Università di Padova – Dipartimento di Matematica

Scuole di Dottorato in Matematica Pura, Matematica Computazionale e Informatica

Seminario Dottorato 2013/14

Preface

This document offers a large overview of the nine months' schedule of Seminario Dottorato 2013/14. 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, June 28th, 2014

Corrado Marastoni, Tiziano Vargiolu

Abstracts (from Seminario Dottorato's web page)

Wednesday 23 October 2013

An overview on the complex Monge-Ampere equation

HA Ly Kim (Padova, Dip. Mat.)

In this seminar we present a tutorial on the Dirichlet problems for the complex Monge-Ampere equation. After a large initial part devoted to a general introduction to the subject, we shall briefly concentrate on the study of Holder continuity of the solutions, showing how it can be handled for classical complex domains and also for some domains where the classical regularity properties fail to be true.

Wednesday 6 November 2013

Option pricing in a defaultable model: a characteristic function approach

Stefano PAGLIARANI (Padova, Dip. Mat.)

We consider a defaultable stock (i.e. a financial risky asset) whose predefault dynamics follows a stochastic differential equation driven by a Levy process. Under suitable assumptions on the default time, the price of a contingent claim (i.e. a financial derivative) is obtained in terms of the characteristic function (i.e. the Fourier transform) of the terminal log price. We characterize it as the solution of a complex valued infinite dimensional system of first order ordinary differential equations, which can be seen as an ordinary differential equation in a suitably chosen Banach space. By using this, we provide an explicit eigenfunction expansion for the characteristic function and use it to price contingent claims by means of standard Fourier inversion techniques. Finally, we present numerical results to demonstrate accuracy and efficiency of the method.

Wednesday 20 November 2013

What is a Hopf algebra ?

Nicolás ANDRUSKIEWITSCH (Cordoba, Argentina)

I will explain from scratch the notion of Hopf algebra, how and when it appeared, the basic examples and some applications.

Wednesday 4 December 2013

Some applications of potential theory to Markov chains

Alessandra BIANCHI (Padova, Dip. Mat.)

The link between potential theory and probability started in the last century with the work of Kakutani concerning the analysis of the Dirichlet problem. Since then, this connection has been explored by many authors and it has found applications in different contexts of probability. In this talk I will review some of these classical results and focus on applications to Markov chains.

Wednesday 18 December 2013

An introduction to the Clemens-Schmid exact sequence

Genaro HERNANDEZ MADA (Padova, Dip. Mat.)

In this talk, we give a very elementary introduction to the Clemens-Schmid exact sequence. In a classical setting, this is a topological result about certain families of complex varieties or manifolds. Therefore, the purpose of the talk is to explain all the concepts involved. If time allows, we will introduce the elements to understand in which cases we can obtain an arithmetic version of this result.

Wednesday 15 January 2014

An introduction to stochastic ergodic control

Marco CIRANT (Padova, Dip. Mat.)

In this talk we give an introduction to stochastic ergodic control problems, where an agent aims at minimizing a long-time average cost by controlling his own state. We will show, through a toy example, the main features of the problem and how it is possible to produce an optimal control by solving a suitable elliptic nonlinear partial differential equation. In the final part of the talk we will explore briefly how the minimization problem for a single agent can be considered more in general for a continuum of identical agents. This research field is called Mean Field Games and has attracted the experts' attention in the last ten years.

Wednesday 29 January 2014

An introduction to geometry over the field with one element

Federico BAMBOZZI (Padova, Dip. Mat.)

In this talk we first give a brief overview of the motivations behind the research on geometry over the field with one element. We then show one possible way to define affine schemes over the field with one element in analogy with the classical theories of algebraic varieties over the complex numbers and of schemes by Grothendieck. We end the talk by giving some examples of schemes over $\mathbb{F}_1$.

Wednesday 12 February 2014

Reachability problems via level set approach

Athena PICARELLI (ENSTA ParisTech and INRIA Saclay-Ile de France)

Given a controlled dynamical system, the characterization of the backward reachable set, i.e. the set of initial states from which it is possible to reach a given target set, can be very interesting in many applications. However realistic models may involve some constraints on state and/or control variables (for taking into account physical or economical constraints, obstacles, etc.) and this can make the characterization of this set much more complicated. After an introduction to the notion of backward reachability in the deterministic as well in the stochastic framework, aim of the talk is to present a technique, based on a level set approach, for characterizing and numerically computing the reachable set also if state constraints are taken into account.

Wednesday 26 February 2014

Minimum time function for linear control systems

NGUYEN Van Luong (Padova, Dip. Mat.)

In this talk, we will first introduce basic notions on linear control systems and minimum time function for linear control systems. We will end with some recent results on regularity of minimum time function for linear control systems.

Wednesday 12 March 2014

Market models with optimal arbitrage

CHAU Ngok Huy (Padova, Dip. Mat.)

In this talk, we will introduce basic notions on financial mathematics, classical no arbitrage theory and some results on markets with arbitrage. We present a systematic method to construct market models where the optimal arbitrage strategy exists and is known explicitly.

Wednesday 26 March 2014

Shape optimization and polyarmonic operators

Davide BUOSO (Padova, Dip. Mat.)

We will start by introducing the general shape optimization problem, giving motivations for its importance in applications. Then we will turn to the problem of shape optimization for eigenvalues of elliptic operators (in particular, poyharmonic operators), which has regained popularity since 1993 with the paper by Buttazzo and Dal Maso. We will recall the most important classical results, giving the main ideas behind the proofs, together with the last ones. Finally, we will move our attention to the problem of criticality with respect to shape deformations for eigenvalues of polyharmonic operators. After explaining the techniques involved, we will provide a characterization of criticality and show that balls are always critical.

Wednesday 9 April 2014

Geometric modeling and splines: state of the art and outlook

Michele ANTONELLI (Padova, Dip. Mat.)

We will give an introductory presentation of the research field of geometric modeling and its applications, with specific attention to the use of splines for the representation of curves and surfaces. In particular, we will start by introducing basic notions of geometric modeling leading up to the definition of splines, which are piecewise functions with prescribed smoothness at the locations where the pieces join. Splines will be exploited for the representation of parametric curves and surfaces, and we will present their application in the context of computer-aided geometric design for shape description by means of approximation and interpolation methods. Finally, we will discuss some open problems in this topic and we will sketch some recent approaches for addressing them.

Wednesday 30 April 2014

An introduction to representation theory of groups

Martino GARONZI (Padova, Dip. Mat.)

Label the faces of a cube with the numbers from 1 to 6 in some order, then perform the following operation: replace the number labeling each given face with the arithmetic mean of the numbers labeling the adjacent faces. What numbers will appear on the faces of the cube after this operation

is iterated many times? This is a sample problem whose solution is a model of the application of the theory of representations of groups to diverse problems of mathematics, mechanics, and physics that possess symmetry of one kind or another. In this introductory talk I will present the tools from representation theory needed to solve this problem. I will also point out the connection with harmonic analysis by expressing Fourier analysis as an instance of representation theory of the circle group (the multiplicative group of complex numbers with absolute value 1) and by stating a version of Heisenberg's uncertainty principle for finite cyclic groups.

Wednesday 7 May 2014

Extrapolation techniques and applications to row-action methods

Anna KARAPIPERI (Padova, Dip. Mat.)

The talk will be divided in three parts. First we will introduce extrapolation methods and notions related to them, such as kernel and convergence acceleration. These definitions will be well understood by the examples of Aitken's process, Shanks' transformation and various generalizations. Afterwards, we will pass to row-action methods that have several interesting properties (i.e. no changes to the original matrix and no operations on the matrix as a whole). We will focus on Kaczmarz and Cimmino method. At the end we will see how extrapolation methods can be used for accelerating the convergence of the aforementioned row-action methods.

Wednesday 21 May 2014

A visual introduction to tilting

Jorge VITÓRIA (Verona)

The representation theory of a quiver (i.e., an oriented graph) can sometimes be understood by... another quiver! Such pictures of complex concepts (such as categories of modules or derived categories) are a source of intuition for many phenomena, among which lie the tools for classification and comparison of representations: tilting theory. The aim of this talk is to give an heuristic view (example driven) of some ideas in this area of Algebra.

Wednesday 4 June 2014

Introduction to representation growth

Michele ZORDAN (Bielefeld)

This seminar is intended as an accessible introduction to representation zeta functions. Given a

group, representation zeta functions are Dirichlet generating functions encoding the numbers of its irreducible representations sorted by dimension. This analytic tool allows the use of analytic methods to compute the rate of growth of the numbers of irreducible representations as their dimension grows. Much akin to the Riemann's zeta function, these representation zeta functions are often Euler's product of local factors. The computation of these factors, therefore, holds the key to understanding the representation growth of the group. In this talk I shall introduce the subject with appropriate examples and discuss the methods that given a group allow us to compute the local factors.

Wednesday 18 June 2014

Jacobi matrices, orthogonal polynomials and Gauss quadrature. An introduction and some results for the non-hermitian case.

Stefano POZZA (Padova, Dip. Mat.)

This seminar is divided in two parts. In the first one we will give an introduction to the matter. We will present the concept of orthogonal polynomials and we will focus on the properties of the Jacobi matrix linked to them. Then we will see their application for the approximation of integrals (Gauss quadrature). In the second part we will see how the first part can be extended to the non-hermitian case. In particular we will present formal orthogonal polynomials and we will see the spectral properties of a generally complex Jacobi matrix. This will lead us to some results about the extension of the Gauss quadrature in the complex plane.

An overview on the complex Monge-Ampère equation

Ly Kim Ha (*)

Abstract. In this seminar we present a tutorial on the Dirichlet problem for the complex Monge-Ampère equation

$$\begin{cases} (dd^c u)^n = h\beta_n \geq 0 & \text{in } \Omega \\ u = \phi & \text{on } \partial\Omega. \end{cases}$$

After a large initial part devoted to a general introduction to the subject, we shall briefly concentrate on the study of Hölder continuity of the solutions, showing how it can be handled for classical complex domains and also for some domains where the classical regularity properties fail to be true.

1 Introduction

The seminar will survey some boundary regularity theorems for complex Monge-Ampère equation which are proved by pluripotential theory methods. In 1976, using Pluripotential theory, Bedford and Taylor [2] defined the complex Monge-Ampère operator $(dd^c)^n$ in the sense of currents. We can think that currents are forms whose coefficients are distributions. The well-known result in their paper is that the unique solution of the Dirichlet problem for the complex Monge-Ampère equation is Hölder continuous. In particular, when Ω is a bounded, strongly pseudoconvex domain in $\mathbb{C}^n$ with C^2 boundary $b\Omega$. Then, if $\phi \in Lip^{\alpha}(b\Omega)$, $0 \leq h^{\frac{1}{n}} \in Lip^{\frac{\alpha}{2}}(\overline{\Omega})$, where $0 < \alpha \leq 1$, in the pluripotential sense, the Dirichlet problem

$$\begin{cases} (dd^c u)^n = h\beta_n \geq 0 & \text{in} \quad \Omega \\ u = \phi & \text{on} \quad \partial\Omega. \end{cases} \tag{1}$$

has an unique solution $u \in Lip^{\frac{\alpha}{2}}(\overline{\Omega})$.
The main purpose is to investigate the Hölder continuity of the unique solution of (1) on more general domains in $\mathbb{C}^n$. Actually, in [16], the reader can find the preliminary of these

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: lykimha35@yahoo.com.vn . Seminar held on October 23rd, 2013.

such equations on Kähler manifolds. This content will be skipped in the seminar.
The seminar is organized as follows. First, we view the geometric notations of pseudoconvexity and D'Angelo type following [15, 9]. Then, in the second section, we will study the pluripotential definition of complex Monge-Ampère equations. A potential condition, namely f-Property, comes from the paper by Khanh and Zampieri is introduced in the third chapter. And the influence of this condition on the boundary regularity for (1) is explained in the after section. Finally, some examples will illustrate the main results.

2 Pseudoconvexity and D'Angelo type

Definition 2.1

(a) A real function u on $\omega \subset \mathbb{C}$ with values in $[-\infty, \infty)$ is subharmonic when u is upper semicontinuous, not identically $-\infty$ on any connected component of ω, and for every disc $D(x_0, r) \subset\subset \omega$, the following sub-mean value property holds

$$u(x_0) \leq \frac{1}{2\pi r} \int_{bD(x_0,r)} u(x) d\sigma(x),$$

where $d\sigma(x)$ is the element of the arc.
In $\mathbb{C}^n$, the notion of subharmonicity is extended to plurisubharmonicity. In particular

(b) A function u defined in a open set Ω in $\mathbb{C}^n$, taking values in $[-\infty, +\infty)$, and not identically $-\infty$ on any connected component of Ω, is called plurisubharmonic (in short, p.s.h) if it is u.s.c in Ω and subharmonic on any intersection of a complex line with Ω, i.e, for every $z_0 \in \Omega$ and $z \in \mathbb{C}^n$, the function $t \mapsto u(z_0 + tz)$ is subharmonic in a neighborhood of 0 in the complex plane.
We denote by $\mathcal{P}(\Omega)$ the space of p.s.h functions on Ω.

A remark that $u \in \mathcal{P}(\Omega) \cap C^2(\Omega)$, if and only if the complex Hessian of u, say $H(u)$, is positive semi-definite, that is

$$\sum_{j,k} \frac{\partial^2 u}{\partial z_j \partial \overline{z}_k} w_j \overline{w}_k \geq 0, \quad w \in \mathbb{C}^n.$$

As the mention before, we will focus on geometry on domains in $\mathbb{C}^n$ in this section. Everything is started from the notions of pseudoconvexity.

Definition 2.2 (Levi form)
Let Ω be an open, bounded domain in $\mathbb{C}^n$, $n \geq 2$, and ρ be a C^2 defining function for Ω. We define the tangent space of type $(1, 0)$ at the point $p \in b\Omega$ by

$$T_p^{1,0}(b\Omega) = \{w \in \mathbb{C}^n : \sum_{j=1}^{n} w_j \frac{\partial \rho}{\partial z_j}(p) = 0\}.$$

The Levi form of ρ at the point p is

$$\mathcal{L}_p(\rho, w) = \sum_{j,k=1}^{n} \frac{\partial^2 \rho}{\partial z_j \partial \bar{z}_k} w_j \bar{w}_k.$$

Remark 2.3 The sign of the Levi form associated to Ω is independent of the defining function up to a positive factor.

Definition 2.4 (Notions of Pseudoconvexity)
Let Ω be a bounded domain in $\mathbb{C}^n$ with $n \geq 2$. Then, Ω is called to be:

(a) Levi-pseudoconvex: if $\mathcal{L}_p(\rho, w) \geq 0$, for all $w \in T_p^{1,0}(b\Omega)$, and all $p \in b\Omega$.

(b) Lelong-pseudoconvex: if it admits a function $\phi \in \mathcal{P}(\Omega) \cap C(\Omega)$ such that: $\forall a \in \mathbb{R}, \{z \in \Omega : \phi(z) \leq a\}$ is compact.

(c) hyperconvex: if it admits a negative function $\phi \in \mathcal{P}(\Omega) \cap CB(\Omega)$ such that: $\forall a < 0, \{z \in \Omega : \phi(z) \leq a\}$ is compact.

(d) strictly pseudoconvex: if $\mathcal{L}_p(\rho, w) > 0$, for all $w \in T_p^{1,0}(b\Omega)$, and all $p \in b\Omega$.

Remark 2.5

(a) If $b\Omega$ is C^2, Levi-pseudoconvexity $\Leftrightarrow$ Lelong-pseudoconvexity, hence we call, in short, weakly pseudoconvexity.

(b) In general, hyperconvexity $\Rightarrow$ LeLong pseudoconvexity. The converse holds if $b\Omega$ is Lipschitz (Demaily 1987).

Example 2.1 Hartogs triangle, $\{(z, w) \in \mathbb{C}^2 : |z| < |w| < 1\}$ is LeLong-pseudoconvex, but neither convex nor hyperconvex.
More details, see the well-known books by S. Krantz [17], S. Kobayashi [15].

In the famous book [9], the author introduced the notion of type on manifolds of real hypersurfaces to investigate the subelliptcity of the $\bar{\partial}$-Neumann problem, another linear problem in multi-dimensional complex analysis. This is, probably, a most beautiful condition in the theory of partial differential equations in several complex variables. Here, we will apply the notion to study complex Monge-Ampère equations.

Definition 2.6 (D'Angelo type)

(a) A holomorphic curve at $p \in b\Omega$ is a non-constant holomorphic map ϕ from a neighborhood of 0 in $\mathbb{C}$ to $\mathbb{C}^n$ such that $\phi(0) = p$.

(b) A germ of holomorphic curve ϕ at $p \in b\Omega$ is by the following equivalence relation

$$[\phi]_p = \{\psi : \mathbb{C} \to \mathbb{C}^n \text{ holomorphic curve at } p \text{ such that}$$

$$\exists U \subset \mathbb{C}, 0 \in U, \phi|_U = \psi|_U\}$$

(c) The corresponding quotient is $\mathbb{C}(n,p) := \{[\phi]_p\}$.

(d) Let ϕ be a holomorphic function of one complex variable. The order of vanishing at the origin $\text{ord}_0\phi = s$ of ϕ if

$$\frac{d}{dt}\phi(0) = ... = \frac{d^{s-1}}{dt^{s-1}}\phi(0) = 0 \neq \frac{d^s}{dt^s}\phi(0).$$

If $\phi = (\phi_1, ..., \phi_n)$, we define $\text{ord}_0\phi = \min_{1\leq j\leq n} \text{ord}_0\phi_j$.

(e) Let ρ be a defining function of $\Omega \subset \mathbb{C}^n$, the D'Angelo 1^{st}-type at $p \in b\Omega$

$$\Delta_1(b\Omega, p) = \sup_{\phi\in\mathbb{C}(n,p)} \frac{\text{ord}_0\phi^*\rho}{\text{ord}_0\phi}.$$

If $\Delta_1(b\Omega, p) = m < \infty$, we call p a point of finite D'Angelo 1^{st}-type of m.

Remark 2.7

(a) $\Delta_1(b\Omega, p)$ is independent of the choice of the defining function ρ.

(b) For domains in $\mathbb{C}^2$, finite D'Angelo 1^{st}-type and finite Hörmander type are equivalent conditions.

(c) Ω is strictly pseudoconvex at $p \in b\Omega$ if and only if $\Delta_1(b\Omega, p) = 2$ (See in the book by John D'Angelo [9]).

Now, we have some examples to illustrate the notions of pseudoconvexity and D'Angelo type.

Example 2.2

(a) Let define a complex ellipsoid by

$$\Omega = \{(z_1, ..., z_n) \in \mathbb{C}^n : \sum_{j=1}^{n} |z_j|^{2m_j} < 1\}$$

.
Here m_j, $j = 1, .., n$, are positive integers. This domain is strongly pseudoconvex if $m_j = 1$, for all $j = 1, ..., n$ and in this case, Ω, of course, is of finite D'Angelo type 2. When $m_j \geq 2$, for some $j = 1, ..., n$, Ω is weakly pseudoconvex and of finite type of $m = \max_{j=1,...,n} m_j$.

(b) If the domain $\Omega = \{(z_1, z_2, z_3) \in \mathbb{C}^3 : 2\Re(z_3) + |z_1^2 - z_2^3|^2 + |z_1^4|^2 - |z_1 z_2^m|^2\}$, with $m \geq 6$, then it is of finite type $2m$, but not pseudoconvex.

