mind its weak formulation (7).

For the sake of simplicity, we assume Ω to be a bounded simply connected domain of $\mathbb{R}^3$ of class C^2, although these hypothesis can be weakened, considering general bounded domains of class $C^{1,1}$.

Proposition 1 *The zero eigenspace is*

$$\mathrm{Ker}_\mathcal{E} = \{\, E = \nabla\rho : \Delta\rho = 0,\ \rho \text{is constant on the every c.c. of the boundary } \partial\Omega \,\} \subset H_\mathcal{E}.$$

Its dimension is

$$\dim \mathrm{Ker}_\mathcal{E} = \#(\text{connected components of } \partial\Omega) - 1.$$

Sketch of the proof. Suppose E is an eigenfield corresponding to the eigenvalue 0. Then $\mathrm{curl}\, E = 0$, as one can see by setting $F = E$ in (7). Hence $E = \nabla\eta$ for some $\eta \in H^1(\Omega)$ (see e.g. Thm. 2.9 of [7]). Moreover, the divergence-free condition that E must satisfy implies

that η is a harmonic function in Ω. Finally, from the boundary condition $E\times\nu = \nabla\eta\times\nu = 0$ we have that $\nabla\eta$ is parallel to the normal vector ν at any boundary point. Thus η is constant in any connected component of $\partial\Omega$.

The reverse inclusion is simpler, while the dimension formula derives from arguments regarding the maximum principle for harmonic functions. □

Denoting with $(H_{\mathcal{E}})^*$ the (topological) dual space of the Hilbert space $H_{\mathcal{E}}$, we introduce the following continuous operators:

$$J : L^2(\Omega)^3 \to (H_{\mathcal{E}})^*, \qquad J[u](v) := \int_\Omega u\cdot v\,dx \quad \text{for all } u\in L^2(\Omega)^3, v\in H_{\mathcal{E}},$$

$$B : H_{\mathcal{E}} \to (H_{\mathcal{E}})^*, \qquad B[u](v) := \int_\Omega u\cdot v\,dx + \int_\Omega \operatorname{curl} u\cdot\operatorname{curl} v\,dx \quad \text{for all } u,v\in H_{\mathcal{E}}.$$

The operator B is coercive on $H_{\mathcal{E}}$ (it corresponds to its inner product) and thus it can be inverted thanks to the Riesz representation theorem. Hence we can construct the following injective continuous linear operator from $H_{\mathcal{E}}$ to itself

$$\mathcal{L} : H_{\mathcal{E}} \overset{\iota}{\hookrightarrow} L^2(\Omega)^3 \overset{J}{\to} (H_{\mathcal{E}})' \xrightarrow{B^{-1}} H_{\mathcal{E}}, \qquad \mathcal{L} : B^{-1}\circ J\circ\iota,$$

where ι denotes the embedding of $H_{\mathcal{E}}$ into $L^2(\Omega)^3$. Since the embedding ι is compact (see [17]) and B is symmetric in the sense that exchanging u with v we obtain the same result, it is not difficult to show that $\mathcal{L}$ is a self-adjoint compact operator, with trivial kernel. Moreover

Proposition 2 (cf., Lemma 2.15 of [14]) *Let $E\in H_{\mathcal{E}}$. Then (λ, E), $\lambda\neq 0$, is a (weak) eigenpair of*

$$(\mathcal{E})\begin{cases} \operatorname{curl}\operatorname{curl} E = \lambda E & \text{on } \Omega,\\ \operatorname{div} E = 0 & \text{on } \Omega,\\ E\times\nu = 0 & \text{on } \partial\Omega. \end{cases}$$

if and only if

$$\mathcal{L}[E] = (\lambda+1)^{-1}E.$$

From the Spectral theorem (see Theorem 4 in the Appendix) we deduce that spectrum of the operator $\mathcal{L}$ is discrete, and the eigenvalues of $\mathcal{L}$ form an infinite sequence of positive eigenvalues converging to 0. Therefore, by Proposition 2, the positive eigenvalues of the electric problem $(\mathcal{E})$ form an increasing, divergent sequence

$$\lambda_1 \le \lambda_2 \le \cdots \le \lambda_j \le \ldots \uparrow +\infty, \tag{8}$$

where each eigenvalue is repeated according to its multiplicity (that is, the dimension of the corresponding eigenspace), which is finite. Finally, each positive eigenvalue can be represented by means of the following min-max formula:

$$\lambda_j = \min_{\substack{V\subset(\ker_{\mathcal{E}})^\perp\\ \dim V = j}} \max_{F\in V\setminus\{0\}} \frac{\int_\Omega |\operatorname{curl} F|^2}{\int_\Omega |F|^2}.$$

2.4 Extremum problems

There is a certain interest, not only in mathematics but also in more applied fields such as engineering and physics, in the study of the behaviour of the eigenvalues and eigenfunctions upon the parameters which enter the equations. For example the *shape* of the region, that is the set Ω in which we set the problem.

Motivated by the question "*are there optimal shapes for the eigenvalues?* ", we are interested in studying the map:

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

One may wonder whether this mapping is continuous, differentiable, or even analytic. Moreover, it is certainly interesting to pose the following constrained problems:

$$\min_{\mathrm{Vol}[\Omega]=const.} \lambda_j[\Omega] \quad \text{or} \quad \max_{\mathrm{Vol}[\Omega]=const.} \lambda_j[\Omega],$$

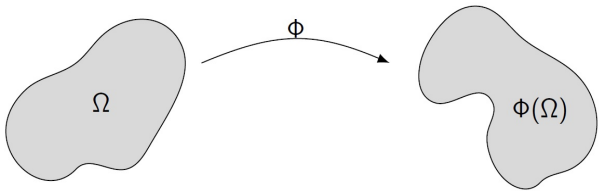
$$\min_{\mathrm{Per}[\Omega]=const.} \lambda_j[\Omega] \quad \text{or} \quad \max_{\mathrm{Per}[\Omega]=const.} \lambda_j[\Omega].$$

We wonder whether there are *critical* sets Ω for the minimization/maximization problems above, subject to fixed volume or perimeter constraint, and if such sets can be characterized. It is worthy to observe that these constraints allow the problem to be well-posed, making the problem "compact", in some sense. If we didn't have any constraint, we cannot expect to have neither maxima nor minima.

To answer these questions we proceed as follows. We fix a bounded domain $\Omega \subset \mathbb{R}^3$ of class C^2, and we introduce the set of admissible perturbations

$$\mathcal{A}_\Omega = \{\, \Phi \in C^2(\overline{\Omega}; \mathbb{R}^3) : \Phi \text{ is injective}, \det D\Phi(x) \neq 0 \ \forall x \in \overline{\Omega} \,\}.$$

Observe that $\mathcal{A}_\Omega$ is open in the topology of $C^2(\overline{\Omega}; \mathbb{R}^3)$. We then perturb the initial domain Ω by means of a diffeomorphism $\Phi \in \mathcal{A}_\Omega$,



and study the electric problem $\Phi(\mathcal{E})$ on the new perturbed domain $\Phi(\Omega)$, namely

$$\Phi(\mathcal{E}) \quad \begin{cases} \operatorname{curl}\operatorname{curl} E = \lambda E, & \text{on } \Phi(\Omega), \\ \operatorname{div} E = 0, & \text{on } \Phi(\Omega), \\ E \times \nu = 0, & \text{on } \partial\Phi(\Omega). \end{cases} \tag{9}$$

With a similar argument used previously to show (8), one can see that the positive eigenvalues (which depend on Φ) of the electric problem $\Phi(\mathcal{E})$ form an increasing, divergent sequence

$$\lambda_1[\Phi] \leq \lambda_2[\Phi] \leq \cdots \leq \lambda_j[\Phi] \leq \ldots \uparrow +\infty,$$

where each eigenvalue is repeated according to its multiplicity.

The quest of finding critical shapes leads us to study the behaviour of the maps

$$\mathcal{A}_\Omega \subset C^2(\overline{\Omega};\mathbb{R}^3) \to \mathbb{R}, \quad \Phi \mapsto \lambda_j[\Phi].$$

If our goal is to find perturbations Φ that are critical for the eigenvalues, we would like to compute their differentials. But there is a first significant problem: other than continuity, we cannot expect these maps to depend in a differentiable manner on the perturbations $\Phi \in \mathcal{A}_\Omega$. Indeed, apart from the case of a one-parameter family of perturbations $\{\Phi_\varepsilon\}_\varepsilon$, with the parameter $\varepsilon \in I \subset \mathbb{R}$ varying in an interval of the real line (which is in any case very interesting, and far from simple), the differentiability fails in general (see the example in the Appendix). For this reason, we will instead focus on the so-called symmetric functions of the eigenvalues, and, at the price of restricting the set of admissible perturbations, we will not only gain differentiability, but also analyticity.

Given a finite set of indices $S \subset \mathbb{N}$, we set

$$\Lambda_{S,h}[\Phi] := \sum_{\substack{j_1,\dots,j_h \in S \\ j_1<\dots<j_h}} \lambda_{j_1}[\Phi]\cdots\lambda_{j_h}[\Phi], \qquad h = 1,\dots,|S|. \tag{10}$$

We would like the symmetric functions $\Lambda_{S,h}[\Phi]$ of the eigenvalues to depend in a differentiable manner w.r.t. the perturbations Φ, in order to deal with extremum problems of this type:

$$\min_{\mathrm{Vol}[\Phi(\Omega)]=const.} \Lambda_{S,h}[\Phi], \qquad \max_{\mathrm{Vol}[\Phi(\Omega)]=const.} \Lambda_{S,h}[\Phi],$$

$$\min_{\mathrm{Per}[\Phi(\Omega)]=const.} \Lambda_{S,h}[\Omega], \qquad \max_{\mathrm{Per}[\Phi(\Omega)]=const.} \Lambda_{S,h}[\Omega].$$

To have that, we need to restrict the admissible perturbations to

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

which is nothing but the set of admissible diffeomorphisms that fix a finite set of eigenvalues and their multiplicities. Then $\mathcal{A}_\Omega[S]$ is an open set $C^2(\overline{\Omega};\mathbb{R}^3)$, and the functions $\Lambda_{S,h}[\Phi]$ depend real-analytically on $\Phi \in \mathcal{A}_\Omega[S]$ (see [14, Thm. 4.5], where the Fréchet differentials are computed as well).

Set now

$$\mathcal{V}(\alpha) := \{\, \Phi \in \mathcal{A}_\Omega : \mathrm{Vol}(\Phi(\Omega)) = \alpha \,\}$$

and

$$\mathcal{P}(\alpha) := \{\, \Phi \in \mathcal{A}_\Omega : \mathrm{Per}(\Phi(\Omega)) = \alpha \,\}.$$

Theorem 3 [14, Thm. 5.13] *Let Ω be a bounded domain in $\mathbb{R}^3$ of class C^2. Let $\Phi \in \mathcal{A}_\Omega$ be such that $\Phi(\Omega)$ is a ball. Let λ be an eigenvalue of the electric problem $\Phi(\mathcal{E})$, and let S be the set of $j \in \mathbb{N}$ such that $\lambda_j[\Phi] = \lambda$. Then Φ is both a critical point for $\Lambda_{S,h}$ with volume constraint $\mathcal{V}[\mathrm{Vol}(\Phi(\Omega))]$ and a critical point for $\Lambda_{S,h}$ with perimeter constraint $\mathcal{P}[\mathrm{Per}(\Phi(\Omega))]$, for all $h = 1,\dots,|S|$.*

The previous theorem affirms that "balls are critical domains for the symmetric functions on the eigenvalues under the constraint of fixed volume/fixed perimeter". Its proof is based on Thm. 5.10 of [14], which gives a method to check whether a perturbation Φ is a critical point or not for the symmetric functions of the eigenvalues, in the form of a necessary and sufficient condition. This, in turn, partially answers the question regarding a characterization of critical shapes addressed earlier.

3 Appendix

Throughout this read, we used some notations that may not be familiar with the reader, and this is why here we try to be more precise.

With $C^k(\overline{\Omega}) = C(\overline{\Omega}) \cap C^k(\Omega)$ we denote the set of functions in Ω continuous up to the boundary, which are also k-times continuously differentiable in Ω.

With $C^k(\Omega)^3 = C^k(\Omega; \mathbb{R}^3)$ we denote the set of real vector fields from Ω to $\mathbb{R}^3$, which are k-times continuously differentiable in Ω.

With $H^{1/2}(\partial\Omega)$ we denote the range of the trace operator

$$\mathrm{tr} : H^1(\Omega) \to L^2(\partial\Omega).$$

For more details about the trace operator we refer to chapter 5 of [3].

With $H^{-1/2}(\partial\Omega)$ we denote the topological dual of $H^{1/2}(\partial\Omega)$.

The following theorem is a one of the fundamental results of functional analysis and spectral theory. Among the many versions, we present the one that we need.

Theorem 4 (Spectral theorem) *Let $\mathcal{L} : H \to H$ be a compact self-adjoint linear operator on a Hilbert space H. Then there is a finite or infinite sequence $\{\mu_n\}_{n=1}^m$ (i.e. $m \in \mathbb{N}$ or $m = \infty$) of real positive eigenvalues with finite multiplicity and a corresponding orthonormal sequence $\{e_n\}_{n=1}^m$ in H such that:*

- $\mathcal{L}\, e_n = \mu_n e_n$ *for all* $n \in \{1, \dots, m\}$;
- $\mathrm{Ker}(\mathcal{L}) = Span(\{e_n\}_{n=1}^m)^\perp$;
- *if* $m = \infty$, *then* $\lim_{n\to\infty} \mu_n = 0$.

It is the infinite-dimensional generalization of the classical spectral theorem of linear algebra, concerning the diagonalization of a symmetric matrix.

3.1 Example

Why do we deal the symmetric functions of the eigenvalues instead of each single eigenvalue? The problem is that, even if there are nice results on the differentiability of the eigenvalues when the dependance is w.r.t. a one-parametric family of perturbations, when we pass to two or more parameters, the differentiability is lost, even in finite dimensional spaces. This will be the setting of the next example.

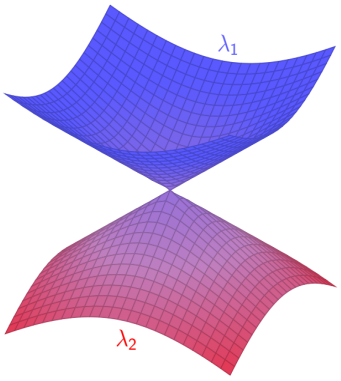
Consider the matrix

$$A(t,r) = \begin{pmatrix} t & r \\ r & -t \end{pmatrix}$$

depending on two parameters $t, r \in \mathbb{R}$. It is a linear operator acting from $\mathbb{R}^2$ to itself. Its eigenvalue are easily computed:

$$\lambda_1[t,r] = \sqrt{t^2+r^2}, \quad \lambda_2[t,r] = -\sqrt{t^2+r^2}.$$

As one can clearly see, the eigenvalues depend continuously on the parameters, but at the point $(t,r) = (0,0)$ they fail to be differentiable.



Taking instead the symmetric functions of the eigenvalues, i.e.

$$\lambda_1[t,r] + \lambda_2[t,r] = 0, \quad \lambda_1[t,r]\lambda_2[t,r] = -t^2 - r^2,$$

we recover the differentiability. The dependence on the parameters t and r is actually analytical. To see more about this subject the reader can take a look at chapter 2 of [11].

References

[1] Alberti, G.S., Capdeboscq, Y., *Elliptic regularity theory applied to time harmonic anisotropic Maxwell's equations with less than Lipschitz complex coefficients.* SIAM J. Math. Anal. 46 (2014), no. 1, 998–1016.

[2] Brezis, H., "Functional analysis, Sobolev spaces and partial differential equations". Springer, New York, 2011.

[3] Burenkov, V.I., "Sobolev Spaces on Domains". Teubner-Texte zur Mathematik, 137. B. G. Teubner Verlagsgesellschaft mbH, Stuttgart, 1998.

[4] Costabel, M., *A coercive bilinear form for Maxwell's equations.* J. Math. Anal. Appl. 157 (1991), no. 2, 527–541.

[5] Costabel, M., Dauge, M., *Maxwell and Lamé eigenvalues on polyhedra.* Math. Methods Appl. Sci. 22 (1999), 243–258.

[6] Dautray, R., Lions, J.-L., "Mathematical Analysis and Numerical Methods for Science and Technology: Vol. 3, Spectral Theory and Applications". Springer-Verlag, Berlin, 1990.

[7] Girault, V., Raviart P.-A., "Finite Element Approximation of the Navier-Stokes Equations". Lecture Notes in Mathematics - No. 749, Springer, Berlin, 1981.

[8] Hanson, G.W., Yakovlev, A.B., "Operator Theory for Electromagnetics". Springer-Verlag New York, 2002.

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

[10] Jimbo, S., "Hadamard variation for electromagnetic frequencies". Geometric properties for parabolic and elliptic PDEs, 179–199, Springer INdAM Ser., 2, Springer,Milan, 2013.

[11] Kato, T., "Pertubation Theory for Linear Operators". Reprint of the 1980 edition. Classics in Mathematics, Springer-Verlag, Berlin, 1995.

[12] Kirsch, A., Hettlich, F., "The Mathematical Theory of Time-Harmonic Maxwell's Equations, Expansion, Integral, and Variational Methods". Applied Mathematical Sciences - Vol. 190, Springer International Publishing, Cham, Switzerland, 2015.

[13] Lamberti, P.D., Lanza de Cristoforis, M., *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) 19–42.

[14] Lamberti, P.D., Zaccaron, M., *Shape sensitivity analysis for electromagnetic cavities.* Mathematical Methods in the Applied Sciences, to appear (2021).

[15] Lanza de Cristoforis, M., *Higher order differentiability properties of the composition and of the inversion operator.* Indag. Math. (N.S.) 5 (1994), no. 4, 457–482.

[16] Monk, P., "Finite Element Methods for Maxwell's Equations". Clarendon Press, Oxford, 2003.

[17] Weber, C., *A local compactness theorem for Maxwell's equations.* Math. Methods Appl. Sci. 2 (1980), 12–25.

[18] Yin, H.M., *An eigenvalue problem for curlcurl operators.* Can. Appl. Math. Q. 20 (2012), no. 3, 421-434.

[19] Zhang, Z., *Comparison results for eigenvalues of curlcurl operator and Stokes operator.* Z. Angew. Math. Phys. (2018), 69–104.

Optimal stopping theory and American options

Francesco Rotondi [(*)]

Abstract. The optimal stopping problem is a classical one within stochastic calculus theory. Formally, given a gain process, the optimal stopping problem is about finding the stopping time that maximizes the expected gain. Optimal stopping theory closely relates to the American derivatives valuation problem. American derivatives are financial contracts characterized by a payoff process that depends on an underlying stochastic process (usually the price of a traded asset). The holder of an American derivative chooses when to cash in the payoff, trying to do so optimizing the expected gain. Therefore, the fair price of this derivate depends on its optimal stopping time. The goal of this note is twofold: first, I describe and then show how to solve the optimal stopping problem in a discrete time setting. Then, I show how to apply these techniques for the valuation of American derivatives in the Black-Scholes model and in the Vasicek one.

1 Optimal Stopping Theorem

This first section recalls a few simple definitions from stochastic processes theory and introduces the optimal stopping problem. This problem is then explicitly solved in a discrete time setting. Finally, the problem is solved also for a class of continuous time processes by means of a limit result. For a further analysis of the optimal stopping theory applied to mathematical finance problems see Chapter 16 of Björk (2009).

1.1 The statement of the problem

Let $(\Omega, \mathcal{F}, \mathbf{F}, \mathbb{P})$ be a filtered probability space where $\mathbf{F} = \{\mathcal{F}_t\}_{t\geq 0}$ satisfies the usual conditions. We first recall a standard definition.

Definition 1 A random variable $\tau \geq 0$ is called a **stopping time** with respect to $\mathbf{F}$ if

$$\{\tau \leq t\} \in \mathcal{F}_t \quad \forall t \geq 0.$$

A stopping time is a random variable that represents the instant in time at which a random event of interest realises. It is called a stopping time because, according to the

[(*)]Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: rotondi.francesco@unibocconi.it. Seminar held on 17 March 2021.

standard interpretation, as soon as the event of interest realizes, we stop and take some other actions. According to the definition, the stopping time is non anticipative. This means that at any t we can actually say whether the event of interest has already occurred or not.
A stopping time is usually associated to a **stopping rule**, which is a stochastic process that tells us when to stop, namely when the event of interest realises. Indeed, the following proposition holds.

Proposition 2 *A random variable $\tau \geq 0$ is a stopping time w.r.t. to $\mathbf{F}$ $\iff$ the stochastic process $\{\mathcal{T}_t\}_{t\geq 0}$, defined as*

$$\mathcal{T}_t := \begin{cases} 1 & \text{if } t \leq \tau \\ 0 & \text{if } t > \tau \end{cases}$$

is adapted to the filtration $\mathbf{F}$.

These definitions are quite general as long as we do not define properly the event of interest that triggers the stop. To do so, we need to introduce another stochastic process: let $\{X_t\}_{t\geq 0}$ be a real-valued integrable stochastic process. We call process X the **reward process** or the **payoff process**. The interpretation of this process is the following: we assume that we are playing a game that starts at $t = 0$. At each point in time $t > 0$ we can observe the current value of $X(t)$ and decide whether to stop the game. if we do so, we cash in $X(t)$ euros. If asked to play this game, it makes sense to find an optimal stopping rule that maximizes the expected reward. This rational approach to the game is formalized by the following problem.

Definition 3 The **optimal stopping problem** reads

$$\sup_{0\leq\tau\leq T} \mathbb{E}\left[X_\tau\right]$$

where τ is a stopping time.

In words, when standing at zero, we want to find a stopping time (or a stopping rule) that tells us when to stop process X so to maximize our expected reward.
In a few cases, the optimal stopping problem has a trivial solution:

- if X is a submartingale, then the optimal stopping time is $\tau^* = T$ as the expected value of X is increasing over time;
- on the contrary, if X is a supermartingale, then $\tau^* = 0$ as the expected value of X is decreasing over time and we are better off by stopping right at the beginning before the expected reward decreases;
- if X is a martingale, then $\tau^* \in [0, T]$ is always optimal as the expected value of X is constant over time.

Besides these few trivial cases, the optimal stopping problem is not easily solvable. However, when time is discrete, there is a simple algorithm that does the job.

1.2 Discrete Time

Let $t \in \Pi = \{0, 1, \dots, T\}$. In order to describe the solution to the optimal stopping problem we need to introduce the following definition.

Definition 4 The **optimal value process** $\{V_t\}_{t\in\Pi}$ is defined as

$$V_t = \sup_{\tau \geq t, \tau \in \Pi} \mathbb{E}\left[X_\tau \middle| \mathcal{F}_t \right].$$

To solve the optimal stopping problem we employ a **dynamic programming** approach.
Fix a point in time $t \in \Pi$. We compare the following three strategies:

[S1] use the optimal stopping time τ^*;

[S2] stop right away at t;

[S3] don't stop at t and use the optimal stopping time at $t+1$.

The value of the first two strategies at t follows by definition:

[S1] the optimal value process V_t;

[S2] the current reward/payoff value X_t.

As for the third strategy, its value at $t+1$ is equal to V_{t+1}. However, we have to evaluate this unknown value that realizes at $t+1$ when standing at t. The value, at time t, of using the optimal stopping time τ^* at $t+1$ is

$$\mathbb{E}\left[X_{\tau^*} \middle| \mathcal{F}_t \right] = \mathbb{E}\left[\mathbb{E}\left[X_{\tau^*} \middle| \mathcal{F}_{t+1} \right] \middle| \mathcal{F}_t \right] = \mathbb{E}\left[V_{t+1} \middle| \mathcal{F}_t \right],$$

thanks to the law of iterated expectations. Therefore, the value at t of the third strategy is

[S3] the conditional expected value of the optimal value process at $t+1$, $\mathbb{E}\left[V_{t+1} \middle| \mathcal{F}_t \right]$.

