Università di Padova – Dipartimento di Matematica

Scuole di Dottorato in Matematica Pura e Matematica Computazionale

Seminario Dottorato 2010/11

Preface

This document offers a large overview of the nine months' schedule of Seminario Dottorato 2010/11. Our "Seminario Dottorato" (Graduate Seminar) is a double-aimed activity. At one hand, the speakers (usually Ph.D. students or post-docs, but sometimes also senior researchers) are invited to think how to communicate their own researches to a public of mathematically well-educated but not specialist people, by preserving both understandability and the flavour of a research report. At the same time, people in the audience enjoy a rare opportunity to get an accessible but also precise idea of what's going on in some mathematical research area that they might not know very well.
Let us take this opportunity to warmly thank the speakers once again, in particular for their nice agreement to write down these notes to leave a concrete footstep of their participation. We are also grateful to the collegues who helped us, through their advices and suggestions, in building an interesting and culturally complete program.

Padova, 26 June 2011

Corrado Marastoni, Tiziano Vargiolu

Abstracts (from Seminario Dottorato's web page)

Wednesday 20 October 2010

The order complex of the coset poset of a finite group

Massimiliano PATASSINI (Univ. Padova, Dip. Mat.)

In this seminar we want to speak about a topological aspect of some objects in finite group theory. Let G be a finite group and let $C(G)$ be the coset poset of G, i.e. $C(G) = \{Hg : H < G, g \in G\}$. In order to give a topological interpretation to this object, we introduce the concept of order complex of $C(G)$. The order complex was studied by Kennet Brown, who pointed out a connection between the Dirichlet polynomial of G and the reduced Euler characteristic of the order complex of $C(G)$. In our talk, we first give an overview of the concepts of coset poset, order complex and Möbius function. Next we introduce the work of Kennet Brown concerning the order complex of the coset poset of a soluble group. Last we give an idea of our result about the non-contractibility of the order complex of the coset poset of a classical group.

Wednesday 3 November 2010

The concept of supermodularity in aggregation functions and copulas

Maddalena MANZI (Univ. Padova, Dip. Mat.)

In many domains we are faced with the problem of aggregating a collection of numerical readings to obtain a typical value, not only in mathematics or physics, but also in majority of engineering, economical, social and other sciences. So, aggregation functions are used to obtain a global score for each alternative taking into account the given criteria, even if the problems of aggregation are very broad and heterogeneous. For example, there is a lot of contributions about the aggregation of finite or infinite number of real inputs, topics treating of inputs from ordinal scales, or also the problem of aggregating complex inputs (such as probability distributions, fuzzy sets). In this talk I will discuss the way to construct, in particular, supermodular aggregation functions, which can be analyzed under various aspects: algebraic, analytical, probabilistic. So, in the first part I will introduce the general concept of supermodularity, which comes from lattice theory, and we will see several basic examples. Then, we will be able to apply this theory to aggregation functions and, in particular, to a subclass of aggregation functions, i.e. the family of copulas. In the last part, we will see some results obtained with a particular intersection with fuzzy set theory. So, a basic background will be given also in this direction. The talk will be based on some joint works with M. Cardin, M. Kalina, E. P. Klement and R. Mesiar.

Wednesday 17 November 2010

Semiconcave type results of the minimum time function

NGUYEN Khai Tien (Univ. Padova, Dip. Mat.)

We will give an overview of "semiconcave type" results of the minimum time function in the case of nonlinear control systems under general controllability assumptions. Moreover, in this connection we will show some regularity results of a function whose hypograph satisfies an exterior sphere condition.

Wednesday 1 December 2010

A simple model for Financial Indexes with some application

Alessandro ANDREOLI (Univ. Padova, Dip. Mat.)

Mathematical finance is applied mathematics concerned with financial markets. Two of its major subjects are: 1. Mathematically modeling the prices of assets and indexes; 2. Option Pricing. We first recall the Black & Scholes model for asset prices, then present an easy model that overcomes some weak points of Black & Scholes and other models (in particular the absence of multiscaling effects and of volatility autocorrelation decay). Finally, we give an overview of the option-pricing problem.

Wednesday 15 December 2010

An introduction to Coxeter group theory

Mario MARIETTI (Univ. Padova, Dip. Mat.)

Coxeter groups arise in many parts of algebra, combinatorics and geometry, providing connections between different areas of mathematics. The purpose of this talk is to give an overview to Coxeter group theory from algebraic, combinatorial and geometrical viewpoints. Some classical and more recent results will be presented.

Wednesday 19 January 2011

The maximum matching problem and one of its generalizations

Yuri FAENZA (Univ. Padova, Dip. Mat.)

Given a graph $G(V, E)$, a matching M is a subset of E such that each vertex in V appears as the endpoint of at most one edge from M. The maximum matching problem and its weighted

counterpart are among the most important and studied problems in combinatorial optimization. In this talk, we survey a number of classical results on the topic and present more recent results for a non-trivial generalization of the maximum weighted matching problem. The talk will be accessible to a general audience.

Monday 31 January 2011

Factorization in categories of modules

Marco PERONE (Univ. Padova, Dip. Mat.)

We study the regularity of the behaviour of the direct sum decomposition in categories of modules. It is well known that in some easy categories the direct sum decomposition is essentially unique. In the last 15 years some interesting examples were found of categories where the decomposition is not essentially unique but there still is an outstanding regularity. With this seminar we want to present in an elementary way these examples, introducing step by step all the necessary ingredients from monoid theory and module theory.

Wednesday 16 February 2011

Numerical simulation of the electrical behavior of Carbon Nanotubes

Vittorio RISPOLI (Univ. Padova, Dip. Mat.)

We introduce in an elementary way the physical setting used to model the electrical behavior of metallic Carbon Nanotubes (CNTs); our aim is to compute the current induced in a CNT by an external electrical field. In the proposed setting, the temporal evolution of electrons and phonons (the last ones needed to take into account quantum mechanics effects) is described by a system of Boltzmann Equations, a system of hyperbolic equations with collision terms. We will give an overview of the general theory on the numerical treatment of such type of equations and present two schemes with some details.

Wednesday 23 February 2011

From Shafarevich's conjecture to finite flat group schemes

Hendrik VERHOEK (Univ. Roma 2)

I will give an introduction to, and overview of, the work of Fontaine, Abrashkin, Schoof, Brumer-Kramer and Calegari that came forth from Shafarevich's conjecture or question about the nonexistence of non-zero abelian varieties over $\mathbb{Q}$ with everywhere good reduction. First I briefly discuss what this conjecture is about. Then we will see the work of the pioneers Fontaine and Abrashkin

passing by: they independently proved there does not exist such an abelian variety. After that we will consider the work of Schoof, Brumer and Kramer and Calegari related to abelian varieties over $\mathbb{Q}$ with so called semi-stable reduction at some places. Finally I will say something about generalizations and open problems.

Wednesday 9 March 2011

Mean-variance Optimization Problems in Financial Mathematics

Claudio FONTANA (Univ. Padova, Dip. Mat.)

Quadratic optimization criteria are ubiquitous in applied mathematics. In particular, they have been successfully exploited in financial mathematics in the context of hedging and portfolio selection problems, beginning with the Nobel prize-winning work of Markowitz (1952). In this introductory talk, we will survey the main aspects of mean-variance optimization problems, both from a mathematical and a financial point of view. Furthermore, we shall present an abstract and unifying approach for the solution of mean-variance problems, together with the related issue of mean-variance indifference valuation.

Wednesday 23 March 2011

Approximating Goldbach conjecture

Valentina SETTIMI (Univ. Padova, Dip. Mat.)

The Goldbach conjecture is one of the oldest unsolved problems in the entire mathematics and, since its appearance in 1742 to nowadays, a lot of mathematicians dealt with it. In my talk I will give an introduction to the origin of the Goldbach conjecture and then I will describe the most important developments in some problems related to it. In particular I will talk about the ternary Goldbach conjecture, the exceptional set in Goldbach's problem and the Goldbach-Linnik problem. Finally, I will give a short overview of our results which can be seen as approximations to the Goldbach-Linnik problem.

Wednesday 6 April 2011

Robustness for path-dependent volatility models

Mauro ROSESTOLATO (S.N.S. Pisa)

In this introductory talk, we present a 2-dimensional market model in which only one component (say S) is observable in the market, while the other one (say P) is not observable, thus the choice of the starting point $(S(0), P(0))$ is a-priori subject to an error. This is the reason why we are

interested firstly in investigating the dependence of the (S, P)-dynamics with respect to the initial condition, secondly in choosing an initial condition which minimizes the error. The first issue is classical even in the deterministic case. The most intuitive and general approach is to recur to estimates based on Gromwall lemma, while if one instead uses the differentiability with respect to the initial conditions such estimates can be significantly improved. We will review this in the case of ordinary differential equations and see how the results generalize to the current case and how this improvement makes the model applicable in reality. The second issue is treated by techniques based on invariant measures and the clever use of past observations: we will show that, by using the estimates based on the differentiability with respect to the initial conditions, the width of the past window required is linear with respect to the future time horizon. Finally, we present some numerical results..

Wednesday 20 April 2011

The Liouville Theorem for conformal maps: old and new

Alessandro OTTAZZI (Univ. Milano-Bicocca)

In this seminar we discuss a classical result of Liouville for conformal maps in euclidean spaces of dimension at least three. Most of the time will be devoted to present a proof which is not present in this form in the classical literature. This proof works in more generality and in fact we (A. Ottazzi and B. Warhurst) can prove a Liouville theorem for all nilpotent and stratified Lie groups endowed with a sub-Riemannian distance. In the last part of the seminar we shall describe the setting of such groups, and possibly discuss some open problems.

Wednesday 4 May 2011

Large Deviations in Probability Theory

Markus FISCHER (Univ. Padova, Dip. Mat.)

In probability theory, the term large deviations refers to an asymptotic property of the laws of families of random variables depending on a large deviations parameter. A classical example is derived from coin flipping. For each number n, consider the random experiment of tossing n coins. Let $S(n)$ denote the number of coins that land heads up. The quantities $S(n)$ and $S(n)/n$ are random variables, $S(n)/n$ being the empirical mean, here equal to the empirical probability of getting heads. If the coins are fair and tossed independently, then by the law of large numbers $S(n)/n$ will converge to $1/2$ as n tends to infinity. Consequently, given any strictly positive c, the probability that $S(n)/n$ is greater than $1/2+c$ (or less than $1/2-c$) goes to zero as n tends to infinity. But one can say more about the convergence of those probabilities of deviation from the law of large numbers limit. Indeed, the decay to zero is exponentially fast (in the large deviations parameter n) with rates that can be determined exactly. The exponential decay of deviation probabilities is a common property of families of random objects arising in many different contexts.

Wednesday 18 May 2011

A Viscosity approach to Monge-Ampere type PDEs

Marco CIRANT (Univ. Padova, Dip. Mat.)

In this introductory talk we present some results of existence and uniqueness of solutions to the Dirichlet problem for the prescribed gaussian curvature equation, a Monge-Ampere type equation arising in differential geometry. We implement the modern tools of viscosity theory, combined with new ideas of Harvey and Lawson; our point of view is also based upon Krylov's language of elliptic branches. Our case study will be the homogeneous equation, i.e. when the curvature is identically zero, for which we outline the proof of existence and uniqueness of a convex solution (in a weak sense). Then, we sketch how to generalize these kind of results to the non-homogeneous equation and show some open problems related to curvature equations of more general form.

Wednesday 8 June 2011

Identification of Reciprocal Processes and related Matrix Extension Problem

Francesca Paola CARLI (Univ. Padova, D.E.I.)

Stationary reciprocal processes defined on a finite interval of the integer line can be seen as a special class of Markov random fields restricted to one dimension. This kind of processes are potentially useful for describing signals which naturally live in a finite region of the time (or space) line. Non-stationary reciprocal processes have been extensively studied in the past especially by Jamison, Krener, Levy and co-workers. The specialization of the non-stationary theory to the stationary case, however, does not seem to have been pursued in sufficient depth in the literature. Moreover, estimation and identification of reciprocal stochastic models starting from observed data seems still to be an open problem.
This talk addresses these problems showing that maximum likelihood identification of stationary reciprocal processes on the discrete circle leads to a covariance extension problem for block-circulant covariance matrices. This generalizes the famous covariance band extension problem for stationary processes on the integer line. We show that the maximum entropy principle leads to a complete solution of the problem. An efficient algorithm for the computation of the maximum likelihood estimates is also provided.

Wednesday 15 June 2011

On the essential dimension of groups

Dajano TOSSICI (Univ. Milano-Bicocca)

At the end of the nineteenth century many authors (Klein, Hermite, Hilbert for instance) studied the problem of reducing the number of parameters of a generic polynomial of fixed degree. This problem was motivated by the problem of finding a formula, in terms of the usual algebraic operations and of radicals, for the roots of polynomial equations. It is very well known that, later, Galois proved that for polynomial equations of degree greater than 5 this formula does not exists. In 1997 Buhler and Reichstein rewrote and generalized this problem in a more modern context. They introduced the notion of essential dimension of a finite group G, which, very roughly speaking, computes the number of parameters needed to describe all Galois extensions with Galois group G. If we consider the symmetric group S_n then one obtains the number of parameters needed to write a generic polynomial of degree n.
In the talk, after recalling the classical problem described above and the precise definition of essential dimension of a group, we illustrate several examples and open problems. At the very end of the talk, if time is left, we quickly give an overview of results we obtained in collaboration with Angelo Vistoli about essential dimension of a group scheme, which is a generalization of the concept of group in the context of algebraic geometry.

The order complex of the coset poset of a finite group

Massimiliano Patassini (*)

Abstract. Let G be a finite group. Let $\mathcal{C}(G)$ be the coset poset of G, i.e. $\mathcal{C}(G) = \{Hg : H < G, g \in G\}$. This object was studied by Kennet Brown, who pointed out a connection between the Dirichlet polynomial of G, $P_G(s) = \sum_{H \leq G} \frac{\mu_G(H)}{|G:H|^s}$ (where μ_G is the Möbius function of the subgroup lattice of G) and the reduced Euler characteristic $\tilde{\chi}(\Delta)$ of the order complex Δ of $\mathcal{C}(G)$. Indeed, we have: $\tilde{\chi}(\Delta) = -P_G(-1)$. In this seminar we first give an overwiev of the concepts of coset poset, order complex and Möbius function. Next we introduce the work of Kennet Brown concerning the order complex of the coset poset of a soluble group. Last we give an idea of our result about the non-contractibility of the order complex of the coset poset of a classical group.

Consider a polyhedron in the three-dimensional Euclidean space. The Euler characteristic of the polyhedron is given by

$$\chi = V - E + F$$

where V is the number of vertices, E is the number of edges, and F is the number of faces of the polyhedron. The number χ describes the shape of the polyhedron as a topological space.

How to extend this concept? Starting from a poset, we construct a suitable topology, we define an Euler characteristic and we study how they are related.

The theory of poset topology evolved from the seminal 1964 paper of Gian-Carlo Rota on the Möbius function of a partially ordered set (poset). This theory provides a deep and fundamental link between combinatorics and other branches of mathematics. In particular, we are interested in the connection with group theory, developed in the work of Kennet Brown (see [2]).

So, what is poset topology? By the topology of a poset we mean the topology of a certain simplicial complex associated with the poset, called the order complex of the poset.

(*) Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: mpatassi@math.unipd.it . Seminar held on 20 October 2010.

1 The Coset poset and the Dirichlet polynomial of a finite group

1.1 The coset poset of a group

Let G be a finite group. We denote by $\mathcal{C}(G)$ the set of proper right cosets of G, i.e.

$$\mathcal{C}(G) = \{Hg : H < G, g \in G\}.$$

This is called the *coset poset* of G. Indeed, $\mathcal{C}(G)$ is a poset ordered by inclusion. In particular, note that if $H_1g_1 \subseteq H_2g_2$ then $H_2g_2 = H_2g_1$ and $H_1 \leq H_2$.

EXAMPLE. Let $G = \langle x, y : x^2 = y^2 = 1, (xy)^2 = 1\rangle \cong C_2 \times C_2$. The proper subgroups of G are 1, $\langle x\rangle = \{1, x\}$, $\langle y\rangle = \{1, y\}$, $\langle xy\rangle = \{1, xy\}$. So, the coset poset is:

$$\mathcal{C}(G) = \{\{1\}, \{x\}, \{y\}, \{xy\}, \{1, x\}, \{y, xy\}, \{1, y\}, \{x, xy\}, \{1, xy\}, \{x, y\}\}.$$

A poset is said to be *bounded* if it has a maximum element (called top element) and a minimum element (called bottom element). The coset poset is not bounded, but we can obtain a bounded poset adding the elements G and $\varnothing$ to $\mathcal{C}(G)$. So we obtain the poset $\tilde{\mathcal{C}}(G) = \mathcal{C}(G) \cup \{G, \varnothing\}$, which is called the *bounded extension* of $\mathcal{C}(G)$.

1.2 Möbius functions and Dirichlet polynomials

An useful tool to investigate some properties of a poset P is the Möbius function. Let P be a bounded poset with top element $\hat{1}$. The *Möbius function* on P is defined by

$$\mu_P(x) = \begin{cases} 1 & \text{if } x = \hat{1}, \\ -\sum_{x<y\leq\hat{1}} \mu_P(y) & \text{if } x < \hat{1}, \end{cases}$$

for $x \in P$. In Figure 1 there is an example of the values of the Möbius function for a bounded poset. Note that if x is a maximal element of a poset P (i.e. an element $x \in P$ such that if $y > x$, then $y = \hat{1}$), then $\mu_P(x) = -1$. Moreover, if x is not intersection of maximal element of a poset P, then $\mu_P(x) = 0$ (see [3]).

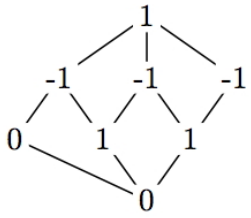


Figure 1: Möbius function for a bounded poset.

Now, let $\mathcal{S}(G)$ be the poset of subgroups of G. Note that $\mathcal{S}(G)$ is a bounded poset with top element G and bottom element $\{1\}$. Denote by μ_G the Möbius function $\mu_{\mathcal{S}(G)}$. We can define two Dirichlet finite series associated to the posets $\mathcal{S}(G)$ and $\tilde{\mathcal{C}}(G)$, in the following way:

$$P_G(s) = \sum_{H\leq G} \frac{\mu_G(H)}{|G : H|^s}$$

and

$$P_{\tilde{\mathcal{C}}(G)}(s) = \sum_{x \in \tilde{\mathcal{C}}(G) - \{\varnothing\}} \frac{\mu_{\tilde{\mathcal{C}}(G)}(x)}{(|G| : |x|)^s},$$

where $|x|$ denotes the size of x and $|G : H|$ is the index of H in G. In particular, the first is called the *Dirichlet polynomial* of G and its multiplicative inverse is called the Probabilistic zeta function of G (see [1] and [5]).

By definition of Möbius function, we have that

$$\mu_{\tilde{\mathcal{C}}(G)}(\varnothing) = - \sum_{x \in \tilde{\mathcal{C}}(G) - \{\varnothing\}} \mu_{\tilde{\mathcal{C}}(G)}(x) = -P_{\tilde{\mathcal{C}}(G)}(0).$$

In [2, §9], Brown noted that

$$P_G(s) = P_{\tilde{\mathcal{C}}(G)}(s+1),$$

hence, in particular,

$$P_G(-1) = P_{\tilde{\mathcal{C}}(G)}(0) = -\mu_{\tilde{\mathcal{C}}(G)}(\varnothing). \quad (\dagger)$$

In the next sections we want to show that the number $\mu_{\tilde{\mathcal{C}}(G)}(\varnothing)$ has an important topological meaning. In particular, we need to give a topological structure to a poset. This can be done associating a simplicial complex to $\mathcal{C}(G)$.

2 Simplicial and order complexes

2.1 Simplicial complexes

An abstract *simplicial complex* Δ on finite vertex set V is a nonempty collection of subsets of V such that

- $\{v\} \in \Delta$ for all $v \in V$,
- if $G \in \Delta$ and $F \subseteq G$, then $F \in \Delta$.

The elements of Δ are called *faces* (or simplices) of Δ. We say that a face F has *dimension* d and write $\dim F = d$ if $d = |F| - 1$. In particular, by definition $\varnothing \in \Delta$ and we have $\dim(\varnothing) = -1$. The *dimension* $\dim \Delta$ of Δ is the maximum of the dimensions of the faces.

Now we want to give a topological structure to the simplicial complex. We need some definitions.

Let $P_0, ..., P_m$ be points of the Euclidean space $\mathbb{R}^n$. The *convex hull* $\langle P_0, ..., P_m \rangle$ of $P_0, ..., P_m$ is the set of points of the form

$$\sum_{i=0}^{m} \lambda_i P_i$$

where $\sum_{i=0}^{m} \lambda_i = 1$ and $\lambda_i \geq 0$ for all $i \in \{0, ..., m\}$. The points $P_0, ..., P_m$ are *affinely independent* if $P_1 - P_0, ..., P_m - P_0$ are linearly independent vectors of $\mathbb{R}^n$.

TOPOLOGICAL STRUCTURE OF A SIMPLICIAL COMPLEX. Let Δ be a simplicial complex with vertex set $V = \{v_0, ..., v_k\}$. Let $E_0, ..., E_k$ be affinely independent points in $\mathbb{R}^k$. Let $\Gamma : \Delta - \{\varnothing\} \to \mathbb{R}^k$ be the map such that $\Gamma(F) = \langle E_{v_{i_1}}, ..., E_{v_{i_l}} \rangle$ for each $F \in \Delta - \{\varnothing\}$, where $F = \{v_{i_1}, ..., v_{i_l}\}$.

A *geometric realization* $|\Delta|$ of V is a topological space homeomorphic to $\bigcup_{F \in \Delta} \Gamma(F)$. This gives a topological structure to the simplicial complex Δ. When we say that Δ has a certain topological property (such as contractibility, homotopy,...) we mean that its geometric realization $|\Delta|$ has this property.

In Figure 2 we give a geometric realization of the simplicial complexes with three vertices.

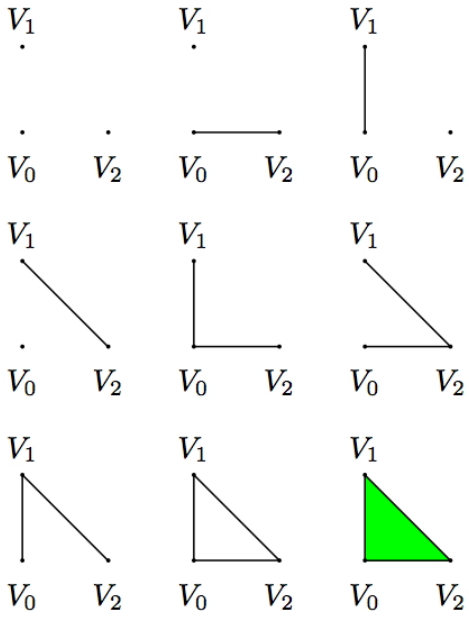


Figure 2: Geometric realization of the simplicial complexes with three vertices.

2.2 The order complex

Let P be a poset. In order to have a topology for P, we construct a simplicial complex $\Delta(P)$ associated to P, called the *order complex of* P. The vertices of $\Delta(P)$ are the elements of P and the faces of $\Delta(P)$ are the chains (i.e. totally ordered subsets) of P.

EXAMPLE. Let $G = \langle x, y : x^2 = y^2 = (xy)^2 = 1 \rangle \cong C_2 \times C_2$. The geometric realization of the order complex $\Delta(\mathcal{C}(G))$ is given in Figure 3, where $z = xy$.

EXAMPLE. Let $G = \langle x : x^8 = 1 \rangle \cong C_8$. Let $A = \langle x^4 \rangle$ and $B = \langle x^2 \rangle$. The coset poset of C_8 is:

$$\mathcal{C}(C_8) = \{\{1\}, \{x\}, \{x^2\}, \{x^3\}, \{x^4\}, \{x^5\}, \{x^6\}, \{x^7\}, A, Ax, Ax^2, Ax^3, B, Bx\}.$$

The geometric realization of the order complex $\Delta(\mathcal{C}(G))$ is given in Figure 4.

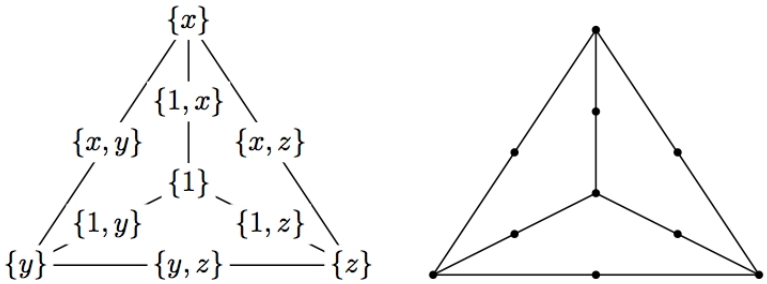


Figure 3: Geometric realization of the order complex $\Delta(\mathcal{C}(C_2 \times C_2))$.

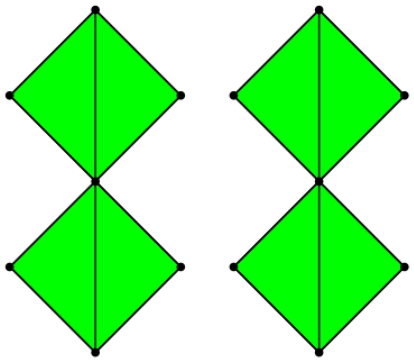


Figure 4: Geometric realization of the order complex $\Delta(\mathcal{C}(C_8))$.

2.3 Contractibility and Euler characteristic

An important concept in topology is the homotopy.

Let X and Y be two topological spaces. Two continuous function $f, g : X \to Y$ are said to be *homotopic* if there exists a continuous function $H : X \times [0,1] \to Y$ such that, if $x \in X$ then $H(x,0) = f(x)$ and $H(x,1) = g(x)$.

Two topological spaces X and Y are *homotopy equivalent* or *of the same homotopy type* if there exist continuous maps $f : X \to Y$ and $g : Y \to X$ such that $f \circ g$ is homotopic to id_Y and $g \circ f$ is homotopic to id_X.

A topological space X is said to be *contractible* if the identity map id_X is homotopic to a constant function. Roughly speaking, a topological space is contractible if it can be continuosly shrunk to a point.

Example. The order complex $\Delta(\mathcal{C}(C_2 \times C_2))$ is homotopy equivalent to a bouquet of three circles and the order complex $\Delta(\mathcal{C}(C_8))$ is homotopy equivalent to two points.

Now we introduce an useful tool which aid us to see if a simplicial complex is contractible. Let Δ be a simplicial complex. The *reduced Euler characteristic* of Δ is

$$\tilde{\chi}(\Delta) = \sum_{d=-1}^{\dim \Delta} (-1)^d \alpha_d,$$

where α_d is the number of simplices of dimension d.

It turns out that if Δ is contractible, then $\tilde{\chi}(\Delta) = 0$.

EXAMPLE. The reduced Euler characteristic of $\Delta(\mathcal{C}(C_2 \times C_2))$ is $\tilde{\chi}(\Delta(\mathcal{C}(C_2 \times C_2))) = -1 + 10 - 12 = -3$. As well, $\tilde{\chi}(\Delta(\mathcal{C}(C_8))) = -1 + 14 - 20 + 8 = 1$.

2.4 Euler characteristic and Dirichlet polynomials

By a result of Philip Hall, we have that if P is a poset, then

$$\tilde{\chi}(\Delta(P)) = \mu_{\hat{P}}(\hat{0}),$$

where $\hat{P}$ is the poset obtained from P adding a bottom element $\hat{0}$ and a top element $\hat{1}$ (even when P has already a top and a bottom element).

This result, together with (†), shows that for a finite group G we have:

$$P_G(-1) = P_{\tilde{\mathcal{C}}(G)}(0) = -\mu_{\tilde{\mathcal{C}}(G)}(\varnothing) = -\tilde{\chi}(\Delta(\mathcal{C}(G))))$$

EXAMPLE. Let $G = \langle x, y : x^2 = y^2 = (xy)^2 = 1 \rangle \cong C_2 \times C_2$. We know, by the previous example, that $\tilde{\chi}(\Delta(\mathcal{C}(C_2 \times C_2))) = -3$. We want to compute $P_G(-1)$. The subgroups of G are $G, \langle x \rangle, \langle y \rangle, \langle xy \rangle, 1$. It is straightforward to show that $\mu_G(G) = 1$, $\mu_G(\langle x \rangle) = \mu_G(\langle y \rangle) = \mu_G(\langle xy \rangle) = -1$ and $\mu_G(1) = -\sum_{1<H\leq G} \mu_G(H) = -(-1 - 1 - 1 + 1) = 2$. Hence we get

$$P_G(s) = 1 - \frac{3}{2^s} + \frac{2}{4^s},$$

so $P_G(-1) = 3$ as we wanted.

3 The work of Kennet Brown

Note that the knowledge of the value $P_G(-1)$ gives us information about the contractibility of the order complex of the coset poset $\mathcal{C}(G)$. Indeed, if $P_G(-1) \neq 0$, then $\Delta(\mathcal{C}(G))$ is not contractible.

In [2], Kennet Brown studied the order complex of the coset poset for the soluble groups and he proved the following.

Proposition 1 [2, Proposition 8] *Let G be a finite soluble group and let d be the number of non-Frattini chief factors of G. If G is a cyclic group of prime power order, then $\Delta(\mathcal{C}(G))$ has the homotopy type of p points (where p is the prime divisor of $|G|$), otherwise $\Delta(\mathcal{C}(G))$ has the homotopy type of a bouquet of $(d-1)$-spheres and the number of spheres is $|\tilde{\chi}(\Delta(\mathcal{C}(G)))| = |P_G(-1)|$.*

EXAMPLE. $G = \langle x, y : x^2 = y^2 = (xy)^2 = 1\rangle \cong C_2 \times C_2$ is a soluble group. A chief series for G is $1 \leq \langle x\rangle \leq G$. Clearly $\langle x\rangle$ is non-Frattini since it is complemented by $\langle y\rangle$. Hence G has 2 non-Frattini chief factors. Since $|P_G(-1)| = 3$, by Proposition 1, the order complex $\Delta(\mathcal{C}(G))$ has the homotopy type of a bouquet of three 1-spheres (i.e. circles), as we proved above.

The non-soluble case is more difficult. Indeed, the coset poset becomes very huge and complicated. Supported by some computational evidences, Brown propoused the following conjecture.

Conjecture 2 *Let G be a finite group. Then $P_G(-1)$ does not vanish. Hence the order complex of the coset poset of G is not contractible.*

4 The order complex of the coset poset of a classical group

In the first part of our PhD thesis we prove the following result.

Theorem 3 *Let G be a classical group and assume that G does not contain non-trivial graph automorphisms. Then $P_G(-1)$ does not vanish, hence the order complex associated to the coset poset of G is not contractible.*

The main idea of the proof is the following. It is quite easy to reduce to the case when G is a simple group of characteristic p. In particular, a vector space V over a field of characteristic p and a form κ (which is zero, unitary, symplectic or orthogonal) are associated to G.

Write

$$P_G(s) = \sum_{k\geq 1} \frac{a_k(G)}{k^s} \quad \text{where } a_k(G) = \sum_{H\leq G, |G:H|=k} \mu_G(H).$$

Clearly we have:

$$P_G(s) = P_G^{(p)}(s) + R_G^{(p)}(s),$$

where

$$P_G^{(p)}(s) = \sum_{p\nmid k} \frac{a_k(G)}{k^s} \text{ and } R_G^{(p)}(s) = \sum_{p|k} \frac{a_k(G)}{k^s}.$$

In order to prove our claim, we show that $|P_G^{(p)}(-1)|_p < |R_G^{(p)}(-1)|_p$, where $|k|_p$ is the greatest power of p dividing the integer k (we set $|0|_p = 0$).

By definition of $a_k(G)$, we have that if $a_k(G) \neq 0$, then there exists a subgroup H of G such that

- $|G : H| = k$ and
- $\mu_G(H) \neq 0$, hence H is an intersection of maximal subgroups of G.

In this case we say that the subgroup H is *contributing* for $a_k(G)$.

4.1 The p-part of $P_G^{(p)}(-1)$

Let k be an integer such that $a_k(G) \neq 0$ and p does not divide k. By definition, if H is a contributing subgroup for $a_k(G)$, then

- $|G : H| = k$, hence $|G : H|_p = 1$, so H contains a Sylow p-subgroup of G;
- H is an intersection of maximal subgroups of G.

These conditions imply that H is a parabolic subgroup of G. The structure of the parabolic subgroups is well known and we can gather enough information in order to find the p-part of $P_G^{(p)}(-1)$. For example, if $G = \mathrm{PSL}_n(q)$, then we have that $|P_G^{(p)}(-1)|_p = q^{n-1}|2|_p$.

The proof of this fact requires a certain amount of combinatorics and some results on root systems.

4.2 The p-part of $R_G^{(p)}(-1)$

Let k be an integer such that $a_k(G) \neq 0$ and p divides k. By definition, if H is a contributing subgroup for $a_k(G)$, then

- $|G : H| = k$, hence $|G : H|_p > 1$;
- H is an intersection of maximal subgroups of G.

We consider two cases:

(A) H is contained in a maximal subgroup M of G such that $|G : M|_p > 1$.

(B) if a maximal subgroup M of G contains H, then $|G : M|_p = 1$, i.e. M is a maximal parabolic subgroup of G.

Suppose that case (A) holds. Using the results on the maximal subgroups of the classical groups we can prove the following.

Proposition 4 *Let G be a classical simple group defined over a field of characteristic p. If M is a maximal subgroup of G such that $|G : M|_p > 1$, then $|G : M|_p^2 > |P_G^{(p)}(-1)|_p$. Thus, if $a_k(G) \neq 0$ and p divides k, then $|k|_p^2 > |P_G^{(p)}(-1)|_p$.*

Assume that case (B) holds. There exists a 1-1 correspondence between the set of non-trivial totally singular proper subspaces of V (when we say that a vector subspace W of V is totally singular, we mean that the restriction κ_W of the form κ is the zero form) and the parabolic maximal subgroups of G, given by $W \mapsto \mathrm{Stab}_G(W) = \{g \in G : g(W) = W\}$.

We denote by $\mathcal{L}_H$ the set of non-trivial totally singular proper subspaces W of V such that $\mathrm{Stab}_G(W) \geq H$. In particular, the image of the restriction of Stab_G to $\mathcal{L}_H$ is the set of parabolic maximal subgroups of G containing H.

We say that $\mathcal{L}_H$ **fulfills the property** $\mathcal{P}$ if there exists $W \in \mathcal{L}_H$ such that for each $U \in \mathcal{L}_H$, we have $W \leq U$ or $W \geq U$.

Theorem 5 *Assume that case (B) holds.*

- *If $\mathcal{L}_H$ fulfills the property $\mathcal{P}$, then $\mu_G(H) = 0$.*
- *If $\mathcal{L}_H$ does not fulfill the property $\mathcal{P}$, then $|G : H|_p^2 > |P_G^{(p)}(-1)|_p$.*

EXAMPLE. Let $G = \mathrm{PSL}_4(q), q$ odd, let $V = \langle e_1, ..., e_4 \rangle$. Let $W_1 = \langle e_1 \rangle$ and $W_2 = \langle e_2 \rangle$. Assume that $H = M_1 \cap M_2$, $M_i = \mathrm{Stab}_G(W_i)$ for $i = 1, 2$.

Then case (B) holds. Note that $|G : H|_p = q$ and $|P_G^{(p)}(-1)|_p = q^3$, so $|G : H|_p^2 \not> |P_G^{(p)}(-1)|_p$. But $\mathcal{L}_H = \{W_1, W_2, W_1 + W_2\}$ and $\mathcal{L}_H$ fulfills the property $\mathcal{P}$.

Let $M_3 = \mathrm{Stab}_G(W_1 + W_2)$. The lattice generated by the maximal subgroups over H and the values of the Möbius function are in Figure 5. So we have that $\mu_G(H) = 0$.

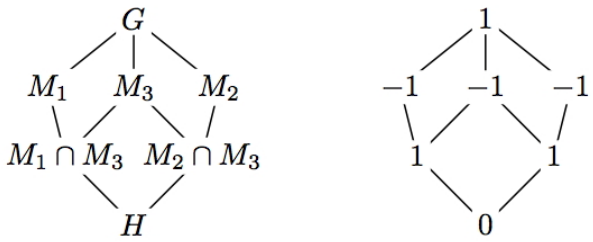


Figure 5: Lattice and Möbius numbers over H.

In general, the idea is: find a *redundant* element W in $\mathcal{L}_H$, i.e. for each $M \subseteq \mathcal{L}_H$ such that $W \in M$,

$$\bigcap_{U \in M} \mathrm{Stab}_G(U) = H \Rightarrow \bigcap_{U \in M - \{W\}} \mathrm{Stab}_G(U) = H.$$

If $\mathcal{L}_H$ has the property $\mathcal{P}$, then there exists a redundant element.

In the previous example, the subspace $W_1 + W_2$ is redundant.

4.3 The last argument

We have shown that if p divides k and $a_k(G) \neq 0$, then $|k|_p^2 > |P_G^{(p)}(-1)|_p$. Now we apply a result on the coefficient of the Dirichlet polynomial of G.

Lemma 6 (See [4]) *Let G be a perfect group and let k be a positive integer. Then k divides $a_k(G)$.*

Since a simple group is perfect, we obtain that

$$|R_G^{(p)}(-1)|_p = \left| \sum_{p|k} a_k(G) k \right|_p \geq \min\{|a_k(G)k|_p : p|k\} \geq \min\{|k|_p^2 : p|k\} > |P_G^{(p)}(-1)|_p.$$

This completes the proof of our main theorem.

References

[1] N. Boston, *A probabilistic generalization of the Riemann zeta function.* Analytic Number Theory 1 (1996), 155–162.

[2] K. S. Brown, *The coset poset and the probabilistic zeta function of a finite group.* J. Algebra 225 (2000), 989–1012.

[3] P. Hall, *The Eulerian Functions of a group.* Quart. J. Math. 7 (1936), 134–151.

[4] T. Hawkes, M. Isaacs and M. Özaydin, *On the Mbius function of a finite group.* Rocky Mountain Journal 19 (1989), 1003–1034.

[5] A. Mann, *Positively finitely generated groups.* Forum Math. 8 (1996), 429–459.

The concept of supermodularity in aggregation functions and copulas

Maddalena Manzi (*)

Abstract. The mathematical concept of supermodularity is known in the literature for functions on a general lattice, but in this paper we define supermodularity for aggregation functions. In particular supermodularity is the main axiom for bivariate copulas, which are a subclass of supermodular aggregation functions. A simple way to generalize the axiom of supermodularity for bivariate copulas is given by the concept of ultramodularity. So, we characterize Archimedean ultramodular copulas and we develop some connections with Choquet integral.
(Keywords: Copula, supermodularity, ultramodularity, aggregation function, Choquet integral.)

1 Introduction

Supermodular functions are extensively investigated in different research areas, both pure and applied. The supermodular property also goes by a variety of names such as L-superadditive (where L is mnemonic for lattice), superadditive and quasimonotone. Our aim is to apply this concept to aggregation functions, but, first of all, we recall the basic definitions and properties both for aggregation functions and copulas. Moreover, we will focus our attention to the main problem that we have when we want to deal with multivariate copulas as aggregation functions.
Aggregation operators (also referred to as *means* or *mean operators*) correspond to particular mathematical functions used for information fusion, the broad area that studies methods to combine data or information supplied by multiple sources. Generally, we consider mathematical functions that combine a finite number of inputs, called arguments, into a single output. So, aggregation has for purpose the simultaneous use of different pieces of information provided by several sources, in order to come to a conclusion or a decision. They are applied in many different domains and in particular aggregation functions play important role in different approaches to decision making, where values to be aggregated are typically preference or satisfaction degrees and thus belong to the unit interval $[0, 1]$. For more details, see [7].

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: mmanzi@math.unipd.it. Seminar held on 3 November 2010.

As it has been shown in [13], we can define an aggregation operator as a function

$$A : \bigcup_{n\in\mathbb{N}} [0,1]^n \to [0,1]$$

that satisfies:

- (Idempotency) $A(x) = x \qquad \forall x \in [0,1]$;
- (Boundary conditions) $A(0,\dots,0) = 0$ and $A(1,\dots,1) = 1$;
- (Monotonicity) $A(x_1,\dots,x_n) \le A(y_1,\dots,y_n)$ if $(x_1,\dots,x_n) \le (y_1,\dots,y_n)$.

Idempotency and monotonicity imply that aggregation operators are functions that yield a value between the minimum and the maximum of the input values. Formally, they are operations that satisfy internality:

$$\min_i x_i \le A(x_1,\dots,x_n) \le \max_i x_i.$$

The paper is organized as follows. In the following section, modular, supermodular and ultramodular aggregation functions are introduced and some basic results are recalled. The last section deals with some special ultramodular aggregation functions, especially with ultramodular copulas and some connections with Choquet integral are proposed.

2 Modular, supermodular, and ultramodular aggregation functions

The concept of modularity and supermodularity was introduced for functions from a lattice L into $\mathbb{R}$.

Definition 2.1 Let $(L,\wedge,\vee)$ be a lattice.

(i) A function $f\colon L \to \mathbb{R}$ is called *modular* if, for all $x, y \in L$,

$$f(x\vee y) + f(x\wedge y) = f(x) + f(y). \tag{1}$$

(ii) A function $f\colon L \to \mathbb{R}$ is called *supermodular* if, for all $x, y \in L$,

$$f(x\vee y) + f(x\wedge y) \ge f(x) + f(y). \tag{2}$$

In the context of aggregation functions, the following characterization of modularity is easily obtained [9]:

Proposition 2.2 *An n-ary aggregation function $A\colon [0,1]^n \to [0,1]$ is modular if and only if there are non-decreasing functions $f_1, f_2, \dots, f_n\colon [0,1] \to [0,1]$ with $f_i(0) = 0$ and $\sum_{i=1}^n f_i(1) = 1$ such that, for all $(x_1,\dots,x_n) \in [0,1]^n$,*

$$A(x_1,\dots,x_n) = \sum_{i=1}^{n} f_i(x_i).$$

The following characterization of supermodular functions $f\colon [0,1]^n \to [0,1]$ is due to [1]:

Proposition 2.3 *An n-ary function $f\colon [0,1]^n \to [0,1]$ is supermodular if and only if each of its two-dimensional sections is supermodular, i.e., for each $\mathbf{x} \in [0,1]^n$ and all $i, j \in \{1, 2, \ldots, n\}$ with $i \neq j$, the function $f_{\mathbf{x},i,j}\colon [0,1]^2 \to [0,1]$ given by $f_{\mathbf{x},i,j}(u,v) = f(\mathbf{y})$, where $y_i = u$, $y_j = v$ and $y_k = x_k$ for $k \in \{1, 2, \ldots, n\} \setminus \{i, j\}$, is supermodular.*

Well-known examples of supermodular n-dimensional aggregation functions (with $n \geq 2$) are *modular aggregation functions* as characterized in Proposition 2.2 and *copulas* as introduced in [16] (see also [8, 15]).

In the case $n = 2$, the supermodularity is even used as an axiom for copulas:

Definition 2.4 An aggregation function $C\colon [0,1]^2 \to [0,1]$ is called a 2-*copula* (or, briefly, a *copula*) if it is supermodular and has 1 as neutral element, i.e., if $C(x,1) = C(1,x) = x$ for all $x \in [0,1]$.

Copulas play an important role in the representation of supermodular binary aggregation functions. The following result is taken from [6]:

Proposition 2.5 *An aggregation function $A\colon [0,1]^2 \to [0,1]$ is supermodular if and only if there are non-decreasing functions $g_1, g_2, g_3, g_4\colon [0,1] \to [0,1]$ with $g_i(1) = 1$ for $i \in \{1,2,3,4\}$ and $g_1(0) = g_2(0) = 0$, a copula $C\colon [0,1]^2 \to [0,1]$ with $C(g_3(0), g_4(0)) = 0$, and numbers $a, b, c \in [0,1]$ with $a + b + c = 1$ such that, for all $(x,y) \in [0,1]^2$,*

$$A(x,y) = a \cdot g_1(x) + b \cdot g_2(y) + c \cdot C(g_3(x), g_4(y)). \tag{3}$$

If 0 is an annihilator of the aggregation function $A\colon [0,1]^2 \to [0,1]$, i.e., if $A(x,0) = A(0,x) = 0$ for all $x \in [0,1]$, then (3) reduces to

$$A(x,y) = C(f(x), g(y)), \tag{4}$$

where $f, g\colon [0,1] \to [0,1]$ are non-decreasing functions with $f(1) = g(1) = 1$ and C satisfies $C(f(0), g(0)) = 0$. Note that then we have $f(x) = A(x,1)$ and $g(x) = A(1,x)$ for all $x \in [0,1]$.

Definition 2.6 An n-ary aggregation function $A\colon [0,1]^n \to [0,1]$ is called *ultramodular* if, for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in [0,1]^n$ with $\mathbf{x} + \mathbf{y} + \mathbf{z} \in [0,1]^n$,

$$A(\mathbf{x} + \mathbf{y} + \mathbf{z}) - A(\mathbf{x} + \mathbf{y}) \geq A(\mathbf{x} + \mathbf{z}) - A(\mathbf{x}). \tag{5}$$

Ultramodularity implies supermodularity of aggregation functions. To see this, for arbitrary $\mathbf{x}, \mathbf{y} \in [0,1]^n$ put first $\mathbf{u} = \mathbf{y} - \mathbf{x} \wedge \mathbf{y}$ and $\mathbf{v} = \mathbf{x} - \mathbf{x} \wedge \mathbf{y}$. Then we get

$$\mathbf{x} \vee \mathbf{y} = \mathbf{x} + \mathbf{y} - \mathbf{x} \wedge \mathbf{y} = \mathbf{x} \wedge \mathbf{y} + \mathbf{u} + \mathbf{v}$$

and, because of (5),

$$\begin{aligned} A(\mathbf{x} \vee \mathbf{y}) + A(\mathbf{x} \wedge \mathbf{y}) &= A(\mathbf{x} \wedge \mathbf{y} + \mathbf{u} + \mathbf{v}) + A(\mathbf{x} \wedge \mathbf{y}) \\ &\geq A(\mathbf{x} \wedge \mathbf{y} + \mathbf{v}) + A(\mathbf{x} \wedge \mathbf{y} + \mathbf{u}) \\ &= A(\mathbf{x}) + A(\mathbf{y}). \end{aligned}$$

In the case of one-dimensional aggregation functions, ultramodularity (5) is just standard convexity. Therefore, ultramodularity can also be seen as an extension of one-dimensional convexity. The following result (Corollary 4.1 of [11]) states the exact relationship between ultramodular and supermodular functions $f\colon [0,1]^n \to [0,1]$:

Proposition 2.7 *A function $f\colon [0,1]^n \to [0,1]$ is ultramodular if and only if f is supermodular and each of its one-dimensional sections is convex, i.e., for each $\mathbf{x} \in [0,1]^n$ and each $i \in \{1,\dots,n\}$ the function $f_{\mathbf{x},i}\colon [0,1] \to [0,1]$ given by $f_{\mathbf{x},i}(u) = f(\mathbf{y})$, where $y_i = u$ and $y_j = x_j$ whenever $j \neq i$, is convex.*

Remark 2.8 Because of Propositions 2.3 and 2.7, for an n-ary aggregation function $A\colon [0,1]^n \to [0,1]$ the following are equivalent:

(a) A is ultramodular;

(b) each two-dimensional section of A is ultramodular;

(c) each two-dimensional section of A is supermodular and each one-dimensional section of A is convex.

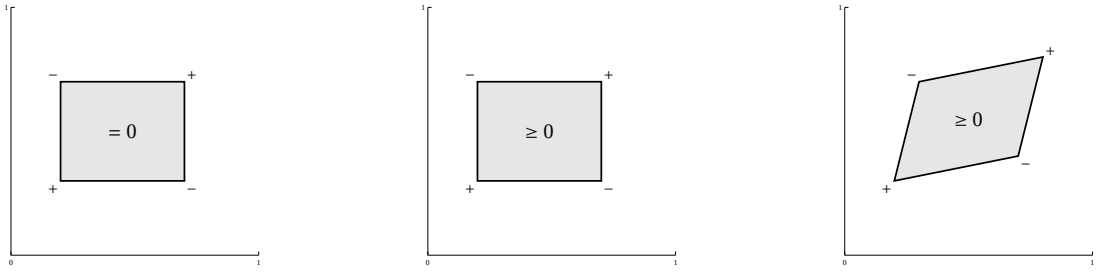


Figure 1: Modularity (left), supermodularity (center), and ultramodularity of a function $f\colon [0,1]^2 \to [0,1]$.

The class of ultramodular aggregation functions is closed under composition (see Theorem 3.2 and Corollary 3.3 in [9]).
A generalization of Proposition 2.5 is the following result (see [9]):

Corollary 2.9 *If A is a bivariate ultramodular aggregation function, then we have*

$$A = \lambda \cdot A_1 + (1-\lambda) \cdot A_2, \tag{6}$$

where A_1 is a modular element, A_2 is a supermodular binary aggregation function with annihilator 0, and $\lambda = 1 - A(1,0) - A(0,1) \in [0,1]$.

3 Special constructions

If an ultramodular binary aggregation function with annihilator 0 has also neutral element 1 then it necessarily is an ultramodular copula, i.e., a copula with convex sections. In statistics, where a copula C describes the dependence structure of a random vector (X, Y), the ultramodularity of C is equivalent to X being stochastically decreasing with respect to Y (and Y being stochastically decreasing with respect to X). Clearly, the set $\mathcal{C}_u$ of all ultramodular binary copulas is convex. The greatest element of $\mathcal{C}_u$ is the product Π, and the smallest element of $\mathcal{C}_u$ is the lower Fréchet-Hoeffding bound W.

3.1 Archimedean copulas

A binary aggregation function $C\colon [0,1]^2 \to [0,1]$ is an Archimedean copula if and only if there is a continuous, strictly decreasing convex function $t\colon [0,1] \to [0,\infty]$ with $t(1) = 0$ such that for all $(x, y) \in [0, 1]$ (see [14])

$$C(x,y) = t^{-1}(\min(t(x) + t(y), t(0))). \tag{7}$$

The function t is called an additive generator of C, and it is unique up to a positive multiplicative constant.

If we want to see whether an Archimedean copula is ultramodular, i.e., has convex horizontal and vertical sections, its symmetry (as a consequence of (7) and boundary conditions tell us that it suffices to check the convexity of all horizontal sections for $a \in]0,1[$. The following results are construction methods of Archimedean ultramodular copulas [9].

Theorem 3.1 *Let $C\colon [0,1]^2 \to [0,1]$ be an Archimedean copula with a two times differentiable additive generator $t\colon [0,1] \to [0,\infty]$. Then C is ultramodular if and only if $\frac{1}{t'}$ is a convex function.*

Theorem 3.2 *Let C be an Archimedean copula with additive generator t, let t' be the left derivative of t on $]0,1]$ and $t'(0) = t'(0^+)$. Then all the one-dimensional sections of C are concave if and only if $t'(0) = \infty$, t' is finite on $]0,1]$, and $\frac{1}{t'}$ is concave.*

For other examples see Section 5 in [9].