(c) Final, the domain

$$\Omega = \{(z_1, z_2) \in \mathbb{C}^2 : 2e^{-\frac{1}{|z_1|^\alpha}} + |z_2 - 1|^2 < 1\}$$

is pseudoconvex, but of infinite D'Angelo type, where $0 < \alpha < 1$.

From the hypothesis of finite D'Angelo type, Catlin (1987) applied the method of weight functions used earlier by Hörmander to prove a sub-elliptic estimate for $\bar{\partial}$-Neumann problem (we do not discuss this problem here). In particular, he showed that

Theorem 2.8 ([4, 5])
Suppose that $\Omega \subset\subset \mathbb{C}^n$ is a pseudoconvex domain with defining function ρ. Moreover, we assume that Ω is of finite D'Angelo type ≥ 2. Then, there exists a neighborhood U of $b\Omega$ and a family of functions $\{\phi_\delta\}_{\delta>0}$ such that

- *ϕ_δ is p.s.h and C^2 for all $\delta > 0$ on U,*
- *$-1 \leq \phi_\delta < 0$ on U,*
- *For any $z \in U \cap \{z \in \Omega : -\delta < \rho(z) < 0\}$ it holds* (here $\epsilon \leq \frac{2^{n-2}}{m^{n-1}}$)

$$\sum_{j,k} (\phi_\delta)_{j,\bar{k}}(z) a_j \bar{a}_k \gtrsim \delta^{-2\epsilon} |a|^2.$$

In Section 4, we will consider this description as a property which characterizes to the notion of D'Angelo' type (including finite and some infinite type) in several complex variables.

3 Currents and Complex Monge-Ampère Equations

First of all, we re-called what a differential form in $\mathbb{C}^n$ means.

Definition 3.1 Let $1 \leq p, q \leq n$ be integers, w is a differential (p, q)-form means

$$w = \sum\nolimits'_{|J|=p,|K|=q} w_{JK} dz_J \wedge d\bar{z}_K,$$

where :

(a) w_{JK}'s are smooth functions or distributions;

(b) $J = (j_1, ..., j_p), K = (k_1, ..., k_q)$ are multi-indexes of non-negative integers, and the length $|J| = j_1 + ... + j_p$, $|K| = k_1 + ... + k_q$;

(c) $dz_J := dz_{j_1} \wedge ... \wedge dz_{j_p}$, $d\bar{z}_K := d\bar{z}_{k_1} \wedge ... \wedge d\bar{z}_{k_q}$;

(d) the symbol $\sum'$ means we take the sum over multi-indexes J, K such that $1 \leq j_1 < ... < j_p \leq n$, $1 \leq k_1 < ... < k_q \leq n$.

Next some complex linear differential operators will be needed. On $\mathbb{C}^n$, we define

$$\frac{\partial}{\partial z_j} = \frac{1}{2}\left(\frac{\partial}{\partial x_j} - i\frac{\partial}{\partial y_j}\right) \quad \text{and} \quad \frac{\partial}{\partial \overline{z}_j} = \frac{1}{2}\left(\frac{\partial}{\partial x_j} + i\frac{\partial}{\partial y_j}\right).$$

By this way, in the distribution sense, we associate $(1,0)$-forms and $(0,1)$-forms

$$\partial u = \sum_{j=1} \frac{\partial u}{\partial z_j} dz_j \quad \text{and} \quad \bar{\partial} u = \sum_{j=1} \frac{\partial u}{\partial \overline{z}_j} d\overline{z}_j.$$

Actually, these operator can be defined on complex (p,q)-differential forms, but in this talk, we only consider in the case of acting on functions. (About L^2 theory for these operator, see Shaw and Chen's book [23]).
We set

$$d = \partial + \bar{\partial} \quad \text{and} \quad d^c = i(\bar{\partial} - \partial),$$

so that $dd^c u = 2i\partial\bar{\partial} u$ is a $(1,1)$-form. Where if $u \in C^2$, by a direct calculation,

$$\partial\bar{\partial} u = \sum_{j,k=1}^{n} \frac{\partial^2 u}{\partial z_j \partial \overline{z}_k} dz_j \wedge d\overline{z}_k,$$

is a $(1,1)$-form.
Now, we can state the main object in this seminar. The Dirichlet problem for the complex Monge-Ampère equation is to solve the following

$$\begin{cases} (dd^c u)^n = h\beta_n & \text{in } \Omega \\ u = \phi & \text{on } b\Omega, \end{cases} \tag{2}$$

where $(dd^c u)^n = \underbrace{dd^c u \wedge dd^c u \wedge ... \wedge dd^c u}_{n \text{ times}}$, $h \geq 0$ is continuous in $\overline{\Omega}$, $\phi \in C(b\Omega)$ and $\beta_n = \left(\frac{i}{2}\right)^n \prod_{j=1}^n dz_j \wedge d\overline{z}_j$.
In the classical sense, $u \in C^2(\Omega)$, the left hand side of (2) is well-defined,

$$(dd^c u)^n = 4^n n! \det\left(\frac{\partial^2 u}{\partial z_j \partial \overline{z}_k}\right) \beta_n$$

Since $\left(\frac{\partial^2 u}{\partial z_j \partial \overline{z}_k}\right) \geq 0$ when $u \in \mathcal{P}(\Omega) \cap C^2(\Omega)$, we say $dd^c u$ is a positive $(1,1)$ form. Then, the right hand side of (1) is necessarily non-negative Borel measure.
However, if we require $u \notin C^2(\Omega)$, the classical definition of the operator $(dd^c)^n$ does not work. We need another argument to understand the operator acting on plurisubharmonic, continuous functions.

Definition 3.2 We denote

(a) $\mathcal{D}_{p,q}(\Omega)$: test forms (means the forms whose coefficients are test functions in Schwartz Space) on Ω equipped with Schwartz's topology.

(b) Current of dimension (p,q) (or of bidegree $(n-p,n-q)$): is a continuous, linear functional on $\mathcal{D}_{p,q}(\Omega)$.

(c) The space of such currents : $\mathcal{D}'_{p,q}(\Omega)$.

(d) Let $T \in \mathcal{D}'_{(p,p)}(\Omega)$, we say T is a positive current if

$$(T,\omega) \geq 0,$$

for any $\omega = i^p \omega_1 \wedge \overline{\omega}_1 \wedge ... \wedge \omega_p \wedge \overline{\omega}_p$, with ω_k's $\in C^\infty_{(1,0)}$.

(e) For two (p,p)-currents S,T, the inequality

$$S \leq T$$

means that $T-S$ is a positive current.

We can think that a such (p,q) current is a $(n-p,n-q)$ form with coefficients in the space of distributions on Ω. This means

$$T = \sum\nolimits'_{|J|=n-p,|K|=n-q} T_{JK} dz_J \wedge d\bar{z}_K,$$

where

$$(T_{JK},\phi) = (T,\phi\alpha_{JK}),$$

and

$$\alpha_{JK} = \lambda dz_{J'} \wedge d\bar{z}_{K'},$$

such that $dz_J \wedge d\bar{z}_K \wedge \alpha_{JK} = \beta_n$.
Moreover, if T' is a positive current on Ω', $f : \Omega \to \Omega'$ is a bi-holomorphic mapping, then the pull-back $T := f^*T'$ defined by

$$(T,\omega) = (T',(f^{-1})^*\omega))$$

is also a positive current in Ω.
We may also define a wedge product of a current and a smooth form ω setting

$$(T \wedge \omega, \phi) := (T, \omega \wedge \phi)$$

for any test form ϕ.
We differentiate currents according to the formula

$$(DT,\phi) = -(T,D\phi),$$

for a first order differential operator D.
Now, let $u \in \mathcal{P}(\Omega) \cap C(\Omega)$, then $dd^c u$ is a bounded, positive of bi-dimension $(1,1)$ current, and $u.dd^c u$ is a well-defined current, so is

$$dd^c u \wedge dd^c u := dd^c(u.dd^c u),$$

in the sense that

$$\int \phi.dd^c u \wedge dd^c u = \int u.dd^c \phi \wedge dd^c u.$$

The latter current is also closed and positive. By this way, we may defined closed positive currents for $u \in \mathcal{P}(\Omega) \cap C(\Omega)$

$$(dd^c u)^m := \underbrace{dd^c u \wedge dd^c u \wedge ... \wedge dd^c u}_{m \text{ times}}.$$

Some important properties of the complex Monge-Ampère operator are:

Property 3.3 Let $u, v \in \mathcal{P}(\Omega) \cap C(\Omega)$, then:

(a) $(dd^c[\max(u,v)])^n \geq \min[(dd^c u)^n, (dd^c v)^n]$.

(b) $(dd^c[u+v])^n \geq (dd^c u)^n + (dd^c v)^n$.

The following "Minimum Principle" is derived for this extension of the complex Monge-Ampère operator $(dd^c)^n$

Theorem 3.4 [Bedford-Taylor 1976] *Let Ω be a bounded open set in $\mathbb{C}^n$. If $u, v \in C(\bar{\Omega})$ are plurisubharmonic, and if $(dd^c u)^n \leq (dd^c v)^n$ in the sense of currents, then*

$$\min_{z \in \bar{\Omega}} \{u(z) - v(z)\} = \min_{z \in b\Omega} \{u(z) - v(z)\}.$$

Now, we have the Dirichlet problem for the Complex Monge-Ampère equation in the sense of currents.
Let Ω be a bounded open set in $\mathbb{C}^n$ with C^2-boundary $b\Omega$. The Dirichlet problem for complex Monge-Ampère is to seek a function u such that

$$\begin{cases} u \in \mathcal{P}(\Omega) \cap C(\overline{\Omega}) \\ (dd^c u)^n = h\, dV \quad (f \geq 0,\ dV \text{ Lebesgue measure }) \\ u(z) = \phi(z) \quad (z \in b\Omega,\ \phi \in C(b\Omega)) \end{cases} \tag{3}$$

We can write $(dd^c u)^n = h\, dV$ by $\det H(u) = h$.

In this seminar, we will discuss how geometric conditions on Ω affect to regularity of the Dirichlet problem. For this, we will recall some fundamental results for boundary regularity .

- In [8], the smoothness of solution of (3) was also established. In particular, on a bounded strongly pseudoconvex domain with smooth boundary, if $\phi \in C^\infty(b\Omega)$, then there exists an unique solution $u \in C^\infty(\overline{\Omega})$ when h is smooth and strictly positive on $\overline{\Omega}$. Their approach followed the continuity method applied to the real Monge-Ampère equations.

- More generally, Blocki also considered the Dirichlet problem (3) on hyperconvex domains in [1]. In the paper, when data $\phi \in C(b\Omega)$ can be continuously extended to a plurisubharmonic function on Ω and the right hand is nonnegative, continuous, then the plurisubharmonic solution exists uniquely and continuously. However, the Hölder continuity for the solution on these domains was not verified.

- In [7], Coman showed how to connect some geometrical conditions on domains in $\mathbb{C}^2$ to the existence of plurisubharmonic upper envelope in Hölder spaces. In particular, the weakly pseudoconvexity of finite type m and the fact that the Perron-Bremermann function belongs to $Lip^{\frac{\alpha}{m}}$ with corresponding data in Lip^α are equivalent. Again, this means the condition of finite type plays the critical role in Hölder regularity for complex Monge-Ampère equations.

- Li [20] studied the problem on domains admitting a non-smooth, uniform and strictly plurisubharmonic defining function. In particular, if Ω admits an uniform and strictly plurisubharmonic defining function in $Lip^{\frac{2}{m}}(\bar{\Omega})$when $0 < \alpha \leq \frac{2}{m}$, and $\phi \in Lip^\alpha(b\Omega)$ and if $h^{\frac{1}{n}} \in Lip^{\frac{\alpha}{m}}$, then the unique existence of the solution for (3) $u \in Lip^\alpha(\overline{\Omega})$ holds. Based on results by Catlin in [6] and by Fornaess-Sibony in [10], there exists a plurisubharmonic defining function in $Lip^{\frac{2}{m}}(\bar{\Omega})$ on pseudoconvex domains of finite type m in $\mathbb{C}^2$ or convex domains of finite type m in $\mathbb{C}^n$.
  The author also gives the example on complex ellipsoid to show that this result is optimal. The critical point in the proof is based on the observation by Catlin and the main result of Fornaess and Sibony in $\mathbb{C}^2$ that a such domain exists under some assumptions

The next parts will concentrate to answer the question that: what is the boundary behaviour (here, we mean Hölder continuity up to the boundary) of the solution u of the problem (3) when Ω DOES NOT satisfy the condition of finite D'Angelo type? This requires that we have to :

(a) provide a geometric condition which generalizes the condition of D'Angelo type.

(b) introduce "suitable" Hölder spaces with the above condition.

(c) construct a solution admitting the new Hölder continuity.

4 f-Property and Main result

The f-Property which was introduced by Khanh and Zampieri generalizes the existence of the Catlin's family of weights on pseudoconvex domains of finite D'Angelo type (2.8).

Definition 4.1 (f-Property) For a smooth monotonic increasing function $f : [1 + \infty) \to [1, +\infty)$ with $\frac{f(t)}{t^{1/2}}$ decreasing, we say that Ω has the f-Property if there exist a neighborhood U of $b\Omega$ and a family of functions $\{\phi_\delta\}$ such that

(i) functions ϕ_δ are plurisubharmonic, $-1 \leq \phi_\delta \leq 0$ and C^2 on U;

(ii) $i\partial\bar{\partial}\phi_\delta \gtrsim f(\delta^{-1})^2 Id$ and $|D\phi_\delta| \lesssim \delta^{-1}$ for any $z \in U \cap \{z \in \Omega : -\delta < r(z) < 0\}$, where r is a defining function of Ω.

Remark 4.2 In [4, 5], Catlin proved that every smooth, pseudoconvex domain Ω of finite type m in $\mathbb{C}^n$ is of the f-Property with $f(t) = t^\epsilon$ with $\epsilon = m^{-n^2 m^{n^2}}$. Specially, if Ω is strongly pseudoconvex, or else it is pseudoconvex of finite type in $\mathbb{C}^2$, or else decoupled or convex in $\mathbb{C}^n$ then $\epsilon = \dfrac{1}{m}$ where m is the finite type (cf. [6, 12, 21, 22]).

Remark 4.3 The relation of the general type (both finite and infinite type) and the f-Property has been studied by Khanh and Zampieri [12, 19]. Moreover, they proved if $P_1, ..., P_n : \mathbb{C} \to \mathbb{R}^+$ are functions such that $\Delta P_j(z_j) \geq \dfrac{F(|x_j|)}{x_j^2}$ or $\dfrac{F(|y_j|)}{y_j^2}$ for any $j = 1, ..., n$, then the pseudoconvex ellipsoid

$$C = \{(z_1, \ldots, z_n) \in \mathbb{C}^n : \sum_{j=1}^{n} P_j(z_j) \leq 1\}$$

has f-Property with $f(t) = (F^*(t^{-1}))^{-1}$. Here we denote F^* is the inverse function of F.

In this seminar, using the f-Property we prove a "weak" Hölder regularity for the solution of the Dirichlet problem of complex Monge-Ampère equation. For this purpose we recall a suitable definition of the Hölder continuous spaces in [13].

Definition 4.4 Let f be an increasing function such that $\lim\limits_{t\to+\infty} f(t) = +\infty$. For $\Omega \subset \mathbb{C}^n$, define the f-Hölder space on $\overline{\Omega}$ by

$$\Lambda^f(\overline{\Omega}) = \{u : \|u\|_\infty + \sup_{z,w\in\overline{\Omega}} f(|z - w|^{-1}) \cdot |u(z) - u(w)| < \infty\}$$

and set

$$\|u\|_f = \|u\|_\infty + \sup_{z,w\in\overline{\Omega}} f(|z - w|^{-1}) \cdot |u(z) - u(w)|.$$

Note that the f-Hölder space includes the standard Hölder space $\Lambda_\alpha(\overline{\Omega})$ by taking $f(t) = t^\alpha$ (so that $f(|h|^{-1}) = |h|^{-\alpha}$) with $0 < \alpha < 1$.
The main result in this section is following.

Theorem 4.5 (In joint work with T.V.Khanh [11])
Let f be in Definition 4.1 such that $g(t)^{-1} := \int_t^\infty \frac{da}{af(a)} < \infty$. Assume that Ω is a bounded, pseudoconvex domain admitting the f-Property. Then, for any $0 < \alpha \leq 1$, if $\phi \in \Lambda^{t^\alpha}(b\Omega)$, and $h \geq 0$ on Ω with $h^{\frac{1}{n}} \in \Lambda^{g^\alpha}(\overline{\Omega})$, then the following Dirichlet problem of complex Monge-Ampère equation

$$\begin{cases} \det(u_{ij}) = h & in \quad \Omega, \\ u = \phi & on \quad \partial\Omega. \end{cases} \tag{4}$$

has an unique plurisubharmonic solution $u \in \Lambda^{g^\alpha}(\overline{\Omega})$.

An example in the last section will show that this result is sharp in some sense.

5 Some special domains

To end this seminar, we will exhibit some special domains to show that our estimate is sharp.

Example 5.1 (On strongly pseudoconvex domains Let Ω be a bounded, strongly pseudoconvex domain in $\mathbb{C}^2$ with the symmetry $(z, w) \in \Omega$ if and only if $(w, z) \in \Omega$ (e.g, the unit ball $\mathbb{B}_0$).
Let $h \geq 0$ satisfy $h^{1/2} \in C_0^\infty(\Omega)$, $\text{supp}(h) \cap \{z = 0\} = \varnothing$, and $h(z, w) = h(w, z)$.
Let u be the solution to the Dirichlet problem, $u = 0$ on $b\Omega$, $(\partial\bar{\partial}u)^2 = h\beta_2$ in Ω, u plurisubharmonic. Then u cannot belong to $C^2(\bar{\Omega})$ [3].
Moreover, in the case $\Omega = \mathbb{B}_0$, we have the best regularity that $u \in C^{1,1}(\bar{\Omega})$ in [2].

The following example illustrates that on domains of infinite type, we do not have any Hölder continuous (in the strong sense) plurisubharmonic solution of (3) on $\bar{\Omega}$.

Example 5.2 (Non-existence [20]) Let define

$$E = \{(z_1, z_2) \in \mathbb{C}^2 : |z_1|^2 + \exp(2 - |z_2|^{-2}) < 1\}.$$

Then, E is a bounded pseudoconvex domain in $\mathbb{C}^2$, with smooth boundary bE. Let

$$u(z) = |z_2|^2(1 - |z_1|^2)^k - \frac{10k}{2 - \log(1 - |z_1|^2)}.$$

Then, $u \in C(\overline{E}) \cap \mathcal{P}(E)$, smooth in E, and $u \notin C^\epsilon(\overline{E})$, for any $\epsilon > 0$.
Moreover, $u(z) = \phi(z) = |z_2|^2(1 - |z_1|^2)^k - 10k|z_2|^2 \in C^\infty(bE)$, and

$$\sqrt{\det H(u)} \in C^{\frac{k}{2}-1-\epsilon} \quad \text{for any } k \geq 4, \text{ and } \epsilon > 0.$$

Example 5.3 (On weakly pseudoconvex domains of infinite type [11]) We consider

$$E=\left\{(z_1,z_2)\in\mathbb{C}^2:\rho(z_1,z_2)=\exp(1-\frac{1}{|z_1|^s})+|z_2|^2<1\right\}, \tag{5}$$

where $0<s<1$. It is well-known that this ellipsoid satisfies the f-property with $f(t)=(1+\log(t))^{\frac{1}{s}}$ (see [12, 18] for details) and hence we define $g(t)=(1+\log(t))^{\frac{1}{s}-1}\approx\left(\int_t^\infty\frac{da}{af(a)}\right)^{-1}$. Then for any datum $\phi\in\Lambda^{t^\alpha}(bE)$, $0<\alpha\leq 1$, and $h\geq 0, h^{1/2}\in\Lambda^{g^\alpha}(\overline{E})$, the Dirichlet problem for the complex Monge-Ampère equation has its unique solution in $\Lambda^{g^\alpha}(\bar{E})$. We will prove that the index g^α can not be improved, i.e., there does not exist $\tilde{g}$ such that $\lim\limits_{t\to\infty}\frac{\tilde{g}(t)}{g(t)}=\infty$ and $u\in\Lambda^{\tilde{g}^\alpha}(\overline{E})$. Indeed, let

$$u(z)=(1-\log(1-|z_2|^2))^{-\frac{\alpha}{s}}\quad\text{for any}\quad z\in\overline{E}.$$

It is easy to check that u is plurisubharmonic and smooth in E. Then u is a solution of the following problem

$$\begin{cases}\det[u_{ij}]=0 & \text{in}\quad E,\\ u=|z_1|^\alpha & \text{in}\quad bE,\end{cases}$$

and hence $u\in\Lambda^{g^\alpha}(\bar{E})$ by Theorem 4.5. The following claim explains why the index function g can not be improved.