By definition, strategy [S1] (the optimal one) dominates both [S2] and [S3]:

$$V_t \geq X_t \tag{1}$$
$$V_t \geq \mathbb{E}\left[V_{t+1} \middle| \mathcal{F}_t \right]. \tag{2}$$

Anyway, at t, it is optimal to:

- either stop, if $V_t = X_t \geq \mathbb{E}\left[V_{t+1} \middle| \mathcal{F}_t \right]$;
- or not to stop, if $V_t = \mathbb{E}\left[V_{t+1} \middle| \mathcal{F}_t \right] \geq X_t$.

Therefore, we learnt how to express the optimal value at t, V_t, as a function of the observable current value X_t and of a conditional expected value that can be computed at t if V_{t+1} is known. If we known V_T, then we can set up a backward recursion.

Proposition 5 *The optimal value process $\{V_t\}_{t\in\Pi}$ solves the following* ***backward recursion****:*

$$\begin{aligned} V_T &= X_T \\ V_t &= \max\{X_t, \mathbb{E}[V_{t+1}|\mathcal{F}_t]\} \quad \forall t \in \{0,\dots,T-1\} \end{aligned}$$

From this backward recursion, we can retrieve the underlying optimal stopping time.

Proposition 6 *At time zero, an optimal stopping time τ^* is*

$$\tau^* = \min\{t \geq 0 : V_t = X_t\},$$

where the process V is computed using the backward recursion in Proposition 5. Analogously, at t, an optimal stopping time τ_t^* is

$$\tau_t^* = \min\{s \geq t : V_s = X_s\}.$$

It turns out that the optimal value process V has an interesting equivalent characterization. In order to describe it, we need two more definitions.

Definition 7 The process $\{S_t\}_{t\in\Pi}$ **dominates** the process $\{Y_t\}_{t\in\Pi}$ if, $\forall t \in \Pi$,

$$S_t \geq Y_t \quad \mathbb{P}\text{-almost surely.}$$

Definition 8 Let $\{X_t\}_{t\in\Pi}$ be such that $\mathbb{E}[X_t] < \infty \ \forall t \in \Pi$. The **Snell envelope** of X is the smallest supermartingale that dominates X.

Then, the following Theorem holds.

Theorem 9 *In the optimal stopping problem, the optimal value process V is the Snell envelope of the "payoff" process X.*

Proof. From $V_t \geq X_t$ in (1), we know that V dominates X.
From $V_t \geq \mathbb{E}[V_{t+1}|\mathcal{F}_t]$ in (2), we know that V is a supermartingale.
Assume now that $\{S_t\}_{t\in\Pi}$ is a supermartingale that dominates X. At T, as S dominates X, $S_T \geq X_T = V_T$, so V is smaller. By (backward) induction, assume $S_{t+1} \geq V_{t+1}$. Since S is a supermartingale, $S_t \geq \mathbb{E}[S_{t+1}|\mathcal{F}_t]$ and $S_t \geq \mathbb{E}[V_{t+1}|\mathcal{F}_t]$. As S dominates Z, $S_t \geq Z_t$. Therefore,

$$S_t \geq \max\{X_t, \mathbb{E}[V_{t+1}|\mathcal{F}_t]\} = V_t$$

so V is smaller than $S \ \forall t \in \Pi$. □

1.3 Continuous time

When time is continuous, the backward recursion of Proposition 5 is no longer applicable. However, the following convergence result tells us that if we know a discrete stochastic process X^d that converges in distribution to the continuous time process X^c, then so do their Snell envelopes.

Theorem 10 *If* $\{X_t^d\}_{t\in\Pi} \xrightarrow{d} \{X_t^c\}_{t\in[0,T]}$, *then the Snell envelope of* X^d *converges to the Snell envelope of* X^c.

Proof. See Mulinacci and Pratelli (1998). □

In other words, if we face a continuous time optimal stopping problem and we know how to discretize the continuous time reward process, we can solve the related discretize problem and then take the limit as the time step of the partition goes to zero.

2 American options: the Black-Scholes model

Consider an arbitrage-free financial market and let $\{B_t\}_{t\geq 0}$, with $B_t = e^{rt}$, be the price process of the risk-free asset. In this section we assume that the risk-free interest rate is constant. Let $\{S_t\}_{t\geq 0}$ be the price process of a traded security.
Fix a maturity $T > 0$. Let $\{X_t\}_{t\geq 0}$ be a payoff process with $X_t = \varphi(S_t)$. A **European derivative** is a binding financial contract that entitles the holder to receive, at maturity T, the amount of money X_T, that depends on the unknown value of the underlying S_T. As most of the times this amount is non negative, the contract comes at an initial price when signed. The fair no-arbitrage price of a European derivative on S, with maturity T and payoff $X_T = \varphi(S_T)$ is

$$\pi_t^E = \mathbb{E}^{\mathbb{Q}}\left[e^{-r(T-t)}X_T \middle| \mathcal{F}_t\right]$$

where $\mathbb{Q}$ is a risk-neutral measure[(1)]. For European derivatives the instant of time in which the holder cashes in the payoff is fixed (and usually equal to T).
On the contrary, the holder of an **American derivative** can choose when to stop and cash in the current payoff X_t. In other words, the holder of an American derivative can exercise it at any time before maturity. The seller of such a contract must assume that the buyer (i.e., the holder) of the contract will exercise it when its (discounted) expected payoff is maximum. Therefore, the fair no-arbitrage price of a American derivative on S, with maturity T and payoff function $X_t = \varphi(S_t)$ is

$$\pi_t^A = \sup_{t\leq\tau\leq T} \mathbb{E}^{\mathbb{Q}}\left[e^{-r(\tau-t)}X_\tau \middle| \mathcal{F}_t\right].$$

Recalling Definition 8, we can claim that the price process $\left\{\pi_t^A\right\}_{t\in[0,T]}$ of an American derivative is the Snell envelope of the (discounted) payoff process X.

[(1)] See Chapters 3-4 of Björk (2009) for an introduction to no arbitrage pricing theory.

Consider now the Black and Scholes (1973) model. Namely, assume that S is a geometric Brownian motion under $\mathbb{Q}$:

$$\frac{\mathrm{d}S_t}{S_t} = (r - \bar{q})\mathrm{d}t + \sigma \mathrm{d}W_t^{\mathbb{Q}}$$
$$S_t = S_0 e^{\left(r - \bar{q} - \frac{\sigma^2}{2}\right)t + \sigma W_t^{\mathbb{Q}}}$$

where $\bar{q}$ is the continuous dividend yield of the security and σ its volatility. Clearly, the discounted payoff process $X = e^{-rT}\varphi(S)$ is a continuous time process. In order to find the price of an American derivative within this market model, we need to come up with a discretization of S. The most handful discretization of the geometric Brownian motion is the binomial model.

2.1 The binomial model

In their celebrated paper, Cox et al. (1979) introduce the **binomial model**, a discrete time process that converges in distribution to a geometric Brownian motion. Given the uniform partition $\Pi = \{0, 1, \ldots, T\}$, with time step $\Delta t = 1$, the binomial tree $\{\tilde{S}_t\}_{t\in\Pi}$ is a lattice discretization of S built as follows:

- $\tilde{S}_0 = S_0$.
- $\tilde{S}_t = \begin{cases} \tilde{S}_{t-1}u & \text{with probability } q \\ \tilde{S}_{t-1}d & \text{with probability } 1-q \end{cases}$ for all $t = 1, \ldots, T$.
- $u > d$, q are chosen in such a way to match the first two moments of $\ln S$:

  $$\mathbb{E}^{\mathbb{Q}}[\ln \tilde{S}_t] = \mathbb{E}^{\mathbb{Q}}[\ln S_t] \text{ and } Var^{\mathbb{Q}}[\ln \tilde{S}_t] = Var^{\mathbb{Q}}[\ln S_t].$$

  As a result, we have

  $$u = e^{\sigma\sqrt{\Delta t}}, \quad d = \frac{1}{u}, \quad q = \frac{e^{(r-\bar{q})\Delta t} - d}{u - d},$$

  where $\Delta t = 1$ is the time step of the uniform partition Π.

The resulting discretized process looks indeed like a tree. Figure 1 shows a two step binomial tree. Notice that, since $u = \frac{1}{d}$, the tree is recombining and the possible values of the tree at a given step increases only linearly in the number of steps. Using the backward algorithm of Proposition 5, we can easily retrieve the value process of an American derivative along this two-step binomial tree as well as the related optimal stopping rule. Assume that the payoff process of an American derivative on S with maturity T is $X_t = \varphi(S_t)$. At $T = 2$, $\pi_2^A = \varphi(\tilde{S}_2)$. At $t = 1$, if $\tilde{S}_1 = S_0 u$,

$$\pi_1^A(S_0 u) = \max\left\{\varphi(S_0 u), e^{-r\Delta t}(q\pi_2^A(S_0 u^2) + (1-q)\pi_2^A(S_0 ud))\right\},$$

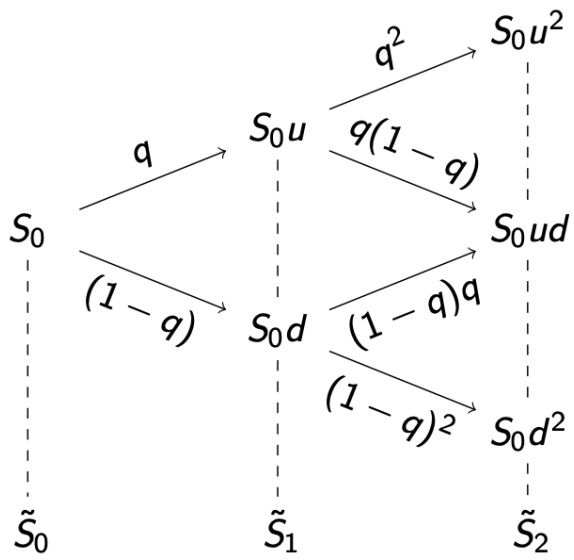


Figure 1 Two-step binomial tree.

while, if if $\tilde{S}_1 = S_0 d$, then

$$\pi_1^A(S_0 d) = \max\left\{\varphi(S_0 d), e^{-r\Delta t}(q\pi_2^A(S_0 ud) + (1-q)\pi_2^A(S_0 d^2))\right\}.$$

Finally, at inception $t = 0$,

$$\pi_0^A(S_0) = \max\left\{\varphi(S_0), e^{-r\Delta t}(q\pi_1^A(S_0 u) + (1-q)\pi_1^A(S_0 d))\right\}.$$

Keeping track, node by node, of the nodes in which the **immediate payoff**, $\varphi(S_t)$ is higher than the **continuation value**, $\mathbb{E}\left[\pi_{t+1}^A \middle| \mathcal{F}_t\right]$ allows us to retrieve the optimal stopping time/rule, also called the **optimal exercise policy**. This policy is uniquely associated to the **critical price**, $\tilde{S}_t^* : \Pi \to \mathbb{R}^+$, which is defined as

$$\tilde{S}_t^* := \max / \inf \left\{\tilde{S}_t \text{ s.t. } \varphi(S_t) = \pi_t^A\right\},$$

where the max is used for a payoff which is decreasing in S and the min is used for a payoff which is increasing in S. In words, the optimal stopping rule prescribes to stop as soon as the value of the underlying touches the critical price.
When taking the limit of the binomial discretization as the time step Δt goes to zero, we are able to get the price at inception of the American derivative under investigation and also the continuous version of the critical price, $\tilde{S}^*$, that tells us the optimal exercise policy of the derivative.

Consider now an American put option issued at the money, namely, set $\varphi(S_t) = (S_0 - S_t)^+ := \max\{0, S_0 - S_t\}$. The limit of the critical price is shown in Figure 2.

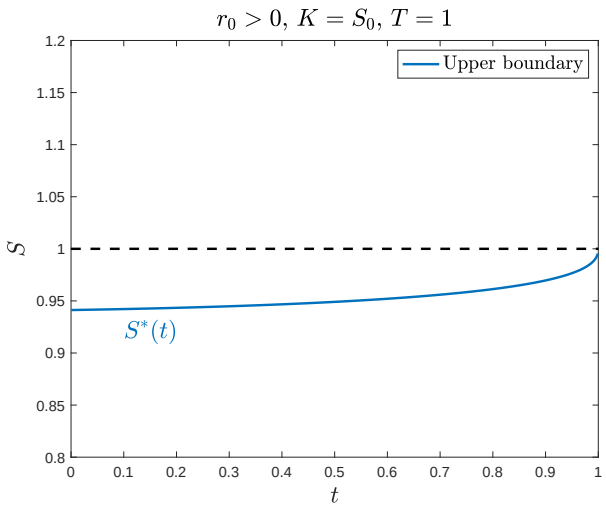


Figure 2 Critical price of an at-the-money American put option within the Black-Scholes model.

As soon as the value of the underlying falls below the blue boundary[(2)] the holder of the derivative is better off by exercising it right away.

3 American options: the Vasicek model

Battauz and Rotondi (2020) extend the well-known results of the previous section to a stochastic interest rates environment. More precisely, we consider a Vasicek (1977) model where, under $\mathbb{Q}$, the price process of the locally riskless asset is $B_t = e^{\int_0^t r_v \mathrm{d}v}$; the locally riskless short-term interest rate $\{r_t\}_{t\in[0,T]}$ solves

$$\begin{aligned} \mathrm{d}r_t &= \kappa\left(\theta - r_t\right)\mathrm{d}t + \sigma_r \mathrm{d}W_t^r \\ r_t &= r_0 e^{-\kappa t} + \theta(1 - e^{-\kappa t}) + \sigma_r \int_0^t e^{-\kappa(t-s)} \mathrm{d}W_s^r, \end{aligned}$$

and the price process of the traded security solves

$$\begin{aligned} \frac{\mathrm{d}S_t}{S_t} &= (r_t - \bar{q})\mathrm{d}t + \sigma_S \mathrm{d}W_t^S \\ S_t &= S_0 e^{\int_0^t r_v \mathrm{d}v - \left(\frac{\sigma_S^2}{2} + \bar{q}\right)t + \sigma_S W_t^S}. \end{aligned}$$

Moreover, the two Brownian innovations that drive the market/equity and the interest rate risk in the market are correlated, $\mathrm{d}\langle W^r, W^S\rangle = \rho \mathrm{d}t$. The parameters in the SDE solved by r_t, which is an Ornstein–Uhlenbeck process, have the following interpretation: κ is the speed of mean reversion, as it is assumed that r, starting at r_0 reverts back to, θ, the

[(2)] The critical price is also referred to as the upper/lower boundary as the American valuation problem can be also stated through a variational inequality approach, solved by a free boundary.

long-run mean; σ_r is the volatility of the interest rate.
This setting allows us to assess the impact of interest rate risk in the valuation of American derivatives as long as we know how to jointly discretize the two processes (S, r).

3.1 The quadrinomial tree

Inspired by the structure of the binomial tree recalled in the previos section, we derive a lattice discretization of $\{(S_t, r_t)\}_{t\in[0,T]}$ and we call it the **quadrinomial tree**. Given a uniform partition $\Pi = \{0, 1, \dots, T\}$, with time step $\Delta t = 1$, the quadrinomial tree is built as follows

- $(\tilde{S}_0, \tilde{r}_0) = (S_0, r_0)$.
- $(\tilde{S}_t, \tilde{r}_t) = \begin{cases} (\tilde{S}_{t-1}u_S, \tilde{r}_{t-1}u_r) & \text{with probability } q^{uu} \\ (\tilde{S}_{t-1}u_S, \tilde{r}_{t-1}d_r) & \text{with probability } q^{ud} \\ (\tilde{S}_{t-1}d_S, \tilde{r}_{t-1}u_r) & \text{with probability } q^{du} \\ (\tilde{S}_{t-1}d_S, \tilde{r}_{t-1}d_r) & \text{with probability } q^{dd} \end{cases}$
  for all $t = 1, \dots, T$.
- u_S, d_S, u_r, d_r, q^{uu}, q^{ud}, q^{du}, q^{dd} are chosen in such a way to match the first two moments of $\ln S$ and r and their covariance:

$$\mathbb{E}^{\mathbb{Q}}[\ln \tilde{S}_t] = \mathbb{E}^{\mathbb{Q}}[\ln S_t], \quad Var^{\mathbb{Q}}[\ln \tilde{S}_t] = Var^{\mathbb{Q}}[\ln S_t]$$

$$\mathbb{E}^{\mathbb{Q}}[\tilde{r_t}] = \mathbb{E}^{\mathbb{Q}}[r_t], \quad Var^{\mathbb{Q}}[\tilde{r_t}] = Var^{\mathbb{Q}}[r_t], \quad \mathbb{E}^{\mathbb{Q}}[\ln \tilde{S}_t \tilde{r_t}] = \mathbb{E}^{\mathbb{Q}}[S_t r_t].$$

  As a result we have:

$$\begin{aligned} u_S &= e^{\sigma_S\sqrt{\Delta t}}, \quad d_S = \tfrac{1}{u_S} \\ u_r &= \sigma_r\sqrt{\Delta t}, \quad d_r = -u_r \end{aligned}$$

$$\begin{aligned} q_{uu} &= \frac{\mu_Y\mu_r\Delta t + \mu_Y u_r + \mu_r u_S + (1+\rho)\sigma_r\sigma_S}{4\sigma_r\sigma_S} \\ q_{ud} &= \frac{-\mu_Y\mu_r\Delta t + \mu_Y u_r - \mu_r u_S + (1-\rho)\sigma_r\sigma_S}{4\sigma_r\sigma_S} \\ q_{du} &= \frac{-\mu_Y\mu_r\Delta t - \mu_Y u_r + \mu_r u_S + (1-\rho)\sigma_r\sigma_S}{4\sigma_r\sigma_S} \\ q_{dd} &= \frac{\mu_Y\mu_r\Delta t - \mu_Y u_r - \mu_r u_S + (1+\rho)\sigma_r\sigma_S}{4\sigma_r\sigma_S}. \end{aligned}$$

  where $\mu_Y := \left(r(t) - q - \frac{\sigma_S^2}{2}\right)$ and $\mu_r := \kappa(\theta - r(t))$. Notice that, in the quadrinomial tree, the transition probabilities are time-dependent and state contingent. This is due to the trend-stationarity of r: jumps that move away from θ must have smaller probability than the ones that move closer to θ.
  Moreover, the four transition probabilities are not necessarily strictly positive, although they tend to become negative for extremely unlikely values of the parameters. See the Appendix of Battauz and Rotondi (2020) for all the related details.

The quadrinomial tree looks like a bivariate binomial tree.

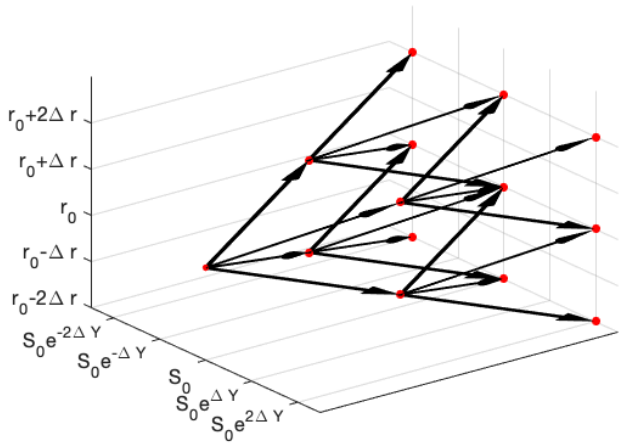


Figure 3 Two-step quadrinomial tree.

Figure 3 shows a two-step quadrinomial tree.

Using the quadrinomial tree, conditional expected values can be computed as in the standard binomial tree averaging out four possible outcomes (with probabilities that have to be computed node by node).
Consider an American derivative with payoff $\varphi(S)$. When valuing it option along the quadrinomial tree keeping track of the optimal stopping time/rule we can derive the analogous of the critical price introduced in the previous section for the univariate Black-Scholes model. The univariate critical price becomes now a **critical surface**, defined as

$$\left(\tilde{S}_t^*, \tilde{r}_t^*\right) := \max / \min \left\{ \left(\tilde{S}_t, \tilde{r}_t\right) \text{ s.t. } \varphi(S_t) = \pi_t^A \right\}.$$

where the max is used for a payoff which is decreasing in S and the min is used for a payoff which is increasing in S. In words, the critical surface is the set of all the state variables couples that trigger the optimal exercise of the American derivative.

Consider now an American put option issued at the money, namely, set $\varphi(S_t) = (S_0 - S_t)^+ := \max\{0, S_0 - S_t\}$. The shape of the critical surface can change significantly with the sign of r and $\bar{q}$, the continuous dividend yield of the equity, as both impact S' drift, $\mu := r - \bar{q}$.
Assume $\bar{q} = 0$. If the underlying pays no dividend the expected drift of S coincides with r. This splits the domain of r in two complementary regions according to the sign of r, as can be seen in the right panel of Figure 4 (that displays the free boundary section at $t = \frac{T}{2}$). In the left region where r and μ are both negative, the investor is willing to wait and postpone the exercise as much as possible in order to gain from both the negative discount

rate and the implied expected depreciation of S. In the right region, on the contrary, where r and μ are both positive, we have the standard tradeoff between a positive discount rate (that makes the investor willing to exercise the option as soon as possible) and a negative expected drift of S (that makes the investor willing to wait for a larger payoff).

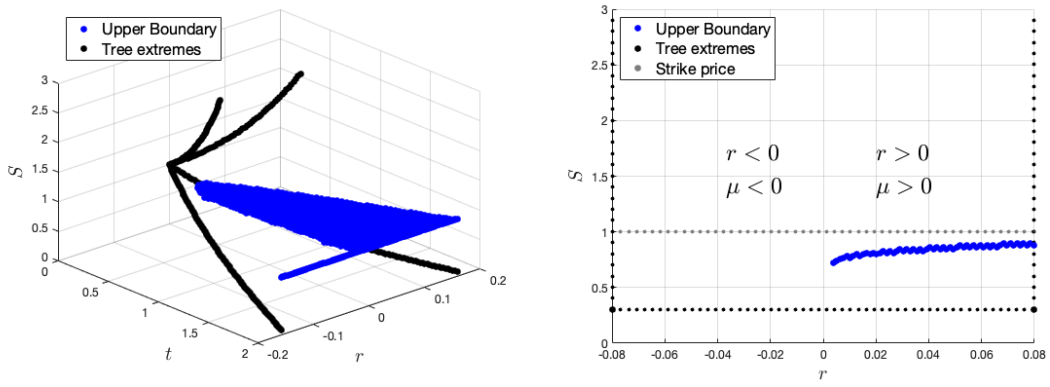


Figure 4 $\bar{q} = 0$ case. Left panel: critical surface of an at-the-money American put option within the Vasicek model. Right panel: r-section of the surface in the left panel at $t = \frac{T}{2} = 1$.

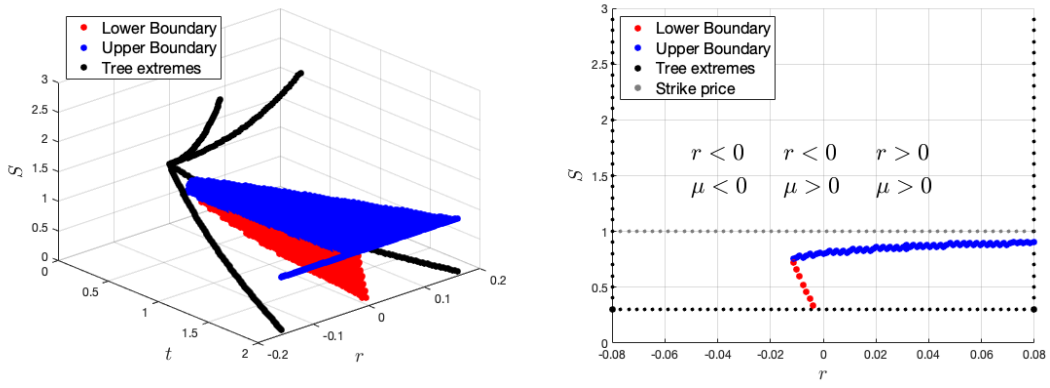


Figure 5 $\bar{q} < 0$ case. Left panel: critical surface of an at-the-money American put option within the Vasicek model. Right panel: r-section of the surface in the left panel at $t = \frac{T}{2} = 1$.