3.2 Connections with Choquet integral

When constructing ultramodular aggregation functions, we can focus on special types of aggregation functions. However, in some cases the ultramodularity can be a contradictory or rather restrictive requirement. For instance, *disjunctive aggregation functions* (such as *triangular conorms* [10]) cannot be ultramodular. As an example of the second type we recall the *Choquet integral* [2, 3] and present the necessary details.

If $n \in \mathbb{N}$ and $X = \{1, \ldots, n\}$ then, for a capacity m on X, i.e., a non-decreasing function $m\colon 2^X \to [0,1]$ with $m(\emptyset) = 0$ and $m(X) = 1$, and $\mathbf{x} \in [0,1]^n$ the Choquet

integral [2] is given by

$$\begin{aligned}\mathrm{Ch}(m,\mathbf{x}) &= \int_0^1 m(\{x_i \geq u\})\,du \\ &= \sum_{i=1}^{n} x_{\pi(i)}\big(m(\{\pi(i),\ldots,\pi(n)\}) - m(\{\pi(i+1),\ldots,\pi(n)\})\big),\end{aligned}$$

where $\pi\colon X \to X$ is a permutation of X with $x_{\pi(1)} \leq x_{\pi(2)} \leq \cdots \leq x_{\pi(n)}$ and, by convention, $\{\pi(n+1),\pi(n)\} = \emptyset$.

For a fixed capacity m, the function $\mathrm{Ch}_m\colon [0,1]^n \to [0,1]$ given by $\mathrm{Ch}_m(\mathbf{x}) = \mathrm{Ch}(m,\mathbf{x})$ is an aggregation function, a so-called *Choquet integral-based aggregation function.*

Proposition 3.3 *Let $\mathrm{Ch}_m\colon [0,1]^n \to [0,1]$ be a Choquet integral-based aggregation function based on a capacity m on $X = \{1,\ldots,n\}$. Then we have:*

(i) *Ch_m is superadditive, i.e., for all $\mathbf{x},\mathbf{y} \in [0,1]^n$ with $\mathbf{x}+\mathbf{y} \in [0,1]^n$ we have*

$$\mathrm{Ch}_m(\mathbf{x}+\mathbf{y}) \geq \mathrm{Ch}_m(\mathbf{x}) + \mathrm{Ch}_m(\mathbf{y}),$$

if and only if the capacity m is supermodular.

(ii) *Ch_m is ultramodular if and only if the capacity m is modular, i.e., Ch_m is a weighted arithmetic mean.*

4 Concluding remarks

We have discussed ultramodular aggregation functions, by noting that bivariate copulas are closely linked to the convexity of one-dimensional functions (e.g., additive generators of Archimedean copulas are convex). Supermodularity and ultramodularity are also connected to measure theory. For example it is known that a Choquet integral operator based on a fuzzy measure m is superadditive if, and only if, the fuzzy measure m is supermodular. In particular connections between fuzzy measures and supermodular aggregation functions are important for constructing supermodular aggregation functions in the multivariate case, which allows to extend several properties of copulas as well. In fact, in the multivariate case there are a lot of unsolved problems, in particular with regard to the multivariate decomposition of aggregation functions in a sum of copulas.

References

[1] H. W. Block, W. S. Griffith, and T. H. Savits, *L-superadditive structure functions.* Adv. in Appl. Probab. 21/4 (1989), 919–929.

[2] G. Choquet, *Theory of capacities.* Ann. Inst. Fourier, Grenoble, 5 (1955), 1953–1954.

[3] D. Denneberg, "Non-additive measure and integral". Volume 27 of *Theory and Decision Library. Series B: Mathematical and Statistical Methods*. Kluwer Academic Publishers Group, Dordrecht, 1994.

[4] F. Durante and P. Jaworski, *Invariant dependence structure under univariate truncation*. Submitted for publication.

[5] F. Durante, R. Mesiar, P. L. Papini, and C. Sempi, *2-increasing binary aggregation operators*. Inform. Sci. 177/1 (2007), 111–129.

[6] F. Durante, S. Saminger-Platz, and P. Sarkoci, *On representations of 2-increasing binary aggregation functions*. Inform. Sci. 178/23 (2008), 4534–4541.

[7] J. Fodor and M. Roubens, "Fuzzy preference modelling and multicriteria decision support". Theory and Decision Library. Series D: Systems Theory, Knowledge Engineering and Problem Solving. 14. Dordrecht: Kluwer Academic Publishers, 1994.

[8] H. Joe, "Multivariate Models and Dependence Concepts". Chapman & Hall, London, 1997.

[9] E. P. Klement, M. Manzi, and R. Mesiar, *Ultramodular aggregation functions and a new construction method for copulas*. Submitted for publication.

[10] E. P. Klement, R. Mesiar, and E. Pap, "Triangular Norms". Volume 8 of *Trends in Logic. Studia Logica Library*. Kluwer, Dordrecht, 2000.

[11] M. Marinacci and L. Montrucchio, *Ultramodular functions*. Math. Oper. Res. 30/2 (2005), 311–332.

[12] R. Mesiar, V. Jágr, M. Juráňová, and M. Komorníková, *Univariate conditioning of copulas*. Kybernetika (Prague) 44 (2008), 807–816.

[13] R. Mesiar and S. Saminger, *Domination of ordered weighted averaging operators over t-norms*. Soft Computing 8/1-2 (2004), 562–570.

[14] R. Moynihan, *On τ_T semigroups of probability distribution functions II*. Aequationes Math. 17 (1978), 19–40.

[15] R. B. Nelsen, "An Introduction to Copulas". Volume 139 of *Lecture Notes in Statistics*. Springer, New York, second edition, 2006.

[16] A. Sklar, *Fonctions de répartition à n dimensions et leurs marges*. Publ. Inst. StatistUniv. Paris 8 (1959), 229–231.

Semiconcavity type results of the minimum time function

Khai Tien Nguyen (*)

The minimum time problem is classical in control theory. Given a nonempty closed target $\mathcal{S}$ and a control system

$$\left\{ \begin{array}{ll} \dot{y}(t) \;=\; f(t, y(t), u(t)) & a.e. \\ u(t) \;\in\; \mathcal{U} & a.e. \\ y(0) \;=\; x, & \end{array} \right. \tag{1}$$

where the function $f : \mathbb{R} \times \mathbb{R}^N \times \mathcal{U} \to \mathbb{R}^N$ is smooth enough and the control set $\mathcal{U}$ is a compact nonempty subset of $\mathbb{R}^M$, for each admissible control $u(\cdot) \in \mathcal{U}_{ad}$, i.e. $u(\cdot)$ is measurable and takes value in $\mathcal{U}$, there exists a unique solution $y^{x,u}(\cdot)$ of (1) which is the trajectory starting from x under the control $u(\cdot)$. The minimum time needed to steer x to $\mathcal{S}$, regarded as a function of x, is called the minimum time function and is denoted by

$$T_{\mathcal{S}}(x) := \inf \{\theta_{\mathcal{S}}(x, u) \mid u(\cdot) \in \mathcal{U}_{\mathrm{ad}}\},$$

where $\theta_{\mathcal{S}}(x, u) := \inf \{t \geq 0 \mid y^{x,u}(t) \in \mathcal{S}\}$. In general, $T_{\mathcal{S}} \in [0, \infty]$. The controllable set $\mathcal{R}$ consists of all points $x \in \mathbb{R}^N$ such that $T_{\mathcal{S}}(x)$ is finite. The regularity of the minimum time function is related on one hand to the controllability properties of system (1), on the other one to the regularity of the target and of the dynamics, together with suitable relations between them.

Such topics were studied by several authors (see, e.g., [1, 2, 5, 6, 7, 8, 9, 15, 26] and reference therein) under different viewpoints. In particular, it is well known that in general the minimum time function T is not everywhere differentiable. It is also well known that suitable controllability conditions imply the Hölder continuity of T (see, e.g., [1, Chapter IV] and references therein). However, the latter fact does not provide information on differentiability. In a 1995 paper (see [7] and also Chapter 8 in the book [8]), Cannarsa and Sinestrari found a connection between the control system and the target which actually implies the semiconcavity (or the semiconvexity) of T. Semiconcave functions are – essentially – $\mathcal{C}^2$-perturbations of concave functions and therefore inherit several regularity

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: khai@math.unipd.it. Seminar held on 17 November 2010.

properties from convexity. Several features of semiconcavity were thoroughly studied (see Chapters 3, 4, 5 in [8] and references therein), thus providing a rich set of information on the structure of the minimum time function and suggesting semiconcavity/semiconvexity as a good regularity class for such value functions. The main result in [7] shows that if the target satisfies a *uniform internal ball condition* and the control system is smooth enough, then T is semiconcave, provided a strong *controllability assumption*, called Petrov condition, holds. A partially symmetric result, contained in [7], states that if the target is convex and the control system is linear, then T is semiconvex, provided, again, Petrov condition holds. The latter requires that the minimized Hamiltonian at all boundary points of $\mathcal{S}$, computed along unit normal vectors, be bounded away from zero locally uniformly, i.e., for all $R > 0$ there exists $\mu > 0$ such that for all $x \in \partial\mathcal{S} \cap B(0, R)$,

$$\min_{u\in\mathcal{U}} \langle f(x, u), \zeta\rangle < -\mu, \quad \text{for all } \zeta \in N_{\mathcal{S}}(x),\ \|\zeta\| = 1. \tag{2}$$

In an entirely different setting, a class of sets which includes both convex and $\mathcal{C}^2$-sets was studied independently by several authors (including Federer [18], Canino [4], Clarke, Stern and Wolenski [12], Poliquin, Rockafellar and Thibault [25]) under different names, for example *sets with positive reach* [18], *φ-convex sets* [4], *proximally smooth sets* [12], and *prox-regular sets* [25]. Such sets, which in this thesis will be called sets with positive reach, are characterized by a strong external sphere condition every normal vector must be realized by a locally uniform ball. By observing that a convex set satisfies the same type of external sphere condition with an arbitrarily large radius, it is natural to expect that sets with positive reach enjoy locally several properties that convex sets enjoy globally. In particular, this holds for the metric projection, which is unique in a neighborhood of a set with positive reach K. This fact is used in proving all the regularity properties which are satisfied by sets with positive reach (see, e.g., [18, Section 4]). Semiconcave functions and sets with positive reach, through the hypograph, are linked together (see, e.g., Theorem 5.2 in [12], where semiconvex functions are called *lower–$\mathcal{C}^2$*): a locally Lipschitz function is semiconcave if and only if its hypograph has positive reach. Of course an entirely symmetric characterization for semiconvex functions can be expressed using the epigraph. Trying to generalize to functions whose hypo/epigraph has positive reach some regularity properties enjoyed by semiconcave/convexity functions was therefore a natural challenge. Some results on this line were obtained in [13, 14], including the a.e. twice differentiability together some results on the structure of singularities.

In several control problems, controllability assumptions weaker than Petrov condition hold, and therefore the minimum time function is not locally Lipschitz. A natural question therefore is trying to understand whether the structure of the minimum time function remains unchanged if in the above setting the controllability assumptions are weakened. In other words it is natural to investigate whether the hypograph/epigraph of T has positive reach if T is supposed to be only continuous.

We first assume that the nonlinear control system is (essentially) $\mathcal{C}^2$, the target $\mathcal{S}$ satisfies an internal sphere condition, and T is continuous, and study the hypograph of T in the complement of $\mathcal{S}$. Since the internal sphere property is closed with respect to the union operator, one can see intuitively that the reachable set $\mathcal{R}^t$, which is the set of

points reachable from $\mathcal{S}$ in time less than t, inherits such property from $\mathcal{S}$. By combining this fact and the Hamiltonian function, a regularity result on the hypograph of T can be obtained. The corresponding theorem is as follows:

Theorem 1 *Under the above assumptions, the hypograph of T satisfies an external sphere condition.*

From this theorem, we obtain that if T is Lipschitz then T is semiconcave (see [23]). However, here the situation is more complicated than in the Lipschitz case: the main results depend on the pointedness of the normal cone to the hypograph. Indeed, from a representation of generalized supergradient of T, we prove that

Theorem 2 *Together with the above assumptions, if the normal cone to the hypograph is pointed in the complement of $\mathcal{S}$, then the hypograph of T has positive reach.*

Several counterexamples (see e.g, [22]), though, show that the external sphere condition is in general weaker than positive reach. In particular, in Example 2 in [17], we constructed a minimum time function with a constant dynamics and a $C^{1,1}$ target such that its hypograph satisfies an external sphere condition but has not positive reach everywhere. On the other hand, the pointedness assumption for the normal cone to the hypograph of a continuous function is hard to verify since it is related to the representation formula for its generalized supergradient (this problem is studied in [16]). Therefore, the problem of understanding whether some concavity features are preserved under the external sphere condition appears natural. Our main result reads -essentially- as follows:

Theorem 3 *Let $\Omega \subset \mathbb{R}^N$ be open and let $f : \Omega \longrightarrow \mathbb{R}$ be continuous. Assume that the hypograph of f satisfies the weak external sphere condition. Then there exists a closed set Γ with zero Lebesgue measure such that the hypograph of the restricted function $f_{\Omega\setminus\Gamma}$ has positive reach.*

Consequently, a function satisfying the assumption of the above theorem enjoys several regularity properties inherited by functions whose hypograph has positive reach. Therefore, using Theorem 1 and Theorem 3 the pointedness assumption of the hypograph of T in Theorem 2 is removed and the a.e. twice differentiability of T for a class of nonlinear control system is also obtained.

References

[1] M. Bardi, I. Capuzzo-Dolcetta, "Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations". Birkhäuser, Boston, 1997.

[2] U. Boscain, B. Piccoli, "Optimal syntheses for control systems on 2-D manifolds". Springer, Berlin, 2004.

[3] A. Bressan, B. Piccoli, "Introduction to the mathematical theory of control". AIMS, Spring-

field, 2007.

[4] A. Canino, *On p-convex sets and geodesics.* J. Differential Equations 75 (1988), 118–157.

[5] P. Cannarsa, P. Cardaliaguet, *Perimeter estimates for reachable sets of control systems.* J. Convex Anal. 13 (2006), 253–267.

[6] P. Cannarsa, H. Frankowska, *Interior sphere property of attainable sets and time optimal control problems.* COCV 12 (2006), 350–370.

[7] P. Cannarsa, C. Sinestrari, *Convexity properties of the minimum time function.* Calc. Var. 3 (1995), 273–298.

[8] P. Cannarsa, C. Sinestrari, "Semiconcave functions, Hamilton–Jacobi Equations, and Optimal Control". Birkhäuser, Boston, 2004.

[9] P. Cardaliaguet, *On the regularity of semipermeable surfaces in control theory with application to the optimal exit-time problem.* SIAM J. Control Optim. 35 (1997), 1638–1671.

[10] F. H. Clarke, "Optimization and Nonsmooth Analysis". Classics in Applied Mathematics, 5. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[11] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern and P. R. Wolenski, "Nonsmooth Analysis and Control Theory". Springer, New York, 1998.

[12] F. H. Clarke, R. J. Stern, P. R. Wolenski, *Proximal smoothness and the lower–$\mathcal{C}^2$ property.* J. Convex Anal. 2 (1995), 117–144.

[13] G. Colombo, A. Marigonda, *Differentiability properties for a class of non-convex functions.* Calc. Var. 25 (2006), 1–31.

[14] G. Colombo, A. Marigonda, *Singularities for a class of non-convex sets and functions, and viscosity solutions of some Hamilton-Jacobi equations.* J. Convex Anal. 15 (2008), 105-129.

[15] G. Colombo, A. Marigonda, P. R. Wolenski, *Some new regularity properties for the minimal time function.* Siam J. Control Optim. 44 (2006), 2285–2299.

[16] G. Colombo, A. Marigonda, P. R. Wolenski, *A representation formula for the generalized gradient for a class of non-Lipschitz functions.* Submitted.

[17] G. Colombo, Khai T. Nguyen, *On the structure of the minimum time function.* Siam J. Control Optim. 48 (2010), 4776–4814.

[18] H. Federer, *Curvature Measures.* Trans. Amer. Math. Soc. 93 (1959), 418–491.

[19] H. Federer, "Geometric Measure Theory". Springer, 1969.

[20] Khai T. Nguyen, *Hypographs satisfying an external sphere condition and the regularity of the minimum time function.* Journal of Mathematical Analysis and Applications (JMAA) 372 (2010), 611–628.

[21] Khai T. Nguyen, D. Vittone, *Rectifiability for special singularities of non-Lipschitz functions.* To appear in JCA.

[22] C. Nour, R. J. Stern and J. Takche, *Proximal smoothness and the exterior sphere condition.* J. Convex Anal. 16 (2009), 501–514.

[23] C. Nour, R. J. Stern and J. Takche, *The θ-exterior sphere condition, φ-convexity, and local semiconcavity.* Nonlinear Anal. 73 (2010), 573–589.

[24] R. A. Poliquin and R. T. Rockafellar, *Prox-regular functions in variational analysis.* Trans. Am. Math. Soc. 348 (1996), 1805–1838.

[25] R. A. Poliquin, R. T. Rockafellar and L. Thibault, *Local differentiability of distance functions.* Trans. Amer. Math. Soc. 352 (2000), 5231–5249.

[26] P. R. Wolenski and Z. Yu, *Proximal analysis and the minimal time function.* SIAM J. of Control and Opt. 36 (1998), 1048–1072.

A Simple Model for Financial Indexes with some Applications

Alessandro Andreoli (*)

Abstract. Two major subjects of Mathematical finance are: mathematically modeling the prices of assets and indices, and option pricing. We propose a simple stochastic model for time series which is analytically tractable, easy to simulate and which captures some relevant stylized facts of financial indexes, including *scaling properties*. We show that the model fits the *Dow Jones Industrial Average* time series in the period 1935-2009 with a remarkable accuracy.
Despite its simplicity, the model has several interesting features. The volatility is not constant and displays high peaks. The empirical distribution of the *log-returns* (increments of the logarithm of the index) is non-Gaussian and may exhibit heavy tails. Log-returns corresponding to disjoint time intervals are uncorrelated but not independent: the correlation of their absolute values decays exponentially fast in the distance between the time intervals for large distances, while it has a slower decay for moderate distances. Moreover, the distribution of the log-returns obeys *scaling relations* that are detected on real time series, but are not satisfied by most available models.
Finally, we give a short overview about traditional option pricing.

1 Modeling Financial Indexes

In this section we give a short overview on modeling prices of indices.

The first partially successful mathematical attempt to model stocks and indices is due to Paul Samuelson, in 1965, who proposed to represent their prices with a *Geometric Brownian motion*. In this way, the Brownian Motion models the logarithm of the prices (*log-prices*).

A slight modification of this idea is the Black & Scholes model: the prices S_t of a stock price (or index) follows the dynamics:

$$\mathrm{d}S_t = S_t(r\mathrm{d}t + \sigma\mathrm{d}W_t), \tag{1.1}$$

where σ (the *volatility*) and r (the *interest rate*) are constant, and $(W_t)_{t\geq 0}$ is a standard Brownian motion.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: aandreol@math.unipd.it. Seminar held on 1 December 2010.

Black & Scholes model met an incredible success, both because on a first approssimation has a good agreement with empirical data, and both because it is reall easy do deal with it, theoretically and numerically.

Despite its success, Black & Scholes is not constistent with a number of other *stylized facts*, that are empirically detected in many real time series. Some of these facts are the following:

- the volatility is not constant: in particular, it may have high peaks, that may be interpreted as *shocks* in the market;
- the empirical distribution of the increments $X_{t+h} - X_t$ of the logarithm of the price (the *log-returns*) has tails heavier than Gaussian;
- log-returns corresponding to disjoint time-interval are uncorrelated, but not independent: in fact, the correlation between the absolute values $|X_{t+h} - X_t|$ and $|X_{s+h} - X_s|$ has a slow decay in $|t - s|$, up to moderate values for $|t - s|$. This phenomenon is known as *clustering of volatility*.

In order to have a better fit with real data, many different models have been proposed to describe the volatility and the price process. In discrete-time, autoregressive models such as ARCH, GARCH and generalizations [12, 7, 3, 8] have been widely used. In continuous time, the basic model (1.1) has been modified by letting $\sigma = \sigma_t$ be a stochastic process, often the solution of a stochastic differential equation driven by a general Lévy process. A systematic account of these *stochastic volatility models* can be found in [4]. Continuous time versions of GARCH include the *generalized Ornstein-Uhlenbeck processes* and the COGARCH (GARCH in continuous time) [17, 18].

Other models, whose effectiveness to model real data is the subject of current research, include jumps in the prices and leverage; these models involve several parameters, whose estimation raises a number of interesting statistical issues (see e.g. [9, 2, 15, 10]).

More recently (see [11, 6, 22]), other stylized facts of financial indexes have been pointed out, concerning the *scaling properties* of the empirical distribution of the log-returns. Consider the time series of an index $(s_i)_{1 \le i \le T}$ over a period of $T \gg 1$ days and denote by p_h the *empirical distribution* of the (detrended) log-returns corresponding to an interval of h days:

$$p_h(\cdot) := \frac{1}{T-h} \sum_{i=1}^{T-h} \delta_{x_{i+h} - x_i}(\cdot)\,, \qquad x_i := \log(s_i) - \overline{d}_i\,, \tag{1.2}$$

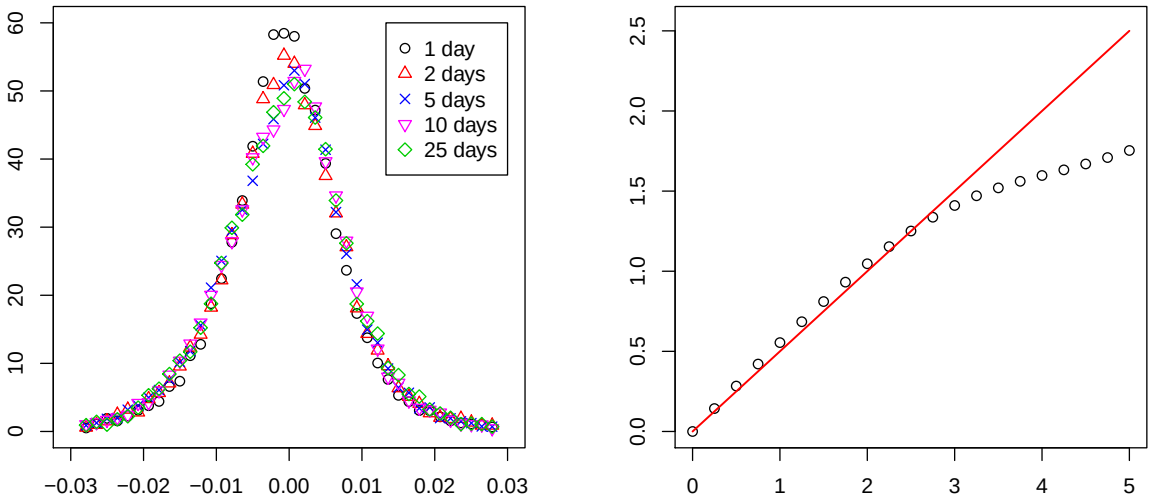
where $\overline{d}_i$ is the local rate of linear growth of $\log(s_i)$ and $\delta_x(\cdot)$ denotes the Dirac measure at $x \in \mathbb{R}$. The statistical analysis of various indexes, such as the *Dow Jones Industrial Average* (DJIA) or the *Nikkei 225*, shows that, for h within a suitable time scale, p_h obeys approximately a diffusive scaling relation (cf. Figure 1(A)):

$$p_h(\mathrm{d}r) \simeq \frac{1}{\sqrt{h}}\, g\!\left(\frac{r}{\sqrt{h}}\right)\mathrm{d}r, \tag{1.3}$$

where g is a probability density with tails heavier than Gaussian. If one considers the q-th empirical moment $m_q(h)$, defined by

$$m_q(h) := \frac{1}{T-h}\sum_{i=1}^{T-h} |x_{i+h} - x_i|^q = \int |r|^q \, p_h(\mathrm{d}r)\,, \tag{1.4}$$

from relation (1.3) it is natural to guess that $m_q(h)$ should scale as $h^{q/2}$. This is indeed what one observes for moments of small order $q \le \bar{q}$ (with $\bar{q} \simeq 3$ for the DJIA). However, for moments of higher order $q > \bar{q}$, the different scaling relation $h^{A(q)}$, with $A(q) < q/2$, takes place, cf. Figure 1(B) (see also [11]). This is the so-called *multiscaling of moments*.



(A) *Diffusive scaling of log-returns.* (B) *Multiscaling of moments.*

Figure 1. *Scaling properties of the DJIA time series (opening prices 1935-2009).*

(A) The empirical densities of the log-returns over 1, 2, 5, 10, 25 days show a remarkable overlap under diffusive scaling.

(B) The scaling exponent $A(q)$ as a function of q, defined by the relation $m_q(h) \approx h^{A(q)}$ (cf. (1.4)), bends down from the Gaussian behavior $q/2$ (red line) for $q \ge \bar{q} \simeq 3$. The quantity $A(q)$ is evaluated empirically through a linear interpolation of $(\log m_q(h))$ versus $(\log h)$ for $h \in \{1, \dots, 5\}$.

In the next section we will define a *simple* continuous-time stochastic model which agrees with *all* mentioned stylized facts. This is a non-trivial point, despite of the variety of models that can be found in the literature. For example, the celebrated and widely used GARCH [3] exhibits non-constant volatility and non-Gaussian distribution of log-returns; the correlation of the absolute values of the log-return is positive and decays exponentially fast, in contrast with empirical evidences indicating a somewhat slower decay; finally,

multiscaling of moments is not present, at least for the range of values of the parameters that most often occur in practice.

Also models very recently proposed, as the one in [9], however extremely accurate to fit the statistics of the empirical volatility and with the multiscaling of moments feature, fail on some other aspects (for example, the one proposed in [9] exhibits a decay of correlations of absolute log-returns that is purely exponential).

2 The Model & The Main Results

In this section we present the model and the main theoretical results we managed to obtained about.

Definition 1 A *point process* τ (on $\mathbb{R}$) is a mapping from a probability space $(\Omega, \mathcal{F}, P)$ to the locally finite subsets of $\mathbb{R}$.

Given two real numbers $D \in (0, 1/2]$, $\lambda \in (0, \infty)$ and a probability ν on $(0, \infty)$ (these may be viewed as our parameters), our model is defined upon the following three sources of alea:

- a standard Brownian motion $W = (W_t)_{t \geq 0}$;
- a Poisson point process $\mathcal{T} = (\tau_\backslash)_{\backslash \in \mathbb{Z}}$ on $\mathbb{R}$ with intensity λ;
- a sequence $\Sigma = (\sigma_n)_{n \geq 0}$ of independent and identically distributed positive random variables. The marginal law of the sequence will be denoted by ν (so that $\sigma_n \sim \nu$ for all n) and for conciseness we denote by σ a variable with the same law ν.

It is worth stressing from now that the first two moments of the law ν, i.e. $E(\sigma)$ and $E(\sigma^2)$, are enough to determine the features of our model that are relevant for real-world times series We assume that $W, \mathcal{T}, \pm$ are defined on some probability space $(\Omega, \mathcal{F}, \mathcal{P})$ and that they are independent. By convention, we label the points of $\mathcal{T}$ so that $\tau_0 < 0 < \tau_1$. We will actually need only the points of $\mathcal{T} \cap [\tau_0, \infty)$, that is the variables $(\tau_n)_{n \geq 0}$. We recall that the random variables $-\tau_0$, τ_1, $(\tau_{n+1} - \tau_n)_{n \geq 1}$ are independent and identically distributed $Exp(\lambda)$, so that $1/\lambda$ is the mean distance between the points in $\mathcal{T}$. Although some of our results would hold for more general distributions of $\mathcal{T}$, we stick for simplicity to the (rather natural) choice of a Poisson process.

We are now ready to define our model $X = (X_t)_{t \geq 0}$. For $t \in [0, \tau_1]$ we set

$$X_t := \sigma_0 \big(W_{(t-\tau_0)^{2D}} - W_{(-\tau_0)^{2D}} \big) , \tag{2.1}$$

while for $t \in [\tau_n, \tau_{n+1}]$ (with $n \geq 1$) we set

$$X_t := X_{\tau_n} + \sigma_n \left(W_{(t-\tau_n)^{2D} + \sum_{k=1}^{n} (\tau_k - \tau_{k-1})^{2D}} - W_{\sum_{k=1}^{n} (\tau_k - \tau_{k-1})^{2D}} \right) . \tag{2.2}$$

In words: at the epochs τ_n the time inhomogeneity $t \mapsto t^{2D}$ is "refreshed" and the volatility is randomly updated: $\sigma_{n-1} \rightsquigarrow \sigma_n$. A possible financial interpretation of this mechanism is

that jumps in the volatility correspond to shocks in the market. The reaction of the market is not homogeneous in time: if $D < 1/2$, the dynamics is fast immediately after the shock, and tends to slow down later, until a new jump occurs. For $D = 1/2$ our model reduces to a simple random volatility model $\mathrm{d}X_t = \sigma_t \, \mathrm{d}W_t$, where $\sigma_t := \sum_{k=0}^{\infty} \sigma_k \, \mathbf{1}_{[\tau_k, \tau_{k+1})}(t)$ is a (random) piecewise constant process.

Using the scale invariance of Brownian motion, we now give an alternative definition of our model X, that is equivalent in law with (2.2) but more convenient for the theoretical tractability of the process. For $t \geq 0$, define

$$i(t) \, := \, \sup\{n \geq 0 : \, \tau_n \leq t\} \, = \, \#\{\mathcal{T} \cap (0, t]\} \, , \tag{2.3}$$

so that $\tau_{i(t)}$ is the location of the last point in $\mathcal{T}$ before t. Now we introduce the process $I = (I_t)_{t \geq 0}$ by

$$I_t \, := \, \sigma_{i(t)}^2 \left(t - \tau_{i(t)}\right)^{2D} \, + \, \sum_{k=1}^{i(t)} \sigma_{k-1}^2 \left(\tau_k - \tau_{k-1}\right)^{2D} \, - \, \sigma_0^2 \left(-\tau_0\right)^{2D} , \tag{2.4}$$

with the agreement that the sum in the right hand side is zero if $i(t) = 0$. We can then redefine our basic process $X = (X_t)_{t \geq 0}$ by setting

$$X_t := W_{I_t} \, . \tag{2.5}$$

Note that I is a strictly increasing process with absolutely continuous paths, and it is independent of the Brownian motion W. Thus this model may be viewed as an independent random time change of a Brownian motion.

We now state our main results concerning the process X. They correspond to the basic stylized facts that we have mentioned in the previous section: diffusive scaling of the distributions of log-returns (Theorem 2 below); multiscaling of moments (Theorem 3 and Corollary 4); clustering of volatility (Theorem 5 and Corollary 6).

The first result states that for small h the increments $(X_{t+h} - X_t)$ have an approximate diffusive scaling, in agreement with (1.3).

Theorem 2 *As $h \downarrow 0$ we have the convergence in distribution*

$$\frac{(X_{t+h} - X_t)}{\sqrt{h}} \xrightarrow[h \downarrow 0]{d} f(x) \, \mathrm{d}x \, , \tag{2.6}$$

where f is a mixture of centered Gaussian densities, namely

$$f(x) = \int_0^{\infty} \nu(\mathrm{d}\sigma) \int_0^{\infty} \mathrm{d}t \, \lambda e^{-\lambda t} \, \frac{t^{1/2-D}}{\sigma \sqrt{4D\pi}} \, \exp\left(-\frac{t^{1-2D} x^2}{4D\sigma^2}\right) . \tag{2.7}$$

We stress that the function f appearing in (2.6)–(2.7), which describes the asymptotic rescaled law of the increment $(X_{t+h} - X_t)$ in the limit of small h, has a different tail behavior

from the density of $(X_{t+h} - X_t)$ for fixed h. For instance, when σ has finite moments of all orders, it follows that the same holds for $(X_{t+h} - X_t)$. However,independently of the law ν of σ, the density f *has always polynomial tails*: $\int_{\mathbb{R}} |x|^q f(x)\,\mathrm{d}x = \infty$ for $q \geq q^*$.

This feature of f has striking consequences on the scaling behavior of the moments of the increments our model. If we set for $q \in (0, \infty)$

$$m_q(h) := E(|X_{t+h} - X_t|^q)\,, \tag{2.8}$$

from the convergence result (2.6) it would be natural to guess that $m_q(h) \approx h^{q/2}$ as $h \downarrow 0$, in analogy with the Brownian motion case. However, this turns out to be true only for $q < q^*$. For $q \geq q^*$, the faster scaling $m_q(h) \approx h^{Dq+1}$ holds instead, the reason being precisely the fact that the q-moment of f is infinite for $q \geq q^*$. This transition in the scaling behavior of $m_q(h)$ goes under the name of *multiscaling of moments* and is discussed in detail, e.g., in [11]. Let us now state our result.

Theorem 3 [Multiscaling of moments] *Let $q > 0$, and assume $E(\sigma^q) < +\infty$. Then the quantity $m_q(h) := E(|X_{t+h} - X_t|^q) = E(|X_h|^q)$ is finite and has the following asymptotic behavior as $h \downarrow 0$:*

$$m_q(h) \sim \begin{cases} C_q\, h^{\frac{q}{2}} & \text{if } q < q^* \\ C_q\, h^{\frac{q}{2}} \log(\frac{1}{h}) & \text{if } q = q^* \\ C_q\, h^{Dq+1} & \text{if } q > q^* \end{cases}, \qquad \text{where } \; q^* := \frac{1}{(\frac{1}{2} - D)}\,.$$

The constant $C_q \in (0, \infty)$ is given by

$$C_q := \begin{cases} E(|W_1|^q)\, E(\sigma^q)\, \lambda^{q/q^*}\, (2D)^{q/2}\, \Gamma(1 - q/q^*) & \text{if } q < q^* \\ E(|W_1|^q)\, E(\sigma^q)\, \lambda\, (2D)^{q/2} & \text{if } q = q^* \\ E(|W_1|^q)\, E(\sigma^q)\, \lambda \left[\int_0^\infty ((1+x)^{2D} - x^{2D})^{\frac{q}{2}}\, \mathrm{d}x \, + \, \frac{1}{Dq+1} \right] & \text{if } q > q^* \end{cases}, \tag{2.9}$$

where $\Gamma(\alpha) := \int_0^\infty x^{\alpha-1} e^{-x} \mathrm{d}x$ denotes Euler's Gamma function.

Corollary 4 *The following relation holds true:*

$$A(q) := \lim_{h \downarrow 0} \frac{\log m_q(h)}{\log h} = \begin{cases} \dfrac{q}{2} & \text{if } q \leq q^* \\ Dq + 1 & \text{if } q \geq q^* \end{cases}. \tag{2.10}$$

Our last theoretical result concerns the correlations of the absolute value of two increments, a quantity which is usually called *volatility autocorrelation*. We start determining the behavior of the covariance.

Theorem 5 *Assume that $E(\sigma^2) < \infty$. The following relation holds as $h \downarrow 0$, for all $s, t > 0$:*

$$Cov(|X_{s+h} - X_s|, |X_{t+h} - X_t|) = \frac{4D}{\pi}\, \lambda^{1-2D}\, e^{-\lambda|t-s|} \left(\phi(\lambda|t-s|)\, h \, + \, o(h) \right), \tag{2.11}$$

where

$$\phi(x) := Cov\big(\sigma\, S^{D-1/2}\,,\, \sigma\,\big(S+x\big)^{D-1/2}\big) \tag{2.12}$$

and $S \sim Exp(1)$ *is independent of* σ*.*

We recall that $\rho(Y,Z) := Cov(Y,Z)/\sqrt{Var(Y)Var(Z)}$ is the correlation coefficient of two random variables Y, Z. As Theorem 3 yields

$$\lim_{h\downarrow 0} \frac{1}{h}\, Var(|X_{t+h} - X_t|) \;=\; (2D)\,\lambda^{1-2D}\, Var(\sigma\,|W_1|\,S^{D-1/2})\,,$$

where $S \sim Exp(1)$ is independent of σ, W_1, we easily obtain the following result.

Corollary 6 [Volatility autocorrelation] *Assume that* $E(\sigma^2) < \infty$*. The correlation of the increments of the process* X *has the following asymptotic behavior as* $h \downarrow 0$*:*

$$\lim_{h\downarrow 0} \rho(|X_{s+h} - X_s|, |X_{t+h} - X_t|) \;=\; \rho(t-s) \;:=\; \frac{2}{\pi\, Var(\sigma\,|W_1|\,S^{D-1/2})}\, e^{-\lambda|t-s|}\, \phi(\lambda|t-s|)\,, \tag{2.13}$$

where $\phi(\cdot)$ *is defined in (2.12) and* $S \sim Exp(1)$ *is independent of* σ *and* W_1*.*

This shows that the volatility autocorrelation of our process decays exponentially fast for time scales greater than the mean distance $1/\lambda$ between the epochs τ_k. For shorter time scales, a relevant contribution is given by the function $\phi(\cdot)$, that decays faster than polynomial but slower than exponential.

So, we have found a model that has the following features:

- it's quite easy to describe.
- it's theoretically and numerically tractable.
- takes into account many features that are charateristic of real time series.

3 Arbitrage Theory & Option Pricing

In this section we present some classical theory on option pricing.

The *general assumptions* for the market $\mathcal{M}$ we consider are:

- Short positions and any fractional (real) holdings, are allowed,
- No bid-ask spread, *i.e.* the selling price is equal to the buying price of all assets,
- There are no transaction costs of trading,
- The market is completely liquid, *i.e.* it is always possible to buy and sell unlimited quantities on the market. In particular it is possible to borrow unlimited amounts from the bank (by selling bonds short),

- there is a *risk-free* asset, *i.e.* an asset with deterministic evolution, $dB_t = r(t)B_t dt$.

Under our assumptions, the market is then made of $N-1$ riskly assets (stochastic processes) and one risk-free asset (deterministic process).

Definition 7 A *Portfolio* is any linear combination of the N assets. A *Self-Financing Portfolio* is a portfolio with no exogenous infusion or withdrawal of money; in other words, the purchase of a new portfolio must be financed solely by selling assets already in the portfolio.

Definition 8 An *Arbitrage* possibility on a financial market is a self-financed portfolio such that:

- Its initial value is zero.
- Its value at time T is greater or equal zero almost surely.
- Its value at time T has strictly positive probability of being strictly greater than zero.

We say that the market is *arbitrage-free* if there are no arbitrage possibilities.

Definition 9 A *Simple claim* with date of maturity T on the underlying asset S is any random variable $\phi(S_T)$ with ϕ measurable. The function ϕ is called the *contract function*.

Example: A (European) *Call Option* with strike price K and maturity time T on the underlying S is a contract of this form:

$$\phi(S_T) = (S_T - K)^+.$$

The aim of option pricing is to give to the options $X = \phi(S_T)$ reasonable prices $\Pi_t(X)$, at any time t. If the market (without these options) is arbitrage-free, there are two main approachs to do this:

- to price the options in such a way that the market remains arbitrage-free once that we add these new prices: in other words we want $B_t, S_t^1, ..., S_t^{N-1}, \Pi_t(X)$ to be arbitrage-free.
- if at time t there exist a self-financing portfolio whose value at time T is identically equal to the value of X, then $\Pi_t(X)$ should be the value of this portfolio at time t (We call *Hedging portfolio* for X a portfolio with this property).

The mathematical problems linked to the two approaches are:

- Characterization for a market to be arbitrage-free.
- Existence and uniqueness of the price.

In the following we will see two theorems that completely solve the option pricing problem, at least for very special markets (namely, for the *complete* markets).

Definition 10 A process $(Z_t)_{t\in T}$ is an $(\mathcal{F}_t)-$*martingale* if

- $(Z_t)_{t\in T}$ is $(\mathcal{F}_t)-$adapted
- $E[|Z_t|] < \infty$ for all t
- $Z_s = E[Z_t|\mathcal{F}_s]$ for all $s \leq t$

Definition 11 Two measures P and Q on the same space $(X, \mathcal{F})$ are *equivalent* if the following condition holds:

$$P(A) = 0 \Leftrightarrow Q(A) = 0,\ A \in \mathcal{F}$$

Definition 12 if P is the "real world" measure of the market model, we say that Q is an *equivalent martingale measure (EMM)* if:

- P and Q are equivalent
- the discounted price processes $\frac{S_t}{B_t}$ of the $N-1$ assets are martingales under the measure Q.

We now presents a theorem that gives a way for pricing the options according to the first approach we proposed:

Theorem 13 [The First Foundamental Theorem of Finance] *The market model (with "real world" probability P) is arbitrage-free if and only if there exists (at least) an equivalent martingale measure Q.*

As a consequence of the First Foundamentak Theorem, we have the following. If

$$\mathcal{M} = \left\{B_t, S_t^1, ..., S_t^{N-1}\right\}$$

is an arbitrage-free market, then ther exists at least one Equivalent Martingale Measure Q; if we set

$$\Pi_t(X) = B_t\, E^Q\left[\frac{X}{B_T}|\mathcal{F}\right],$$

by definition the discounted price of X is a martingale under Q, and then

$$\mathcal{M}' = \left\{B_t, S_t^1, ..., S_t^{N-1}, \Pi_t(X)\right\}$$

is an arbitrage-free market too, again by the First Foundamental Thoerem.

Remark 14 The First foundamental theorem assures that *at least* one EMM exists, but there could be *more than one*, and then many possible prices for X.

Theorem 13 gives then a partial answer to the problem of otpion pricing: we know what prices we can give to X consistently with the market, but we have to many of them! Let us now try using the second approach.

First of all, note that if there exists an hedging portfolio for X, his value should be the price of X (otherwise there would be an arbitrage.); but, for the same reason, *all the hedging portfolios for X should have the same value.* Then, if X is hedgeable, all the $E.M.M$ should give the same price for it, so X has an *unique* price. This leads to the definition of *completeness* for markets and to the Second Foundamental Theorem of Finance:

Definition 15 A market-model is *complete* if every option X is headgeable.

Theorem 16 [The Second Fundamental Theorem of Finance] *Assuming absence of arbitrage, a market model is complete if and only if the martingale measure Q is unique.*

Remark 17 From the First and the Second Fundamental Theorems, it follows immediately that for a market-model *arbitrage-free and complete* there exists an *unique* non-arbitrage price for all options.

Then, the problem of Option Pricing in complete (and arbitrage free) markets is theoretically solved. What about the incomplete ones?

Many solutions have been proposed in literature:

- *SuperHedging*,
- to choose the EMM that best suits the prices already present on the market,
- other

But this is still an open problem.

References

[1] L. Accardi, Y. G. Lu, *A continuous version of de Finetti's theorem.* Ann. Probab. 21 (1993), 1478–1493.

[2] Y. Aït-Sahalia, J. Jacod, *Testing for jumps in discretely observed processes.* Ann. Statistics 37 (2009), 184-222.

[3] R. T. Baillie, *Long memory processes and fractional integration in econometrics.* J. Econometrics 73 (1996), 5–59.

[4] O. E. Barndorff-Nielsen, N. Shephard, *Non Gaussian Ornstein-Uhlenbeck based models and some of their uses in financial economics.* J. R. Statist. Soc. B 63 (2001), 167–241.

[5] P. Billingsley, "Probability and Measure". Third Edition, John Wiley and Sons,1995.

[6] F. Baldovin, A. Stella, *Scaling and efficiency determine the irreversible evolution of a market.* PNAS 104, n. 50 (2007), 19741–19744.

[7] T. Bollerslev, *Generalized Autoregressive Conditional Heteroskedasticity.* J. Econometrics 31 (1986), 307–327.

[8] T. Bollerslev, H.Ø. Mikkelsen, *Modeling and pricing long memory in stock market volatility.* J. Econometrics 31 (1996), 151–184.

[9] T. Bollerslev, U. Kretschmer, C. Pigorsch, G. Tauchen, *A discrete-time model for daily S & P500 returns and realized variations: Jumps and leverage effects.* J. Econometrics 150 (2009), 151–166.

[10] T. Bollerslev, V. Todorov, *Jump Tails, Extreme Dependencies, and the Distribution of Stock Returns.* CREATES Research Paper 2010-64 (2010).

[11] T. Di Matteo, T. Aste, M. M. Dacorogna, *Long-term memories of developed and emerging markets: Using the scaling analysis to characterize their stage of development.* J. Banking Finance 29 (2005), 827–851.

[12] R. F. Engle, *Autoregressive Conditional Heteroscedasticity with Estimates of Variance of United Kingdom Inflation.* Econometrica 50 (1982), 987–1008.

[13] D. A. Freedman, *Invariants Under Mixing Which Generalize de Finetti's Theorem: Continuous Time Parameter.* Ann. Math. Statist. 34 (1963), 1194–1216.

[14] J. C. Hull, "Options, Futures and Other Derivatives". Pearson/Prentice Hall, 2009.

[15] J. Jacod, V. Todorov, *Testing for common arrivals of jumps for discretely observed multidimensional processes.* Ann. Statistics 37 (2009), 1792-1838.

[16] I. Karatzas, S. E. Shreve, "Brownian Motion and Stochastic Calculus". Springer, 1988.

[17] C. Kluppelberg, A. Lindner, R. A. Maller, *A continuous time GARCH process driven by a Lévy process: stationarity and second order behaviour.* J. Appl. Probab. 41 (2004) 601–622.

[18] C. Kluppelberg, A. Lindner, R. A. Maller, *Continuous time volatility modelling: COGARCH versus Ornstein-Uhlenbeck models.* In *From Stochastic Calculus to Mathematical Finance*, Yu. Kabanov, R. Lipster and J. Stoyanov (Eds.), Springer (2007).

[19] Wolfram Research, Inc., *Mathematica.* Version 7.0, Champaign, IL, 2008.

[20] B. Øksendal, "Stochastic Differential Equations". Springer-Verlag, 2003.

[21] R Development Core Team (2009), "R: A language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL: `http://www.R-project.org`.

[22] A. L. Stella, F. Baldovin, *Role of scaling in the statistical modeling of finance.* Pramana 71 (2008), 341–352.

[23] L. Weiss, *The Stochastic Convergence of a Function of Sample Successive Differences.* Ann. Math. Statist. 26 (1955), 532–536.

An introduction to Coxeter group theory

Mario Marietti (*)

Abstract. Coxeter groups arise in many parts of algebra, combinatorics and geometry, providing connections between different areas of mathematics. The purpose of this talk is to give an elementary overview to Coxeter group theory from algebraic, combinatorial and geometrical viewpoints. Some classical and more recent results will be presented.

1 Overview

Coxeter group theory derives much of its appeal from its interactions with several areas of mathematics such as algebra, combinatorics, and geometry. In Coxeter group theory, a crucial role is played by Bruhat order. It arises not only in the Bruhat decomposition (this motivates the terminology although it would be more appropriate to call it Chevalley order), but also in many other contexts such as in connection with inclusions among Schubert varieties, with the Verma modules of a complex semisimple Lie algebra, and in Kazhdan–Lusztig theory.

These notes mirror the content and the spirit of the talk. First, I give an elementary introduction to the subject that can be accessible to a public of non specialist people, referring who is interested to the classical references [2], [3], [8], [10]. Then, I present some more recent research results, preserving understandability. More precisely, after giving some motivating examples, I give the definition and the basic properties of Coxeter systems and Bruhat order. Then I consider special matchings and some of the results in the theory that flows from these recently introduced combinatorial objects, which have several applications (see [4], [5], [12]). In particular, I give a combinatorial characterization of Coxeter groups (partially ordered by Bruhat order) among all posets.

2 Motivating examples

We begin by briefly giving some motivating basic examples, which should be kept in mind during the talk.

(*)Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: marietti@mat.uniroma1.it. Seminar held on 15 December 2010.

2.1 Dihedral groups

Let V be the euclidean plane, P_m be a regular m-sided polygon centered at the origin, and $I_2(m)$ be the group of orthogonal transformations preserving P_m. The group $I_2(m)$, which is called *dihedral*, has order $2m$ and, more precisely, contains:

- m rotations (through multiples of $2\pi/m$),
- m reflections.

The dihedral group $I_2(m)$ is generated by two adjacent reflections s_1 and s_2 whose products s_1s_2 and s_2s_1 have order m. For example, $I_2(6)$ is generated by s_1 and s_2, the reflections through the lines in Figure 1, and the products s_1s_2 and s_2s_1 are rotations through $2\pi/6$ and therefore have order 6.

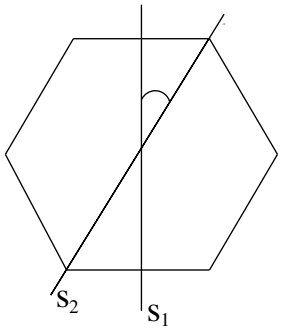


Figure 1: $I_2(6)$ is generated by s_1 and s_2.

2.2 Reflection Groups

Let V be a (real) euclidean space with a positive definite symmetric bilinear form $< -, - >$. An *orthogonal reflection* s_α is a linear trasformation sending a vector $\alpha \neq 0$ to its negative and fixing pointwise the hyperplane orthogonal to α. It has the following expression:

$$s_\alpha(v) = v - \frac{2 < v, \alpha >}{< \alpha, \alpha >}\alpha$$

and it is immediate to verify that it is an orthogonal transformation.

A group generated by orthogonal reflections is called a *reflection group*. Hence a finite reflection group is obtained by a set Φ of nonzero vectors in an euclidean space V whose associated reflections generate a finite group. For example, we obtain a finite reflection group taking as Φ any root system (in the sense of Lie theory).

2.3 Weyl groups

Let

- G be a reductive algebraic group (for example, GL_n, the group of $n \times n$ invertible matrices),

- B be a Borel subgroup of G (for example, the subgroup of upper triangular matrices of GL_n),
- T be a maximal torus of B (for example, the subgroup of diagonal matrices of B),
- $N(T)$ be the normalizer of T in G.

Definition 2.1 The *Weyl group of* G is the group $W = N(T)/T$.

The Weyl group of G is generated by involutions (elements of order 2).

3 Coxeter groups

3.1 Coxeter systems

As a matter of fact, it is a popular abuse of language to use the term Coxeter group instead of the more appropriate Coxeter system: the subject of study is not merely a group but a group with attached a distinguished set of generators which play a fundamental role in the theory. A *Coxeter system* is a pair (W, S) where

- $S = \{s_1, \dots, s_n\}$ is a finite set,
- W is the group generated by S with relations only of the form $(s_i s_j)^{m_{ij}} = id$, with $m_{ii} = 1$ and $m_{ij} = m_{ji} \in \mathbb{N}_{\geq 2} \cup \{\infty\}$ if $i \neq j$.

We make the convention that no relation occurs for the pair (s_i, s_j) if $m_{ij} = \infty$. The matrix (m_{ij}) and the cardinality of S are, respectively, the *Coxeter matrix* and the *rank* of the Coxeter system (W, S). Since, for all $i = 1, \dots, n$, $m_{ii} = 1$, we have that every generator is an involution.