Claim: Assume $u\in\Lambda^{\tilde{g}^\alpha}(E)$, then $\lim\limits_{t\to\infty}\dfrac{\widetilde{g}(t)}{g(t)}<\infty$.

Proof of Claim. For small $\epsilon>0$, let $z_\epsilon=(0,1-\epsilon)$ and $w_\epsilon=(0,1-2\epsilon)$. Since $u\in\Lambda^{\tilde{g}^\alpha}(E)$, it follows

$$|u(z_\epsilon)-u(w_\epsilon)|\lesssim(\tilde{g}^\alpha)^{-1}(|z_\epsilon-w_\epsilon|^{-1})=(\tilde{g}(\epsilon^{-1}))^{-\alpha}. \tag{6}$$

By the basis inequality $|x^{-\alpha}-y^{-\alpha}|\geq|x-y|^{-\alpha}$ for any $x,y\in[0,\infty)$ and $\alpha>0$, we obtain

$$|u(z_\epsilon)-u(w_\epsilon)|=|(f(\epsilon^{-2}))^{-\alpha}-(f((2\epsilon)^{-2}))^{-\alpha}|\geq|f(\epsilon^{-2})-f((2\epsilon)^{-2})|^{-\alpha}. \tag{7}$$

On the other hand,

$$\begin{aligned}f(\epsilon^{-2})-f((2\epsilon)^{-2})&=\int_{(2\epsilon)^{-2}}^{\epsilon^{-2}}f'(t)dt\\&=\frac{1}{s}\int_{(2\epsilon)^{-2}}^{\epsilon^{-2}}\frac{g(t)}{t}dt\\&\leq\frac{1}{s}\frac{g(t)}{t}\Big|_{t=\epsilon^{-2}}\times\int_{(2\epsilon)^{-2}}^{\epsilon^{-2}}dt\\&\lesssim g(\epsilon^{-2})\ \approx\ g(\epsilon^{-1}),\end{aligned} \tag{8}$$

where the first inequality follows by $\dfrac{g(t)}{t} = \dfrac{(1+\log t)^{\frac{1}{s}-1}}{t}$ is decreasing in a neighborhood of infinity. From (6), (7) and (8) we get

$$\tilde{g}(\epsilon^{-1}) \lesssim g(\epsilon^{-1}) \quad \text{for any } \epsilon > 0.$$

This proves the claim and explains why the index function g can not be improved.

References

[1] Z. Blocki, *The complex Monge-Ampère operator in hyperconvex domains.* Ann. Scuola Norm. Sup. Pisa Cl. sci. 23 (1996), 721–747.

[2] E. Bedford and B. A. Taylor, *The Dirichlet Problem for a Complex Monge-Ampère Equation.* Inventiones math. 37 (1976), 1–44.

[3] E. Bedford and J. E. Fornaess, *Counterexamples to regularity for the complex Monge-Ampère equation.* Inventiones math. 50 (1979), 129–134.

[4] D. Catlin, *Necessary conditions for subellipticity of the $\bar{\partial}$-Neumann problem.* Ann. of Math. 117/1 (1983), 147–171.

[5] D. Catlin, *Subelliptic estimates for the $\overline{\partial}$-Neumann problem on pseudoconvex domains.* Ann. of Math. 126/1 (1987), 131–191.

[6] D. Catlin, *Estimates of invariant metrics on pseudoconvex domains of dimension two.* Math. Z. 200/3 (1989), 429–466.

[7] D. Coman, *Domains of finite and Hölder continuity of the Perron-Bremermann function.* Proc. Amer. Math. Soc. 125/12 (1997), 3569–3574.

[8] L. Caffarelli, J. J. Kohn, L. Nirenberg, and J. Spruck, *The Dirichlet problem for nonlinear second-order elliptic equations II. Complex Monge-Ampère equations and uniformly elliptic equations.* Comm. Pure Appl. Math. 38 (1985), 209–252.

[9] J. D'Angelo, "Several Complex Variables and the Geometry of Real Hypersurfaces". CRC Press, Boca Raton, 1993.

[10] J E. Fornaess and N. Sibony, *Construction of P.S.H. functions on weakly pseudoconvex domains.* Duke Math. J. 58/3 (1989), 633–655.

[11] L. K. Ha, T. V. Khanh, *Boundary regularity for the Complex Monge-Ampère equation on pseudoconvex domains of infinite type.* Preprint.

[12] Tran Vu Khanh, "A general method of weights in the $\bar{\partial}$-Neumann problem". Ph.D. thesis, in `arxiv:1001.5093v1`, 2010.

[13] T. V. Khanh, *Supnorm and f-Hölder estimates for $\bar{\partial}$ on convex domains of general type in $\mathbb{C}^2$.* J. Math. Anal. Appl. 403 (2013), 522–531.

[14] T. V. Khanh, *Boundary behavior of the Kobayashi metric near a point of infinite type.* `arXiv:1302.0789` (2013).

[15] S. Kobayashi, "Hyperbolic complex spaces". Grundlehren der Mathematischen Wissenschaften, Springer-Verlag, Berlin, 1998.

[16] S. Kolodziej, "The Complex Monge-Ampère Equation and Pluripotential Theory". Memoirs of the A.M.S., 2005.

[17] S. Krantz, "Function Theory of Several Complex Variables (2nd Edition)". Wadsworth, Belmont, California, 1992.

[18] T. V. Khanh and G. Zampieri, *Regularity of the $\overline{\partial}$-Neumann problem at point of infinite type.* J. Funct. Anal. 259/11 (2010), 2760–2775.

[19] T. V. Khanh and G. Zampieri, *Necessary geometric and analytic conditions for general estimates in the $\bar{\partial}$-Neumann problem.* Invent. Math. 188/3 (2012), 729–750.

[20] S.-Y. Li, *On the existence and regularity of Dirichlet problem for complex Monge-Ampere equations on weakly pseudoconvex domains.* Calc. Var. 20 (2004), 119-132.

[21] J. D. McNeal, *Local geometry of decoupled pseudoconvex domains.* In *Complex analysis (Wuppertal, 1991)*, Aspects Math., E17, 223–230. Vieweg, Braunschweig, 1991.

[22] J. D. McNeal, *Convex domains of finite type.* J. Funct. Anal. 108/2 (1992), 361–373.

[23] S. C. Chen, M. C. Shaw, "Partial Differential Equations in Several Complex Variables". Studies in Adv. Math., AMS Int. Press 19, 2001.

Option pricing in a defaultable model: a characteristic function approach

Stefano Pagliarani (*)

This paper is based on joint work with Agostino Capponi(†) and Tiziano Vargiolu(‡) .

Abstract. We consider a stock (i.e. a financial risky asset) whose predefault dynamics follows a stochastic differential equation driven by a Brownian motion. Then we introduce the concept of *default time* by means of a hazard rate intensity function given by a negative power of the stock price. We recover the characteristic function of the terminal log-price as the solution of a complex valued infinite dimensional system of first order ordinary differential equations. In particular, we provide an explicit eigenfunction expansion for the characteristic function in a suitably chosen Banach space, and use it to price defaultable bonds and stock options by means of standard Fourier inversion techniques.

1 Introduction to option pricing

Consider a stock (a financial asset) whose price is described by a non-negative stochastic process $S = (S_t)_{t\geq 0}$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A financial derivative of European type (European option) is a contract that pays you back a certain amount of money at fixed time maturity $T > 0$ depending on the terminal value of S_T. Precisely, a European option is characterized by a measurable function $h : \mathbb{R}_0^+ \to \mathbb{R}$ called payoff function, and its value at time T is given by $h(S_T)$.

More generally, there exist financial contract whose payoff h at time $T > 0$ depends on the whole path $(S_t)_{0\leq t\leq T}$ of the underlying stock. Nevertheless, throughout this report we will only consider the case of terminal payoffs (only dependent on the terminal value S_T). Next we list a couple of example of financial derivatives.

Example 1.1 Call (put) option: this contract gives you the right to buy (sell) the stock S, in a future moment $T > 0$, at a fixed price $K > 0$ called strike. The payoff for

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: stefanop@math.unipd.it . Seminar held on November 6th, 2013.

(†) Department of Applied Mathematics and Statistics, Johns Hopkins University, 21218, Baltimore, MD.

(‡) Dipartimento di Matematica, Università di Padova, Padova, Italy.

such an option is then:

$$h(s) = (s - K)_+, \qquad (h(s) = (K - s)_+).$$

Example 1.2 Defaultable Bond: the payoff of this contract is

$$h(s) = \begin{cases} 1, & s > 0, \\ 0, & s = 0. \end{cases}$$

Here, S_t can be viewed as the reference asset value of a company, and the bond h can be viewed as a 1\$ debit issued by the company, which will be always repaid at time T, unless $S_T = 0$ (i.e. default of the company).

In order to face the problem of determining a fair price for a European option h, we first need to assign a model for the evolution of the underlying asset price S_t. We assume here that $S_t > 0$ for any $t > 0$, and denote by X the log-price process, i.e. $X_t = \log(S_t)$, then the dynamics of X is typically described by a *stochastic differential equation* (SDE) of the type

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t, \quad X_0 = x,$$

or equivalently

$$X_t = x + \int_0^t \mu(s, X_s)ds + \int_0^t \sigma(s, X_s)dW_s.$$

Here, the function μ and σ are called *drift* and *volatility* functions respectively. The reader who is not familiar with stochastic integration and stochastic calculus can refer to several textbooks where the theory of Brownian integration and Ito's calculus is presented (e.g. [6], [5]).

One of the most dominant concepts in the modern mathematical finance, and for sure the most important one in option pricing, is the concept of *arbitrage*. Roughly speaking, an arbitrage is *a financial operation that allows to realize a profit with positive probability with no risk of possible losses.*

Example 1.3 A simple non-arbitrage principle: *if $X_T = Y_T$ almost surely at some time $T > 0$, then $X_t = Y_t$ almost surely for any $t < T$.* Indeed, assume that $X_t > Y_t$ for some $t < T$. Then, one could realize an arbitrage by means of the so called *short selling*, i.e. the practice of selling securities or other financial instruments that are not currently owned, and subsequently repurchasing them ("covering"). In particular, at time t one could short sell X_t and buy Y_t, this way realizing a risk-free profit, given that the two prices at time T, X_T and Y_T, will be worth the same with probability 1.

The general *arbitrage theory* is quite complex. The reader can refer to [2] for an exhaustive presentation of the topic. Hereafter we will take as given the following result.

Proposition 1.4 *Let $u(t, x)$ be the non-arbitrage price at time t of an option with payoff h, given $X_t = x$. Then we have*

$$u(t, x) = \mathbb{E}_{\mathbb{Q}}\left[h(X_T) \middle| X_t = x\right]. \tag{1.1}$$

Here the expectation is taken with respect to a new probability measure $\mathbb{Q}$ which is called risk neutral measure, and the dynamics of X under $\mathbb{Q}$ becomes

$$dX_t = -\frac{\sigma^2(t, X_t)}{2}dt + \sigma(t, X_t)dW_t. \tag{1.2}$$

Note that, the pricing formula (1.1) is independent on the drift μ of the underlying process X in the original (real world) probability measure $\mathbb{P}$. Now, under mild assumptions on the function σ, it is possible to prove that u also solves the backward Cauchy problem

$$\begin{cases} \mathcal{L}\, u = (\partial_t + \mathcal{A})u = 0, \quad t < T, \\ u(T, x) = h(x), \end{cases} \tag{1.3}$$

where the elliptic operator

$$\mathcal{A} = \frac{\sigma^2(t, x)}{2}(\partial_{xx} - \partial_x)$$

is the generator of the process X under the risk-neutral probability measure $\mathbb{Q}$. Therefore, we have

$$u(t, x) = \int_{\mathbb{R}} p(t, x; T, y)h(y)dy, \qquad t < T, \qquad x \in \mathbb{R}, \tag{1.4}$$

where $p(t, x; T, y)$ can be viewed both as the fundamental solution of the operator $\mathcal{L}$ (analytical point of view), and as the transition density of the process X starting from (t, x) and ending at (T, y) (probabilistic point of view).

Example 1.5 the Black and Scholes model (1973). The dynamics of the log-price of the underlying asset is given by

$$dX_t = -\frac{\sigma^2}{2}dt + \sigma dW_t, \quad \sigma > 0.$$

The pricing operator $\mathcal{L}$ becomes

$$\mathcal{L} = \frac{\sigma^2}{2}(\partial_{xx} - \partial_x) + \partial_t, \qquad \text{(Heat operator)}$$

whose fundamental solution is the Gaussian density function

$$p(t, x; T, y) = \Gamma(t, x : T, y) = \frac{1}{\sqrt{2\pi(T-t)}\sigma}e^{-\frac{(x-y)^2}{2\sigma^2(T-t)}}$$

Thus the price $u(t, x) = \int_{\mathbb{R}} p(t, x; T, y)h(y)dy$ can be computed either explicitly, or by means of numerical integration.

For a general volatility function $\sigma(t, x)$, p is not explicitly known. Here comes the necessity to develop indirect alternative methods to compute the conditional expectation (1.1) (or the solution of the PDE (1.3) from the analytical point of view).

2 Fourier methods in option pricing

Fourier methods are rather used in finance in order to compute expectations such as the one in (1.1), and are based on the the knowledge of the characteristic function of the underlying process X (equivalently the Fourier transform of p), typically defined as

$$\begin{cases} \hat{p}(t,x;T,\xi) := \mathcal{F}(p(t,x;T,\cdot))(\xi) = \int_{\mathbb{R}} e^{i\xi y} p(t,x;T,y)dy = \mathbb{E}\left[e^{i\xi X_T^{t,x}}\right] =: \varphi_T^{t,x}(\xi), \\ 0 \le t < T, \quad x, \xi \in \mathbb{R}. \end{cases} \tag{2.1}$$

Here, the suffix in $X_T^{t,x}$ indicates that the process starts at time $t < T$ from the initial point $x \in \mathbb{R}$. Furthermore, the two notations $\hat{p}(t,x;T,\xi)$ and $\varphi_T^{t,x}(\xi)$ refer respectively to the Fourier transform of $p(t,x,T,\cdot)$ and the characteristic function of $X_T^{t,x}$. The easier way to represent the pricing integral (1.4) in terms of the Fourier transform $\hat{p}$ is rather simple, and relies on the following

Lemma 2.1 *Let $f, \hat{f} \in L^1(\mathbb{R})$ with f real function. Then $\hat{f}(-\xi) = \overline{\hat{f}(\xi)}$ and thus $\xi \mapsto e^{-ix\xi}\hat{f}(\xi)$ is an even function. Then we have*

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-ix\xi}\hat{f}(\xi)d\xi = \frac{1}{\pi}\int_0^\infty e^{-ix\xi}\hat{f}(\xi)d\xi.$$

Assume now that $h, \hat{h} \in L^1(\mathbb{R})$. Then the pricing function $u(t,x)$ in (1.4) reads as

$$\begin{aligned} u(t,x) &= \int_{\mathbb{R}} h(y)p(t,x;T,y)dy = \int_{\mathbb{R}} h(y)\left(\tfrac{1}{\pi}\int_0^\infty e^{-iy\xi}\varphi_T^{t,x}(\xi)d\xi\right)dy \\ &= \tfrac{1}{\pi}\int_0^\infty \varphi_T^{t,x}(\xi)\int_{\mathbb{R}} h(y)e^{-iy\xi}dy\,d\xi = \tfrac{1}{\pi}\int_0^\infty \varphi_T^{t,x}(\xi)\hat{h}(-\xi)d\xi \end{aligned}$$

Unfortunately, in the practice, the payoff h is generally not in $L^1(\mathbb{R})$. This problem is typically overcome by properly dumping the payoff h, which is equivalent to translating the Fourier transform in the complex plane. We define the *dumped* payoff h_α as

$$h_\alpha := e^{-\alpha x}h(x), \qquad \alpha \in \mathbb{R}.$$

We have the following classical result that can be found on several textbooks, e.g. [6].

Theorem 2.2 *Assume there exist $\alpha \in \mathbb{R}$ such that*

(a) $h_\alpha, \hat{h}_\alpha \in L^1(\mathbb{R})$,

(b) $\mathbb{E}_{\mathbb{Q}}\left[e^{\alpha X_T^{t,x}}\right] < \infty$.

Then we have

$$u(t,x) = \frac{1}{\pi}\int_0^\infty \hat{h}(\xi + i\alpha)\,\varphi_T^{t,x}(-\xi - i\alpha)\,d\xi. \tag{2.2}$$

Note that, this simple inversion formula is only one of the several Fourier techniques existing in the literature of mathematical finance (see for instance the so called *Fourier-cosine series expansions (COS method)* by [4]).

3 Defaultable models: an eigenvalues expansion for the characteristic function

In a recent work by Capponi et al. [3], the authors proposed a defaultable model, where the predefault dynamics of the underlying asset follows an exponential Lévy process. After a suitable change of measure that lead the defaultable pricing back to the usual undefaultable case ($S_t > 0$), and carry out an eigenvalues expansion for the characteristic function of the underlying. Eventually they combined such an expansion with the Fourier pricing technique illustrated in the previous section to price European vulnerable derivatives. In this section we present a summary of the techniques proposed in [3] in a simplified version of the original model (purely diffusion case). The main purpose of this exposition is to give an overview and an intuition of the main ideas behind this approach. Therefore, in what follows we will prefer sometimes simplicity over rigor to maintain the text accessible to the reader who is not familiar with the mathematical tools of stochastic calculus.

We start by defining the default time τ of the company as the random time:

$$\tau = \inf\left\{ t \geq 0 : \int_0^t \lambda_u du \geq \zeta \right\},$$

where $(\lambda_t)_{t>0}$ is a positive stochastic process called *default intensity* and where ζ is an independent exponential random variable. Hereafter we drop the previous assumption that $S_t > 0$ in order to include the case $S_t = 0$, which correspond to the default of the firm to which the stock S refer. In particular, we set

$$S_t = \tilde{S}_t \cdot \mathbb{1}_{[0,\tau)}(t),$$

where $\tilde{S}_t$ represents now the predefault value of the reference asset of a company. The idea is now to link the default intensity λ_u to $\tilde{S}_u$ so that: the higher the asset value the lower the default intensity, or also, in the other way around, the lower the asset value the higher the default intensity. In particular, we set

$$\lambda_t = \tilde{S}_t^{-p}, \qquad t \geq 0.$$

To conclude the definition of the model, we still need to specify the dynamics of the predefault asset price $\tilde{S}$. In analogy to Section 1, we do this by defining the predefault log-prices $X_t := \log \tilde{S}_t$, with X as in (1.2) (Black and Scholes).