This generates the "standard boundary" shown in the left panel of Figure 4.
Assume now $\bar{q} < 0$[(3)]. In this case, the drift of S is equal to r plus a positive quantity. As a result, μ may be positive also when r is mildly negative. This splits the domain of r into three complementary regions, as shown in the right panel of Figure 5: the one in which r and μ are both negative, the one in which r is negative but μ is positive and the last

[(3)] A negative dividend yield is common if the underlying is commodity which entails a cost of carry or if the underlying is the price of a foreign security, see Battauz et al. (2018).

one in which r and μ are both positive. In the first region, the option is again optimally exercised at maturity as in the previous example. In the middle section a non standard lower critical surface (or boundary) appears. The **lower critical surface** is defined as

$$\left(\tilde{S}_{*t}, \tilde{r}_{*t}\right) := \inf \left\{ \left(\tilde{S}_t, \tilde{r}_t\right) \text{ s.t. } (S_0 - \tilde{S}_t)^+ = \pi_t^A \right\}.$$

and it represents the set of notes below which it is not optimal to exercise the American put option. Below this lower critical surface, indeed, the expected drift of the underlying is pushes it up making the option lose value and therefore the holder of the option is tempted to exercise it. On the other hand, the negative interest rate suggests her to cash in later on rather than right now. This unusual tradeoff delivers the non standard double surfaces. In the last region where both r and μ are positive, we find the standard behaviour already outlined in the previous case.

References

Battauz, A., De Donno, M., and Sbuelz, A. (2018), *On the exercise of american quanto options.* Working paper.

Battauz, A. and Rotondi, F. (2020), *American options and stochastic interest rates.* Working paper.

Björk, T. (2009), "Arbitrage theory in continuous time". Oxford Finance, 3 edition.

Black, F. and Scholes, M. (1973), *The pricing of options and corporate liabilities.* Journal of Political Economy, 81(3):637–654.

Cox, J., Ross, S., and Rubinstein, M. (1979), *Option pricing: a simplified approach.* Journal of Financial Economics, 7(3):229–263.

Mulinacci, S. and Pratelli, M. (1998), *Functional convergence of snell envelopes: application to american option approximations.* Finance and Stochastics, 2(3):311–327.

Vasicek, O. (1977), *An equilibrium characterization of the term structure.* Journal of Financial Economics, 5(2):177–188.

On the simulation of planar homogeneous flows

Francesca Tedeschi (*)

Abstract. In this report we are going to consider Non-Equilibrium Molecular Dynamics (NEMD), a well-established simulation technique for molecular fluids undergoing simple shear and planar extensional flows, see [12]. Periodic boundary conditions shall be used and, since the simulation box deforms with the flow, image particles become arbitrarily close causing a breakdown in the simulation at a certain time. Kraynik and Reinelt [6] (KR) got a theoretical result in 1992 to avoid this problem for planar elongational flow that requires you to carefully choose the initial simulation box and periodically remap the simulation box in a way that conserves image locations. Our particular interest is to generalize the KR method to the case of planar mixed flows (linear combination between planar Couette and elongation) in order to guarantee infinite simulation times and thus, provide a reliable method to simulate also these conditions of motion with NEMD techniques.

1 Introduction

Molecular Dynamics (MD) techniques simulate the interactions between the molecules of a fluid and compute the trajectories of particles belonging to certain statistical ensembles. Through the outputs of these simulations we are able to extract local information (about atoms positions, bonds and averaged quantities, such as pressure, temperature and energies) and build the Cauchy stress tensor of the macroscopic variational problem. Our approach takes into account the fact that the response of a complex material can depend strongly on the local flow type and is thus necessary to properly sample the space of kinematic parameters, performing MD simulations in different conditions of motion.

The Figure 1 represents the flow of a fluid in the contraction channel: it is a paradigmatic example of complex geometry that gives rise to different kinematic conditions in the various regions: simple shear (planar Couette) motion is displayed far from the contraction, elongation is developing along the centerline, rotational motion is occupying the corners of the geometry and mixed motion is spread all around.

We would like to take into account these effects on the dynamic, due to the presence of holes, barriers, obstacles in the Cauchy stress tensor (at the continuum scale). Thus our tensor is not given *a priori* through a constitutive law (as usually) but is reconstructed, point by point, using the decomposition proposed in [4] and the information

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: tedeschi@math.unipd.it . Seminar held on 31 March 2021.

(about viscosity and other fluid properties) for the different local flow types, collected in MD simulations.

In the classical way of thinking, rheological measurements of real fluids are recovered through viscometric flows, equivalent to simple shear motions. For example in Figure 2 (left) is reported the parallel plate rheometer which is a capstone in rheology for fluid properties investigation: the material is here confined between the two shearing plates by capillary forces, without flowing out, and the torque is measured on the rotating disc. Also the simulation techniques are, in this case, almost old and well developed, [8]. However planar extension (and mixed motion too) cannot be ignored because plays a crucial role in complex flows and is much more investigated recently, see [5]. This flow type is really important in polymer processing and many other industrial processes, such as the extrusion process. In Figure 2 (right) I represented this application of the 2D extensional motion. It is widely used in engineering field to create an object which has a fixed cross-sectional area. For making the object, the raw material is pushed into a die to provide it with the desired shape. There exist algorithms and experimental tools such as rheometer for the study of this flow type ([1] and [11]), but they are not so updated and developed and need to be investigated more. So, many efforts in the last years has been directed towards the creation of methods usable in presence of motions other than simple shear. Something is already available, [2]. In this report I will show in details the application of the Krayinik and Reinelt boundary conditions to the case of intermediate motions, as first theoretical step in building a MD algorithm to perform simulations in those conditions.

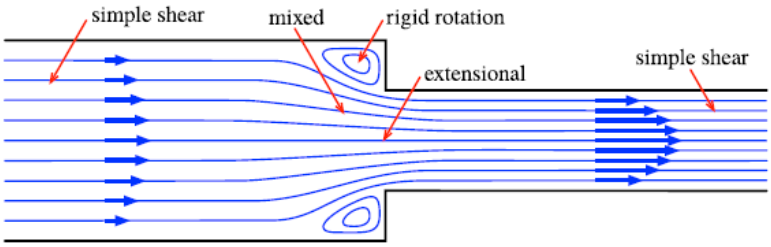


Figure 1 The flow through a contraction channel is an example of complex flow because of the appearance of different flow types. (Taken from [4])

2 Balance equations of fluid dynamics

From continuum mechanics, we assume the *principle of local conservation of mass*: the mass is preserved through the motion, there is neither loss nor accumulation of mass along the pathway, the mass is following exactly trajectories of geometrical points. Because of this hypothesis we get the *equation of continuity* with $\rho : \Omega \times \mathbb{R}^+ \to \mathbb{R}$, $\rho = \rho(\mathbf{x}, t)$ density of mass function

$$\frac{\partial \rho}{\partial t} + \mathbf{v} \cdot \nabla \rho + \rho \nabla \cdot \mathbf{v} = \frac{d\rho}{dt} + \rho \nabla \cdot \mathbf{v} = 0.$$

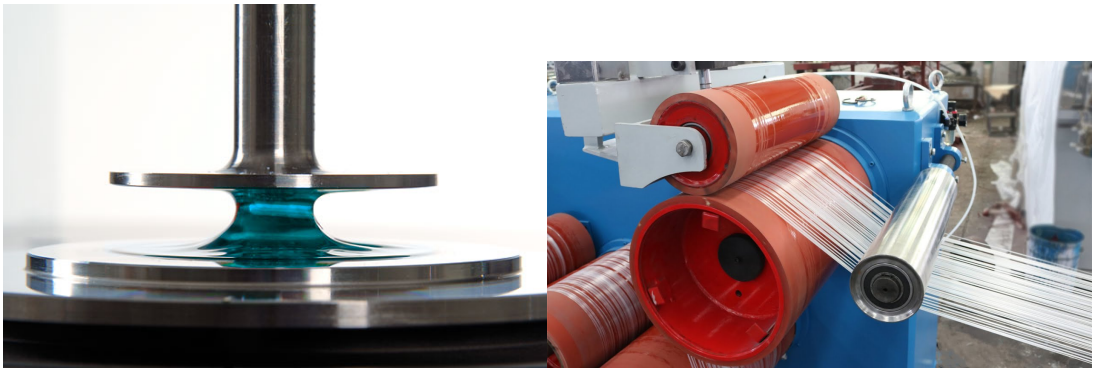


Figure 2 Parallel plates rheometer measuring the shear viscosity of a fluid (left) and extrusion process as example of application of planar extensional flow (right).

If $\mathbf{b} : \Omega \times \mathbb{R}^+ \to \mathbb{R}^3$, $\mathbf{b} = \mathbf{b}(\mathbf{x}, t)$ is a force density for unit mass, from the *principle of conservation of linear momentum* (analogue of the Newton's equation for the classical mechanics) we obtain the *equation of motion*

$$\rho \frac{d\mathbf{v}}{dt} = \rho \mathbf{b} + \nabla \cdot \boldsymbol{\sigma}$$

where $\boldsymbol{\sigma}$ is the *Cauchy stress tensor* of internal forces, prescribed with a *constitutive law*

$$\boldsymbol{\sigma} = f\left(\rho, \nabla\rho, \dots, \mathbf{v}, \nabla\mathbf{v}, \dots\right).$$

This relation codifies how the fluid internally reacts to external forces imposed with some velocity gradients $\nabla\mathbf{v}$. At this point, the only things we know is that it depends on the two unknowns of the problems, the density ρ and the velocity $\mathbf{v}$ and their derivatives.

In our work we are interested in considering only *incompressible* fluids subject to *homogeneous* flows meaning that the density is a constant function in space and time $\rho = \overline{\rho}$. Moreover if we neglect the influences of external volume forces $\mathbf{b} = 0$, we have the following PDEs

$$\begin{cases} \nabla \cdot \mathbf{v} = 0 \\ \rho \dfrac{d\mathbf{v}}{dt} = \nabla \cdot \boldsymbol{\sigma} \\ \boldsymbol{\sigma} = f\left(\mathbf{v}, \nabla\mathbf{v}, \dots\right) \end{cases}$$

where the constitutive law no longer depends on the density ρ.

In Table 1 we report some basic examples of constitutive laws and their correspondent equation of motion. There are many other choices that can be made for the Cauchy stress tensor.

	Constitutive law	Equation of motion
Perfect	$\boldsymbol{\sigma} = -p\mathbf{I}$	$\rho\frac{d\mathbf{v}}{dt} = -\nabla p$
Newtonian	$\boldsymbol{\sigma} = -p\mathbf{I} + 2\eta\mathbf{D}, \quad \eta(\lvert\dot{\varepsilon}\rvert) = \text{const}$	$\rho\frac{d\mathbf{v}}{dt} = -\nabla p + \nu\Delta\mathbf{v}$
Power-law	$\boldsymbol{\sigma} = -p\mathbf{I} + 2\eta(\lvert\dot{\varepsilon}\rvert)\mathbf{D}, \quad \eta(\lvert\dot{\varepsilon}\rvert) = k\lvert\dot{\varepsilon}\rvert^{s-2}$	$\rho\frac{d\mathbf{v}}{dt} = -\nabla p + k\nabla\cdot(\lvert\dot{\varepsilon}\rvert^{s-2}\mathbf{D})$

Table 1 Three examples of constitutive laws: Perfect, Newtonian and Non-Newtonian (Power-law) fluids.

For the *Perfect fluid* the stress tensor $\boldsymbol{\sigma}$ is constant and, in particular, is a multiple of the identity through p, the Lagrange multiplier due to the incompressibility constraint. This means that no forces are arising inside the fluid in reaction to a compression from outside. With this model we are neglecting viscous forces, thus the fluid is moving as a rigid body, governed by the Euler equation of motion.

The first attempt to consider the influence of viscosity is through *Newtonian fluids* represented by a linear relation between the stress $\boldsymbol{\sigma}$ and the *strain rate tensor* $\mathbf{D} = \text{Sym}(\nabla\mathbf{v})$, which codifies the type of deformation imposed, with η kinematic viscosity and

$$|\dot{\varepsilon}| = \|\mathbf{D}\| = \sqrt{\frac{\text{tr}\,(\mathbf{D}^2)}{2}}$$

the *strain rate*, measuring the intensity of the deformation. The correspondent equation of motion is the Navier-Stokes equation where ν is the dynamic viscosity, instead.

The third example is a more refined case of viscid fluid, the *Power-law* fluid, characterized by a non linear stress-strain relation, because of the non-constant (power-law) viscosity. $k > 0$ is a dimensional constant, while $s > 1$ is a non-dimensional constant. In particular:

- if $s < 2$ the viscosity is decreasing with respect to $\|\mathbf{D}\|$, *shear-thinning fluids*
- if $s > 2$ the viscosity is increasing with respect to $\|\mathbf{D}\|$, *shear-thickening fluids*.

There are plenty of other models for viscid fluids that takes into account many different properties, see [7]. However, real fluids are so complex, in a way that they disregard any classification and any characterization because choosing a method implies to look at some properties, ignoring many others. There is no, at the moment, a very complete method to describe real fluids. Is for these reasons that we are trying to build a numerical representation of the stress tensor $\boldsymbol{\sigma}$, instead of prescribing a constitutive law, based on data produced with MD simulation techniques, that are as realistic as possible.

3 Velocity gradient tensor

We are going to focus our attention to 2-dimensional flows and to suppose that $\nabla \mathbf{v} \neq 0$ which would be the case of a rigid body, not interesting for fluid dynamics.

Historically, Stokes was the first to look at the velocity gradient as the sum of its symmetric and skew components, respectively $\mathbf{D} = \text{sym}(\nabla \mathbf{v})$ and $\mathbf{W} = \text{skew}(\nabla \mathbf{v})$, giving to them the meaning of *strain rate tensor* and *vorticity tensor*.

$$\nabla \mathbf{v} = \mathbf{D} + \mathbf{W}.$$

We prefer to represent the two tensors on the basis of the eigenvectors $\hat{\mathbf{d}}_1, \hat{\mathbf{d}}_2$ of $\mathbf{D}$

$$\mathbf{D} = \dot{\varepsilon}[\hat{\mathbf{d}}_1\hat{\mathbf{d}}_1 - \hat{\mathbf{d}}_2\hat{\mathbf{d}}_2] \qquad \mathbf{W} = \dot{\varepsilon}\beta_3[\hat{\mathbf{d}}_2\hat{\mathbf{d}}_1 - \hat{\mathbf{d}}_1\hat{\mathbf{d}}_2]$$

because in this way $\nabla \mathbf{v}$ is clearly represented by two kinematic parameters that are the *strain rate* $\dot{\varepsilon}$, already introduced in Section 2, and

$$\beta_3 = \frac{(\nabla \times \mathbf{v})}{2\dot{\varepsilon}} \cdot \hat{\mathbf{d}}_3$$

that is the *flow parameter*, quantifying the amount of vorticity along the out-of-flow direction and measuring the relative importance of the rotational part with respect to the pure deformation part.

According to the parameter β_3 we classify *local (homogeneous) flows* as Table 2 shows

$\beta_3 = 0$	**planar extension**
$\beta_3 = 1$	**simple shear**
$\beta_3 \to \infty$	**rigid rotational motion**
$0 < \beta_3 < 1$	**intermediate motions**

Table 2 Local (homogeneous) flow classification through the parameter β_3.

and Figure 3 displays the vector fields and the correspondent streamlines, for the flow types ranging from planar extensional (hyperbolic lines), passing through simple shear (straight lines) to rigid rotational motion (circular lines).

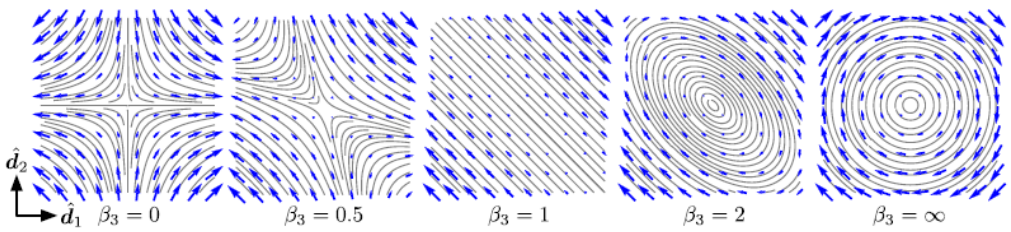


Figure 3 Representation of the vector field in the different flow types. (Figures taken from [4])

The deformation of a system subject to a constant (in space and time) velocity gradient $\mathbf{A} = \nabla\mathbf{V}$ is described by the following autonomous linear ODE

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{A}\,\mathbf{x} \\ \mathbf{x}(0) = \mathbf{x}_0 \end{cases}$$

where $\mathbf{x} \in \mathbb{R}^2$ is the "material" vector position of a generic point in the system and $\mathbf{x}_0 \in \mathbb{R}^2$ is its initial condition. The flux of the ODE is represented by the infinitely differentiable function

$$\begin{aligned} \Phi : \mathbb{R}^+ \times \mathbb{R}^2 &\rightarrow \mathbb{R}^2 \\ (t, \mathbf{x}_0) &\rightarrow \Phi^{\mathbf{A}}(t, \mathbf{x}_0) = \exp(\mathbf{A}t)\mathbf{x}_0 \end{aligned}$$

where $\mathrm{F}(t) = \exp(\mathbf{A}t)$ is the deformation matrix and has the following expression for each type of flow

- **simple shear** $\quad (\nabla\mathbf{v})\,t = \begin{bmatrix} 0 & 2\dot{\varepsilon}t \\ 0 & 0 \end{bmatrix} \rightarrow \mathrm{F}(t) = \begin{bmatrix} 1 & 2\dot{\varepsilon}t \\ 0 & 1 \end{bmatrix}$

- **elongational** $\quad (\nabla\mathbf{v})\,t = \begin{bmatrix} \dot{\varepsilon}t & 0 \\ 0 & -\dot{\varepsilon}t \end{bmatrix} \rightarrow \mathrm{F}(t) = \begin{bmatrix} \exp(\dot{\varepsilon}t) & 0 \\ 0 & \exp(-\dot{\varepsilon}t) \end{bmatrix}$

- **intermediate** $\quad (\nabla\mathbf{v})\,t = \begin{bmatrix} \dot{\varepsilon}t & -\dot{\varepsilon}\beta_3 t \\ \dot{\varepsilon}\beta_3 t & -\dot{\varepsilon}t \end{bmatrix}$

We refer to Section 5 for the expression of the deformation matrix in the mixed case, which is quite complex.

Thus, to sum up, we use MD techniques to realize simulations with imposed globally constant $\nabla\mathbf{v}$, associated to a specific flow type, and this gives rise to a global homogeneous flow. The simulation box is deformed through the deformation matrices presented above and results about the stress tensor components are collected and associated to each point in the macroscopic solver that displays such a flow type. Imposing $\nabla\mathbf{v}$ in our case is equivalent to impose a specific value for $\dot{\varepsilon}$ and β_3, kinematic parameters, and from the values of $\boldsymbol{\sigma}$ we will recover the relation f of

$$\boldsymbol{\sigma} = f(\dot{\varepsilon}, \beta_3)\,.$$

4 Non-Equilibrium Molecular Dynamics (NEMD)

Molecular Dynamics (MD) simulations aim to predict the time-dependent trajectories in a system of N interacting particles, through the integration of the Newton's equations of motion.

It has been introduced by Alder and Wainwright between 1957 and 1959 with the hard sphere model, but were Rahman and Stillinger that carried out the first molecular dynamic simulation of a realistic system (liquid water) in 1972. It is used to understand the molecular behavior of a material because it is able to simulate (even for short times) on the atomic scale.

The term Non-Equilibrium (NE) refers to the fact that the average velocity of the system is not zero $\overline{\mathbf{v}} \neq 0$ and we are providing energy to the system from external sources to maintain active this condition, thus we are not at the thermodynamic equilibrium.

In MD simulations each particle is a *point mass* $\mathrm{m_i}$ with initial positions $\mathbf{x}_i(0)$ and velocities $\mathbf{v}_i(0)$, but can build even more complex systems characterized by single atoms, group of atoms or macro-/ meso- particle.

Particles interact via *empirical force laws*, that can act at distance through an interaction potential denoted by $\mathrm{U_{non\text{-}bond}}$ or through covalent bonds, thus for direct contact with potential $\mathrm{U_{bond}}$, so we find the total interaction potential through the sum of the two contributions

$$\mathrm{U} = \mathrm{U_{bond}} + \mathrm{U_{non-bond}} \quad \rightarrow \mathbf{F}_\mathrm{i}(t) = -\nabla_\mathrm{i}\mathrm{U}(\mathbf{x}_1(t), \mathbf{x}_2(t), \ldots, \mathbf{x}_\mathrm{N}(t))$$

where $\mathbf{F}_\mathrm{i}(t)$ is the force acting on particle i and $\mathbf{x}_j(t)$ is the position of the j-th particle.

Moreover, the interactions may occur between a different number of particles: can be pair-wise (LJ, WCA, Coulombic, ...), many-body (EAM, Tersoff, REBO, ...) or molecular (spring, torsions, FENE, ...).

As already said, Molecular Dynamics integrate the *Newton's equations* of motion, three for each particle

$$\mathbf{A}_\mathrm{i}(t) = \frac{1}{\mathrm{m_i}} \cdot \mathbf{F}_\mathrm{i}(t) \quad \forall \mathrm{i} = 1, \ldots, \mathrm{N}$$

and this means that we have a set of 3N coupled ODEs to be solved. The meaningful properties are finally retrieved via time-averaging ensemble snapshots.

5 Periodic Boundary conditions (PBCs)

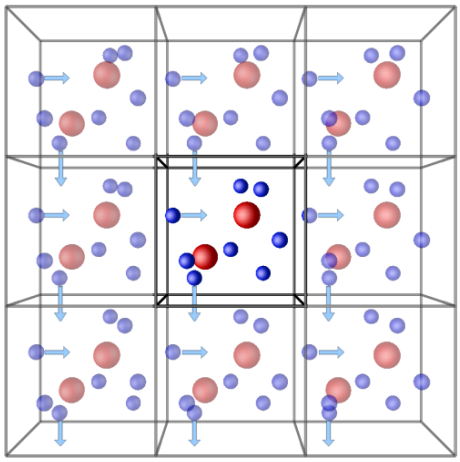
We want to simulate a *bulk flow* of a fluid with $\sim 10^{23}$ atoms, but the simulated system has at most 10^7 atoms. So, the question that now arises is: *how do we approximate a system virtually infinite through a finite system?* The answer is achieved through Periodic Boundary Conditions (PBCs).

The computational box represents a portion of fluid (containing a finite number N of particles) and the software is surrounding it by periodic copies of itself, called *images*, that follow exactly the same trajectories of the main particles, see Figure 4.

This is obviously advantageous because it requires fewer computational resources, compared to considering cells composed of completely different particles. For Molecular Dynamics issues refer to [3]. PBCs are used because they guarantee the following principles:

- the forces in the main cell are computed using all the particles, also copies that belong to different boxes

- each particle interacts only with one copy of another particle
- if N is sufficiently large, the statistics observed is *independent of N* and of the *periodicity*.



(Figures taken from https://lammps.sandia.gov/)

Figure 4 Scheme of PBCs conditions.

If we consider a particle and its images we obtain a lattice of points, that should meet two requirements for the simulation to be reliable and to be carried out for an indefinite time: *compatibility* and *reproducibility*.

Since a particle must not interact with its copies, the minimum separation D of all lattice points (identical image particles) must exceed the cutoff range $\mathrm{r_{cut}}$ of interactions $D \gg \mathrm{r_{cut}}$. This is the compatibility condition.