All pieces of information are encoded in the *Coxeter graph* of the Coxeter system (W, S). This is the labeled graph having S as the set of vertices and where $\{s_i, s_j\}$ is an edge labeled m_{ij} if and only if $m_{ij} > 2$ (labels 3 are omitted). Notice that the generators s_i and s_j commute if $m_{ij} = 2$. For example, if (W, S) is the Coxeter system with Coxeter matrix

$$\begin{pmatrix} 1 & 6 & 3 & 2 \\ 6 & 1 & 4 & 2 \\ 3 & 4 & 1 & 9 \\ 2 & 2 & 9 & 1 \end{pmatrix}$$

the Coxeter graph of (W, S) is the one depicted in Figure 2.

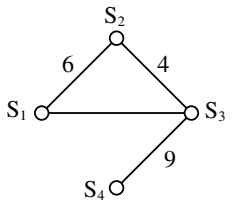


Figure 2: The Coxeter graph of (W, S).

Coxeter groups with connected Coxeter graphs are called *irreducible*. A non irreducible Coxeter group is the direct product of the Coxeter groups associated to its connected components and hence we may often reduce to the case of irreducible Coxeter systems.

3.2 Word problem

By definition, $W = F/N$, where F is the free group generated by S and N is the normal subgroup of F generated by the relations. Thus elements in W are equivalence classes of words in the alphabet S (the inverses of the generators are not needed since $s^{-1} = s$, for all $s \in S$). The problem of telling when two words represent the same element or not has been uniformly solved, for all Coxeter systems, by Tits (see [14]). The answer is the simplest one may expect: one can always transform an arbitrary word to an equivalent one by making only the most obvious types of modifications (the modifications given by the defining relations).

Theorem 3.1 [Tits' word Theorem] *Two words represent the same element if and only if they are linked by a sequence of moves of the following types:*

- *deleting a pair* $s_i s_i$ *(*nil move*),*
- *inserting a pair* $s_i s_i$*,*
- *replacing* $\underbrace{s_i s_j s_i \ldots}_{m_{ij} \text{ letters}}$ *by* $\underbrace{s_j s_i s_j \ldots}_{m_{ij} \text{ letters}}$ *(*braid move*).*

Example 3.2 Let (W, S) be the Coxeter system whose Coxeter graph is given in Figure 2. Consider the two expressions $s_3 s_1 s_3 s_4 s_1$ and $s_1 s_3 s_4$. They represent the same element since

$$\underbrace{s_3 s_1 s_3} s_4 s_1 \overset{\text{braid mv}}{=} s_1 s_3 s_1 \underbrace{s_4 s_1} \overset{\text{braid mv}}{=} s_1 s_3 \underbrace{s_1 s_1} s_4 \overset{\text{nil mv}}{=} s_1 s_3 s_4$$

3.3 Properties of Coxeter groups

Let (W, S) be a Coxeter system. Given an element $w \in W$, let

$$\ell(w) = \min\{k : w \text{ is a product of } k \text{ generators}\}$$

be the *length of* w. Any word that represents w with exactly $\ell(w)$ geneators is a *reduced expression* of w. Other statistics on W are the right and left descens, which are defined as follows:

- $D_R(w) = \{s \in S : \ell(ws) < \ell(w)\}$ *right descents of* w
- $D_L(w) = \{s \in S : \ell(sw) < \ell(w)\}$ *left descents of* w

The main reason why Coxeter groups have remarkable combinatorial properties is the fact that they satisfy the Exhange Property and, as a consequence, the Deletion Property.

Exchange Property Suppose $w = s_1 s_2 \cdots s_k$, $s_j \in S$, and $s \in S$. If $\ell(ws) < \ell(w)$ then $ws = s_1 \cdots s_{i-1}\, \hat{s_i}\, s_{i+1} \cdots s_k$ for some i.

Deletion Property Suppose $w = s_1 s_2 \cdots s_k$, $s_i \in S$. If $\ell(w) < k$ then it holds $w = s_1 \cdots s_{i-1}\, \hat{s_i}\, s_{i+1} \cdots s_{j-1}\, \hat{s_j}\, s_{j+1} \cdots s_k$ for some i, j.

4 Geometric representation

We cannot expect a faithful representation of an arbitrary Coxeter group W as a group generated by orthogonal reflections. We can do it if we consider (not necessarily orthogonal) reflections. A *reflection* s_α of a vector space V is a linear trasformation sending a vector $\alpha \neq 0$ to its negative and fixing pointwise a hyperplane.

Given a Coxeter system (W, S) (where $S = \{s_1, \ldots, s_n\}$), we can construction a faithful representation of W as a group generated by reflections in the following way. Let $V = \mathbb{R}^{|S|}$ with a basis $\alpha_{s_1}, \ldots, \alpha_{s_n}$ which is in bijection with S, and consider the following reflections:

$$\sigma(s_i)(x) := x - 2(x, \alpha_{s_i})\alpha_{s_i}$$

where $(-,-)$ is the symmetric bilinear form defined by

$$(\alpha_{s_i}, \alpha_{s_j}) = \begin{cases} -\cos(\frac{\pi}{m_{ij}}), & \text{if } m_{ij} < \infty \\ -1, & \text{if } m_{ij} = \infty. \end{cases}$$

Then the group W is isomorphic to the discrete subgroup of $GL(V)$ generated by the $\sigma(s)$. It is a quick calculation to verify that it preserves $(-,-)$.

From standards facts about group representations, it follows the following result.

Theorem 4.1 *Let (W, S) be a Coxeter system and $(-,-)$ be is the symmetric bilinear form defined above. Then the following conditions are equivalent:*

(a) *W is finite,*

(b) *W is a finite reflection group,*

(c) *$(-,-)$ is positive definite.*

By the previous theorem, finite Coxeter groups are precisely the finite reflection groups, which are classified. The finite irreducible Coxeter systems can be divided into four infinite classes, those of type A_n, B_n, D_n, $I_2(m)$, and six sporadic groups, those of type E_6, E_7, E_8, F_4, H_3, H_4 (the subscript is the rank). See, for example, [2] or [10] for the definitions.

5 Presentation of Coxeter systems of type A: the symmetric groups

As we have seen, the geometric representation is an important tool in the theory. However, it is often useful to have a more concrete presentation. We now discuss such presentations for the Coxeter systems of type A: Coxeter systems of type B and D have analogous presentations.

Let (W, S) be the Coxeter system of type A_n, hence $S = \{s_1, \dots, s_n\}$ and the Coxeter matrix is given by $m_{ij} = \begin{cases} 1, & \text{if } i = j; \\ 3, & \text{if } i = j \pm 1; \\ 2, & \text{else.} \end{cases}$ The group W is isomorphic to the symmetric group S_{n+1} through the map $s_i \longmapsto (i, i+1)$, where $(i, i+1)$ denotes the simple transposition switching i and $i+1$ and fixing all the other j. By abuse of language, it is customary to refer to W as the symmetric group S_{n+1} (having always in mind the distinguished set of generators and the map). With this presentation, the statistics we introduced read as follows. If $\pi \in S_n$, then

- $\ell(\pi) = inv(\pi) := |\{(i, j) : i < j, \pi(i) > \pi(j)\}|$ *length= # inversions*
- $D_R(\pi) = \{i : \pi(i) > \pi(i+1)\}$ *right descents*
- $D_L(\pi) = \{i : \pi^{-1}(i) > \pi^{-1}(i+1)\}$ *left descents*

Example 5.1 $\pi = 31265487 \in S_8$

- $\ell(\pi) = inv(\pi) = |\{(1,2), (1,3), (4,5), (4,6), (5,6), (7,8)\}| = 6$
- $D_R(\pi) = \{1, 4, 5, 7\}$
- $D_L(\pi) = \{2, 4, 5, 7\}$

6 Bruhat order

6.1 Properties of Bruhat order

In Coxeter group theory, a crucial role is played by Bruhat order. The remarkable aspects of Bruhat order make the theory appealing for combinatorialists.

Let (W, S) be a Coxeter system and $T = \cup_{w \in W} wSw^{-1}$ be its set of reflections. There are several equivalent definitions of the Bruhat order. For example, the Bruhat order can be defined as the transitive closure of the covering relation $\lhd$:

$$u \lhd v \Leftrightarrow \begin{cases} v = tu, \ t \in T \\ \ell(v) = \ell(u) + 1 \end{cases}$$

A Coxeter group W partially ordered by Bruhat order has a rich combinatorial structure. We now list a few of its properties:

(a) W is ranked with the length ℓ as rank function,

(b) W has a bottom element, which is always the identity,

(c) if it is finite, W has a maximum, which is usually denoted by w_0,

(d) W is Eulerian, meaning that, for all $u \leq v$, its Möbius function μ satisfies $\mu(u,v) = (-1)^{\ell(v)-\ell(u)}$ or, equivalently, $|\{z \in [u,v] : \ell(z) \text{ is even}\}| = |\{z \in [u,v] : \ell(z) \text{ is odd}\}|$,

(e) W admits BGG-labelings (see [12] for the definition).

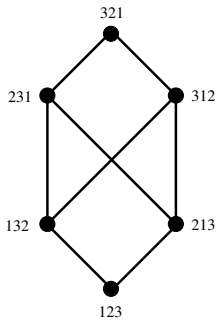


Figure 3: The symmetric group S_3.

6.2 An occurance of the Bruhat order

The Bruhat order has an algebraic-geometric origin. Let $G = GL_n(\mathbb{C})$, B the subgroup of invertible upper triangular matrices, T the subgroup of invertible diagonal matrices, W the *Weyl group of* G, which is isomorphic to the symmetric group S_n (recall that the Weyl group is the group $N(T)/T$, where $N(T)$ is the normalizer of T in G). The quotient G/B is an irreducible projective variety, the *flag variety*: it is in bijection with the set $\{V_0 \subset V_1 \subset \cdots \subset V_n : V_i \text{ subspace of } \mathbb{C}^n \text{ with } \dim V_i = i\}$ whose points are called flags. By the Bruhat decomposition $G = \sqcup_{w \in W} BwB$, we have an induced decomposition $G/B = \sqcup_{w \in W} \Omega_w$, where $\Omega_w := BwB/B$ are affine spaces indexed by the elements $w \in W$ which are called *Schubert cells*. The Schubert cell Ω_w has dimension $\ell(w)$. The Zariski closure $X_w := \overline{\Omega_w}$ of the Schubert cell indexed by w is the *Schubert variety* indexed by w. Since Ω_w is B-invariant, also X_w is B-invariant, and hence it is a union of Schubert cells. The Bruhat order determines which are these cells.

Theorem 6.1 *The Bruhat order determines the inclusions of Schubert varieties:*

$$X_u \subseteq X_v \Leftrightarrow u \leq v.$$

Equivalently,

$$X_w = \sqcup_{u \leq w} \Omega_u = \cup_{u \leq w} X_u$$

(the first one is a disjoint union).

6.3 Bruhat order for the symmetric group: tableau criterion

For the symmetric group S_n, the Bruhat order reads as follows. Let $\sigma = \sigma_1\sigma_2\cdots\sigma_n$ and $\pi = \pi_1\pi_2\cdots\pi_n$ be two permutations. Fix i and take the increasing rearrangements of the first i numbers of σ and π. Then $\sigma \leq \pi$ in the Bruhat order if and only if the sequence associated to σ is component-wise smaller than the sequence associated to π, for all i.

Example 6.2 Let $\sigma = 21453$ and $\pi = 53412$. We have $\sigma \leq \pi$ since

σ :

1	2	3	4	5
1	2	4	5	
1	2	4		
1	2			
2				

$\leq$ π :

1	2	3	4	5
1	3	4	5	
3	4	5		
3	5			
5				

Handy presentations are usuful also for Bruhat order.

7 Special matchings

7.1 Special matchings

Recall that a matching of a graph $G = (V, E)$ is an involution $M : V \to V$ such that $\{M(v), v\} \in E$, for all $v \in V$. Let P be a partially ordered set. A matching M of the Hasse diagram of P is a *s*pecial matching of P if

$$u \lhd v \Longrightarrow M(u) \leq M(v),$$

for all $u, v \in P$ such that $M(u) \neq v$.

Figure 4 gives two matchings of the Boolean algebra of rank 3, the first of which is special while the second is not.

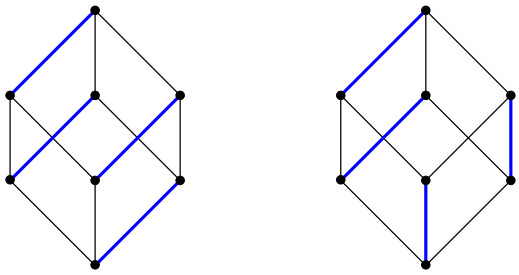


Figure 4: Two matchings.

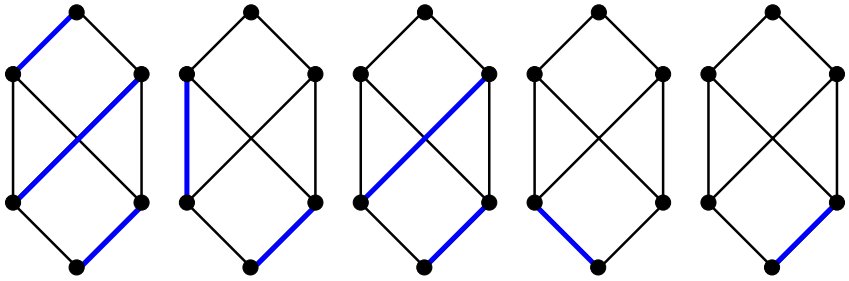
Applications of special matchings are also in Kazhdan–Lusztig theory (see the works of Brenti, Caselli, Delanoy, Du Cloux, M,).

7.2 Zircons

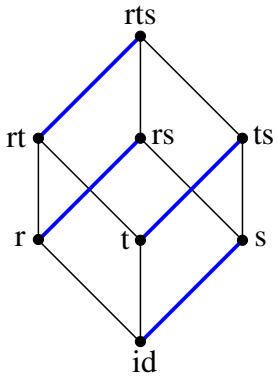
The following definition was given in [11]-[12] (see [9] for an equivalent definition).

Definition 7.1 A locally finite ranked poset Z with bottom element $\hat{0}$ is a *zircon* if, $\forall z \in Z \setminus \{\hat{0}\}$, the interval $[\hat{0}, z]$ admits a special matching.

Example 7.2 S_3 is a zircon.



All Coxeter groups partially ordered by Bruhat order are connected zircons. In fact, let (W, S) be any Coxeter system. Then W is a locally finite ranked poset with the length function as rank function. Fix $w \in W \setminus \{e\}$ and $s \in D_R(w)$. Then the involution $\rho_s : [e, w] \to [e, w]$ defined by $\rho_s(u) = us$ for all $u \in [e, w]$ is a special matching of w. Similarly, if $s \in D_L(w)$, the involution $\lambda_s : [e, w] \to [e, w]$ defined by $\lambda_s(u) = su$ for all $u \in [e, w]$ is a special matching of w. Hence, for every element w in W with $\ell(w) > 1$, there exist at least 2 special matchings of $[id, w]$.



On the other hand, there are "many" zircons that are not Coxeter groups. For example, the zircon in Figure 5 cannot be an interval of the type $[e, w]$ in a Coxeter group since it admits only one special matching.

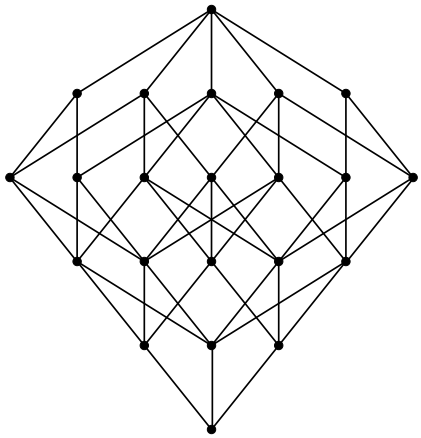


Figure 5: A zircon with only one special matching.

7.3 Properties of zircons

Zircons behave like Coxeter groups: many of the properties of the Coxeter groups extend to zircons. It is often the case that the proofs for zircons are simpler than the corresponding proofs for Coxeter groups: in particular, the proof of part a. of the following result, as far as we know, is the shortest among the many different arguments which prove the Eulerianity of Coxeter groups and part b., which generalizes the fact that Coxeter groups admit BGG labelings, has an elementary proof using special matchings (see [12]).

Theorem 7.3 [M] *Let Z be a zircon. Then*

(a) *Z is Eulerian: the Möbius function of Z is*

$$\mu(u, v) = (-1)^{\rho(v)-\rho(u)}$$

for all $u \leq v \in Z$.

(b) *BGG-labelings of Z are in bijection with the subsets of $Z \setminus \hat{0}$.*

Eulerianity for Coxeter groups was first conjectured [15] and later proved [16] by Verma. Other arguments come from the shellability (Björner-Wachs), Kazhdan-Lusztig theory (Kazhdan-Lusztig), equalities in the 0-Hecke algebra (Stembridge). Bernstein-Gelfand-Gelfand [1] proved that every finite Coxeter group admits a BGG-labeling.

8 Characterizations

The problem of characterizing Coxeter groups among the groups generated by involutions and the problem of characterizing the Bruhat order among the partial orders on a fixed Coxeter group have been solved by Matsumoto [13] and Deodhar [6]-[7], respectively.

Theorem 8.1 [Matsumoto] *Let W be a group and S be a set of generators of order 2. Then the following are equivalent.*

(a) (W, S) *is a Coxeter system.*

(b) (W, S) *has the Exchange Property.*

(c) (W, S) *has the Deletion Property.*

Theorem 8.2 [Deodhar's Subword Property] *Let (W, S) be a Coxeter system and $\leq$ be a partial order on W. Then the following are equivalent.*

(a) *The order $\leq$ is the Bruhat order.*

(b) *We have $u \leq v$ if and only if some reduced expression for v has a subword which is a reduced expression for u.*

(c) *We have $u \leq v$ if and only if every reduced expression for v has a subword which is a reduced expression for u.*

We now give a characterization of Coxeter groups partially ordered by Bruhat order among all posets. In other words, we give a necessary and sufficient condition for an abstract poset to be isomorphic to a Coxeter group partially ordered by Bruhat order. This result is proved by studying the combinatorics of words in the alphabet of special matchings and, in particular, giving a combinatorial version of Tits' Word Theorem. As a matter of fact, the special matchings of a zircon play the role that Coxeter generators play in Coxeter group theory (see [12]).

Theorem 8.3 [M] *Let Z be a ranked poset with a bottom element $\hat{0}$. The following are equivalent.*

(a) *There exists a Coxeter group which, under Bruhat order, is isomorphic to Z.*

(b) *Z is a zircon having a set $\mathcal{R}$ of special matchings such that:*

- *for all $z \in Z \setminus \{\hat{0}\}$, there exists $M \in \mathcal{R}$ such that $M(z) \lhd z$,*
- *given $M, M' \in \mathcal{R}$, all orbits under the action of the group generated by M and M' have the same cardinality (possibly ∞).*

If all orbits under the action of the group generated by two special matchings M and M' have the same cardinality, we denote this cardinality by $2\ m(M, M')$. As a matter of fact, the proof of the previous result is constructive. If the two equivalent conditions of Theorem 8.3 are satisfied, the Coxeter group W has $\mathcal{R}$ as set of Coxeter generators and the integers $m(M, M')$ as Coxeter matrix.

For example, the poset in Figure 6 is isomorphic to the symmetric group S_4 (the application of Theorem 8.3 is given by picture).

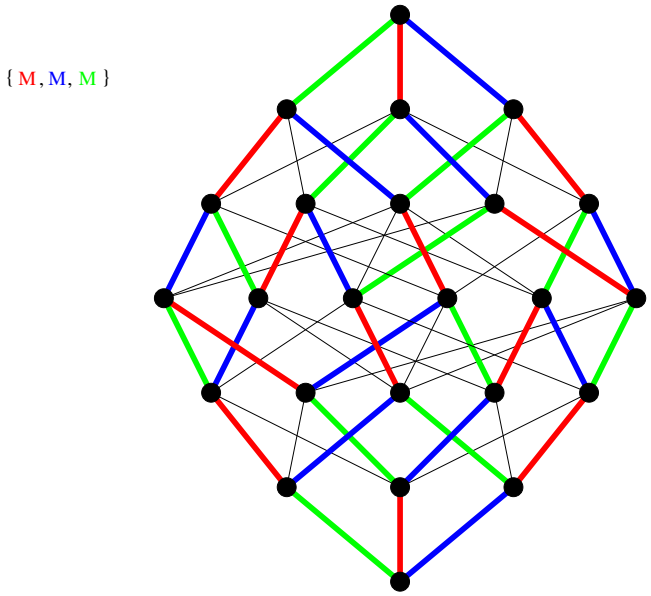
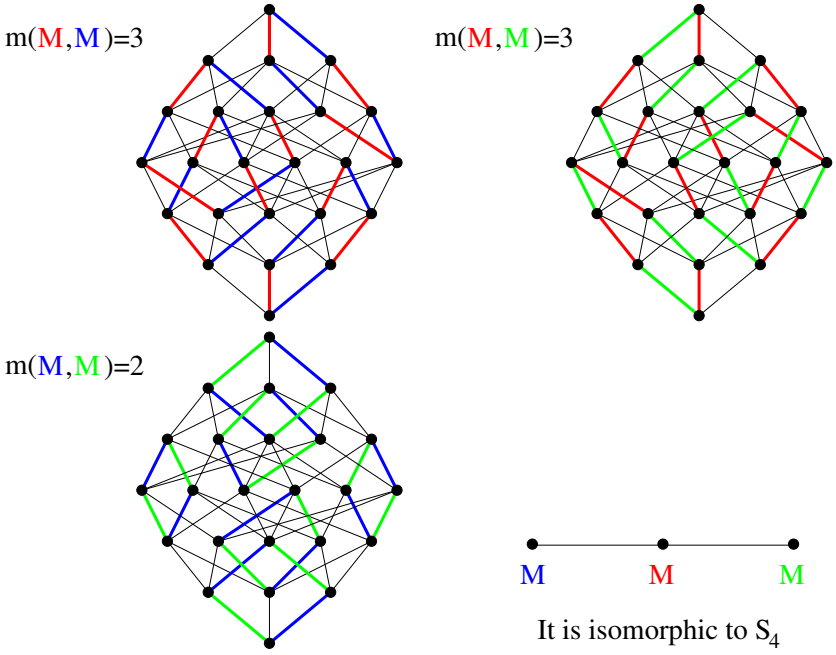


Figure 6: A zircon and three of its special matchings.



References

[1] I. N. Bernstein, I. M. Gelfand, S. I. Gelfand, *Differential operators on the base affine space and a study of* $\mathfrak{g}$*-modules*. In: Lie Groups and their Representations, Summer School of the Bolyai János Math. Soc., Halsted Press (1975), 21–64.

[2] A. Björner and F. Brenti, "Combinatorics of Coxeter groups". Graduate Text in Mathematics, no. 231, Springer, New York, 2005.

[3] N. Bourbaki, "Groupes et Algèbres de Lie". Ch. 4-6, Hermann, Paris, 1968.

[4] F. Brenti, F. Caselli and M. Marietti, *Special matchings and Kazhdan-Lusztig polynomials*. Advances in Math. 202 (2006), 555–601.

[5] F. Brenti, F. Caselli and M. Marietti, *Diamonds and Hecke algebra representations*. Int. Math. Res. Not. 34 (2006), article ID 29407.

[6] V. V. Deodhar, *Some characterizations of Bruhat ordering on a Coxeter group and determination of the relative Möbius function*. Invent. Math. 39 (1977), 187–198.

[7] V. V. Deodhar, *Some characterizations of Coxeter groups*. Enseign. Math. 32 (1986), 111–120.

[8] H. Hiller, "Geometry of Coxeter Groups". Research Notes in Mathematics, no. 54, Pitman Advanced Publishing Programm, 1982.

[9] A. Hultman, *Fixed points of zircon automorphism*. Order 25 (2008), no. 2, 85–90.

[10] J. E. Humphreys, "Reflection Groups and Coxeter Groups". Cambridge Studies in Advanced Mathematics, no. 29, Cambridge Univ. Press, Cambridge, 1990.

[11] M. Marietti, "Kazhdan-Lusztig theory: Boolean elements, special matchings and combinatorial invariance". Ph.D. Thesis, Università degli Studi di Roma "La Sapienza", Italy, 2003.

[12] M. Marietti, *Algebraic and combinatorial properties of zircons*. J. Algebraic Combin. 26 (2007), no. 3, 363–382.

[13] H. Matsumoto, *Générateurs et relations des groupes de Weyl généralisés*. C. R. Acad. Sci. Paris 258 (1964), 3419–3422.

[14] J. Tits, *Le problème des mots dans les groupes de Coxeter*. Symposia Mathematica, vol. 1, INDAM, Roma, 1969, 175–185.

[15] D.-N. Verma, *Structure of certain induced representations of complex semisimple Lie algebras*. Bull. Am. Math. Soc. 74 (1968), 160–166.

[16] D.-N. Verma, *Möbius inversion for the Bruhat order on a Weyl group*. Ann. Sci. École Norm. Sup. 4 (1971), 393–398.

The maximum matching problem and one of its generalizations

Yuri Faenza (*)

Abstract. Given a graph $G(V, E)$, a matching M is a subset of E such that each vertex in V appears as the endpoint of at most one edge from M. The maximum matching problem is among the most important and studied problems in combinatorial optimization. In this short note, we survey a number of classical results on the topic and present more recent results for a non-trivial generalization of the maximum weighted matching problem, i.e. the maximum weighted stable set problem in claw-free graphs.

1 Introduction

Matching problems are among the oldest and most studied in combinatorial optimization and polyhedral combinatorics. They have been the subject of a large number of important studies, whose results and techniques often extended well beyond matching and turned out to be of more general interest. The corpus of results on matching has now reached significative size, deepness, and is often enriched by elegant proofs. Despite these fairly old roots, the matching's well has not run dry, and in recent years several important results and intriguing open questions arose from matching problems or their generalizations.

In this short note, we present some results on the classical *maximum matching problem* and some of its generalizations, ascending a hierarchy of increasing complexity, up to the *maximum weighted stable set problem in claw-free graphs*. In particular, we focus on "theoretically fast" algorithms for the problems above, i.e. whose running time is bounded by a fixed polynomial of the size of the vertex and edge sets of the input graph. As we shall see, a number of structural results on graphs and matchings will be of great help for deriving those algorithms. For the amount of space is limited, we shall not deal with the huge amount of results on matching that are not related to the problems mentioned above, and even on those, we shall cover only a subset of the results from the literature: an interested reader may refer, among others, to the classical text [14], to the more recent book [21], or to the survey [9]. In order to formally state the problems we deal with, we start with some definitions.

(*)Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: faenza@math.unipd.it . Seminar held on 19 January 2011.

2 Definitions

Let $G(V, E)$ be a (undirected) graph, where V denotes the set of *vertices* of G and E is a set of unordered pairs of vertices from V. The elements of E are called *edges*. For the sake of shortness, we shall denote an edge $\{u, v\}$ by uv. Two vertices $u, v \in V$ such that $uv \in E$ are called *adjacent* (in G), and u, v are the *endpoints* of uv. The graphs we deal with are simple (repetitions in E are not allowed), loopless (no vv edge belongs to E, for any vertex $v \in V$), finite (V is a finite set, and consequently so is E). A *matching* of G is a set $M \subseteq E$ such that each vertex of V appears in at most one edge from M. Given a graph G and a matching M of G, a vertex $v \in V$ is called $M-$*covered* if there exist $u \in V$ such that $uv \in M$; $M-$*exposed* otherwise. Given a graph, a number of different questions (some of which arising from real-world settings) can be asked on the set of matchings of G. For instance, one may ask if G has a *perfect matching*, i.e. a matching M such that each vertex of V is $M-$covered, or to find a (inclusionwise) maximal matching that contains the smallest possible number of edges. In this short note we deal with the maximum matching problem, defined as follows:

The Maximum Matching problem (*MM*)
Given: a graph $G(V, E)$;
Find: a matching M of G of maximum cardinality.

An immediate generalization of the latter problem can be obtained by assigning a *weight* $w_e \in \mathbb{R}$ to each edge $e \in E$, and asking for a matching of maximum weight. (In the following, given a set S and a function $f : S \to \mathbb{R}$, we define $f(S) := \sum_{s \in S} f(s)$).

The Maximum Weighted Matching problem (*MWM*)
Given: a graph $G(V, E)$ and a weight function $w : E \to \mathbb{R}$;
Find: a matching M of G such that $w(M) = \max\{w(M') : M' \text{ is a matching of } G\}$.

Given a graph $G(V, E)$, its *line graph* $L(G)$ is a graph whose vertex set is made of a vertex v_e for each edge $e \in E$, and given $e, f \in E$, two vertices v_e, v_f are adjacent in $L(G)$ if and only if e and f share an endpoint. A *stable set* of a graph G is a set $S \subseteq V$ such that no two vertices of S are adjacent in G. As one easily checks, $M \subseteq E$ is a matching in G if and only if the set $L(M)$ obtained by replacing each edge e of M with the vertex v_e is a stable set of $L(G)$ (See Figure 1 for an example).

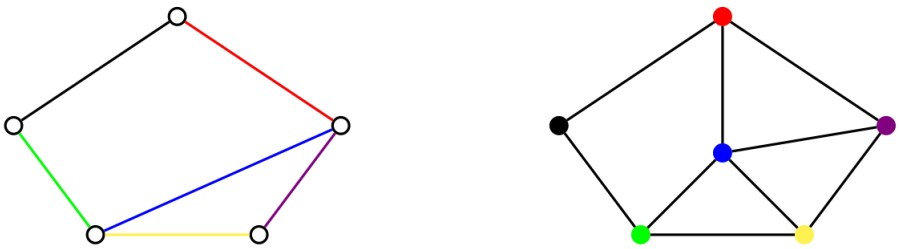


Figure 1: A graph G (on the left) and its line graph $L(G)$. An edge of G of a given color corresponds to a vertex of $L(G)$ of the same color, thus the matching {yellow, red} of G corresponds to the stable set {yellow, red} of $L(G)$.

Hence, the following problem is a generalization of *MWM*:

The Maximum Weighted Stable Set problem (*MWSS*)
Given: a graph $G(V, E)$ and a weight function $w : V \to \mathbb{R}$;
Find: a stable set S of G such that $w(S) = \max\{w(S') : S' \text{ is a stable set of } G\}$.

It is well-known [12] that *MWSS* in general graphs is NP-Hard, thus it is conjectured (and strongly believed) that a polynomial time algorithm for solving this problem does not exist. From what argued above, it follows that *MWSS* in line graphs (*MWSSL*) is equivalent to *MWM*. We now introduce a tractable generalization of the latter. An *induced subgraph* $H(U, F)$ of a given graph $G(V, E)$ is a graph such that $U \subseteq V$ and $e \in F$ if and only if *a)* $e \in E$ and *b)* the endpoints of e belong to U. A *claw* is the graph depicted in Figure 2. A graph is called *claw-free* if it has no induced subgraph that is isomorphic to a claw. The following relation between line and claw-free graphs can be easily shown true.

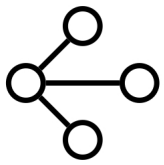


Figure 2: A claw.

Lemma 1 *Each line graph is a claw-free graph, while not every claw-free graph is a line graph.*

Thus, *MWSS* in claw-free graphs (*MWSSC*) is a generalization of *MWSSL* and consequently of *MWM*. Note that a similar argument holds for the *cardinality* or *unweighted* case, i.e. when all weights are equal to 1: the maximum cardinality stable set problem in claw-free graphs (*MSSC*) is a generalization of the maximum cardinality stable set problem in line graphs (*MSSL*), which in its turn is equivalent to *MM*.

3 The Maximum Matching problem

A *path* in a graph $G(V, E)$ is a sequence $v_1, \ldots, v_k$ of distinct vertices from V such that $v_i v_{i+1} \in E$ for each $i = 1, \ldots, k-1$. Given a graph $G(V, E)$ and a matching M on G, a path $P = v_1, \ldots, v_k$ is called $M-$*augmenting* if k is even, $v_i v_{i+1} \in M$ if and only if i is even, and v_1, v_k are $M-$exposed. Hence, in particular, the edges $v_1 v_2$ and $v_{k-1} v_k$ belong to $E \setminus M$, and $|P \cap M| = |P \setminus M| - 1$. It is easy to check that, given a matching M and an $M-$augmenting path P, the set $M \triangle E(P)$ is a matching of cardinality $|M| + 1$, where we denoted by $E(P)$ the set of edges between consecutive vertices of P, and by $\triangle$ the symmetric difference operator. Thus, a necessary condition for a matching M in a graph G to be of maximum cardinality is that there is no $M-$augmenting path in G. This condition turns out to be also sufficient as shown by Petersen and independently by Berge.

Figure 3: An $M-$augmenting path: matching edges are red.

Theorem 2 [1, 19] *Given a matching M in a graph G, M is a matching of maximum cardinality if and only if there is no $M-$augmenting path in G.*

The previous theorem immediately suggests a general scheme for solving MM in a graph.

Input: a graph G. *Output*: a matching of G of maximum cardinality.
Set $M = \emptyset$.
while there is an $M-$augmenting path P in G **do**
  set $M = M \triangle P$;
end while
stop: M is a solution to MM.

As the size of the matching increases at each step by exactly one, the algorithm above outputs after at most $\frac{V}{2}$ iterations a matching of maximum cardinality. In order to fully define the procedure, we are left with specifying how to find an $M-$augmenting path, or decide that such a path does not exist. In particular, it is not clear that a polynomial time algorithm for this problem exists. We now provide such an algorithm, starting with graphs with a special structure.

3.1 Finding $M-$augmenting paths

A graph $G(V, E)$ is *bipartite* if there exists a bipartition U, W of V such that there is no edge between any two vertices of U and no edge between any two vertices of W. It turns out that $M-$augmenting paths in bipartite graphs can be found more easily than in general graphs. In order to sketch an algorithm for this problem, we need to deal with *digraphs*, which are graphs whose set E is composed of ordered pairs of vertices (thus, an edge of a digraph is now denoted by (u, v)). A *path* P in a digraph $D(V, E)$ is defined to be a sequence $v_1, \ldots, v_k$ of distinct vertices from V such that, for each $i = 1, \ldots, k-1$, $(v_i, v_{i+1}) \in E$. We also say that the path P above *connects* v_1 *to* v_k, and for each pair $S, T \subseteq V$ such that $v_1 \in S$ and $v_k \in T$, that P *connects* S *to* T.

The Path Problem on Digraphs (PPD)
Given: a digraph $D(V, E)$ and subsets S, T of V;
Find: a path that connects S to T, or determine that no such a path exists.

PPD is a well-known problem in combinatorial optimization; in particular, there exists an $O(|E|)$ algorithm for solving it (see e.g. [21]). We now show that the problem of finding an $M-$augmenting path in a bipartite graph with respect to a given matching M can be reduced to PPD.

Let $G(V, E)$ be a graph and M be a matching in G. Define $D(V, E')$ to be the digraph with edge set

$$E' = \{(u, v) : \exists z \in V \text{ with } uz \in E, vz \in M\}.$$

It is not difficult to see that each path in D that connects the set S of $M-$exposed nodes to the set T of vertices that are adjacent to some vertex of S, corresponds to some $M-$augmenting path in G. More precisely, given a path $P = v_1, \ldots, v_k$ in D as above, we can construct an $M-$augmenting path P' in G as follows:

(a) for $i = 1, \ldots, k-1$ replace vertex v_i with v_i, z where z is the vertex of V such that $v_i z \in E$, $v_{i+1} z \in M$ (note that such a z exists by definition of E', and it is unique by definition of matching);

(b) as last vertex of the path, add a vertex $s \in S$ that is adjacent to v_k (such a vertex exists since, by definition, $v_k \in T$),

and vice versa, by inverting the operations above, a path in D that connects S to T can be constructed from an $M-$augmenting path in G. See Figure 4 for an example.

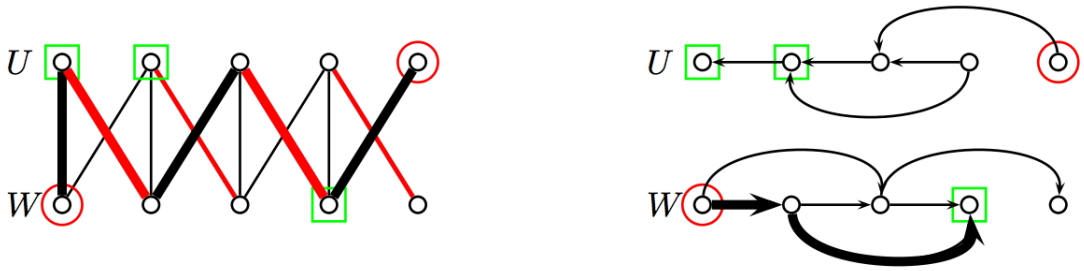


Figure 4: A bipartite graph G (on the left) with a matching (in red) and an augmenting path (in bold); on the right, the corresponding digraph D and path connecting S to T (in bold). In both graphs, vertices from S have a red circle around, while vertices from T are surrounded by a green box.

As argued above, for a given instance one needs to find at most $\frac{|V|}{2}$ augmenting paths. Thus, the following result by Kuhn holds true.

Theorem 3 [13] *The Maximum Matching problem in a bipartite graph $G(V, E)$ can be solved in time $O(|V||E|)$.*

More recently, faster algorithms (e.g. [11]) for solving *MM* in bipartite graphs appeared (see [21] for a complete list).

When considering general graphs, things can get nastier. In fact, the equivalence above is not anymore true — it is a simple exercise to find a path that connects S to T in D that does not correspond to an $M-$augmenting path in some non-bipartite graph G. Thus, it requires some extra mathematical (and computational) work to come up with a polynomial time algorithm for finding an augmenting path in the general case, but as Edmonds showed the following holds true.

Theorem 4 [4] *The Maximum Matching problem in a graph $G(V, E)$ can be solved in $O(|V|^2|E|)-$time.*

The complexity of the algorithm from the theorem above can be lowered to $O(|V|^3)$ [8], and even below (see [21] for a complete list).

4 The Maximum Weighted Matching problem

In the weighted case, a dichotomy similar to the one for the cardinality case holds true: in bipartite graphs, *MWM* can be solved more easily than in general graphs. In particular, combinatorial arguments are not sufficient for the latter, and the algorithms rely on primal-dual methods for linear programming. We skip details, and only state theorems.

Theorem 5 (Edmonds and Karp [6], Tomizawa [23]) *The Maximum Weighted Matching problem in a bipartite graph $G(V, E)$ can be solved in $O(|V|(|E| + |V| \log |V|))-$time.*

Theorem 6 (Edmonds [5]) *The Maximum Weighted Matching problem in a graph $G(V, E)$ can be solved in time $O(|V|^2|E|)$.*

It has been shown that the algorithm from the latter theorem can be implemented as to run in $O(|V|^3)-$time, and below that (see [21]).

5 The Maximum Stable Set problem in Claw-free graphs

As mentioned in Section 2, *MWSSC* is a generalization of *MWM*. Thus, it is not clear a priori that a polynomial time algorithm for *MWSSC* exists — recall that the stable set problem in general graphs is NP-Hard. But, in fact, it does. Quite a number of studies have been devoted to this subject: the first polynomial time algorithms trace back to the early 1980s, and recent work on the subject used new techniques to improve the complexity bound by a significant amount. We now outline the main results for this problem, starting with the *cardinality* or *unweighted* case, i.e the case when all weights are equal to 1.

5.1 The cardinality case

A first stream of algorithms for *MWSSC* exploit the fact that Theorem 2 can be extended to stable sets in claw-free graphs. Indeed, given a graph $G(V, E)$ and a stable set S of G, a path $P = v_1, \ldots, v_k$ is an $S-$augmenting path if k is odd, $v_i \in S$ if and only if i is even, and $(S \setminus \{v_2, v_4, \ldots, v_{k-1}\}) \cup \{v_1, v_3, \ldots, v_k\}$ is a stable set of G. Note in particular, that the latter is a stable set of cardinality $|S| + 1$. Berge observed the following.

Theorem 7 [2] *Given a stable set S in a claw-free graph G, S is a stable set of maximum cardinality if and only if there is no $S-$augmenting path in G.*

This immediately suggests that, similarly to the matching case, one could find a maximum stable set in a claw-free graph by iteratively searching for an $S-$augmenting path. Again, it is not clear how to find such a path in a claw-free graph. Minty shows that the latter problem can be reduced to the solution of a *MWM* in an auxiliary graph, which as reported in Theorem 6 can be solved in polynomial time. Thus, we obtain the following.

Theorem 8 [15] *A stable set of maximum cardinality in a claw-free graph can be found in polynomial time.*

A similar result to the one above, again using techniques derived by matching, has been obtained by Sbihi [20]. Minty gave no explicit bound on the complexity of his algorithm, but it has been shown that it can be implemented as to run in time $O(|V|^5)$ [22], with V being the set of vertices of the input graph.

A different algorithm is based on *reduction* techniques. Let $G(V, E)$ be a claw-free graph, R be an operator that maps G into a graph $R(G)$ such that one can easily (e.g. in polynomial time) obtain a stable set of maximum cardinality in G from a stable set of maximum cardinality in $R(G)$. Now suppose that one can show that after a polynomial (in $|V|$ and $|E|$) number of applications of the operator R, the graph obtained is a line graph G', and recall that the maximum stable set problem in a line graph is equivalent to a maximum matching problem (see Section 2). As a maximum matching problem can be solved in polynomial time by Theorem 4, one can solve MSS in G' and deduce a MSS for G in polynomial time. This is the key idea of the algorithm proposed by Lovász and Plummer, who proved the following:

Theorem 9 [14] *A stable set of maximum cardinality in a claw-free graph $G(V, E)$ can be found in $O(|V|^4)-$time.*

5.2 The weighted case: first algorithms

It is not obvious how to extend to the weighted case the algorithms presented in the previous section. Minty claimed that his procedure could indeed be generalized to the weighted case, but his argument was buggy, and the algorithm may fail to return the optimal solution. This mistake was discovered by Nakamura and Tamura [16], who suitably modified Minty's procedure to turn it into a correct polynomial time algorithm for *MWSSC*. A simpler modification of Minty's algorithm, again leading to a polynomial time algorithm, was proposed by Schrijver [21].

Theorem 10 [16, 21] *A maximum weighted stable set in a claw-free graph can be found in polynomial time.*

Schrijver's modification of Minty's algorithm can be implemented as to run in $O(|V|^5 \log |V| + |V|^4|E|)-$time [22], with V (resp. E) being the vertex set (resp. edge set) of the input graph.

Recently, other algorithms have been proposed for the problem. The one by Nobili and Sassano [17] combines reduction techniques and detection of $S-$augmenting paths, and can be implemented as to run in $O(|V|^4 \log |V|)-$time, with V being again the vertex set of the input graph. The state of the art algorithm for *MWSSC* is based on a *decomposition* technique introduced by Oriolo, Pietropaoli and Stauffer [18]. We are now presenting both the technique and the algorithm, starting with some definitions.

5.3 Intermezzo: solving *MWSS* via decomposition

A *strip* is a triple (G, A, B) where G is a graph, and A, B are (possibly empty) *cliques* of G, where a clique of G is an induced subgraph $H(U, F)$ of G with the property that each pair of vertices from U are adjacent. We call A, B the *extremities* of the strip, and remark that $A = B$ is possible. Let $\{G_i(V_i, E_i)\}_{i=1}^k$ be a family of vertex disjoint graphs, and $\mathcal{F} = \{(G_i, A_i, B_i)\}_{i=1}^k$ a family of strips. The *composition* of $\mathcal{F}$ with respect to a partition $\mathcal{P}$ of the multi-set $\{A_i\}_{i=1}^k \cup \{B_i\}_{i=1}^k$, is the graph $G(V, E)$ obtained as follows:

- $V = \cup_{i=1}^k V_i$;
- $E = E' \cup_{i=1}^k E_i$, where $uv \in E'$ if and only if there exist (possibly coincident) indices $i, i' \in \{1, \ldots, k\}$ such that u belongs to an extremity of G_i, say A, and v belongs to an extremity of $G_{i'}$, say B, and A and B belong to the same set P from the partition $\mathcal{P}$.

Thus, the composition of strips is an operation that, from (usually simple) graphs, constructs a new (usually more complex) one. On the other hand, a *decomposition* of a graph G is a set of strip that can be composed as to obtain G. See Figure 5 for an example of the composition of strips. The concept of composition and decomposition have been introduced in structural graph theory, but turned out to have interesting algorithmic property. In fact, the following theorem shows that, if one can solve in polynomial time *MWSS* in graphs $G_1, \ldots, G_k$, then also *MWSS* in graphs obtained as the composition of strips $(G_1, A_1, B_1), \ldots, (G_k, A_k, B_k)$ can be solved in polynomial time.

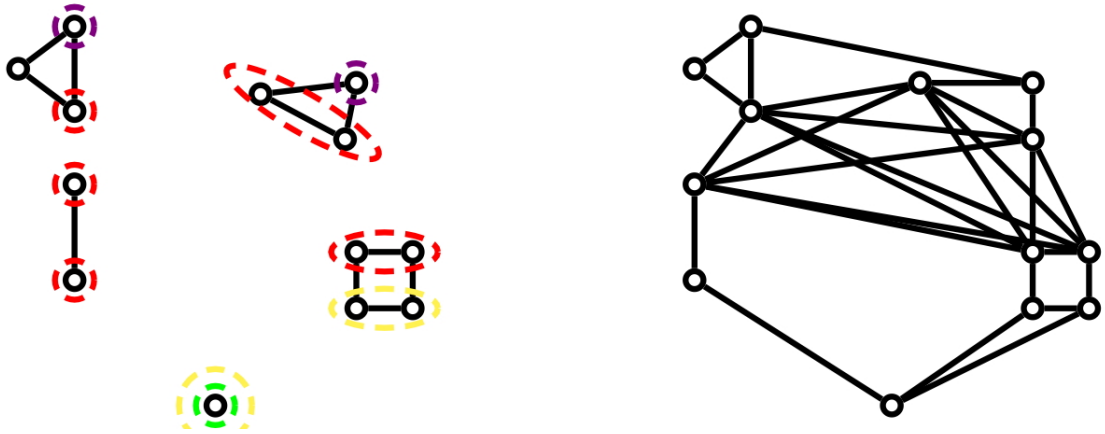


Figure 5: On the left, a family of five strips: the extremities of each strip are surrounded by a dashed ellipse, and extremities in the same set of the partition $\mathcal{P}$ are surrounded by ellipsis of the same color. On the right: the graph obtained by the composition of the strips w.r.t. $\mathcal{P}$.

Theorem 11 (Oriolo, Pietropaoli, and Stauffer [18]) *Let $G(V, E)$ be the composition of strips (G_i, A_i, B_i), $i = 1, \ldots, k$, with respect to some partition $\mathcal{P}$. If, for each $i = 1, \ldots, k$,* MWSS *in $G_i(V_i, E_i)$ can be computed in time $O(p_i(V_i))$, then* MWSS *in G can be solved in time $O(\sum_{i=1,\ldots,k} p_i(V_i) + O(|V||E| + |V|^2 \log |V|))$.*

5.4 A decomposition approach to *MWSSC*

In the previous paragraph, we argued that knowing a decomposition of a graph can be a useful tool in solving *MWSS*. We now show that such a decomposition is actually at hand for claw-free graphs. An important recent result in graph theory states the following.

Theorem 12 (Chudnovsky and Seymour [3]) *Each claw-free graph:*

- *either belongs to the class of graphs $\mathcal{G}$,*
- *or can be obtained as the composition of strips from the class $\mathcal{H}$.*

We call a result as the one above a *decomposition theorem*. Of course one cannot understand its importance without defining the families $\mathcal{G}$ and $\mathcal{H}$. Details on those family are not important to us. What matters is that $\mathcal{G}$ is a family of well-known graphs with a highly structured shape, while the strips from $\mathcal{H}$ are "very easy" (e.g. they have a bounded number of vertices, or can be obtained by a strip with a bounded number of vertices by iteratively performing some transformation on the graph). Thus, in particular, a broad class of claw-free graphs can be written as the composition of strips with a relatively simple structure, and the remaining ones are, in a way, "well-known". (We shall not say more on the family $\mathcal{G}$ and $\mathcal{H}$; the interested reader may refer to [3].)

Unfortunately, Theorem 12 does not immediately imply the existence of a polynomial time procedure for recognizing whether an input claw-free graph G belongs to $\mathcal{G}$ and, in case it does not, obtaining a decomposition of G into strips from $\mathcal{H}$. (Such a polynomial time algorithm has been very recently obtained by Hermelin, Mnich, Van Leeuwen, and Woeginger [10] but it is not clear that using it will lead to a fast algorithm for *MWSSC*). Nevertheless, the "decomposition approach" seems promising, thus one could search for a friendlier decomposition theorem for claw-free graphs. Oriolo, Pietropaoli and Stauffer's [18] proposed one such result for a subclass of claw-free graphs, and apply it together with Theorem 11 in order to obtain an algorithm for *MWSSC* that run in time $O(|V|^6)$ (V being again the vertex set of the input graph).

Even though the complexity of their algorithm does not improve over the state of the art for *MWSSC*, it shows the potential of combining decomposition results (often available in structural graph theory) with algorithms for "composing" solution to combinatorial optimization problems (like Theorem 11). It is possible that this approach may turn out to be efficient for problems other than stable set, or in classes of graphs other than claw-free. However, already for *MWSSC*, their approach can be refined in order to obtain a more efficient algorithm. Indeed, one can show that the following algorithmic decomposition theorem for claw-free graphs with stability number at least 4 holds true (the *stability number* of a graph G is the maximum size of a stable set of G):

Theorem 13 (Faenza, Oriolo, and Stauffer [7]) *Each claw-free graph with stability number at least 4:*

- *is either quasi-line, distance claw-free without articulation cliques,*
- *or can be obtained as the composition of strips from a class $\mathcal{H}''$.*

Moreover, one can distinguish between the two cases (and, if possible, obtain the strip decomposition) in $O(|V|^3)-$time.

(We omit details on the class of graphs mentioned above and on algorithms for solving *MWSS* in such classes: the interested reader may refer to [7]). By exploiting the theorem above, Faenza, Oriolo, and Stauffer [7] obtained the following $O(|V|^3)-$time algorithm for the maximum weighted stable set problem in a claw-free graph $G(V, E)$, currently being the theoretically fastest algorithm from the literature:

- Recognize in $O(|V|^3)-$time if G has stability number $\alpha \leq 3$;
- If $\alpha \leq 3$, solve *MWSS* in G by enumerating in $O(|V|^3)$ all its stable sets, and picking the one of maximum weight;
- Else $\alpha \geq 4$: use Theorem 13 to recognize in $O(|V|^3)$ that G is quasi-line, distance claw-free without articulation cliques, or to obtain a decomposition of G into strips from $\mathcal{H}''$;
- If G is quasi-line, distance claw-free without articulation cliques, compute *MWSS* in G in $O(|V|^3)-$time.
- Else solve *MWSS* in each strip from $\mathcal{H}''$ in $O(|V|^2)-$time, and use Theorem 11 to solve *MWSS* in the input graph in $O(|V|^3)-$time.