Consider now a European option with terminal payoff $h(S_T)$, with

$$h(s) = \begin{cases} \tilde{h}(s), & s > 0 \\ R, & s = 0. \end{cases}$$

Here, the function $\tilde{h}$ is called promised payoff, whereas R is called recovery payoff, i.e. the amount of money obtained at maturity if the default of the company as already occurred ($S_T = 0$). A typical example is the defaultable bond presented in Example 1.2. In analogy to Proposition 1.4, by the non-arbitrage theory and other theoretical results (see the Key Lemma in [1]), we have the following

Proposition 3.1 *Let $u(t,x)$ be the non-arbitrage price at time t of an option with payoff h, given $X_t = x$. Then we have*

$$u(t,x) = \mathbb{E}_{\mathbb{Q}}\left[e^{-\int_t^T e^{-pX_u}du}\, \tilde{h}(X_T) \middle| X_t = x \right], \tag{3.1}$$

where the above expectation is taken with respect to a risk neutral measure $\mathbb{Q}$, under which the dynamics of X reads as

$$dX_t = \left(e^{-pX_t} - \frac{\sigma^2}{2} \right) dt + \sigma dW_t.$$

Note that the payoff in expectation (3.1) contains a path-dependent term, $e^{-\int_t^T e^{-pX_u}du}$. The next lemma transforms the expectation (3.1) into another one where the payoff only depends on the terminal value X_T. From now on we will consider for simplicity $t = 0$.

Lemma 3.2 [Change of measure] *For any $T > 0$ and $x \in \mathbb{R}$ we have*

$$u(x) = u(0,x) = \mathbb{E}_{\tilde{\mathbb{Q}}}\left[e^{-X_T}\, \tilde{h}\left(e^{X_T}\right) | X_0 = x \right], \tag{3.2}$$

where $\tilde{\mathbb{Q}}$ is a new probability measure, equivalent to $\mathbb{Q}$, defined as

$$\frac{d\tilde{\mathbb{Q}}}{d\mathbb{Q}} = \exp\left(-\frac{\sigma^2}{2}t + \sigma W_t \right).$$

under which the dynamics of X_t becomes

$$dX_t = \left(e^{-pX_t} + \frac{\sigma^2}{2} \right) dt + \sigma d\tilde{W}_t.$$

Our goal is now to study the characteristic function of X under the new measure $\tilde{\mathbb{Q}}$ and then use the Fourier representation in Theorem 2.2 to compute the expectation (3.2). We set $\varphi(T,\xi) := \varphi_T^{0,x}(\xi)$, where $\varphi_T^{0,x}$ is the characteristic function of $X_T^{0,x}$ as defined in (2.1). We have the following fundamental result.

Lemma 3.3 *For any $\xi \in \mathbb{C}$ we have $|\varphi(T,\xi)| < \infty$, and*

$$\begin{cases} \frac{d}{dT}\varphi(T,\xi) = f(\xi)\varphi(T,\xi) + i\xi\,\varphi(T,\xi+i), \quad T > 0 \\ \varphi(0,\xi) = e^{i\xi x}, \end{cases}$$

where

$$f(\xi) = \frac{\sigma^2}{2}\left(ip\xi - p^2\xi^2 \right).$$

The key point of the previous result is that $\varphi(T,\xi)$ depends on $\varphi(T,\xi+i)$. In general: $\varphi(T,\xi+ni)$ depends on $\varphi(T,\xi+(n+1)i)$ for any $n \in \mathbb{N}_0$. We then set $\xi \in \mathbb{C}$ and define

$$\varphi_n(T) := \varphi(T,\xi+ni), \quad f_n := f(\xi+ni), \quad n \in \mathbb{N}_0.$$

Therefore, by the previous lemma we obtain that the functions φ_n satisfy the infinite-dimensional system of ordinary differential equations:

$$\begin{cases} \frac{d}{dT}\varphi_n(T) = f_n\varphi_n(T) + i(\xi + ni)\varphi_{n+1}(T), \quad T > 0, \\ \varphi_n(0) = e^{i(\xi+ni)x}. \end{cases} , \qquad n \in \mathbb{N}_0.$$

Equivalently, we can set the complex-valued sequence $\Phi_t := \big(\varphi_n(t)\big)_{n\geq 0}$, and obtain

$$\begin{cases} \frac{d}{dT}\Phi_T = A\,\Phi_T, \quad T > 0, \\ \Phi_0 = \big(e^{i(\xi+ni)x}\big)_{n\geq 0}, \end{cases} \tag{3.3}$$

where A is the infinite dimensional bi-diagonal matrix

$$A = (a_{n,n}, a_{n,n+1})_{n\in\mathbb{N}_0}, \qquad a_{n,n} = f_n, \quad a_{n,n+1} = i(\xi + ni). \tag{3.4}$$

The main idea now is to consider the infinite dimensional system (3.3) as an abstract Cauchy problem defined on a suitable Banach space, in order to study the spectrum of the linear operator A, and carry out an eigenvalues expansion for the solution Φ_T. More precisely, let us consider the measure μ on $\mathbb{N}_0$ given by

$$\mu(\{n\}) = e^{-f((n+1)i)\log(n+1)}, \qquad n \in \mathbb{N}_0,$$

and we denote by X the Banach space $L^2(\mathbb{N}_0, \mu)$, i.e. the space of the complex valued sequences $(u(n))_{n\in\mathbb{N}_0}$ endowed with the norm

$$\|u\|^2 = \sum_{n=0}^{\infty} e^{-f((n+1)i)\log(n+1)}|u(n)|^2 < \infty.$$

Proposition 3.4 *Let $A : \mathcal{D}(A) \to X$ be the bi-diagonal linear operator in (3.4), defined on its natural domain $\mathcal{D}(A) := \{u \in X | Au \in X\}$. Then, for any $T \geq 0$ we have $\Phi_T \in \mathcal{D}(A)$.*

$$\begin{cases} \frac{d}{dT}\Phi_T = A\,\Phi_T, \quad T > 0 \\ \Phi_0 = \big(e^{i(\xi+ni)x}\big)_{n\geq 0} \end{cases} , \qquad A = (a_{n,n}, a_{n,n+1})_{n\in\mathbb{N}_0}$$

The operator A is unfortunately neither bounded nor closed, and thus we cannot use the classical C_0-semigroup theory to characterize the semigroup e^{TA} determining the evolution of Φ_T. Nevertheless, we can still study its spectrum and find an eigenvalues expansion for Φ_T. We have the following result.

Theorem 3.5 *For any $T > 0$ and $\xi \in \mathbb{R}$, we have*

$$\varphi(T, \xi) = \Phi_T(0) = \sum_{n=0}^{\infty} e^{a_{n,n}T} \sum_{m=n}^{\infty} e^{i(\xi+im)x} b_{m,n}, \tag{3.5}$$

with

$$b_{m,n} = \left(\prod_{l=0}^{n-1} \frac{a_{l,l+1}}{a_{n,n} - a_{l,l}}\right)\left(\prod_{l=n}^{m-1} \frac{a_{l,l+1}}{a_{n,n} - a_{l+1,l+1}}\right). \tag{3.6}$$

We conclude with a practical example where we combine the above representation for the characteristic function of X_T, together with the Fourier inversion formula (2.2), in order to price a defaultable bond, as the one introduced in Example 1.2 ($h(S_T) = \mathbb{1}_{(0,\infty)}(S_T)$). Note that, for practical implementation, a truncation of the double series (3.5)-(3.6) is required. i.e.

$$\varphi(T,\xi) = \Phi_T(0) = \sum_{n=0}^{N} e^{a_{n,n}T} \sum_{m=n}^{M} e^{i(\xi+im)x} b_{m,n}$$

for some $N, M \in \mathbb{N}$. The results are reported in Table 1.

T	Price	$M = N$	Error bound
1/12	0.995	5	0.001
1/2	0.969	4	0.002
2	0.883	4	0.004

Table 1: 1-st column: time to maturity; 2-nd column: bond-price; 3-rd column: $M = N$ needed for the bond prices, computed by (3.5)-(3.6)-(2.2), to enter the 99% Monte-Carlo confidence band; 4-th column: error bound of the 99% Monte-Carlo confidence band. The parameters of the model were set to $p = 1, \sigma = 0.3$.

References

[1] Bielecki, T., and Rutkowski, M, "Credit Risk: Modelling, Valuation and Hedging". Springer, New York, NY, 2001.

[2] Björk, T., "Arbitrage Theory in Continuous Time". Oxford Finance Series, Hardback, August 2009.

[3] Capponi, A., and Pagliarani, S., and Vargiolu T., *Pricing vulnerable claims in a Lévy driven model*. To appear in Finance and Stochastics (2014).

[4] Fang, F., and Oosterlee, C. W., *A novel pricing method for European options based on Fourier-cosine series expansions*. SIAM Journal on Scientific Computing 31 (2008), 826–848.

[5] Karatzas, I., and Shreve S.E., "Brownian Motion and Stochastic Calculus". Springer-Verlag, New York, 1992.

[6] Pascucci, A., "PDE and Martingale Methods in Option Pricing". Vol. 2 of Bocconi & Springer Series, Springer, Milan, 2011.

Some applications of potential theory to Markov chains

Alessandra Bianchi (*)

Abstract. The link between potential theory and probability started in the last century with the work of Kakutani concerning the analysis of the Dirichlet problem. Since then, this connection has been explored by many authors and it has found applications in different contexts of probability. The purpose of this talk is to revisit the classical results of potential theory from a probabilistic point of view, giving a probabilistic meaning of the objects they contemplate and exploring the many possible applications of this approach in the study of Markov chains.

1 Dirichlet problem in $\mathbb{R}^d$

1.1 Harmonic functions

We start by giving some useful notation e definition. In what follows, $\mathcal{U}$ will denote an open subset of the d-dimensional vector space $\mathbb{R}^d$. For x in $\mathbb{R}^d$ we will write x_1, x_2, ..., x_d for its coordinates and we will use the notation $B_2(x,r)$ for the open Euclidean ball of centre x and radius $r > 0$.
For $f \in \mathcal{C}^1(\mathcal{U},\mathbb{R})$ we will denote by ∇f the gradient of f, that is

$$\nabla f : x \in \mathcal{U} \mapsto \nabla_x f := \left(\frac{\partial f}{\partial x_k}(x)\right)_{1\le k\le d} \in \mathbb{R}^d \tag{1.1}$$

while for $\phi \in \mathcal{C}^1(\mathcal{U},\mathbb{R}^d)$ we will denote by $\operatorname{div}\phi$ the divergence of ϕ, that is,

$$\operatorname{div}\phi : x \in \mathcal{U} \mapsto \operatorname{div}_x\phi := \sum_{k=1}^{d} \frac{\partial \phi_k}{\partial x_k}(x) \in \mathbb{R} \tag{1.2}$$

For $f \in \mathcal{C}^2(\mathcal{U},\mathbb{R})$ we define the *Laplacian* of f as

$$\Delta f : x \in \mathcal{U} \mapsto \Delta_x f := \sum_{k=1}^{d} \frac{\partial^2 f}{\partial x_k^2}(x) \in \mathbb{R} \tag{1.3}$$

(*)Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: bianchi@math.unipd.it. Seminar held on December 4th, 2013.

and notice $\Delta f = \operatorname{div}(\nabla f)$.

Definition 1.1 Given a function $f \in \mathcal{C}^2(\mathcal{U}, \mathbb{R})$, we say that f is *harmonic* on $\mathcal{U}$ if it satisfies the *Laplace equation on* $\mathcal{U}$

$$\forall x \in \mathcal{U},\ \Delta_x f = 0 \tag{1.4}$$

Examples With

$$r : (x_1, \dots, x_d) \in \mathbb{R}^d \mapsto \sqrt{x_1^2 + \cdots + x_d^2} \tag{1.5}$$

a harmonic function on $\mathbb{R}^d \setminus \{0\}$ is r if $d = 1$, $\ln r$ if $d = 2$ and r^{2-d} if $d \geq 3$.

In other words, f is harmonic on $\mathcal{U}$ if $\operatorname{div}_x(\nabla f) \equiv 0$ for all $x \in \mathcal{U}$, or equivalently, if ∇f is a null divergence field. By Stokes' lemma, this implies that for a harmonic function f on $\mathcal{U}$ and a closed oriented smooth surface $\partial\mathcal{V} \subset \mathcal{U}$ the flux of the vector field ∇f through $\partial\mathcal{V}$ is zero. As a first consequence, we get:

Proposition 1.2 (Mean-value property) *If f is harmonic on $\mathcal{U}$ then f satisfies the* mean-value property *(m.v.p.), that is:*

$$\forall r > 0,\ \forall x \in \mathcal{U},\ \overline{B_2(x,r)} \subset \mathcal{U} \Rightarrow f(x) = \int_{\partial B_2(x,r)} f \frac{d\sigma}{|\partial B_2(x,r))|} \tag{1.6}$$

where $|\partial B_2(x,r)|$ denotes the surface area of $\partial B_2(x,r)$.

The mean-value property characterizes harmonic functions and gives additional information on their regularity. It is not difficult to prove the following proposition.

Proposition 1.3 *If $f \in \mathcal{C}(\mathcal{U})$ has the m.v.p. then*

(i) *$f \in \mathcal{C}^\infty(\mathcal{U})$;*

(ii) *f is harmonic on $\mathcal{U}$.*

The mean-value property leads also to the following:

Proposition 1.4 (Maximum principle) *If f is harmonic on $\mathcal{U}$ then, for all compact sets $K \subset \overline{\mathcal{U}}$ such that f can be extended by continuity on K, $f|_K$ reaches its maximum (and its minimum) on ∂K.*

The maximum principle opens the door to uniqueness properties of the solution of the Dirichlet problem.

Definition 1.5 (Dirichlet problem) Given $g \in \mathcal{C}(\partial\mathcal{U}, \mathbb{R})$ we say that f is a solution of the Dirichlet problem on $\mathcal{U}$ with boundary condition g if f in $\mathcal{C}^2(\mathcal{U})$ and $\mathcal{C}(\overline{\mathcal{U}})$ satisfies the Laplace equation on $\mathcal{U}$ and coincides with g on $\partial\mathcal{U}$.

Proving the existence of a solution of a Dirichlet problem turns out to be a rather difficult task when one stay inside the framework of plain functional analysis. It is time to turn to Markov processes.

1.2 Solutions by probabilistic methods

For simplicity we will now assume that $\mathcal{U}$ is a bounded open domain. For extensions and generalizations to unbounded domains of the results presented here we refer to [6, Section 4.2]. We denote by P_x the law of a d-dimensional Brownian motion W starting from $x \in \mathbb{R}^d$ and by τ_A the hitting time of any set A:

$$\tau_A = \inf \{t \geq 0 : \ W(t) \in A\} \tag{1.7}$$

Kakutani's idea [5] was to present the candidate

$$h : x \in \overline{\mathcal{U}} \mapsto E_x \left[g \left(W(\tau_{\partial \mathcal{U}}) \right) \right] \tag{1.8}$$

as solution of the Dirichlet problem on $\mathcal{U}$ with boundary condition g. Since $\overline{\mathcal{U}}$ is a compact set we have, for all x in $\overline{\mathcal{U}}$,

$$P_x(\tau_{\partial \mathcal{U}} < +\infty) = 1 \tag{1.9}$$

so that h is well defined. We clearly have $h|_{\partial \mathcal{U}} = g$ and h has the m.v.p. Indeed, for any $x \in \mathcal{U}$ and $r > 0$ such that $\overline{B_2(x,r)} \subset \mathcal{U}$, we have, by the strong Markov property at time $\tau_{\partial B_2(x,r)}$ and using radial symmetry:

$$h(x) = E_x \left[g \left(W(\tau_{\partial \mathcal{U}}) \right) \right] \tag{1.10}$$

$$= E_x \left[E_x \left[g \left(W(\tau_{\partial \mathcal{U}}) \right) \Big| W(\tau_{\partial B_2(x,r)}) \right] \right] \tag{1.11}$$

$$= \int_{y \in \partial B_2(x,r)} E_y \left[g \left(W(\tau_{\partial \mathcal{U}}) \right) \right] dP_x(W(\tau_{\partial B_2(x,r)}) = y) \tag{1.12}$$

$$= \int_{\partial B_2(x,r)} h(y) \frac{d\sigma(y)}{|\partial B_2(x,r)|} \tag{1.13}$$

As a consequence, h is harmonic on $\mathcal{U}$ and to get a solution of the Dirichlet problem one has only to check the continuity of h on $\overline{\mathcal{U}}$. This question is intimately linked to the notion of *regularity*.

Definition 1.6 For any set A we define

$$\tau_A^+ := \inf \{t > 0 : \ W(t) \in A\} \tag{1.14}$$

and we say that $\mathcal{U}$ *has a regular border* when

$$\forall a \in \partial \mathcal{U}, \ P_a \left(\tau_{\mathcal{U}^c}^+ = 0 \right) = 1 \tag{1.15}$$

Proposition 1.7 *A bounded open domain $\mathcal{U}$ has a regular border if and only if, for all g in $\mathcal{C}(\partial \mathcal{U})$ the function h defined in (1.8) is continuous on $\overline{\mathcal{U}}$,i.e., is solution of the associated Dirichlet problem.*

We refer to [6] for the proof (of a stronger result).

One implication of the above proposition is that, under the hypothesis of regularity of the set $\mathcal{U}$, the Dirichlet problem has (unique) solution h. When the regularity hypothesis is not satisfied, we may wonder if there are solutions different from h. The answer is no, as it is stated by the following proposition:

Proposition 1.8 *If the Dirichlet problem on a bounded open domain $\mathcal{U}$ with boundary condition g has a solution f, then f coincides with h defined in (1.8).*

The content of the above propositions shows an interplay between analytic and probabilistic quantities related to potential theory. We clarify this interplay with the following example, which provides an easy application of the potential approach to the study of properties of the Brownian motion.

1.3 Recurrence and transience of the Brownian motion

In dimension two and for $b > a > 0$, consider the Dirichlet problem on

$$\mathcal{U} = B_2(0,b) \setminus \overline{B_2(0,a)} \tag{1.16}$$

with boundary conditions 1 on $\partial B_2(0,a)$ and 0 on $\partial B_2(0,b)$.
Since $\ln r$ is harmonic in open domains of $\mathbb{R}^2$ (see Eq. (1.5)), one can easily check that

$$f = \frac{\ln b - \ln r}{\ln b - \ln a}\,, \tag{1.17}$$

is a solution of the Dirichlet problem. By the maximum principle, being $\overline{\mathcal{U}}$ a compact set, it turns out that f is indeed the unique solution.