Reproducibility means instead that the lattice "repeats itself periodically" with the deformation. The exponential deformation of the system periodically returns to a state where replacing some of the original particles with their images the initial state boundaries are recovered. In correspondence to the time period the box should be re-initialized to allow the simulation to be still carried out. Reproducibility, usually, guarantees compatibility.

In simple shear the squared lattice is reproduced at periods of time $\tau_p = 1$ but can be re-initialized at any time, in fact, in the initial squared box there is always exactly a copy of each particle.

KR found reproducibility for the planar extensional motion, using a squared simulation box, for a period $\tau_p = 0.962424$ and tilting the initial box of an angle $\vartheta = 0.55357435$.

Following the pathway indicated by [6] we found the conditions of reproducibility of a lattice under mixed conditions of motion. We are going to illustrate in details the algorithm. Supposing $\dot{\varepsilon} = 1$, the velocity gradient tensor is represented by the matrix

$$(\nabla \mathbf{V})_{\mathcal{B}} = \begin{bmatrix} 1 & -\beta_3 \\ \beta_3 & -1 \end{bmatrix} \qquad \text{with basis} \quad \mathcal{B} = \{\mathbf{d}_1, \mathbf{d}_2\}$$

whose eigenvectors are $\mathbf{v}_1 = \left(\frac{1+\sqrt{1-\beta_3^2}}{\beta_3}, 1\right)$, $\mathbf{v}_2 = \left(\frac{1-\sqrt{1-\beta_3^2}}{\beta_3}, 1\right)$, subsequently normalized $\hat{\mathbf{v}}_1$, $\hat{\mathbf{v}}_2$.

Since $(\nabla \mathbf{V})_{\mathcal{B}}$ is diagonalizable, if $\mathbf{D}$ is the correspondent diagonal matrix, we have

$$(\nabla \mathbf{V})_{\mathcal{B}} = \mathbf{S}(\mathbf{D})\mathbf{S^{-1}}.$$

Now, taking as basis the two orthonormal eigenvectors $\mathcal{B}' = \{\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_1^{\perp}\}$ and let $\mathbf{R} = \mathcal{M}_{\mathcal{B}\mathcal{B}'}$ be the matrix of change basis, we obtain the velocity gradient matrix in the new basis (still diagonalizable) with matrix $\mathbf{S}' = \mathbf{RS}$

$$(\nabla \mathbf{V})_{\mathcal{B}'} = \begin{bmatrix} \sqrt{1-\beta_3^2} & -2\beta_3 \\ 0 & -\sqrt{1-\beta_3^2} \end{bmatrix}$$

$$\begin{aligned} (\nabla \mathbf{V})_{\mathcal{B}'} &= \mathbf{R}(\nabla \mathbf{V})_{\mathcal{B}}\mathbf{R}^{\mathsf{T}} \\ &= \mathbf{RS}(\mathbf{D})\mathbf{S^{-1}}\mathbf{R}^{\mathsf{T}} \\ &= \mathbf{S}'(\mathbf{D})\mathbf{S}'^{\mathbf{-1}}. \end{aligned}$$

Therefore the evolution operator (matrix deformation) in the mixed flow is

$$\mathrm{F(t)} = \exp\left((\nabla \mathbf{V})_{\mathcal{B}'} t\right) = \begin{bmatrix} \exp\left(t\sqrt{1-\beta_3^2}\right) & -\frac{\beta_3 \exp\left(-t\sqrt{1-\beta_3^2}\right)\left(-1+\exp\left(2t\sqrt{1-\beta_3^2}\right)\right)}{\sqrt{1-\beta_3^2}} \\ 0 & \exp\left(-t\sqrt{1-\beta_3^2}\right) \end{bmatrix}$$

In our 2D problem, one particle and its images are represented by a mathematical lattice of N points in $\mathbb{R}^2$, where each point has vector position

$$\mathbf{l}_{\mathrm{i}}(0) = \mathrm{N_{i1}}\mathbf{l}_1(0) + \mathrm{N_{i2}}\mathbf{l}_2(0) \qquad \mathrm{N_{ij}} \in \mathbb{Z},\ \forall \mathrm{i} = 1, \ldots, \mathrm{N}$$

that is a linear combination of two vector basis $\mathbf{l}_1(0), \mathbf{l}_2(0)$ with integer coefficients $\mathrm{N_{i1}}, \mathrm{N_{i2}}$ with $\mathrm{i} = 1, \ldots, \mathrm{N}$. The vector basis are the dimensions of the simulation box and are grouped together in the matrix

$$\mathbf{L}(0) = \begin{bmatrix} \mathbf{l}_1(0) & \mathbf{l}_2(0) \end{bmatrix}.$$

We look for a rectangular box, tilted of an angle ϑ, thus the basis vectors have the following components

$$\mathbf{l}_1(0) = \begin{bmatrix} a\cos\vartheta \\ a\sin\vartheta \end{bmatrix} \qquad \mathbf{l}_2(0) = \begin{bmatrix} -\sin\vartheta \\ \cos\vartheta \end{bmatrix}$$

where a is the *aspect ratio* and ϑ the *tilting angle*, called "magic angle" in a popular sense.

A grid of points, undergoing mixed motion, is *reproducible* if there exist a time $\exists\, \tau_p \in \mathbb{R}^+$ *s.t.* $\forall \mathrm{i} = 1, \ldots, \mathrm{N}\ \exists\, (\mathrm{N_{i1}}, \mathrm{N_{i2}}) \in \mathbb{Z}^{2\mathrm{N}}$ for some j

$$\mathbf{l}_{\mathrm{i}}(\tau_p) = \mathbf{l}_{\mathrm{j}}(0) = \mathrm{N_{i1}}\mathbf{l}_1(0) + \mathrm{N_{i2}}\mathbf{l}_2(0)$$

At the same time, the evolution at time τ_p of the lattice point individuated by $\mathbf{l}_\mathrm{i}$ is

$$\begin{aligned}\mathbf{l}_\mathrm{i}(\tau_p) &= \exp\left((\nabla\mathbf{V})_{\mathcal{B}'}\tau_p\right)\mathbf{l}_\mathrm{i}(0)\\ &= \mathbf{S}'\exp\left(\mathbf{D}\tau_p\right)\mathbf{l}'_\mathrm{i}(0)\end{aligned}$$

by placing $\mathbf{l}'_\mathrm{i}(t) = \mathbf{S}'^{-\mathbf{1}}\mathbf{l}_\mathrm{i}(t)$ and $\mathbf{L}'(0) = \mathbf{S}'^{-\mathbf{1}}\mathbf{L}(0)$, therefore a lattice is reproducible if $\exists\tau_p \in \mathbb{R}^+$ s.t. $\forall \mathrm{i} = 1,\dots,\mathrm{N}$, $\exists\,(\mathrm{N_{i1}},\mathrm{N_{i2}}) \in \mathbb{Z}^{2\mathrm{N}}$:

$$\exp\left(\mathbf{D}\tau_p\right)\mathbf{l}'_\mathrm{i}(0) = \mathrm{N_{i1}}\mathbf{l}'_1(0) + \mathrm{N_{i2}}\mathbf{l}'_2(0).$$

Previous equation must be valid for all lattice points, generated by the basis vectors, thus it is sufficient to impose the condition on the generators, that means

$$\exp\left(\mathbf{D}\tau_p\right)\mathbf{l}'_\mathrm{i}(0) = \mathrm{N_{i1}}\mathbf{l'}_1(0) + \mathrm{N_{i2}}\mathbf{l'}_2(0) \qquad \mathrm{i} = 1,2$$

and making some manipulations we obtain the following *eigenvalue problem* for each component of $\mathbf{l}'(0)$

$$(\mathbf{N} - \lambda_\mathrm{z}\mathbf{I})(\mathbf{l'}_\mathrm{z}(0)) = 0 \qquad\qquad \mathrm{z} = \mathrm{x},\mathrm{y}$$

where $\mathbf{l'}_x(0) = \mathbf{l}'(0)\cdot\mathbf{e}_1$ and $\mathbf{l'}_y(0) = \mathbf{l}'(0)\cdot\mathbf{e}_2$, $\mathbf{N} \in \mathrm{SL}(2;\mathbb{Z})$, $\lambda_\mathrm{x} = \exp(\tau_p\sqrt{1-\beta_3^2})$ and $\lambda_\mathrm{y} = \frac{1}{\lambda_\mathrm{x}}$.

Through some restrictions on the sum of eigenvalues $k = \lambda_\mathrm{x} + \frac{1}{\lambda_\mathrm{x}}$ and the coefficients $\mathrm{N_{ij}}$ we have been able to find the conditions on a, ϑ and τ_p to have reproducibility (and compatibility) of the lattice points.

To sum up, for $0 < \beta_3 < 1$ we obtain

$$a = \frac{2\sqrt{1-\beta_3^2}}{\sqrt{4+\beta_3^2} - \sqrt{5}\beta_3}$$

$$\vartheta = \frac{1}{2}\left(\arccos\beta_3 - \arcsin\left(\frac{\sqrt{1-\beta_3^2}}{\sqrt{5}}\right)\right)$$

$$\tau_p = \frac{\log\left(\frac{3+\sqrt{5}}{2}\right)}{\sqrt{1-\beta_3^2}}$$

and therefore choosing carefully an initial simulation box, with aspect ratio a tilted with an angle ϑ, we will obtain reproducibility in mixed conditions of motion at a period of time τ_p, thus at that time the simulation box will be re-initialized and the simulation can proceed for the same period of time and, in a wide perspective, for an indefinite amount of time.

While for $\beta_3 = 0$ the method is still valid, $\beta_3 = 1$ represents a singular limit of the method. In Figure 5 are reported the results in the case of $\beta_3 = 0.6$.

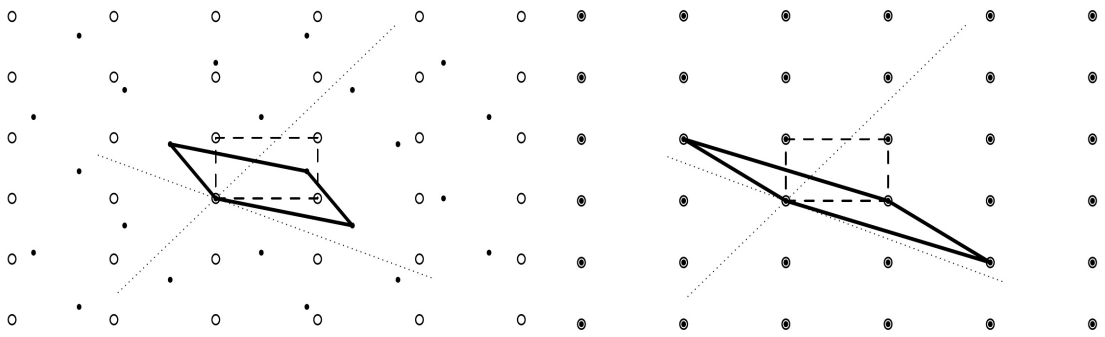


Figure 5 Box deformation at time $t \neq \tau_p$ (left) and at $t = \tau_p$ (right) with $\beta_3 = 0.6$, $a = 2.143563$, $\vartheta = 0.28070777$ and $\tau_p = 1.20302956$. (Figures done with asymptote https://asymptote.sourceforge.io/)

6 Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS)

LAMMPS is a classical Molecular Dynamics software, used especially in Material Sciences and Biophysics to investigate properties of liquids or solids, simulating their motions under different conditions of flow and geometry, subjecting the fluid to forces of different nature.

It is *Open source* (under GPL licence) written in highly portable C++ language. It is a particle simulator at different length and time scales (from electrons, to atomistic, coarse-grained and continuum scale) and uses spatial decomposition of simulation domain for *parallelism*, doing energy minimization, computing the non-equilibrium dynamics. There exist many other MD softwares with similar characteristics: AMBER, NAMD, GROMACS,...

MD data are useful to provide local measurements of material properties (in contrast to experimental measurements, usually affected by errors), such as viscosity, normal stress differences, elasticity that are often *coupled* with simulators on other different scales: QM, CFD, ... to test some macroscale algorithm or to provide a complete multi-scale tool for fluid dynamics simulations. For more detailed explanation of coupling and multi-scale methods see [10] and [9].

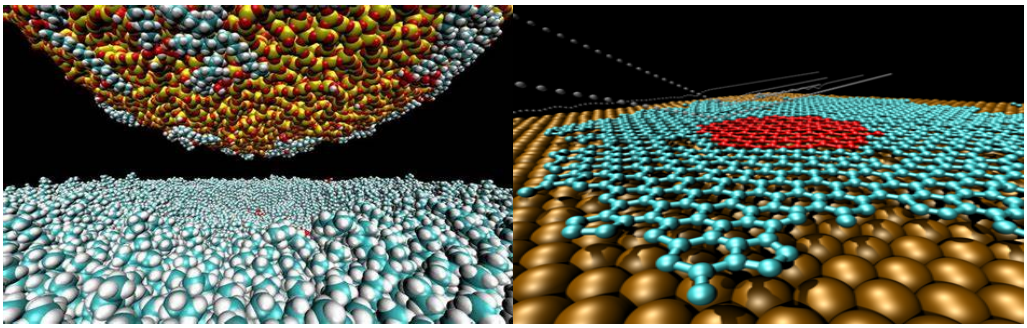


Figure 6 Representation of the molecular structures of Graphene. (Figures taken from https://lammps.sandia.gov/)

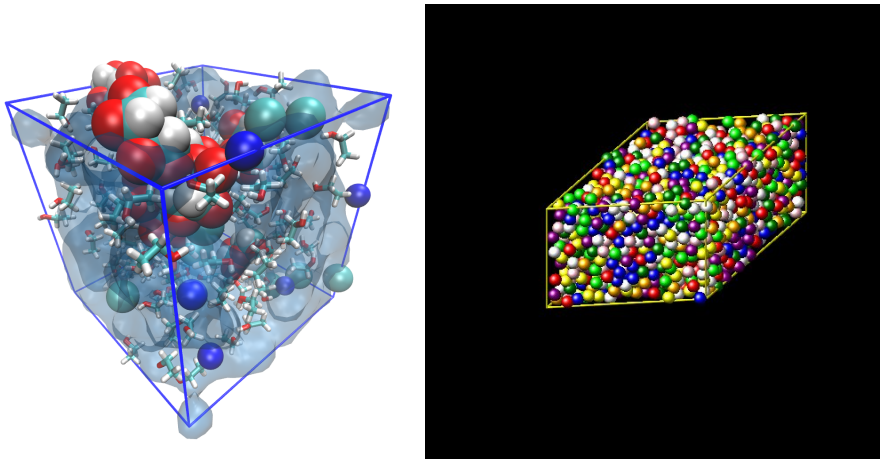


Figure 7 Molecular structure of polymeric fluids (Figures taken from https://www.h2awsm.org/)

References

[1] A. Baranyai and P.T. Cummings, *Steady state simulation of planar elongation flow by nonequilibrium molecular dynamics.* The Journal of Chemical Physics, 110/1 (1999), 42–45.

[2] M. Dobson, *Periodic boundary conditions for long-time nonequilibrium molecular dynamics simulations of incompressible flows.* The Journal of Chemical Physics, 141/18 (2014), 184103.

[3] D. Frenkel and B. Smit, "Understanding molecular simulation: from algorithms to applications. Volume 1". Elsevier, 2001.

[4] G.G. Giusteri and R. Seto, *A theoretical framework for steady-state rheometry in generic flow conditions.* Journal of Rheology, 62/3 (2018), 713–723.

[5] T.A. Hunt, *Periodic boundary conditions for the simulation of uniaxial extensional flow of arbitrary duration.* Molecular Simulation, 42/5 (2016), 347–352.

[6] A. Kraynik and D. Reinelt, *Extensional motions of spatially periodic lattices.* International Journal of Multiphase Flow 18/6 (1992), 1045–1059.

[7] L.D. Landau and E.M. Lifshitz, "Fluid Mechanics, volume 6 of Course of Theoretical Physics". Pergamon, second edition, 1987. URL http://www.worldcat.org/isbn/9781483161044.

[8] A. Lees and S. Edwards, *The computer study of transport processes under extreme conditions.* Journal of Physics C: Solid State Physics, 5/15 (1972), 1921.

[9] S. Stalter, L. Yelash, N. Emamy, A. Statt, M. Hanke, M. Lukáčová-Medvidová, and P. Virnau, *Molecular dynamics simulations in hybrid particle-continuum schemes: Pitfalls and caveats.* Computer Physics Communications 224 (2018), 198–208.

[10] F. Tedeschi, G.G. Giusteri, L. Yelash, and M. Lukáčová-Medvidová, *A multi-scale method for complex flows of non-newtonian fluids.* arXiv preprint arXiv:2103.10161 (2021).

[11] B. Todd and P.J. Daivis, *Nonequilibrium molecular dynamics simulations of planar elongational flow with spatially and temporally periodic boundary conditions.* Physical Review Letters 81/5 (1998), 1118.

[12] B. Todd and P.J. Daivis, *Homogeneous non-equilibrium molecular dynamics simulations of viscous flow: techniques and applications.* Molecular Simulation 33/3 (2007), 189–229.

Topics in Numerical Linear Algebra for High-Performance Computing

Monica Dessole (*)

Abstract. As computer architectures evolve, numerical algorithms have to face high resolution simulations and data integration that are now key to many research fields. Solution methods for real world problems require a constantly increasingly computational effort, demanding for more and more resources and tailored algorithms. In this talk we will introduce some recent high-performance algorithmic developments about solving dense linear systems, possibly numerical rank-deficient, and we will present some parallel techniques for the solution of large-scale sparse systems of linear equations.

1 Introduction

Real world problems are large scale and their numerical solution boils down to linear algebra problems, examples range from PDEs simulation to optimization and more. Some recent applications need (almost) real-time responses from algorithms, motivating the development of high-performance general purpose codes. It is important to stress that the effort for high quality product code is way larger than the effort for prototype code. As an example, the widely used MATLAB's backslash requires about 200 000 lines of code. Efficient numerical linear algebra code is build on top of Basic Linear Algebra Subprograms (BLAS), which are grouped in three categories:

- BLAS1: vector-vector operations;
- BLAS2: matrix-vector operations;
- BLAS3: matrix-matrix operations,

In the context of high-performance coding, the classical operation count alone may not be a good indicator of the efficiency of an algorithm. In order to close this cap, it is fundamental to design algorithms according to modern computer architectures while ensuring robustness. It is well known that increasing the BLAS3 fraction of work by grouping together vector operations results in a better efficiency on modern processing architectures,

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: mdessole@math.unipd.it . Seminar held on 21 Aprile 2021.

while, when dealing with parallel architectures, one should devise algorithms which are scalable, namely that have a good the capacity of handle a growing amount of work, and which can guarantee a high achieved occupancy, that is the fraction of active computing units (cores) over the upper limit. The former task will be discusses in Section 2, while we will focus on the latter in Section 3.

2 Dense linear algebra

We seek for the solution of a linear systems of equation, or at least an approximate solution. Let A be matrix of size $m \times n$, $\mathbf{b}$ a vector of length m. We seek for $\mathbf{x}$ such that

$$A\mathbf{x} = \mathbf{b} \quad (\text{or } A\mathbf{x} \approx \mathbf{b}).$$

If the matrix A is invertible, then the solution is $\mathbf{x} = A^{-1}\mathbf{b}$, but in finite precision we do not explicitly compute A^{-1}. Instead, we try to derive an "easier" equivalent system with the same solution by means of matrix decompositions. In this talk, the QR factorization is be the common thread, and therefore we here recall some properties. We denote the singular values of a matrix A by

$$\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{\min}(A) = \sigma_{\min(m,n)}(A) \geq 0.$$

Suppose matrix A has full column rank, implying $m \leq n$. Then there exists an orthogonal matrix Q of order m such that

$$A = QR = Q \begin{pmatrix} R_{11} \\ \mathbb{O} \end{pmatrix}, \quad Q^T Q = QQ^T = \mathbb{I},$$

where R_{11} is upper triangular $n \times n$ with positive diagonal elements. We have

$$A = Q \begin{pmatrix} R_{11} \\ \mathbb{O} \end{pmatrix} = (Q_1 \; Q_2) \begin{pmatrix} R_{11} \\ \mathbb{O} \end{pmatrix} = Q_1 R_{11},$$

which yields the so-called thin QR factorization, which is unique. Consider the column partitioning $A = (\mathbf{a_1}, \ldots, \mathbf{a}_n)$ and $Q = (\mathbf{q_1}, \ldots, \mathbf{q}_m)$, then

$$\operatorname{span}(\mathbf{a_1}, \ldots, \mathbf{a}_k) = \operatorname{span}(\mathbf{q_1}, \ldots, \mathbf{q}_k), \quad k \leq n.$$

In particular, if $\mathcal{R}(A) = \operatorname{span}(\mathbf{a_1}, \ldots, \mathbf{a}_n)$, we have

- $\mathcal{R}(Q_1) = \mathcal{R}(A)$,
- $\mathcal{R}(Q_2) = \mathcal{R}(A)^{\perp}$.

The QR decomposition is usually computed as a product of Householder reflectors, which are orthogonal matrices of the form

$$H = \mathbb{I} - 2\frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}}.$$

Given a vector $\mathbf{x} \in \mathbb{R}^n$, if we set $\mathbf{v} = \mathbf{x} - \|\mathbf{x}\|_2 \mathbf{e}_1$, then $H\mathbf{x} = \|\mathbf{x}\|_2 \mathbf{e}_1$, where $\mathbf{e}_1$ is the first element of the canonical basis of $\mathbb{R}^n$. Therefore, we can exploit Householder reflectors to zero out column by column the entries of the matrix A below the diagonal, see Figure 1.

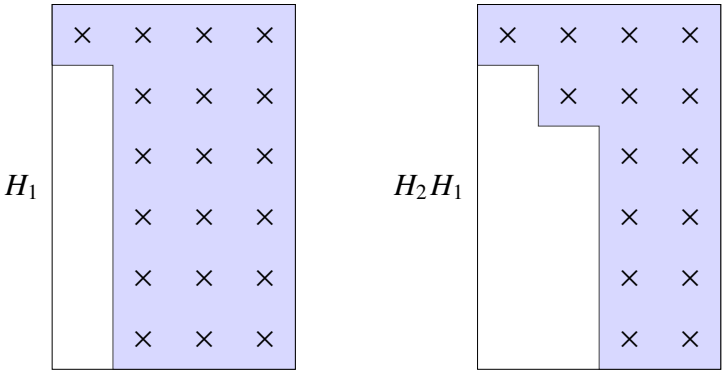


Figure 1 Householder QR.

2.1 Rank-deficient problems

Suppose now that A is a matrix of size $m \times n$, with rank $\text{rank}(A) = r < \min(m, n)$, and let $\mathbf{b}$ be a vector of length m. We consider the least squares problem, i.e. we aim at finding $\mathbf{x}$ that solves

$$\min_x \|A\mathbf{x} - \mathbf{b}\|^2. \tag{1}$$

When the matrix of the objective function A is not full column rank, the QR decomposition is not unique, and neither the solution is: if $\mathbf{x}^\star$ solves (1), then

$$\|A(\mathbf{x}^\star + \mathbf{y}) - \mathbf{b}\|^2 = \|A\mathbf{x}^\star - \mathbf{b}\|^2$$

for any $\mathbf{y} \in \mathcal{N}(A) = \{\mathbf{x} : \ A\mathbf{x} = 0\} \neq \emptyset$. First, it is necessary to give some additional constraints to identify the solution. If one looks for the minimum ℓ_2-norm solution, then the gold standard is the SVD decomposition. However, there are alternatives based on modifies QR decompositions: If we are able to find $r = \text{rank}(A)$ linearly independent columns of A, namely $\mathbf{a}_{j_1} \ldots, \mathbf{a}_{j_r}$, then

$$A\Pi = (Q_1 \ Q_2) \begin{pmatrix} R_{11} & R_{12} \\ \mathbb{O} & \mathbb{O} \end{pmatrix},$$

where R_{11} is upper triangular of order r, and Π is a permutation matrix that moves $\mathbf{a}_{j_1} \ldots, \mathbf{a}_{j_r}$ to the leftmost positions. Thus, a solution is given by

$$\mathbf{x}^\star = \Pi \begin{pmatrix} R_{11}^{-1} Q_1 \mathbf{b} \\ 0 \end{pmatrix}.$$

This solution has nonzero entries only in positions $j_1, \ldots, j_r$, and in general it is not the minimum ℓ_2-norm solution. A solution of this kind is usually referred as *basic* solution. A Rank-Revealing QR (RRQR) factorization is

$$A\Pi = QR = Q \begin{pmatrix} R_{11} & R_{12} \\ \mathbb{O} & R_{22} \end{pmatrix},$$

where Q is orthogonal, R_{11} is upper triangular of order r and

$$\sigma_{\min}(R_{11}) \gg \|R_{22}\| = \mathcal{O}(\varepsilon),$$

where ε is the working precision. For a fixed rank r, the best of such family of factorization is clearly the solution of the maximization problem

$$\max_{\Pi} \sigma_{\min}(R_{11}). \tag{2}$$

However, the problem above clearly has a combinatorial nature, and we belive it cannot be solved in a polynomial time. Therefore, we rather require

$$\sigma_{\min}(R_{11}) \geq \frac{\sigma_r(A)}{p(n)}, \tag{3}$$

where $p(n)$ is a low degree polynomial in n. The state-of-art algorithm to solve this problem is the so-called QR with column pivoting [7], and it is a greedy strategy for approximate solving problem (2).