6 Concluding remarks

As the results presented in this short note witness, when moving from algorithms for *MM* to those for its generalizations *MWM* and *MWSSC*, one needs deeper knowledge of graphs and more sophisticated algorithmic tools. This increasing complexity is mirrored in the running time of the routines for solving those problems. Indeed, the state of the art algorithms for *MM*, *MWM*, and *MWSSC* run respectively in $O(\sqrt{|V|}|E| \log_{|V|} \frac{|V|^2}{|E|})$, $O(|V||E| + |V|^2 \log |V|)$ (see [21]), and $O(|V|^3)$ (see [7]), where once again we considered $G(V, E)$ to be the input graph. One may ask whether those complexity bound are tight, or there is room for an improvement. As the the complexity bound for *MM* and *MWM* are due to algorithms that appeared more than 15 years ago, such an improvement for those problem seems to be unlikely. On the other hand, the state of the art algorithm for *MWSSC* appeared in the literature as late as this year, thus it is more likely that its complexity bound can be improved. Since the root graph of a line graph $G(V, E)$ has $O(|V|)$ edges and vertices, *MWSS* in line graphs can be solved in $O(|V|^2 \log |V|)$. Can *MWSSC* can be solved within this time bound ?

References

[1] C. Berge, *Two theorems in graph theory.* Proceedings of the National Academy of Sciences of the United States of America 43 (1957), 842–844.

[2] C. Berge, *Balanced matrices.* Mathematical Programming 2 (1972), 19–31.

[3] M. Chudnovsky and P. Seymour, *Claw free Graphs IV. Global structure.* Journal of Combinatorial Theory. Ser B 98 (2008), 1373–1410.

[4] J. Edmonds, *Paths, trees and flowers.* Canadian Journal of Mathematics 17 (1965), 449–467.

[5] J. Edmonds, *Maximum matching and a polyhedron with 0,1-vertices.* Journal of Research of the National Bureau of Standards 69 (1965), 125–130.

[6] J. Edmonds, R.M. Karp, *Theoretical improvements in algorithmic efficiency for network flow problems.* In *Combinatorial Structures and Their Applications.* Gordon and Breach, New York (1970), 93–96.

[7] Y. Faenza, G. Oriolo, and G. Stauffer, *An algorithmic decomposition of claw-free graphs leading to an $O(n^3)-$algorithm for the weighted stable set problem.* Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms (SODA) (2011), 630–646.

[8] H. N. Gabow, *Data structures for weighted matching and nearest common ancestor with linking.* In *Proceeding of the First Annual ACM-SIAM Symposium on Discrete Algorithms*, New York (1990), 321–325.

[9] B. Gerards, *Matching.* In M.O. Ball, T.L. Magnanti, C.L. Monma and G.L. Nemhauser eds, *Network Models*, Handbooks in Operations Research and Management Science, Volume 7. North-Holland, Amsterdam (1995), 135–224.

[10] D. Hermelin, M. Mnich, E. J. van Leeuwen, G. Woeginger, *Domination When the Stars are Out.* Proceedings of ICALP 2011, to appear.

[11] J.E. Hopcroft, R.M. Karp, *A $n^{5/2}$ algorithm for maximum matchings in bipartite graphs.* In *Conference Record 1971 Twelfth Annual Symposium on Switching and Automata Theory* (East Lansing, Michigan, 1971), IEEE. New York (1971), 122–125.

[12] Richard M. Karp, *Reducibility Among Combinatorial Problems.* In R. E. Miller and J. W. Thatcher (editors), *Complexity of Computer Computations.* New York: Plenum (1972), 85–103.

[13] H.W. Kuhn, *The Hungarian method for the assignment problem.* Naval Research Logistics Quarterly 2 (1955), 83–97.

[14] L. Lovász and M.D. Plummer, "Matching theory". North Holland, Amsterdam, 1986.

[15] G. J. Minty, *On maximal independent sets of vertices in claw-free graphs.* Journal on Combinatorial Theory 28 (1980), 284–304.

[16] D. Nakamura and A. Tamura, *A revision of Minty's algorithm for finding a maximum weighted stable set of a claw-free graph.* Journal of the Operations Research Society of Japan, 44/2 (2001), 194–204.

[17] P. Nobili, A. Sassano, *A reduction algorithm for the weighted stable-set problem in claw-free graphs.* Talk at AIRO 2009.

[18] G. Oriolo, U. Pietropaoli and G. Stauffer, *A new algorithm for the maximum weighted stable set problem in claw-free graphs.* In A. Lodi, A. Panconesi and G. Rinaldi, editors, *Proceedings Thirteenth IPCO Conference* (2008), 77–96.

[19] J. Petersen, *Die Theorie der regulären graphs.* Acta Mathematica 15 (1891), 193–220.

[20] N. Sbihi, *Algorithme de recherche d'un stable de cardinalité maximum dans un graphe sans étoile.* Discrete Mathematics 29 (1980), 53–76.

[21] A. Schrijver, "Combinatorial optimization. Polyhedra and efficiency". Algorithms and Combinatorics 24. Springer Berlin, 2003 (3 volumes).

[22] M. Senatore, "Analisi della complessità di algoritmi risolutivi per il problema del massimo insieme stabile in grafi claw-free". Master's thesis, Università di Roma Tor Vergata, 2009.

[23] N. Tomizawa, *On some techniques useful for solution of transportation network problems.* Networks 1 (1971), 173–194.

Factorization in categories of modules

Marco Perone (*)

1 Introduction

Given a ring R, the class of isomorphism classes of right R-modules forms a commutative monoid V_R under direct sum. This means that we can describe the behaviour of direct sum decomposition of R-modules with the monoid V_R. Similarly, if $\mathcal{C}$ is a class of R-modules closed under isomorphism and direct sum, we can study the behaviour of the direct sum decomposition in $\mathcal{C}$ by mean of the commutative monoid $V(\mathcal{C}) \subseteq V_R$ of isomorphism classes of modules in $\mathcal{C}$.

In particular, studying the direct sum decomposition in a class $\mathcal{C}$ of R-modules is equivalent to study the factorization of elements in the monoid $V(\mathcal{C})$.

Definition 1.1 Let M be a monoid. An element $a \in M$ is *invertible*, or a *unit*, if there exists an element $b \in M$ such that $a + b = b + a = e$, where e denotes the identity element of M.

An element $a \in M$ is an *atom* if it is not invertible and $a = b + c$ implies b invertible or c invertible.

We say that a monoid M is *atomic* if every element of M can be written as a finite sum of atoms.

To study the factorization in a monoid M means:

- identify the atoms of M;
- determine when the sum of two finite families of atoms coincide;
- determine all the possible factorizations for every element of the monoid.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: pasafama@gmail.com . Seminar held on 31 January 2011.

2 Free commutative monoids and the Krull-Schmidt-Azumaya Theorem

The easiest case in which we can investigate factorization is the following.

Example 2.1 A monoid M is a *free commutative monoid* if it is isomorphic to the direct sum $\mathbb{N}_0^{(X)}$ of copies of the monoid of non-negative integers. We can see the elements of M as functions $s\colon X \to \mathbb{N}_0$ such that $s(x) \neq 0$ only for finitely many $x \in X$. Then, given two elements $s_1, s_2\colon X \to \mathbb{N}_0$ of M, we have

$$(s_1 + s_2)(x) = s_1(x) + s_2(x).$$

- An element $s\colon X \to \mathbb{N}_0$ of M is an atom if an only if there exists $x \in X$ such that $s(x) = 1$ and $s(y) = 0$ for every $x \neq y \in X$.
- Given two families $s_1, \ldots, s_n$ and $t_1, \ldots, t_m$ of atoms of M, the equality $s_1 + \ldots + s_n = t_1 + \ldots + t_m$ implies that $m = n$ and there exists a permutation $\sigma \in \mathcal{S}_n$ such that $s_i = t_{\sigma(i)}$ for every $i = 1, \ldots, n$.
- Every element $s \in M$ admits a decomposition as a sum of atoms, unique up to a permutation of the summands.

It is clear that in free commutative monoids the factorization of every element is essentially unique, because the only thing that can change between two factorization of the same element is the order of the summands.

Let $\mathcal{C}$ be a class of R-modules closed under isomorphism and direct sum. If we want to study the factorization in $V(M)$, the first step is to identify the atoms.

Definition 2.2 A non-zero right R-module M is *indecomposable* if $M = M_1 \oplus M_2$ implies $M_1 = 0$ or $M_2 = 0$.

Example 2.3 If R is a division ring, a right R-vector space V is indecomposable if and only if it has dimension 1.

Example 2.4 A finitely generated abelian group is indecomposable if and only if it is isomorphic to $\mathbb{Z}$ or to $\mathbb{Z}/p^n\mathbb{Z}$ for some prime number p and some positive integer n.

In fact in the two cases of the previous examples, it happens that the monoid $V(\mathcal{C})$ is a free commutative monoid. Hence the direct sum decomposition in these classes has an extremely regular behaviour.

Example 2.5 If R is a division ring and $\mathcal{C}$ is the class of all finitely dimensional right R-vector spaces, then the monoid $V(\mathcal{C})$ is isomorphic to the free commutative monoid $\mathbb{N}_0$.

Example 2.6 Let G be a finitely generated abelian group. Then there exist prime numbers $p_1, \ldots, p_n$, an integer $k \geq 0$ and positive integers $k_1, \ldots, k_n$ such that

$$G \cong \mathbb{Z}^k \oplus \mathbb{Z}/p_1^{k_1}\mathbb{Z} \oplus \ldots \oplus \mathbb{Z}/p_n^{k_n}\mathbb{Z}.$$

Moreover, the prime numbers $p_1, \ldots, p_n$ and the integers $k, k_1, \ldots, k_n$ are uniquely identified by G.

We can state the result of the last example saying that the monoid $V(\mathcal{C})$, where $\mathcal{C}$ is the class of finitely generated abelian groups, is a free commutative monoid.

Now we want to find more general classes $\mathcal{C}$ of modules that provide free commutative monoids $V(\mathcal{C})$. To do this we need to restrict to classes of modules having the appropriate endomorphism rings.

Proposition 2.7 *Let R be a ring and x an element of R. Then the following are equivalent.*

- *$x \in M_R$ for every maximal right ideal M_R.*
- *$x \in {}_RM$ for every maximal left ideal ${}_RM$.*
- *$1 - xy$ is right invertible for any $y \in R$.*
- *$1 - yx$ is left invertible for any $y \in R$.*
- *$1 - yxz$ is two-sided invertible for any $y, z \in R$.*
- *$Mx = 0$ for any simple right R-module M.*

The set of all the elements $x \in R$ satisfying the conditions of the Proposition form an ideal $J(R)$, called the *Jacobson radical* of R.

Definition 2.8 A ring R is said to be *local* if it satisfies one of the following equivalent conditions:

- $R/J(R)$ is a division ring;
- R has a unique left ideal;
- R has a unique right ideal;
- the sum of two non-invertible elements of R is non-invertible.

The following theorem explains the key role played by local endomorphism rings in the theory of factorization of modules.

Theorem 2.9 [Krull-Schmidt-Azumaya] *Let M_i, $i \in I$, and N_j, $j \in J$, right R-modules with local endomorphism ring. Then*

$$\oplus_{i \in I} M_i \cong \oplus_{j \in J} N_j$$

if and only if there is a bijection $\sigma \colon I \to J$ such that $M_i \cong N_{\sigma(i)}$ for every $i \in I$.

The Theorem implies the following.

Theorem 2.10 *Let R be a ring and let $\mathcal{C}$ be the class of modules that are finite direct sums of modules with local endomorphism ring. Then the monoid $V(\mathcal{C})$ is a free commutative monoid.*

The examples that follow provide classes of modules with local endomorphism ring, and therefore of classes $\mathcal{C}$ of modules such that the monoid $V(\mathcal{C})$ is a free commutative monoid.

Example 2.11 Any indecomposable module of finite length has local endomorphism ring.

Example 2.12 The endomorphism ring of a simple module is a division ring.

Example 2.13 Every artinian module with simple socle has local endomorphism ring.

3 Krull monoids

To every commutative monoid M, we can associate a preorder $\leq$ defined as follows: for elements $a, b \in M$, we have $a \leq b$ if and only if there exists an element $c \in M$ such that $a + c = b$. This is called the *algebraic preorder* of M. It is clear that the algebraic preorder of a monoid keeps track of some properties of the factorization of the monoid, so it is really interesting to understand this preorder for a given monoid M.

Example 3.1 It is clear that in a free commutative monoid $M = \mathbb{N}_0^{(X)}$, two elements $a, b \in M$ satisfy $a \leq b$ if and only if every component of a is less or equal that the corresponding component of b with respect to the usual order of $\mathbb{N}_0$.

We now introduce a class of monoids whose algebraic preorder is controlled by the algebraic preorder of a free commutative monoid. This means that we can recover the algebraic preorder of the monoid M looking at the algebraic preorder of a free commutative monoid. This is provided by a *divisor homomorphism*, that is a morphism of monoids $f\colon M \to N$ such that $f(a) \leq f(b)$ implies $a \leq b$ for every $a, b \in M$.

Definition 3.2 A commutative monoid M is a *Krull monoid* if there is a divisor homomorphism $\varphi\colon M \to \mathbb{N}_0^{(I)}$ into a free commutative monoid $\mathbb{N}_0^{(I)}$.

In other words, M is a Krull monoid if and only if there exists a family of morphisms of monoids $f_i\colon M \to \mathbb{N}_0$, $i \in I$, such that, for every a and b in M:

- $f_i(a) = 0$ for almost all $i \in I$;
- $a \leq b$ if and only if $f_i(a) \leq f_i(b)$ for every $i \in I$.

If we restrict to monoids of the form $V(\mathcal{C})$ for some class $\mathcal{C}$ of R-modules, we notice that the zero module is the only invertible element of the monoid, i.e. $V(\mathcal{C})$ is a reduced monoid.

We say that a monoid M is cancellative if, given elements $a, b, c \in M$, we have that $a+c = b+c$ implies $a = b$. It is the case that a reduced Krull monoid must be cancellative. Therefore, if we want a monoid of the form $V(\mathcal{C})$ to be a Krull monoid, we need to look for classes $\mathcal{C}$ of R-modules that cancel from direct sum. To find these classes we need to restrict to the appropriate type of endomorphism rings.

Definition 3.3 A ring R is *semilocal* if there exist positive integers $k_1, \ldots, k_n$ and division rings $D_1, \ldots, D_n$ such that

$$R/J(R) \cong M_{k_1}(D_1) \times \ldots \times M_{k_n}(D_n).$$

If we restrict to semilocal endomorphism rings, our next theorem guarantees that modules cancel from direct sum.

Theorem 3.4 *Any R-module M_R with semilocal endomorphism ring cancels from direct sums.*

It turns out that assuming the endomorphism rings to be semilocal is enough to obtain Krull monoids.

Theorem 3.5 *Let $\mathcal{C}$ be a class of modules closed under isomorphism and finite direct sum, such that every module in $\mathcal{C}$ has semilocal endomorphism ring. Then the monoid $V(\mathcal{C})$ is a reduced Krull monoid.*

Since the endomorphism ring of every artinian module and of every uniserial module is semilocal, we immediately obtain the following examples.

Example 3.6 Given the class $\mathcal{C}$ of artinian modules over any ring R, the monoid $V(\mathcal{C})$ is a Krull monoid.

Example 3.7 Let $\mathcal{C}$ be the class of modules that are finite direct sums of uniserial modules over any ring R. Then the monoid $V(\mathcal{C})$ is a Krull monoid.

4 The 2-Krull-Schmidt Property

It is clear that in Krull monoids that are not free, one loses the uniqueness of the decomposition up to one permutation. Anyhow, inside the family of Krull monoids we can find classes of monoids with a very regular behaviour of the factorization. We are particularly interested in the following property, since, as we will see, it appears in some important categories of modules.

Definition 4.1 Let M be an atomic commutative monoid and let A be the set of atoms of M. Given two equivalence relations $\sim$ and $\equiv$ on A, we say that the *2-Krull-Schmidt Property* holds for M with respect to $\sim$ and $\equiv$ if and only if, for atoms $a_1, \ldots, a_n, b_1, \ldots, b_m \in A$,

we have

$$\sum_{i=1}^{n} a_i = \sum_{j=1}^{m} b_j$$

if and only if $m = n$ and there exist two bijections $\sigma, \tau \in \mathcal{S}_n$ such that $a_i \sim b_{\sigma(i)}$ and $a_i \equiv b_{\tau(i)}$ for every $i = 1, \dots, n$.

To provide an example of monoids satisfying the 2-Krull-Schmidt Property, we associate an atomic commutative monoid to every graph $G = (V, E)$, where V is the set of vertices and E the set of edges of G, in the following way. Consider the free commutative monoid $\mathbb{N}_0^{(V)}$ having as free set of generators the set $\{\delta_v \mid v \in V\}$. If $e = \{v, w\}$ is an edge of G, define $\delta_e = \delta_v + \delta_w$. Define $M(G)$ to be the submonoid of $\mathbb{N}_0^{(V)}$ generated by all the elements $\delta_e \in \mathbb{N}_0^{(V)}$, where e ranges in E.

Example 4.2 If G is a bipartite graph, then the 2-Krull-Schmidt Property holds for $M(G)$.

In fact, the example above is in some sense universal for monoids satisfying the 2-Krull-Schmidt Property. In fact, the following holds.

Theorem 4.3 *The following are equivalent for an atomic commutative monoid M.*

- *The 2-Krull-Schmidt Property holds for M.*
- *There exist a complete bipartite graph G and an injective monoid homomorphism $\varphi\colon M \to M(G)$ that sends atoms to atoms.*

As we did above, to obtain a category of modules $\mathcal{C}$ such that the 2-Krull-Schmidt holds for the monoid $V(\mathcal{C})$ we need to restrict to the appropriate endomorphism rings. In this case, the class of rings that we are looking for is the following.

Definition 4.4 A ring R is said to have *type* n if $R/J(R)$ is isomorphic to the product of n division rings.

We say that an R-module M_R has type n if its endomorphism ring has type n.

A useful criterion to determine if a ring has finite type is the following.

Proposition 4.5 *A ring is of finite type if and only if it has finitely many right ideals and they are all two-sided.*

We will be interested just in rings of type ≤ 2. We notice that rings of type 1 are exactly local rings. Hence modules of type 1 are necessarily indecomposable. There are also interesting examples of indecomposable modules of type 2.

Examples 4.6 The following holds.

- Every artinian module with heterogeneous socle of length 2 has type ≤ 2.
- Every uniserial module has type ≤ 2.

- Every cyclically presented module over a local ring has type ≤ 2.

The next theorem explains why we consider modules of type ≤ 2 to realize the 2-Krull-Schmidt Property. The weak (DSP) condition that we require in the hypotheses is a technical condition that assures us that there are enough objects in the category.

Theorem 4.7 *Let $\mathcal{D}$ be a class of indecomposable right R-modules of type ≤ 2 satisfying weak (DSP) and $\mathcal{C}$ the class of modules that are finite direct sums of modules in $\mathcal{D}$. Exactly one of the two following conditions hold:*

- *Either there exist two right R-modules M_1 and M_2 in $\mathcal{D}$ of type 2 such that $M_1 \oplus M_2$ has three non-isomorphic direct sum decompositions.*
- *Or the 2-Krull-Schmidt Property holds for the monoid $V(\mathcal{C})$.*

Now we provide some concrete examples of categories of modules where the 2-Krull-Schmidt Property is satisfied.

We say that two modules U and V are in the same *monogeny class*, and we write $[U]_m = [V]_m$, if there exist a monomorphism $U \to V$ and a monomorphism $V \to U$. Similarly, U and V are in the same *epigeny class*, and we write $[U]_e = [V]_e$, if there exist an epimorphism $U \to V$ and an epimorphism $V \to U$. It is clear that monogeny and epigeny class define two equivalence relations on the class Mod-R for every ring R.

Theorem 4.8 *Let $U_1, \ldots, U_n, V_1, \ldots, V_m$ be non-zero uniserial R-modules. Then $U_1 \oplus \ldots \oplus U_n \cong V_1 \oplus \ldots \oplus V_m$ if and only if $m = n$ and there exist two permutations $\sigma, \tau \in \mathcal{S}_n$ such that $[U_i]_m = [V_{\sigma(i)}]_m$ and $[U_i]_e = [V_{\tau(i)}]_e$ for every $i = 1, \ldots, n$.*

The 2-Krull-Schmidt Property holds also for the class of cyclically presented modules over a local ring, with respect to the equivalence relations defined by epigeny class and lower part. We say that two cyclically presented modules R/aR and R/bR over a local ring R have the same *lower part*, and we write $[R/aR]_l = [R/bR]_l$, if there exist $u, v \in U(R)$ and $r, s \in R$ such that $au = rb$ and $bv = sa$.

Theorem 4.9 *Let $U_1, \ldots, U_n, V_1, \ldots, V_m$ be non-zero cyclically presented modules over a local ring R. Then $U_1 \oplus \ldots \oplus U_n \cong V_1 \oplus \ldots \oplus V_m$ if and only if $m = n$ and there exist two permutations $\sigma, \tau \in \mathcal{S}_n$ such that $[U_i]_l = [V_{\sigma(i)}]_l$ and $[U_i]_e = [V_{\tau(i)}]_e$ for every $i = 1, \ldots, n$.*

5 Infinite Krull-Schmidt Properties

Up to now, we were interested only in finite sums. However, if we go back to the Krull-Schmidt-Azumaya Theorem, we notice how it concerns not only finite direct sums of modules with local endomorphism ring, but also infinite ones. In the same fashion one can ask if the regularity in the category of serial modules over any ring or in the category of cyclically presented modules over a local ring continues to hold when we pass from finite direct sums to infinite ones. More generally, one can try to search for a behaviour analogous to the 2-Krull-Schmidt Property where also infinite sums are allowed. It is clear

that we can not do this in monoids, since in this setting we can not perform infinite sums. Hence we need to define a new algebraic structure.

Definition 5.1 Let M be a class. If $\aleph$ is a cardinal number, we can define the class $M^{\aleph} = \{ f \colon \aleph \to M \mid f \text{ is a function } \}$. An $\aleph$-*operation* on M is a function $p_{\aleph} \colon M^{\aleph} \to M$.

We define a *commutative infinitary monoid* to be a class M together with an $\aleph$-operation $p_{\aleph}$ for every cardinal number $\aleph$ such that:

- $p_1 \colon M^1 \to M$ is the canonical bijection that sends the map $f \colon 1 \to M$, defined by $f(1) = m$, to the element $m \in M$;
- if $\aleph_i, i \in I$, and $\aleph$ are cardinal numbers, $\gamma_i \colon \aleph_i \to \aleph$, $i \in I$, are injective maps such that $\aleph = \dot{\bigcup}_{i \in I} \gamma_i(\aleph_i)$ and $\aleph_I = |I|$, then, for any $f \in M^{\aleph}$, $p_{\aleph}(f) = p_{\aleph_I}(\Gamma)$, where $\Gamma \in M^I$ is the function from I to M defined by $\Gamma(i) = p_{\aleph_i}(f\gamma_i)$, for any $i \in I$.

The second axiom provides, in one instance, the existence of an identity element, the commutativity and the associativity of M as a commutative infinitary monoid.

The easiest examples of commutative infinitary monoids are given by the class of cardinal numbers with the sum of cardinals and by the classes $V(\mathcal{C})$ of isomorphism classes of a family of modules closed under isomorphism and infinite direct sum.

In this setting we can define the infinitary version of the 2-Krull-Schmidt Property.

Definition 5.2 The *Infinite 2-Krull-Schmidt Property* holds for an atomic commutative infinitary monoid M if there exist two equivalence relations $\sim$ and $\equiv$ on the class A of atoms of M such that, given two families $\{ a_i \mid i \in I \}$ and $\{ b_j \mid j \in J \}$ of atoms of M, we have

$$\sum_{i \in I} a_i = \sum_{j \in J} b_j$$

if and only if there exist two bijections $\sigma, \tau \colon I \to J$ such that $a_i \sim b_{\sigma(i)}$ and $a_i \equiv b_{\tau(i)}$ for every $i \in I$.

Similarly to the finite case, given a graph G, let $M_{\infty}(G)$ be the submonoid of the free commutative infinitary monoid with basis the class of vertices of G generated by the edges of G. As in the finite case, the monoids of the form $M_{\infty}(G)$, where G is a complete bipartite graph, turn out to be universal with respect to the Infinite 2-Krull-Schmidt Property.

Proposition 5.3 *Let M be an atomic commutative infinitary monoid. Then the following are equivalent.*

- *The Infinite 2-Krull-Schmidt Property holds for M;*
- *There exist a complete bipartite graph G and an injective morphism of commutative infinitary monoids $M \to M_{\infty}(G)$ that sends atoms to atoms.*

If we go back to the examples of classes of modules where the 2-Krull-Schmidt Property holds that we presented above, we have that the Infinite 2-Krull-Schmidt Property holds only in the second case.

Theorem 5.4 *Let $\{U_i \mid i \in I\}$ and $\{V_j \mid j \in J\}$ be two families of cyclically presented modules over a local ring R. Then*

$$\oplus_{i\in I}U_i \cong \oplus_{j\in J}V_j$$

if and only if there exist two bijections $\sigma, \tau \colon I \to J$ such that $[U_i]_l = [V_{\sigma(i)}]_l$ and $[U_i]_e = [V_{\tau(i)}]_e$ for every $i \in I$.

In the case of uniserial modules, we have to restrict to a subclass to ensure that the Infinite 2-Krull-Schmidt Property holds.

We say that an R-module U is *quasi-small* if, whenever U is isomorphic to a direct summand of a direct sum $\oplus_{\lambda\in\Lambda}M_\lambda$ of arbitrary modules M_λ, there is a finite subset $F \subseteq \Lambda$ such that U is isomorphic to a direct summand of $\oplus_{\lambda\in F}M_\lambda$.

Theorem 5.5 *Let $\{U_i \mid i \in I\}$ and $\{V_j \mid j \in J\}$ be two families of non-zero quasi-small uniserial modules over an arbitrary ring R. Then*

$$\oplus_{i\in I}U_i \cong \oplus_{j\in J}V_j$$

if and only if there exist two bijections $\sigma, \tau \colon I \to J$ such that $[U_i]_m = [V_{\sigma(i)}]_m$ and $[U_i]_e = [V_{\tau(i)}]_e$ for every $i \in I$.

Anyway, if we have a closer look at the class of serial modules over a ring R, we find out that their behaviour with respect to direct sum decomposition is not far from the Infinite 2-Krull-Schmidt Property. We just have to relax a bit our requirements.

Definition 5.6 Let M be an atomic commutative infinitary monoid and let A be the class of atoms of M. Suppose we are given two subclasses A' and A'' of A such that $A' \cup A'' = A$, an equivalence relation $\sim$ on A' and an equivalence relation $\equiv$ on A''. We say that the *Infinite Quasi 2-Krull-Schmidt Property* holds for M with respect to the equivalence relations $\sim$ and $\equiv$ if for any couple of families $\{\, a_i \mid i \in I \,\}$ and $\{\, b_j \mid j \in J \,\}$ of atoms of M, we have that

$$\sum_{i\in I} a_i = \sum_{j\in J} b_j$$

if and only if there exist two bijections $\sigma \colon I' = \{\, i \in I \mid a_i \in A' \,\} \to J' = \{\, j \in J \mid b_j \in A' \,\}$ and $\tau \colon I'' = \{\, i \in I \mid a_i \in A'' \,\} \to J'' = \{\, j \in J \mid b_j \in A'' \,\}$ such that $a_i \sim b_{\sigma(i)}$ for every $i \in I'$ and $a_i \equiv b_{\tau(i)}$ for every $i \in I''$.

This is exactly the phenomenon that happens for serial modules. The following theorem in fact states that the Infinite Quasi 2-Krull-Schmidt Property holds for the class of serial modules over any ring with respect to the equivalence relations given by monogeny class and epigeny class.

Theorem 5.7 *Let $\{U_i \mid i \in I\}$ and $\{V_j \mid j \in J\}$ be non-empty families of non-zero uniserial modules. Let $I' = \{i \in I \mid U_i$ is quasi-small$\}$ and $J' = \{j \in J \mid V_j$ is quasi-small$\}$. Then*

$$\oplus_{i\in I}U_i \cong \oplus_{j\in J}V_j$$

if and only if there exist a bijection $\sigma\colon I \to J$ *and a bijection* $\tau\colon I' \to J'$ *such that* $[U_i]_m = [V_{\sigma(i)}]_m$ *for any* $i \in I$ *and* $[U_i]_e = [V_{\tau(i)}]_e$ *for any* $i \in I'$.

References

[1] Amini, A. and Amini, B. and Facchini, A., *Direct summands of direct sums of modules whose endomorphism rings have two maximal right ideals.* To appear in J. Pure Appl. Algebra (2011).

[2] Amini, A. and Amini, B. and Facchini, A., *Equivalence of diagonal matrices over local rings.* J. Algebra 320 (2008), 1288–1310.

[3] Amini, A. and Amini, B. and Facchini, A., *Weak Krull-Schmidt for infinite direct sums of cyclically presented modules over local rings.* Rend. Semin. Mat. Univ. Padova 122 (2009), 39–54.

[4] Azumaya, G., *Corrections and supplementaries to my paper concerning Krull-Remak-Schmidt's Theorem.* Nagoya Math. J. 1 (1950), 117–124.

[5] Camps, R. and Dicks, W., *On semilocal rings.* Israel J. Math. 81 (1993), 203–211.

[6] Facchini, A., "Module theory. Endomorphism rings and direct sum decompositions in some classes of modules". Birkhäuser Verlag, Basel, 1998.

[7] Facchini, A., *Krull monoids and their application in module theory.* In "Algebras, rings and their representations", World Scientific 2006, pp. 53–71.

[8] Facchini, A., *Krull-Schmidt fails for serial modules.* Trans. Amer. Math. Soc. 348 (1996), 4561–4575.

[9] Facchini, A. and Girardi, N., *Couniformly presented modules and dualities.* In "Advances in Ring Theory", Birkhäuser Verlag 2010, pp. 149–163.

[10] Facchini, A. and Halter-Koch, F., *Projective modules and divisor homomorphisms.* J. Algebra Appl. 2 (2003), 435–449.

[11] Facchini, A. and Herbera, D., *Two results on modules whose endomorphism ring is semiloca.* Algebr. Represent. Theory 7 (2004), 575–585.

[12] Facchini, A. and Herbera, D., *Local Morphisms and Modules with a Semilocal Endomorphism Ring.* Algebr. Represent. Theory 9 (2006), 403–422.

[13] Facchini, A. and Herbera, D. and Levy, L.S. and Vámos, P., *Krull-Schmidt Fails for Artinian Modules.* Proc. Amer. Math. Soc. 123/12 (1995), 3587–3592.

[14] Facchini, A. and Perone, M., *Maximal Ideals in Preadditive Categories and Semilocal Categories.* To appear in J. Algebra Appl. 10/1 (2011), 1–27.

[15] Facchini, A. and Příhoda, P., *Factor Categories and Infinite Direct Sums.* Int. Electron. J. Algebra 5 (2009), 135–168.

[16] Facchini, A. and Příhoda, P., *Endomorphism rings with finitely many maximal right ideals.* To appear in Comm. Algebra (2011).

[17] Facchini, A. and Příhoda, P., *The Krull-Schmidt Theorem in the case two.* To appear in Algebr. Represent. Theor. (2011).

[18] Facchini, A. and Wiegand, R., *Direct-sum decompositions of modules with semilocal endomorphism ring.* J. Algebra 274 (2004), 689–707.

[19] Příhoda, P., *On uniserial modules that are not quasi-small.* J. Algebra 299 (2006), 329–343.

[20] Příhoda, P., *A version of the Weak Krull-Schmidt Theorem for infinite direct sums of uniserial modules.* Comm. Algebra 34 (2006), 1479–1487.

[21] Příhoda, P., *Weak Krull-Schmidt theorem and direct sum decompositions of serial modules of finite Goldie dimension.* J. Algebra 281 (2004), 332–341.

[22] Puninski, G., *Some model theory over a nearly simple uniserial domain and decompositions of serial modules.* J. Pure Appl. Algebra 163 (2001), 319–337.

[23] Warfield Jr., R.B., *An infinite Krull-Schmidt theorem for infinite sums of modules.* Proc. Amer. Math. Soc. 22 (1969), 460–465.

[24] Warfield Jr., R.B., *Serial rings and finitely presented modules.* J. Algebra 37 (1975), 187–222.

Numerical solution of electrons and phonons coupled dynamics in Carbon Nanotubes

Vittorio Rispoli (*)

Abstract. A model for electrons transport properties in Carbon Nanotubes is introduced, including also the effects of the coupling of electrons with optical phonons. The derived equations form a system of bi-dimensional hyperbolic conservation laws with collision terms on the right hand side. The system is solved by a method of line scheme, with WENO reconstruction and a TVD Runge-Kutta scheme for time integration.
(MSC: 65M08, 65L99, 35L50. Keywords: Electrons transport, WENO, TVD Runge-Kutta.)

1 Introduction

Among all nanosized components for nanotechnological innovations, a major role is played by *Carbon NanoTubes* (CNTs) because of their great physical properties; they are tubes made of carbon atoms whose surface is only one carbon atom thick and formed by a lattice of carbon atoms arranged in a hexagonal lattice [1].

Carbon Nanotubes remarkable electrical properties make them, in many cases, the best candidates for innovative electronic applications. We will consider a very accurate model for the study of SWCNTs electrical properties: we will be able to numerically simulate, thanks to a deterministic solver, the dynamics of the electrons inside the tube generated by an applied bias and, thus, to compute the current.

A SWCNT diameter d is usually a few nanometers while its length L can go from a few hundreds nanometers up to some micron. For this reason ($d \ll L$) CNTs can be considered as one-dimensional objects. Depending on how the hexagonal lattice of carbon atoms is arranged around the tube, SWCNTs show different physical, say mechanical or electrical, behaviors.

Thinking of a SWCNT as a rolled-up graphene sheet [1], its geometrical structure can be easily described. From a crystallographic point of view, it is characterized by the wrapping vector C_h (also called *chirality* vector); this is the vector determining the folding of the graphene sheet. Chirality vector C_h connects two equivalent atoms (two atoms in

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: rispoli@math.unipd.it . Seminar held on 16 February 2011.

the graphene lattice that will coincide when the tube is rolled-up) and, given two lattice vectors a_1 and a_2 as in Figure 1, it can be written as $C_h = n\,a_1 + m\,a_2 \equiv (n, m)$.

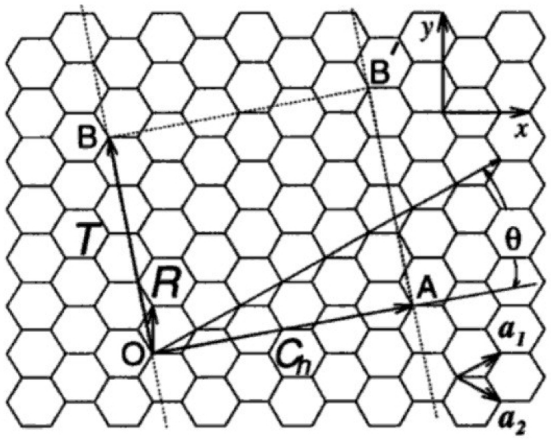


Figure 1: Geometric characterization of a CNT on a graphene sheet.

Given the chiral vector, it is possible to characterize all the geometric and also symmetry properties of a SWCNT [1].

Early theoretical calculations showed that the electronic properties of carbon nanotubes are very sensitive to their geometric structure [1]. Electronic properties can be derived from those of graphene; although graphene is a semi-metal, theory has predicted that carbon nanotubes can be metals or semiconductors with an energy gap that depends on the tube diameter and helicity, i.e., on the indices (n, m). This can be simply understood within a *zone-folding* picture by combining analytic results on the electronic structure of graphene with the requirements that the wave functions in the tubes must satisfy the proper boundary conditions around the tube circumference. This approach is made relatively simple in nanotubes because of the special shape of the graphene Fermi surface.

When forming a tube, owing to the periodic boundary conditions imposed in the circumferential direction, only a certain set of momentum vectors in the graphene Brillouin Zone (BZ) are allowed; the allowed set depends on the diameter and helicity of the tube. Within the zone-folding approximation, the general rules for the electronic behavior (whether they are metallic or not) of single-wall carbon nanotubes are as follows: a nanotube defined by the (n, m) indices will be *metallic* if $n - m \equiv 0 \pmod 3$ or *semiconductor* if $n - m \not\equiv 0 \pmod 3$. Consequently, most carbon nanotubes are semiconductors and only a fraction $1/3$ are metallic or semi-metallic (i.e. zero-gap semi-conductors).

In the first studies regarding nanotubes electrical behavior, nanotubes were considered as one-dimensional quantum wires with ballistic electron transport. However, when considering high-field transport measurements, current's magnitude found in real experiments had lower values than the predicted ones. The reason is that the scattering of electrons

with optical phonons destroys the ballistic behavior [2] and the interaction of electrons with phonons lowers current values significantly; indeed, the simulation of the generation of optical phonons during high-field electron transport, which can be directly detected by Raman scattering experiments, is essential to understand the reduction of the conductivity at high fields.

In the past, the high-field transport in metallic SWCNTs was studied either at a macroscopic level or by solving the semi-classical Boltzmann Equation (BE), as in [2]. In the latter case, the dynamics of electrons was treated in a kinetic way, while phonons where kept in equilibrium at a fixed lattice temperature. In order to model the effect of hot phonons on the distribution of electrons it was necessary to introduce a kinetic model for both electrons and optical phonons and, thanks to this model, it was possible to investigate the transient behavior of interacting electrons and phonons [3].

Our work find its placement in this context: starting from previous successful simulations, we gave a deeper description of some of the parameters present in the physical model. We considered a self consistent computation of electrons-phonons coupling factors and found great agreement with experimental data.

2 Transport Model

We are interested in investigating metallic SWCNTs, in particular those for which $n = m$ usually called armchair nanotubes. Electrical properties of SWCNTs arise from the confinement of electrons, which allows motion in only two directions and from the requirements for energy and momentum conservation. These constraints lead to a reduced phase space for scattering processes.

The allowed electronic states are characterized by two equivalent points K and $K' = 2K$ in the reciprocal space. In this case, the electronic energies are well approximated by the linear dispersion relations:

$$\varepsilon_i(k) = \hbar v_i \, k, \qquad i = 1, 2,$$

where $v_1 = +v_F$ and $v_2 = -v_F$ are the positive and negative Fermi velocities, $\hbar$ denotes the reduced Plank constant and k stands for the electron momentum along the tube axis [4].

Electrons in corresponding states according to K and K' can be considered as equivalent in our transport model. For this reason and since electrons can move in only two directions, it is sufficient to introduce only two distribution functions $f_i = f_i(t, x, \varepsilon)$ for right ($i = 1$) and left ($i = 2$) moving electrons; distributions f_i depend on time t, position x along the tube axis and electric energy density $\varepsilon = \varepsilon_i(k)$. We assume the tube diameter d is $1\,nm < d < 3\,nm$; this is the validity range for diameter values and allows us to neglect higher energy sub-bands for electrons with energies < 0.5 eV.

Boltzmann equations governing the evolution of the distribution functions f_i are:

$$\frac{\partial f_i}{\partial t} + v_i \, \frac{\partial f_i}{\partial x} - e_0 v_F E \, \frac{\partial f_i}{\partial \varepsilon} = \mathcal{C}_i, \qquad i = 1, 2, \tag{1}$$

where E is the electric field along the axis and e_0 is the electron charge (considering the negative value).

Collision operators for electrons are defined by:

$$\mathcal{C}_i = \mathcal{C}_i^{ac} + \sum_{\eta=1}^{3} \mathcal{C}_i^{\eta}, \tag{2}$$

where

$$\mathcal{C}_i^{ac} = \frac{v_F}{l_{ac}}(f_j - f_i), \quad j \neq i,$$

models interactions of electrons with acoustic phonons and

$$\begin{aligned}\mathcal{C}_i^{\eta} &= \gamma_\eta \Big\{ g_\eta(q_i^-) f_j^- (1 - f_i) + \left[g_\eta(q_i^+) + 1 \right] f_j^+ (1 - f_i) \\ &\quad - g_\eta(q_i^+) f_i (1 - f_j^+) - \left[g_\eta(q_i^-) + 1 \right] f_i (1 - f_j^+) \Big\}\end{aligned}$$

for $\eta = 1, 2, 3$ model back ($\eta = 1, 2$) and forward ($\eta = 3$) scattering with phonons. In the above formula γ_η denotes electron-phonon coupling (EPC) constants. We used the abbreviations $f_i = f_i(t, x, \varepsilon)$ and $f_i^\pm = f_i(t, x, \varepsilon \pm \hbar\omega_\eta)$, where $f_i^\pm$ model the emission or absorption of a phonon energy quantum $\hbar\omega_\eta$.

Modes $\eta = 1, 2, 3$ refer, respectively, to K-phonons, longitudinal optical Γ-phonons and transverse optical Γ-phonons. Further, for $\eta = 1, 2$, $q_i^\pm = \mp(2\varepsilon \pm \hbar\omega_\eta)/\hbar v_i$, while for $\eta = 3$, $q_i^+ = q_i^- = q_i = \omega_3/v_i$. Constant l_{ac} stands for the acoustic mean free path (MFP); electron scattering at impurities can be taken into account by choosing the elastic MFP $l_e = (1/l_{ac} + 1/l_{im})^{-1}$ instead of l_{ac}.

The time evolution of phonons distribution functions $g_\eta(t, x, q)$ is governed by the BEs:

$$\frac{\partial g_\eta}{\partial t} + \nu_\eta \frac{\partial g_\eta}{\partial x} = \mathcal{D}_\eta, \quad \eta = 1, 2, 3, \tag{3}$$

where q is the one-dimensional wave vector and ν_η represent phonons velocities.

Collision operators for phonons are:

$$\mathcal{D}_\eta = \mathcal{D}_\eta^{ep} + \mathcal{D}_\eta^{pp}. \tag{4}$$

The first term takes into account electron-phonon interaction. For $\eta = 1, 2$ (back scattering) it has the following form:

$$\begin{aligned}\mathcal{D}_\eta^{ep} &= 2 \sum_{i=1}^{2} \gamma_\eta \{ (g_\eta + 1) f_i(\varepsilon_i^+)(1 - f_j(\varepsilon_i^-)) \\ &\quad - g_\eta \, f_j(\varepsilon_i^-)(1 - f_i(\varepsilon_i^+)) \}, \quad j \neq i,\end{aligned}$$

where, $f_i(\varepsilon_i^\pm) = f_i(t, x, \varepsilon_i^\pm)$, with $\varepsilon_i^\pm = \hbar(v_i q \pm \omega_\eta)/2$.
For $\eta = 3$ (forward scattering), electron-phonon collision operator reads

$$\begin{aligned}\mathcal{D}_3^{ep} &= \gamma_3 \sum_{i=1}^{2} J_i \, \delta_{q,q_i} \int_{\mathbb{R}} \big\{ (g_3 + 1) f_i(\varepsilon) \left[1 - f_i(\varepsilon^-) \right] \\ &\quad - g_3 f_i(\varepsilon^-) \left[1 - f_i(\varepsilon) \right] \big\} \, d\varepsilon,\end{aligned}$$

with $\varepsilon^- = \varepsilon - \hbar\omega_3$ and $J_i = 4L/(hv_F)$ denoting the density of states for electrons of type i with respect to the tube length L. The Kronecker δ_{q,q_i} in the collision operator reflects the fact that only phonons with the wave vectors $q_i = \omega_3/v_i$ are emitted and absorbed by forward scattering of electrons. Phonon-phonon interactions are modeled by

$$\mathcal{D}_\eta^{pp} = -\frac{1}{\tau_\eta}\left[g_\eta(t,x,q) - g_\eta^0\right],$$

where τ_η denotes the relaxation time and g_η^0 is the Bose-Einstein distribution at a fixed temperature T; k_B is the Boltzmann constant.

The introduced BEs (1) and (3) represent a kinetic transport model which includes both the dynamics of electrons and optical phonons. Together they form a system of hyperbolic conservation laws with source terms at the right hand side (conservation laws with source terms are usually called Balance Laws). Computing the general solution of this system is a very hard task since most of the difficulties one could encounter in solving this type of problems are present in this case: multidimensionality (excluding the time variable, the phase-space is bi-dimensional), coupled equations and the presence of source terms. We will present now the numerical setting in which we will solve the system and the obtained results.

3 Numerical experiments

For the numerical approximation of our system, first we choose a fixed uniform discretization for the phase-space variables x, ε and q. For a $L\,nm$ long SWCNT, we define $\Delta x = L/N_x$, where N_x is the chosen number of grid points; in our computations, we considered variable tube lengths: $L = 150\,nm$, $L = 300\,nm$ and $L = 600\,nm$. For all calculations, we assumed a tube's diameter $d = 2\ nm$.

For the ε variable it is necessary to make a different choice: the discretization length $\Delta\varepsilon$ of the energy variable is chosen so that the phonon energies $\hbar\omega_\eta$ are integer multiples of $\Delta\varepsilon$ for $\eta = 1, 2, 3$, which means:

$$\hbar\omega_\eta = \sigma_\eta \Delta\varepsilon,$$

with $\sigma_\eta \in \mathbb{N}$. The energy grid is then determined by the values $\varepsilon_n = -\widehat{\varepsilon} + n\Delta\varepsilon$ for $n = 0, \ldots, N_\varepsilon$, where N_ε is the number of energy grid points, with the maximal energy given by

$$\widehat{\varepsilon} = \Delta\varepsilon \frac{N_\varepsilon}{2}.$$

From these electrons energy grid points, we define the discretization for the wave vector of the phonon modes $\eta = 1, 2, 3$ in the following way:

$$\Delta q = \frac{2\Delta\varepsilon}{\hbar v_F},$$

which gives the grid points

$$q_m^\eta = -\widehat{q}^\eta + m\Delta q, \qquad \text{for } m = 1, \ldots, N_q,$$

where $N_q = N_\varepsilon - \sigma_\eta$ and

$$\widehat{q}^\eta = \Delta q \frac{N_q}{2} = \Delta q \frac{N_\varepsilon - \sigma_\eta}{2}.$$

This choice for the discretization of the ε and q variables, ensures that the energy and momentum relations

$$\varepsilon(k') = \varepsilon(k) \pm \hbar\omega_\eta \quad \text{and} \quad k' = k \pm q$$

are satisfied at the discrete level in each individual back-scattering process. Hence, collision operators $\mathcal{C}_i$ and $\mathcal{D}_\eta$ can be evaluated *exactly* in terms of the discretized distribution functions $f_i(t, x, \varepsilon_n)$ and $g_\eta(t, x, q_m^\eta)$. This is a major advantage since no approximations (for example, of extrapolation type) are needed to compute the collision operators. For the values of all other parameters refer to [3].

The determination of electrons transport in a SWCNT, according to our model, is a low dimensional problem, in contrast to, for example, classical semiconductor devices simulations. Distribution functions $f_i(t, x, \varepsilon)$ and $g_\eta(t, x, q)$ depend both on two phase-space variables, therefore the kinetic equations can be solved very efficiently by means of a deterministic solver. For the solution of the BEs we proceed as follows.

We adopted a method of lines approach: when dealing with hyperbolic differential equations, an efficient way to numerically solve them is to approximate first the derivatives of the phase space variables (which are x and ε in this case) and then integrate in time the resulting system of ODEs.

The idea of the method of lines is the following: given $u = u(t, x)$ and a hyperbolic conservation law $u_t + f_x(u) = 0$, if we approximate $f_x(u) \approx L(u)$, then we obtain $u_t = -L(u)$ and the time evolution can then be simulated by ODE solvers. In the case of a BL $u_t + f_x(u) = C(u)$, one of the possible strategies is to integrate in time the resulting ODE with right hand side given by $u_t = -L(u) + C(u)$.

Both the approximation of the derivatives, which is usually called *reconstruction*, and the integration in time have to be computed using numerical methods specifically designed for hyperbolic laws.

The left hand side of equations (1) and (3) form a system of bi-dimensional hyperbolic conservation laws:

$$\begin{cases} \partial_t f_i + v_i\, \partial_x f_i + e_0 v_i E\, \partial_\varepsilon f_i = 0, & i = 1, 2 \\ \partial_t g_\eta + \nu_\eta\, \partial_x g_\eta = 0, & \eta = 1, 2, 3 \end{cases} \tag{5}$$

and specific numerical methods available in literature can be used in order to solve the general system. An efficient way to approximate derivatives with respect to x and ε is to use a conservative high-order scheme, which could be a finite differences or a finite volumes scheme. The derivative $\partial_\varepsilon f_i$, for instance, can be approximated by

$$\partial_\varepsilon f_i(t, x, \varepsilon_n) = \frac{1}{\Delta\varepsilon} \left[\widehat{f}_{i,n+\frac{1}{2}}(t, x) - \widehat{f}_{i,n-\frac{1}{2}}(t, x) \right] \tag{6}$$

where $\widehat{f}_{i,n+1/2}$ is the numerical flux function, which is needed to combine in the proper way the reconstruction of the cell averages of the function. To obtain good accuracy and to avoid unphysical solutions, it is advisable to use high-order methods such as, e.g., high-order versions of *Weighted Essentially Non Oscillatory* (WENO) schemes [5]. WENO

schemes are improved and more accurate versions of *Essentially Non Oscillatory* (ENO) schemes.

Once the approximation of the derivatives is obtained, we can integrate in time the obtained system of ODEs, where at the right side we have the sum of the reconstruction term plus the collision operator. Such system of ODEs has to be solved by proper schemes, suitable for Conservation Laws. A class of schemes usually adopted for the time integration is that of Total Variation Diminishing (TVD) methods: they satisfy the requirement that that the total variation of the numerical solution does not increase. This property is fundamental to compute the solution of hyperbolic equations, because of the discontinuities that (generically) arise, even from smooth initial data, during the time evolution. In our computations we used the explicit, optimal, third-order version of a TVD Runge-Kutta type scheme [6]. It is optimal in the sense that it is third-order accurate in the presence of a discontinuity while it gain fifth-order accuracy where the function is smooth.

To compute the solution of the system, we have to impose initial conditions; since we are dealing with a hyperbolic problem and the domain is limited in space, $x \in [0, L]$, we also need inflow and outflow boundary conditions; these are generally called Initial-Boundary conditions or Initial-Values Problems (IVP).

Regarding boundary conditions, it is necessary to assign inflow conditions both at the left contact for right moving particles and at the right contact for left traveling particles; this means we impose:

$$\begin{aligned} f_1(t,\,0,\varepsilon) &= t_1^2\, f_0(\varepsilon) + (1-t_1^2) f_2(t,0,\varepsilon) \\ f_2(t,L,\varepsilon) &= t_2^2\, f_0(-\varepsilon) + (1-t_2^2) f_1(t,L,-\varepsilon) \end{aligned}$$

for all time $t > 0$ for f_1 and f_2, where $f^0(\varepsilon) = 1/[1 + e^{(\varepsilon/(k_B T))}]$ is the Fermi-Dirac distribution. Regarding g_η's, we have

$$g_\eta(t, 0, q) = g_\eta^0, \qquad \forall t > 0,$$

for η such that $\nu_\eta > 0$. Here $g_\eta^0 = 1/[e^{\hbar\omega_\eta/(k_B T)} - 1]$ is the Bose-Einstein distribution at a fixed lattice temperature T. Since $\nu_3 = 0$, for g_3 we only need initial conditions. On respective opposite boundaries, values are self-consistently determined by the time evolution of the system.