On the other hand $\mathcal{U}$ has a regular border. Indeed a Brownian motion that starts from x in $\partial\mathcal{U}$ crosses $\partial\mathcal{U}$ infinitely many times during any time interval $[0;t]$ with $t > 0$. As a consequence of the Prop 1.7, the function h defined in (1.8) is solution of the problem and coincides with f, that reads

$$\forall x \in \mathbb{R}^2,\ a \le r(x) \le b \Rightarrow P_x\left(\tau_{\partial B_2(0,a)} < \tau_{\partial B_2(0,b)}\right) = \frac{\ln b - \ln r(x)}{\ln b - \ln a} \tag{1.18}$$

We then obtained an explicit formula for the probability of exiting the region $\mathcal{U}$ from $\partial B_2(0,a)$ instead of $\partial B_2(0,b)$. Moreover, formula (1.18) gives the recurrence of the two-dimensional Brownian motion: send b to infinity to get

$$\forall x \notin B_2(0,a),\ P_x(\tau_{\partial B_2(0,a)} < +\infty) = 1\,. \tag{1.19}$$

The analogous study in dimension $d \ge 3$, gives

$$\forall x \in \mathbb{R}^d,\ a \le r(x) \le b \Rightarrow P_x\left(\tau_{\partial B_2(0,a)} < \tau_{\partial B_2(0,b)}\right) = \frac{r(x)^{2-d} - b^{2-d}}{a^{2-d} - b^{2-d}} \tag{1.20}$$

and then, sending b to infinity, the transience of the Brownian motion:

$$\forall x \notin \overline{B_2(0,a)},\ P_x(\tau_{\partial B_2(0,a)} < +\infty) = \left(\frac{r}{a}\right)^{2-d} < 1\,. \tag{1.21}$$

Consider now the Dirichlet problem on the punctured ball

$$\mathcal{U} = B_2(0,b) \setminus \{0\}\,. \tag{1.22}$$

Our candidate solution h, defined in (1.8), is obtained by sending a to 0 in (1.18), which gives

$$\forall x \in \overline{B_2(0,b)} \setminus \{0\},\ P_x\left(\tau_{\{0\}} < \tau_{\partial B_2(0,b)}\right) = 0 \tag{1.23}$$

The function we get, $h = \mathbb{1}_{\{0\}}$, is not continuous and from Prop. 1.8 we can conclude that the Dirichlet problem on $\mathcal{U}$ has not solution.

2 Generalization to Markov chains

2.1 Discrete Laplacian and simple random walk

Going from $\mathbb{R}^d$ to $\mathbb{Z}^d$ we lose the differential tool: derivatives have to be replaced by their discrete version and one has to find correspondence with the ∇ and Δ operators. Without entering to much into the details, for a real valued function f on $\mathbb{Z}^d$ we define the discrete Laplacian of f by

$$\Delta f : x \in \mathbb{Z}^d \mapsto \Delta_x f := \sum_{\substack{y \in \mathbb{Z}^d \\ x \sim y}} (f(y) - f(x))\,, \tag{2.1}$$

where we write $x \sim y$ for nearest-neighbor points in $\mathbb{Z}^d$, that is points x, y with Hamming distance $= 1$. Note that this is coherent with second order Taylor developments: for f in $\mathcal{C}^2(\mathbb{R}^d)$ and a unitary vector u

$$\frac{\partial^2 f}{\partial u^2}(x) = \lim_{\substack{h \to 0 \\ h \in \mathbb{R}}} \frac{f(x+hu) + f(x-hu) - 2f(x)}{h^2}\,. \tag{2.2}$$

For $\mathcal{U} \subset \mathbb{Z}^d$, the external border of $\mathcal{U}$ is

$$\partial_+ \mathcal{U} := \left\{ y \in \mathbb{Z}^d \setminus \mathcal{U} :\ \exists x \in \mathcal{U} \text{ with } x \sim y \right\}\,. \tag{2.3}$$

A function f defined on $\mathcal{U} \cup \partial_+ \mathcal{U}$ is *harmonic* on $\mathcal{U}$ if

$$\forall x \in \mathcal{U},\ \Delta_x f = 0 \tag{2.4}$$

Observe that (2.4) expresses a local mean-value property: it is equivalent to

$$\forall x \in \mathcal{U},\ \frac{1}{2d}\Delta_x f = \left(\frac{1}{2d} \sum_{d_1(x,y)=1} f(y) \right) - f(x) = 0 \tag{2.5}$$

The set of harmonic functions on $\mathbb{Z}^d$ is the kernel of the generator of the (continuous time) simple random walk, $\frac{1}{2d}\Delta$, just like the set of harmonic functions on $\mathbb{R}^d$ was the kernel of the generator of Brownian motion, $\frac{1}{2}\Delta$.

Like in the continuous case the mean-value property of harmonic functions gives a maximum principle (for which the notion of compactness is replaced by that of finiteness), and this maximum principle can be used to show uniqueness properties for the solutions of Dirichlet problems.

Given $\mathcal{U} \subset \mathbb{Z}^d$ and g a real valued function on its external border, we say that f is a solution of the Dirichlet problem on $\mathcal{U}$ with boundary condition g if f is harmonic on $\mathcal{U}$ and f coincides with g on $\partial_+\mathcal{U}$. Just like we used Brownian motion to prove the existence of a solution for some Dirichlet problem, we can do the same with simple random walks on $\mathbb{Z}^d$. For example:

Proposition 2.1 *For any finite subset $\mathcal{U}$ of $\mathbb{Z}^d$ and any real valued function g on $\partial_+\mathcal{U}$, there is a unique solution of the Dirichlet problem on $\mathcal{U}$ with boundary condition g. This solution is the function h defined by*

$$\forall x \in \mathcal{U} \cup \partial_+\mathcal{U},\ h(x) := E_x\Big[g\left(\zeta(\tau_{\mathcal{U}^c})\right)\Big] \tag{2.6}$$

where E_x stands for the expectation under the law of a simple random walk ζ that start from x.

Remark If $\mathcal{U}$ is not a finite set, then the above statement has to be reformulated among the set of bounded solutions of the Dirichlet problem.

2.2 Electrical networks and Markov chains

An *electrical network* is a connected undirected weighted graph with positive weights, with no more than one edge between any pair of vertices and with finite total weight on each vertex. More formally it is a pair $(\mathcal{X}, c)$ with $\mathcal{X}$ a countable set and c a real valued non-negative symmetric function on $\mathcal{X} \times \mathcal{X}$ such that

$$\forall x \in \mathcal{X},\ \mu(x) := \sum_{y \in \mathcal{X}} c(x,y) < +\infty \tag{2.7}$$

and such that, for all distinct x and y in $\mathcal{X}$, there exist $x = z_1, z_2, \ldots, z_n = y$ in $\mathcal{X}$ with

$$\forall k \in \{1; \ldots; n-1\},\ c(z_k, z_{k+1}) > 0 \tag{2.8}$$

We call *nodes* the elements of $\mathcal{X}$, we say that two nodes x and y are connected when $c(x,y) > 0$ and we call *edges* the elements of $\mathcal{E}$, defined as the set of ordered pairs of connected nodes:

$$\mathcal{E} := \{(x,y) \in \mathcal{X} \times \mathcal{X} :\ c(x,y) > 0\} \tag{2.9}$$

The quantity $c(x,y)$ is called *conductance* between two nodes x and y while its inverse

$$r(x,y) := \frac{1}{c(x,y)} \in]0; +\infty] \tag{2.10}$$

is the *resistance* between x and y.

We call *potential* any real valued function on $\mathcal{X}$. If we impose a potential $g(x)$ on each node x outside a subset $\mathcal{U}$ of $\mathcal{X}$, an *equilibrium potential* V associated with the constraint

$$\forall x \in \mathcal{U}^c,\ V(x) = g(x) \tag{2.11}$$

has to satisfy Ohm's and Kirchoff's laws.

Ohm's law: The *current* i associated with V is

$$i : (x,y) \in \mathcal{E} \mapsto i(x,y) = \frac{V(x) - V(y)}{r(x,y)} \tag{2.12}$$

Kirchoff's law: For all x in $\mathcal{U}$

$$\sum_{\substack{y \in \mathcal{X} \\ (x,y) \in \mathcal{E}}} i(x,y) = 0 \tag{2.13}$$

In other words, defining the operator $\mathcal{L}$ acting on potential f as

$$\mathcal{L}f : x \in \mathcal{X} \mapsto \mathcal{L}_x f := \sum_{y \in \mathcal{X}} \frac{c(x,y)}{\mu(x)} (f(y) - f(x))\,, \tag{2.14}$$

the Kirchoff's law states

$$\forall x \in \mathcal{U},\ -\mathcal{L}_x V = 0\,. \tag{2.15}$$

The operator $\mathcal{L}$ is the generator of ξ, discrete time random walk on the network with transition probabilities

$$p(x,y) = \frac{c(x,y)}{\mu(x)} \qquad x, y \in \mathcal{X} \tag{2.16}$$

Notice that (2.15) expresses once again a local mean-value property and as a consequence one can deal with the question of existence and uniqueness of an equilibrium potential associated with $\mathcal{U}$ and g by using the maximum principle that follows from the mean value property and using Kakutani's solution. For example if $\mathcal{X}$ is finite there exists a unique equilibrium potential

$$V(x) = E_x \left[g(\xi(\tau_{\mathcal{U}^c})) \right], \quad x \in \mathcal{X} \tag{2.17}$$

Remark The Markov chain ξ we associated with $(\mathcal{X}, c)$ is ergodic and reversible with respect to the measure μ. Conversely, *any* reversible ergodic Markov chain ξ on $\mathcal{X}$ is the random walk associated with some electrical network on $\mathcal{X}$. If μ is a reversible measure and $p(.,.)$ gives the transition probabilities of ξ, we just define c through (2.16) to build the network corresponding for which the transition probabilities of the associated random walk are given by $p(.,.)$.

3 Applications

The reader who is interested in this topic, can find more details in [4], [1], [3], [9], [8], [6] and [7].

3.1 Capacities and related variational principles

Consider A and B subsets of $\mathcal{X}$ that satisfy

$$A \cap B = \emptyset \text{ and } \forall x \in \mathcal{X},\ P_x(\tau_{A\cup B} < +\infty) = 1 \tag{3.1}$$

with P the law of the random walk ξ associated with the network. Assuming that A and B are disjoint subsets of $\mathcal{X}$, condition (3.1) certainly holds when ξ is recurrent, or $\mathcal{U} := \mathcal{X} \setminus (A \cup B)$ is finite.

Consider a potential V such that $V_{|A} = 1$ and $V_{|B} = 0$. Condition (3.1) ensures that the Kakutani's solution of the Dirichlet problem on $\mathcal{U}$ is well defined and is given by

$$h_{A,B} : x \in \mathcal{X} \mapsto P_x(\tau_A < \tau_B) \tag{3.2}$$

The function $h_{A,B}$ is the *equilibrium potential* associated to the above boundary conditions and is the only bounded solution of the related Dirichlet problem.

Another quantity of interest, in the electrical network framework as well as in many probabilistic applications, is the *capacity* of a pair of disjoint subsets $A, B \subset \mathcal{X}$, given by

$$C_{A,B} := \sum_{a\in A} \mu(a) P_a(\tau_A^+ > \tau_B^+) = \sum_{b\in B} \mu(b) P_b(\tau_B^+ > \tau_A^+) \tag{3.3}$$

where, for any $S \subset \mathcal{X}$,

$$\tau_S^+ := \min\{n > 0 :\ \xi(n) \in S\}\ . \tag{3.4}$$

The capacity is related to the equilibrium potential $h_{A,B}$, through the following equation

$$C_{A,B} = \mathcal{D}(h_{A,B}) \tag{3.5}$$

where, for any potential f

$$\mathcal{D}(f) := \frac{1}{2} \sum_{x,y\in\mathcal{X}} c(x,y)[f(x) - f(y)]^2 \tag{3.6}$$

Remark $\mathcal{D}(.)$ is the quadratic form associated with the bilinear *Dirichlet form* $\mathcal{D}(.,.)$. In the electrical network context, $\mathcal{D}(f)$ represents the energy dissipated per time unit.

In fact it turns out that the capacity satisfies the following variational principle:

Proposition 3.1 (Dirichlet's principle)

$$C_{A,B} = \min\{\mathcal{D}(f) :\ f|_A \equiv 1, f|_B \equiv 0\} \tag{3.7}$$

and this minimum is reached in $h_{A,B}$ only.

In fact, it is also possible to provide a variational principle that characterizes the capacity as a maximum over a suitable set of functions (*Thomson's principle*). Without entering into the details, let's just stress the fact that, altogether, these two variational principles are powerful tools to control capacities between sets. To clarify the importance of having a control capacities, in the next subsections we provide a few applications to the study of Markov chains.

3.2 Application to recurrence

Let ξ be a reversible ergodic Markov chain on $\mathcal{X}$, and $(\mathcal{X}, c)$ an associated electrical network. The random walk is *recurrent* if

$$\forall x, y \in \mathcal{X},\ P_x(\tau_y < +\infty) = 1 \tag{3.8}$$

otherwise it is *transient*. If $\mathcal{X}$ is finite ξ is necessarily recurrent, while in general one has to check property (3.8) or some equivalent condition. It turns out that recurrence is related to capacities through the following proposition:

Proposition 3.2 *The Markov chain ξ is recurrent if and only if there exists an increasing sequence of finite connected subsets K_n such that $\mathcal{X} = \cup_n K_n$ and*

$$\lim_{n\to+\infty} C_{a,K_n^c} = 0 \tag{3.9}$$

for some $a \in \mathcal{X}$. (See [4] for a more general statement.)

The following example, taken from [4], shows the powerful of the above result.

Example For the simple random walk on $\mathbb{Z}^2$ the conductance of each edge is $1/4$. We set, for all $n \geq 1$,

$$K_n := [-(n-1); n-1]^2 \tag{3.10}$$

and we consider the potentials

$$\begin{array}{rcccl} f_n & : & \mathbb{Z}^2 & \longrightarrow & [0;1] \\ & & x & \longmapsto & \begin{cases} 1 - \frac{\ln(1+\|x\|_\infty)}{\ln(1+n)} & \text{if } x \in K_n \\ 0 & \text{if } x \in K_n^c \end{cases} \end{array} \tag{3.11}$$

We have

$$\mathcal{D}(f_n) = \frac{1}{\ln^2(1+n)} \sum_{k=1}^{n} \frac{(8k-8)\vee 1}{4} \left[\ln(k+1) - \ln k\right]^2 \tag{3.12}$$

$$\leq \frac{1}{\ln^2(1+n)} \sum_{k=1}^{n} 2k \frac{1}{k^2} \tag{3.13}$$

$$\leq 2\frac{1+\ln(n+1)}{\ln^2(1+n)}\,. \tag{3.14}$$

By Prop. 3.2 and Eq. (3.7), we conclude that the random walk is recurrent.

3.3 Application to convergence to equilibrium

We recall that for ξ Markov chain on a finite state space $\mathcal{X}$ with transition probability matrix M, ξ is reversible with respect to the probability measure μ if and only if P is a self-adjoint operator on $\ell^2(\mu)$. In this case P has only real eigenvalues

$$1 = \lambda_0 > \lambda_1 \geq \cdots \geq \lambda_{N-1} \geq -1\,. \tag{3.15}$$

The rate of convergence to equilibrium in $\ell^2(\mu)$ (see [10]) is then given by the *spectral gap*

$$\lambda := 1 - \lambda_1 \tag{3.16}$$

that is the smallest non-zero eigenvalue of $I - P = -\mathcal{L}$ (with $\mathcal{L}$ being the generator of the Markov chain). Equivalently, the spectral gap is characterized by the following variational principle

$$\lambda = \min_{\mathrm{Var}(f)\neq 0} \frac{\mathcal{D}(f)}{\mathrm{Var}(f)}\,. \tag{3.17}$$

Any test function f gives an upper bound on the spectral gap. Restricting the minimum to characteristic functions we get

$$\lambda \leq \min_{A\subset\mathcal{X}} \frac{\sum_{x\in A, y\in\partial_+ A} c(x,y)}{\mu(A)(1-\mu(A))} \leq 2 \min_{\mu(A)\leq\frac{1}{2}} \frac{C_{A,A^c}}{\mu(A)} \;=\; 2I \tag{3.18}$$

where the *isoperimetric constant* I is defined by the last equation. Actually I gives also a *lower* bound on λ:

Lemma (Cheeger):

$$\frac{I^2}{2} \leq \lambda \leq 2I \tag{3.19}$$

If instead of restricting the minimum to characteristic functions, that are particular cases of equilibrium potentials, we restrict the minimum to general equilibrium potential $h_{A,B}$, we get

$$\lambda \leq \min_{A\cap B=\emptyset} \frac{C_{A,B}}{\mathrm{Var}(h_{A,B})} \leq \min_{A\cap B=\emptyset} \frac{C_{A,B}}{\mu(A)\mu(B)}\,, \tag{3.20}$$

where we used $\mathrm{Var}(h_{A,B}) \geq \mu(A)\mu(B)$.

There are examples where the upper bounds obtained through capacities in fact provide the correct behavior of the spectral gap, and thus of the convergence to equilibrium of the Markov chain (see [4]).

References

[1] N. Berger, *Transience, recurrence and critical behavior for long-range percolation.* Commun. Math. Phys. 226 (2002), 531–558.

[2] J. L. Doob, "Classical potential theory and its probabilistic counterpart". Grundlehren der Mathematischen Wissenschaften, Springer-Verlag, New York, xxiv+846 pp, 1984.

[3] P. G. Doyle and J. Snell, "Random walks and electric networks". Carus Mathematical Monographs 22, Mathematical Association of America, Washington, DC, xiv+159 pp., 1984.

[4] A. Gaudillière, "Condenser physics applied to Markov chains. A brief introduction to potential theory". Lecture notes, available at `http://arxiv.org/abs/0901.3053` (2009).

[5] S. Kakutani, *Markov processes and the Dirichlet problem.* Proc. Jap. Acad.. 21 (1945), 227–233.

[6] I. Karatzas and S. E. Shreve, "Brownian motion and stochastic calculus". Springer-Verlag, New York, xxiv+470 pp., 1991.

[7] G. Lawler, "Intersections of random walks". Birkhäuser, Boston, MA, 219 pp., 1991.

[8] R. Lyons an Y. Peres, "Probability on Trees and Networks".
Available at `http://php.indiana.edu/~rdlyons/prbtree/prbtree.html`, 2014.

[9] J. Norris, "Markov Chains". Cambridge University Press, Cambridge, xvi+237 pp., 1998.

[10] L. Saloff-Coste, *Lectures on finite Markov chains.* Lecture Notes in Math. 1665, Springer, Berlin (1997), 301–413.

An introduction to Stochastic Ergodic Control

Marco Cirant (*)

This note is a brief introduction to stochastic control theory, with particular attention to ergodic problems. Such theoretical framework can be implemented to treat a wide variety of models coming from physics, economics and social sciences. The approach presented here, relying deeply on the so-called *dynamic programming principle*, was formulated in the fifties, and since then the theory has grown up hugely. We refer to [1, 2], which treat in much more details this kind of problems, but the literature is indeed very rich, crossing the fields of partial differential equation, numerical analysis and applied mathematics. We describe the main features of ergodic control, and then we move to Mean Field Games, a much more recent area of research of increasing interest, showing how it is connected to control theory and in which extent it is a possible generalization.

1 Statement of the problem

In control theory, an agent typically aims at minimizing some cost functional by controlling his own state. His goal is to compute a strategy that is optimal, in the sense that any other possible strategy increases the cost he pays. Mathematically, we denote the agent's state by X, which we suppose to be a point of the euclidean space $\mathbb{R}^d$. The state X evolves in time $t \geq 0$ according to a differential equation, and the agent can affect his own state through a control $\alpha_t \in \mathbb{R}^d$ (the subscript will denote explicitly the time dependance). Precisely, X is a solution of the *controlled stochastic differential equation*

$$X_t = x + \int_0^t [b(X_s) + \alpha_s]ds + \sqrt{2}B_t \tag{1}$$

The agent has an initial state x, then his state X is subject to some drift b depending on X itself, the control α implemented and a random noise B_t, which in this note will be a Brownian motion. We skip some technical details here, but point out that α should be a process adapted to B_t, meaning that the agent cannot predict the future evolution of the

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: cirant@math.unipd.it. Seminar held on January 15th, 2014.