Algorithm 1 QR with column pivoting $QRP(A)$

1: **for** $s = 1, \ldots, n-1$ **do**
2: $\quad j = \text{argmax}_i \|\mathbf{c}_i\|$
3: $\quad$ Exchange columns s and $s+j$ of $R^{(s-1)}$
4: $\quad$ Compute and apply the Householder reflector to get $R^{(s)}$
5: **end for**

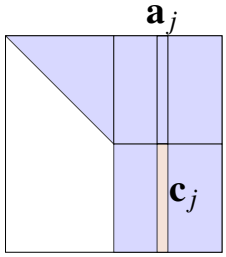


Figure 2 Column pivoting strategy.

The procedure is shown in Algorithm 1. At the s-th step, we have already reduced to triangular form $s-1$ columns which form the block $R_1^{(s-1)}$ by means of the orthogonal

transformation $Q^{(s)} = (Q_1^{(s)}\ Q_2^{(s)})$, where the columns of $Q_1^{(s)}$ span the same subspace as that spanned by the columns of $R_1^{(s-1)}$. We identity the s-th column pivot as the column $\mathbf{a}_j$ with the largest norm of $\mathbf{c}_j$, where $(Q_1^{(s)})^T\mathbf{a}_j = \mathbf{b}_j + \mathbf{c}_j$, with $\mathbf{b}_j \in \mathbb{R}^{s-1}$ and $\mathbf{c}_j \in \mathbb{R}^{m-s+1}$, as shown in Figure 2. Then we compute and apply the Householder transformation, by means of BLAS2 operation.

Remark 1 The choice of the pivot in step 2 may have the following geometric interpretation. Consider the block column partitions of the partial factorization ar step s, namely $R^{(s)} = (R_1^{(s)}\ R_2^{(s)})$, $Q^{(s)} = (Q_1^{(s)}\ Q_2^{(s)})$. Then the most linearly independent column $\mathbf{a}_j$ can be identified as the column with the largest projection on the orthogonal complement on the space spanned by the columns already processed, namely $\mathcal{R}(R_1^{(s)})^\perp = \mathcal{R}(Q_2^{(s)})$. In formulae, the column $\mathbf{a}_j$ solves

$$\max_{j\geq s} \left\| \mathcal{P}_{\mathcal{R}(R_1^{(s)})^\perp} \mathbf{a}_j \right\| = \max_{j\geq s} \left\| Q_2^{(s)} \mathbf{c}_j \right\|.$$

However, the product $Q_2^{(s)}\mathbf{c}_j$ is not directly available, then we settle for the solution of $\max_{j\geq s} \|\mathbf{c}_j\|$.

2.2 QR decomposition with Deviation Maximization pivoting

We would like to substitute step 2 of Algorithm 1 with a block pivoting strategy, in order to exploit BLAS3 operations when applying block Householder vectors. A possible strategy is to pick k columns with pairwise large angles, within those columns which are the most linearly independent from $\mathcal{R}(R_1^{(s)})$. This is precisely what is done by the Deviation Maximization (DM) strategy [4]. Let us introduce the notion of cosine matrix.

Definition 1 Let $C = (\mathbf{c}_1\ \dots\ \mathbf{c}_k)$ be an $m \times k$ matrix with non-zero columns. The **cosine matrix** Θ has entries

$$\theta_{ij} = \frac{\mathbf{c}_i^T \mathbf{c}_j}{\|\mathbf{c}_i\|\|\mathbf{c}_j\|} = \cos(\alpha_{ij}), \quad \alpha_{ij} = \alpha(\mathbf{c}_i, \mathbf{c}_j), \qquad 1 \leq i, j \leq k.$$

Before describing the procedure, let us provide the following result.

Lemma 1 *Take $C = (\mathbf{c}_1\ \dots\ \mathbf{c}_k)$, such that $\|\mathbf{c}_j\| \geq \tau \max_i \|\mathbf{c}_i\|$, $1 \geq \tau > 0$, for all $1 \leq j \leq k$. If the cosine matrix Θ associated to C is a strictly diagonally dominant matrix with*

$$\delta = \min_i (1 - \sum_{j\neq i} |\theta_{ij}|) > 1 - \tau^2 \geq 0.$$

then

$$\sigma_{\min}(C) \geq \sqrt{\delta + \tau^2 - 1}\ \max_i \|\mathbf{c}_i\|,$$

The Lemma above essentially says that the linear independence of the column vectors $C = (\mathbf{c}_1 \ \dots \ \mathbf{c}_k)$ can be controlled by tuning two thresholds, namely τ and δ. The resulting procedure, which we call QRDM, is shown in Algorithm 2.

Algorithm 2 QR with DM pivoting $QRDM(A)$

1: **while** $n_s < n$ **do**
2: $j_1 = \text{argmax}_i \|c_i\|$
3: Choose columns $j_2, \dots, j_k$ within $\{c_i : \|c_i\| \geq \tau \|c_{j_1}\|\}$ with large pairwise angles
4: Exchange columns $n_s, \dots, n_s + (k-1)$ and $n_s + j_1, \dots, n_s + j_k$ of $R^{(s-1)}$
5: Compute and apply the block Householder reflector to get $R^{(s)}$
6: $s = s + 1$, $n_s = n_s + k$
7: **end while**

This algorithm allows us to compute the new factor $R^{(s)}$ by means of BLAS3 operations. Moreover, step 3 requires to compute the cosine matrix $\Theta = \text{diag}(\|c_i\|)^{-1} C^T C \text{diag}(\|c_i\|)^{-1}$, where i belongs set of candidate columns $\{c_i : \|c_i\| \geq \tau \|c_{j_1}\|\}$, and it can be carried out by means of BLAS3 operations as well. We then look for a suitable submatrix of Θ that satisfies Lemma 1, namely

$$\delta = \min_i \left(1 - \sum_{j \neq i} |\theta_{ij}| \right) > 1 - \tau^2 > 0, \quad i, j \in \{j_1, \dots, j_k\}.$$

We now try to derive estimates in the form (3) in order to quantify the quality of the factorizarions obtained with QRP and QRDM algorithms. Let $\bar{\sigma}^{(s)}$ be the smallest singular value of the computed $R_{11}^{(s)}$ block at step s, that is

$$\bar{\sigma}^{(s)} = \sigma_{\min} \left(R_{11}^{(s)} \right).$$

We have the following result for Algorithm 1.

Theorem 1 *The standard pivoting guarantees*

$$\bar{\sigma}^{(s+1)} \geq \sigma_{s+1}(A) \frac{\bar{\sigma}^{(s)}}{\sigma_1(A)} \frac{1}{\sqrt{2(n-s)(s+1)}}.$$

This shows that QRP indeed gives a RRQR decompotisions. However, even if the leading s columns have been carefully selected, so that $\bar{\sigma}^{(s)}$ is an accurate approximation of $\sigma_s(A)$, there could be a potentially dramatic loss of accuracy in the estimation of the successive singular value $\sigma_{s+1}(A)$. In fact, it is well known that QRP algorithm failure may occur, as well as for other greedy algorithms, but it is very unlikely in practice, and in fact this algorithm is implemented in state-of-art libraries, such as LAPACK. We have an analogous result for Algorithm 2.

Theorem 2 *The DM pivoting guarantees*

$$\bar{\sigma}^{(s+1)} \geq \sigma_{n_{s+1}}(A)\frac{\bar{\sigma}^{(s)}}{\sigma_1(A)}\frac{1}{\sqrt{2(n-n_{s+1})n_{s+1}}}\frac{\sqrt{\delta+\tau^2-1}}{k^2 n_s}.$$

Therefore, even if efficiency may improve by adopting the QRDM strategy, we cannot guarantee more accuracy than that provided by the QRP strategy.

2.3 NonNegativite Least Squares problem

Let us now introduce a slightly different family of problems. Let A be a matrix of size $m \times n$ and $\mathbf{b}$ a vector of length m. The NonNegativite Least Squares problem (NNLS) consists in finding $\mathbf{x}^\star$ that solves

$$\min \|A\mathbf{x} - \mathbf{b}\|^2 \quad \text{subject to } \mathbf{x} \geq 0.$$

Within the existing algorithms, many require the objective function to be strictly convex, that is the Hessian matrix A^TA should be positive definite. However, if A has not full column rank then A^TA is only positive semidefinite. An active set method due to Lawson and Hanson is not affected by this issue. Recall first that a constraint is said to be active at $\mathbf{x}$ if

$$x_i = 0,$$

otherwise it is called passive, namely

$$x_i > 0.$$

The Lawson-Hanson algorithm is based on the following observation. If the active set $Z = \{j : \ x_j^\star = 0\}$ and its complement $P = \{j : \ x_j^\star > 0\}$ are known at a solution $\mathbf{x}^\star$, then $\mathbf{x}^\star$ solves the unconstrained least squares problem

$$\mathbf{x}_P^\star = \text{argmin}_{\mathbf{x}}\|A_P\mathbf{x} - \mathbf{b}\|^2, \quad \mathbf{x}_Z^\star = 0,$$

where $A_P = \{a_{ij}\}$, $1 \leq i \leq m$, $j \in P$. The procedure, which we briefly call LH, is presented in Algorithm 3.

Algorithm 3 Lawson-Hanson active set $LH(A, \mathbf{b})$

```
P = ∅, Z = {1, ..., n}, x = 0
while the optimum has not been reached do
    Move from Z to P the index j* = argmax(A^T(Ax − b))_j
    Update: x_P = argmin‖A_P y − b‖, x_Z = 0
    while min(x_P) ≤ 0 do
        Move from P to Z the unfeasible indices {j ∈ P : x_j ≤ 0}
        Downdate: x_P = argmin‖A_P x − b‖, x_Z = 0
    end while
end while
```

A least squares problem needs to be solved each time we reach step 4 and step 7. An efficient implementation can be obtained by exploiting QR updates/downdates. However, since one index $j^\star$ is selected in step 4, we can only exploit rank-1 QR updates in step 4 by means of BLAS2 operations. In order to exploit low rank updates by means of QR decomposition, we use the Deviation Maximization pivoting in step 3 to identify a subset J of indices. We call the overall procedure Lawson-Hanson with Deviation Maximization, or briefly LHDM.

2.4 Application to Tchackaloff-Caratheodory regression

Let us consider the application that mainly motivates the development of LHDM, see [5]. Let $X = \{x_1, \dots, x_M\}$ be a set on a compact $K \subset \mathbb{R}^d$, $d \geq 2$, and suppose X is $\mathbb{P}_n^d$-determining, that is

$$p \in \mathbb{P}_k^d,\ p(X) = 0 \Rightarrow p \equiv 0 \text{ on } K.$$

The following result is due to Tchakaloff and it is a cornerstone in quadrature theory.

Theorem 3 *Let $\mathbf{u} = \{u_1, \dots, u_M\} \geq \mathbf{0}$ be the weights of a measure supported at X. If $M = card(X) > N_k = dim(\mathbb{P}_k^d)$, we can find $\{t_1, \dots, t_m\}$ points of X and weights $\mathbf{w} = \{w_1, \dots, w_m\} \geq 0$ such that*

$$\sum_{i=1}^{M} u_i\, p(x_i) = \sum_{j=1}^{m} w_j\, p(t_j), \quad \forall p \in \mathbb{P}_k^d, \quad m \leq N_k.$$

In other words, for any discrete measure supported at X, there exists a subset of X and proper weights such that we have the same quadrature error on the polynomials up to a suitable degree. The same theorem can be applied to obtain polynomial regression at degree n, by taking uniform weights $\mathbf{u} = (1, \dots, 1)^T$ and $k = 2n$. This is interesting in applications because the compression ratio M/m may be large. The proof is based on Caratheodory theorem and it is not constructive, thus we need to formulate an algebraic version of Tchakaloff theorem. Fix a polynomial basis $span(p_1, \dots, p_{N_k}) = \mathbb{P}_k^d$ of $\mathbb{P}_n^d$. Then Tchakaloff theorem can be stated as the existence of a sparse nonnegative solution $\mathbf{w}$ with at most $m \leq N_k$ non-zero components to the underdetermined linear system

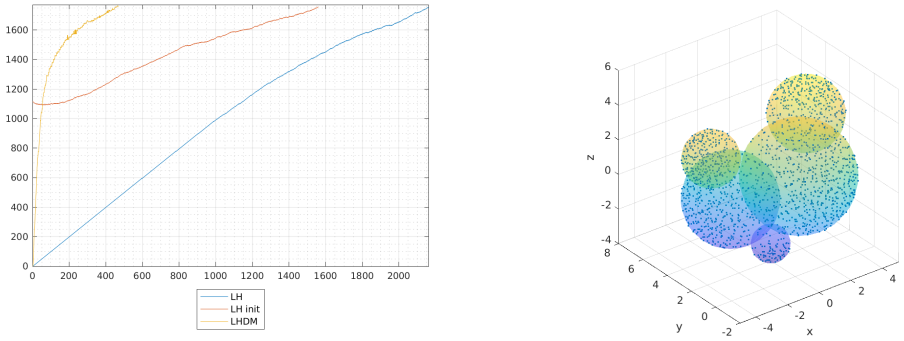
$$V_k^T \mathbf{w} = \mathbf{b}, \quad V_k = (p_j(x_i)) \in \mathbb{R}^{M \times N_k}, \mathbf{b} = V_k^t \mathbf{u} \in \mathbb{R}^M.$$

A way to solve this linear system is to use the Non Negative Least Squares (NNLS) formulation: $\min \|V_k^T \mathbf{w} = \mathbf{b}\|_2^2$, subject to $\mathbf{w} \geq 0$ Tchakaloff-Carathodory regression has been implemented in the package dCATCH [6]. We show numerical results of regression at degree 10 ($k = 2 \cdot 10 = 20$) on a compact subset of $\mathbb{R}^3$ obtained as the union of some balls with nonempty intersection. Figure 3b shows the domain with 1763 compressed Tchakaloff points, extracted from 18,915 original point. Figure 3a shows the evolution of the cardinality of the passive set P among the iterations of three different executions of Lawson-Hanson, namely the standard LH (blue line), the LHDM (yellow line), and finally a version of the Lawson-Hanson algorithm with nonempty initialization of the passive set

by means of unconstrained least squares, which we briefly call LH-init (orange line). The numerical results are shown in Table 1: *cpts* is the number of compressed Tchakaloff points and *momerr* is the final moment residual; $compr = M/mean(cpts)$ is the mean compression ratio obtained by the three methods listed; t_{LH}/t_{LHDM} (t_{LHI}/t_{LHDM}) is the ratio between the execution time of LH (LH-init) and that of LHDM.

Test		compr.	LH			LH-init			LHDM	
k	M	ratio	t_{LH}/t_{LHDM}	cpts	momerr	t_{LHI}/t_{LHDM}	cpts	momerr	cpts	momerr
20	18915	11	2.7	1755	3.4e-8	3.2	1758	3.2e-8	1755	1.5e-8

Table 1 Results for the multibubble numerical test.



(a) (b)

Figure 3 Multibubble test case, regression degree $n = 10$.

3 Direct Sparse Linear Algebra

Let A be a sparse matrix of order n, $\mathbf{b}$ a vector of length n. We look for a solution $\mathbf{x}$ of the sparse linear system $A\mathbf{x} = \mathbf{b}$. A matrix A is said to be sparse when the number of nonzero element is way smaller then the total number of entries. Standard matrix decompositions (LU, QR...) usually yield dense factors. Therefore, we either try to incrementally reach a solution by means of iterative methods, or we have to carefully design a solver trying to minimize the fill-in. Define the sparsity pattern as the set of pairs of indices corresponding to nonzero entries, namely

$$\mathcal{S}(A) = \{(i, j) : \ a_{ij} \neq 0\}.$$

A fill-in is a non-zero entry in the factors of a matrix decomposition of A in a position (i, j) that is not in $\mathcal{S}(A)$. Sparse systems do not only come from numerical methods for PDEs, in fact we are going to focus of a family of sparse matrices with block structure that arise

from optimal control problems. An Optimal Control Problem consists in finding a control $\mathbf{u}(s)$ that minimises the functional

$$J(\mathbf{x}, \mathbf{u}) = \int_{s_a}^{s_b} f(\mathbf{x}(s), \mathbf{u}(s))\, ds \tag{4}$$

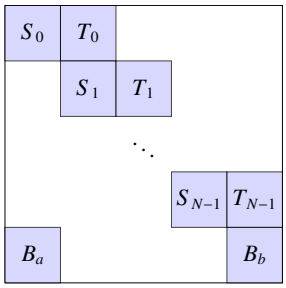
subject to

$$\begin{aligned} &\mathbf{A}(\mathbf{x}(s))\dot{\mathbf{x}}(s) + \mathbf{b}(\mathbf{x}(s), \mathbf{u}(s)) = 0, && s \in (s_a, s_b) \\ &\mathbf{c}(\mathbf{x}(s_a), \mathbf{x}(s_b)) = 0 && s \in (s_a, s_b) \\ &\mathbf{d}(\mathbf{x}(s), \mathbf{u}(s)) \leq 0 && s \in (s_a, s_b) \end{aligned}$$

Indirect methods (first optimize, then discretize) yield to a large nonlinear system $\Psi(\mathbf{Z}) = 0$, where $\mathbf{Z}$ contains the discretized unknowns. Solution methods, e.g. Newton, require the solution of a sequence of linear systems with matrix

$$J_{ij}^{(k)} = \frac{\partial \Psi_i}{\partial Z_j}(\mathbf{Z}^{(k)}).$$

The Jacobian $J^{(k)}$ has a block structure and it is a so called Bordered Almost Block Diagonal (BABD)



where S_i, T_i, B_a, B_b are square blocks of order n. Such a matrix has $2(N+1)n^2$ nonzeros over $(N+1)^2n^2$ entries. Let us describe a parallel algorithm due to Wright [8], the Structured Orthogonal Factorization (SOF). First divide the BABD system into P slices and assign each slice to one processor. Each processor will perform the same local factorization on each slice. Let us consider the first slice and suppose it has five block row equations. We proceed as follows:

- Find Q_0 orthogonal such that

$$\begin{bmatrix} T_0 \\ S_1 \end{bmatrix} = Q_0 \begin{bmatrix} U_0 \\ \mathbb{O} \end{bmatrix},$$

  and apply Q_0^T to to the whole subsystem, see Figure 4a. We get with the equivalent system in Figure 4b.

- With reference to Figure 4b find Q_1 orthogonal such that

$$\begin{bmatrix} \overline{W}_1 \\ S_2 \end{bmatrix} = Q_1 \begin{bmatrix} U_1 \\ \mathbb{O} \end{bmatrix},$$

  and apply Q_1^T to the whole subsystem to get a new equivalent system.

- Repeat until all rows have been processed.

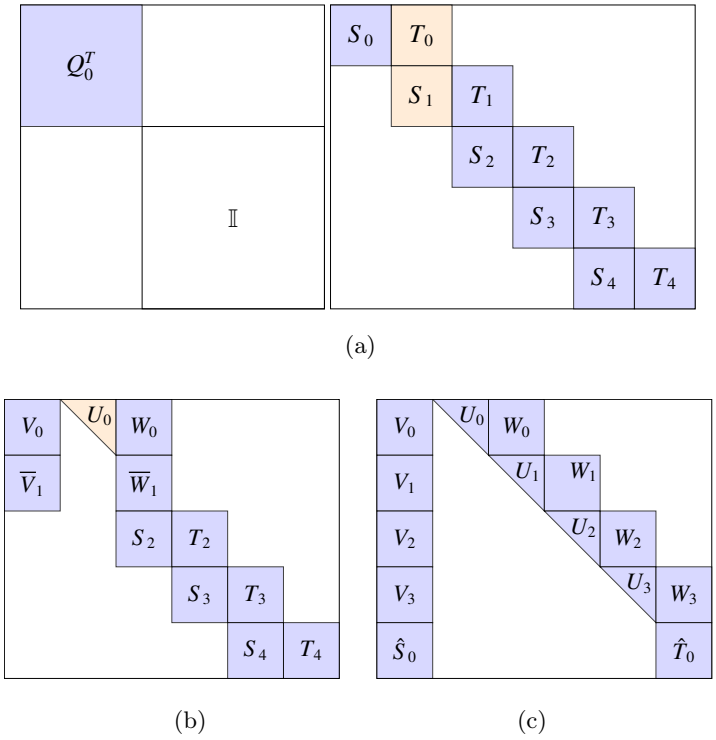


Figure 4 Local SOF, multiplication by the first orthogonal factor (a), equivalent system at step 1 (b), equivalent system at the last step (c).

At the last step, we get the equivalent subsystem shown in Figure 4c. Notice that the such a system can be solved in two steps:

(a) solve the last block equation, involving only two block unknowns, namely

$$\hat{S}_0\mathbf{x}_0 + \hat{T}_0\mathbf{x}_5 = \hat{\mathbf{b}}_0;$$

(b) retrieve the remaining unknowns by solving with back-substitution the following triangular systems for $i = 4, 3, 2, 1$, namely

$$U_{i-1}\mathbf{x}_i = \hat{\mathbf{b}}_{i-1} - V_{i-1}\mathbf{x}_0 - W_{i-1}\mathbf{x}_{i+1}.$$

Thus, by concatenating all local factorizations, we get an overall equivalent system that can be solved in two steps:

(a) solve the subsystem formed by the last block row of each slice, which has itself a BABD structure and therefore can be recursively reduced by SOF, until a sufficiently small system is reached and solved with, e.g., LU decomposition;

(b) retrieve the missing unknowns by back-substitution.

We can represent data dependencies of this algorithm by means of a Directed Acyclic Graph (DAG), in which each vertex represent a block equation or a block unknown and each edge represent a dependency. Figure 5 shows the data dependencies DAG of the recursive SOF on a BABD system with $N+1=9$ block equations divided into $P=4$ slices.

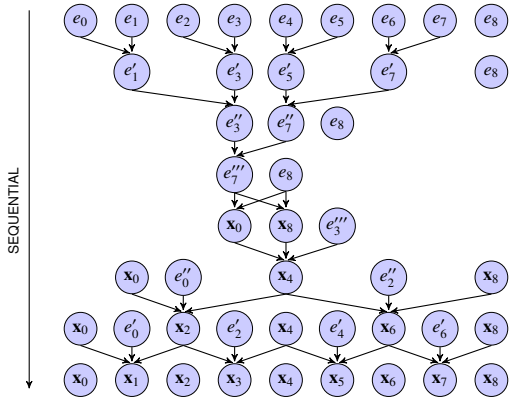


Figure 5 DAG showing data dependencies in SOF.