The aim of our simulations is to compute the system response to the application of an electric potential V, i.e. the value of the generated current $I = I(V)$.

We defined the (mean) current at time t as:

$$I(t) = \frac{1}{N_x} \sum_{i=1}^{N_x} J(t, x_i), \tag{7}$$

where

$$I(t,x) = \frac{4e_0}{h} \int_{\mathbb{R}} f_2(t,x,\varepsilon) - f_1(t,x,\varepsilon)\, d\varepsilon\,. \tag{8}$$

We used the resulting computed current as stopping criteria for our simulations: $|I(t + \Delta t) - I(t)| < \varepsilon_I \; I(t)$. Computations stopped when the previous relation was satisfied for $\varepsilon_I \; = \; 10^{-3}$.

Until phonons are assumed in thermal equilibrium (thermalized) at room temperature, which means $n_{q\eta} \approx 0$, the contribution of forward scattering has to be neglected (as was the case, e.g., for the model considered in [2]. Thus, to compare with experiments, one could only consider backscattering and obtain a simple scaling between scattering length and diameter: $l = 65 \; d$.

Already in [7], authors have pointed out that the assumption of thermalized phonon does not hold: only a significant phonon occupation n can explain the small value of the measured scattering length l. With a high phonon occupation, both phonon emission and absorption processes are equally relevant, so to take into account more complex scattering processes a deeper models should be considered.

A better approximation was used in [4]; their model took into account the role of the time evolution of one more phonon mode, related to forward scattering, and also considered different values for scattering lengths. The quantities l_η, for $\eta = 1, 2, 3$, determining the electron-phonon coupling coefficients for the different phonon modes were taken as constants having different values depending on phonon modes: for any phonon mode $\eta = 1, 2, 3$, the following relations were considered: $l_1 = 92.0 \; d$ and $l_2 = l_3 = 225.6 \; d$.

What was proposed in [7], is that it is possible to estimate the scattering lengths depending on phonon occupation n, assuming the latter is independent of q and η. The scattering lengths are then obtained using a fixed phonon occupation n_0:

$$l = \frac{65 \, d}{(1 + n_0)}, \tag{9}$$

with n_0 in the $2.7 \sim 5$ range to reconcile the scattering lengths derived from the computed and the measured EPCs. Results obtained with these parameters are consistent with the observation that high-bias saturation currents in SWCNTs on a substrate are significantly higher than those in suspended SWCNTs. Indeed, the effective temperature of optical phonons in suspended SWCNTs is expected to be higher due to the absence of a thermally conductive substrate for heat sinking, lowering current values significantly.

What we did was to consider non constant values for EPC values. We considered a relation similar to (9) but assuming a varying phonon occupation n according to the computed mean phonons distributions $\langle g_\eta \rangle$: $n = n(\langle g_\eta \rangle)$.

What we observed from the computed distributions was a low transverse optical phonons population (the one given by g_3) having, anyway, very high peaks near the boundaries in the neighborhood of the scattering momenta q_i^3. This led us to search for another improvement, considering different values of scattering lengths for longitudinal optical ($\eta = 2$) and transversal optical ($\eta = 3$) phonon modes and considering also the dependence on q. This is also theoretically justified by the fact that distribution g_3 takes very small values almost everywhere inside the tube but has very high picks in the neighborhood of $q_i^3 = \hbar\omega_3/\hbar v_F$; it makes sense, thus, to consider longer paths for longitudinal phonons with respect to those for transverse phonons; we included, for transverse

phonons, the dependence on q. In this context, we are still assuming n depends on η and q but not on x. With these assumptions, we found great agreement between the computed results and the experimental data; simulation results are shown in Figure 2, compared with experimental data. In Figure 2 we show the obtained results for Current-Voltage characteristics for the proposed model.

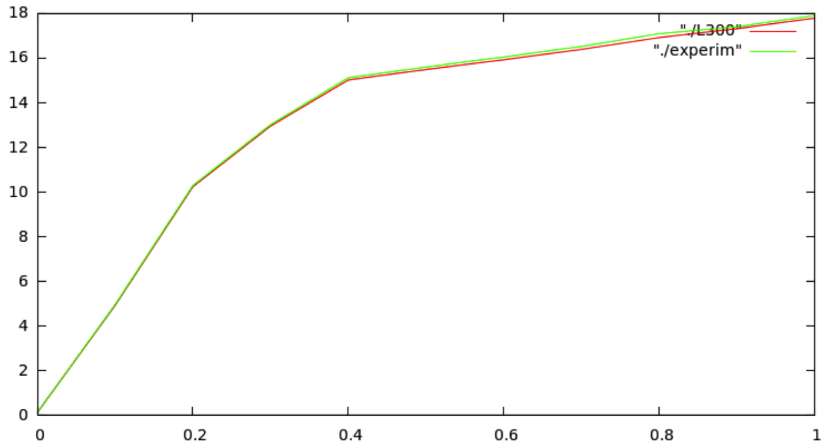


Figure 2: IV curves for computed and experimental data.

In the low-bias regime results are similar to those with ballistic transport, i.e. a linear response is observed, but when the strong emission of optical phonons starts, calculation based on equilibrium phonons significantly overestimate the current [3]. We observe that the sudden drop of conductance at $V = 0.2\,V$ is accurately reproduced by the performed transport calculations.

Our study concerns electronic properties of carbon nanotubes. To model the evolution of electrons and phonons distributions during the current generation, a system of kinetic Boltzmann equations is given. We *present* a more accurate description of the physical model: we did not assume constant values for the lengths of the scattering between electrons and phonons, as was in previous works, but compute them in a self consistent way. Obtained results show great agreement with the available experimental data. In our approximation, lengths are functions of phonon distributions, depending in particular on the momentum variable; more accurate characterization could be obtained depending, for example, also on the space variable.

Attempts were made to simulate, also, electrical behavior of large diameter nanotubes (i.e. $d > 6\,nm$) but results show a different model should be considered in this case. Large diameter nanotubes are often used in practical applications so this will be a very interesting subject for future work.

References

[1] R. Saito, G. Dresselhaus, and M. Dresselhaus, "Physical properties of carbon nanotubes". Imp. College Press, 1998.

[2] C. Kane, C. Dekker, and Z. Yao, *High-field electrical transport in single-wall carbon nanotubes.* Phys. Rev. Lett. 84/13 (2000), 2941–2944.

[3] C. Auer, F. Schürrer, and C. Ertler, *Hot phonon effects on the high-field transport in metallic carbon nanotubes.* Phys. Rev. B 74 (2006), 165409–165419.

[4] C. Auer, F. Schürrer, and C. Ertler, *Deterministic solution of Boltzmann equations governing the dynamics of electrons and phonons in cabon nanotubes.* In *Applied and Industrial Mathematics in Italy, II*, World Scientific, Singapore (2006), 89–100.

[5] C. Shu and G. Jiang, *Efficient implementation of weighted ENO schemes.* J. Comp. Phys. 126 no. 130 (1996), 202–228.

[6] C. Shu and S. Gottlieb, *Total variation diminishing Runge-Kutta schemes.* Math. Comput. 67 no. 227 (1998), 73–85.

[7] M. Lazzeri, S. Piscanec, F. Mauri, A. Ferrari, and J. Robertson, *Electron transport and hot phonons in carbon nanotubes.* Phys. Rev. Lett. 85 (2005), 236802–236805.

From Shafarevich's conjecture to finite flat group schemes

Hendrik Verhoek (*)

1 Introduction

The goal of the talk was to explain the theorem of Fontaine and Abrashkin, giving an affirmative answer to Shafarevich's conjecture posed at the ICM in 1962, that says that there do not exist non-zero abelian varieties over $\mathbb{Q}$ that have good reduction everywhere.

Theorem 1.1 [Abrashkin, Fontaine] *There do not exist non-zero abelian varieties over* $\mathbb{Q}$, $\mathbb{Q}(i)$, $\mathbb{Q}(\zeta_3)$, $\mathbb{Q}(\zeta_5)$, $\mathbb{Q}(\sqrt{\pm 2})$ *and* $\mathbb{Q}(\sqrt{7})$ *that have good reduction everywhere.*

In the first sections I will explain the statement of the theorem, what abelian varieties are and what it means that abelian varieties have good reduction everywhere. Then I will give an indication of the proof, though with some definitions skipped due to space limitations.

2 Abelian varieties

We start with the concept of an algebraic variety, which one can think of as a set of solutions to some polynomial equations with coefficients in some fixed field k. Depending on these polynomial equations, these varieties can have a very rich structure. Sometimes, besides being a set of solutions to equations, the solutions form a group: one can add solutions, substract them to get another solution and there is a trivial solution (the identity). A variety that admits such a group structure is called a group variety.

One can also endow algebraic varieties with topologies. Often one endows it with the so-called Zariski topology which reflects well the algebraic nature of the variety. With such a topology it makes sense to talk about whether a variety is connected. It makes sense to talk about the connected component of the identity when the variety is a group variety. Also there is a notion of completeness for varieties whose definition goes beyond what I want to say in this talk.

(*) Università Roma 2, Rome, Italy. Web page of the author: `http://www.mat.uniroma2.it/~verhoek/` E-mail: `hendrikverhoek@gmail.com` . Seminar held on 23 February 2011.

Definition 2.1 An abelian variety is a complete connected group variety over some field k.

A very important property of abelian varieties is that they are non-singular. Remember that a variety is a set of solutions of equations. If the number of variables occurring in these equations is n, then we can embed the variety in n-dimensional space, where space depends on the actual variety and the field k. A point on a variety is called singular if the dimension of the tangent space at that point is larger than the codimension of the variety embedded in this space. In the talk we saw a few examples of singular points for a curve, namely cuspidal points and nodal points.

All points on an abelian variety are non-singular. An abelian variety of dimension 1 is called an elliptic curve and it is for such an abelian variety that we will give an example, since it is the set of solutions of only one equation.

Example 2.2 Let $k = \mathbb{Q}$, the rational numbers. An equation of the form $y^2 = x^3 + ax + b$, or better yet its projectivization $zy^2 = x^3 + axz^2 + bz^3$, defines a curve in the projective plane $\mathbb{P}^2$. If $-16(4a^3 + 27b^2) \neq 0$ it is non-singular and the curve is an elliptic curve. It is a group variety with identity point $(x : y : z) = (0 : 1 : 0)$.

3 Good and bad reduction

What does it mean that an abelian variety over $\mathbb{Q}$ has good or bad reduction at a prime number p? An abelian variety A over $\mathbb{Q}$ is a variety and hence corresponds to solutions to some equations with coefficients in $\mathbb{Q}$. By transforming these equations one can make them have coefficients in the integers $\mathbb{Z}$. Now for each prime p we can reduce mod p and consider the solutions of these equations in $\mathbb{F}_p$. These are varieties again, but not necessarily abelian varieties. Such a family of varieties, ranging over all primes p, is a model for A, often denoted by $\mathcal{A}$. When we reduce the equation with coefficients in $\mathbb{Z}$ modulo the prime p, we're considering what is called the fiber of A at p. The variety A, which remember is over $\mathbb{Q}$, is called the generic fiber of $\mathcal{A}$. There are special nice models for A, for example a Néron model. A Néron model $\mathcal{A}$ has the nice properties that it is non-singular in every fiber, and this model has a global group law, extending the one in the generic fiber. We denote from now on by $\mathcal{A}$ the Néron model of A. Now we're able to define what good and bad reduction mean:

Definition 3.1 The abelian variety A over $\mathbb{Q}$ has good reduction at p when the fiber at p of $\mathcal{A}$ is again an abelian variety. If this is not the case A is said to have bad reduction at p.

Although the above definition is mathematically correct, a better way to think about a prime p of bad reduction is that for all 'proper' models of A, there are singular points in the fiber at p. Proper is similar to the notion of completeness mentioned above, and basically means that such a model has no points missing in the fiber at p, the obtained variety in the fiber at p is connected. A Néron model over need not be proper.

Here's an explicit example of good and bad reduction, again for the case of an elliptic

curve:

Example 3.2 Take $E : y^2 = x^3 + 2$. This is an equation already having integral coefficients. Hence we have a model of E over $\mathbb{Z}$ and it is 'proper'. The primes of bad reduction are 2 and 3. It is easy to see that the fiber at 2 of this model has bad reduction: it gives a singular variety over the field $\mathbb{F}_2$ since there is a singular point, a cusp. On the other hand, this equation modulo the prime 5 has no singular solutions and is a prime of good reduction.

It is always true that an abelian variety over whatever field has good reduction at almost all primes, or in other words, there are only finitely many primes at which the abelian variety has bad reduction. To ask that an abelian variety has good reduction at all primes is a very strong condition. In fact, it is so strong that Shafarevich conjectured they don't exist over $\mathbb{Q}$. Shafarevich's conjecture has been proven independently by both Abrashkin [1] and Fontaine [2] in the mid eighties.

Theorem 3.3 [Abrashkin, Fontaine] *There do not exist non-zero abelian varieties over* $\mathbb{Q}$, $\mathbb{Q}(i)$, $\mathbb{Q}(\zeta_3)$, $\mathbb{Q}(\zeta_5)$, $\mathbb{Q}(\sqrt{\pm 2})$ *and* $\mathbb{Q}(\sqrt{7})$ *that have good reduction everywhere.*

4 Proving the Theorem

This section is more complicated than the previous sections. The previous sections suffice to understand the statement of the Theorem. This one explains the proof.

In order to prove the Theorem 3.3, we introduced so called group schemes. They are a generalization of groups and we'll see them as representable functors from R-algebras (where R is any ring) to groups. Due to space limitations this is not the right place to give a full introduction of these objects. Therefore we restrict to mentioning that group schemes are called commutative if the image of the functor are commutative groups. It is finite flat if the algebra that represents the functor is finite and flat over R. In the example below we will see two instances of group schemes: the constant group scheme and the group scheme of the n-th roots of units. We briefly recall what they are.

To describe the constant group schemes, let Γ be any group in the regular, abstract sense that we're all used to. The constant group scheme C_Γ over R is the functor that associates to any connected R-algebra S the group Γ. The constant group scheme associated to the cyclic group of order n is denoted by $\mathbb{Z}/n\mathbb{Z}$. The group scheme of the n-th roots of units associates to any R-algebra S the group of n-th roots of units of S, that is, $\{s \in S : s^n = 1\}$. This group scheme is denoted by μ_n.

We indicate how to prove Theorem 3.3 in steps. Let K be a number field, a finite extension of $\mathbb{Q}$, with ring of integers O_K. Let S be a finite set of prime ideals in O_K. Denote by O_S the ring of S-integers of K. Let ℓ be a rational prime such that none of the primes in S divides ℓ. After each step we say what happens when $K = \mathbb{Q}$ to answer Shafarevich's conjecture.

Definition 4.1 Let C be a subcategory of the category of finite flat commutative group schemes over O_S of ℓ-power order. We suppose that C is closed under taking products,

subquotients and Cartier duality.

Example 4.2 Let $K = \mathbb{Q}, O_K = \mathbb{Z}$ and $S = \emptyset$. We take $\ell = 2$ and we let $C :=$ category of all finite flat commutative group schemes over $\mathbb{Z}$ of 2-power order.

Next we will find so-called simple objects in C. The generic fiber of a finite flat commutative group scheme J over O_S is a group scheme over K. The group scheme J_K is étale and just an abelian group $J(\overline{K})$ together with the Galois action

$$\rho_J : G_K \to \mathrm{Aut}(J(\overline{K})).$$

The representation ρ_J factors through a finite Galois extension $K(J)/K$. By considering the generic fiber J_K we get quite some information about the group scheme J. A group scheme is called simple if it has no non-trivial closed flat subgroup schemes. A simple object in C is a simple group scheme. One can show that the representation ρ_J of a simple group scheme J is irreducible and that simple objects are annihilated by ℓ. We find all simple objects by considering the maximal ℓ-torsion extension of C which is the compositum of fields $K(J)$ where the J are in C and are annihilated by ℓ. The extension L/K need not be finite in general.

Example 4.3 Group schemes annihilated by 2 are for example $\mu_2, \mathbb{Z}/2\mathbb{Z}$ and an extension of $\mathbb{Z}/2Z$ by μ_2:

$$0 \to \mu_2 \to G \to \mathbb{Z}/2\mathbb{Z} \to 0,$$

where G has an associated (Hopf) algebra

$$\prod_{i=1}^{2} \mathbb{Z}[X_i]/(X_i^2 - (-1)^i).$$

We leave out the description of the group law.

For suitable categories C one can prove that L is finite. Suitable means: small ℓ and low ramification at the primes in S. If L is finite and small, then we can determine all simple objects of C. For this we use the following result of Fontaine:

Theorem 4.4 [Fontaine] *Let e be the absolute ramification index of $F/\mathbb{Q}$ at ℓ, let J be a finite flat commutative group scheme over O_F (the ring of integers of F) annihilated by ℓ. Let Δ be the discriminant of the extension $F(J)/F$. Then $\Delta^{\frac{1}{[F(J):\mathbb{Q}]}} < \Delta_{F/\mathbb{Q}}^{\frac{1}{[F(J):\mathbb{Q}]}} \ell^{e(1+\frac{1}{\ell-1})}$.*

Example 4.5 In our case $\ell = 2$, $F = \mathbb{Q}$ and $e = 1$. This means that the discriminant Δ of $L/\mathbb{Q}$ satisfies:

$$\Delta^{1/[L:\mathbb{Q}]} < 4.$$

Using the tables of Odlyzko this implies : $[L : \mathbb{Q}] \leq 4$. We can prove that $L = \mathbb{Q}(i)$ and then that all simple group schemes in C are isomorphic to either μ_2 or $\mathbb{Z}/2\mathbb{Z}$.

Once we determined all simple objects using the maximal ℓ-torsion extension, we want to compute extensions between them. The goal is to prove that the category C satisfies

the following two conditions, in which the notion étale occurs. We say that a group scheme $J \in C$ is étale if $K(J)/K$ is unramified at ℓ:

condition (1) : For all simple non-étale group schemes T in C and all simple étale group schemes E in C, the group $\mathrm{Ext}^1_C(T, E)$ is trivial.

condition (2) : The compositum F of all $K(E)$, where E is a simple étale group scheme in C, is finite and the maximal abelian extension R of F that is unramified outside S and at most tamely ramified at primes over S, is a cyclic extension.

Example 4.6 In fact, condition 1 holds since any extension

$$0 \to \mathbb{Z}/2\mathbb{Z} \to G \to \mu_2 \to 0$$

splits. Condition 2 is verified since the class number of $\mathbb{Q}$ is one.

We omit the proof of the following theorem:

Theorem 4.7 *Let A be an abelian variety such that $A[\ell^n]$ are objects in C for all n. If conditions* (1) *and* (2) *hold for the category C, then $A[\ell]$ cannot be filtered by étale group schemes or group schemes of multiplicative type.*

Example 4.8 We saw that both conditions hold for C. Since A is supposed to have good reduction everywhere, the group scheme $A[2^n]$ is a finite flat commutative group scheme over $\mathbb{Z}$ of 2-power order. Hence $A[2^n]$ is an object in C. But C has as only simple objects the constant group scheme $\mathbb{Z}/2\mathbb{Z}$ which is étale, and μ_2 which is of multiplicative type. Contradiction. There are no abelian varieties over $\mathbb{Q}$ with good reduction everywhere.

References

[1] V. A. Abrashkin, *Galois modules of group schemes of period p over the ring of Witt vectors.* Izv. Akad. Nauk SSSR Ser. Mat., 51/4, (1987), 691–736, 910.

[2] Jean-Marc Fontaine, *Il n'y a pas de variété abélienne sur* $\mathbb{Z}$. Invent. Math. 81/3 (1985), 515–538.

Mean-variance optimisation problems in financial mathematics

Claudio Fontana (*)

Abstract. This short note surveys the main aspects of mean-variance portfolio optimisation, both from a mathematical and a financial point of view, in the context of classical financial economics and modern stochastic finance. We also present an abstract approach to mean-variance portfolio problems, allowing us to obtain general characterisations of optimal portfolios in a simple and model-independent way, under a minimal no-arbitrage condition.

1 Introduction

Many optimisation problems in applied mathematics are formulated in terms of *quadratic* optimality criteria, mainly due to their analytical tractability. In particular, in the context of portfolio optimisation, this has been successfully exploited, giving rise to an extensive literature dealing with *mean-variance portfolio optimisation*, beginning with the seminal work of Markowitz [10].

From a financial point of view, means and variances admit a natural interpretation in terms of *expected return* and *risk* and mean-variance portfolio optimisation essentially consists in maximising the expected return while minimising the risk. Clearly, in order to make this approach work, one needs a sufficient understanding of the financial market as well as proper mathematical tools to deal with the formulation and solution of mean-variance optimisation problems. In the course of the last decades, this interaction between mathematics and finance has not only led to the development of sound techniques for investment and asset pricing, but also to significant advances in mathematics, especially in stochastic analysis and related fields.

This short note aims at discussing some of the essential features of mean-variance portfolio optimisation problems in the context of classical financial economics as well as modern mathematical finance. Of course, we cannot properly survey the huge relevant literature and, hence, we limit ourselves to some simple and fundamental facts. We also briefly illustrate a general and unifying approach to mean-variance problems which has

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; and Politecnico di Milano, Dip. di Matematica "F. Brioschi", p.zza L. da Vinci 32, I-20133, Milano (Italy). E-mail: fontana@math.unipd.it . Seminar held on 9 March 2011.

been recently proposed in [5]. In a nutshell, this approach relies on an abstract description of the financial market and allows for general characterisations of mean-variance optimal portfolios, under the minimal no-arbitrage condition of *no approximate riskless profits in* L^2. At the same time, such an abstract approach allows to show that many of the classical properties of mean-variance optimal portfolios do not depend on the specific model under consideration but are natural outcomes of the mean-variance criteria themselves.

This note is structured as follows. Section 2 briefly surveys classical mean-variance portfolio selection theory. Section 3 moves to modern mathematical finance and deals with general mean-variance hedging and portfolio optimisation problems. Finally, Section 4 studies mean-variance optimisation problems in an abstract setting, allowing for general characterisations of optimal portfolios.

2 Classical mean-variance problems in financial economics

As mentioned in the introduction, mean-variance portfolio selection goes back to the seminal work of Markowitz [10]. In the traditional and simplest formulation, one considers a single-period model (i.e. $t = 0$ is *today* and $t = 1$ is *tomorrow*), where the returns on n risky assets are described by an $\mathbb{R}^n$-valued random vector R, with mean $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$. In this context, a *portfolio* is simply represented by a deterministic vector w in $\mathbb{R}^n$, with w_i denoting the proportion of wealth invested in the i-th asset. The classical Markowitz mean-variance portfolio selection problem is then formulated as follows, for $m \in \mathbb{R}$:

$$w'\Sigma w = \min! \qquad \text{over all } w \in \mathbb{R}^n \text{ s.t. } \mu' w = m \text{ and } \mathbf{1}' w = 1 \tag{1}$$

where $\mathbf{1} := (1, \dots, 1)' \in \mathbb{R}^n$. Problem (1) consists in minimising the variance of the random portfolio return $R'w$ given a constraint on its expected value $\mu' w = m$, for a fixed $m \in \mathbb{R}$. From a financial point of view, this amounts to minimising the risk given a constraint on the required expected return. Problem (1) can be explicitly solved by elementary linear algebra, leading to the following classical result (see e.g. [6, Chapter 3]).

Proposition 1 *For any $m \in \mathbb{R}$, the solution $w^*(m) \in \mathbb{R}^n$ to Problem (1) is explicitly given as follows:*

$$w^*(m) = p(m)\,\hat{w} + \big(1 - p(m)\big) w_{\mathrm{mv}}$$

where $\hat{w} := \frac{\Sigma^{-1}\mu}{\mathbf{1}'\Sigma^{-1}\mu}$, $w_{\mathrm{mv}} := \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}$ *and* $p(m) := \frac{m(\mu'\Sigma^{-1}\mathbf{1})(\mathbf{1}'\Sigma^{-1}\mathbf{1}) - (\mu'\Sigma^{-1}\mathbf{1})^2}{(\mu'\Sigma^{-1}\mu)(\mathbf{1}'\Sigma^{-1}\mathbf{1}) - (\mu'\Sigma^{-1}\mathbf{1})^2}$.

In particular, Proposition 1 shows that the solution $w^*(m)$ to Problem (1) is always given by a linear combination of the two fixed elements $\hat{w}$ and w_{mv} and only the amounts invested in them depend on the constraint m (*two-fund separation*). Furthermore, the element w_{mv} can be easily shown to be the *minimum-variance* strategy, i.e. the element which minimises $\mathrm{Var}[R'w] = w'\Sigma w$ over all $w \in \mathbb{R}^n$. By relying on Proposition 1, we can also compute $(m, \mathrm{Var}[R'w^*(m)])_{m\in\mathbb{R}}$, thus obtaining the so-called *mean-variance efficient frontier*, which represents in the mean-variance plane the performance of all portfolios which solve Problem (1) for different values of expected return.

Classical mean-variance portfolio selection has also been extended in several directions. For instance, we mention alternative formulations of Problem (1), with the variance being replaced by more refined and asymmetric measures of risk, as in [3], or models taking into account constraints or other restrictions on portfolio strategies, as in the book [11]. Also, multi-period discrete-time settings have been studied via recursive techniques, as in [7] and [16]. Mean-variance portfolio selection has also been applied to asset pricing, leading to the celebrated *CAPM* model, see e.g. Chapter 7 of [8]. In the next section we shall see how the basic mean-variance optimisation Problem (1) has evolved in the context of modern mathematical finance.

3 Mean-variance problems in stochastic finance

In the last two decades, mean-variance portfolio optimisation problems have also drawn the attention of researchers in the mathematical finance community. In a nutshell, modern mathematical finance deals with the *dynamic* modeling of financial quantities via stochastic processes, hence the name *stochastic finance*.

More formally, suppose that we are given a filtered probability space $\left(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, P\right)$ and an $\mathbb{R}^n$-valued *semimartingale* $S = (S_t)_{t\geq 0}$, representing the dynamic random evolution of the price of n risky assets. The activity of trading in the financial market is described by the concept of *trading strategy*, here represented by an $\mathbb{R}^n$-valued *predictable S-integrable* process $\theta = (\theta_t)_{t\geq 0} \in \Theta$, with Θ denoting the set of all *admissible self-financing* trading strategies satisfying suitable technical conditions. The gains generated by investing according to a strategy $\theta \in \Theta$ are then given by the *stochastic integral* $\int \theta\, dS = \left(\int_0^t \theta_u\, dS_u\right)_{t\geq 0}$. Consequently, the value at a fixed time horizon $T \in (0,\infty)$ of a portfolio starting from an initial capital $x \in \mathbb{R}$ and generated by the strategy $\theta \in \Theta$ is given by $V_T(x,\theta) := x + \int_0^T \theta_u\, dS_u$.

We can formulate the two following Problems, for some fixed $x \in \mathbb{R}$, $\alpha \in (0,\infty)$ and $H \in L^2$:

Problem (I) $E\left[V_T(x,\theta)\right] - \alpha \mathrm{Var}\left[V_T(x,\theta)\right] = \max!$ over all $\theta \in \Theta$

Problem (II) $E\left[\left|H - V_T(x,\theta)\right|^2\right] = \min!$ over all $\theta \in \Theta$

Problem (I) corresponds to the portfolio optimisation problem faced by an agent with mean-variance preferences and *risk-aversion* coefficient α. If we interpret $H \in L^2$ as the random value at time T of a *contingent claim* or *derivative*, Problem (II) consists in finding the portfolio which best approximates H in the L^2-norm. In the literature, Problem (II) has been called the *mean-variance hedging* problem.

For reasons of space, we do not attempt an overview of the extensive relevant literature, for which we refer to the survey papers [12], [14] and [15]. Mathematically, solving Problem (II) amounts to projecting the random variable H in L^2 onto the space of stochastic integrals $V_T(x,\Theta) := \{V_T(x,\theta) : \theta \in \Theta\}$. Hence, the existence of a solution to Problem (II) is equivalent to the closedness in L^2 of the space $V_T(x,\Theta)$. This question has been dealt with and answered in [2] and [1], thus showing that financial problems can indeed

motivate the development of deep and significant results in stochastic analysis. Clearly, being able to obtain an explicit characterisation of the solution to Problem (II) depends on the specific model under consideration and, in general, is rather difficult, requiring the application of martingale methods and/or stochastic optimal control, see e.g. the recent paper [9]. Furthermore, it is worth mentioning that Problem (I) can be solved by relying on the solution to Problem (II). This relation between Problem (I) and Problem (II) will also be exploited in the abstract approach outlined in next section.

4 A unifying and abstract approach to mean-variance problems

This section presents an abstract approach to mean-variance portfolio optimisation problems. Loosely speaking, the setting considered in this section lies on a middle ground between the classical approach outlined in Section 2 and the more sophisticated semimartingale setting of Section 3. On the one hand, the framework we are going to introduce is essentially a single-period model. On the other hand, we avoid any specific description of the underlying financial market and we only work with an abstract subspace of L^2 representing the set of all attainable final wealths. This allows us to obtain general model-independent characterisations of mean-variance optimal portfolios, together with their fundamental economic properties, and at the same time connect several seemingly different mean-variance optimisation problems. This section is based on [5], to which we refer the interested reader for full details. In a related context, see also [13] and [4, Chapter 1].

Let $(\Omega, \mathcal{F}, P)$ be a given probability space and let L^2 be the space of all real-valued square-integrable random variables, endowed with the usual scalar product $(X, Y) := E[XY]$. Let $\mathcal{G}$ be a given non-empty subset of L^2 and denote by $\mathcal{G}^\perp$ its orthogonal complement in L^2, i.e. $\mathcal{G}^\perp := \{X \in L^2 : (X, Y) = 0 \text{ for all } Y \in \mathcal{G}\}$. We also denote by $\overline{\mathcal{G}}$ the closure of $\mathcal{G}$ with respect to the L^2-norm. Finally, let B be a real-valued random variable in L^2 such that $B > 0$ P-a.s. The financial interpretation of this abstract setup is as follows. Fix a time horizon $T \in (0, \infty)$ and let $t = 0$ be the initial time. The set $\mathcal{G}$ represents the set of all undiscounted cumulated gains from trade (evaluated at time T), generated by self-financing trading strategies starting from zero initial capital. The element B represents the strictly positive value (at the final time T) of a *numeraire* asset, which can be interpreted as a *savings account*. The set $\mathcal{A} := \mathbb{R}B + \mathcal{G} = \{cB + g : c \in \mathbb{R}, g \in \mathcal{G}\}$ represents the set of all attainable undiscounted final wealths.

Let us now introduce the following standing Assumption.

Assumption 1 The two following conditions hold:

(a) $\mathcal{G}$ is a linear subspace of L^2;

(b) There are *no approximate riskless profits in* L^2, meaning that $\overline{\mathcal{G}}$ does not contain 1.

Part (a) of Assumption 1 amounts to considering a frictionless financial market without restrictions on trading. The condition $1 \notin \overline{\mathcal{G}}$ of *no approximate riskless profits in* L^2

represents an abstract and minimal no-arbitrage condition. Clearly, it can be equivalently formulated as $\mathbb{R} \cap \overline{\mathcal{G}} = \{0\}$.

Let us now consider four major mean-variance portfolio optimisation problems, here denoted as Problems (A)-(D) and formulated in the following abstract terms. We let $Y \in L^2$ represent the final undiscounted value of a generic financial position/liability, $\alpha \in (0, \infty)$ a given risk-aversion coefficient, $\mu \in \mathbb{R}$ a target minimal expected value and $\sigma^2 \in (0, \infty)$ a target maximal variance.

Problem (A) $\quad E[g - Y] - \alpha \mathrm{Var}[g - Y] = \max!$ $\quad$ over all $g \in \mathcal{G}$

Problem (B) $\quad \mathrm{Var}[g - Y] = \min!$ $\quad$ over all $g \in \mathcal{G}$ such that $E[g - Y] \geq \mu$

Problem (C) $\quad E[g - Y] = \max!$ $\quad$ over all $g \in \mathcal{G}$ such that $\mathrm{Var}[g - Y] \leq \sigma^2$

Problem (D) $\quad E\left[|Y - g|^2\right] = \min!$ $\quad$ over all $g \in \mathcal{G}$

Problem (A) corresponds to the portfolio optimisation problem faced by an agent with mean-variance preferences and risk-aversion coefficient α. Problems (B) and (C) correspond to abstract versions of the classical Markowitz mean-variance portfolio selection problems, here extended with the inclusion of the random liability Y. Finally, Problem (D) consists in finding the optimal mean-variance hedge for Y. It can be easily seen that there is no loss of generality in introducing the following additional standing Assumption.

Assumption 2 $1 \notin \mathcal{G}^{\perp}$, i.e. $\{g \in \mathcal{G} : E[g] \neq 0\} \neq \emptyset$.

Denote by π the orthogonal projection in L^2 onto $\mathcal{G}^{\perp}$ and note that $\left(\mathcal{G}^{\perp}\right)^{\perp} = \overline{\mathcal{G}}$, since $\mathcal{G}$ is a linear. This yields the direct sum decomposition $L^2 = \overline{\mathcal{G}} \oplus \mathcal{G}^{\perp}$, meaning that any $Y \in L^2$ can be uniquely decomposed as follows:

$$Y = g^Y + N^Y = g^Y + \pi(Y) \qquad \text{with } g^Y \in \overline{\mathcal{G}} \text{ and } N^Y = \pi(Y) \in \mathcal{G}^{\perp} \tag{2}$$

Remark The optimal values of Problems (A)-(D) do not depend on whether we optimise over $\mathcal{G}$ or $\overline{\mathcal{G}}$. This can be easily checked due to the fact that $g_n \to g$ in L^2 as $n \to \infty$ implies that $E[g_n - Y] \to E[g - Y]$ and $\mathrm{Var}[g_n - Y] \to \mathrm{Var}[g - Y]$ as $n \to \infty$, for any $Y \in L^2$. In view of this Remark, we henceforth consider Problems (A)-(D) as optimisation problems over the closed subspace $\overline{\mathcal{G}}$.

By relying on (2), we are already in a position to solve Problem (D). In fact, a simple application of the projection theorem gives:

$$g^Y = \underset{g \in \overline{\mathcal{G}}}{arg\,min} \, \|Y - g\|_{L^2}$$

The following variance-minimisation problem plays a crucial role in the solution of Problems (A)-(C).

Problem (MV) $\quad \mathrm{Var}[Y - g] = \min!$ $\quad$ over all $g \in \overline{\mathcal{G}}$

Proposition 2 *For $Y \in L^2$, Problem (MV) admits in $\overline{\mathcal{G}}$ the unique solution*

$$g^Y_{\mathrm{mv}} := \underset{g\in\overline{\mathcal{G}}}{arg\,min}\mathrm{Var}[Y-g] = g^Y - a^*_Y\big(1-\pi\,(1)\big), \qquad \textit{where } a^*_Y := \frac{E[N^Y]}{E[\pi\,(1)]}.$$

Let us now introduce the notation $R^Y_{\mathrm{mv}} := g^Y_{\mathrm{mv}} - Y$, where "$R$" stands for the final "result" of an abstract financial position. Then, for any $g \in \overline{\mathcal{G}}$, we can write as follows:

$$g - Y = g - g^Y_{\mathrm{mv}} + g^Y_{\mathrm{mv}} - Y = R^Y_{\mathrm{mv}} + g - g^Y_{\mathrm{mv}}$$

Furthermore, due to the variance-optimality of $g^Y_{\mathrm{mv}} \in \overline{\mathcal{G}}$ and the linearity of $\overline{\mathcal{G}}$, the element R^Y_{mv} enjoys the following fundamental zero-covariance property:

$$\mathrm{Cov}\left(R^Y_{\mathrm{mv}}, g\right) = 0 \qquad \text{for all } g \in \overline{\mathcal{G}}$$

Since $g - g^Y_{\mathrm{mv}} \in \overline{\mathcal{G}}$, for any $g \in \overline{\mathcal{G}}$, this implies that we can write as follows:

$$\mathrm{Var}\left[g-Y\right] = \mathrm{Var}\left[R^Y_{\mathrm{mv}} + g - g^Y_{\mathrm{mv}}\right] = \mathrm{Var}\left[R^Y_{\mathrm{mv}}\right] + \mathrm{Var}\left[g - g^Y_{\mathrm{mv}}\right]$$

This shows that in the analysis of Problems (A)-(C) we can isolate the minimum variance element R^Y_{mv}. Furthermore, since $g^Y_{\mathrm{mv}} \in \overline{\mathcal{G}}$ and $\overline{\mathcal{G}}$ is a linear space, the mapping $g \mapsto g' := g - g^Y_{\mathrm{mv}}$ is a bijection of $\overline{\mathcal{G}}$ to itself. These two observations suggest that we can reduce the general versions of our abstract mean-variance problems to the particular case $Y \equiv 0$. This fact allows us to easily derive the solution to Problem (A), denoted by $g^Y_{\mathrm{opt},A}(\gamma)$, where $\gamma := \frac{1}{\alpha}$ is the risk-tolerance coefficient corresponding to the risk-aversion coefficient α.

Proposition 3 *For $Y \in L^2$ and $\gamma \in [0,\infty)$, Problem (A) has a unique solution $g^Y_{\mathrm{opt},A}(\gamma) \in \overline{\mathcal{G}}$. It is explicitly given by*

$$g^Y_{\mathrm{opt},A}(\gamma) = \underset{g\in\overline{\mathcal{G}}}{arg\,min}\big\{\mathrm{Var}[g-Y] - \gamma E[g-Y]\big\} = g^Y_{\mathrm{mv}} + g^0_{\mathrm{opt},A}(\gamma),$$

where $g^0_{\mathrm{opt},A}(\gamma) \in \overline{\mathcal{G}}$ is the solution to Problem (A) for $Y \equiv 0$, explicitly given by

$$g^0_{\mathrm{opt},A}(\gamma) = \underset{g\in\overline{\mathcal{G}}}{arg\,min}\big\{\mathrm{Var}[g] - \gamma E[g]\big\} = \frac{\gamma}{2}\frac{1}{E[\pi\,(1)]}\big(1-\pi\,(1)\big).$$

Furthermore, the condition of *no approximate riskless profits in L^2* is not only sufficient for ensuring the existence of a solution to Problem (A), but it is also necessary for the solvability of (A). The solutions to Problems (B) and (C) can be recovered from the general solution to Problem (A) by choosing a suitable risk-aversion coefficient α, which will depend on the constraints μ and σ. More precisely, we have the two following Propositions.

Proposition 4 *Let $Y \in L^2$ and $\mu \in \mathbb{R}$. If $\mu > E[R^Y_{\mathrm{mv}}]$, then Problem (B) admits a unique solution $g^Y_{\mathrm{opt},B}(\mu) \in \overline{\mathcal{G}}$. It is explicitly given by*

$$g^Y_{\mathrm{opt},B}(\mu) = g^Y_{\mathrm{mv}} + g^0_{\mathrm{opt},B}\big(\mu - E[R^Y_{\mathrm{mv}}]\big),$$

where $g^0_{\mathrm{opt},B}(m)$ *is the solution to Problem (B) for* $Y \equiv 0$ *and constraint* m*, explicitly given by*

$$g^0_{\mathrm{opt},B}(m) = g^0_{\mathrm{opt},A}\left(2m\frac{E\left[\pi\left(1\right)\right]}{E\left[1-\pi\left(1\right)\right]}\right) = \frac{m}{E[1-\pi\left(1\right)]}\big(1-\pi\left(1\right)\big).$$

If $\mu \leq E[R^Y_{\mathrm{mv}}]$*, then Problem (B) has* g^Y_{mv} *as unique solution.*

Proposition 5 *Let* $Y \in L^2$ *and* $\sigma^2 \in [0,\infty)$*. If* $\sigma^2 \geq \mathrm{Var}[R^Y_{\mathrm{mv}}]$*, then Problem (C) admits a unique solution* $g^Y_{\mathrm{opt},C}(\sigma^2) \in \overline{\mathcal{G}}$*. It is explicitly given by*

$$g^Y_{\mathrm{opt},C}(\sigma^2) = g^Y_{\mathrm{mv}} + g^0_{\mathrm{opt},C}\big(\sigma^2 - \mathrm{Var}[R^Y_{\mathrm{mv}}]\big),$$

where $g^0_{\mathrm{opt},C}(v)$ *is the solution to Problem (C) for* $Y \equiv 0$ *and constraint* v*, explicitly given by*

$$g^0_{\mathrm{opt},C}(v) = g^0_{\mathrm{opt},A}\left(2\frac{\sqrt{v}E\left[\pi\left(1\right)\right]}{\sqrt{\mathrm{Var}\left[1-\pi\left(1\right)\right]}}\right) = \sqrt{\frac{v}{\mathrm{Var}[1-\pi\left(1\right)]}}\big(1-\pi\left(1\right)\big).$$

If $\sigma^2 < \mathrm{Var}[R^Y_{\mathrm{mv}}]$*, Problem (C) cannot be solved.*

Referring the reader to [5] for a more detailed analysis, we make a few important remarks on the results of Propositions 3-5.

Remarks **1.** Note that the solutions to Problems (A)-(D) all share the same fundamental structure:

$$g^Y_{\mathrm{opt},i} = g^Y_{\mathrm{mv}} + c^Y_{\mathrm{opt},i}\big(1-\pi\left(1\right)\big) \qquad \text{for } i \in \{A,B,C,D\} \tag{3}$$

for some $c^Y_{\mathrm{opt},i} \in \mathbb{R}$, $i \in \{A,B,C,D\}$, and where $g^Y_{\mathrm{opt},D} := g^Y$. Property (3) represents an abstract counterpart of the classical two-fund separation result already observed in Section 2. This property allows us to easily obtain a model-independent description of the mean-variance efficient frontier and an abstract counterpart of the traditional *CAPM* formula.

2. The element $1-\pi\left(1\right)$ can be characterised as the unique element of $\overline{\mathcal{G}}$ in the Riesz representation of the continuous linear functional $E\left[\cdot\right]$ on $\overline{\mathcal{G}}$. Furthermore, in our abstract mean-variance theory, the element $1-\pi\left(1\right)$ plays the role of a generalised *market portfolio*, determining the slope of the mean-variance efficient frontier. The terms $c^Y_{\mathrm{opt},i}$ in (3) also admit a representation as "*beta factors*". In fact, it can be shown that $c^Y_{\mathrm{opt},i} = \mathrm{Cov}\left(g^Y_{\mathrm{opt},i} - Y, 1-\pi\left(1\right)\right)/\mathrm{Var}\left[1-\pi\left(1\right)\right]$, for $i \in \{A,B,C,D\}$.

3. Several investment situations can be represented by letting $Y = -cB + (H - hB) - H_0$, for $c,h \in \mathbb{R}$ and $H, H_0 \in L^2$. This describes the net financial balance (outflows minus incomes) at the final time T faced by an agent who, at the starting time $t = 0$, has an initial endowment c and sells for a compensation h the contingent claim H, to be paid at T. In addition, the agent has a position H_0 (evaluated at T), which can be interpreted as a random endowment. Furthermore, the compensation h for selling the claim H can

be determined endogenously in the model, as an application of *mean-variance indifference valuation* rules.

References

[1] Choulli, T., Krawczyk, L. and Stricker, C., *$\mathcal{E}$-martingales and their applications in mathematical finance.* Annals of Applied Probability 26/2 (1998), 853–876.

[2] Delbaen, F., Monat, P., Stricker, C., Schachermayer, W. and Schweizer, M., *Weighted norm inequalities and hedging in incomplete markets.* Finance and Stochastics 1 (1997), 181–227.

[3] Fishburn, P.C., *Mean-risk analysis with risk associated with below-target returns.* American Economic Review 67/2 (1977), 116–126.

[4] Fontana, C., "Mean-variance Problems with Applications to Credit Risk Models". MAS Finance Thesis, ETH Zürich and University of Zürich, 2010.

[5] Fontana, C. and Schweizer, M., *The mathematics and financial economics of mean-variance portfolio optimisation.* Preprint, ETH Zürich (2011).

[6] Huang, C. and Litzenberger, R.H., "Foundations for Financial Economics". North-Holland, New York, 1988.

[7] Li, D. and Ng, W.L., *Optimal dynamic portfolio selection: Multiperiod mean-variance formulation.* Mathematical Finance 10/3 (2000), 387–406.

[8] Luenberger, D.G., "Investment Science". Oxford University Press, New York, 1998.

[9] Mania, M., Jeanblanc, M., Santacroce, M. and Schweizer, M., *Mean-variance hedging via stochastic control and BSDEs for general semimartingales.* NCCR FINRISK working paper No. 675, ETH Zürich (2011).

[10] Markowitz, H., *Portfolio selection.* Journal of Finance 7 (1952), 77–91.

[11] Markowitz, H., "Mean-variance Analysis in Portfolio Choice and Capital Markets". Basil Blackwell, Oxford - New York, 1987.

[12] Pham, H., *On quadratic hedging in continuous time.* Mathematical Methods of Operations Research 51 (2000), 315–339.

[13] Schweizer, M., *From actuarial to financial valuation principles.* Insurance: Mathematics and Economics 28 (2001), 31–47.

[14] Schweizer, M., *A guided tour through quadratic hedging approaches.* In: Jouini, E., Cvitanic, J. and Musiela, M. (eds.), "Option Pricing, Interest Rates and Risk Management", 538–574. Cambridge University Press, Cambridge, 2001.

[15] Schweizer, M., *Mean-variance hedging.* In: Cont, R. (ed.), "Encyclopedia of Quantitative Finance", 1177–1181. Wiley, Chichester, 2010.

[16] Steinbach, M.C., *Markowitz revisited: Mean-variance models in financial portfolio analysis.* SIAM Review 43/1 (2001), 31–85.

Approximating the Goldbach Conjecture

Valentina Settimi (*)

Abstract. The Goldbach conjecture is one of the oldest unsolved problems in the entire mathematics and, since its appearance in 1742 to nowadays, a lot of mathematicians dealt with it. In my talk I will give an introduction to the origin of the Goldbach conjecture and then I will describe the most important developments in some problems related to it. In particular I will talk about the ternary Goldbach conjecture, the exceptional set in Goldbach's problem and the Goldbach-Linnik problem. Finally, I will give a short overview of our results which can be seen as approximations to the Goldbach-Linnik problem.

1 Origin of the Goldbach Conjecture

In a letter to Euler dated 7 June of 1742, Goldbach stated the following conjecture

$$\begin{gathered}\textit{if } N \textit{ is an integer such that } N = p_1 + p_2, \textit{ with } p_1 \textit{ and } p_2 \textit{ primes,}\\ \textit{then, for every } 2 \leq k \leq N,\ N = p_1 + \ldots + p_k, \textit{ with } p_1, \ldots, p_k \textit{ prime.}\end{gathered}$$

We have to keep in mind that in Goldbach's time the number 1 was considered to be a prime, in contrast with the modern definition. In the margin of the same letter, Goldbach stated another conjecture

$$\begin{gathered}\textit{if } N \textit{ is a integer greater than } 2,\\ \textit{then } N = p_1 + p_2 + p_3, \textit{ with } p_1,\ p_2 \textit{ and } p_3 \textit{ primes.}\end{gathered}$$

In his reply letter, dated 30 June of the same year, Euler wrote another conjecture which is now ascribed to Goldbach

$$\textit{if } N \textit{ is a positive even integer, then } N = p_1 + p_2, \textit{ with } p_1 \textit{ and } p_2 \textit{ primes.}$$

Today, these conjectures are known to be equivalent (see, *e.g.*, Pintz [7]).

1.1 Modern versions

The conjectures above can be rewritten using the modern language of primes, that is without considering 1 to be a prime number. The first conjecture is strictly connected to

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: vsettimi@math.unipd.it . Seminar held on 23 March 2011.

the primality of 1 and therefore its modern version has no interest. On the contrary, the modern version of the second conjecture is the so called **ternary Goldbach conjecture**

(TGC) *if N is an odd integer greater than 5,*
then $N = p_1 + p_2 + p_3$, with p_1, p_2 and p_3 primes,

while the modern version of the third conjecture is the famous **Goldbach conjecture**

(GC) *if N is an even integer greater than 2,*
then $N = p_1 + p_2$, with p_1 and p_2 primes.

Despite its very simple statement, GC is extremely hard to prove and nowadays, after more than 250 years, it is still an open problem. Nevertheless, it can be approached in two main ways:

(a) Numerical check: the record is up to 2×10^{18}, due to Oliveira e Silva [10] (2010).

(b) Approximations: the study of problems closely related to GC.

2 Some approximations to Goldbach conjecture

2.1 Ternary Goldbach conjecture

We remark that the modern conjectures above have more restrictive hypothesis than original ones and so they are stronger. In particular the Goldbach conjecture and the ternary Goldbach conjecture are not equivalent, but only

$$\text{GC} \Rightarrow \text{TGC},$$

which is trivial, considering $N - 3 = p_1 + p_2$. For this reason TGC can be considered an approximation to GC.

The first important result about TGC is due to Hardy-Littlewood [1] (1923), who proved it for any sufficiently large odd integer and under *Generalized Riemann Hypothesis* (GRH). The main tool they used is the *circle method*. We shortly recall here what the GRH and the circle method are.

Generalized Riemann Hypothesis: We need the following definitions.

Definition Given $q \in \mathbb{N} \setminus \{0\}$, χ is *Dirichlet character* $\pmod q$ iff

(i) $\chi : \mathbb{Z} \to \mathbb{C}$;
(ii) $\chi(mn) = \chi(m)\chi(n) \quad \forall m, n \in \mathbb{N}$;
(iii) $\chi(n + q) = \chi(n) \quad \forall n \in \mathbb{N}$;
(iv) $\chi(n) = 0 \Leftrightarrow (n, q) > 1$.

Definition For every $s \in \mathbb{C}$ such that $\Re(s) > 1$ and for every Dirichlet character χ, we define the *Dirichlet L-function* as

$$L(s, \chi) = \sum_{n \geq 1} \frac{\chi(n)}{n^s},$$

which, by analytic continuation, can be extended to the whole $\mathbb{C}$.

We remark that the Dirichlet L-functions are generalizations of the *Riemann ζ-function* (which is defined as $\zeta(s) = \sum_{n \geq 1} \frac{1}{n^s}$ for every $s \in \mathbb{C}$ with $\Re(s) > 1$, and than extended to $\mathbb{C}$ by analytic continuation). In fact, for every $s \in \mathbb{C}$, we have that $L(s, \chi_0) = \zeta(s)$, where χ_0 is the Dirichlet character modulo 1.

Finally, let us call *non-trivial zeros* of $\zeta(s)$ or $L(s, \chi)$ those zeros having $0 < \Re(s) < 1$.

Riemann Hypothesis (RH). The non-trivial zeros of $\zeta(s)$ lie on the critical line $\Re(s) = \frac{1}{2}$.

Generalized Riemann Hypothesis (GRH). The non-trivial zeros of $L(s, \chi)$, for all χ, lie on the critical line $\Re(s) = \frac{1}{2}$.