Brownian motion. Moreover, some (mild) conditions should be guaranteed in order for X_t to be well-defined for all $t \geq 0$.

The second key ingredient of a control problem is the cost functional associated to the agent. In our setting, this can be formulated as

$$J(x, \alpha) = \liminf_{T \to \infty} \frac{1}{T} \mathbb{E} \int_0^T [L(\alpha_t) + f(X_t)]\, dt. \tag{2}$$

The function L is the cost paid for using the control α, f is the cost for being at position X_t; the overall cost is then averaged in the $[0, T]$ time interval and renormalized by dividing by T. Taking the limit as $T \to \infty$ means to consider the cost paid by the agent in the *long time* horizon: such type of cost functional is called *ergodic*. This operation has the effect of discarding phenomena that take place in short time periods. The operator $\mathbb{E}$ averages among all possible trajectories of the Brownian motion, so $J(x, \alpha)$ is just a real number depending on the initial state $X_0 = x$ and the control α which is implemented.

As an example, we might think of an agent which chooses his velocity α_t at every time t, paying $L(\alpha) = |\alpha|^2$ (in many real life applications, the effort of moving at speed α is quadratically proportional to $|\alpha|$); he moves around in $\mathbb{R}^d$, trying to stay as close as possible to minimum points of the potential f.

Our goal is now to find a way of constructing an optimal control, namely to find an α^* such that

$$J(x, \alpha^*) = \min_{\alpha} J(x, \alpha), \tag{3}$$

the minimum being taken in some suitable set of admissible (meaningful) controls.

1.1 Feedback controls

So far, the kind of controls α we considered depend on time, i.e. they assign a velocity α_t to every time $t \geq 0$. A more *useful* kind of controls, at least from the point of view of possible applications, are the so-called *feedback* strategies, where the chosen velocity depends only on the current state X_t. Suppose that we are given a measurable function $\bar{\alpha} : \mathbb{R}^d \to \mathbb{R}^d$; the solution of

$$X_t = x + \int_0^t [b(X_s) + \bar{\alpha}(X_s)] ds + \sqrt{2} B_t \tag{4}$$

enables us to define

$$\alpha_t := \bar{\alpha}(X_t).$$

Such a control, that we might indicate by $\bar{\alpha}_t$ with a slight abuse of notation, is said to be of feedback type; from the practical point of view, the agent who implements such kind of strategies, just observes his own state (and not the elapsed time t) and moves accordingly.

2 Ergodicity

Ergodic theory is devoted to the study of long time behavior of dynamical systems. At this stage, we are particularly interested in what happens to X_t solving (1) as $t \to \infty$, as our functional J is an averaged integral taken on long time intervals. We would like our trajectory to be stable as time passes by, not oscillating nor escaping at infinity. This "convergence" property is called *ergodicity* of X_t.

More precisely, denote with $m(y,t) = \mathbb{P}(X_t \in dy)$ the law of X_t, which captures how the process X_t is distributed on $\mathbb{R}^d$. The process is ergodic if

$$m(\cdot, t) \to \bar{m}(\cdot) \quad \text{locally uniformly}$$

for some continuous function $\bar{m}$ defined on $\mathbb{R}^d$ and for all initial states m_0. This tells us that no matter how the initial distribution of the agent is, after some time it "stabilizes", and the steady state is always the same.

We would like to find an optimal control (of feedback type) such that the corresponding system is ergodic. Obtaining sufficient conditions on $\bar{\alpha}$ such that X_t is ergodic has been a line of research over the last forty years. We refer to [3] for a presentation of the ideas and main results concerning the topic. As we need it in the sequel, we point out that the invariant measure of a process (if it exists), has a density that satisfies the *Kolmogorov* equation

$$\Delta m(x) - \operatorname{div}((b(x) + \alpha(x))m(x)) = 0 \quad \forall x \in \mathbb{R}^d. \tag{5}$$

3 A PDE approach

There are two main approaches for solving the minimization problem (3). The first one relies in the *Pontryagin's maximum principle*, which was formulated in 1956 by the Russian mathematician Pontryagin and his students. It provides a necessary condition for an optimum, in the sense that, if some control α realizes the minimum in (3), then it must solve a system of differential equations. The reader who is in interested the details of this principle may give a look at [9].

Another approach consists in finding an optimality condition which involves a partial differential equation. This method, introduced by Bellman and co-workers in the fifties, relies in breaking the optimization problem into simpler subproblems. Taking the limit of this procedure leads to an *Hamilton-Jacobi-Bellman* partial differential equation, whose solutions provide natural candidates for optimal controls of feedback type for (3).

Before writing down the HJB equation, we need to define the Hamiltonian function

$$H(x,p) := \sup_{a \in \mathbb{R}^d} \{-L(a) - a \cdot p\} - b(x).$$

Then, the basic formulation of the link between the optimization problem and the HJB equation is stated in the following way.

Theorem 3.1 *Suppose that $u \in C^2(\overline{\Omega}), \lambda \in \mathbb{R}$ solves*

$$-\Delta u(x) + H(x, Du(x)) + \lambda = f(x) \quad \forall x \in \mathbb{R}^d. \tag{HJB}$$

Then, the feedback α^* defined by

$$\alpha^*(x) = D_pH(-Du(x)) \quad \forall x \in \Omega \tag{6}$$

is optimal:

$$\lambda \leq J(x_0, \alpha) \quad \forall \alpha, x_0,$$
$$\lambda = J(x_0, \alpha^*) \quad \forall x_0.$$

We first observe that in (HJB) two unknowns appear: the function u and the constant $\lambda \in \mathbb{R}$. In the equation every datum of the problem is present: the Laplacian operator Δ is a consequence of the Brownian motion in (1), the Hamiltonian includes the dependance on b, L and the cost f appears on the right hand side.

The proof of this theorem is quite standard and the ideas involved can be found in any textbook on stochastic control theory. It relies on the Ito formula, which is used to expand functions of the stochastic process X_t, and the definition of the Hamiltonian function H.

Solutions are in general not unique, so there might be more than one optimal control; a pair u, λ provides at the same time a feedback strategy and the minimum value attained by the functional J. The existence of such a pair is non-trivial at all; the unbounded state space $\mathbb{R}^d$ turns out to be an additional difficulty, but sufficient conditions can be stated for the solvability of (HJB).

Another issue we have to deal with is that the feedback strategy synthesized using u can be *non-admissible*. We recall that a control should be such that the corresponding trajectory X_t is well-defined for all positive times. Moreover, we would like X_t to be ergodic, i.e. to have a "nice" long-time behavior. This requires additional assumptions on the data L, b, f.

In the literature, each point briefly presented here has been discussed and analyzed, and recently it has become of interest again. We refer to [10, 4] and references therein.

4 Mean Field Games

In the last section, we shortly present the theory of Mean Field Games, proposed independently by Lasry, Lions [5] and Caines, Huang and Malhame [6] to model and analyze complex decision processes involving a very large number of indistinguishable rational agents who have individually a very small influence on the system and are, on the other hand, influenced by the mass of the other agents. Mean Field Games systems are obtained by taking the limit of equilibria of games with N players as $N \to \infty$, under the symmetry assumption that players are indistinguishable. They capture, heuristically, equilibria of a population with a very large number of agents which aim at minimizing some common cost. We will not focus on this (very deep) derivation, but rather we will show the optimal control interpretation that can be given a-posteriori.

In a Mean Field Game, the cost every agent aim at minimize depends not only on his position x, but also on the distribution of other players. Suppose that such a distribution is denoted by m, then J has the same form of (2), but

$$f = f(x, m).$$

In an optimal situation, the agent implements an optimal control. As we saw before, he solves the (HJB) equation and implements the feedback strategy $\alpha(x) = -Du(x)$ (see (6)); we also mentioned that, supposing that his trajectory is ergodic, his invariant measure, or in other words his stationary distribution he converges to as $t \to \infty$, solves the Kolmogorov equation (5). This argument gives rise to a system of partial differential equation of the form

$$\text{(MFG)} \qquad \begin{cases} -\Delta u(x) + H(x, Du(x)) + \lambda = f(x, m(x)) \quad \forall x \in \mathbb{R}^d \\ \Delta m(x) - \operatorname{div}((b(x) - Du(x))m(x)) = 0, \end{cases}$$

where the coupling between the two equations is realized by the cost term $f(x, m)$. The idea is that the distribution of agents m is optimal with respect to a cost criterion which depends on the distribution itself. We are dealing with a problem that is not anymore strictly of optimal control type, but it is in some sense more general. The fixed-point structure of (MFG) suggests a strategy for finding solutions, that can be obtained by means of fixed-point theorems and a-priori estimates.

Mean Field Games theory is turning out to be a prolific area or research, involving challenging mathematical problems and interesting applications to many fields. It can be implemented to analyze systems with a large number of "thinking particles", and to observe how the individual behavior reflects into macroscopic phenomena. For a deeper introduction we refer to [7], and to [8] for applications.

References

[1] Bardi, Martino and Capuzzo-Dolcetta, I., "Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations". Systems & Control: Foundations & Applications, Birkhäuser Boston Inc., Boston, MA, 1997. (With appendices by M. Falcone and P. Soravia).

[2] Fleming, W. H. and Soner, H. M., "Controlled Markov processes and viscosity solutions". Stochastic Modelling and Applied Probability 25, Springer, New York, 2006.

[3] Khasminskii, R., "Stochastic stability of differential equations". Stochastic Modelling and Applied Probability 66, Springer, Heidelberg, 2012. (With contributions by G. N. Milstein and M. B. Nevelson).

[4] Arapostathis, A. and Borkar, V. S. and Ghosh, M. K., "Ergodic control of diffusion processes". Encyclopedia of Mathematics and its Applications 143, Cambridge University Press, Cambridge, 2012.

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

[6] Huang, M. and Malhamé, R. P. and Caines, P. E., *Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle.* Communications in Information and Systems 6/3 (2006), 221–251.

[7] Cardaliaguet, P., "Notes on Mean Field Games".
In `https://www.ceremade.dauphine.fr/~cardalia/MFG100629.pdf` .

[8] Gueant, O. and Lasry, J.-M. and Lions, P.-L., "Mean field games and applications".
In `http://www.oliviergueant.com/documents.html` .

[9] Fleming, W. H. and Rishel, R. W., "Deterministic and stochastic optimal control". Applications of Mathematics, No. 1, Springer-Verlag, Berlin-New York, 1975.

[10] Cirant, M., *On the solvability of some ergodic control problems in* $\mathbb{R}^d$. Submitted (2014).

Geometry over $\mathbb{F}_1$: the field with one element

Federico Bambozzi [(*)]

1 Motivations for the geometry over the field with one element

First of all let's make clear that there is no field with one element in the algebraic sense. Finite fields are classified with the following theorem:

Theorem 1.1 *Let K be a finite field, then there exists a prime number p and an integer $q = p^n$ such that $K \cong \mathbb{F}_q$.*

So when people talk about the field with one element don't talk about a field in the strict algebraic sense but they refer to a more subtle idea that we will try to explain in the following. We underline also that nowadays there is no common agreement on what $\mathbb{F}_1$ is and what the right theory of $\mathbb{F}_1$-geometry is. This is due to the fact that none of the proposed approaches have been proved able to lead to the desired results, even if there has been much work and progress in last decades. We can list some of the mathematicians linked to the most famous approaches: Deitmar [2]; Toen-Vaquie [10]; Durov [4]; Soulé [8]; Connes-Consani [1]; Haran [6].

In these notes we will describe the Dietmar approach which is one of the first in chronological order and also one of the more natural. Even if the proposed theories don't lead to equivalent categories of spaces over $\mathbb{F}_1$, nevertheless the notion of projective geometry over $\mathbb{F}_1$ agrees in all and is a common denominator of the theories. So we will start by introducing the classical problems that lead to the idea of projective geometry over the field with one element.

The idea of the field with one element was first hinted in an article of Tits [9] and was so revolutionary that was not taken seriously at that time. The aim of Tits was to find a geometric interpretation Chevalley groups analogous to that of classical Lie groups over $\mathbb{C}$. We will not enter in the details of the huge topic, we use the analogy of the simple example of $\mathrm{GL}^n(\mathbb{C})$ (and its subgroups) that can be seen as particular symmetries of $\mathbb{C}^n$. In particular we are interested in groups that can be realized as symmetries of projective spaces. So, let's start by recalling what a projective space is.

[(*)]Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: f.bambozzi@gmail.com. Seminar held on January 29th, 2014.

Definition 1.2 The n-dimensional *projective space* over a field K, denoted $\mathbb{P}^n(K)$, is the set of lines through the origin of K^{n+1}.

More precisely we can write

$$\mathbb{P}^n(K) = \frac{K^{n+1} - \{0\}}{\sim}$$

where $\sim$ is the equivalence relation

$$(x_0, \dots, x_n) \sim (y_0, \dots, y_n) \iff (y_0, \dots, y_n) = (\lambda x_0, \dots, \lambda x_n), \lambda \in K^\times,$$

where two vectors are identified if they generates the same line through the origin. If $K = \mathbb{C}, \mathbb{R}$ then K^{n+1} is a topological space and $\mathbb{P}^n(K)$ can be equipped with the quotient topology and the group

$$\mathrm{PGL}^n(K) = \frac{GL^{n+1}(K)}{\Delta_K}, \Delta_K \doteq \{ \text{ diagonal matrices}\}$$

acts on $\mathbb{P}^n(K) = \frac{K^{n+1}}{\sim}$ by the action on K^{n+1}, and this group coincides with the set of autocollineations of $\mathbb{P}^n(K)$ i. e. the geometric automorphism of $\mathbb{P}^n(K)$.

But projective geometries can also be introduced axiomatically. We will give these axioms for the case of finite projective geometries, i. e. geometries with a finite set of points, since it is the only relevant case for our discussion.

Definition 1.3 A *projective geometry of order q* is the data of

- a finite set P, whose elements are said *points*;
- a family $\mathcal{L} \subset \mathscr{P}(P)$, whose elements are said *sub-spaces*;
- a function $\dim : \mathcal{L} \to \mathbb{Z}_{\geq -1}$, called *dimension*;

satisfying the following axioms:

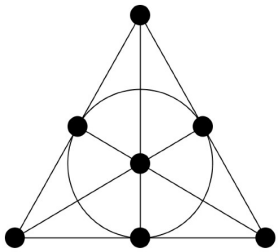
- the set $\mathcal{L}$ forms a lattice when partially ordered by inclusion;
- for $S, T \in \mathcal{L}$ and $S \subset T$, then $\dim(S) < \dim(T)$;
- the $\varnothing, P \in \mathcal{L}$:
- for all $x \in P$, the singleton $\{x\} \in \mathcal{L}$;
- $S \in \mathcal{L}$, $\dim(S) = -1 \iff S = \varnothing$ and $\dim(S) = 0 \iff S = \{x\}$;
- for $S, T \in \mathcal{L}$, $\dim(S) + \dim(T) = \dim(S \vee T) + \dim(S \wedge T)$;
- if $S \in \mathcal{L}$ and $\dim(S) = 1$, then $|S| = q + 1$;

the dimension of P is called the dimension of the geometry.

The previous set of axioms is not minimal, but seems, in our opinion, a very clear way to introduce the concept. Moreover, this axioms lead to a very strong classification of possible projective geometries.

Theorem 1.4 *Every projective geometry of order q, where q is a prime power, and satisfying the Desargues theorem is equivalent to $\mathbb{P}^n(\mathbb{F}_q)$ as constructed in the first definition.*

Hence, what we called degree of the projective geometry can be thought as the characteristic of the field over which the geometry is defined. The most famous example of finite projective geometry is the Fano plane, often drawn with the following diagram.



The Fano plane is isomorphic to $\mathbb{P}^2(\mathbb{F}_2)$, so it's a degree 2 geometry. In this picture the lines are identified by subsets of three points (because the geometry has degree 2) joint by straight lines, provided to think at the circle in the centre as a straight line. Finite (projective) geometries provide an environment where to test Tits conjectures and in fact to all of these geometries are associated some groups like $\mathrm{PGL}^n(\mathbb{F}_q), \mathrm{PSL}^n(\mathbb{F}_q), \ldots$. In the case of Fano plane one has that $\mathrm{PGL}^2(\mathbb{F}_2) = \mathrm{PSL}(3,2)$ is the finite simple group of order 168.

The crucial observation for the geometry over $\mathbb{F}_1$ is the following: the axiomatic definition of projective space make sense also for $q = 1$! So we can in this way, think to be able to define projective geometries over an hypothetical field with characteristic one, which we conjecturally can denote with $\mathbb{F}_1$. Moreover, there is the following classification result.

Theorem 1.5 *Any projective geometry of order 1 is equivalent to $\mathscr{P}(\{0,\ldots,n\})$, the boolean algebras of subsets of $\{0,\ldots,n\}$, for some $n \in \mathbb{N}$. Furthermore, in this geometry for any $S \in \mathscr{P}(\{0,\ldots,n\})$ one has $\dim S = |S| - 1$.*

So we will denote with $\mathbb{P}^n(\mathbb{F}_1) = \mathscr{P}(\{0,\ldots,n\})$ and call it the *projective space over* $\mathbb{F}_1$ of dimension n. And then, in Tits conjectures the group of automorphisms of $\mathbb{P}^n(\mathbb{F}_1)$ would be the symmetric group S_{n+1}, i. e. S_{n+1} would be the projective general linear group of dimension n and the alternating group A_{n+1} would be the analogous of the projective special linear group. Thus we are linking and trying to unify the combinatoric theory of symmetric groups and the theory of finite group of Lie type by adding new geometrical meaning to combinatorics.

There are other connections between combinatorics and the hypothetical geometry over the field with one element. For any $q \in \mathbb{N}$, the *Gaussian Binomial* coefficient is defined recursively by the formulas

$$\binom{n}{k}_q \doteq \binom{n-1}{k}_q + q^{n-k}\binom{n-1}{k-1}_q, \text{with } \binom{n}{n}_q = \binom{n}{0}_q = 1.$$

Then, for any $q \in \mathbb{N}$, we define the *q-analog* of $n \in \mathbb{N}$

$$[n]_q \doteq 1 + q + \ldots + q^{n-1}$$

and the *q-factorial* of $n \in \mathbb{N}$

$$[n]_q! \doteq [1]_q[2]_q \ldots [n]_q.$$

The following formula holds:

Proposition 1.6

$$\binom{n}{k}_q = \frac{[n]_q!}{[n-k]_q![k]_q!}.$$

And we have also the following remarkable proposition.

Proposition 1.7 *Let q be a prime power or $q = 1$ then the number of k-dimensional subspaces of $\mathbb{P}^n(\mathbb{F}_q)$ is*

$$\binom{n+1}{k+1}_q.$$

Thus, geometry over $\mathbb{F}_1$ hope give geometrical interpretation to (at least some) combinatorics formulae.

Finally, to end this motivational introduction we recall what is the most spectacular possible application of $\mathbb{F}_1$-geometry. In [3] Christopher Deninger proved that if a suitable category of motives exists, and he gave precise property that this theory must satisfies, then the proof of the Riemann hypothesis for zeta functions over finite fields given by Deligne, can be used to prove Riemann hypothesis through his theory of generalized determinants. Later Yuri Manin speculated that the "compactification" of $\operatorname{Spec}\mathbb{Z}$ needed by Deninger could be interpreted in the framework of $\mathbb{F}_1$-geometry. This gave a strong impetus to the development of $\mathbb{F}_1$-geometry several years ago which now seems faded away due to the little progress obtained in this direction.