On the x-axis we have the amount of parallel work at each step, while on the y-axis we have sequentiality in time. In this case, the DAG takes a double cone shape, from which it is evident that the middle steps lack of parallelism causing a tremendous waste of resources on parallel machines. Our goal is therefore to rearrange computations in order to avoid the bottleneck.
Let us introduce an alternative technique, called recursive doubling. The idea is to fully decouple odd and even block unknowns at each step by deriving a double number of smaller linear systems. Suppose that $N+1$ is even, and couple each block equation with the successive one, obtaining a partition into slices consisting of two block equations each, as shown in Figure 6a. We can then obtain the solution by solving the system (5) of half the size of the original one, involving only odd block unknowns, and then retrieve the even block unknowns by back-substitution (6).

$$\hat{A}^{(e)}\mathbf{x}^{(e)} = \hat{\mathbf{b}}^{(e)}, \qquad \mathbf{x}^{(e)} = (\mathbf{x}_i), i = 0, 2, 4, \ldots \tag{5}$$

$$U_{i-1}^{(e)}\mathbf{x}_i = \hat{\mathbf{b}}_{i-1}^{(e)} - V_{i-1}^{(e)}\mathbf{x}_{i-1} - W_{i-1}, \qquad \mathbf{x}_{i+1}, i = 1, 3, 5, \ldots \tag{6}$$

On the other hand, if we couple each block equation with the previous one, as shown in Figure 6b, we can then obtain the solution by solving the system (7) of half the size of the original one, involving only even block unknowns, and then retrieve the even block unknowns by back-substitution (8).

$$\hat{A}^{(o)}\mathbf{x}^{(o)} = \hat{\mathbf{b}}^{(o)}, \qquad \mathbf{x}^{(o)} = (\mathbf{x}_i), i = 1, 3, 5, \ldots \tag{7}$$

$$U_{i-1}^{(o)}\mathbf{x}_i = \hat{\mathbf{b}}_{i-1}^{(o)} - V_{i-1}^{(o)}\mathbf{x}_{i-1} - W_{i-1}, \qquad \mathbf{x}_{i+1}, i = 0, 2, 4, \ldots \tag{8}$$

Notice that we can compute the reduced systems (5) and (7) in parallel, and apply recursively the same technique until we reach a large number of small and independent linear systems we can solve directly obtaining all block unknowns at once. This approach do not even require to compute equations (6) and (8) for back-substitution.

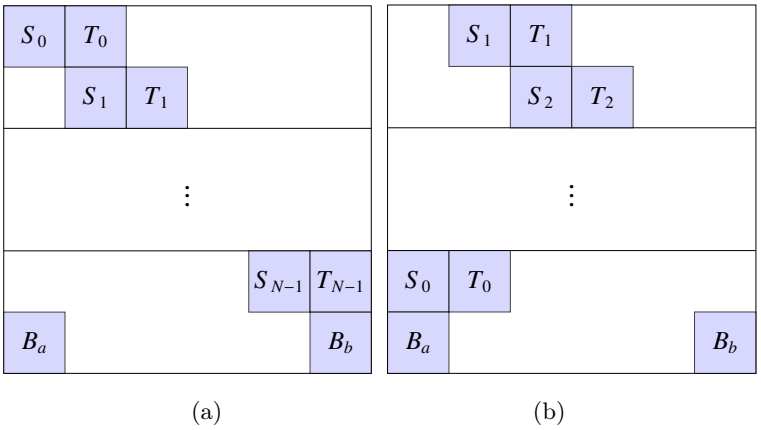


Figure 6 Recursive doubling, even (a) and odd (b) coupling.

Figure 7 shows the DAG representing data dependencies for the recursive doubling technique on a BABD system with $N + 1 = 8$ block equations divided into $P = 4$ slices.

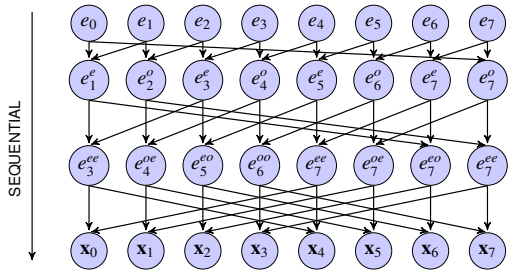


Figure 7 DAG showing data dependencies in Recursive Doubling.

This shows quite clearly that the amount of work is now constant among the algorithmic steps, thus the work can be equally distributed among the processors. In order to run this

algorithm with actual parallelism we need at least $P = N + 1$ processors, however in a real application it is often the case that $N \gg P$, thus we implement an hybrid version which we call PARAllel Structured Orthogonal Factorization (PARASOF) [3]. The procedure is organized as follows:

(a) apply SOF to reduce the BABD system to a suitable number P of block equations;

(b) apply recursive doubling;

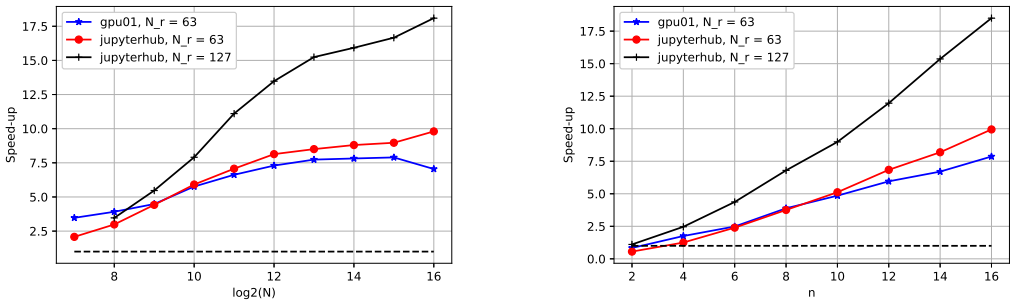
(c) apply backward substitution of SOF.

Last, we present numerical tests on BABD systems of various sizes with dense randomly generated blocks. We compare our CUDA [2] implementation of PARASOF with `babdcr` [1], a FORTAN90 package for the solution of BABD systems. All tests are performed on two different machines:

(a) **jupyterhub**,3.70GHz CPU, Nvidia TITAN V GPU with 5120 CUDA cores;

(b) **gpu01**, 3.50GHz CPU, Nvidia GeForce GTX1060 GPU with 1280 CUDA cores.

Notice that each CUDA core has not to be intended as a stand-alone processor: in fact, CUDA cores are designed to execute small computations and they physically operate in groups of 32 cores called *warps*. Figure 8 shows the speed-up of PARASOF over `babdcr`, that is the ratio between the execution time of `babdcr` over that of PARASOF, in function of the number of block equations $N + 1$ for a fixed block size $n = 16$ (Figure 8a), and in function of the block size n for a fixed block-size $N \approx 2^{16}$ (Figure 8b). In the legend, N_r denotes the size of the reduced system solved with recursive doubling in PARASOF. The blue and red lines show results for $N_r + 1 = 64$ on **gpu01** and **jupyterhub** respectively, revealing similar speed-ups. However, if we choose a larger reduced size $N_r + 1 = 128$, we are able to exploit all the processors of **jupyterhub** (black line), and we achieve a maximum speed-up of about 17×.



(a) Speed-up, fixed block size $n = 16$.

(b) Speed-up, fixed block number $N \approx 2^{16}$.

Figure 8 $t_{\texttt{babdcr}}/t_{PARASOF}$, N_r = size of reduced system solved with recursive doubling.

4 Conclusions

Efficient and accurate code is fundamental for applications, and we showed some examples of how algorithms should be carefully designed not only to robustly solve the algebraic problem, but also to achieve high-performances. We did not stress that the Vandermonde-like matrices in Tchakaloff-Caratheodory regression suffer from the so-called *curse of dimensionality*, in fact

$$V_k \in \mathbb{R}^{M \times N_k}, \quad M \gg N_k = \binom{n+d}{d} \sim \frac{n^d}{d!}.$$

As n and/or d get larger, the problem quickly runs out of memory. Suitable slitting method should be devised for large scale NNLS. An interesting application of Lawson-Hanson algorithm with Deviation Maximization that will be investigated concerns the reduction of the trajectory space in model order reduction methods for large scale parameter dependent PDE problems by means of Tchakaloff compression.

References

[1] P. Amodio and G. Romanazzi, *Algorithm 859: BABDCR - a Fortran 90 package for the solution of bordered ABD linear systems.* ACM Trans. Math. Softw. 32 (2006), 597–608.

[2] N. Corporation. CUDA C Programming Guide, 2019. `https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html`. Version 10.1.

[3] M. Dessole and F. Marcuzzi, *A massively parallel algorithm for Bordered Almost Block Diagonal Systems on GPUs.* Numerical Algorithms, 2020.

[4] M. Dessole and F. Marcuzzi, *Deviation Maximization for Rank-Deficient Problems.* Manuscript in preparation, 2021.

[5] M. Dessole, F. Marcuzzi, and M. Vianello, *Accelerating the Lawson-Hanson NNLS solver for large-scale Tchakalo regression designs.* Dolomites Research Notes on Approximation 13 (2020), 20–29.

[6] M. Dessole, F. Marcuzzi, and M. Vianello, *dCATCH/A Numerical Package for d-Variate Near G-Optimal Tchakalo Regression via Fast NNLS.* Mathematics 8 (2020).

[7] G. Golub and C. Van Loan, "Matrix Computations (4th ed.)". Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013. ISBN 9781421407944.

[8] S.J. Wright, *Stable parallel algorithms for two-point boundary value problems.* SIAM J. Sci. Statist. Comput 13 (1992), 742–764.

An intuitive introduction to p-adic geometry

Yukihide Nakada (*)

Abstract. The field of p-adic numbers forms a bridge from number theory to analysis, mixing number-theoretic constructions with analytic ideas. And just as the complex numbers made possible complex analytic geometry, the p-adic numbers opened up a new, strange geometry over a number-theoretic field.
In this summary we give a very brief and informal introduction to the p-adic numbers and to p-adic geometry, also called *rigid geometry*. A standard reference for the p-adic numbers is [3], but a more accessible introduction is [2]. For rigid geometry, we reference [1].

1 Introduction to p-adic numbers

1.1 The p-adic norm

Trying to understand p-adic geometry without knowing about p-adic numbers is like trying to understand PDEs without knowing about the reals, so we start with a brief introduction to the construction and basic characteristics of our base field.

First recall the traditional construction of the real numbers $\mathbb{R}$ that we all see in real analysis:

(a) We take as given the integers $\mathbb{Z}$ with its absolute $\|\cdot\|$ which, intuitively, measures the distance of any integer from 0.

(b) We formally construct the rational numbers $\mathbb{Q}$ (in your favorite way). The absolute value on $\mathbb{Z}$ extends uniquely to a norm on $\mathbb{Q}$ if we add the natural assumption that it be multiplicative, since

$$\left\|\frac{a}{b}\cdot b\right\| = \left\|\frac{a}{b}\right\|\cdot\|b\| \Rightarrow \left\|\frac{a}{b}\right\| = \frac{\|a\|}{\|b\|}$$

(c) We then *complete* the rational numbers with respect to this norm to get $\mathbb{R}$. Filling out the rational numbers in this way makes possible topological and analytic notions like continuity, differentiability, and convergence of power series.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: yukihide.nakada@gmail.com . Seminar held on 5 May 2021.

This is great, but for a number theorist this norm doesn't capture anything we want about $\mathbb{Z}$ or $\mathbb{Q}$. Number theorists, generally speaking, care divisilibity and prime numbers and other such properties that aren't respected by the absolute value. For example,

$$\|17 - 16\| = 1$$

but 17 is prime and $16 = 2^4$; they're as close as integers can get, but as far as divisility is concerned they have nothing to do with each other! What's a number theorist to do?

That being said, it'd be disingenuous to say that we can't use traditional analytic in number theory. A subfield of number theory which does exactly this is *analytic number theory*, which manages to use traditional analytic tools to understand the large-scale behavior of the integers.
It's orthogonal to p-adic geometry so we'll only mention in passing a neat illustrative result from the theory: if the *partition function* $p(n)$ denotes the number of ways of writing an integer as a sum of smaller integers (for example, $4 = 1+1+1+1 = 2+1+1 = 3+1 = 2+2$ so $p(4) = 5$), then

$$p(n) \sim \frac{1}{4n\sqrt{3}} \exp\left(\pi\sqrt{\frac{2n}{3}}\right)$$

as $n \to \infty$.
The textbook [4] is a very accessible introduction to the field, and [5] is a fun but remarkably rigorous young-adult novel (!) which proves the above asymptoptic result, among other things.

At the turn of the 20th century some number theorists introduced a norm that better respected the structures that number theorists were interested in. From here on out, fix a prime number p.

Definition 1.1 The *p-adic valuation* $v_p(n)$ of an integer $n \in \mathbb{Z}$ is the number of factors of p in n:

$$v_p(n) = \max\{k : p^k \mid n\}.$$

For example, $v_3(18) = 2$ since $18 = 2 \cdot 3^2$ has two factors of 3, $v_2(18) = 1$ since it only has one factor of 2, and $v_5(18) = 0$.

Definition 1.2 The *p-adic norm* $\|n\|_p$ for an integer $n \in \mathbb{Z}$ is defined to be

$$\|n\|_p = \left(\frac{1}{p}\right)^{v_p(n)}$$

The base of the exponent is fixed and < 1, so the norm is smaller when the p-adic valuation is larger. For example, $v_3(18) = (1/3)^2 = 1/9$ and $v_5(18) = 1$.

It's not too hard to prove that this defines a norm on $\mathbb{Z}$. Furthermore, the additivity of the valuation implies that this norm is mulplicative on $\mathbb{Z}$, just like the traditional absolute value. Because of this, we can uniquely extend $\|\cdot\|_p$ to a norm

$$\begin{aligned} \|\cdot\|_p &: \mathbb{Q} \to \mathbb{R} \\ \left\|\frac{a}{b}\right\|_p &= \frac{\|a\|_p}{\|b\|_p} \end{aligned}$$

Alright, we've defined another norm on $\mathbb{Q}$ besides the standard absolute value. You may ask, what does this norm look like? It's notoriously difficult to visualize because it behaves nothing like any geometric norm we're familiar with, mostly because it satisfies what is called the *strong triangle inequality*

$$\|a+b\|_p \leq \max\{\|a\|_p, \|b\|_p\}.$$

This has the consequence that every triangle is isoceles, any point of a sphere can be taken as its center, and other such nonsense. For now it is best to think of it as a formal construction.

1.2 The field of p-adic numbers

With another norm in hand, we can replace the absolute value with $\|\cdot\|_p$ in our the construction of $\mathbb{R}$:

(a) We start with our normed ring $(\mathbb{Z}, \|\cdot\|_p)$.

(b) This extends uniquely to a norm $(\mathbb{Q}, \|\cdot\|_p)$ on $\mathbb{Q}$.

(c) This norm defines a metric and we complete $\mathbb{Q}$ with respect to this metric.

We call the result the field of *p-adic numbers*.

Definition 1.3 Let p be a prime. Then the *field of p-adic numbers* $\mathbb{Q}_p$ is the completion of $\mathbb{Q}$ with respect to the p-adic norm $\|\cdot\|_p$ (or, more precisely, its induced metric).

This is already a remarkable construction! It blends an analytic construction, completion, with a number-theoretic norm. The result is a field over which we can do analysis but which taps into number-theoretic properties of $\mathbb{Z}$ and $\mathbb{Q}$. And just as $\mathbb{R}$ and $\mathbb{C}$ open the door to real manifolds and complex geometry, the field $\mathbb{Q}_p$ is the starting point of a new geometric theory.

> One can also put geometry aside and dive into analysis over $\mathbb{Q}_p$. We can, for example, speak of *p-adic L-functions*, which are p-adic analogues of the Riemann zeta function and more generally L-functions. They are fascinating objects of study which weave analytic properties of power series with deep properties of $\mathbb{Q}$. For an introduction to analysis over $\mathbb{Q}_p$, take a look at [2].

One construction we haven't yet touched on is the analogue of the movement from $\mathbb{R}$ to $\mathbb{C}$; since we'll need this later we introduce it now. As we all know, one of the issues with $\mathbb{R}$ is that not every real polynomial has a real root; they are only guaranteed to have roots in the field extension $\mathbb{C}$. Being the smallest field containing $\mathbb{R}$ in which we can find a root for any polynomial over $\mathbb{R}$, we call $\mathbb{C}$ the *algebraic closure* of $\mathbb{R}$. It is a finite extension of $\mathbb{R}$ since $\mathbb{C}$ as an $\mathbb{R}$-vector space has basis $\{1, i\}$.

The algebraic closure $\overline{\mathbb{Q}_p}$ is similarly defined to be the smallest extension of $\mathbb{Q}_p$ containing a root for every polynomial over $\mathbb{Q}_p$, and will be our analogue of $\mathbb{C}$. (If you thought $\mathbb{Q}_p$ was hard enough to imagine, $\overline{\mathbb{Q}_p}$ is even worse: unlike $\mathbb{C}$, it is not a finite extension of $\mathbb{Q}_p$! There are simply too many elements that we have to add to $\mathbb{Q}_p$.)

> This is not entirely accurate: to make matters worse, while $\overline{\mathbb{Q}_p}$ inherits a metric from $\mathbb{Q}_p$, it's no longer complete: completeness is only guaranteed for finite extensions. To remedy this we complete the already huge field $\overline{\mathbb{Q}_p}$ with respect to this norm to obtain a complete field denoted $\mathbb{C}_p$. A result of Krasner is that $\mathbb{C}_p$, by some miracle, remains algebraically closed. It is called the field of p-adic complex numbers and is the field over which much of contemporary p-adic geometry is done.

1.3 Peripheral Remarks

Before we go into p-adic geometry I want to justify the construction. From a skeptical point of view, we can ask "but why?". It's a pretty construction, but why this norm instead of another number-theoretic one? Does it tap into something fundamental about the integers, or is it an arbitrary man-made construction? I'll try to respond to these philosophical questions with some theorems which I think illustrate that the p-adics are part of the fabric of the metaphysics of $\mathbb{Q}$.

The first of these results is *Ostrowski's Theorem*:

Theorem 1.4 (Ostrowski's Theorem) *Every nontrivial absolute value is equivalent to either the usual absolute value or a p-adic absolute value for some p.*

Here two absolute values are said to be equivalent if they induce the same topology. This result says that the reals $\mathbb{R}$ and the p-adics $\mathbb{Q}_p$ constitute all of the completions of $\mathbb{Q}$. They're the only candidates for a field extending $\mathbb{Q}$ over which we can do analysis.

Secondly we give a simple example where information over $\mathbb{Q}_p$ trickles down into knowledge over $\mathbb{Q}$. We call $\mathbb{Q}$ a *global* field, since it containers all of the prime numbers, and we call the $\mathbb{Q}_p$ *local* fields since they're 'concentrated' at a prime. Recall that a quadratic form is a homogeneous polynomial of degree 2, e.g.

$$q(x, y) = 3x^2 - 5xy + 9y^2.$$

Via the embedding $\mathbb{Q} \hookrightarrow \mathbb{R}$, a solution $(x, y) \in \mathbb{Q}^2$ to a quadratic form with integral coefficients is automatically a solution $(x, y) \in \mathbb{R}^2$. Similarly the embeddings $\mathbb{Q} \hookrightarrow \mathbb{Q}_p$ provide a p-adic solution to a quadratic form for every p once we have a solution over $\mathbb{Q}$. In other words, a global solution induces a solution over every local field.

The following result is an example of a so-called *local-global principle*, which says that local information can provide global information:

Theorem 1.5 (Hasse's Principle) *A quadratic form $q(x, y)$ over $\mathbb{Z}$ has a solution over $\mathbb{Q}$ if and only if it has a solution in $\mathbb{R}$ and in $\mathbb{Q}_p$ for every prime p.*

Such a theorem is extremely useful for the study of quadratic forms, since in local fields we have access to analytic tools which make it significantly easier to find zeroes of polynomials! Newton's method, for instance, is a tool for finding zeroes of polynomials over $\mathbb{R}$ which isn't available over $\mathbb{Q}$. The field of p-adics has an analogue, called *Hensel's Lemma*, which similarly allows for successive approximations of zeroes of polynomials defined over $\mathbb{Q}_p$.

Similar local-global principles don't hold in general for forms of higher degrees, but there are local-global principles for other algebraic objects such as algebraic groups.

2 p-adic Geometry

2.1 Introduction

Geometry over $\mathbb{Q}_p$ has to work around a frustrating truth: unlike the natural topology of $\mathbb{R}$, the topology of $\mathbb{Q}_p$ is *extremely bad*, owing to the strong triangle inequality:

- Every open disk $B^-(x; r) = \{y \in \mathbb{Q}_p : \|x - y\|_p < r\}$ is closed.
- Every closed disk $B(x; r) = \{y \in \mathbb{Q}_p : \|x - y\|_p \leq r\}$ is open.
- These imply that the topology on $\mathbb{Q}_p$ is totally disconnected: every point is its own connected component.

The last fact ruins the prospect of naively carrying over techniques from real and complex geometry to p-adic geometry. For instance there are no (nonconstant) lines in $\mathbb{Q}_p$ since there are no nonconstant continuous maps $[0, 1] \to \mathbb{Q}_p$ where the former is given the real topology. Also, local functions don't glue to global functions: one can cover any open subset $U \subseteq \mathbb{Q}_p$ by *disjoint* open disks $B^-(x_i; r_i)$. Assign to each of them the constant function $f_i : B^-(x_i; r_i) \to \mathbb{Q}_p, y \mapsto c_i$ for some constant c_i. Then these functions tautologically agree on their intersection, but there is no global function U which restricts to f_i on each $B^-(x_i; r_i)$.

Geometry over $\mathbb{Q}_p$, therefore, is going to require a new approach to geometry which wasn't required of $\mathbb{R}$ or $\mathbb{C}$.

2.2 Sheaves and Geometry

Instead of asking 'what sorts of shapes and manifolds can we build over this new field?', the modern algebro-geometric approach is instead to ask, 'what kinds of *functions* do we want over our spaces?'

A geometric object in the context of a geometric theory is defined more by the functions on it than the object itself. Real and complex geometry get their identifies not primarily through their base spaces but by the kinds of functions you're interested in:

- $\mathbb{R}^n$ + differentiable functions = real analysis
- $\mathbb{R}^n$ + continuous functions = real topology
- $\mathbb{C}^n$ + holomorphic functions = complex analysis
- $\mathbb{C}^n$ + regular maps = complex algebraic geometry

and so on. To full specify a geometric object requires both the topological space and the functions on it.

The modern mathematical formalism is to specify the context of a topological space X by pairing it with a *structure sheaf* $\mathcal{O}_X$. The structure sheaf pins a topological space to a geometric theory. Intuitively, it is an assignment

$$\begin{aligned}\mathcal{O}_X : \{\text{Open subsets of } X\} &\to \text{Sets} \\ U &\mapsto \{\text{Functions } f : U \to k\}\end{aligned}$$

where k is the base field of X, e.g. $\mathbb{R}, \mathbb{C}, \mathbb{Q}_p$. For example, for a real manifold $\mathcal{O}_X(U)$ would be the set of holomorphic functions $U \to \mathbb{R}$; for a topological space the set of continuous functions $U \to \mathbb{R}$; and as an algebraic variety the set of rational functions $U \to \mathbb{R}$.

All of these examples have the property, easy to check, that if $f_i \in \mathcal{O}_X(U_i)$ and $f_j \in \mathcal{O}_X(U_j)$ agree on their intersection $U_i \cap U_j$ then there is a function $f_{ij} \in \mathcal{O}_X(U_i \cup U_j)$ which restricts to f_i and f_j, and that if two functions $f, g \in \mathcal{O}_X(U)$ agree on a covering $U = \cup U_i$ then $f = g$. These are the *sheaf conditions*.

In other words, a geometric space is really characterized by a topological space X alongside a structure sheaf $\mathcal{O}_X$. In p-adic geometry, we will see that the functions we're interested in are *convergent power series* $U \to \mathbb{Q}_p$. From this perspective the quest for a good p-adic geometry consists of finding a good class of topological spaces X and defining for them structure sheaves $\mathcal{O}_X$ of convergent power series which satisfy the sheaf conditions.