Circle Method: It is used to approach many additive problems and it can be roughly summarized as follows:

(a) Turning an additive problem over integers (*e.g.*, the ternary Goldbach problem) into an analytic problem, by means of Fourier analysis.

(b) Dissecting the obtained integration inteval into major and minor arcs which respectively give the expected main term and the expected error term.

To better explain how the method works, we shortly sketch here its application to TGC: the weighted counting function for TGC is

$$r_3(N) = \sum_{p_1+p_2+p_3=N} \log(p_1) \log(p_2) \log(p_3).$$

If we define the exponential sum $S(\alpha) = \sum_{p \leq N} \log(p) e^{2\pi i p \alpha}$, then, by the Fourier coefficients formula, we obtain the following fundamental relation

$$r_3(N) = \int_0^1 S(\alpha)^3 e^{-2\pi i N \alpha} \mathrm{d}\alpha,$$

which allows us to change over our additive problem into an analytic one. The last step is to dissect the integration interval $[0, 1]$ into *major arcs* $\mathfrak{M}$ and *minor arcs* $\mathfrak{m}$ in such a way that: $\mathfrak{M}$ are around the peaks of $S(\alpha)$ and therefore they give the main term in the asymptotic estimation of $r_3(N)$, while $\mathfrak{m}$ give the error term, since $|S(\alpha)|$ can be suitably bounded whenever $\alpha \in \mathfrak{m}$.

Some year later, Vinogradov [12] (1937) succeeded in proving, for any sufficiently large odd integer, TGC *unconditionally*, that is without assuming GRH. The key point in his proof is a better estimation on minor arcs.

2.2 Exceptional Set in Goldbach's Problem

If we denote by $\mathfrak{P}$ the set of all prime numbers, then the *exceptional set* for the Goldbach conjecture is

$$\mathcal{E} = \{N \in \mathbb{N} : 2|N;\ \nexists p_1, p_2 \in \mathfrak{P} \text{ s.t. } N = p_1 + p_2\},$$

which naturally leads to the following problem.

Problem *Given $X \in \mathbb{R}$, finding an upper bound for $E(X) = |\mathcal{E} \cap [1, X]|$.*

It can be considered an approximation to the Goldbach conjecture, since

$$\text{GC} \quad \Leftrightarrow \quad \mathcal{E} = \{2\} \quad \Leftrightarrow \quad E(X) = 1,\ \forall X \geq 2.$$

The first important result regarding the size of the exceptional set is due to Hardy-Littlewood [2] (1923), who proved that, under GRH

$$E(X) \ll_\epsilon X^{1/2+\epsilon} \qquad \forall \epsilon > 0.$$

As for the unconditional side, we recall the fundamental result by Montgomery-Vaughan [6] (1975):

$$E(X) \ll X^{1-\delta} \qquad \exists \delta > 0.$$

The key point in Montgomery-Vaughan's proof is a careful analysis of the *Siegel zero*, which, in plain words, is a potential counterexample to the GRH.

Recently, Pintz [8] (2009) announced that the Montgomery-Vaughan estimate holds for $\delta = 1/3$. The key point in his proof is a careful analysis of the *generalized exceptional zeros* which are, roughly speaking, generalizations of the Siegel zero.

2.3 Goldbach-Linnik Problem

The Goldbach-Linnik (or just G.-Linnik) problem originates from the following theorem by Linnik, proved in 1951 under GRH and two years later unconditionally.

Theorem *If N is a sufficiently large even integer, then there exist $p_1, p_2 \in \mathfrak{P}$ and $\nu_1, \ldots, \nu_s \in \mathbb{N}$ such that*

$$N = p_1 + p_2 + 2^{\nu_1} + \ldots + 2^{\nu_s},$$

where s is an unspecified absolute constant.

Goldbach-Linnik Problem. *Finding the smallest allowed value of s.*

It can be considered an approximation to the Goldbach conjecture, since

$$s = 0 \quad \Leftrightarrow \quad \text{GC, for sufficiently large } N.$$

To this day, the best upper bounds for s are $s = 7$ under GRH and $s = 13$ unconditionally, obtained by Heath–Brown-Putcha [3] (2002).

Recently Languasco-Pintz-Zaccagnini [4] (2007) studied a variation of the G.-Linnik problem: given $s \geq 1$, finding an asymptotic formula for the number of representations of a positive even integer (less than a large parameter X) as sum of two primes and k powers

of 2, which holds for almost all positive even integers. The important point in this work is that, for every $s \geq 1$, the number of exceptional values for the asymptotic formula is $\ll_k X^{3/5}(\log X)^{10}$. In fact:

(a) as said in the previous section, by Pintz [8], the size of the exceptional set for the Goldbach conjecture is $\ll X^{2/3}$, and $3/5 < 2/3$. It means that, just adding a single power of 2, a better estimation can be obtained;

(b) the exponent $3/5$ is the best possible level, according to the state of the art: to lower it, we have to refine, in the exponents, the famous estimation in Theorem 3.1 of Vaughan [11].

3 Our results

Our results can be considered as variations of G.-Linnik problem:

(a) Generalization of Languasco-Pintz-Zaccagnini [4] with $g \geq 3$ instead of 2. That is studying the formula
$$N = p_1 + p_2 + g^{\nu_1} + \ldots + g^{\nu_s}.$$

(b) Diophantine approximation to G.-Linnik Problem, combined with *Waring-Goldbach Problem* (*i.e.* the problem of representing an integer as sum of prime powers). That is studying the formula
$$\lambda_1 p_1 + \lambda_2 p_2^2 + \lambda_3 p_3^2 + \mu_1 2^{\nu_1} + \ldots + \mu_s 2^{\nu_s},$$
where the coefficients $\lambda_i, \mu_i \in \mathbb{R} \setminus \{0\}$ satisfy some suitable relations.

3.1 Generalization of Languasco-Pintz-Zaccagnini: study of $N = p_1+p_2+g^{\nu_1}+\ldots+g^{\nu_s}$

We start by fixing the integer $s \geq 1$ and $g \geq 3$. Our N has to satisfy some (standard) *arithmetic conditions*:

(AC) $\qquad N$ even, if g even; $\qquad\qquad N \equiv s \pmod 2$, if g odd.

We now fix a large real parameter X and we set $N \in [1, X]$ and $L = \log_g X$. The relevant counting function for our problem, which depends on s and g, is

$$\begin{aligned} r(N) &= |\{(p_1, p_2, \nu_1, \ldots, \nu_s) \in \mathfrak{P}^2 \times [1, L]^s : N = p_1 + p_2 + g^{\nu_1} + \ldots + g^{\nu_s}\}| \\ &= \sum_{1 \leq p_1, p_2 \leq X} \sum_{\substack{1 \leq \nu_1, \ldots, \nu_s \leq L \\ p_1+p_2+g^{\nu_1}+\ldots+g^{\nu_s}=N}} 1. \end{aligned}$$

So the associated weighted function, which still depends on s and g, is

$$R(N) = \sum_{1 \leq m_1, m_2 \leq X} \sum_{\substack{1 \leq \nu_1, \ldots, \nu_s \leq L \\ m_1+m_2+g^{\nu_1}+\ldots+g^{\nu_s}=N}} \Lambda(m_1)\Lambda(m_2),$$

with $\Lambda(n)$ the *Von Mangoldt function*: $\Lambda(n) = \log p$, if there exists $p \in \mathfrak{P}$ and $k \in \mathbb{N} \setminus \{0\}$ such that $n = p^k$, and $\Lambda(n) = 0$ otherwise. Using this notation, our theorem is the following:

Theorem A *Let $\eta > 0$ be an arbitrarily small constant, then there exists a positive constant $\mathbf{C} = \mathbf{C}(g, s, N)$ such that*

$$|R(N) - \mathbf{C}NL^s| \leq \eta NL^s,$$

for every N satisfying (AC), apart from at most $\mathcal{O}_g\big(X^{3/5}(\log X)^{10}\big)$ exceptions.

The relevant point is that $\mathcal{O}_g\big(X^{3/5}(\log X)^{10}\big)$ is optimal according to the state of the art, as said before. Moreover $\mathbf{C}$ can be bounded from above with dependency only on g. We finally remark that in the proof we use both the circle method and generalized exceptional zeros mentioned before.

3.2 Diophantine approximation to the G-Linnik problem: study of $\lambda_1 p_1 + \lambda_2 p_2^2 + \lambda_3 p_3^2 + \mu_1 2^{\nu_1} + \ldots + \mu_s 2^{\nu_s}$

The coefficients λ_i, μ_i have to satisfy the following (standard) relation:

$$\text{(CR)} \qquad \begin{array}{ll} \lambda_i, \mu_i \in \mathbb{R} \setminus \{0\} & \lambda_1 < 0 \text{ and } \lambda_2, \lambda_3 > 0; \\ \lambda_2/\lambda_3 \notin \mathbb{Q}; & \lambda_i/\mu_i \in \mathbb{Q}, \text{ for every } 1 \leq i \leq 3. \end{array}$$

Let a_i/q_i denote the reduced representation of $\lambda_i/\mu_i \in \mathbb{Q}$. Using this notation, our theorem is the following.

Theorem B *For every real number x and for every integer $s \geq s_0$, then*

$$| \lambda_1 p_1 + \lambda_2 p_2^2 + \lambda_3 p_3^2 + \mu_1 2^{\nu_1} + \cdots + \mu_s 2^{\nu_s} + x | < \eta$$

has infinitely many solutions in $p_i \in \mathfrak{P}$ and $\nu_i \in \mathbb{N} \setminus \{0\}$, where

(i) *the coefficients λ_i, μ_i verify (CR);*

(ii) *$\eta > 0$ is a sufficiently small constant such that $\eta < \min\big(\big|\frac{\lambda_1}{a_1}\big|; \frac{\lambda_2}{a_2}; \frac{\lambda_3}{a_3}\big)$;*

(iii) *$s_0 = s_0(q_1, q_2, q_3, \lambda_1, \lambda_2, \lambda_3, \eta, \epsilon)$ explicit constant, with $\epsilon > 0$ arbitrarily small.*

The key point here is to find an allowed value for s_0 as smaller as possible. Under this point of view, our result can be considered as a refinement of the analogous work by Li-Wang [5] (2005), since we find a s_0 which is smaller than their one by about 90%. With respect to Li-Wang our main gain comes from use for the first time a standard tool for exponential sums over primes, to deal with exponential sums over prime squares.

References

[1] G. H. Hardy, J. E. Littlewood, *Some problems of 'Partitio Numerorum'; III: on the expression of a number as a sum of primes.* Acta Math. 44 (1923), 1–70.

[2] G. H. Hardy, J. E. Littlewood, *Some problems of 'Partitio numerorum'; V: A further contribution to the study of Goldbach's problem.* Proc. London Math. Soc. 22 (1923), 46–56.

[3] D. R. Heath-Brown, J.-C. Puchta, *Integers represented as a sum of primes and powers of two.* Asian J. Math. 6 (2002), 535-565.

[4] A. Languasco, J. Pintz, A. Zaccagnini, *On the sum of two primes and k powers of two.* Bull. London Math. Soc. 39 (2007), 771–780.

[5] W. P. Li, T. Z. Wang, *Diophantine approximation by a prime, squares of two primes and powers of two.* Pure Appl. Math. (Xi'an) 21 (2005), 295–299.

[6] H. L. Montgomery, R. C. Vaughan, *The exceptional set in Goldbach's problem.* Acta Arith. 27 (1975), 353–370.

[7] J. Pintz, *Recent results on the Goldbach conjecture.* Lecture Notes in Math. (2006), 220–254.

[8] J. Pintz, *Landau's problems on primes.* J. Théor. Nombres Bordeaux 21 (2009), 357–404.

[9] J. Pintz, I. Z. Ruzsa, *On Linnik's approximation to Goldbach's problem, I.* Acta Arith. 109 (2003), 169-194.

[10] T. Oliveira e Silva, *Goldbach conjecture verification.* Available in the following webpage: `http://www.ieeta.pt/~tos/goldbach.html`.

[11] R. C. Vaughan, "The Hardy-Littlewood method. Second Edition". Cambridge University Press 125, 1997.

[12] I. M. Vinogradov, *The representation of an odd number as a sum of three primes.* Dokl. Akad. Nauk SSSR 15 (1937), 169–172.

[13] A. Languasco, V. Settimi, *On a Diophantine problem with one prime, two squares of primes and s powers of two.* ArXiV, submitted (2011).

[14] V. Settimi, *On the sum of two primes and k powers of an integer $g > 2$.* Preprint (2011).

Robustness for path-dependent volatility models

Mauro Rosestolato (*)

Based on joint work with Tiziano Vargiolu (University of Padova)
and Giovanna Villani (La Caixa, Barcelona)

1 Introduction

The Black and Scholes model is based upon the assumption that the behaviour of the logarithm of the asset price is well represented by a Gaussian process with stationary independent increments. This assumption is mathematically given by imposing that the drift and the volatility are deterministic functions. The important role played by the volatility in the Black-Scholes formula and the fact that a constant volatility assumption is not consistent with observations of actual financial markets are both well known, and for these reasons several proposals have been made to introduce some sort of stochastic dependency in the volatility parameter, either with a deterministic dependency on the current stock price or with a dedicated dynamics driven by a new source of uncertainty.

One of the models which better fits to market data is the so-called Hobson-Rogers model, introduced in [10] and studied with respects to various features in [1, 4, 5, 6, 7, 8, 13] which consists in the following. For a risky asset whose price is denoted by the process $S = (S_t)_{t\in\mathbb{R}}$, define the discounted log-price process Z_t at time t as $Z_t = \log(S_t \mathrm{e}^{-rt})$ where r is the (constant) risk-free interest rate, and the *offset function* of order 1, denoted by $P = (P_t)_t$, by

$$P_t = \int_0^\infty \lambda \mathrm{e}^{-\lambda u}(Z(t) - Z(t-u))du \tag{1}$$

the constant λ being a parameter of the model which describes the rate at which past information is discounted. We assume that Z satisfies the SDE (stochastic differential equation)

$$dZ_t = -\frac{1}{2}\sigma^2(P_t)dt + \sigma(P_t)\ dW(t) \tag{2}$$

where $\sigma(\cdot)$ is a strictly positive function and $(W_t)_{t\in\mathbb{R}}$ is a so-called two-sided Brownian motion [3] under a risk-neutral probability measure $\mathbb{P}$ (see [1, 2, 10] and the references therein for details).

(*)Scuola Normale Superiore, Pisa (Italy). E-mail: **mauro.rosestolato@sns.it** . Seminar held on 6 April 2011.

This model can be seen as a "good" model because no new Brownian motions (or other sources of uncertainty) have been introduced in the specification of the price process. This means that the market is complete and any contingent claim is hedgeable in this way: if we calculate the stochastic differential of P, we obtain

$$dP_t = dZ(t) - \lambda P_t \, dt \tag{3}$$

so (Z, P), as well as (S, P), is a 2-dimensional Markov process (see [10]), and we can easily employ the Kolmogorov equation when pricing a contingent claim with final payoff $h(S_T)$. In fact, its price $V_t = \mathbb{E}[h(S_T)|\mathcal{F}_t]$ is of the form $V_t = F(t, S_t, P_t)$, where F is the solution of the Kolmogorov equation

$$\begin{cases} F_t + rsF_s - \lambda pF_p + \left(\frac{1}{2}s^2F_{ss} + sF_{ps} + \frac{1}{2}F_{pp} - \frac{1}{2}F_p\right)\sigma^2(p) = rF \\ \\ F(p, s, T) = h(s). \end{cases} \tag{4}$$

In conclusion this model allows to construct a process for the price, but we can see that some difficulties arise. The problem of pricing a contingent claim with the Hobson-Rogers model is equivalent to solve the PDE (4), once the initial conditions $S(0) = s$, $P(0) = p$ are specified. While the price $S(0)$ is observed in the market, in order to calculate the true value $P(0)$ one would have to observe the asset in all its past, which is impossible. A possible approach to circumvent this problem in the Hobson-Rogers model is to use the model with a misspecification $\widetilde{\Sigma}_0 := (\widetilde{P}_0, Z_0)$ instead of the true initial values $\Sigma_0 := (P_0, Z_0)$: we thus obtain, as solution of Equations (2) and (3), a misspecified process $\widetilde{\Sigma}_t := (\widetilde{P}_t, \widetilde{Z}_t)$ instead of the "true" process $\Sigma_t := (P_t, Z_t)$; we then search for an initial condition $\widetilde{P}_0$ which minimizes the error of pricing the contingent claim $h(S_T)$. This approach has been carried out in detail in [2] via L^2-estimates of the solutions of Equations (2) and (3) with respects to the initial condition, and the result in that paper is that

$$\mathbb{E}\left[\sup_{0\le u\le T}|\Sigma_u - \widetilde{\Sigma}_u|^2\right] \le K\mathbb{E}[|P_0 - \widetilde{P}_0|^2]\mathrm{e}^{cT^2+dT}$$

where K, c and d are suitable constant depending on λ and the function σ. The L^2-error of P_0 is then estimated by linking it to the L^2-error of P_{-R}, where $R > 0$ is assumed to be an observation interval of the past price of the stock $(S_t)_{t\in[-R,0]}$, which we assume to be available, and one has $\mathbb{E}[|P_0 - \widetilde{P}_0|^2] = \mathrm{e}^{-\lambda R}\mathbb{E}[|P_{-R} - \widetilde{P}_{-R}|^2]$. This latter L^2-error is then assumed to be equal to the variance V of the invariant measure of P: in fact, if the dynamics (3) of P is ergodic, then we have that $\mathbb{E}[|P_{-R} - \widetilde{P}_{-R}|^2]$ converges to V as $R \to +\infty$, so if R is big enough we can approximate $\mathbb{E}[|P_{-R} - \widetilde{P}_{-R}|^2]$ with V.

This entails that when pricing a European (possibly path-dependent) contingent claim with maturity T and final payoff $h(S(\cdot))$ we have

$$\left|\mathbb{E}[h(S_T)] - \mathbb{E}[h(\widetilde{S}_T)]\right|^2 \le KJ^2\mathrm{e}^{-\lambda R}V\mathrm{e}^{cT^2+dT} \tag{5}$$

where J is the Lipschitz constant of the functional $z(\cdot) \to h(\mathrm{e}^{z(\cdot)})$. If one wants to obtain prices with a given precision, the estimate (5) gives a quadratic dependence of R on the

maturity T, which produces very long and unlikely observation times: in the examples in [2], for a maturity of $T = 3$ months one has to observe $R \simeq 4$ years of the historical prices of S, while for a maturity of $T = 5$ years this observation window becomes $R \simeq 100$ years long. If h is a simple European claim, then an analogous estimate holds which has all the previous drawbacks.

In this note, we present a L^1-estimates (instead of the L^2-estimates of [2]) of the form

$$\mathbb{E}\left[\sup_{0\le u\le T}|\Sigma_u - \widetilde{\Sigma}_u|\right] \le K(T)\mathbb{E}[|P_0 - \widetilde{P}_0|]\mathrm{e}^{dT} \tag{6}$$

where K is a function with subexponential growth. This is done by obtaining the differential $\partial\Sigma_t$ of the sample paths of Σ with respects to the initial condition Σ_0 and using Lagrange's theorem, which entails $\Sigma_t^{(p,z)} - \Sigma_t^{(\widetilde{p},z)} = \int_{\widetilde{p}}^{p} \partial_1\Sigma_t^{(\zeta,z)}\,\mathrm{d}\zeta$. Since for $\partial_1\Sigma_t^{(\zeta,z)}$ we obtain estimates of the kind $\mathbb{E}[\|\partial_1\Sigma_t^{(\zeta,z)}\|] \le K\mathrm{e}^{dt}$, by integrating we get the desired estimate (6), which is a great improvement of the result in [2].

2 Dependence with respect to the initial data

We obtain L^1-estimates on $\Sigma := (P, Z)$ by the use of differentiation of stochastic processes and of Lagrange's theorem. The use of this latter technique and the requirement for a L^1-estimate will allow us to obtain log-linear estimates of the kind of (6) instead of the original log-quadratic ones present in [2].

The starting point is to see that the process $\Sigma := (P, Z)$ is differentiable with respect to the initial value, and the derivative process with respect to P_0 satisfies the SDE in the following theorem.

Theorem 1 *Assume that σ and σ^2 are differentiable, with locally Lipschitz derivatives bounded respectively by L_1 and L_2, and call $\Sigma = \Sigma^{P_0,Z_0}$ the solution to Equations (2) and (3) with initial condition $\Sigma_0 := (P_0, Z_0) \in L^2(\Omega, \mathcal{F}_0, \mathbb{P}; \mathbb{R}^2)$. If $(P_0, Z_0) = (p, z) \in \mathbb{R}^2$, then $\Sigma^{p,z}$ is differentiable with respect to the initial value, and the derivative process with respect to $P_0 = p$ satisfies the SDE*

$$\begin{cases} d\partial_1 P_t = -\left(\frac{1}{2}\left(\sigma^2\right)'(P_t) + \lambda\right)\partial_1 P_t\,dt + \sigma'(P_t)\,\partial_1 P_t\,dW_t \\ d\partial_1 Z_t = -\frac{1}{2}\left(\sigma^2\right)'(P_t)\,\partial_1 P_t\,dt + \sigma'(P_t)\,\partial_1 P_t\,dW_t \end{cases} \tag{7}$$

with initial conditions $\partial_1 P_0 = 1$, $\partial_1 Z_0 = 0$, where for a generic process $X = (P, Z)$ we indicate $\partial_1 X_t := \partial X_t^{p,z}/\partial p$.

Dealing with the 2-dimensional processess Σ and $\partial_1\Sigma$, we will use the norm $\mathbb{E}[\|\cdot\|_1]$, where $\|x\|_1 := |x_1| + |x_2|$ for all $x \in \mathbb{R}^2$.

Theorem 2 *Under the assumptions of Theorem 1, the following inequalities hold:*

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le u\le t}\|\partial_1\Sigma_u\|_1\right] &\le (5+\lambda t)\,\mathrm{e}^{\left(\frac{L_1^2}{2}+\frac{L_2}{2}-\lambda\right)^+ t}\\ \mathbb{E}\left[\|\partial_1\Sigma_t\|_1\right] &\le (3+\lambda t)\,\mathrm{e}^{\left(\frac{L_2}{2}-\lambda\right)^+ t}\ .\end{aligned}$$

Theorem 3 *Under the assumptions of Theorem 1, for each initial conditions η, $\widetilde{\eta}\in L^2(\Omega,\mathcal{F}_0,\mathbb{P})$ and $z\in\mathbb{R}$, the following inequalities hold*

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le u\le t}\left\|\Sigma_u^{(\eta,z)}-\Sigma_u^{(\widetilde{\eta},z)}\right\|_1\right] &\le (5+\lambda t)\,\mathrm{e}^{(\frac{1}{2}L_1^2+\frac{1}{2}L_2-\lambda)^+ t}\mathbb{E}\left[|\eta-\widetilde{\eta}|\right]\ ,\\ \mathbb{E}\left[\left\|\Sigma_t^{(\eta,z)}-\Sigma_t^{(\widetilde{\eta},z)}\right\|_1\right] &\le (3+\lambda t)\,\mathrm{e}^{(\frac{1}{2}L_2-\lambda)^+ t}\mathbb{E}\left[|\eta-\widetilde{\eta}|\right]\ .\end{aligned}$$

We now apply the results of Theorem 3 to the pricing error of European derivative assets $h(S(\cdot))$, possibly path-dependent, which are Lipschitz with respect to the log-return, i.e. such that the application $C^0([0,T])\ni f\to h(e^{f(\cdot)})\in\mathbb{R}$ is globally Lipschitz. In the case of a European claim which is a function of the final price, we require that $\mathbb{R}\ni x\to h(e^x)\in\mathbb{R}$ is globally Lipschitz. For some examples of such assets see [2].

Theorem 4 *Suppose the assumptions of Theorem 1 hold.*

1. *Let $h:C^0[0,T]\to\mathbb{R}$ be the payoff of a claim such that the functional $C^0[0,T]\to\mathbb{R}: f\mapsto h(\mathrm{e}^f)$ is globally Lipschitz (with respect to the sup-norm $\|\cdot\|_{C^0}$), with Lipschitz constant J. Then*

$$\left|\mathbb{E}\left[h\left(S(\cdot)\right)\right]-\mathbb{E}[h(\widetilde{S}(\cdot))]\right|\le J\,(5+\lambda T)\,\mathrm{e}^{\left(\frac{L_1^2}{2}+\frac{L_2}{2}-\lambda\right)^+ T}\mathbb{E}\left[\left|P(0)-\widetilde{P}(0)\right|\right] \tag{8}$$

2. *Let $h:\mathbb{R}\to\mathbb{R}$ be the payoff of a European claim such that the function $\mathbb{R}\to\mathbb{R}: x\mapsto h(\mathrm{e}^x)$ is globally Lipschitz with constant J. Then*

$$\left|\mathbb{E}\left[h\left(S(T)\right)\right]-\mathbb{E}[h(\widetilde{S}(T))]\right|\le J\,(3+\lambda T)\,\mathrm{e}^{\left(\frac{L_2}{2}-\lambda\right)^+ T}\mathbb{E}\left[\left|P(0)-\widetilde{P}(0)\right|\right] \tag{9}$$

The previous results only allow to obtain pricing errors of derivative assets which are Lipschitz with respect to the log-return, condition which is rather non-natural in the financial literature. As a simple example, notice that a plain vanilla call option $h(S(T)):=(S(T)-K)^+$ does not satisfy the previous Lipschitz condition, while instead being globally Lipschitz. Another example is the floating strike Asian option, with payoff $h(S(\cdot)):=(S_T-\int_0^T S_t\,dt)^+$. Nevertheless, our analysis can be extended to contingent claims which are globally Lipschitz in the natural variable S.

3 Using past information

The aim of the L^1-estimates of the previous section is to choose $\widetilde{P}_0$ in order to minimise the final error. As in [2], we assume to know all the past values of the price S_t for $t \in [-R, 0]$, where $R > 0$ is thus the width of the past observation window, while the process P remains unobserved also in the past.

It turns out that we can make the uncertainty on P decay exponentially with respect to the width R of the observation window. Again, we represent this uncertainty by defining the process $\widetilde{P}$, starting from the misspecified condition $\widetilde{P}_{-R}$ and following the dynamics

$$d\widetilde{P}_t = -\lambda \widetilde{P}_t \, dt + dZ_t, \quad t \in (-R, 0] \tag{10}$$

while the process P always follows the dynamics given by Equation (3). Notice that this time, as we can observe Z in the interval $[-R, 0]$, we have no uncertainty on this process.

Lemma 1 *Suppose that $\widetilde{P}$ and P have the dynamics (10) and (3), respectively, and that at time $-R$ their values are $\widetilde{P}_{-R}$ and P_{-R}, respectively. Then*

$$P_0 - \widetilde{P}_0 = \mathrm{e}^{-\lambda R} \left(P_{-R} - \widetilde{P}_{-R} \right) .$$

Now we are in the position of solving the following problem: for a given $\varepsilon > 0$ we want to find a minimum observation time R_0 such that the error when pricing a contingent claim h is less than ε. We present a result on European claims, possibly path-dependent, which are Lipschitz with respect to the log-return Z.

Corollary 1 *Suppose that σ and σ^2 admit locally Lipschitz first partial derivatives, bounded by L_1 and L_2 respectively, and let $V_1 = \mathbb{E}[|D_{-R} - \widetilde{D}_{-R}|]$.*

1. *If $h : C^0[0, T] \to \mathbb{R}$ is the payoff of a path-dependent claim such that the function $C^0[0, T] \to \mathbb{R} : f \mapsto h(\mathrm{e}^f)$ is globally Lipschitz with constant J, and $R > R_0$, where*

$$R_0 := \left(\frac{L_1^2}{2\lambda} + \frac{L_2}{2\lambda} - 1 \right)^+ T + \frac{1}{\lambda} \log \frac{J \, (5 + \lambda T) \, V_1}{\varepsilon} \tag{11}$$

   then

$$\left| \mathbb{E} \left[h \left(S(\cdot) \right) \right] - \mathbb{E}[h(\widetilde{S}(\cdot))] \right| < \varepsilon \ . \tag{12}$$

2. *If $h : \mathbb{R} \to \mathbb{R}$ is the payoff of a European claim such that the function $\mathbb{R} \to \mathbb{R} : x \mapsto h(\mathrm{e}^x)$ is globally Lipschitz with constant J, then in order for (12) to hold it is sufficient that $R > R_0$, where now*

$$R_0 := \left(\frac{L_2}{2\lambda} - 1 \right)^+ T + \frac{1}{\lambda} \log \frac{J \, (3 + \lambda T) \, V_1}{\varepsilon} \tag{13}$$

We can prove similar results for h globally Lipschitz with respect to the natural variable S.

We are now going to consider a determination of σ that satisfy our assumptions, thereby calculating explicitly the density f and then the width R of the past window: this will be done comparing the old robustness results from [2] with the ones presented in this note.

Suppose that

$$\sigma(P) = \min\left\{\sqrt{a+bP^2}, N\right\} \ ,$$

where $a > 0$, $b > 0$ and $N > 0$ are constants, with $a < N^2$. We know from [2] that the unique invariant measure for the process P has density given by the formula

$$f(x) = \begin{cases} K_1 \mathrm{e}^{-\frac{\lambda(N^2-a)}{bN^2} - \frac{N^2}{4\lambda}} N^{\frac{2\lambda}{b}} \mathrm{e}^{-x} (a+bx^2)^{-\frac{\lambda}{b}-1} & \text{if } |x| \le \sqrt{\frac{N^2-a}{b}} \\ \frac{K_1}{N^2} \mathrm{e}^{-\frac{\lambda}{N^2}\left(x+\frac{N^2}{2\lambda}\right)^2} & \text{if } |x| \ge \sqrt{\frac{N^2-a}{b}} \end{cases}$$

where K_1 is a convenient constant. As in [8] and [2], we take

$$a = 0.04, \quad b = 0.2, \quad \lambda = 1, \quad N = 1$$

so we have

$$L_1 = \sup_{x\in\mathbb{R}} \left|\frac{\partial\sigma}{\partial x}\right| = \frac{\sqrt{b(N^2-a)}}{N} = 0.438178$$

and

$$L_2 = \sup_{x\in\mathbb{R}} \left|\frac{\partial\sigma^2}{\partial x}\right| = 2\sqrt{b(N^2-a)} = 0.876356$$

and we obtain

$$V_1 = \mathbb{E}[|P - m_P|] = 0.116144 \ .$$

We want to find R such that the pricing error is less than $\varepsilon = 10^{-2}$, both for a path-dependent contingent claim as well as for a European one, both with Lipschitz constant $J = 1$. By taking different maturities, we find the results in Table 1: we indicate with R_{HV} the observation window obtained with the original estimates of [2] and with R the observation window obtained with the estimates (11) and (13) of Corollary 1.

	path-dependent		European	
T	R_{HV}	R	R_{HV}	R
0.25	3.971	4.110	3.611	3.631
0.5	5.157	4.157	4.438	3.705
1.0	8.943	4.244	7.504	3.839
2.0	22.167	4.398	19.288	4.062
3.0	42.927	4.532	38.608	4.244
4.0	71.223	4.649	65.464	4.398
5.0	107.055	4.755	99.856	4.532

Table 1: Time to wait (in years) for a precision $\varepsilon = 0.01$: with R the estimate with the current method, with R_{HV} with the one in [2].

We can see a huge improvement of the new results presented here over those in [2], which is evident especially for longer maturities: in fact, while in order to price a 5-years contingent claim with an error of less than $\varepsilon = 10^{-2}$ with the old estimates from [2] one needed an observation window of more than a century, with the results of this note one knows that the necessary time window is really less than 5 years long.

References

[1] Blaka Hallulli, V., T. Vargiolu, *Financial models with stochastic dependence on the past: a survey.* In *Applied industrial mathematics in Italy*, editors M. Primicerio et al., Series on Advances in Mathematics for Applied Sciences 69, World Scientific (2005), 348–359.

[2] Blaka Hallulli, V., T. Vargiolu, *Robustness of the Hobson-Rogers model with respect to the offset function.* In Proceedings of the Ascona '05 Seminar on Stochastic Analisys, Random Fields and Application, R. C. Dalang, M. Dozzi, F. Russo, editors, Birkhäuser (2007), 469–492.

[3] K. Burdzy, *Some path properties of iterated Brownian motion.* Seminar on Stochastic Processes, E. Cinlar (ed.) et al., Birkhäuser, Boston (1993), 67–87.

[4] Chiarella, C., Kwon, K., *A complete Markovian stochastic volatility model in the HJM framework.* Asia-Pacific Financial Markets 7/4 (2000), 293–304.

[5] Di Francesco M., Pascucci A., *On the complete model with stochastic volatility by Hobson and Rogers.* Proc. R. Soc. Lond. A Vol. 460 (2004), 3327–3338.

[6] Figà-Talamanca, G., Guerra, M.L., *Fitting prices with a complete model.* Journal of Banking and Finance 30/1 (2006), 247–258.

[7] Foschi, P., A. Pascucci, *Calibration of a path-dependent volatility model: empirical tests.* Comput. Statist. Data Anal., Volume 53 (2009), 2219–2235.

[8] Foschi, P., A. Pascucci, *Path dependent volatility.* Decisions in Economics and Finance 31/1 (2007), 1–20.

[9] Has'minskiĭ, R.Z., "Stochastic stability of differential equations". Alphen den Rijn, Sijthoff & Noordhoff, 1980.

[10] Hobson, D.G., L.C.G. Rogers, *Complete models with stochastic volatility.* Mathematical Finance 8/1 (1998), 27–48.

[11] Kawai, R., *Sensitivity analysis and density estimation for the Hobson-Rogers stochastic volatility model.* International Journal of Theoretical and Applied Finance 12/3 (2009), 283–295.

[12] Rosestolato, M., T. Vargiolu, G. Villani, *Robustness for path-dependent volatility models.* Preprint (2010).

[13] Sekine, J., *Marginal distributions of some path-dependent stochastic volatility model.* Statistics and Probability Letters 78 (2008), 1846–1850.

The Liouville Theorem for conformal maps: old and new

Alessandro Ottazzi (*)

1 Introduction

This seminar is devoted to a classical theorem of Liouville, dated back to 1850. This theorem deals with the problem of mapping domains onto given domains in a controlled way. It is well known that in $\mathbb{R}^2$ every simply connected domain which is not the whole plane can be deformed onto the disc in a conformal way (this is the statement of the Riemann mapping theorem, dated back to 1851). The theorem of Liouville states that in the euclidean spaces of dimension at least 3 it is not possible to prove the Riemann mapping theorem, because there are only a few conformal maps.

Liouville proved the theorem for $n = 3$, with the extra assumption that the conformal maps have C^4 regularity. In 1960, Nevanlinna [7] proved the theorem for every $n \geq 3$, assuming the same regularity. In 1962, Gehring [2] gave a definition of conformal map in a weak sense, assuming that the maps were homeomorphisms only. He proved that these maps are in fact smooth, and therefore Liouville theorem follows with low regularity assumption. Later on, other versions of the theorem were proved dropping the injectivity assumption, see [3, 10].

In this seminar we sketch a proof of the Liouville theorem that holds under the hypothesis that mappings are at least of class C^2. The proof that we give is somehow a consequence of [5, Theorem 1, p. 333]. In a work in collaboration with B. Warhurst [8], we show that a Liouville type theorem holds in the context of Carnot groups. Our proof is based on a generalization of the argument that we present here, by means of Tanaka prolongation theory [11].

An extensive survey of the Liouville theorem in the setting of euclidean spaces and Riemannian manifolds can be found in the book [3]. For generalizations concerning the sub-Riemannian metric setting, see [1, 6, 9].

(*)Università Milano-Bicocca, Milano (Italy). E-mail: alessandro.ottazzi@unimib.it . Seminar held on 20 April 2011.

2 The Liouville Theorem

The conformal transformations are defined as those diffeomorphisms that preserve angles between the tangent vectors of any two intersecting curves in a domain. If the overlapping direction of the two tangent vectors is also preserved, the conformal map is called orientation preserving.

Definition 2.1 Let $\mathcal{U}$ be an open and connected subset of $\mathbb{R}^n$. We say that a C^1 homeomorphism $f : \mathcal{U} \to \mathbb{R}^n$ is a conformal map if for every $x \in \mathcal{U}$ we have

$$< Df(x)v, Df(x)w >= |\lambda(x)| < v, w >,$$

for every $v, w \in \mathbb{R}^n$. Here $Df(x)$ denotes the Jacobian matrix of f at x.

From the definition above, we easily see that

$$Df(x)^{tr} Df(x) = |\lambda(x)| I.$$

Note that $\det Df(x)^{tr} Df(x) = (\det Df(x))^2 = |\lambda(x)|^n$, whence $|\lambda(x)| = |\det Df(x)|^{2/n}$. The condition of conformality becomes equivalent to the following:

$$Df(x)^{tr} Df(x) = |\det Df(x)|^{2/n} I. \tag{2.2}$$

The equation above is usually referred to as Cauchy-Riemann system.

Exercise. If $n = 2$, show that condition (2.2) is equivalent to the Cauchy-Riemann equations for f, viewed as a complex valued function, and for the conjugate of f. In other words, conformal maps on domains of $\mathbb{R}^2$ coincide with holomorphic and anti-holomorphic homeomorphisms.

Note that (2.2) is equivalent to ask that for every $x \in \mathcal{U}$, the matrix $Df(x)$ lies in the conformal group

$$\mathrm{CO}(n) = \{A \in \mathrm{GL}(n, \mathbb{R}) \,:\, A^{tr} A = \lambda I, \text{ for some } \lambda > 0\}.$$

Using chain rule one sees that if f and g are conformal maps and if the image of g is contained in the domain of f, then $f \circ g$ is conformal.

The main purpose of this lecture is to demonstrate the following theorem, whose first version dates back to 1850. There exist several proofs of this result, in a variety of smoothness assumptions. Here we give a proof that it is not in the classical literature and that holds for C^2 conformal maps. This approach is of interest because it generalizes to wide contexts, the nature of which we shall discuss later.

We shall divide the proof into two steps and we focus mostly on the first one.

Theorem 2.3 *Let f be a twice differentiable conformal map from a domain $\mathcal{U} \subset \mathbb{R}^n$ into $\mathbb{R}^n$, with $n \geq 3$. Then f is the restriction to $\mathcal{U}$ of a Möbius transformation.*

Remark 2.4 Recall that Möbius transformations form a Lie transformation group generated by the following maps.

- Translations: $x \mapsto x + y$.
- Rotations: $x \mapsto Ax$, with $A \in \mathrm{O}(n)$.
- Dilations: $x \mapsto ax$, for some $a > 0$.
- Inversion on the sphere: $x \mapsto R^2 \frac{x-x_0}{\|x-x_0\|^2} + x_0$, with $x_0 \in \mathbb{R}^n$ and $R > 0$.

It is an instructive exercise to verify that all the maps listed above are conformal, namely that they satisfy (2.2). In particular, the first three preserve orientation, whereas inversion on the sphere is orientation reversing.

By means of Bruhat decomposition of semisimple Lie groups one can also show that the maps above form the Lie group

$$\mathrm{O}(1, n+1) = \{A \in \mathrm{GL}(n+2, \mathbb{R}) \,:\, A^{tr} J A = J\},$$

where

$$J = \begin{bmatrix} 1 & & \\ & & \\ & & -I_{n+1} \end{bmatrix}.$$

Proof. We divide the proof into two steps. In the first part we show that 1-parameter groups of C^∞ conformal maps are in one-to-one correspondence with the Lie algebra

$$\mathfrak{so}(1, n+1) = \{A \in \mathfrak{gl}(n+2, \mathbb{R}) \,:\, A^{tr} J + J A = 0\}.$$

In the second part we prove that if f is any twice differentiable conformal map, then it coincides with the action of an element in $\mathrm{O}(1, n+1)$.

First part. We start with a 1-parameter group of C^∞ conformal maps fixing the identity. This means that $f_t \circ f_s = f_{t+s}$, $f_0 = id$, and (2.2) holds for every x in $\mathcal{U}$ and for a neighborhood of 0 in the time variable.
Consider the vector field defined as $U(x) = \frac{d}{dt} f_t(x)|_{t=0}$ and write $U(x) = \sum_{i=1}^n u_i \partial_i$, with $\partial_i = \frac{\partial}{\partial x_i}$, for every $i = 1, \dots, n$. We call conformal such a vector field. We differentiate the left hand side of (2.2) to obtain

$$\begin{aligned} \frac{d}{dt} Df_t(x)^{tr} Df_t(x)|_{t=0} &= \frac{d}{dt} Df_t(x)^{tr}|_{t=0} + \frac{d}{dt} Df_t(x)|_{t=0} \\ &= \mathcal{S}(U)(x)^{tr} + \mathcal{S}(U)(x), \end{aligned} \tag{2.5}$$

where $\mathcal{S}(U)(x) = \partial_j u_i(x)$ is the Ahlfors operator. The right hand side of (2.2) yields

$$\begin{aligned} \frac{d}{dt} |\mathrm{det} Df_t(x)|^{2/n}|_{t=0} I &= \frac{2}{n} |\mathrm{det} Df_t(x)|^{2/n-1}|_{t=0} \cdot \frac{\mathrm{det} Df_t(x)}{|\mathrm{det} Df_t(x)|}|_{t=0} \cdot \frac{d}{dt} \mathrm{det} Df_t(x)|_{t=0} I \\ &= \frac{2}{n} \mathrm{trace}(\mathcal{S}(U)(x)) I. \end{aligned} \tag{2.6}$$

Exercise. Prove that $\frac{d}{dt}\det Df_t(x)|_{t=0} = \mathrm{trace}(\mathcal{S}(U)(x))$.

By putting equal (2.5) and (2.6) we finally obtain

$$\mathcal{S}(U)(x)^{tr} + \mathcal{S}(U)(x) = \frac{2}{n}\mathrm{trace}(\mathcal{S}(U)(x))I,$$

for every $x \in U$. In particular, $\mathcal{S}(U)(x) \in \mathfrak{co}(n)$, where

$$\mathfrak{co}(n) = \{A \in \mathfrak{gl}(n,\mathbb{R}) \,:\, A^{tr} + A = \lambda I, \text{ for some } \lambda \in \mathbb{R}\}$$

is the Lie algebra of $\mathrm{CO}(n)$. This condition characterizes conformal vector fields, in terms of a system of differential equations, that explicitly are

$$\partial_i u_j = -\partial_j u_i \qquad \partial_i u_i = \partial_j u_j,$$

with $i, j = 1, \dots, n$ and $i \neq j$. Note that since $\{\partial_i u_j(x)\}_{ij} \in \mathfrak{co}(n)$ for every $x \in \mathcal{U}$, then the higher order derivatives still do. Namely, $\{\partial^k_{jj_1\dots j_{k-1}} u_i(x)\}_{ij} \in \mathfrak{co}(n)$, for every fixed $j_1 \dots j_{k-1}$. We formalize these observations by means of the following definition. Set

$$\begin{aligned}(2.7)\qquad \mathfrak{h}^k = \{&T : \mathbb{R}^n \times \cdots \times \mathbb{R}^n \to \mathbb{R}^n \text{ multilinear and symmetric such that} \\ &\forall v_1, \dots, v_k \in \mathbb{R}^n, \text{the map } v \mapsto T(v, v_1, \dots, v_k) \text{ is in } \mathfrak{co}(n)\}.\end{aligned}$$

Notice that the map

$$(x_{j_1}, \dots, x_{j_{k+1}}) \mapsto \partial^{k+1}_{x_{j_1},\dots,x_{j_{k+1}}} u_i(x)$$

is in $\mathfrak{h}^k$ for every $x \in \mathcal{U}$. Moreover, the definition of $\mathfrak{h}^k$ is inductive. Indeed, $(v_1, \dots, v_k) \mapsto T(v, v_1, \dots, v_k)$ lies in $\mathfrak{h}^{k-1}$ for every $v \in \mathbb{R}^n$. In particular, if $\mathfrak{h}^j = \{0\}$ then $\mathfrak{h}^{j+l} = \{0\}$ for every $l \geq 0$. The space $\mathfrak{h}^k$ is called the $k-th$ Singer and Sternberg prolongation of $\mathfrak{co}(n)$ (note that $\mathfrak{h}^0 = \mathfrak{co}(n)$). If the sequence $\{\mathfrak{h}^k\}_{k\geq 0}$ is finite, then the space $\mathfrak{g} = \oplus_{k\geq 0}\mathfrak{h}^k$ is a Lie algebra, and it is isomorphic to the space of conformal vector fields. Namely, if $\bar{x} \in \mathcal{U}$, we write the formal Taylor series at $\bar{x}$ for the coefficients of U:

$$u_i \sim_{\bar{x}} u_i(\bar{x}) + \sum_{k=1}^{\infty} \frac{1}{k!} \sum_{j_1,\dots,j_k=1}^{n} a^i_{j_1,\dots,j_k}(x_{j_1} - \bar{x}_{j_1}) \cdots (x_{j_k} - \bar{x}_{j_k}),$$

for every $i = 1, \dots, n$ and where we denoted $a^i_{j_1,\dots,j_k} = \partial^k_{j_1,\dots,j_k} u_i(\bar{x})$. Therefore, if the prolongation sequence is finite, we conclude that the expansion above is a polynomial for every i, and so the conformal vector fields vary in a finite dimensional space, whose structure is defined by $\mathfrak{g}$.
Some linear algebra shows that $\mathfrak{g} = \mathfrak{h}^0 + \mathfrak{h}^1$ [4, page 9], because $\mathfrak{h}^2 = \{0\}$. Furthermore, $\mathfrak{g} = \mathfrak{so}(1, n+1)$.

Remark 2.8 If $n = 2$, the prolongation is infinite. This reflects the fact that in $\mathbb{R}^2$ the Liouville theorem is false. In fact, in $\mathbb{R}^2$ the Riemann mapping theorem holds, and it represents the counterpart of Liouville's. Roughly speaking, there are so many conformal maps in $\mathbb{R}^2$ that every proper simply connected subset of the plane can be conformally deformed to the unit disc.

Second part. In the first part we showed that every conformal vector field defined on a domain in $\mathbb{R}^n$ with $n \geq 3$ is the restriction of a globally defined polynomial vector field. The space of C^∞ globally defined conformal vector fields, say $\mathrm{Conf}(\mathbb{R}^n)$, is finite dimensional and isomorphic to $\mathfrak{so}(1, n+1)$. Let now f be a C^2 conformal map. Normalizing with translations, we may assume that $f(0) = 0$. For a vector field $U \in \mathrm{Conf}(\mathbb{R}^n)$, we claim that the push-forward f_*U is again a conformal vector field. Indeed, by definition $(f_*U)(x) = Df(f^{-1}(x))U(f^{-1}(x))$. If $U(x) = \frac{d}{dt}h_t(x)|_{t=0}$, then

$$(f_*U)(x) = \frac{d}{dt}(f \circ h_t \circ f^{-1})(x)|_{t=0}.$$

Since the composition of conformal maps is still conformal, the composition $f \circ h_t \circ f^{-1}$ defines a flow of conformal maps. Thus f_*U is a C^1 conformal vector field. A standard argument of mollification shows that f_*U is in fact C^∞ and therefore lies in $\mathrm{Conf}(\mathbb{R}^n)$. Call $\tau : \mathfrak{so}(1, n+1) \to \mathrm{Conf}(\mathbb{R}^n)$ the isomorphism whose existence was established in the first part. Then U lies in the image of τ and so does f_*U. Hence $\alpha := \tau^{-1} \circ Df \circ \tau$ is an automorphism of $\mathfrak{so}(1, n+1)$. So f is uniquely determined as the solution of the system

$$\begin{cases} Df = \tau \circ \alpha \circ \tau^{-1} \\ f(0) = 0. \end{cases}$$

We conclude that f is defined by an element in $\mathrm{Aut}(\mathfrak{so}(1, n+1)) = \mathrm{O}(1, n+1)$. □

3 Conformal maps on Carnot groups

Let $\mathbb{G}$ be a stratified nilpotent Lie group with identity e. This means that its Lie algebra $\mathfrak{g}$ admits an s-step stratification

$$\mathfrak{g} = V_1 \oplus \cdots \oplus V_s,$$

where $[V_j, V_1] = V_{j+1}$, for $1 \leq j \leq s$, and with $V_s \neq \{0\}$ and $V_{s+1} = \{0\}$. To avoid degeneracies, we assume $\mathfrak{g}$ to have at least dimension two, which is reasonable to our purposes.

Given a point $p \in \mathbb{G}$ we denote by l_p the left translation by p. An element X in the Lie algebra $\mathfrak{g}$ can be considered as a tangent vector at the identity. Such a vector induces the left invariant vector field that at a point $p \in \mathbb{G}$ is given by $(l_p)_*|_e(X)$. This vector field will still be denoted by X, unless confusion might arise. The set of all left invariant vector fields with the bracket operation is isomorphic to $\mathfrak{g}$ and it inherits the stratification of $\mathfrak{g}$. The sub-bundle $\mathcal{H} \subseteq TG$ where $\mathcal{H}_p = (l_p)_*|_e(V_1)$ is called the *horizontal distribution*. A scalar product $\langle\, , \,\rangle$ on V_1 defines a left invariant scalar product on each $\mathcal{H}_p$ by setting

$$\langle v, w \rangle_p = \langle (l_{p^{-1}})_*|_p(v), (l_{p^{-1}})_*|_p(w) \rangle \tag{3.1}$$

for all $v, w \in \mathcal{H}_p$. The left invariant scalar product gives rise to a left invariant sub-Riemannian metric d on $\mathbb{G}$. The Carathéodory-Chow-Rashevsky Theorem shows that the bracket generating property implies that any two points in $\mathbb{G}$ can be joined by a

horizontal path, i.e., an absolutely continuous path whose tangents belong to the horizontal distribution. The sub-Riemannian metric is then defined by setting

$$d(p,q) := \inf \int_0^1 \sqrt{\langle \dot\gamma(t), \dot\gamma(t)\rangle_{\gamma(t)}}dt,$$

where the infimum is taken along all horizontal curves $\gamma : [0,1] \to \mathbb{G}$ such that $\gamma(0) = p$ and $\gamma(1) = q$. We call $(\mathbb{G}, d)$ a *Carnot group*. Notice that euclidean spaces are in particular (abelian) Carnot groups.

Let now f be a diffeomorphism from an open and connected subset $\mathcal{U}$ of $\mathbb{G}$ into $\mathbb{G}$. Define $D_{\mathbb{G}} f(p) = (l^{-1}_{f(p)} \circ f \circ l_p)_{*e}$. We say that f is conformal if $D_{\mathbb{G}} f(p)|_{V_1} \in CO(m)$, where m denotes the dimension of V_1. This definition of conformality agrees with the one we gave in the previous section in the euclidean setting. In the noncommutative case this notion expresses the fact that the angle preserving property is asked only for curves that remain tangent to the horizontal distribution. The following theorem holds.