2 Categories

In this section we review some material on the language of category theory as far as we need to explain some basic geometric ideas which are nicely expressed with category theory.

Definition 2.1 A *category* **C** is formed by the following data:

- a *class* (which can be a proper class) of objects $\mathrm{ob}(\mathbf{C})$;
- for any $X, Y \in \mathrm{ob}(\mathbf{C})$ a *set* $\mathrm{Hom}_{\mathbf{C}}(X, Y)$ of morphisms from X to Y;
- an associative composition law

$$\mathrm{Hom}_{\mathbf{C}}(Y, Z) \times \mathrm{Hom}_{\mathbf{C}}(X, Y) \to \mathrm{Hom}_{\mathbf{C}}(X, Z)$$

  for any $X, Y, Z \in \mathrm{ob}(\mathbf{C})$;
- for any $X \in \mathrm{ob}(X)$ a morphism $\mathrm{Id}_X \in \mathrm{Hom}_{\mathbf{C}}(X, X)$, called the *identity* of X.

Examples of categories are everywhere in mathematics. We give a very brief description of some natural examples.

Example 2.2

- The category of sets, **Sets** whose class of objects is the class of all sets and the sets of morphism are given by functions between sets.
- The category of rings, **Rings** whose class of objects is the class of all rings and the sets of morphism are given by ring homomorphism between ring.
- The category of topological spaces, **Top** whose class of objects is the class of all topological spaces and the sets of morphism are given by continuous maps.
- The category of open sets in a topological space X, $\mathbf{Ouv}(X)$ whose class of objects is the class of all open subsets $U \subset X$ and the sets of morphism are given by injections.
- Any partially ordered set can be seen as a category: given a poset E one can associate to it a category $\mathbf{E}$ whose class of objects is given by E and for morphism one define $\mathrm{Hom}_{\mathbf{E}}(i, j) = \{*\} \iff i \leq j$ and $\mathrm{Hom}_{\mathbf{E}}(i, j) = \varnothing$ otherwise.
- One can continue to add structure over sets and consider as morphisms functions which preserve all this structures: for example the category of topological vector spaces or locally convex topological vector spaces over a valued field with continuous linear maps is an example of a category endowed with both an algebraic and a topological structure.

The notion of morphism between categories is that of functor, which is something more than simply a function which associates to an object of a category an object of a second category, as the following definition shows.

Definition 2.3 A (covariant) *functor* $F : \mathbf{C} \to \mathbf{D}$ between two categories is an correspondence of the following type:

- for any $X \in \mathrm{ob}(\mathbf{C})$ an object $F(X) \in \mathrm{ob}(\mathbf{D})$;
- for any $f : X \to Y$ a morphism $F(f) : F(X) \to F(Y)$;

such that

- for any $\mathrm{Id}_X \in \mathrm{ob}(\mathbf{C})$, $F(\mathrm{Id}_X) = \mathrm{Id}_{F(X)}$;
- for any $X \xrightarrow{f} Y \xrightarrow{g} Z$, $F(g) \circ F(f) = F(g \circ f)$.

One can give the definition of composition of functors in a natural way, whose details are left to the reader. In this way one can also define the category of all categories whose morphisms are functor between category. To give a precise definition of this objects is quite complicated because one has to avoid paradoxes coming from circular references. Thus the precise definition of the category of all categories and the description of his rich structure is beyond the scope of this notes.

Anyway, functors are ubiquitous in mathematics.

Example 2.4

- The easiest example of functors are the forgetful functors as for example the followings: $U : \mathbf{Rings} \to \mathbf{Grp}$ which associate to any ring the underlying abelian group and to any morphism of rings the same map viewed as morphism of abelian groups or $V : \mathbf{Grp} \to \mathbf{Sets}$ which associates to any group the underlying set and to any group homomorphism the same function viewed as map of bare sets. This kind of functors are called forgetful because they are obtained simply by discarding some structures. Clearly composing V with U one obtain the forgetful $V \circ U : \mathbf{Rings} \to \mathbf{Sets}$.
- The forgetful functors have a companion which is the "free-object" functor, i. e. the functor $S : \mathbf{Sets} \to \mathbf{Grp}$ (and $T : \mathbf{Sets} \to \mathbf{Rings}$) which associates to the set X the free group (resp. the free ring, i. e. free $\mathbb{Z}$ algebra) over X.
- Functors often comes in pairs analogous the above pairs $S : \mathbf{Sets} \rightleftarrows \mathbf{Rings} : V \circ U$ called adjoint pairs. These pair are sources of rich structures.
- The association $X \mapsto \pi^n(X)$ where X is a pointed topological space and $\pi^(X)$ is the n-th homotopy group of X is functorial.
- In the same way $X \mapsto H^n(X)$ where X is a topological space, H a cohomology theory and H^n the n-th cohomology module of X is a functorial association.

The examples of categories we saw so far are all examples of *concrete categories*, i. e. categories that admit a forgetful (faithful) functor to the category of sets. These categories are particularly nice because they can be thought as categories of structured sets, i. e. categories whose objects are sets over which we fix an addition structure that we impose to be preserved by morphisms. In general there can exist, quite strange, categories which don't admit such a kind of functor a forgetful functor and cannot be thought as made of structured sets.

Example 2.5

- The most famous example of this phenomena is homotopy category of the category of topological spaces i. e. the category obtained from the category of topological space by considering has morphisms equivalence classes of homotopy equivalent continuous maps. Peter Freyd proved in [5] that this category is not concrete.
- An easy example is obtained by considering the class of sets with morphisms the relations between sets i. e. multivalued functions.
- More sophisticated examples of non-concrete categories are the derived category of an abelian category and the category of perverse sheaves which are very much used nowadays in geometry.

Definition 2.6 Let $\mathbf{C}$ be a category, a *sub-category* $\mathbf{D}$ of $\mathbf{C}$ is the data of

- a sub-class $\mathrm{ob}(\mathbf{D}) \subset \mathrm{ob}(\mathbf{C})$;
- for any $X, Y \in \mathrm{ob}(\mathbf{D})$ an inclusion of $\mathrm{Hom}_{\mathbf{D}}(X, Y) \subset \mathrm{Hom}_{\mathbf{C}}(X, Y)$;

such that this data forms a category with the restriction of the composition of $\mathbf{C}$ $\mathbf{D}$ is endowed with the composition inherited by $\mathbf{C}$.

Definition 2.7 A *full sub-category* $\mathbf{D}$ of $\mathbf{C}$ is a sub-category such that $\mathrm{Hom}_{\mathbf{D}}(X, Y) = \mathrm{Hom}_{\mathbf{C}}(X, Y)$ for all $X, Y \in \mathrm{ob}(\mathbf{D})$.

Therefore to give a full sub-category of $\mathbf{C}$ is enough to describe the class of its objects.

Example 2.8

- The sub-category of abelian groups is a full sub-category of **Grp**.
- The sub-category of Banach spaces in the category locally convex topological vector space is full which itself is full in the category of topological vector spaces.

The next is the last categorical notion that we need to recall.

Definition 2.9 Let $\mathbf{C}$ be a category. The *dual category* of $\mathbf{C}$ is the category $\mathbf{C}^\circ$ defined by

- $\mathrm{ob}(\mathbf{C}^\circ) \doteq \mathrm{ob}(\mathbf{C})$;
- $\mathrm{Hom}_{\mathbf{C}^\circ}(X, Y) \doteq \mathrm{Hom}_{\mathbf{C}}(Y, X)$;
- dual composition law: if $h = g \circ f$ in $\mathbf{C}$ then $h^\circ = f^\circ \circ g^\circ$ in $\mathbf{C}^\circ$.

If we represent morphisms of a category by directed arrows joining its objects, one says that $\mathbf{C}^\circ$ is obtained from $\mathbf{C}$ by "reversing the arrows". But one must be warned that this is purely formal procedure, it is often not clear what one get after applying this definition to a (also well known) category.

Example 2.10

- Dual category of an ordered set (thought as a category) is the opposite ordered set.
- Dual of **Sets** is not so clear a priori. It can be described, but not in a immediate way.
- If $\mathbf{C}$ is a concrete category then is seldom the case that $\mathbf{C}^\circ$ is a concrete category. In the next section we will see some important geometrical example where this is true. The "concretization" of the dual category of some important category, like the category of commutative rings, was a big advancement in the mathematics of last century and the concretization problems are still open for other important categories, like for example the category of all rings.

3 A recurrent theme in the construction of geometric spaces

It's recurrent theme in geometry to construct global (and so geometrically complicated) geometrical spaces by glueing (in some specified way) affine models whose geometry is better understood. For example this theme can be seen at work in the definition of the following geometrical spaces: topological manifolds, smooth manifolds, algebraic varieties, schemes, complex analytic spaces, rigid (or Berkovich) non-archimedean analytic spaces, and much more. In this section we give some details about the philosophy behind this classical theories.

We can give the following categorical-theoretic construct of topological manifolds: we can start from the following sub-category $\mathscr{M}$ of **Top**:

- objects of $\mathscr{M}$ are open subsets of $\mathbb{R}^n$;
- morphisms are continuous maps.

VarTop is the full sub-category of **Top** obtained by glueing objects of $\mathscr{M}$. There is a precise way to state the notion of glueing in categorical terms. This is obtained as a colimit of a diagram of maps, but we don't give here a precise definition of what this mean. So, this object, constructed as a colimit, solve a universal mapping problem in the category of topological spaces. The solution of this universal mapping theorem is equivalent to the gluing data defined by the charts (and maximal atlases) on a topological manifold. By requiring that morphisms to be C^n functions ($n = 0, 1, \ldots, \infty$) we get the theory of C^n-manifolds.

In other theories the category of local models has more structure. In classical algebraic geometry one take as affine models sets of the form

$$S = \{x \in \mathbb{C}^n | f_1(x) = 0, \ldots, f_r(x) = 0, f_i \in \mathbb{C}[X_1, \ldots, X_n]\}.$$

Let's call these *affine algebraic sets*.

Not all continuous maps are admissible maps between algebraic sets. Only polynomial functions are to be allowed since the aim is to do algebraic geometry. The problem of giving a better understanding to the category so obtained is solved in the following way. One has the association

$$S \mapsto \mathbb{C}[S] \doteq \frac{\mathbb{C}[X_1, \ldots, X_n]}{(f_1(x), \ldots, f_r(x))}$$

and $\mathbb{C}[S]$ is said *the coordinate ring of* S, and the following proposition.

Proposition 3.1 *To give a polynomial map $S \to T$ between affine algebraic sets is equivalent to give a ring homomorphism $\mathbb{C}[T] \to \mathbb{C}[S]$.*

Which easily leads to the following corollary:

Corollary 3.2 *The category of affine algebraic sets is (equivalent to) the dual category of finitely generated $\mathbb{C}$-algebras, denoted $\mathbf{Alg}^f_{\mathbb{C}}$.*

So, we can represent $(\mathbf{Alg}^f_{\mathbb{C}})^\circ$ as a (non-full) sub-category of **Top**. The objects so obtained are the building blocks of the theory of algebraic varieties over $\mathbb{C}$, that are defined to be appropriate gluing of algebraic sets. We also notice that in this situation is very difficult to have an insight and geometrical intuition on what these glueing should looks like. So the categorical way to define glueing now is crucial to obtain and understanding and a precise description because the calculation of colimits which is a standard task in category theory and can be done explicitly in many situations.

There is a more formal way to introduce affine models in classical algebraic geometry. By mean of the Hilbert's Nullstellensatz one can see

$$S \cong \mathrm{Max}\,(\mathbb{C}[S])$$

where $\mathrm{Max}\,(\mathbb{C}[S])$ is the set of maximal ideals of $\mathbb{C}[S]$ (on $\mathrm{Max}\,(\mathbb{C}[S])$ one can think to put Zariski topology, but we don't need to enter in these details). Grothendieck used this idea to enlarge the category of affine models: represent $(\mathbf{CommRings})^\circ$ as a (non-full) sub-category of **Top**. This category is called the category of *affine scheme*. The duality $\mathbf{CommRings} \to (\mathbf{CommRings})^\circ$ defined by Grothendieck is called the *spectrum functor*

$$A \mapsto \mathrm{Spec}\,(A) \doteq \{I \subset A | I \text{ is prime ideal}\}$$

and $\mathrm{Spec}\,(A)$ is equipped with the Zariski topology.

In the theory of complex analytic spaces, one starts form the category of open sets of $U \subset \mathbb{C}^n$ and $U \mapsto \mathcal{O}_X(U)$ ring of holomorphic functions and consider quotients by coherent ideals, thus obtaining a duality with *affine analytic sets*:

$$S \doteq \{x \in U \subset \mathbb{C}^n | f_1(x) = 0, \ldots, f_n(x) = 0, f_i \in \mathcal{O}_X(U)\}.$$

This is (more or less) the theory of Stein spaces and Stein algebras, which are linked by a duality analogous to the one seen above. Again the abstract methods of category theory are very helpful to define global analytic spaces as suitable glueing (i. e. suitable colimits) of affine analytic sets.

In non-archimedean analytic geometry the same theme is obtained considering the sub-category of rings identified by the quotients of the ring

$$K\langle X_1, \dots, X_n\rangle \doteq \left\{ \sum a_i X^i \in K[[X_1, \dots, X_n]] | |a_i| \to 0 \text{ for } |i| \to \infty \right\}.$$

called *affinoid algebras* over K (where K is field complete with respect to a non-archimedean absolute value). This category is denoted by $\mathbf{Aff}_K$ and $(\mathbf{Aff}_K)^\circ$ can be represented as a (non-full) sub-category of **Top**, in several different way:

$$A \mapsto \mathrm{Max}\,(A), A \mapsto \mathcal{M}(A), A \mapsto \mathcal{H}(A).$$

where $\mathrm{Max}\,(A)$ is the maximal spectrum of A equipped with the G-topology of Tate, $\mathcal{M}(A)$ is the Berkovich spectrum of A and $\mathcal{H}(A)$ is the Huber spectrum of A. Also in this case the notion of global analytic space is defined by suitable glueing of these models but the technical details of these constructions are more involved than the previous one.

Finally we want to remark that this philosophy is very useful but not a panacea: for example for non-commutative rings doesn't work (naively)! After many unfruitful attempts to extend Grothendieck functor from the category of commutative rings to the category of all rings, recently a negative result in this way was found in [7]:

Theorem 3.3 (Reyes 2011)
There is no functor which extend Spec *to the category of all rings and which preserve the fundamental properties of* Spec .

4 Deitmar approach to $\mathbb{F}_1$-geometry

At this point it's easy to explain Deitmar's approach to $\mathbb{F}_1$-algebraic geometry: it is simply to apply the recurrent theme explained in previous section to the category of commutative monoids. It is quite surprisying that a simple category like the category of monoids can carry so much geometrical informations, but we will see that in fact this is the case. So let's start by recall what are monoids.

Definition 4.1 A set M equipped with a binary operation $\cdot : M \times M \to M$ is said to be a *monoid* if

- the operation is associative;
- there is an identity.

Namely, the definition of a monoid is obtained by dropping the request of existence of inverses in the definition of a group.

Definition 4.2 A morphism $M \to N$ between monoids is a map of sets which preserve the operations.

From now on all monoids are commutative and **Mon** will denote the category of commutative monoids. Thus, **Mon** is a concrete category.

Definition 4.3 A subset $I \subset M$ is called *ideal* if

$$IM \doteq \{xm | x \in I, m \in M\} = I.$$

An ideal $I \subset M$ is called *prime* if $M - I$ is a submonoid of M.

Definition 4.4 Let M be a monoid, it's *spectrum over* $\mathbb{F}_1$ is defined

$$\mathrm{Spec}_{\mathbb{F}_1}(M) \doteq \{\mathfrak{p} \subset M | \mathfrak{p} \text{ is a prime ideal}\}.$$

$\mathrm{Spec}_{\mathbb{F}_1}(M)$ can be equipped with the Zariski topology whose definition is similar to the classical one.

We remark that with this definition, $\varnothing \in \mathrm{Spec}_{\mathbb{F}_1}(M)$ for all monoids, hence the spectrum is never empty.

Proposition 4.5 *$M \mapsto \mathrm{Spec}_{\mathbb{F}_1}(M)$ is functorial, i. e. any morphism of commutative monoids $M \to N$ gives a morphism $\mathrm{Spec}_{\mathbb{F}_1}(N) \to \mathrm{Spec}_{\mathbb{F}_1}(M)$.*

It follow easily that:

Corollary 4.6 *The category* **Mon**$^\circ$ *is equivalent to a (non-full) sub-category of* **Top**.

We call the objects of **Mon**$^\circ$, represented by mean of their spectra over $\mathbb{F}_1$, *affine schemes over* $\mathbb{F}_1$.

Remark 4.7 The properties of affine schemes over $\mathbb{F}_1$ are sometimes similar and sometimes different from classical affine schemes. The major difference is that any commutative monoid has a unique maximal idea, hence commutative monoids are more similar to local rings then to rings.

Finally we can define the field with one element.

Definition 4.8 We define

$$\mathbb{F}_1 = \{1\}$$

with its unique monoid structure. It follows that $\mathrm{Spec}_{\mathbb{F}_1}(\mathbb{F}_1) = \{\varnothing\}$.

Tits hinted to use as $\mathbb{F}_1$ the trivial ring $\overline{0} \doteq \{0 = 1\}$, but this seems to be not a good choice. For example the category of modules over $\overline{0}$ is trivial, i. e. it has only one object. Instead the category of modules over $\mathbb{F}_1$ is equivalent to the category of sets, a much richer and interesting object to study.

Let's use the following notation: for any monoid M and any set X, we define the monoid

$$M[X] \doteq \{mx_1^{d_1} \dots x_n^{d_n} | m \in M, x_i \in X, d_i \in \mathbb{N}, n \in \mathbb{N}\}$$

and call it the *free monoid* over M with base X.

The n-dimensional affine space over $\mathbb{F}_1$ is defined

$$\mathbb{A}^n_{\mathbb{F}_1} \doteq \mathrm{Spec}_{\mathbb{F}_1}(\mathbb{F}_1[X_1, \dots, X_n]).$$

The category of schemes of $\mathbb{F}_1$ is obtained by "glueing" affine scheme over $\mathbb{F}_1$, and also in this case the best way to give a precise meaning to this kind of operations is to use the abstract language of category theory. "Glueing" in the right way some copies of $\mathbb{A}^n_{\mathbb{F}_1}$ we obtain the $\mathbb{P}^n_{\mathbb{F}_1}$ of Tits, but now on $\mathbb{P}^n_{\mathbb{F}_1}$ we have additional geometrical structure, for example:

- $\mathbb{P}^n_{\mathbb{F}_1}$ has a structural sheaf of monoids;
- on $\mathbb{P}^n_{\mathbb{F}_1}$ there is the notion of line bundles;
- these line bundles, or better their isomorphism classes, form the Picard group which is isomorphic to $\mathbb{Z}$.

Finally, given a monoid M we can get functorially a commutative ring

$$\mathbb{Z}[M] \doteq \{\sum n_i m_i | n_i \in \mathbb{Z}, m_i \in M\}$$

like the free functor we saw so far. The geometric interpretation of this algebra operation is given by the following proposition.

Proposition 4.9 *For any affine $\mathbb{F}_1$-scheme X we get an affine scheme $X \otimes_{\mathbb{F}_1} \mathbb{Z}$ (by base change). Moreover the base change "glue" nicely, i. e. for any $\mathbb{F}_1$-scheme we get a scheme $X \otimes_{\mathbb{F}_1} \mathbb{Z}$.*

Finally, this base change satisfies the following reasonable properties.