But like we said before, even constant functions in general don't glue to global functions because of the weird topology of $\mathbb{Q}_p$, and this is tantamout to saying that the sheaf condition is impossible to satisfy as it is. This problem and finding the right spaces will constitute the rest of this summary.

2.3 The Tate Algebra

For notational simplicity, let $|\cdot|$ denote the p-adic norm on $\mathbb{Q}_p$. Let

$$\mathbb{B}^n(\overline{\mathbb{Q}_p}) = \{(x_1, \dots, x_n) \in \overline{\mathbb{Q}_p}^n \,:\, |x_i| < 1\}$$

denote the unit ball in $\overline{\mathbb{Q}_p}^n$.

Lemma 2.1 *A formal power series*

$$f = \sum_{\nu \in \mathbb{N}^n} c_\nu X^\nu = \sum_{\nu \in \mathbb{N}^n} c_{\nu_1 \dots \nu_n} X_1^{\nu_1} \cdots X_n^{\nu_n} \in \mathbb{Q}_p[\![X_1, \dots, X_n]\!]$$

converges on $\mathbb{B}^n(\overline{\mathbb{Q}_p})$ *if and only if* $\lim_{|\nu| \to \infty} |c_\nu| = 0$.

This simple lemma, which connects convergence to the divisibility of the coefficients, hints that convergent power series are a good setting to unify the analytic and number-theoretic threads of $\mathbb{Q}_p$.

Definition 2.2 The $\mathbb{Q}_p$-algebra $T_n = \mathbb{Q}_p\langle X_1, \ldots, X_n\rangle$ of all formal power series

$$f = \sum_{\nu \in \mathbb{N}^n} c_\nu X^\nu \in \mathbb{Q}_p[\![X_1, \ldots, X_n]\!], \qquad \lim_{|\nu| \to \infty} |c_\nu| = 0$$

is called the *Tate Algebra* of *restricted*, or *strictly convergent* power series.

(A $\mathbb{Q}_p$-algebra is a fancy term for a vector space where elements can be multiplied together as well as added).

Functions in the Tate algebra have properties familiar from complex analysis. For example, there is a natural norm on T_n, and with respect to this norm functions in the Tate algebra satisfy the maximum modulus principle. There is also a Weierstrass preparation theorem for Tate algebras.

For the more algebraic readers, it's interesting to note another example of the link between the analytic and algebraic character of the Tate algebra: the Weierstrass preparation theorem is the necessary tool to prove some important algebraic properties of T_n, such as the existence of Noether Normalization, being Noetherian, Jacobson, and having Krull dimension n. One notices that these are precisely the sorts of properties of a ring which give its spectrum or maximal spectrum nice *geometric* properties.

2.4 Affinoid Algebras and Affinoid Spaces

Let $M(T_n) = \mathrm{SpecMax}(T_n)$ denote the set of maximal ideals of T_n. For those unfamiliar with maximal ideals, it's enough to know that it's a basic construction in algebraic geometry which takes an algebraic object and spits out a more geometric set. This is quite concrete in the example of the Tate algebra, since one can prove a bijection (of sets)

$$M(T_n) \cong \mathbb{B}^n(\overline{\mathbb{Q}_p})/\sim$$

where $\sim$ is an equivalence relation (conjugacy by the natural Galois action, for those who have the background).

This justifies squinting a bit and replacing $\mathbb{B}^n(\overline{\mathbb{Q}_p})$ with $M(T_n)$. Then, by our previous lemma, we can think of T_n as precisely of the power series which convergence on all of $M(T_n)$. We can thus see a bit of our goal in sight: $M(T_n)$ should be equipped with a topology so that the structure sheaf $\mathcal{O}_{M(T_n)}$ satisfies

$$\mathcal{O}_{M(T_n)}(M(T_n)) = T_n$$

Those familiar with algebraic geometry and the usual spectrum construction on a ring will recognize this property.

Knowing that the maximal spectrum construction links the Tate algebra to a geometric set we pull a move typical of modern algebraic geometry: we reverse perspectives and consider the *algebra* the starting point of our construction and call 'geometric' the sets which arise out of the maximal spectrum construction!

Definition 2.3

(a) A $\mathbb{Q}_p$-algebra A is called an *affinoid $\mathbb{Q}_p$-algebra* if there exists an isomorphism $A \cong T_n/\mathfrak{a}$ for some ideal $\mathfrak{a}$.

(b) The maximal spectrum $M(A) := \mathrm{SpecMax}(A)$ of an affinoid algebra is called an *affinoid space*.

These all have a natural topology, and on an affinoid space A we can then construct a presheaf $\mathcal{O}_A$ on A which maps an open subset of $U \subseteq A$ to convergent power series on $U \subseteq A$, so-called because they don't satisfy the sheaf axioms. Which brings us to the problem which has been plaguing us this whole time.

2.5 The Tate Topology

These affinoid spaces are candidate geometric spaces on which to put a topology. Alas these constructions don't solve the fundamental issue of finding an appropriate topology. We can endow affinoid spaces with a topology inherited in the end from $\mathbb{Q}_p$, but the same issues crop up and affinoid spaces with the natural topology don't have a good enough topology on which the sheaf conditions can be satisfied.

To get around this we pull a trick which to anyone seeing it for the first time seems like a magic (or a scam). Alexander Grothendieck noticed that the properties of topological spaces such as coverings and intersections involved in the formalism of sheaves could be abstracted into axioms. He then defined what is now called a *Grothendieck topology* to be any set (rather, a category) that satisfied these axioms.

In the context of p-adic geometry, John Tate realized that while the canonical topology on affinoid algebras was too fine for a good geometric theory (there were too many open sets), if we only consider special open subsets called *affinoid subdomains* then even though these don't constitute a topology they still constitute a Grothendieck topology, now called the *Tate topology*.

Tate proved the crucial

Theorem 2.4 (Tate's Acyclicity Theorem) *Let A be an affinoid space. Then the presheaf $\mathcal{O}_A$ of affinoid functions is a sheaf on the Tate topology.*

Huzzah! This means that every affinoid space equipped with the Tate topology can be equipped with a structure sheaf $\mathcal{O}_A$. We can now justifiably call affinoid spaces geometric spaces! We are now ready to define the objects of study in p-adic geometry.

2.6 Rigid Spaces

Affinoid spaces are the analogues of open disks in real and complex geometry. One builds real and complex manifolds from scratch as topological spaces which admit a covering by real or complex disks. Similarly, a *rigid space* is a 'topological' space which admits a covering by affinoid spaces:

Definition 2.5 A rigid analytic space is a space X equipped with a Grothendieck topology and a structure sheaf $\mathcal{O}_X$ such that X admits a covering by subsets X_i where $(X_i, \mathcal{O}_X|_{X_i})$ is isomorphic to an affinoid space.

Starting from the number-theoretic analogue to the field of complex numbers, this geometric theory is the counterpart to the theory of complex manifolds.

It gave algebraic geometers a landscape in which they could use analytic techniques to study the p-adic analogues of traditionally complex constructions such as elliptic curves. It would take too much extra machinery to describe the payoff of our constructions in any detail, so we finish by outlining one parallel between complex analysis and rigid geometry.

Example 2.6 In complex analysis there is a so-called GAGA correspondence between complex algebraic varieties and complex manifolds. Let X be a projective complex algebraic variety, e.g. the zero locus of a homogeneous polynomial $p : \mathbb{C}^n \to \mathbb{C}$. There is an 'analytification functor' $X \mapsto X^{\mathrm{an}}$ which interprets the same variety X as a complex analytic manifold. A priori there is no reason to believe that the theories of X that these separate perspectives produce should be related, but Serre proved that there's an equivalence of categories between so-called coherent sheaves on X and coherent sheaves on X^{an}. In plainer language, the sheaf theory of projective complex algebraic varieties is the same as the sheaf theory of projective complex manifolds, and the equivalence allows an exchange of techniques between complex algebraic geometry and complex analytic geometry, yielding some wonderful results.

One of the results of rigid geometry is that there is an analogous equivalence between projective algebraic varieties over $\mathbb{Q}_p$ and projective rigid analytic varieties, in fact proven along the same ideas. Just as in complex analysis, this opens up a whole range of analytic tools to be newly incorporated into the study of algebraic varieties over $\mathbb{Q}_p$.

References

[1] S. Bosch, "Lectures on Formal and Rigid Geometry". Lecture Notes in Mathematics. Springer International Publishing, 2014.

[2] N. Koblitz, "p-adic Numbers, p-adic Analysis, and Zeta-Functions". Graduate Texts in Mathematics. Springer New York, 2012.

[3] J. Neukirch and N. Schappacher, "Algebraic Number Theory". Grundlehren der Mathematischen Wissenschaften. Springer Berlin Heidelberg, 2013.

[4] D. J. Newman, "Analytic Number Theory". Graduate Texts in Mathematics. Springer, 1998.

[5] H. Yūki and T. Gonzalez, "Math Girls". Bento Books, 2011.

An introduction to singular control problems through an electricity market model

Almendra Awerkin Vargas (*)

Abstract. In the electricity market it has been observed that the production of renewable energy decreases the electricity price. Imagine that we want to sell renewable energy, then it raises the question: at which electricity price it is optimal to increase the current renewable installed power in order to obtain the maximum profit of selling the produced energy net the installation costs? We will see that the mathematics used to model this problem is called a singular control and, always leaning in our electricity market example, we will briefly introduce the main concepts and results that define this branch of control theory. We conclude answering our question considering the Italian market case.

1 Entering in the market

Let us imagine that we want to sell renewable energy, so we have to consider some basic aspects of the electricity market, the power grid and our investment, in order to conduct a "good business". Let us point them out:

1. In the Italian market everybody can sell energy, in the sense that all of us can install solar panels on our roofs and sell the produced energy, introducing it in the power grid.
2. There is a maximum amount of power that we are able to install, which is a physical constraint of a power system, related with the maintenance of the stability and the good quality of the service.
3. The investment in installation of the energy generator device (solar panels, wind turbines) is *irreversible*, in the sense that the invested money cannot be returned if we decide to take off the device.

Also, let us assume

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: awerkin@math.unipd.it . Seminar held on 19 May 2021.

4. that in the electricity market the increments in renewable energy production reduce the price of the electricity.

Therefore, considering the above aspects, we ask ourselves: at which price it is optimal to increase the installed power level in order to obtain the maximum profits of selling that energy net the installation cost of the generator device?

2 The model

The electricity market evolves randomly therefore we need to set up our problem in a probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, where a standard Brownian motion $(W_t)_{t\geq 0}$ is defined.

We identify the current level of installed power I_t at time t as as our control variable. Due to the *irreversibility* of our investment, I_t can not decrease and therefore for every $s < t$, we will have $I_s \leq I_t$. Also we have to respect the constraint of maximum allowed installed power, which we will call θ, therefore for every $t \geq 0$ we should have $I_t \leq \theta$. We define the set of *admissible controls* as follows

$$\mathcal{I} = \{I : \Omega \times [0,\infty) \to [0,\infty) : t \to I_t \text{ is } \mathcal{F}_t\text{-measurable, cadlag, non decreasing, } I_0 = y \leq I_t \leq \theta\}.$$

It has been observed that the electricity price follows a mean reverting behavior and assuming that it is affected by renewable energy production we suppose that the dynamic of the electricity price evolves according to the following SDE

$$\begin{cases} dX_t^x = \kappa\left(\zeta - X_t^x - \beta I_t\right) dt + \sigma dW_t \ t \in (0,\infty), \\ X_0^x = x \end{cases} . \tag{1}$$

with $\kappa, \beta, \sigma > 0$ and $\zeta, x \in \mathbb{R}$. The parameter ζ represents the mean value of the process without increments in the renewable installed power, σ the degree of volatility, κ the rate at which the variable reverts towards the mean and β the proportion of power influencing the price.

Now, we write our utility functional describing the total expected profits from selling electricity in the market, net of the total expected costs of installation

$$\mathcal{J}(x,y,I) = \mathbb{E}\left[\int_0^\infty e^{-\rho t} a X_t^x I_t dt - c\int_0^\infty e^{-\rho t} dI_t\right], \tag{2}$$

where $e^{-\rho t}$ is the discounting rate which describes the depreciation of the money, c is the cost of installing one power unit of technology, a is a convector factor of power into energy and dI_t represent the cumulative installed power made up to time t.

Our objective is to follow an optimal installation strategy in order to maximize our utilities, that is to say, to find $I^* \in \mathcal{I}$ such that

$$V(x,y) = \mathcal{J}(x,y,I^*) = \sup_{I\in\mathcal{I}} \mathcal{J}(x,y,I). \tag{3}$$

where V is known as the *value function*.

3 Singular control

Now, we explain how to solve our control problem and why it is call *Singular Control Problem*. Imagine that we can write our cumulative installation as an absolutely continuous control $dI_t = u_t dt$ and forget for a while the constraint θ. Define the respective set of admissible controls,

$$\mathcal{U} = \{u : \Omega \times [0,\infty) \to [0,\infty), u \text{ progressible measurable } \}$$

and we solve our optimal control problem using the the *dynamic programming principle*, which states that for every $u \in \mathcal{U}$

$$V(x,y) \geq \mathbb{E}\left[\int_0^{\Delta t} e^{-\rho t}\left(X_t^x u_t - c u_t\right) dt + e^{-\rho \Delta t} V(X_{\Delta t}^x, u_{\Delta t})\right]. \tag{4}$$

Using *Ito formula*, which generalized the chain rule, allow us to write $V(X_t^x, u_t)$ in term of its derivatives. Substituting it in the above expression we end up with the so call *Hamilton-Jacobi-Bellman* equation which is a partial differential equation whose solution can be identified with the value function $V(x,y)$. For or modified model $dI_t = u_t dt$, we have

$$\mathcal{L}w(x,y) - \rho w(x,y) + xy + \sup_{a \geq 0}\{u(w_y(x,y) - c)\} = 0 \tag{5}$$

where $\mathcal{L}$ is the differential operator $\mathcal{L}u(x,y) = \kappa\left(\left(\zeta - \beta y\right) - x\right) u_x(x,y) + \frac{\sigma^2}{2} u_{xx}(x,y)$. For the supremum in Equation (5) we get

$$\sup_{u \geq 0}\{u(w_y(x,y) - c)\} = \begin{cases} \infty \text{ if } w_y(x,y) - c > 0 \\ 0 \text{ if } w_y(x,y) - c = 0 \\ 0 \text{ if } w_y(x,y) - c < 0 \end{cases}. \tag{6}$$

We obtain as byproduct

1. $w_y(x,y) - c \leq 0$
2. If $w_y(x,y) - c < 0$ then $\mathcal{L}w(x,y) - \rho w(x,y) + axy = 0$.

Nevertheless, we do not have a clear answer of which strategy we have to follow in the case $w_y(x,y) - c = 0$. Actually, we are this situation because of the linear dependence on the control of the utility functional. Indeed the way to solve this problem is to consider our cumulative installation dI_t as the variation of a monotone process, which we ask to be cadlag in order to have some regularity. In this setting such a control becomes singular with respect to the Lebesgue measure and is called a *singular control problem.*

3.1 Characterization of the optimal strategy

Let us come back to our original problem (3). At the initial time $t = 0$, we can consider the following strategies:

1. Let evolve the system during a time interval Δt without increase the installed power and then our utility functional continues optimal. This is actually the *Dynamic Programming* approach, and we will end up with the following inequation

$$0 \geq \mathcal{L}w(x,y) - \rho w(x,y) + axy \ , \ (x,y) \in \mathbb{R} \times [0,\theta),$$

   with boundary condition $w(x,\theta) = R(x,\theta)$, where $R(x,\theta)$ is the utility obtained just considering the initial installation, i.e. without increments in the power level $dI \equiv 0$.

2. Increment immediately the installation and then continue optimal, which lead us to

$$0 \geq w_y(x,y) - c.$$

Since only one of these two options can be true at time zero, we have

$$\max\{\mathcal{L}w(x,y) - \rho w(x,y) + axy, w_y(x,y) - c\} = 0, \tag{7}$$

with boundary condition $w(x,\theta) = R(x,\theta)$. This expression is called *variational inequality*.

According to our reasoning the state space should be divided in two regions: a non installation region $\mathbb{W}$ associated with the first strategy and an installation region $\mathbb{I}$, associated with the second one.

$$\mathbb{W} = \{(x,y) \in \mathbb{R} \times [0,\theta) : axy - \rho w(x,y) + \mathcal{L}w(x,y) = 0, w_y(x,y) - c < 0\}, \tag{8}$$

$$\mathbb{I} = \{(x,y) \in \mathbb{R} \times [0,\theta) : axy - \rho w(x,y) + \mathcal{L}w(x,y) \leq 0, w_y(x,y) - c = 0\}. \tag{9}$$

It can be proved that this two region are separated by an injective, non decreasing function $F : [0,\theta] \to \mathbb{R}$ called *the free boundary*, such that

$$\mathbb{W} = \{(x,y) \in \mathbb{R} \times [0,\theta) : x < F(y)\}, \tag{10}$$

$$\mathbb{I} = \{(x,y) \in \mathbb{R} \times [0,\theta) : x \geq F(y)\}. \tag{11}$$

If we reach to write an expression for the free boundary, we will be able to describe the optimal installation strategy and solve our *singular control problem*. In the following figure we can see graphically the free boundary, the waiting and the installation region:

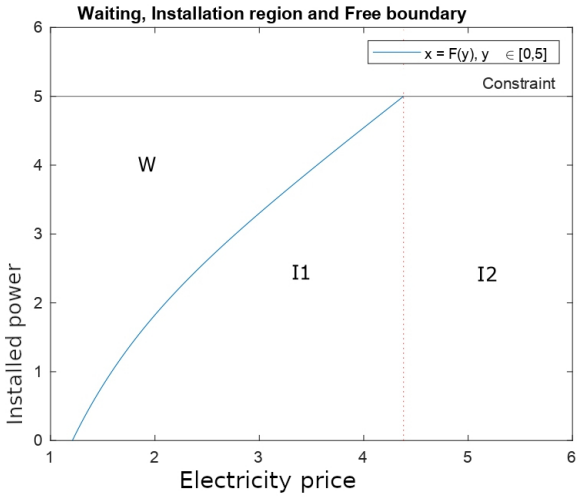


Figure 1 Free Boundary.

The red dotted line corresponds with the value of the free boundary at the constraint $\hat{x} = F(\theta)$. The optimal strategy can be describe as follows: when the electricity price X_t^x is lower than $F(I_t)$, i.e., when we are in the waiting region $\mathbb{W}$ (see (10)), no installation should be done and it is necessary to wait until the price X_t^x crosses $F(I_t)$ to optimally increase the installed power level. When the electricity price X_t^x is between $F(0)$ and $F(\theta)$ (region $\mathbb{I}_1$ in Figure 1), enough power should be installed to move the pair price-installation in the up-direction until reaching the free boundary F. In the extreme case when $X_t^x \geq F(\theta)$ (region $\mathbb{I}_2$ in Figure 1) the energy producer should install instantaneously the maximum allowed power θ.

To obtain an expression for the *free boundary* we start observing that for $(x, y) \in \bar{\mathbb{W}} \cap \mathbb{I}$, i.e., when $x = F(y)$, our candidate value function should satisfy

$$axy - \rho w(x, y) + \mathcal{L}w(x, y) = 0 \text{ and } w_y(x, y) - c = 0.$$

Doing the computations described in [2], at the end of the day, when there is significant impact price $\beta > 0$, we will end up with an ODE for the free boundary which can be solve numerically [2]. On the other hand, when there is no price impact $\beta = 0$, our free boundary is constant and is obtained solving an algebraic equation [1].

4 N energy producers

As we can imagine, one single producer should not be big enough in order to impact the electricity price, but if we consider the aggregate production of N producer, with N big enough we could expect to observe an influence of the renewable production on the electricity price. Suppose every producer aim to maximize the same utility function that we saw previously in (2), but in this case the price will be affected by the production of the N producers, that is

$$\begin{cases} dX_t^x = \kappa(\zeta - X_t^x - \beta \sum_{i=1}^N I_t^i)dt + \sigma dW_t \ t > 0 \ , \\ X_0^x = x. \end{cases} \tag{12}$$

We can imagine the perspective of a coordinator, which will seek for maximizing the sum of the utilities J_i, $i = 1, \dots, N$ of the N producer, so we look for a strategy $\bar{I} = (I_1, \dots, I_N)$, with $\bar{I}$ on the admissible set

$$\bar{\mathcal{I}} = \{\bar{I} : [0, \infty) \times \Omega \to [0, \infty)^N \text{ , non decreasing, left continuous, adapted process}$$
$$\text{with } I_0^i = y^i, \ \mathbb{P}\text{-a.s.}, \ \textstyle\sum_{i=1}^N I_t^i \leq \theta\},$$

such that

$$V_{SP} = \sup_{\bar{I} \in \mathcal{I}} \sum_{i=1}^N \mathcal{J}_i(I_i) = \sup_{\bar{I} \in \mathcal{I}} \sum_{i=1}^N \mathbb{E}\left[\int_0^\infty e^{-\rho t} a X_t^x I_t^i dt - c \int_0^\infty e^{-\rho t} dI_t^i\right], \tag{13}$$

This problem is called the *social planner problem* and it is much more difficult to track with the approach that we saw previously, but in our particular case due to the linear dependence in the control variable, we can write $\sum_{i=1}^N I_t^i := \nu_t \in \mathcal{I}$ and our problem is equivalent with the single producer case. Actually, the optimal control for the single producer case I^* is a *Pareto optimum* for the *social planner problem*, which means that if we identify by $\bar{I}^*$ the vector whose components satisfy $\sum_{i=1}^N I_t^i := I_t^*$, there is not another vector strategy $\hat{I} \in \bar{\mathcal{I}}$ where at least one component I_i maximize the utility of the producer i.

5 The Italian case

Finally, we test the previous model in the six main economic Italian zone: North, Central North, Central South, South, Sicily and Sardinia, considering the aggregate renewable power production and electricity price by zone. We use real weekly data from 2012 to 2018 of the electricity price and current installed power of photovoltaic and wind sources, to estimate the parameters of the SDE describing the evolution of the electricity price (1). As first result we are able to check if the mean reverting dynamic is enough to describe the evolution of the electricity price and if the renewable production influences the electricity price. We found that the production of photovoltaic energy impacts the North zone, while wind is significant for Sardinia. On the other hand, the Central North zone does not present electricity price impact. In the other zones the residuals of the model were correlated and therefore there was "something" that our model was not able to describe. Secondly, with the estimated values we are able to graph the free boundary for every zone in which our models fits well.

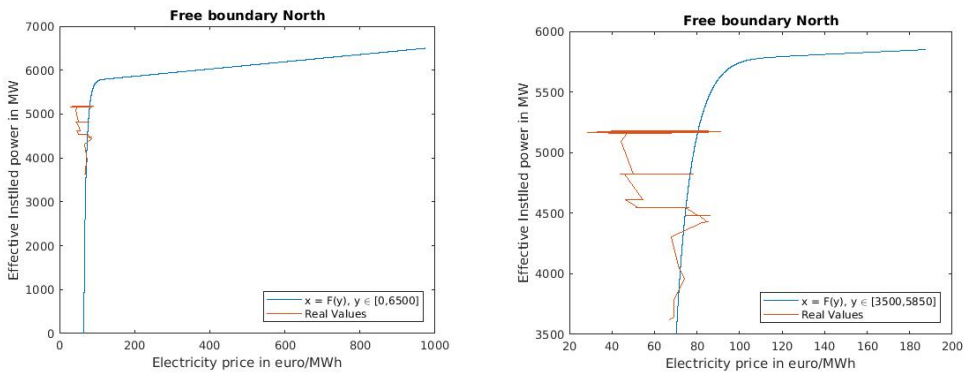


Figure 2 (a) Simulated free boundary and real data for the North. (b) Detail of free boundary and real data for the North.