Theorem 3.2 (A. Ottazzi, B. Warhurst [8]) *The conformal maps defined on a domain of a Carnot group $\mathbb{G} \neq \mathbb{R}^2$ form a finite dimensional space.*

Example 3.3 Let $\mathfrak{h} = \mathrm{span}\{X, Y, Z\}$ be the Lie algebra with nonzero bracket $[X, Y] = Z$. The connected and simply connected nilpotent Lie group that has $\mathfrak{h}$ as Lie algebra is the three dimensional Heisenberg group that we denote by $\mathbb{H}$. Global coordinates on the group are $(x, y, z) = \exp(xX + yY + zZ)$, and the noncommutative product of two points is

$$(x, y, z)(x', y', z') = (x + x', y + y', z + z' + \frac{1}{2}xy' - \frac{1}{2}x'y).$$

A choice of a scalar product on $V_1 = \mathrm{span}\{X, Y\}$ leads to the definition of a sub-Riemannian metric, that in turn it allows us to define the conformal maps. The group of conformal maps in this case is given by $\mathrm{U}(1,2)$, whereas those that are orientation preserving lie in $\mathrm{SU}(1,2)$ (see [6]). We recall that

$$\mathrm{U}(1,2) = \{A \in \mathrm{GL}(3,\mathbb{C}) \,:\, A^*JA = A\},$$

where A^* denotes the conjugate transpose of A. The group $\mathrm{SU}(1,2)$ is the subgroup of members of $\mathrm{U}(1,2)$ of determinant 1.

References

[1] Luca Capogna and Michael Cowling, *Conformality and q-harmonicity in Carnot groups.* Duke Math. J. 135/3 (2006), 455–479.

[2] F. W. Gehring, *Rings and quasiconformal mappings in space.* Trans. Amer. Math. Soc. 103 (1962), 353–393.

[3] Iwaniec, Tadeusz and Martin, Gaven, "Geometric function theory and non-linear analysis". Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, New York, 2001.

[4] Shoshichi Kobayashi, "Transformation groups in differential geometry". Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the 1972 edition.

[5] Kobayashi, Shoshichi and Nomizu, Katsumi, "Foundations of differential geometry. Vol. II". Interscience Tracts in Pure and Applied Mathematics, No. 15 Vol. II. Interscience Publishers John Wiley & Sons, Inc., New York-London-Sydney, 1969.

[6] A. Korányi and H. M. Reimann, *Quasiconformal mappings on the Heisenberg group.* Invent. Math. 80/2 (1985), 309–338.

[7] R. Nevanlinna, *On differentiable mappings.* In *Analytic Functions*, Princeton Math. Ser. 24 , Princeton Univ. Press, Princeton, 1960, pp. 3–9.

[8] Alessandro Ottazzi and Ben Warhurst, *Contact and* 1*-quasiconformal maps on Carnot groups.* J. Lie Theory 21 (2011), 787–811.

[9] Pierre Pansu, *Métriques de Carnot-Carathéodory et quasiisométries des espaces symétriques de rang un.* Ann. of Math. 129/2 (1989), 1–60.

[10] Yuri G. Reshetnyak, *Liouville's conformal mapping theorem under minimal regularity hypotheses.* Sib. Math. J. 8 (1967), 631–634.

[11] Noboru Tanaka, *On differential systems, graded Lie algebras and pseudogroups.* J. Math. Kyoto Univ. 10 (1970), 1–82.

Large Deviations in Probability Theory

Markus Fischer (*)

Abstract. In probability theory, the term large deviations refers to an asymptotic property of the laws of families of random variables depending on a large deviations parameter. The aim of these notes is to give an idea of the theory of large deviations, illustrating it by elementary examples as well as in the context of a class of mean field models.

1 Introduction

Consider the following family of random experiments: Given $n \in \mathbb{N}$, toss n identical coins and count the number of coins that land heads up. Denote that (random) number by S_n. Then S_n/n is the empirical mean, here equal to the empirical probability of getting heads. What can be said about S_n/n for n large?

In order to make that question precise (and find answers), let us specify a mathematical model for the coin tossing experiments. Let $X_1, X_2, \ldots$ be $\{0,1\}$-valued independent and identically distributed (i.i.d.) random variables defined on some probability space $(\Omega, \mathcal{F}, \mathrm{P})$. In particular, any X_i has Bernoulli distribution with common parameter $p \doteq \mathrm{P}(X_1 = 1)$. Interpret $X_i(\omega) = 1$ as saying that coin i at realization $\omega \in \Omega$ lands head up. Then $S_n = \sum_{i=1}^n X_i$.

A first answer to the above question is given by the *law of large numbers*, which applies in its strong and thus also weak version, stating that

$$\frac{S_n}{n} \overset{n\to\infty}{\longrightarrow} p \quad \text{with probability one / in probability.}$$

In particular, by the weak law of large numbers, for all $\epsilon > 0$,

$$\mathrm{P}\left\{S_n/n - p \geq \epsilon\right\} \overset{n\to\infty}{\longrightarrow} 0.$$

We can obtain more information about the asymptotic behavior of those *deviation* probabilities.

First notice that S_n has binomial distribution with parameters n, p, that is,

$$\mathrm{P}\left\{S_n = k\right\} = \tfrac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, \quad k \in \{0, \ldots, n\}.$$

(*)Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: fischer@math.unipd.it . Seminar held on 4 May 2011.

By Stirling's formula, asymptotically for large n,

$$\mathrm{P}\{S_n = k\} \simeq \frac{\sqrt{2\pi n}\, n^n e^{-n}}{\sqrt{2\pi k}\, k^k e^{-k} \sqrt{2\pi(n-k)}(n-k)^{n-k} e^{-(n-k)}} p^k (1-p)^{n-k}.$$

Proceeding formally, if $k \simeq n\,x$ for some $x \in (0,1)$, then

$$\begin{aligned} \log \mathrm{P}\{S_n = k\} &\simeq -\frac{1}{2}\left(\log(2\pi) + \log(x) + \log(1-x) + \log(n)\right) \\ &\quad - n\,x \log(\tfrac{x}{p}) - n(1-x)\log(\tfrac{1-x}{1-p}), \\ \frac{1}{n}\log \mathrm{P}\{S_n = k\} &\simeq -\left(x \log(\tfrac{x}{p}) + (1-x)\log(\tfrac{1-x}{1-p})\right). \end{aligned}$$

The expression $x \log(\frac{x}{p}) + (1-x)\log(\frac{1-x}{1-p})$ gives the *relative entropy* of the Bernoulli distribution with parameter x w.r.t. the Bernoulli distribution with parameter p, which is minimal and zero iff $x = p$. The asymptotic equivalence

$$\frac{1}{n}\log \mathrm{P}\{S_n = k\} \simeq -\left(x \log(\tfrac{x}{p}) + (1-x)\log(\tfrac{1-x}{1-p})\right), \quad k \simeq n\,x,$$

shows that the probabilities of the events $\{S_n/n - p \geq \epsilon\}$ converge to zero *exponentially fast* with rate (up to arbitrary small corrections)

$$-\left((p+\epsilon)\log(\tfrac{p+\epsilon}{p}) + (1-p-\epsilon)\log(\tfrac{1-p-\epsilon}{1-p})\right),$$

which corresponds to $x = p + \epsilon$, the rate of slowest convergence. To be a bit more precise, with $\delta > 0$ small, rewrite the event $\{S_n/n - p \geq \epsilon\}$ as

$$\{S_n/n - p \geq \epsilon\} = \cup_{k=1}^{\infty} \{S_n/n - p \in [\epsilon + (k-1)\delta, \epsilon + k\delta)\}.$$

The exponential rate of decay of the probabilities $\mathrm{P}(S_n/n - p \geq \epsilon)$ as $n \to \infty$ is governed by the rate of decay of $\mathrm{P}(S_n/n - p \in [\epsilon, \epsilon + \delta))$, the probabilities of the sub-events that converge most slowly.

Events like $\{S_n/n - p \geq \epsilon\}$ describe *large deviations* from the law of large numbers limit, in contrast to the *fluctuations* captured by the *central limit theorem*, which says here that $\sqrt{n}(S_n/n - p)$ is asymptotically normal.

In Section 2 we collect basic definitions and results of the theory of large deviations, while Section 3 contains a sketch of two classical theorems, named after Cramér and Sanov, respectively, which deal with empirical means and empirical measures of i.i.d. random variables. In Section 4, we present a large deviation analysis for a class of weakly interacting (or mean field) systems. Most of the material of the first three sections and a number of other important results and examples can be found in the survey article [10]. The coin tossing example and related models are extensively treated in the book [7]. A standard reference on large deviations is the book [5], another the text [8]. An alternative approach, based on the Laplace principle, is developed in [6]. Section 4 is based on our joint work [2].

2 Basic definitions and results

Let $(\xi^n)_{n\in\mathbb{N}}$ be a family of random variables with values in a Polish space $\mathcal{S}$ (i.e., a topological space that allows for a complete and separable metric). Let $I\colon \mathcal{S}\to[0,\infty]$ be a function with compact sublevel sets (i.e., $\{x\in\mathcal{S} : I(x)\le c\}$ compact for all $c\in\mathbb{R}$). Such a function is lower semicontinuous and is called a (good) *rate function*.

Definition 1 The family $(\xi^n)_{n\in\mathbb{N}}$ satisfies a *large deviation principle* with rate function I iff for all $G\in\mathcal{B}(\mathcal{S})$,

$$\begin{aligned}-\inf_{x\in G^\circ} I(x) &\le \liminf_{n\to\infty}\frac{1}{n}\log \mathrm{P}\left\{\xi^n\in G\right\}\\ &\le \limsup_{n\to\infty}\frac{1}{n}\log \mathrm{P}\left\{\xi^n\in G\right\} \le -\inf_{x\in \mathrm{cl}(G)} I(x),\end{aligned}$$

where G° denotes the open interior and $\mathrm{cl}(G)$ the closure of G.

If $\inf_{x\in G^\circ} I(x)=\inf_{x\in\mathrm{cl}(G)} I(x)$, then G is called an *I-continuity set* and

$$\lim_{n\to\infty}\frac{1}{n}\log \mathrm{P}\left\{\xi^n\in G\right\} = -\inf_{x\in G} I(x) \doteq -I(G).$$

Notice the analogy with the portmanteau theorem, which characterizes weak convergence of probability measures.

A large deviation principle is a distributional property. Writing P_n for the law of ξ^n, the family (ξ^n) satisfies a large deviation principle with rate function I iff for all $G\in\mathcal{B}(\mathcal{S})$,

$$-\inf_{x\in G^\circ} I(x) \le \liminf_{n\to\infty}\frac{1}{n}\log \mathrm{P}_n(G) \le \limsup_{n\to\infty}\frac{1}{n}\log \mathrm{P}_n(G) \le -\inf_{x\in\mathrm{cl}(G)} I(x).$$

A large deviation principle gives a rough description of the asymptotic behavior of the probabilities of *rare events*: for all continuity sets G,

$$\mathrm{P}_n(G) = e^{-n(I(G)+o(1))}.$$

When a large deviation principle holds, it is sometimes possible to obtain sharper asymptotics of the form $\mathrm{P}_n(G)=e^{-nI(G)+o(1)}$.

The (good) rate function of a large deviation principle is uniquely determined. If I is the rate function of a large deviation principle, then $\inf_{x\in\mathcal{S}} I(x)=0$ and $I(x^*)=0$ for some $x^*\in\mathcal{S}$. If I has a unique minimizer, then the large deviation principle implies a corresponding law of large numbers.

A large deviation principle is transferred under continuous mappings ("contraction principle"): Let $\mathcal{Y}$ be a Polish space and $\psi\colon\mathcal{S}\to\mathcal{Y}$ be a *continuous* function. If (ξ^n) satisfies a large deviation principle with rate function I, then $(\psi(\xi^n))$ satisfies a large deviation principle with (good) rate function $J(y)\doteq\inf_{x\in\psi^{-1}(y)} I(x)$.

Returning to the coin tossing example, we have that $\xi^n\doteq S_n/n$, $n\in\mathbb{N}$, satisfies a large deviation principle in $\mathcal{S}\doteq\mathbb{R}$ (or $\mathcal{S}\doteq[0,1]$) with rate function

$$I(x)=\begin{cases} x\log(\frac{x}{p})+(1-x)\log(\frac{1-x}{1-p}) & \text{if } x\in[0,1],\\ \infty & \text{otherwise.}\end{cases}$$

The function I is finite and continuous on $[0,1]$, and convex on $\mathbb{R}$.

An alternative (and under mild hypotheses equivalent) characterization of large deviations is the following.

Definition 2 The family (ξ^n) satisfies a *Laplace principle* with rate function I iff for all $F \in \mathbf{C}_b(\mathcal{S})$,

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbf{E}\left[\exp\left(-n \cdot F(\xi^n)\right)\right] = \inf_{x\in\mathcal{S}} \left\{I(x) + F(x)\right\}.$$

In Definition 2 it is equivalent to require that for all $F \in \mathbf{C}_b(\mathcal{S})$,

$$\lim_{n\to\infty} \frac{1}{n} \log \int_{\mathcal{S}} \exp\left(n \cdot F(x)\right) \mathrm{P}_n(dx) = \sup_{x\in\mathcal{S}} \left\{F(x) - I(x)\right\}.$$

If $\mathcal{S} = \mathbb{R}^d$ and we take $F(x) = \theta \cdot x$ (although such F is not bounded), then on the right-hand side above we have the *Legendre transform* of I at $\theta \in \mathbb{R}^d$.

The name "Laplace principle" derives from the analogy with *Laplace's method*: $\lim_{n\to\infty} \frac{1}{n} \log \int_0^1 e^{nf(x)} dx = \max_{x\in[0,1]} f(x)$ for all $f \in \mathbf{C}([0,1])$.

Laplace principle and large deviation principle are equivalent in the sense that a large deviation principle holds if and only if a Laplace principle holds, and the rate function is the same for both principles.

In Section 4, we will prove a Laplace principle starting from a variational representation of the Laplace functionals. Let $\mu \in \mathcal{P}(\mathcal{S})$. Then for all $g\colon \mathcal{S} \to \mathbb{R}$ bounded and *measurable*,

$$-\log \int_{\mathcal{S}} \exp\left(-g(x)\right) \mu(dx) = \inf_{\nu\in\mathcal{P}(\mathcal{S})} \left\{R(\nu\|\mu) + \int_{\mathcal{S}} g(x)\nu(dx)\right\},$$

where $R(.\|.)$ is *relative entropy*, that is,

$$R(\nu\|\mu) = \begin{cases} \int_{\mathcal{S}} \log\left(\frac{d\nu}{d\mu}(x)\right) \nu(dx) & \text{if } \nu \ll \mu, \\ \infty & \text{else.} \end{cases}$$

The infimum in the variational formula is attained at $\nu^* \in \mathcal{P}(\mathcal{S})$ given by

$$\frac{d\nu^*}{d\mu}(x) = \frac{\exp\left(-g(x)\right)}{\int_{\mathcal{S}} \exp\left(-g(y)\right) \mu(dy)}, \quad x \in \mathcal{S}.$$

3 Empirical means and empirical measures of i.i.d. systems

Let $X_1, X_2, \ldots$ be $\mathbb{R}$-valued i.i.d. random variables with $\mathbf{E}[X_1] = c$. As in the case of coin flipping, set $S_n \doteq \sum_{i=1}^n X_i$ and consider the asymptotic behavior of S_n/n, the empirical mean of $X_1, \ldots, X_n$, $n \in \mathbb{N}$. By the law of large numbers, $S_n/n \overset{n\to\infty}{\longrightarrow} c$ with probability one. Let ϕ be the *moment generating function* of $X_1, X_2, \ldots$, that is,

$$\phi(t) \doteq \mathbf{E}\left[e^{t\,X_1}\right], \quad t \in \mathbb{R}.$$

Theorem (Cramér) *Assume that $\phi(t)$ is finite for all $t \in \mathbb{R}$. Then $(S_n/n)_{n\in\mathbb{N}}$ satisfies a large deviation principle with rate function I given by*

$$I(x) \doteq \sup_{t\in\mathbb{R}} \{t \cdot x - \log(\phi(t))\}.$$

The rate function I is the Legendre transform of $\log\phi$, the *cumulant generating function* of the common distribution. If the X_i are $\{0,1\}$-Bernoulli with parameter p, then $\phi(t) = 1 - p + p\,e^t$ and $I(x) = x\log(\frac{x}{p}) + (1-x)\log(\frac{1-x}{1-p})$ for $x \in [0,1]$ ($+\infty$ outside $[0,1]$). If the X_i have normal distribution with mean 0 and variance σ^2, then $\phi(t) = e^{\sigma^2 t^2/2}$, while $I(x) = x^2/(2\sigma^2)$.

Cramér's theorem expresses the rate function I as the convex dual of the cumulant generating function $\log\phi$. Both functions are convex, $\log\phi$ is strictly convex, I is finite, strictly convex and infinitely differentiable on the interior of its support, while $I(x) = \infty$ for $x \notin [\operatorname{essinf} X_1, \operatorname{esssup} X_1]$. Moreover, $I(c) = 0$ (cf. law of large numbers) , $I'(c) = 0$ and $I''(c) = 1/\operatorname{var}(X_1)$ (cf. central limit theorem).

Sketch of proof for Cramér's theorem. Assume that $\mathbf{E}[X_1] = 0$ and consider, for $x > 0$, the probabilities of the events $\{S_n/n \geq x\}$. By Markov's inequality, for any $t > 0$,

$$\mathrm{P}\{S_n \geq n\,x\} = \mathrm{P}\left\{e^{t\,S_n} \geq e^{t\,n\,x}\right\} \leq e^{-t\,n\,x}\,\mathbf{E}\left[e^{t\,S_n}\right] = e^{-t\,n\,x}(\phi(t))^n.$$

Since this inequality holds for any $t > 0$ and $\log(.)$ is non-decreasing, we obtain the upper bound

$$\limsup_{n\to\infty} \frac{1}{n}\log \mathrm{P}\{S_n/n \geq x\} \leq -\sup_{t>0}\{t\,x - \log(\phi(t))\},$$

Now optimize over $t > 0$ in the inequality above (this actually yields the lower bound): the optimal $t = t_x$ is determined by $x = (\log\phi)'(t_x)$, that is,

$$x = \frac{\phi'(t_x)}{\phi(t_x)} = \frac{\mathbf{E}\left[X_1 e^{t_x X_1}\right]}{\mathbf{E}\left[e^{t_x X_1}\right]}.$$

The parameter t_x corresponds to a *change of measure* with exponential density. Under the new measure, the random variable X_1 has expected value x instead of zero. The rare event becomes typical! □

Cramér's theorem gives a large deviation principle for the empirical means of i.i.d. real-valued random variables. A related classical result is Sanov's theorem, which gives a large deviation principle for the empirical measures. Let $\mathcal{S}$ be a Polish space, and let $Y_1, Y_2, \ldots$ be $\mathcal{S}$-valued i.i.d. random variables on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ with common distribution $\mu \in \mathcal{P}(\mathcal{S})$. For $n \in \mathbb{N}$ let μ^n be the *empirical measure* of $Y_1, \ldots, Y_n$, that is,

$$\mu^n(\omega) \doteq \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i(\omega)}, \quad \omega \in \Omega,$$

where δ_y denotes the Dirac measure concentrated in $y \in \mathcal{S}$.

Theorem (Sanov) *The family $(\mu^n)_{n\in\mathbb{N}}$ of $\mathcal{P}(\mathcal{S})$-valued random variables satisfies a large deviation principle with rate function $I\colon \mathcal{P}(\mathcal{S}) \to [0,\infty]$ given in terms of relative entropy by*

$$I(\theta) = R\,(\theta\|\mu)\,.$$

Consider the particular case where $\mathcal{S} = \{s_1, \dots, s_k\}$ is finite. Then any $\mu \in \mathcal{P}(\mathcal{S})$ corresponds to a probability vector $(p_1, \dots, p_k)$. The empirical measure μ^n is determined by the observed frequencies of $s_1, \dots, s_k$ in n trials, which have multinomial distribution:

$$\mathrm{P}(n; f_1, \dots, f_k) = \frac{n!}{f_1! \cdots f_k!} p_1^{f_1} \cdots p_k^{f_k}.$$

An application of Stirling's formula with $f_i \simeq n\,x_i$ yields

$$\log \mathrm{P}(n; f_1, \dots, f_k) = -n \left(\sum_{i=1}^{k} x_i \log(\tfrac{x_i}{p_i}) \right) + o(n).$$

This reasoning can be made into a proof of Sanov's theorem.

Still assuming that $\mathcal{S}$ is finite, let $f : \mathcal{S} \to \mathbb{R}$ be a function. Define the mapping $\Psi\colon \mathcal{P}(\mathcal{S}) \to \mathbb{R}$ by $\Psi(\nu) \doteq \nu(f) = \sum_{s\in\mathcal{S}} f(s)\nu(s)$. Then Ψ is continuous. Let $X_1, X_2, \dots$ be $\mathcal{S}$-valued i.i.d. random variables with common distribution μ. Then by Sanov's theorem and the contraction principle, $(\frac{1}{n}\sum_{i=1}^n f(X_i))_{n\in\mathbb{N}}$ satisfies a large deviation principle with rate function

$$J(s) = \inf_{\nu\in\mathcal{P}(\mathcal{S}):\nu(f)=s} R\,(\nu\|\mu)\,.$$

The expression for the rate function J can be seen as an instance of the maximum entropy principle (also known as Jaynes's principle). If $\mathcal{S}$ is finite, the contraction principle allows to derive Cramér's theorem from Sanov's theorem.

4 Large deviations for a class of mean field systems

For $N \in \mathbb{N}$ we are given a system of N weakly interacting particles. The evolution of the state of particle $i \in \{1, \dots, N\}$ is described in terms of the Itô stochastic differential equation (SDE)

$$dX^{i,N}(t) = b\big(X^{i,N}(t), \mu^N(t)\big)dt + \sigma\big(X^{i,N}(t), \mu^N(t)\big)dW^i(t), \tag{1}$$

where $W^1, W^2, \dots$ are *independent* standard Wiener processes and weak interaction is through $\mu^N(t)$, the *empirical measure* at time $t \in [0,T]$:

$$\mu^N(t) \doteq \frac{1}{N}\sum_{i=1}^{N} \delta_{X^{i,N}(t)},\ t \in [0,T], \qquad\qquad \mu^N \doteq \frac{1}{N}\sum_{i=1}^{N} \delta_{X^{i,N}}.$$

Example (Gradient systems for Brownian particles)

$$dX^{i,N}(t) = -\nabla U\left(X^{i,N}(t)\right)dt - \frac{1}{N}\sum_{j=1}^{N}\nabla_1 V\left(X^{i,N}(t), X^{j,N}(t)\right)dt + dW^i(t),$$

where U environment potential, V symmetric interaction potential with $V(x,x)=0$.

Example (Stabilization through monotone dependence)

$$dX^{i,N}(t) = \hat{b}\left(X^{i,N}(t)\right)dt + \tilde{b}\left(\frac{1}{N}\sum_{j=1}^{N}X^{j,N}(t)\right)dt + \sigma\left(X^{i,N}(t)\right)dW^i(t),$$

$\tilde{b}\colon \mathbb{R}\to(-\infty,0]$ decreasing (component-wise monotonicity in d dimensions).

Our aim is to describe the asymptotic behavior of the N-particle systems as the number of particles N tends to infinity in terms of a Laplace principle for the family $(\mu^N)_{N\in\mathbb{N}}$.

It is well-known [9, for instance] that the empirical measures $(\mu^N)_{N\in\mathbb{N}}$ satisfy a law of large numbers, that is, μ^N converges to μ as N tends to infinity, where μ is the law of the "nonlinear diffusion"

$$dX(t) = b\big(X(t), \mathrm{Law}(X(t))\big)dt + \sigma\big(X(t), \mathrm{Law}(X(t))\big)dW(t), \tag{2}$$

thus $\mu(t) = \mathrm{Law}(X(t))$. The forward equation for μ (or the law of X) is the nonlinear *McKean-Vlasov equation*

$$\frac{d}{dt}\mu(t) = \mathcal{L}(\mu(t))^*\mu(t),$$

where $\mathcal{L}(\mu(t))^*$ is the formal adjoint of the infinitesimal generator $\mathcal{L}(\mu(t))$,

$$\mathcal{L}(\nu)(f)(x) \doteq \big\langle b\big(x,\nu\big), \nabla f(x)\big\rangle + \frac{1}{2}\sum_{j,k=1}^{d}(\sigma\sigma^{\mathsf{T}})_{jk}\big(x,\nu\big)\frac{\partial^2 f}{\partial x_j \partial x_k}(x).$$

In [4], large deviations from the McKean-Vlasov limit for weakly interacting processes of the form

$$dX^{i,N}(t) = b\big(X^{i,N}(t), \mu^N(t)\big)dt + \sigma\big(X^{i,N}(t)\big)dW^i(t)$$

are derived under mild regularity and general growth conditions and the assumption that $\sigma\sigma^{\mathsf{T}}$ be non-degenerate. The techniques include exponential probability bounds, a large deviation principle (LDP) for independent time-inhomogeneous diffusions (freezing of μ^N), time discretization and projective limits (LDP from LDP for approximating systems), and a martingale problem.

Here we establish a Laplace principle, using weak convergence and ideas from stochastic optimal control. In order to establish a Laplace principle, we have to show that for all $F\in\mathbf{C}_b(\mathcal{S})$,

$$\lim_{N\to\infty} -\frac{1}{N}\log\mathbf{E}\left[\exp\left(-N\cdot F(\mu^N)\right)\right] = \inf_{x\in\mathcal{S}}\left\{I(x)+F(x)\right\},$$

where $\mathcal{S} = \mathcal{P}(\mathcal{X})$, the space of probability measures on the path space $\mathcal{X} \doteq \mathbf{C}([0,T], \mathbb{R}^d)$. Recall from Section 2 the variational formula

$$-\log \int_{\mathcal{S}} \exp\left(-g(x)\right) \mu(dx) = \inf_{\nu \in \mathcal{P}(\mathcal{S})} \left\{ R(\nu \| \mu) + \int_{\mathcal{S}} g(x) \nu(dx) \right\},$$

which holds for all g bounded and measurable. This formula takes a less abstract form in the context of Itô processes [1]. Let $\mathcal{U}_N$ be the space of all square-integrable $(\mathcal{F}_t)$-predictable processes $u \colon [0,T] \times \Omega \to \mathbb{R}^{d_0}$ with $d_0 = N \cdot d_1$. Assume strong existence and uniqueness for solutions to (1). Then for all $F \in \mathbf{C}_b(\mathcal{P}(\mathcal{X}))$,

$$-\frac{1}{N} \ln \mathbf{E}\left[\exp\left(-N \cdot F(\mu^N)\right)\right] = \inf_{u \in \mathcal{U}_N} \mathbf{E}\left[\frac{1}{2N} \sum_{i=1}^{N} \int_0^T |u_i(t)|^2 dt + F(\bar{\mu}^N)\right], \tag{3}$$

where $\bar{\mu}^N = \bar{\mu}^{N,u}$ is the empirical measure of $\bar{X}^N = (\bar{X}^{1,N}, \ldots, \bar{X}^{N,N}) = \bar{X}^{N,u}$, the solution to the system of controlled SDEs

$$\begin{aligned} d\bar{X}^{i,N}(t) = b\big(\bar{X}^{i,N}(t), \bar{\mu}^N(t)\big)dt + \sigma\big(\bar{X}^{i,N}(t), \bar{\mu}^N(t)\big)u_i(t)dt \\ + \sigma\big(\bar{X}^{i,N}(t), \bar{\mu}^N(t)\big)dW^i(t). \end{aligned} \tag{4}$$

Thus we have a stochastic optimal control problem for each $N \in \mathbb{N}$. Establishing a Laplace principle now corresponds to showing convergence of control problems (essentially Γ-convergence of associated cost functionals). This can be done using weak convergence methods.

The law of large numbers suggests that weak limit points of $(\bar{\mu}^N)$ should correspond to the laws of solutions to the *controlled* limit SDE

$$\begin{aligned} d\bar{X}(t) = b\big(\bar{X}(t), \mathrm{Law}(\bar{X}(t))\big)dt + \sigma\big(\bar{X}(t), \mathrm{Law}(\bar{X}(t))\big)u(t)dt \\ + \sigma\big(\bar{X}(t), \mathrm{Law}(\bar{X}(t))\big)dW(t), \end{aligned} \tag{5}$$

where W is a d_1-dimensional standard Wiener process on some stochastic basis, u some square-integrable predictable $\mathbb{R}^{d_1}$-valued control process.

Only weak solutions of Eq. (5) are needed. Weak solutions correspond to probability measures on a canonical space. The canonical space here has three components, one for the solution process, one for the control process and one for the Wiener process. A technical difficulty arises with the space of control processes, which should be Polish. Define the canonical space as $\mathcal{Z} \doteq \mathcal{X} \times \mathcal{R}_1 \times \mathcal{W}$, where $\mathcal{X}$, $\mathcal{W}$ are path spaces and $\mathcal{R}_1$ the space of deterministic *relaxed controls* on $\mathbb{R}^{d_1} \times [0,T]$ with finite first moments. With the maximum norm topology on $\mathcal{X}$, $\mathcal{W}$, the topology of weak convergence plus convergence of first moments on $\mathcal{R}_1$, all three spaces are Polish.

Let $(\bar{X}, \rho, W)$ be the coordinate process on $\mathcal{Z}$. Let $\Theta \in \mathcal{P}(\mathcal{Z})$. Define

$$\nu_\Theta(t) \doteq \Theta(\{(\phi, r, w) \in \mathcal{Z} : \phi(t) \in B\}),\ B \in \mathcal{B}(\mathbb{R}^d),\ t \in [0,T].$$

Then Θ corresponds to a *weak solution* of Eq. (5) if and only if W is a standard Wiener process under Θ and, Θ-almost surely,

$$\bar{X}(t) = \bar{X}(0) + \int_0^t b\big(\bar{X}(s), \nu_\Theta(s)\big)ds + \int_{\mathbb{R}^{d_1}\times[0,t]} \sigma\big(\bar{X}(s), \nu_\Theta(s)\big)y\,\rho(dy\times ds) + \int_0^t \sigma\big(\bar{X}(s), \nu_\Theta(s)\big)dW(s).$$

Assume that for some $\nu_0 \in \mathcal{P}(\mathbb{R}^d)$, $\frac{1}{N}\sum_{i=1}^N \delta_{x^{i,N}} \to \nu_0$ as $N \to \infty$. Assume continuity of the coefficients b, σ, strong existence and uniqueness for the prelimit systems, uniqueness for the limit system, as well as a mild stability condition.

Let $\mathcal{P}_\infty$ be the set of all $\Theta \in \mathcal{P}(\mathcal{Z})$ such that

(a) $\int_{\mathcal{R}_1}\int_{\mathbb{R}^{d_1}\times[0,T]}|y|^2\,r(dy\times dt)\Theta_{\mathcal{R}}(dr) < \infty$,

(b) $\Theta_{\mathcal{X}}(\{\phi \in \mathcal{X} : \phi(0) \in B\}) = \nu_0(B)$, $B \in \mathcal{B}(\mathbb{R}^d)$,

(c) Θ corresponds to a weak solution of Eq. (5).

Theorem 1 *The family of empirical measures $(\mu^N)_{N\in\mathbb{N}}$ satisfies a Laplace principle with rate function*

$$I(\theta) = \inf_{\Theta\in\mathcal{P}_\infty:\Theta_{\mathcal{X}}=\theta} \frac{1}{2}\int_{\mathcal{R}}\int_{\mathbb{R}^{d_1}\times[0,T]}|y|^2\,r(dy\times dt)\Theta_{\mathcal{R}}(dr),\ \theta \in \mathcal{P}(\mathcal{X}),$$

with the convention that $\inf\emptyset = \infty$.

Ignoring the question of relaxed controls, the rate function I of Theorem 1 can be written in terms of processes; for $\theta \in \mathcal{P}(\mathcal{X})$ such that $\theta(0) = \nu_0$,

$$I(\theta) = \inf_{u\in\mathcal{U}:\mathrm{Law}(\bar{X}^u)=\theta} \mathbf{E}\left[\frac{1}{2}\int_0^T |u(t)|^2 dt\right],$$

where $\bar{X}^u$ is a solution to Eq. (5) under some control process u such that $\mathrm{Law}(\bar{X}^u) = \theta$. Thus $\bar{X}^u = \bar{X}^{u,\theta}$ solves

$$d\bar{X}^u(t) = b\big(\bar{X}^u(t), \theta(t)\big)dt + \sigma\big(\bar{X}^u(t), \theta(t)\big)u(t)dt + \sigma\big(\bar{X}^u(t), \theta(t)\big)dW(t).$$

Sanov's theorem and work on mean field models (e.g. [3]) suggest the following connection with relative entropy:

$$I(\theta) = R(\theta\|\,\mathrm{Law}(X^\theta)),$$

where $X^\theta = \bar{X}^{0,\theta}$ is the solution to

$$dX^\theta(t) = b\big(X^\theta(t), \theta(t)\big)dt + \sigma\big(X^\theta(t), \theta(t)\big)dW(t).$$

Proof of Theorem 1. The proof consists of two steps, establishing a lower bound and an optimality upper bound, in analogy with Γ-convergence. As to the lower bound, show that for any sequence $(u^N)_{N\in\mathbb{N}}$ with $u^N \in \mathcal{U}_N$,

$$\begin{aligned}\liminf_{N\to\infty} &\left\{ \frac{1}{2}\,\mathbf{E}\left[\frac{1}{N}\sum_{i=1}^{N}\int_0^T |u_i^N(t)|^2 dt\right] + \mathbf{E}\left[F(\bar{\mu}^N)\right]\right\} \\ &\geq \inf_{\Theta\in\mathcal{P}_\infty}\left\{\frac{1}{2}\int_{\mathcal{R}}\int_{\mathbb{R}^{d_1}\times[0,T]} |y|^2\, r(dy\times dt)\Theta_{\mathcal{R}}(dr) + F(\Theta_{\mathcal{X}})\right\}.\end{aligned}$$

For the upper bound (the optimality step), show that for any $\Theta \in \mathcal{P}_\infty$ there is a sequence $(u^N)_{N\in\mathbb{N}}$ with $u^N \in \mathcal{U}_N$ such that

$$\begin{aligned}\limsup_{N\to\infty} &\left\{ \frac{1}{2}\,\mathbf{E}\left[\frac{1}{N}\sum_{i=1}^{N}\int_0^T |u_i^N(t)|^2 dt\right] + \mathbf{E}\left[F(\bar{\mu}^N)\right]\right\} \\ &\leq \frac{1}{2}\int_{\mathcal{R}}\int_{\mathbb{R}^{d_1}\times[0,T]} |y|^2\, r(dy\times dt)\Theta_{\mathcal{R}}(dr) + F(\Theta_{\mathcal{X}}).\end{aligned}$$

Let $u^N \in \mathcal{U}_N$, $N \in \mathbb{N}$, be a sequence of control processes. A difficulty stems from the fact that the space of control processes depends on N. Define $\mathcal{P}(\mathcal{Z})$-valued random variables by

$$Q_\omega^N(B\times R\times D) \doteq \frac{1}{N}\sum_{i=1}^{N}\delta_{\bar{X}^{i,N}(\cdot,\omega)}(B)\cdot\delta_{\rho_\omega^{i,N}}(R)\cdot\delta_{W^i(\cdot,\omega)}(D),$$

$B \in \mathcal{B}(\mathcal{X})$, $R \in \mathcal{B}(\mathcal{R}_1)$, $D \in \mathcal{B}(\mathcal{W})$, $\omega \in \Omega$. The *functional occupation measures* Q^N are related to the variational representation by

$$\begin{aligned}&\frac{1}{2}\,\mathbf{E}\left[\frac{1}{N}\sum_{i=1}^{N}\int_0^T |u_i^N(t)|^2 dt\right] + \mathbf{E}\left[F(\bar{\mu}^N)\right] \\ &\quad= \int_\Omega\left[\int_{\mathcal{R}_1}\left(\frac{1}{2}\int_{\mathbb{R}^{d_1}\times[0,T]}|y|^2\, r(dy\times dt)\right) Q^N_{\omega,\mathcal{R}}(dr) + F(Q^N_{\omega,\mathcal{X}})\right]\mathrm{P}(d\omega).\end{aligned}$$

A key step is to show that any weak limit point of (Q^N) corresponds to a weak solution of Eq. (5) with probability one.

Lemma *Suppose (Q^n) is a weakly convergent subsequence of (Q^N). Let Q be a $\mathcal{P}(\mathcal{P}(\mathcal{Z}))$-valued random variable defined on some probability space $(\tilde{\Omega},\tilde{\mathcal{F}},\tilde{\mathrm{P}})$ such that $Q^n \to Q$ in distribution. Then Q_ω corresponds to a weak solution of Eq. (5) for $\tilde{\mathrm{P}}$-almost all $\omega \in \tilde{\Omega}$.*

The proof of the lemma uses a local martingale problem to show the solution property. In the proof of the lower bound, we may assume that $u^N \in \mathcal{U}_N$, $N \in \mathbb{N}$, are such that

$$\sup_{N\in\mathbb{N}}\mathbf{E}\left[\frac{1}{N}\sum_{i=1}^{N}\int_0^T |u_i^N(t)|^2 dt\right] \leq 2\|F\|.$$

Let Q^N, $N \in \mathbb{N}$, be the corresponding functional occupation measures. Thanks to the bound on the control costs, $(Q^N)_{N\in\mathbb{N}}$ is tight (precompact) as a family of $\mathcal{P}(\mathcal{Z})$-valued random variables (or as subset of $\mathcal{P}(\mathcal{P}(\mathcal{Z}))$). Thanks to the lemma, all limit points of $(Q^N)_{N\in\mathbb{N}}$ are elements of $\mathcal{P}_\infty$ (essentially weak solutions of Eq. (5) with probability one. The lower bound is now a consequence of (a version of) Fatou's lemma, the continuity of F and weak convergence.

To show the upper bound, let $\Theta \in \mathcal{P}_\infty$. Then the coordinate process $(\bar{X}, \rho, W)$ on the canonical space solves Eq. (5) under Θ. Find a family of control processes for the prelimit systems such that the corresponding occupation measures converge to Θ. To this end, take a sequence $(\rho^{i,\infty}, W^{i,\infty})$, $i \in \mathbb{N}$, of i.i.d. copies of (ρ, W). For each N, solve the system of prelimit equations under $(\rho^{i,\infty}, W^{i,\infty})$, $i \in \{1, \dots, N\}$. Define the corresponding functional occupation measures $\tilde{Q}^N$. Apply the same argument as before: $(\tilde{Q}^N)$ is tight; any limit random variable $\tilde{Q}$ on $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathrm{P}})$ takes values in $\mathcal{P}_\infty$ and, by construction and Varadarajan's theorem, for $\tilde{\mathrm{P}}$-almost all $\omega \in \tilde{\Omega}$,

$$\tilde{Q}_{\omega|\mathcal{B}(\mathcal{R}_1\times\mathcal{W})} = \Theta \circ (\rho, W)^{-1}.$$

By weak uniqueness, it follows that $\tilde{Q}_\omega = \Theta \circ (\bar{X}, \rho, W)^{-1} = \Theta$ for $\tilde{\mathrm{P}}$-almost all $\omega \in \tilde{\Omega}$. □

Theorem 1 and its proof can be easily extended to more general weakly interacting processes, in particular to delay systems. The N-particle prelimit model is given by a system of N stochastic delay (or functional) differential equations

$$dX^{i,N}(t) = b\big(t, X^{i,N}, \mu^N(t)\big)dt + \sigma\big(t, X^{i,N}, \mu^N(t)\big)dW^i(t), \tag{6}$$

where b, σ are progressive functionals on $[0,T] \times \mathcal{X} \times \mathcal{P}(\mathbb{R}^d)$. The corresponding uncontrolled limit equation is

$$dX(t) = b\big(t, X, \mathrm{Law}(X(t))\big)dt + \sigma\big(t, X, \mathrm{Law}(X(t))\big)dW(t). \tag{7}$$

Let $\Theta \in \mathcal{P}(\mathcal{Z})$. Then Θ corresponds to a *weak solution* of the controlled analogue of Eq. (7) if and only if W is a standard Wiener process under Θ and, Θ-almost surely,

$$\bar{X}(t) = \bar{X}(0) + \int_0^t b\big(s, \bar{X}, \nu_\Theta(s)\big)ds + \int_{\mathbb{R}^{d_1}\times[0,t]} \sigma\big(s, \bar{X}, \nu_\Theta(s)\big)y\,\rho(dy\times ds) + \int_0^t \sigma\big(s, \bar{X}, \nu_\Theta(s)\big)dW(s).$$

The Laplace principle for the empirical measures (μ^N) is now completely analogous to that of Theorem 1.

References

[1] M. Boué and P. Dupuis, *A variational representation for certain functionals of Brownian motion*. Ann. Probab. 26/4 (1998), 1641–1659.

[2] A. Budhiraja, P. Dupuis, and M. Fischer, *Large deviation properties of weakly interacting processes via weak convergence methods*. Ann. Probab., to appear.

[3] P. Dai Pra and F. den Hollander, *McKean-Vlasov limit for interacting random processes in random media*. J. Stat. Phys. 84/3-4 (1996), 735–772.

[4] D. Dawson and J. Gärtner, *Large deviations from the McKean-Vlasov limit for weakly interacting diffusions*. Stochastics 20/4 (1987), 247–308.

[5] A. Dembo and O. Zeitouni, "Large Deviations Techniques and Applications". Springer, Berlin, 2nd edition, 1998.

[6] P. Dupuis and R. S. Ellis, "A Weak Convergence Approach to the Theory of Large Deviations". John Wiley & Sons, New York, 1997.

[7] R. S. Ellis, "Entropy, Large Deviations, and Statistical Mechanics". Springer, Berlin, 1985.

[8] F. den Hollander, "Large deviations". Volume 14 of *Fields Institute Monographs*. American Mathematical Society, Providence, RI, 2000.

[9] K. Oelschläger, *A martingale approach to the law of large numbers for weakly interacting stochastic processes*. Ann. Probab. 12/2 (1984), 458–479.

[10] S. R. S. Varadhan, *Large deviations. Special invited paper*. Ann. Probab. 36/2 (2008), 397–419.

A Viscosity approach to Monge-Ampère type PDEs

Marco Cirant (*)

1 Monge-Ampère type equations

A Monge-Ampère equation is a second order partial differential equation of the form

$$\det D^2 u(x) = g(x, u, Du),$$

where $g : X \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$, $X \subseteq \mathbb{R}^n$ (we denote with Du the gradient of u and with D^2u its hessian matrix). It belongs to the wide class of *fully non-linear equations*, because it is non-linear with respect to second order derivatives u_{ij}: $\det D^2u$ is a n-degree polynomial of u_{ij}.

Monge-Ampère equations arise in many fields of mathematics; for example, transportation problems (or Monge-Kantorovich problems, concerning optimal transportation and allocation of resources) can be reduced to the resolution of equations of that type.

Differential geometry is also an interesting source of non-linear equations, in particular within the study of curvatures of surfaces. An extensive study has been carried out for many years on curvatures, tools that measure how a surface bends; we will define the notion of *principal curvature* and how Monge-Ampère type equations spring from related problems.

Suppose we are given a continuous function $u : \Omega \to \mathbb{R}$, where Ω is a bounded domain of $\mathbb{R}^n$ (throughout these notes we will use this convention, with X denoting any subset of $\mathbb{R}^n$ instead). The graph of the function u defines a surface in $\mathbb{R}^{n+1}$, given by the points $\{(x, u(x)) \in \mathbb{R}^{n+1} : x \in \Omega\}$.

Suppose now u is twice differentiable at some point $x \in \Omega$. In the differential geometry jargon, principal curvatures at point x are the eigenvalues of the shape operator (an operator that computes in x the degree of bending at different directions): they are n real numbers that describe quantitatively the shape of the surface at that point. We are interested in an analytical equivalent definition:

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: `cirant@math.unipd.it` . Seminar held on 18 May 2011.

Definition 1.1 The *principal curvatures* k_i of the surface $\{(\cdot, u(\cdot))\} \subset \mathbb{R}^{n+1}$ at some point $x \in \Omega$ are[(1)]

$$k_i(x) = \lambda_i \left(\left(I - \frac{Du(x) \otimes Du(x)}{1 + |Du(x)|^2} \right) \frac{D^2 u(x)}{\sqrt{1 + |Du(x)|^2}} \right),$$

$i = 1, ..., n$.

Now, the so-called *Gaussian curvature* collects the informations carried by principal curvatures:

$$K = \prod_{i=1}^{n} k_i.$$

The gaussian curvature is an important quantity associated to the surface (it is actually a function $K : \Omega \to \mathbb{R}$) because of its *intrinsic* nature. It can be computed explicitly:

$$K(x) = \det \left(\left(I - \frac{Du \otimes Du}{1 + |Du|^2} \right) \frac{D^2 u}{\sqrt{1 + |Du|^2}} \right) = \frac{\det D^2 u(x)}{(1 + |Du(x)|^2)^{\frac{n+2}{2}}}.$$

Now, *given a map $K : \Omega \to \mathbb{R}$, is it possible to find a surface defined by the graph of a function u whose gaussian curvature is $K(x)$ at every $x \in \Omega$?* If u satisfies

$$\det D^2 u(x) = K(x)(1 + |Du(x)|^2)^{\frac{n+2}{2}}, \tag{1}$$

it will be a solution to the problem, an answer to our question. (1) is called the *prescribed gaussian curvature equation* and, as we see, it is of Monge-Ampère type.

In these notes we will present a result of existence and uniqueness of solutions for a simpler class of equations:

$$\det D^2 u(x) = f(x), \tag{2}$$

with no explicit dependance of the right hand side upon Du and $f \geq 0$. We will implement modern viscosity techniques in order to solve the associated Dirichlet problem: we will find a solution u to (2), which satisfy also a boundary condition $u|_{\partial\Omega} = \varphi$, where φ is a given datum. We will generalize the machinery presented in [1] by Harvey and Lawson; they produce a clever and elegant reformulation of the "classical" viscosity theory for fully non-linear equations (the interested reader may check the famous User's Guide [2] on this subject).

Measure theory has been widely used to study (2) and the book [3] contains many informations on that. Another interesting reference is the inspiring paper [4]; completely different topological methods are implemented, based upon a-priori estimates on the hessian of solutions.

[(1)] $\lambda_i(A)$ is the i-th eigenvalue of the symmetric matrix A.

2 The Dirichlet problem

We are not going to study directly the Dirichlet problem for the equation $F(x, D^2u) = \det D^2u - f(x) = 0$, but we will focus on its *elliptic branches*. This kind of "geometric" approach is based upon ideas of Krylov ([5]); the simple triggering observation is that if at some point x

$$D^2u(x) \in \{F(x, \cdot) = 0\},$$

then at that point $F(x, D^2u(x)) = 0$.

Let now $f : X \to [0, +\infty)$ be a given function, pick the family of sets of symmetric matrices[(2)]

$$\Theta(x) = \{A \in \mathcal{P} : \det A \geq f(x)\} \quad \forall x \in X. \tag{3}$$

We see that

$$D^2u(x) \in \partial\Theta(x) \Rightarrow \det D^2u = f(x),$$

so (smooth) functions u satisfying the expression

$$D^2u(x) \in \partial\Theta(x)$$

at some point x will be solutions of $F = 0$. Our aim will be to study this new equation, that is strictly linked to the original Monge-Ampère equation, and owns a particular so-called elliptic structure, because $\Theta(x) + \mathcal{P} \subset \Theta(x)$. To be more precise (denoting with $\wp(\mathcal{A})$ the powerset of $\mathcal{A}$),

Definition 2.1 Let $\Theta : X \to \wp(\mathrm{Sym}^2(\mathbb{R}^n))$. We say that

$$D^2u(x) \in \partial\Theta(x) \quad \forall x \in X$$

is an *elliptic branch* defined by Θ on X and associated to the equation

$$F(x, D^2u) = 0$$

if

(i) consistency: $\partial\Theta(x) \subset \{A : F(x, A) = 0\} \quad \forall x \in X$

(ii) positivity: $\Theta(x) + \mathcal{P} \subset \Theta(x) \quad \forall x \in X$

As we saw, the Monge-Ampère equation has (at least) one elliptic branch; from now on we will discuss on elliptic branches in general, without referring to the equations they are associated to. We will come back eventually to Monge-Ampère and to other particular equations we are able to solve through elliptic branches.

If u is at least twice differentiable, D^2u is well-defined and the expression $D^2u(x) \in \partial\Theta(x)$ has a precise meaning, so it is clear when u is a *solution* of the elliptic branch. When

[(2)] We will denote with $\mathrm{Sym}^2(\mathbb{R}^n)$ is the space of $n \times n$ real symmetric matrices and with $\mathcal{P}$ its subspace of non-negative matrices.

the problem we are dealing with is nonlinear, it is convenient to have at our disposal a notion of solution in a weak sense, when u is no more twice differentiable[(3)]. Harvey and Lawson, in the spirit of viscosity theory, formulated such a notion suiting very well the setting of elliptic branches. Let's start with the definition of subsolution in the classical sense (when second order derivatives exist).

Definition 2.2 Let $u \in \mathbb{C}^2(\Omega)$. u will be said a subsolution of $D^2u(x) \in \partial\Theta(x)$ if

$$D^2u(x) \in \Theta(x) \quad \forall x \in \Omega$$

The key step is to define a *dual branch* whose subsolutions work as test functions for subsolutions in weak sense.

Definition 2.3 Let $D^2u \in \partial\Theta(x)$ be an elliptic branch defined by Θ on X. Its *dual elliptic branch* is defined as $D^2u(x) \in \partial\widetilde{\Theta}(x)$, where

$$\widetilde{\Theta}(x) = -(\operatorname{Int}\Theta(x))^c \quad \forall x \in X$$

Subaffine functions play a major role in our weak setting; they are functions that satisfy a *maximum principle* with respect to affine functions, and it is easy to show that they satisfy also the standard maximum principle ($\sup_K u \leq \sup_{\partial K}$ for $K \subset\subset \Omega$).

Definition 2.4 A function $u \in \mathrm{USC}(X)$ will be said *subaffine in* $x \in X$ if there exists a neighborhood Y of x such that for every compact $K \subset Y$ and every affine function a,

$$u \leq a \text{ on } \partial K \;\Rightarrow u \leq a \text{ on } K.$$

The dual family of sets $\widetilde{\Theta}(x)$ defining the dual branch enables a simple characterization of subsolutions of $D^2u \in \partial\Theta(x)$:

$$D^2u(x) \in \Theta(x) \Leftrightarrow D^2u(x) + B \text{ has at least one non-negative eig. } \forall B \in \widetilde{\Theta}(x),$$

and it is easily proved that regular functions whose hessian has at least one non-negative eigenvalue are subaffine. It is motivated and well-posed the

Definition 2.5 $u \in \mathrm{USC}(\Omega)$ is a *weak subsolution*[(4)] of $D^2u \in \partial\Theta(x)$ in Ω if for all $x \in \Omega$

$$\begin{gathered} u + v \text{ is subaffine in } x \\ \text{for all } v \in \mathcal{C}^2(\Omega) \text{ such that } D^2v(x) \in \widetilde{\Theta}(x). \end{gathered}$$

$u \in \mathbb{C}(\Omega)$ is a *weak solution* of $D^2u \in \partial\Theta(x)$ in Ω if

[(3)] For example, $u(x,y) = [\max\{(x^2-1/2)^+, (y^2-1/2)^+\}]^2$ is a solution to $\det D^2u = 0$ almost everywhere in the unitary disk of $\mathbb{R}^2$, and it is not everywhere twice differentiable

[(4)] $\mathrm{USC}(\Omega)$ is the space of upper semicontinuous functions on Ω.

- u is a subsolution of $D^2u \in \partial\Theta(x)$,
- $-u$ is a subsolution of $D^2u \in \partial\widetilde{\Theta}(x)$.

Albeit this definition of weak solution is a bit technical, it turns out to be very reasonable, since we will be able to find a unique solution in this sense to the Dirichlet problem for many non-linear equations; it must be also mentioned that this is actually equivalent to the notion of viscosity solution ([2]).

As we briefly introduced the idea of weak solution, we move to discuss about sufficient conditions for the solvability of Dirichlet problems. Continuity of the map $\Theta : x \mapsto \Theta(x)$ which defines the branch is an important one:

Definition 2.6 An elliptic branch $D^2u \in \partial\Theta(x)$ will be said *continuous* in X if $\Theta : X \to \wp(\mathrm{Sym}^2(\mathbb{R}^n))$ is continuous, provided $\wp(\mathrm{Sym}^2(\mathbb{R}^n))$ with the Hausdorff distance.

It is not known how far the theory can be pushed and how much the structural hypotheses on the elliptic branch can be relaxed, but continuity proves to be a natural one, because it guarantees three important features of subsolutions (in weak sense) that we are going to use in our main result:

Proposition 2.7 *If $D^2u \in \partial\Theta(x)$ is an elliptic branch, continuous in X, the following properties of viscosity subsolutions hold:*

- *MAX: If u, v are subsolutions in X, then $\max\{u, v\}$ is a subsolution in X.*
- *SUP: The upper envelope $\sup_{u\in\mathcal{F}} u$ of a family $\mathcal{F}$ of subsolutions in X is a subsolution in X.*
- *QUASI-TRANSL: Suppose u is a subsolution in X, $\bar{u}$ be its extension to $\mathbb{R}^n$ ($\bar{u} = -\infty$ on $\mathbb{R}^n \setminus X$); then, for all $\epsilon > 0$ there exists $\eta > 0$ such that[5] $\bar{u}_y + \epsilon|x|^2$ is a subsolution in X for all $|y| < \eta$.*

Under the continuity assumption follows the existence and uniqueness of solutions to the Dirichlet problem for elliptic branches.

Theorem 2.8 *Let $\Omega \subset \mathbb{R}^n$ be a smooth, bounded, strictly convex domain, and $D^2u \in \partial\Theta(x)$ be a* continuous *elliptic branch on $\overline{\Omega}$. Then, for each $\varphi \in \mathbb{C}(\partial\Omega)$ there exists a unique $u \in \mathbb{C}(\overline{\Omega})$ which is a solution of the branch and equals φ on $\partial\Omega$.*

On the side of existence, the theorem is relies upon the Perron method, which consists in taking as a solution the upper envelope (the pointwise supremum) of a family of subsolutions; convexity of the boundary ensures that this solution equals the datum φ on $\partial\Omega$. As for uniqueness, a comparison principle is proved making use of tools from convex analysis (maximum principle of Slodkowski [6]) and sup-convolution approximation. These are the keywords of the proof; we follow the scheme of [1] in the more general situation of x-dependent branches.

[5] $f_y(x) = f(x + y)$.

As a corollary we have the solvability of the Dirichlet problem for the Monge-Ampère equation (2).

Corollary 2.9 *Let $\Omega \subset \mathbb{R}^n$ be a smooth, bounded, strictly convex domain, and $f \in \mathbb{C}(\Omega, [0, +\infty))$. Then, for each $\varphi \in \mathbb{C}(\partial\Omega)$ there exists a unique $u \in \mathbb{C}(\overline{\Omega})$ satisfying (weakly)*

$$\begin{cases} \det D^2 u(x) = f(x) & \Omega \\ u = \varphi & \partial\Omega. \end{cases}$$

Indeed, if the right hand side of the equation is continuous, the associated elliptic branch defined by (3) is continuous and Theorem 2.8 is applied.

We notice that the solution u is a-posteriori convex; a theorem of Aleksandrov asserts that a convex function is almost everywhere twice differentiable, so u is a solution to the equation in the classical sense (D^2u is well-defined in the usual way!) in Ω, at least outside a null-measure set.

3 More general equations

Theorem 2.8 is an *abstract* result, meaning that it states existence and uniqueness of weak solutions of a branch satisfying the hypotheses of ellipticity and continuity; it is not strictly related to branches associated to Monge-Ampère type equations, so we may ask whether it is applicable to other kinds of problems.

Non-totally degenerate elliptic equations, with x and D^2u separated, is a wide class that falls in our "domain of solvability":

$$F(D^2u) = f(x). \tag{4}$$

Proposition 3.1 *Suppose that*

1. $F \in \mathbb{C}(\overline{\Omega})$ *is* non-totally degenerate elliptic*:* $\exists \eta > 0$ *s.t.*

$$F(A + rI) \geq F(A) + \eta r, \quad \forall r > 0$$

2. $f \in \mathbb{C}(\overline{\Omega})$.

Then the branch

$$D^2u(x) \in \partial\Theta_{F,f}(x) \quad \forall x \in \overline{\Omega}$$

defined in $\overline{\Omega}$ by

$$\Theta_{F,f}(x) = \{A \in \mathrm{Sym}^2(\mathbb{R}^n) : F(A) \geq f(x)\}$$

is a continuous elliptic branch associated to the equation (4).

By Theorem 2.8 and Proposition 3.1 we have a general result concerning the Dirichlet problem for (4):

Theorem 3.2 *Let $\Omega \subset \mathbb{R}^n$ be a smooth, bounded, strictly convex domain. Then, for each $\varphi \in \mathbb{C}(\partial\Omega)$ there exists a unique $u \in \mathbb{C}(\overline{\Omega})$ satisfying (weakly)*

$$\begin{cases} F(D^2u(x)) = f(x) & \Omega \\ u = \varphi & \partial\Omega. \end{cases}$$

References

[1] Harvey, F. Reese and Lawson, Jr., H. Blaine, *Dirichlet duality and the nonlinear Dirichlet problem*. Communications on Pure and Applied Mathematics 62/3 (2009), 396–443.

[2] Crandall, Michael G. and Ishii, Hitoshi and Lions, Pierre-Louis, *User's guide to viscosity solutions of second order partial differential equations*. Bull. Amer. Math. Soc. (N.S.) 27/1 (1992), 1–67.

[3] Gutiérrez, Cristian E., "The Monge-Ampère equation". Progress in Nonlinear Differential Equations and their Applications, 44. Birkhäuser Boston Inc., Boston, MA, 2001.

[4] Caffarelli, L. and Nirenberg, L. and Spruck, J., *The Dirichlet problem for nonlinear second-order elliptic equations. I. Monge-Ampère equation*. Comm. Pure Appl. Math. 37/3 (1984), 369–402.

[5] Krylov, N. V., *On the general notion of fully nonlinear second-order elliptic equations*. Trans. Amer. Math. Soc. 347/3 (1995), 857–895.

[6] Slodkowski, Zbigniew, *The Bremermann-Dirichlet problem for q-plurisubharmonic functions*. Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) 11/2 (1984), 303–326.

Identification of Reciprocal Processes and related Matrix Extension Problem

Francesca Paola Carli (*)

Abstract. Stationary reciprocal processes defined on a finite interval of the integer line can be seen as a special class of Markov random fields restricted to one dimension. This kind of processes are potentially useful for describing signals which naturally live in a finite region of the time (or space) line. Non–stationary reciprocal processes have been extensively studied in the past especially by Jamison, Krener, Levy and co–workers. The specialization of the non–stationary theory to the stationary case, however, does not seem to have been pursued in sufficient depth in the literature. Moreover, estimation and identification of reciprocal stochastic models starting from observed data seems still to be an open problem. This note addresses these problems showing that maximum likelihood identification of stationary reciprocal processes on the discrete circle leads to a covariance extension problem for block–circulant covariance matrices. Covariane extension problems have gained considerable attention in the past (think for example to the covariance extension problem for stationary processes on the integer line, i.e. for Toeplitz matrices and to general matrix extension problems introduced by A. P. Dempster). Neverthless, the band extension problem for block–circulant matrices does not seem to have been addressed before. We show that the maximum entropy principle leads to a complete solution of the problem. An efficient algorithm for the computation of the maximum likelihood estimates is also provided. This note sketches the results in [3], [4] and [2].

1 Notation and Preliminaries

Throughout this note, we work in the wide-sense setting of zero-mean random variables which have finite second moment. Random variables which have finite second moment are commonly called second order random variables. The set of real or complex-valued second-order random variables defined on the same probability space, say $\mathbf{H}$, is obviously a linear vector space under the usual operations of sum and multiplication by real (or

(*)Department of Information Engineering (DEI), University of Padova, via Gradenigo 6/B, 35131 Padova, Italy – E-mail: carlifra@dei.unipd.it . Seminar held on 8 June 2011. This seminar is based on joint works with A. Ferrante, T. T. Georgiou, M. Pavon, and G. Picci.

complex) numbers. This vector space comes naturally equipped with an inner product

$$\langle \xi, \eta \rangle = \mathbb{E}\, \xi\eta \tag{1}$$

where $\mathbb{E}\,[\,\cdot\,]$ denotes the mathematical expectation (i.e. the inner product is just the correlation of the two random variables). It is well–known that $\mathbf{H}$ is complete with respect to the norm associated with the inner product (1) and is therefore an *Hilbert space*. The correspondence between probabilistic concepts depending only on second–order moments and geometric operations on certain subspaces of the Hilbert space of finite variance random variables was established by Kolmogorov in the early 1940's and will be assumed henceforth.

Following this correspondence, we say that two random vectors $\mathbf{x} = [x_1, \ldots, x_n]$ and $\mathbf{y} = [y_1, \ldots, y_n]$ are orthogonal, which we shall write $\mathbf{x} \perp \mathbf{y}$, if they are componentwise uncorrelated, i.e. if $\langle x_i, y_i \rangle = \mathbb{E}\, x_i y_i = 0$ for all $i = 1, \ldots, n$. The symbol $\hat{\mathbb{E}}\,[\,\cdot \mid \cdot\,]$ will denote orthogonal projection (conditional expectation in the Gaussian case) onto the subspace spanned by a family of finite variance random variables listed in the second argument.

The concept of conditional orthogonality plays a fundamental role in the definition of reciprocal process.

Definition 1.1 Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ be subspaces of zero–mean second–order random variables in a certain common ambient Hilbert space $\mathbf{H}$. $\mathbf{X}$ and $\mathbf{Y}$ are said to be *conditionally orthogonal given* $\mathbf{Z}$, which we shall write as

$$\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$$

if

$$\Big(\mathbf{x} - \hat{\mathbb{E}}\,[\,\mathbf{x} \mid \mathbf{Z}\,]\Big) \perp \Big(\mathbf{y} - \hat{\mathbb{E}}\,[\,\mathbf{y} \mid \mathbf{Z}\,]\Big)\,, \quad \forall\, \mathbf{x} \in \mathbf{X}, \forall\, \mathbf{y} \in \mathbf{Y}\,, \tag{2}$$

i.e., conditional orthogonality is orthogonality after subtracting the projections on $\mathbf{Z}$.

Conditional orthogonality is the same as conditional uncorrelatedness (and hence conditional independence) in the Gaussian case. The intuitive meaning of conditional orthogonality is captured by the following Lemma (see, e.g., [11]).

Lemma 1.1 $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$ *if and only if one of the following equivalent conditions holds*

(i) $\hat{\mathbb{E}}\,[x \mid \mathbf{Y} \vee \mathbf{Z}] = \hat{\mathbb{E}}\,(x \mid \mathbf{Z})$, $x \in \mathbf{X}$

(ii) $\hat{\mathbb{E}}\,[y \mid \mathbf{X} \vee \mathbf{Z}] = \hat{\mathbb{E}}\,(y \mid \mathbf{Z})$, $y \in \mathbf{Y}$

where $\mathbf{X} \vee \mathbf{Z}$ *(*$\mathbf{Y} \vee \mathbf{Z}$*) denotes the smallest closed vector space containing* $\mathbf{X}$ *(*$\mathbf{Y}$*) and* $\mathbf{Z}$*.*

When $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ are generated by finite dimensional random vectors, condition (2) can equivalently be rewritten in terms of the generating vectors, which we shall normally do in the following.

2 Reciprocal Processes on the Discrete Circle

In this section reciprocal processes on the discrete circle are introduced.

Let n be a natural number such that $N > 2n$. This inequality will be assumed to hold throughout. We introduce the notation $\mathbf{y}_{[t-n,\,t)}$ for the nm-dimensional random vector obtained by stacking $\mathbf{y}(t-\ n), \ldots, \mathbf{y}(t-1)$ in that order. Similarly, $\mathbf{y}_{(t,t+n]}$ is the vector obtained by stacking $\mathbf{y}(t+1), \ldots, \mathbf{y}(t+n)$ in that order. Likewise, the vector $\mathbf{y}_{[t-n,\,t]}$ is obtained by appending $\mathbf{y}(t)$ as last block to $\mathbf{y}_{[t-n,\,t)}$, etc.. The sums $t-k$ and $t+k$ are to be understood modulo N. Consider a subinterval $(t_1,\, t_2\,) \subset [1,\, N]$ where $(t_1,\, t_2\,) := \{t \mid t_1 < t < t_2\}$ and $(t_1,\, t_2)^c$ denotes the complementary set in $[1,\, N]$. The following definition does not require stationarity.

Definition 2.1 (Reciprocal process of order n) A process $\{\mathbf{y}(t)\}$ on $\mathbb{Z}_N$ is *reciprocal of order n* if, for any interval $(t_1,\, t_2\,) \subseteq \mathbb{Z}_N$

$$\mathbf{y}_{(t_1,\,t_2\,)} \perp \mathbf{y}_{(t_1-n,\,t_2+n)^c} \Big| \ \big\{\mathbf{y}_{(t_1-n,\,t_1\,]} \vee \mathbf{y}_{[t_2,\,t_2+n\,)}\big\}.$$

Equivalently (see Lemma 1.1), it must hold that

$$\hat{\mathbb{E}}\,[\,\mathbf{y}_{(t_1,\,t_2)} \mid \mathbf{y}(s),\, s \in (t_1,\, t_2)^c\,] = \hat{\mathbb{E}}\,[\,\mathbf{y}_{(t_1,\,t_2)} \mid \mathbf{y}_{(t_1-n,\,t_1\,]} \vee \mathbf{y}_{[t_2,\,t_2+n\,)}\,]\,, \tag{3}$$

for $t_1,\, t_2 \in \mathbb{Z}_N$.

This definition slightly generalized the definition in the literature (see [1, 12, 9 ,10]). In fact, it is given in terms of conditionally orthogonality (instead of conditionally independence). This allows us to deal with not necessarily Gaussian processes, the definition in the literature following as a particularization since, for Gaussian processes, conditional orthogonality is the same as conditional independence. Moreover, in the spirit of the "higher order models" introduced by Frezza (see [6]), we consider general reciprocal processes *of order n*, standard reciprocal processes in the literature following as a particularization for $n = 1$.

Let $\mathbf{y}$ be a *stationary* reciprocal processes on the discrete circle with positive definite covariance matrix (i.e. $\mathbf{y}$ is a nonsingular process). The following representation result holds [3].

Theorem 2.1 (Modeling of stationary reciprocal processes of order n) *A nonsingular stationary process $\mathbf{y}$ taking values in $\mathbb{R}^m$ is reciprocal of order n on $\mathbb{Z}_N$ if and only if it satisfies a linear, constant-coefficients difference equation of the type*

$$\sum_{k=-n}^{n} M_k\, \mathbf{y}(t-k) \ = \ \mathbf{e}(t)\,, \quad t \in \mathbb{Z}_N \tag{4}$$

where the M_k's are $m \times m$ matrices, $M_{-k} = M_k^\top$, $k = 1, \ldots, n$, and $\mathbf{e}(t)$, besides satisfying the orthogonality property

$$\mathbb{E}\,\mathbf{y}\,\mathbf{e}^\top \ = \mathbf{I}_N\,,$$

is a stationary locally correlated process with covariance matrix

$$\mathrm{Var}\{\mathbf{e}\} := \mathbf{M}_N = \begin{bmatrix} M_0 & M_1^\top & \dots & M_n^\top & 0 & \dots & 0 & M_n & \dots & M_1 \\ M_1 & M_0 & M_1^\top & \ddots & M_n^\top & 0 & & 0 & \ddots & \vdots \\ \vdots & & \ddots & & & \ddots & \ddots & & \ddots & M_n \\ M_n & \dots & M_1 & M_0 & M_1^\top & \dots & M_n^\top & \ddots & & 0 \\ 0 & M_n & & \dots & M_0 & \dots & & M_n^\top & \ddots & \vdots \\ \vdots & \ddots & \ddots & & & \ddots & & & \ddots & 0 \\ 0 & & & & & & & & & M_n^\top \\ M_n^\top & \ddots & & & & & & & & \vdots \\ \vdots & \ddots & \ddots & & \ddots & \ddots & & & \ddots & M_1^\top \\ M_1^\top & \dots & M_n^\top & 0 & \dots & 0 & M_n & \dots & M_1 & M_0 \end{bmatrix}$$

i.e. $\mathbf{M}_N$ *is a symmetric block-circulant matrix* banded of bandwidth n *with the model parameters* $\{M_k\}$ *as block–entries.*

An important characterization of stationary reciprocal processes on $\mathbb{Z}_N$ is the following.

Theorem 2.2 (Characterization of Reciprocal Processes) *The* $mN \times mN$ *nonsingular matrix* $\mathbf{\Sigma}_N$ *is the covariance matrix of a* m*–dimensional reciprocal process of order* n *on* $\mathbb{Z}_N$ *if and only if* $\mathbf{\Sigma}_N^{-1}$ *is a positive-definite symmetric block–circulant matrix* banded of bandwidth n.

3 Maximum likelihood Identification of Reciprocal Processes

Assume that T independent realizations of one period of the process $\mathbf{y}$ are available and let us denote by

$$\underline{y} := \begin{bmatrix} y^{(1)} & \dots & y^{(T)} \end{bmatrix}$$

the collection of these T realizations. The problem we are interested in solving is the following.

Problem 3.1 (Identification Problem) *Given the observations* $\underline{y}$ *of a reciprocal process* $\mathbf{y}$ *of (known) order* n*, estimate the parameters* $\{M_k\}$ *of the underlying reciprocal model (4).*

To solve this problem we set up the Gaussian log-likelihood (this does not require to assume that $\mathbf{y}$ has a Gaussian distribution, see [8, p. 112])

$$L(M_0, \dots, M_n) = -\frac{T}{2} \log\det \mathbf{M}_N^{-1} - \frac{1}{2}\mathrm{tr}(\mathbf{M}_N \underline{y}\,\underline{y}^\top)$$

where $\bar{\Sigma}_N$ is the sample covariance $\bar{\Sigma}_N = \frac{1}{T}\underline{y}\,\underline{y}^\top$.

It can be shown [3] that Problem 3.1 is equivalent to the following matrix completion problem, which, from now on, will be referred to as the *block-circulant band extension problem.*

Problem 3.2 (Block-Circulant Band Extension Problem) *Given $n+1$ initial data $m \times m$ matrices $\Sigma_0, \ldots, \Sigma_n$, arranged in a way consistent with a symmetric block circulant structure, i.e. given the partially specified block–circulant matrix*

$$\begin{bmatrix}
\Sigma_0 & \Sigma_1^\top & \ldots & \Sigma_n^\top & ? & \ldots & ? & \Sigma_n & \ldots & \Sigma_1 \\
\Sigma_1 & \Sigma_0 & \Sigma_1^\top & \ddots & \Sigma_n^\top & ? & & ? & \ddots & \vdots \\
\vdots & & \ddots & & & \ddots & \ddots & & \ddots & \Sigma_n \\
\Sigma_n & \ldots & \Sigma_1 & \Sigma_0 & \Sigma_1^\top & \ldots & \Sigma_n^\top & \ddots & & ? \\
? & \Sigma_n & & \ldots & \Sigma_0 & \ldots & & \Sigma_n^\top & \ddots & \vdots \\
\vdots & \ddots & \ddots & & & \ddots & & & \ddots & ? \\
? & & & & & & & & & \Sigma_n^\top \\
\Sigma_n^\top & \ddots & & & & & & & & \vdots \\
\vdots & \ddots & \ddots & & \ddots & \ddots & & & \ddots & \Sigma_1^\top \\
\Sigma_1^\top & \ldots & \Sigma_n^\top & ? & \ldots & ? & \Sigma_n & \ldots & \Sigma_1 & \Sigma_0
\end{bmatrix}$$

complete it in such a way to form a positive definite symmetric block-circulant matrix $\mathbf{\Sigma}_N$ with a (block-circulant) banded inverse of bandwidth n.

This problem recalls the *covariance selection* problem introduced by A. P. Dempster [5] and studied by many authors (see e.g. [7], [13] and references therein). It reads as follows.

Problem 3.3 (Covariance Selection Problem - Dempster) *Given a partially specified symmetric matrix, find a completion $\mathbf{\Sigma}_N$ which agrees with the partially specified one in the given positions, is symmetric positive definite and such that its inverse has zeros in the complementary positions of those assigned.*

At a first sight our problem seems to be a particular instance of the Dempster problem where the given entries lie on the main diagonals and on the NE and SW corners. Notice, however, that the linear constraint that forces the completed matrix to be circulant is *not* present in the Dempster's setting. Neverthless, a key observation in the Dempster's work is the following.

Proposition 3.1 (Dempster) *Assume that Problem 3.3 is feasibile. Among all the positive definite extensions, there exists a unique one whose inverse's entries are zero in all the positions complementary to those where the elements of the covariance are assigned. This extension corresponds to the Gaussian distribution with maximum entropy.*

Inspired by this maximum entropy principle, we switch to consider the following problem. Let $\mathbf{U}_N$ denote the block-circulant "shift" matrix with $N \times N$ blocks,

$$\mathbf{U}_N = \begin{bmatrix}
0 & I_m & 0 & \ldots & 0 \\
0 & 0 & I_m & \ldots & 0 \\
\vdots & \vdots & & \ddots & \vdots \\
0 & 0 & 0 & \ldots & I_m \\
I_m & 0 & 0 & \ldots & 0
\end{bmatrix}.$$

Clearly, $\mathbf{U}_N^\top \mathbf{U}_N = \mathbf{U}_N \mathbf{U}_N^\top = I_{mN}$, i.e. $\mathbf{U}_N$ is orthogonal. Recall that a matrix $\mathbf{C}_N$ with $N \times N$ blocks is block-circulant if and only if it commutes with $\mathbf{U}_N$, namely if and only if it satisfies

$$\mathbf{U}_N^\top \mathbf{C}_N \mathbf{U}_N = \mathbf{C}_N. \tag{5}$$

Moreover, let $\mathfrak{S}_N$ denote the vector space of *symmetric* matrices with $N \times N$ square blocks of dimension $m \times m$ and $\mathbf{T}_n \in \mathfrak{S}_{n+1}$ the Toeplitz matrix of *boundary data*

$$\mathbf{T}_n = \begin{bmatrix} \Sigma_0 & \Sigma_1^\top & \dots & \dots & \Sigma_n^\top \\ \Sigma_1 & \Sigma_0 & \Sigma_1^\top & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \Sigma_1^\top \\ \Sigma_n & \dots & \dots & \Sigma_1 & \Sigma_0 \end{bmatrix}, \tag{6}$$

while E_n denotes the $N \times (n+1)$ block matrix

$$E_n = \begin{bmatrix} I_m & 0 & \dots & 0 \\ 0 & I_m & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & I_m \\ 0 & \dots & & 0 \end{bmatrix}.$$

Recall that the *differential entropy* $H(p)$ of a probability distribution with density p on $\mathbb{R}^n$ is defined by

$$H(p) = -\int_{\mathbb{R}^n} \log(p(x)) p(x) dx. \tag{7}$$

In the case of a zero-mean Gaussian distribution p with covariance matrix $\mathbf{\Sigma}_N$, we get

$$H(p) = \frac{1}{2} \log(\det \mathbf{\Sigma}_N) + \frac{1}{2} n \left(1 + \log(2\pi)\right). \tag{8}$$

The problem we are interested in is the following.

Problem 3.4 (Maximum entropy band extension problem for block-circulant matrices (CMaxEnt))

$$\max \{\log\det \mathbf{\Sigma}_N \mid \mathbf{\Sigma}_N \in \mathfrak{S}_N,\ \mathbf{\Sigma}_N > 0\} \tag{9.a}$$

subject to :

$$E_n^\top \mathbf{\Sigma}_N E_n = \mathbf{T}_n, \tag{9.b}$$

$$\mathbf{U}_N^\top \mathbf{\Sigma}_N \mathbf{U}_N = \mathbf{\Sigma}_N. \tag{9.c}$$

Problem 3.4 amounts to finding the maximum entropy Gaussian distribution with a block-circulant covariance whose first $n+1$ blocks are precisely $\Sigma_0, \ldots, \Sigma_n$. Notice that in this problem we are minimizing a strictly convex function on the intersection of a convex cone (minus the zero matrix) with a linear manifold. Hence, we are dealing with a *convex optimization problem*. Moreover, we are *not imposing* that the inverse of the solution $\mathbf{\Sigma}_N$ of Problem 3.4 should have a banded structure. We shall see that, whenever solutions exist, this property will be *automatically guaranteed*, i.e. Problem 3.4 solves our original Problem 3.1.

3.1 The Maximum Entropy Problem for Block–Circulant Covariance Matrices

The first question to be addressed is feasibility of Problem 3.4, namely the existence of a positive definite, symmetric matrix $\mathbf{\Sigma}_N$ satisfying (9.b)-(9.c). The following result can be established. We refer the reader to [3] and [4] for the proofs, as well as for a discussion and further details about the statements in this Section.

Theorem 3.1 (Feasibility) *Given the sequence $\Sigma_i \in \mathbb{R}^{m\times m}$, $i = 0, 1, \ldots, n$, such that*

$$\mathbf{T}_n = \mathbf{T}_n^\top > 0, \tag{10}$$

1. *there exists $\bar{N}$ such that for $N \geq \bar{N}$, the matrix $\mathbf{T}_n$ can be extended to an $N \times N$ block-circulant, positive-definite symmetric matrix $\mathbf{\Sigma}_N$.*

2. *The set of all positive definite block-circulant completions of $\mathbf{\Sigma}_N$ is delimited by the intersection of the m-order surfaces defined by the positive semidefiniteness of the matrices*
$$c(e^{-j\frac{2\pi}{N}\ell}) = \sum_{k=0}^{N-1} \Sigma_k e^{-j\frac{2\pi}{N}\ell k}, \qquad \text{for } \ell = 0, 1, \ldots, N-1.$$

Our main result is as follows.

Theorem 3.2 (Existence and Uniqueness – Bandedness property) *Let $\mathbf{\Sigma}_N$ be a partially positive definite block–circulant matrix that admits a positive definite block–circulant completion, then the CMaxEnt 3.4 admits a unique solution whose inverse is a block-circulant matrix banded of bandwidth n.*

This is the result we hoped for since it shows that the solution of the CMaxEnt in fact *solves our original maximum likelihood identification problem* 3.1. Moreover, this result, together with the uniqueness property of the solution of the Dempster problem, allows us to conclude that *the solution of the CMaxEnt and of the covariance selection problem with circulant data* (namely with data consistent with the circulant structure) *coincide*. Indeed it can be shown [4] that this equivalence holds true in general, i.e for any number of missing block-bands as well as arbitrary missing elements in a block-circulant structure. This generalization is based on an alternative approach to the proof, which relies on the observation that circulant and block-circulant matrices are stable points of a certain group. We refer the reader to [4] for further details on this.

Finally, we anticipate that the results of this Section lead to an efficient iterative algorithm for the solution of the CMaxEnt which is guaranteed to converge to a unique minimum (see [2]). The proposed algorithm compares very favorably with the best techniques available so far. This solves the circulant band extension problem and hence the maximum likelihood identification of reciprocal processes.

References

[1] S. Bernstein, *Sur les liaisons entre le grandeurs aleatoires.* In *Proc. Intern. Congr. Math.*, Zürich, Switzerland (1932), 288–309.

[2] F. P. Carli, A. Ferrante, M. Pavon, and G. Picci, *An efficient algorithm for maximum-entropy extension of block-circulant covariance matrices.* Preprint, Univ. of Padova (June 7, 2011).

[3] F. P. Carli, A. Ferrante, M. Pavon, and G. Picci, *A maximum entropy solution of the covariance extension problem for reciprocal processes.* To appear in IEEE Transactions on Automatic Control (2011).

[4] F. P. Carli and T. T. Georgiou, *On the covariance completion problem under a circulant structure.* IEEE Transactions on Automatic Control 56/4 (2011), 918–922.

[5] A. P. Dempster, *Covariance selection.* Biometrics 28 (1972), 157–175.

[6] R. Frezza, "Models of Higher-order and Mixed-order Gaussian Reciprocal Processes with Application to the Smoothing Problem". PhD thesis, Applied Mathematics Program, U. C. Davis, 1990.

[7] R. Grone, C. R. Johnson, E. M. Sa, and H. Wolkowicz, *Positive Definite Completions of Partial Hermitian Matrices.* Linear Algebra and Its Applications 58 (1984), 109–124.

[8] E. J. Hannan and M. Deistler, "The Statistical Theory of Linear Systems". Wiley, 1988.

[9] B. Jamison, *Reciprocal processes: The stationary gaussian case.* Ann. Math. Stat. 41 (1970), 1624–1630.

[10] B. C. Levy, R. Frezza, and A. J. Krener, *Modeling and estimation of discrete-time Gaussian reciprocal processes.* IEEE Trans. Automatic Control, AC 35/9 (1990), 1013–1023.

[11] A. Lindquist and G. Picci, *Realization theory for multivariate stationary Gaussian processes.* SIAM J. Control Optim. 23/6 (1985), 809–857.

[12] E. Schrödinger, *Sur la theorie relativiste de l'electron et l'interpretation de la mecanique quantique.* Ann. Inst. H. Poincaré 2 (1932), 269–310.

[13] T. P. Speed and H. T. Kiiveri, *Gaussian markov distribution over finite graphs.* The Annals of Statistics 14/1 (1986), 138–150.

On the essential dimension of groups

Dajano Tossici (*)

1 Introduction

In these notes we give an introduction to some aspects of the theory of essential dimension of groups. This notion, even if already known in particular cases already at the end of the nineteenth century, has been introduced in 1997 by Buhler and Reichstein [2]. Roughly speaking the essential dimension of a finite group G over a field k is the number of parameters to describe all the Galois extension E/F with Galois group G and containing k. In the present paper we mainly concentrate to the case of finite groups but however this notion can be generalized also to algebraic groups or more generally to group schemes: we will say something about this in the last section. Moreover the essential dimension can be used also in much more general contexts. We suggest [1] and [8] as an overview about essential dimension. The present notes are not at all exhaustive but they want just to give an idea of problems and results studied in the field of essential dimension of groups.

In the next section we give an historical motivation to study essential dimension. In the Section §3 we explain the problem in a more modern language. In the Sections §4, 5, 6 we collect some main results and main problems in characteristic zero. In §7 we deal with essential dimension of finite groups in positive characteristic. Finally, in the last section, we report the main results of [9] on essential dimension of group schemes. There will be a stylistic gap respect to the previous sections. Indeed in that section we will not give the precise definitions, which would require much time, but we will just report some statements of [9].

2 Classical problem

One of the main problems for mathematicians of the nineteenth century was the problem of finding a formula for the roots of a polynomial of fixed degree, using only radicals and the usual algebraic operations. Before the work of Galois, which gave a negative answer to this problem when the degree is at least 5, one of the main strategies consisted in simplifying the generic polynomial.

(*) Università Milano-Bicocca, Milano (Italy). E-mail: dajano.tossici@gmail.com . Seminar held on 15 June 2011.

Let k be a field of characteristic 0 end let n be a positive integer. Let us consider the generic polynomial of degree n

$$p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$$

where a_i are variables. Using a (non-degenerate) transformation

$$y = \alpha_{n-1}x^{n-1} + \cdots + \alpha_0,$$

with $\alpha_i \in k(a_0, \ldots, a_{n-1})$ for $i = 0, \ldots, n$, one obtains another polynomial

$$q(y) = y^n + b_{n-1}y^{n-1} + \cdots + b_0.$$

with possibly less parameters algebraically independent over k.

Problem 2.1 *Find the minimal number of parameters needed to define the generic polynomial of degree n.*

Example 2.2

(1) If $n = 2$, let us consider the transformation $y = x + \frac{a_1}{2}$, then we obtain $q(y) = y^2 + \frac{a_1^2}{4} + a_0$. If we set $b_0 := \frac{a_1^2}{4} + a_0$ then we have $q(y) = y^2 + b_0$. So we have reduced the number of parameters in the general polynomial to one parameter.

(2) Using a transformation $y = x + \frac{a_2}{3}$ one can reduce, as above, the general polynomial to a polynomial of type $q(y) = y^3 + b_1y + b_0$ with $b_0, b_1 \in k(a_0, a_1, a_2)$. With the transformation $z = \frac{b_1}{b_0}y$ we obtain the polynomial $r(z) = z^3 + \frac{b_1^3}{b_0^2}z + \frac{b_1^3}{b_0^2}$. So, if we set $c_0 = c_1 = \frac{b_1^3}{b_0^2}$, we obtain $z^3 + c_0z + c_0$. We have again one parameter.

(3) If $n = 4$, using a similar argument as above, one can reduce to a polynomial $r(z) = z^4 + c_2z^2 + c_0z + c_0$. This polynomial depends on two variables. We will see later that in fact one can not do better.

We finish the section with the following definition

Example 2.3 We call $d_k(n)$ the minimal number of parameters required to define the generic polynomial of degree n.

In the next section we will give the above definition in a more modern language. It follows from the examples above that $d_k(2) = d_k(3) = 1$ and $d_k(4) \leq 2$.

3 Formalization of the classical problem

For the next sections, until it will not be specified, k is a field of characteristic 0. Let E/F an extension field, containing k, of degree n. We say that E/F is defined over a field F_0 if F_0 contains k and there exists an extension field E_0/F_0 of degree n such that $E_0F = E$.

It is well known that, since k is of characteristic 0 any extension field E/F (necessarily separable) is obtained adjoining an element $\alpha \in F$ to E. So $E = F[x]/(p(x))$ where $p(x)$ is the minimal polynomial of α, i.e. the polynomial with minimal degree (hence irreducible) which has as root α. So, to say that F is defined over F_0, means that there exists an element $\beta \in F$ such that $F = E(\beta)$ and the minimal polynomial of β has coefficients in F_0.

Definition 3.1 Let E/F be an extension fields as above. We call *essential dimension* of E/F, and we denote it by $ed_k(E/F)$ the integer

$$\min\{trdeg_k F_0 \text{ such that } E/F \text{ is defined over } F_0\}.$$

We recall that if $k \subseteq F_0$ then $trdeg_k F_0$, the transcendence degree of F_0 over k, is the maximal cardinality of a subset of F_0 which consists by elements which do not satisfy any non-trivial polynomial equation with coefficients in k.

Now let $E_n = k(a_1, \ldots, a_n)$, $p(x)$ the generic polynomial of degree n as in the previous section and F_n the field $E_n[x]/(p(x))$. Then it follows, from what said before the definition and in the previous section, that

$$d_k(n) = ed_k(F_n/E_n).$$

Since any extension field of degree can be obtained by specialization by the generic extension M/L one can easily prove that

$$d_k(n) = \max\{ed_k(E/F) \text{ such that the degree of } E/F \text{ is } n\}$$

Now let G be a finite group. And let us suppose that E/F is a Galois extension with Galois group G. One can prove that if E/F comes from an extension E_0/F_0 with F_0 subfield of F then there exists a subfield F_0' of F, with same trascendence degree of F_0 over k, such that E/F comes from a Galois extension E_0'/F_0' with Galois group G (see [2, Lemma 2.2]).

So we arrive to the following definition.

Definition 3.2 Let G be a finite group. Then we define

$$ed_k G := \sup\{ed_k E/F \text{ such that } E/F \text{ is Galois with Galois group } G\}.$$

One can in fact prove that above supremum is in fact a maximum.

The above definition is different from that one given in [2, Theorem 3.1 (b)]. The equivalence is essentially proved in [2, Theorem 3.1 (c)].

Remark 3.3 We have that

$$ed_k(n) = ed_k S_n,$$

where S_n is the symmetric group on a set of n elements. The idea to prove this is to consider the Galois closure G_n of the extension E_n/F_n above. One proves that the extension G_n/F_n has Galois group S_n and its essential dimension is the same of E_n/F_n and of S_n (see [2, Corollary 4.2]).

So, for the rest of the notes we will study the essential dimension of groups.

4 Essential dimension of abelian groups over a field with enough root of unity

We begin with the simplest nontrivial example. Let us suppose that the field k contains a primitive n-th root of unity. And let us consider the cyclic group $\mathbb{Z}/n\mathbb{Z}$. If $k \subseteq F$, it is well known, by Kummer Theory, that any extension field E/F is obtained adjoining the n-th root of an element of $f \in E$ which is not a n-th power in E, i.e. $E = F[x]/(x^n - f)$. So E is defined over $k(f)$. This means that $ed_k\mathbb{Z}/n\mathbb{Z} \leq 1$. On the other hand considering as F the field $k(t)$ of rational funcrions in one variable and $f = t$ we have that $ed_k\mathbb{Z}/n\mathbb{Z}$ is exactly 1.

In the above example it is crucial the hypothesis on the existence of the n-th root of unity. We will see later what happens when this hypothesis is not satisfied. We now recall the following results.

Lemma 4.1

(1) *If $k \in k'$ then $ed_kG \leq ed_{k'}G$*

(2) *If H is a subgroup of G then $ed_kH \leq ed_kG$.*

(3) *If $G = G_1 \times G_2$ then $ed_kG \leq ed_kG_1 + ed_kG_2$.*

Proof. All the assertions are easy to prove. The first one is proved, in a more general context, in [1, Proposition 1.5]. The last two are proved in [2, Lemma 4.1]. □

Example 4.2 It is easy to find examples where there is no equality in the third statement of the above lemma. Let n, m two positive integer numbers coprime and let us suppose that k contains a primitive mn-th root of unity. Then let us consider $G = \mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/m\mathbb{Z}$. Since m and n are coprime then $G \simeq \mathbb{Z}/mn\mathbb{Z}$. From what said at the beginning of the section we have $ed_kG = 1$ which is smaller than $ed_k\mathbb{Z}/n\mathbb{Z} + ed_k\mathbb{Z}/m\mathbb{Z} = 2$.

But there are some cases when one has the equality. For instance there is the following result.

Theorem 4.3 *Let p be a prime number and let us suppose that k contains a primitive p-th of unity. Let $G = H \times \mathbb{Z}/p\mathbb{Z}$ and let us suppose that the center of H is a p-group. Then*

$$ed_kG = ed_kH + 1.$$

Proof. See [2, Corollary 5.5]. □

Using this result one can, for instance, compute the essential dimension of abelian groups if the base field contains enough roots of unity.

Corollary 4.4 *Let G be an abelian group of order n and let us suppose that k contains a primitive n-th root of unity. Then*

$$ed_k G = r,$$

where r is the rank G and it is equal to minimal number of generator of G.

Proof. By definition of rank it follows immediately that $G \simeq \mathbb{Z}/m_1\mathbb{Z} \times \cdots \times \mathbb{Z}/m_r\mathbb{Z}$. Therefore from Lemma 4.1 (3) it follows that $ed_k G \leq r$. On the other hand it follows quite easily by the defintion of the rank of a group that there exist a prime number p such that G contains a subgroup isomorphic to $(\mathbb{Z}/p\mathbb{Z})^r$. By Theorem 4.3 we have that $ed_k(\mathbb{Z}/p\mathbb{Z})^r = r$. So by Lemma 4.1 (2) it follows $ed_k G \geq r$ and we are done. □

5 Essential dimension of symmetric group S_n

In this section we come back to the initial question, i.e. to the computation of $d_k(n)$ which, as we have seen, is equal to the essential dimension of the symmetric group on a set of n elements. We begin to report some results due to Buhler and Reichstein [2]. Some of them follows by the results recalled in the previous section.

Theorem 5.1 *Let n be a positive integer. The following assertion are true.*

(1) $ed_k S_n \leq ed_k(S_{n+1})$.

(2) $ed_k S_n + 1 \leq ed_k S_{n+2}$.

(3) $ed_k S_n \geq [n/2]$.

(4) *If $n \geq 5$ the $ed_k S_n \leq n - 3$.*

Proof. (1) We observe that S_{n+1} contains a subgroup isomorphic to S_n, i.e. the group which fix the first element of the set with n elements. Then the assertion follows from Lemma 4.1 (2).

(2) This follows, as above, from the fact that S_{n+2} has a subgroup isomorphic to $S_n \times \mathbb{Z}/2\mathbb{Z}$, i.e. the direct product of the subgroup which fix the first two elements and the subgroup generated by the transposition which permutes the first two elements and fix all other elements.

(3) We proceed as above remarking that S_n contains a subgroup isomorphic to $(\mathbb{Z}/2\mathbb{Z})^{[n/2]}$, which has essential dimension $[n/2]$ by Corollary 4.4. This subgroup is given by the direct product of the subgroups generated by the transposition $(2k-1 \;\; 2k)$, for $k = 1, \ldots, [n/2]$. One could prove this result also reasoning by induction using the part (2).

(4) This is proved in [2, Theorem 6.5 (c)].

□

As an immediate corollary one obtains the computation of the essential dimension of S_n for small n.

Corollary 5.2 *We have*

- $ed_k S_2 = ed_k S_3 = 1$;
- $ed_k S_4 = ed_k S_5 = 2$;
- $ed_k S_6 = 3$.

We already proved the first assertion in the second section. The essential dimension of S_4 and S_5 was already known (using different terminology) by Klein and Hermite.

From the above theorem follows that $ed_k S_7$ is equal to 3 or to 4.

Theorem 5.3 *The essential dimension of S_7 is 4.*

Proof. See [4]. □

6 Essential dimension of cyclic groups

In this section we recall some results about the essential dimension of cyclic groups in the case the base field does not contain enough root of unity. The main result is the following.

Theorem 6.1 *Let p be a prime number and let us suppose that k contains a primitive root of unity. Then*

$$ed_k(\mathbb{Z}/p^n\mathbb{Z}) = [k(\zeta_{p^n}) : k]$$

where ζ_{p^n} is a primitive p^n-th root of unity.

Proof. The first proof of this result is due to Florence ([5, Theorem 4.1]). The above result is also a particular case of a more general result of Karpenko and Merkurjev ([6, Theorem 4.1]) in which they prove that the essential dimension of a p-group G is the minimal dimension of a faithful representation of G. And in the case of $G = \mathbb{Z}/p^n\mathbb{Z}$ the faithful representation with minimal dimension is given by the k-vector space $k(\zeta_{p^n})$ with the natural action of $\mathbb{Z}/p^n\mathbb{Z}$. □

If one removes the hypothesis on the p-th root of unity the result is not known. However by the result above and by Lemma 4.1 (1) it follows that, for any k, the essential dimension of $\mathbb{Z}/p^n\mathbb{Z}$ over k is at least $[k(\zeta_{p^n}) : k]$. For instance there is the following open problem.

Problem 6.2 *Compute the essential dimension of $\mathbb{Z}/p^n\mathbb{Z}$ over $\mathbb{Q}$.*

We remark that $\mathrm{ed}_{\mathbb{Q}}\mathbb{Z}/2^n\mathbb{Z} = 2^{n-1}$. This follows from the above theorem, since the 2-nd root of unity is -1 which is always in k. Moreover also $\mathrm{ed}_{\mathbb{Q}}\mathbb{Z}/3^n\mathbb{Z}$ is known and it is equal to 3^{n-1} (see [5, Corollary 4.2]).

7 Essential dimension in positive characteristic

We will now consider a field k of characteristic $p > 0$. The definition of essential dimension of a finite groups given in the case of characteristic zero works in fact in general. And all the results of the previous sections work if p do not divide the order of the group. In this section we will consider the case of p-groups in positive characteristic.

The simplest case to compute is the essential dimension of $\mathbb{Z}/p\mathbb{Z}$. It is well known, using Artin-Schreier Theory, that in positive characteristic if E/F is a Galois extension with Galois group $\mathbb{Z}/p\mathbb{Z}$ then

$$E = F[x]/(x^p - x - f),$$

with $f \in E$ and such that it is not of the form $g^p - g$ with $g \in E$. This implies that $ed_k(\mathbb{Z}/p\mathbb{Z}) = 1$.

Using Artin-Schreier-Witt Theory one shows, more generally, that in positive characteristic one can describe any Galois extension with Galois group $\mathbb{Z}/p^n\mathbb{Z}$ using at most n parameters. This shows that $ed_k\mathbb{Z}/p^n\mathbb{Z} \leq n$. And one has the following conjecture due to Ledet.

Conjecture 7.1 *The essential dimension of $\mathbb{Z}/p^n\mathbb{Z}$ over k is n.*

The above conjecture is true for $n = 2$ (see [7, p. 7]) but it is completely open for $n > 2$.

The opposite case to study is the case of abelian elementary p-groups. If the cardinality of k is at least p^n then

$$ed_k(\mathbb{Z}/p\mathbb{Z})^n = 1.$$

Indeed one can shows, generalizing Artin-Schreier Theory, that, under the hypothesis on the field, if E/F is a Galois extension with Galois group $(\mathbb{Z}/p\mathbb{Z})^n$ then $E = F[x]/(x^{p^n} - x - f)$ for some $f \in F$ which is not of the form $g^{p^n} - g$ with $g \in F$.

More generally one can compute, conjecturally, the essential dimension of any abelian p-group. Let $G = (\mathbb{Z}/p^{n_1}\mathbb{Z})^{r_1} \times \cdots \times (\mathbb{Z}/p^{n_k}\mathbb{Z})^{r_k}$ with $n_1 > n_2 > \cdots > n_k$. By Lemma 4.1 (2) we have $ed_kG \geq ed_k(\mathbb{Z}/p^{n_1}\mathbb{Z})^{r_1}$. Moreover $G \subseteq (\mathbb{Z}/p^{n_1}\mathbb{Z})^{r_1+\cdots+r_k}$, therefore we also have $ed_kG \leq ed_k(\mathbb{Z}/p^{n_1}\mathbb{Z})^{r_1+\cdots+r_k}$. But, generalizing the above argument, one can prove that, if the cardinality of k is at least p^n, then $ed_k(\mathbb{Z}/p^{r_1}\mathbb{Z})^n = ed_k(\mathbb{Z}/p^{r_1}\mathbb{Z})$. So $\mathrm{ed}_kG = ed_k(\mathbb{Z}/p^{r_1}\mathbb{Z})$, which conjecturally is r_1.

So one can see that this result is completely orthogonal to the result for abelian groups (with enough roots of unity on the base field) in characteristic zero, where the essential dimension is the rank of the group.

8 Essential dimension of group schemes

In this last section we briefly recall some results obtained by the author in collaboration with Vistoli [9]. We will consider in this section the essential dimension of affine group schemes of finite type over a field. Such an object is the zero locus of polynomials in some affine spaces over a field with some group structure. Important particular cases are

algebraic groups, i.e. groups with a structure of algebraic variety. Group schemes are not in general groups but they behave like them. It is known that in characteristic zero all group schemes are smooth over k, while in positive characteristic one can have non-smooth group schemes.

One can define the notion of essential dimension also for group schemes. Roughly speaking, the essential dimension of a group schemes is the number of parameters needed to define G-torsors. The notion of G-torsor is the equivalent, in algebraic geometry, of what is called, in other contexts, principal homogenous space. And it also generalizes the notion of Galois extensions of fields to inseparable extensions. We stress that this definition works also for group schemes of positive dimension, so not necessarily finite. We will not give here the precise definitions, which could be found in [9].

In the work with Vistoli we give two general bounds for essential dimension of group schemes.

Theorem 8.1 *Let G be an affine group scheme of finite type over a field k of characteristic $p \geq 0$. Then*

$$\mathrm{ed}_k G \geq \dim_k \mathrm{Lie} G - \dim G\,.$$

Proof. See [9, Theorem 1]. □

In fact it is not necessary that G is affine. In the above inequality, the term on the right measures how much the group scheme G is non-smooth: indeed, by definition, it is zero if and only if G is smooth. So the above bound is really interesting in positive characteristic for non-smooth group schemes. And we also remark that there exist non-smooth group schemes of dimension 0.

We also have a fairly general upper bound. Let us recall the definition of a trigonalizable group scheme.

Definition 8.2 Let G be an affine group scheme of finite type over a field k. We say that G is *trigonalizable* if it is a subgroup scheme of the group scheme of invertible upper triangular $n \times n$ matrices over k, for some n.

We observe that any affine commutative group scheme fo finite type over an algebraically closed field is trigonalizable (see [3, IV, §3, 1.1]).

Theorem 8.3 *Let G be a finite trigonalizable group scheme over a field of characteristic $p > 0$, of order p^n. Then $\mathrm{ed}_k G \leq n$.*

The particular case of finite abstract p-groups (which are unipotent over a field of positive characteristic) had already been proved by Ledet. We use in [9] these two results to compute the essential dimension of several group schemes.

References

[1] Grégory Berhuy and Giordano Favi, *Essential dimension: a functorial point of view (after A. Merkurjev)*. Doc. Math. 8 (2003), 279–330 (electronic).

[2] Joe Buhler and Zinovy Reichstein, *On the essential dimension of a finite group*. Compositio Math. 106/2 (1997), 159–179.

[3] Michel Demazure and Pierre Gabriel, "Groupes algéques. Tome I: Géométrie algébrique, généralités, groupes commutatifs". Masson & Cie, Éditeur, Paris, 1970. Avec un appendice *Corps de classes local* par Michiel Hazewinkel.

[4] Alexander Duncan, *Essential dimensions of A_7 and S_7*. Math. Res. Lett. 17/2 (2010), 263–266.

[5] Mathieu Florence, *On the essential dimension of cyclic p-groups*. Inventiones mathematicae 171 (2007), 175–189.

[6] Nikita A. Karpenko and Alexander S. Merkurjev, *Essential dimension of finite p-groups*. Inventiones Mathematicae 172/3 (2008), 491–508.

[7] Arne Ledet, *On the essential dimension of p-groups, Galois Theory and Modular Forms*. Dev. Math. 11, Kluwer Acad. Publ. (2004), 159–172.

[8] Zinovy Reichstein, *On the essential dimension of infinitesimal group schemes*. 26 pages, `http://www.math.ubc.ca/~reichst/pub.html`, to appear in Proceedings of the International Congress of Mathematicians 2010.

[9] Dajano Tossici and Angelo Vistoli, *Essential dimension*. 11 pages, `arXiv:1001.3988`, to appear in American Journal of Mathematics.