In Figure 2a, the point at zero installation level corresponds to $F(0) = 64.9$ €/MWh. The red irregular line corresponds to the realized trajectory $t \to (X(t), Y(t))$, i.e. to the values of electricity price vs effective photovoltaic installed power in the North: from it we can see that, at the beginning of the observation period (2012), the installed power was already around 3600 MW. Instead, the blue smooth line corresponds to the computed free boundary $F(y) = x$, which expresses the optimal installation strategy as we already see. In the detailed Figure 2b we can observe the strategy followed in the North zone: the installation level from 3500 MW until 4500 MW was approximately optimal, in the sense that the pair price-installed power was around the free boundary F, with possibly some missed gain opportunities when, between 4300 and 4500 MW, the price was deep into the installation region; nevertheless, the rise in renewable installation from 4500 MW to 4800 MW was at the end done with a power price which resulted lower than what should be the optimal one. At around 4800 MW, there was an optimal no installation procedure until the price entered again the installation region: again, the consequent installation strategy was executed with some delay, resulting in a non-optimal strategy. At the end of the installation (around 5200 MW), we can see that the pair price-installed power moved again deep into the installation region: we should then expect an increment in installation.

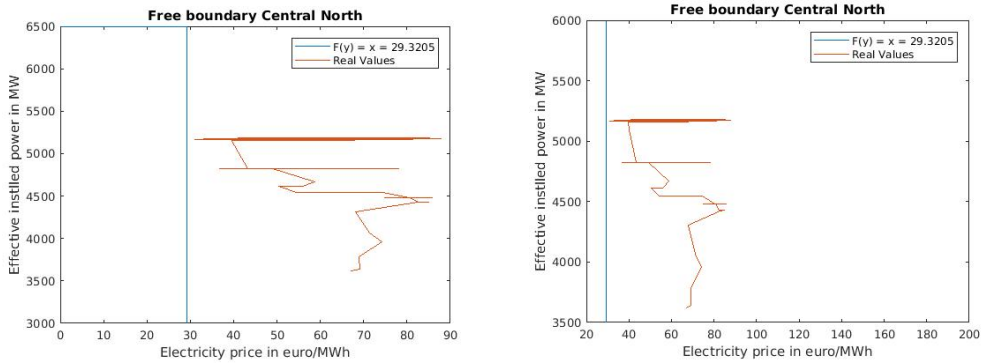


Figure 3 (a) Simulated free boundary and real data for Central North. (b) Detail of free boundary and real data for Central North.

In Figure 3a the vertical blue line corresponds to the constant free boundary $\bar{x} = 29.3205$ €/MWh, while the red irregular line with the realized values of price-installation action that was put in place in the Central North zone. In this case, the optimal strategy is described as follows: for electricity prices less than $\bar{x}$, no increments on the installation level should be done. Conversely, when the electricity price is grater or equal to $\bar{x}$ the producer should increment the installation level up to the maximum level allowed for photovoltaic power (here we posed $\theta = 6500$ MW). As we can clearly see on Figure 3a, the electricity price has always been greater than $\bar{x}$ in the observation period; however, the increments on the installation level was not high enough to arrive to the maximum level $\theta = 6500$ MW, therefore the performed installation was not optimal.

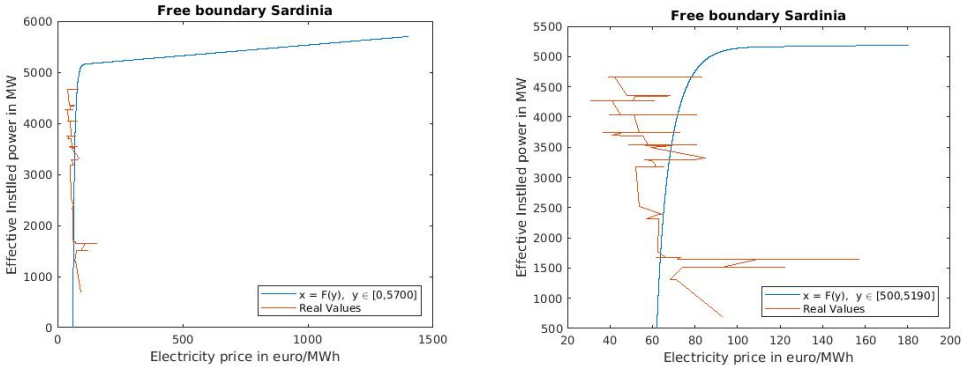


Figure 4 (a) Simulated free boundary and real data for Sardinia. (b) Detail free boundary and real data for Sardinia.

In Figure 4a the point at zero installation level corresponds to $F(0) = 61.5199$ €/MWh. The red irregular line corresponds with the realized values of electricity price vs effective wind installed power in Sardinia, from which we can see that the installed wind power at the beginning of the observation period was already around 600 MW. The blue smooth line corresponds to the simulated free boundary $F(y) = x$, which expresses the optimal installation strategy as was already explained for the North case. In the detailed Figure 4b we can observe the strategy followed in the Sardinia zone: until the level 1600 MW the power price was very deeply into the installation region, but the installation increments were not high enough to be optimal. Optimality came between the levels 1600 MW and 2400 MW, where the performed strategy was to effectively maintain the pair price-installed power around the free boundary F. However, the subsequent increments were not optimal, in the sense that the installed power was often increased in periods where the electricity price was too low, and in other situations the power price entered deeply in the installation region without the installed capacity following that trend, or rather doing it with some delay.

References

[1] Awerkin, A. and Vargiolu, T. (2021), *Optimal installation of renewable electricity sources: the case of Italy*. Available at `https://arxiv.org/abs/2102.04243`.

[2] Koch, T. and Vargiolu, T. (2019), *Optimal installation of solar panels under permanent price impact*. Available at `https://arxiv.org/abs/1911.04223`.

Synchronization and asymptotic dynamics of mechanical systems: an introduction

Sara Galasso (*)

Abstract. Synchronization is a fascinating and eye-catching phenomenon, which spontaneously emerges from the collective behaviour of a huge variety of interacting systems. Examples permeate science, from fireflies to metronomes, from neurons to celestial bodies. In this seminar we shall focus on mechanical systems, having in mind, in particular, systems of coupled pendula. To construct a physical-mathematical model able to describe synchronicity patterns in their evolution, classical tools from dynamical systems theory are essential. We will therefore recall the basic notions of invariant manifold and stability, as well as some results that will allow us to investigate, at an introductory level, the long-time asymptotic behaviour. Along with the theory, we will provide examples and examine simple models which should help us visualize some fundamental mechanisms underlying synchronization.

1 Introduction

These notes are an attempt to introduce the reader to a slice of the world of synchronization. In particular, I will try to describe, at an introductory level, and with the support of a toy model, the emergence of synchronization in the asymptotic dynamics of unforced mechanical systems. The main aim will be to develop an intuitive understanding of simple mechanisms for these phenomena by providing the very basics from the theory of dynamical and mechanical systems.

1.1 Sketch of the model

The main –very simple– model we will rely on is the one sketched in Figure 1 below. It consists of two identical simple pendula which are supported by a common rigid bar that can move in one-dimension. The model takes into account dissipative contributions acting on the support, while it neglects other possible (and physically present but at first approximation less relevant) sources. As we shall see, for generic initial conditions, the asymptotic dynamics is characterised by the oscillation of the two pendula in opposition of phase.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: galasso@math.unipd.it . Seminar held on 9 June 2021.

Even if this model may look oversimplified, it contains the main ingredients needed to understand quite general mechanisms. Moreover, it is actually an easily-realizable physical model which can be used to observe synchronization in coupled pendula and as a very preliminary check for more complicated mathematical predictions.

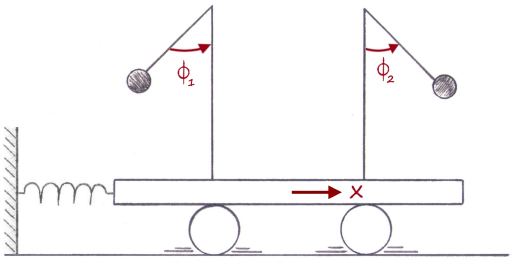


Figure 1 Model of two coupled pendula.

1.2 Scientific context

Synchronization phenomena are present in very wide and variegated contexts, involving physics, biology, neuroscience. Accordingly, each research field investigates these systems through different approaches. Generically, synchronization is referred to the spontaneous emergence of a collective and organized behaviour in the evolution of a composite system. Some examples are the flashing of fireflies, the pacemaker cells of the heart, the tidal locking of planet-moon systems. The interested reader can find an enjoyable introduction in [7].

It is particularly active the research in the field of mechanical systems. In fact, it dates back to the 17^{th}-century the first observation of a synchronization phenomenon by the dutch scientist Christiaan Huygens (1629-1695): he noticed that two pendulum clocks would oscillate in opposition of phase when hanging on a common wooden bar. The toy model we will be working with is a simplified version (see also [9]) and it is often used as starting point for more specific analyses of Huygens' observation, which is still an open problem. For a first linear analysis see [4], while an up-to-date study can be found e.g. in [6]. Here, we will highlight a different aspect, focusing on the role of damping and how it affects the structure of the phase space of the system. In this perspective, we will be considering unforced mechanical systems only.

2 Asymptotic behaviour of dynamical systems

In this section we introduce some of the classical notions and tools used in Dynamical Systems theory to study the long-time asymptotic behaviour of evolutionary processes. Very good references are [3] and [5].

We consider evolutionary processes that are deterministic, finite-dimensional, time-continuous and differentiable, described by an ordinary differential equation of the type

$$\dot{z} = X(z), \quad z \in M \tag{1}$$

where $\dot{z} := \frac{dz(t)}{dt}$, $t \in \mathbb{R}$, with X smooth and autonomous vector field on a n-dimensional differentiable manifold M, the *phase space*. To fix the ideas, in what follows we will assume M to be an open set of $\mathbb{R}^n$, endowed with the Euclidean metric.

We assume X to be Lipschitz. The *flow* of X is a differentiable map $\Phi : \mathbb{R} \times M \to M$, $(t, z) \mapsto \Phi(t, z) =: \Phi_t(z)$ such that $\Phi_t(z)$ is the solution of (1) at time t passing through z. Moreover, Φ is a differentiable action of the group $\mathbb{R}$ on M. For any $z \in M$, the *orbit* of X through z is $\mathcal{O}_z = \{\Phi_t(z) : t \in \mathbb{R}\}$, namely the image of the solution through z. The collection of all the orbits defines a partition of the phase space, called *phase portrait*.

In what follows, we will make use of the notion of *Lyapunov function*, which is a real function decreasing along the solutions of X, and is often usefull in studying stability properties of a dynamical system. More precisely, let $\mathcal{L}_X f$ denote the Lie derivative of a function $f : M \to \mathbb{R}$ along the vector field X, defined as $\mathcal{L}_X f(z) := \lim_{t\to 0} \frac{f(\Phi_t(z))-f(z))}{t}$, then

Definition 1 (Lyapunov function) Let $\Omega \subseteq M$. A differentiable function $\mathcal{W} : M \to \mathbb{R}$ such that $\dot{\mathcal{W}}(z) := \mathcal{L}_X \mathcal{W}(z) \leq 0 \ \forall z \in \Omega$ is called *Lyapunov function* for X on Ω.

In order to study the long-time asymptotic dynamics we need the following definition:

Definition 2 (ω-limit set) Let $z, p \in M$. p is said *ω-limit point of z* if there exists a sequence $\{t_n\}_n$ in $\mathbb{R}$ such that $t_n \to +\infty$ and $\Phi_{t_n}(z) \to p$ as $n \to +\infty$. The set $\omega(z)$ of all the ω-limit points of z is called *ω-limit set of z*.

The ω-limit set contains the information regarding the asymptotic behaviour of a solution as time goes to infinity. Examples are equilibrium points, periodic orbits, limit cycles, strange attractors.

It is, however, typically challenging to explicitly solve (1) and to determine the solutions of the associated Cauchy problem. Moreover, it is often preferable to have a more global understanding of the dynamics on the phase space, for example by knowing the behaviour of open sets of initial conditions, possibly without actually integrating the differential equation. At the end of this section, we will state a theorem, the invariance principle by LaSalle, which allows to locate the ω-limit sets when a Lyapunov function for the system is known.

Before giving the statement, we need some further understanding of the phase space of the system. The first important property a region of the phase space might have is the invariance under the flow of the vector field X.

Definition 3 (Invariant set) A subset $N \subset M$ is *invariant* (resp. *positively invariant*) under the flow of X if $\Phi_t(N) \subseteq N \ \forall t \in \mathbb{R}$ (resp. $t \geq 0$).

Particularly important are invariant sets which have the structure of a submanifold. In which case, it is possible to reduce the study of the dynamics to a lower dimensional space. Example of invariant manifolds are the orbits, stable/unstable/center manifold of fixed points, regular level sets of first integrals.

We will say that a closed invariant set $N \subset M$ is *attracting* if there exists a neighborhood U of N (the *basin of attraction*) such that $\Phi_t(z) \to N \ \forall z \in U$ as $t \to +\infty$.

We have then the following characterisation of the ω-limit sets:

Proposition 1 *If $N \subset M$ is compact and positively invariant, then, $\forall z \in N$, $\omega(z)$ is contained in N, nonempty, closed, connected and invariant under the flow.*

Therefore, ω-limit sets have some nice properties. However, they might have a complicated structure (e.g. the strange attractors) and it is often difficult to determine them. Thus, a good strategy is to look for bigger (attracting) sets in which ω-limit sets are contained. In particular, the invariance property justifies the name of the following important result by LaSalle [5]:

Theorem 1 (LaSalle Invariance Principle) *Let $\Omega \subset M$ be a compact positively invariant set and let $\mathcal{W}$ be a Lyapunov function for X on Ω. Let $\mathcal{I}$ be the largest invariant set in $\{\dot{\mathcal{W}} = 0\}$, then $\omega(z) \subseteq \mathcal{I} \cap \mathcal{W}^{-1}(c)$ $\forall z \in \Omega$ and some $c \in \mathbb{R}$.*

This results gives information about the location of the ω-limit set and about the extension of its region of attraction, whenever a Lyapunov function for the system is known. Moreover, it is often possible to define the "trapping region" Ω as a sublevel set of $\mathcal{W}$.

3 Conservative and dissipative mechanical systems

We want now to focus on mechanical systems (see e.g. [2]). We consider finite-dimensional conservative mechanical systems with holonomic constraints, described by a second-order ODE of the form

$$\ddot{q} = Y(q, \dot{q}), \quad (q, \dot{q}) \in TQ \tag{2}$$

where Q is a n-dimensional differentiable manifold, called *configuration manifold*, and TQ, the phase space, is its tangent bundle, Y is a smooth vector field on TQ.

Note: in the Lagrangian formalism this would mean to be working with a Lagrangian function of the form $L(q, \dot{q}) = T(q, \dot{q}) - V(q)$, where T denotes the kinetic energy and V the potential energy of the system, and equations (2) are the Lagrange equations for the Lagrangian L.

3.1 Dissipation

Real mechanical systems are not conservative, since the total mechanical energy decreases as *dissipation* acts on the system. By this we mean that part of the kinetic energy of

the system is irreversibly transferred into thermal energy, due to the interaction of particles at microscopic scales. The macroscopic manifestation of dissipative mechanisms is called *damping*. At first approximation, dissipative contributions are often neglected when modelling a system. However, damping has a crucial impact on the asymptotic dynamics.

Possible sources of dissipation are the deformation of a material or the friction produced by the relative sliding of two surfaces [1]. Typically, several damping mechanisms take place simultaneously, and each might affect the dynamics with different intensity. Here, we are going to take into account the viscous drag affecting the motion of a body moving within a fluid. A typical example is the air resistance. The most commonly used model for this mechanism is the *viscous damping*, which adds in the equations of motion a term proportional to the velocity of the system:

$$\ddot{q} + \Gamma \dot{q} = Y(q, \dot{q}), \quad (q, \dot{q}) \in TQ \tag{3}$$

where Γ is a positive semi-definite matrix, called *damping matrix*.

3.2 Example: the pendulum

As an example of viscously damped system we consider here the pendulum.

The simple pendulum consists of a point mass constrained on a smooth circle, parametrized by the angle θ. The equation of motion can be written in the form

$$\ddot{\theta} + \sin\theta + \gamma\dot{\theta} = 0, \quad (\theta, \dot{\theta}) \in TQ = S^1 \times \mathbb{R}, \gamma \geq 0,$$

where the term $\gamma\dot{\theta}$ models the viscous damping of the air. In Figure 2, the phase portraits in absence and in presence of damping are compared. We can notice that the inclusion of the damping term modifies strongly the orbits of the system, making, in particular, the origin an attracting point. The convergence toward one equilibrium will be faster as γ increases.

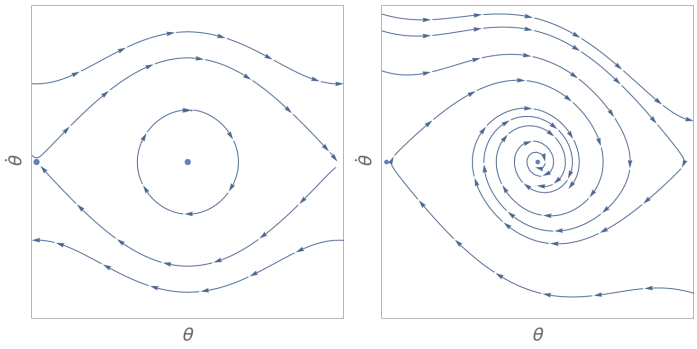


Figure 2 Phase portraits of the undamped and the damped pendulum.

3.3 Partial damping

The inclusion of a damping term makes of the energy $\mathcal{W}$ of the system a good Lyapunov function, in the sense that

$$\dot{\mathcal{W}}(q,\dot{q}) = -\dot{q}\cdot\Gamma\dot{q} \leq 0 \quad \forall (q,\dot{q}) \in TQ$$

and, in particular, $\dot{\mathcal{W}}(q,\dot{q}) = 0$ if $\dot{q} \in \ker(\Gamma)$. As we said, it is often the case that parts of a system are affected differently by the dissipation, and some contributions can be neglected with respect to others, at least at first approximation. We are interested in the situation in which this happens, namely when the damping matrix has a nontrivial kernel: $\ker(\Gamma) \neq \{0\}$. It follows that the set $\{\dot{\mathcal{W}} = 0\}$ is nontrivial, and therefore Theorem 1 implies that the ω-limit set of any bounded solution is contained in that set. If there exists a submanifold $S \subset Q$ such that

(a) TS is compact and invariant,

(b) $T_qS \subseteq \ker(\Gamma) \quad \forall q \in S$,

then, the submanifold TS is attracting and on it the dynamics is conservative. Hence, the asymptotic behaviour of the system is well described by the dynamics on TS.

The formalization of these concepts is an ongoing work. However, these techniques are widely applied to the study of specific problems. The model we are analysing in the following section is an example, as in [8]. We will use it to highlight precisely the possible consequences of such a more general structure of the phase space (see Figure 3).

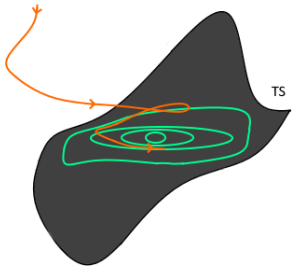


Figure 3 Invariant undamped submanifold and possible asymptotic dynamics.

4 Simple model of synchronized coupled pendula

We are finally ready to investigate the long-time dynamics of the model in Figure 1.

The configuration manifold of the systems is $Q = S^1 \times S^1 \times \mathbb{R}$. Let ϕ_1, ϕ_2 be the angular coordinates of the two pendula and x the linear displacement of the supporting bar. The equations of motion can be written in the following form

$$\begin{cases} \mu\ddot{x} + \gamma\dot{x} + \sum_{i=1}^{2}\left(\ddot{\phi}_i\cos\phi_i - \dot{\phi}_i^2\sin\phi_i\right) + \alpha x = 0 \\ \ddot{\phi}_i + \ddot{x}\cos\phi_i + \sin\phi_i = 0, \quad i = 1,2 \end{cases} \tag{4}$$

where μ, α, γ are positive parameters. Note that the term $\gamma\dot{x}$ models a viscous damping contribution acting on the supporting bar, so that the damping matrix has the form

$$\Gamma = \begin{pmatrix} \gamma & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

A numerical integration of the equations of motion (4) shows that (see Figure 4), for almost every initial condition, the system tends to a configuration in which the supporting bar is at rest and the two pendula oscillate in opposition of phase. And this is in agreement with observations.

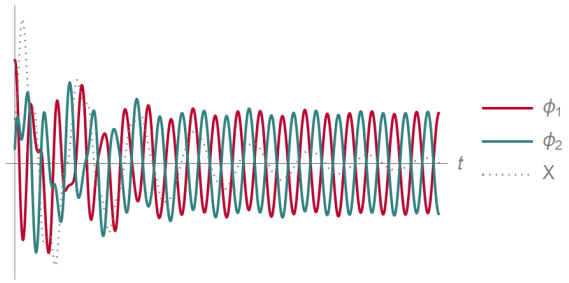


Figure 4 Time evolution of a solution of (4) for $\gamma > 0$.

Hence, the asymptotic behaviour of the system is characterised by the anti-phase synchronization of the two pendula. However, this is really a consequence of the damping acting on the coupling. Following the reasoning of Section 3.3, namely by computing the kernel of the damping matrix Γ and applying LaSalle's principle, it is possible to prove (see e.g. [8]) the following result.

Proposition 2 *The two-dimensional tangent bundle of the submanifold*

$$S = \{(x, \phi_1, \phi_2) \in Q : \phi_1 = -\phi_2, x = 0\}$$

is invariant and attracting.

In conclusion, this simple example shows how the emergence of synchronization in the asymptotic dynamics of unforced mechanical systems is related to the presence of invariant structures in the phase space on which the dissipation does not act. It is therefore of great interest to investigate further the inclusion of suitable damping contributions. Moreover, since, as we have seen, there are typically several sources of dissipation acting on a physical system, the dynamics evolves with multiple time-scales. In this perspective, of particular interest is the study of systems which present families of positively invariant submanifolds characterised by different decay rates.

4.1 On the modelling of coupled pendula

The problem of constructing models of coupled pendula is an active field of research, especially –but not only– in the perspective of explaining the synchronization of Huygens'

clocks and sever other phenomena that emerge in systems of pendulum clocks or similar. The model investigated here is a preliminary step, and leaves space to several generalizations. For example, it is possible to study the case of a generic number of pendula, finding a similar result as the one in Proposition 2. Moreover, the flexibility of the support can be included in the model by increasing the number of degrees of freedom for the coupling.

References

[1] A. Akay and A. Carcaterra, *Damping Mechanisms.* In: CISM International Centre for Mechanical Sciences, Courses and Lectures. Vol. 558. Springer International Publishing (2014), 259–299.

[2] V. I. Arnold, "Mathematical methods of classical mechanics". Vol. 60. Springer Science, 1989.

[3] J. Guckenheimer and P. Holmes, "Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields". Applied Mathematical Sciences. Springer New York, 2013.

[4] D. J. Korteweg, *Les horloges sympathiques de Huygens.* Archives Neerlandaises, Sér. II, tome XI, pp. 273–295. The Hague: Martinus Nijhoff, 1906.

[5] J. P. LaSalle, *The Stability of Dynamical Systems.* CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, 1976.

[6] J. Peña Ramirez, L. A. Olvera, H. Nijmeijer, and J. Alvarez, *The sympathy of two pendulum clocks: Beyond Huygens observations.* In: Scientific Reports 6.1 (2016), 1–16.

[7] A. Pikovsky, M. Rosenblum, and J. Kurths, "Synchronization: A Universal Concept in Nonlinear Sciences". Cambridge Nonlinear Science Series. Cambridge University Press, 2003.

[8] A. Yu Pogromsky, V. N. Belykh, and H. Nijmeijer., *Controlled synchronization of pendula.* In: Proceedings of the IEEE Conference on Decision and Control. Vol. 5 (2003), 4381–4386.

[9] F. Talamucci, *Synchronization of two coupled pendula in absence of escapement.* In: Applied Mathematics and Mechanics (2016).