Proposition 4.10 *There are the following isomorphisms:*

$$\mathbb{A}^n_{\mathbb{F}_1} \otimes_{\mathbb{F}_1} \mathbb{Z} \cong \mathbb{A}^n_{\mathbb{Z}}$$

$$\mathbb{P}^n_{\mathbb{F}_1} \otimes_{\mathbb{F}_1} \mathbb{Z} \cong \mathbb{P}^n_{\mathbb{Z}}.$$

So, from this discussion we can state that geometry over $\mathbb{F}_1$ can (or could) be thought as a sort of "geometry over commutative monoids".

References

[1] A. Connes, C. Consani, *On the notion of geometry over* $\mathbb{F}_1$. Available at `arXiv:0809.2926v2` [math.AG] (2008).

[2] A. Deitmar, "Schemes over $\mathbb{F}_1$". Number fields and function fields – two parallel worlds, Progr. Math., vol. 239, 2005.

[3] C. Deninger,, *Local L-factors of motives and regularized determinants.* Invent. Math. 107 (1992), 135–150.

[4] N. Durov, "A New Approach to Arakelov Geometry". Ph.D. Thesis, available at `arXiv:0704.2030v1` [math.AG], 2007.

[5] P. Freyd, "Homotopy is not concrete". The Steenrod Algebra and its Applications, Springer Lecture Notes in Mathematics Vol. 168, 1970.

[6] M. J. Shai Haran, *Non-additive geometry.* Compositio Math. 143 (2007), 618–688.

[7] M. L. Reyes, *Obstructing extensions of the functor Spec to noncommutative rings.* Israel J. of Math. 192/2 (2012), 667–698.

[8] C. Soulé, *Les variétés sur le corps á un élément.* Mosc. Math. J. 4 (2004), 217–244.

[9] J. Tits, "Sur les analogues algébriques des groupes semi-simples complexes". 1956.

[10] B. Töen and M. Vaquié, *Au-dessous de* Spec $\mathbb{Z}$. Journal of K-Theory (2008), 1–64.

Reachability problems via level set approach

Athena Picarelli [(*)]

Abstract. Given a controlled dynamical system, the characterization of the backward reachable set, i.e. the set of initial states from which it is possible to reach a given target set, can be very interesting in many applications. However realistic models may involve some constraints on state and/or control variables (for taking into account physical or economical constraints, obstacles, etc.) and this can make the characterization of this set much more complicated. After an introduction to the notion of backward reachability in the deterministic as well in the stochastic framework, aim of the talk is to present a technique, based on a level set approach, for characterizing and numerically computing the reachable set also if state constraints are taken into account.

1 Introduction

Consider the following dynamics given by a system of controlled ordinary differential equations (ODEs) in $\mathbb{R}^d$

$$\begin{cases} \dot{X}(s) = b(s, X(s), u(s)) & s \in [t, T] \\ X(t) = x \end{cases} \tag{1}$$

where $T > 0$ is a fixed final instant and the control $u \in \mathcal{U}$ set of the measurable functions with values in a compact set $U \subset \mathbb{R}^m$. Let us introduce the following assumptions:

(D1) $b : [0, T] \times \mathbb{R}^d \times U \to \mathbb{R}^d$ is a continuous function. Moreover, there exists $L > 0$ such that for every $x, y \in \mathbb{R}^d, s \in [0, T], u \in U$:

$$|b(s, x, u) - b(s, y, u)| \leq L|x - y|;$$

(D2) for any $t \in [0, T]$, $x \in \mathbb{R}^d$

$$\big\{b(t, x, U)\big\} \text{ is a convex set.}$$

It is well-known that under assumption (D1), for any choice of $u \in \mathcal{U}$ there is a unique solution of equation (1). We will denote this solution by $X^u_{t,x}(\cdot)$.
Let us introduce a target set $\mathcal{T} \subseteq \mathbb{R}^d$ nonempty and closed. In what follows we aim to

[(*)]SADCO project, INRIA Saclay & ENSTA ParisTech, 828 Boulevard des Maréchaux, 91762 Palaiseau Cedex, France. E-mail: `athena.picarelli@inria.fr`. Seminar held on February 12th, 2014.

characterize and numerically compute the "backward reachable set" for the control system (1)

$$\mathcal{R}_t^{\mathcal{T}} := \left\{ x \in \mathbb{R}^d : \exists u \in \mathcal{U} \text{ such that } X_{t,x}^u(T) \in \mathcal{T} \right\},$$

that is the set of initial points $x \in \mathbb{R}^d$ from which it is possible to reach the target $\mathcal{T}$ at the final time T.
Because of its great involvement in many applications, this kind of problems has been widely studied by several authors with different techniques. Typical example may come from UAVs (unmanned aerial vehicles) control problems (landing procedure, detection of safe zones, etc.) and, more recently, from the development of driving assistance systems. Along the main references we can mention [2], [16] and, more recently [3]. In what follows for characterizing $\mathcal{R}_t^{\mathcal{T}}$ we will apply a level set approach. The idea at the basis of this method is to interpret a set, the backward reachable set $\mathcal{R}_t^{\mathcal{T}}$ for us, as the level set of a suitable continuous function that can be characterize as the solution of a partial differential equation (PDE).
This method has been introduced by Osher and Sethian [15] for fronts propagation problems and then applied by other authors for dealing with a wide class of problem (as in [12] for rendez-vous problems). The main advantage of the level-set method is that it allows to put in relation the characterization of a set with the solution of a PDE for which a wide choice of numerical methods is now available.
In the rest of this introductory section this technique is applied in order to compute the backward reachable set $\mathcal{R}_t^{\mathcal{T}}$. Then the rest of this note is concerned with the extension of this technique to the constrained case.
Let us introduce a function $g_{\mathcal{T}}$ such that:

(D3) $g_{\mathcal{T}} : \mathbb{R}^d \to \mathbb{R}$ is a Lipschitz continuous function and

$$g_{\mathcal{T}}(x) \leq 0 \Leftrightarrow x \in \mathcal{T}.$$

Remark 1.1 If the target set $\mathcal{T}$ is closed, then such a function $g_{\mathcal{T}}$ always exists taking for instance the signed Euclidean distance to $\mathcal{T}$.

Consider the following optimal control problem

$$v(t,x) := \inf_{u \in \mathcal{U}} g_{\mathcal{T}}(X_{t,x}^u(T)) \tag{2}$$

(we refer to v as the value function associated to the optimal control problem (2).
The following result can be easily proved:

Proposition 1.2 *Let assumptions (D1)-(D3) be satisfied. Then*

$$\mathcal{R}_t^{\mathcal{T}} = \left\{ x \in \mathbb{R}^d : v(t,x) \leq 0 \right\}.$$

Proposition 1.2 represent the fundamental point in the level-set method since it establishes a link between the reachability problem and the value function v associated to an optimal control problem in a standard Mayer's form. It comes from the classical theory of optimal control that, applying dynamic programming techniques, v can be characterize as the unique solution, in viscosity sense (see [4] for definitions and main results) of a Hamilton-Jacobi-Bellman (HJB) equation as stated by the following theorem:

Theorem 1.3 *Let assumption (D1)-(D3) be satisfied. Then v is the unique continuous viscosity solution to the following HJB equation*

$$\begin{cases} -\partial_t v + \sup_{u\in U}\{-b(t,x,u)Dv\} = 0 & (t,x) \in [0,T)\times\mathbb{R}^d \\ v(T,x) = g_{\mathcal{T}}(x) & x \in \mathbb{R}^d \end{cases} \tag{3}$$

We can refer to Appendix A in [4] for a general discussion on numerical approximation of equation (3).

When state constraints are taken into account solve the optimal control problem by the dynamic programming approach becomes more complicated. This case will be discussed in Section 2 proposing an efficient technique for overcome the difficulties that typically arise in this case. In Section 3 the stochastic case is presented.

2 State-constrained reachability: the deterministic case

Let us introduce a nonempty and closed set $\mathcal{K} \subseteq \mathbb{R}^d$. The set $\mathcal{K}$ represents the state constraint of our problem, that means that the following condition is required to be satisfied:

$$X^u_{t,x}(s) \in \mathcal{K}, \qquad \forall s \in [t,T]. \tag{4}$$

In this case the backward reachable set is defined by

$$\mathcal{R}_t^{\mathcal{T},\mathcal{K}} := \left\{ x \in \mathbb{R}^d : \exists u \in \mathcal{U} \text{ s.t. } X^u_{t,x}(T) \in \mathcal{T} \text{ and } X^u_{t,x}(s) \in \mathcal{K}, \forall s \in [t,T] \right\}.$$

Applying the level set approach as presented in the introduction we are faced with the following state-constrained optimal control problem:

$$v(t,x) = \inf \left\{ g_{\mathcal{T}}(X^u_{t,x}(T)), u \in \mathcal{U} \text{ and } X^u_{t,x}(s) \in \mathcal{K}, \forall s \in [t,T] \right\},$$

where the infimum is taken only over the controls $u \in \mathcal{U}$ such that the corresponding trajectory satisfies the state constraint (4).
In the state-constrained case the characterization of v in terms of the unique solution of a suitable HJB equation becomes more complicated and it is due mainly to the loss of regularity of v that, in absence of further hypotheses can only be characterize as a bilateral solution of the state-constrained HJB equation. In order to prove uniqueness

some *compatibility assumptions* between the dynamics and the set of state constraints have to be considered. It is the case of the well-known inward pointing conditions introduced by Soner in [17, 18]. Other examples of this kind of conditions are [19] and [13].
The problem is that this kind of condition can be quite restrictive and several simple examples can be produced where they are not satisfied. For this reason in what follows we try to avoid it, taking into account the state constraints by adding an exact penalization term in the definition of the value function v. A complete presentation of this arguments can be found in [9].
Let us introduce a function $g_{\mathcal{K}}$ such that:

(D4) $g_{\mathcal{K}} : \mathbb{R}^d \to \mathbb{R}$ is a Lipschitz continuous function and

$$g_{\mathcal{K}}(x) \leq 0 \Leftrightarrow x \in \mathcal{K}$$

(the existence of such a function is guaranteed, see Remark 1.1 and let us consider the following optimal control problem:

$$w(t,x) := \inf_{u \in \mathcal{U}} \left\{ g_{\mathcal{T}}(X^u_{t,x}(T)) \vee \max_{s \in [t,T]} g_{\mathcal{K}}(X^u_{t,x}(s)) \right\} \tag{5}$$

where the notation $a \vee b := \max(a, b)$ is used.

Remark 2.1 What it is important to notice at this point is that (5) is an *unconstrained* optimal control problem, since the state constraints appear only in the penalization term $\max_{s \in [t,T]} g_{\mathcal{K}}(X^u_{t,x}(s))$ and the infimum is taken over the whole set $\mathcal{U}$.

The following result can be proved:

Proposition 2.2 *Let assumptions (D1)-(D4) be satisfied. Then w is a Lipschitz continuous function and*

$$\mathcal{R}_t^{\mathcal{T},\mathcal{K}} = \left\{ x \in \mathbb{R}^d : w(t,x) \leq 0 \right\}.$$

Proposition 2.2 can be seen as the equivalent of Proposition 1.2 in the introduction. In particular it expresses a link between the state-constrained reachability problem and the unconstrained optimal control problem (5). We conclude this section with the characterization of w as the unique solution of a suitable HJB equation. The optimal control problem (5) is not in a standard form because of the presence of the maximum in the definition of the cost functional. Anyway in the deterministic framework this non-standard formulation does not pose particular difficulties and a Dynamic Programming Principle (DPP) can be obtained (we can also refer to [7] for a different approach).

Lemma 2.3 (DPP) *Assume that (D1)-(D4) hold. For any $t \in [0,T]$ and $h \geq 0$ such that $t + h \leq T$,*

$$w(t,x) = \inf_{u \in \mathcal{U}} \left\{ w(t+h, X^u_{t,x}(t+h)) \vee \max_{s \in [t,t+h]} g_{\mathcal{K}}(X^u_{t,x}(s)) \right\}.$$

Thanks to this result together with the comparison principle that can be proved for equation (6) below (see the appendix in [1], we can finally characterize w:

Theorem 2.4 *Let assumptions (D1)-(D4) be satisfied. Then w is the unique continuous viscosity solution of the following equation*

$$\text{(6)} \qquad \begin{cases} \min\Big(-\partial_t w + \sup_{u\in U}\{-b(t,x,u)Dw\}, w - g_{\mathcal{K}} \Big) = 0 & (t,x) \in [0,T)\times\mathbb{R}^d \\ w(T,x) = g_{\mathcal{T}}(x) \vee g_{\mathcal{K}}(x) & x \in \mathbb{R}^d \end{cases}$$

Equation (6) is referred in literature as an obstacle problem since, roughly speaking, any solution of (6), being in particular a super-solution is required to "overcome" the obstacle given by the function $g_{\mathcal{K}}$.
We conclude the section discussing the reasons that aim the choice of the level function w as a solution of an optimal control problem with a maximum cost. It is in fact evident that other possible kinds of penalization are available so that Proposition 2.2 holds. The simplest one (at least from the theoretical point of view) is given by

$$\text{(7)} \qquad \tilde{w}(t,x) = \inf_{u\in\mathcal{U}}\left\{ g_{\mathcal{T}}(X^u_{t,x}(T)) + \int_t^T g_{\mathcal{K}}(X^u_{t,x}(s))ds \right\},$$

where $g_{\mathcal{T}}$ and $g_{\mathcal{K}}$ are two positive functions satisfying (D3) and (D4). The optimal control problem (7) is in a standard Bolza form. It turns out from the very classical theory of optimal control (see again [4]) that, under our assumptions, $\tilde{w}$ is the unique continuous viscosity solution of the following HJB equation

$$\text{(8)} \qquad \begin{cases} -\partial_t \tilde{w} + \sup_{u\in U}\{-b(t,x,u)D\tilde{w}\} - g_{\mathcal{K}}(x) = 0 & (t,x) \in [0,T)\times\mathbb{R}^d \\ \tilde{w}(T,x) = g_{\mathcal{T}}(x) & x \in \mathbb{R}^d \end{cases}$$

The reasons that lead us to prefer the function w as level set function come from numerics, as the following example shows.

Example 2.5 Let us consider a simple uncontrolled dynamics (harmonic oscillator) in 2 dimensions:

$$\begin{cases} \dot{x} = 2\pi y \\ \dot{y} = -2\pi x. \end{cases}$$

Let the target set $\mathcal{T}$ and the set of state constraints $\mathcal{K}$ be given by

$$\mathcal{T} = B_{\frac{1}{2}}(1,0) \quad \text{and} \quad \mathcal{K} = \mathbb{R}^2 \setminus B_{\frac{1}{2}}(0,0.5)$$

(where $B_r(x_0)$ denotes the ball of radius r centered in x_0). Figures 1 and 2 show the different approximations of the backward reachable set obtained applying respectively the penalization with the maximum term (i.e. the value function w) and the penalization by

the integral term (i.e. the value function $\tilde{w}$).

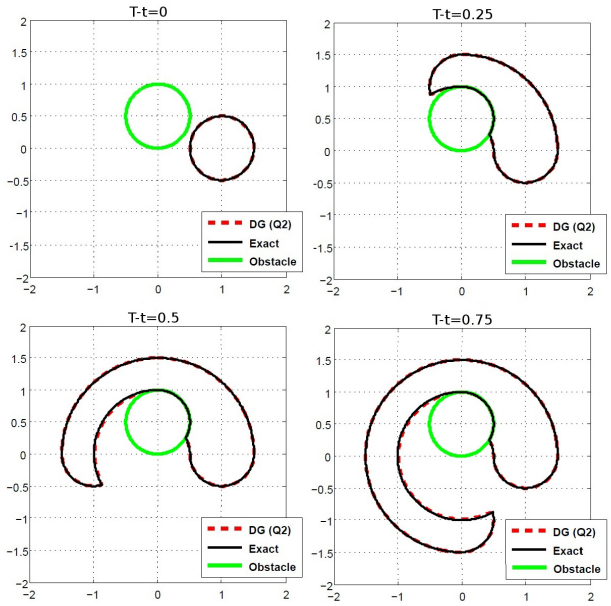


Figure 1: Representation of the zero-level set of the function w (broken line) and of $\mathcal{R}_t^{\mathcal{T},\mathcal{K}}$ (continuous line).

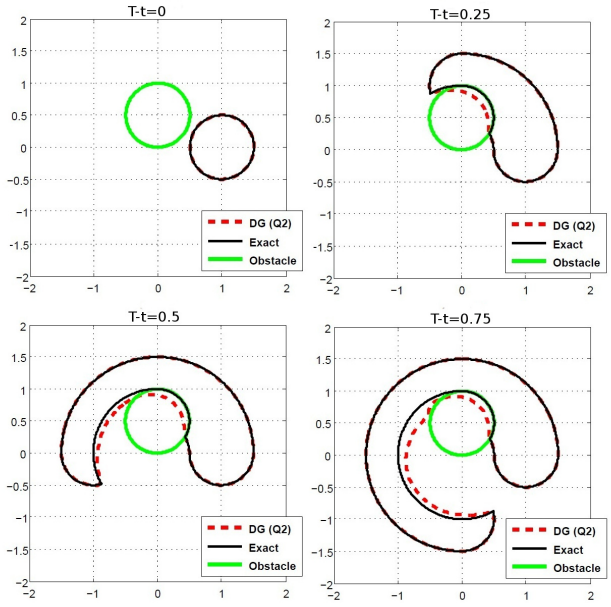


Figure 2: Representation of the zero-level set of the function $\tilde{w}$ (broken line) and of $\mathcal{R}_t^{\mathcal{T},\mathcal{K}}$ (continuous line).

The numerical test presented is taken by [8] where the HJB is solved my the discontinuous Galerkin (DG) method (by the way similar results are obtained also by using different methods).
By this example (and it seems to hold in general) it is evident the best accuracy of the level set approach based on the optimal control problem with maximum cost (see [8] for a discussion about this phenomena).

3 State-constrained reachability: the stochastic case

In this last section the adaptation to the stochastic context of the presented results is discussed. Let $(\Omega, \mathcal{F}, ¶)$ be a probability space with a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ and a Brownian motion $\mathcal{B}(\cdot)$ adapted to $\{\mathcal{F}_t\}_{t\geq 0}$. In this case the dynamics is given by a system of stochastic differential equations (SDEs) in $\mathbb{R}^d$:

$$(9) \qquad \begin{cases} dX(t) &= b(s, X(s), u(s))\mathrm{d}s + \sigma(s, X(s), u(s))\mathrm{d}\mathcal{B}(s) \qquad s > 0,\ x \in \mathbb{R}^d, \\ X(0) &= x, \end{cases}$$

where the control u belongs to $\mathcal{U}$ set of the progressively measurable processes with values in the compact set U.
The following classical assumptions for the coefficients b (drift) and σ (volatility) are taken into account:

(S1) $b : [0,T] \times \mathbb{R}^d \times U \to \mathbb{R}^d$ is a continuous function. Moreover, there exists $L > 0$ such that for every $x, y \in \mathbb{R}^d, s \in [0,T], u \in U$:

$$|b(s,x,u) - b(s,y,u)| + |\sigma(s,x,u) - \sigma(s,y,u)| \leq L|x-y|;$$

(S2) for any $t \in [0,T]$, $x \in \mathbb{R}^d$

$$\left\{(b, \sigma\sigma^T)(t,x,U)\right\} \text{ is a convex set.}$$

It is well-known (see [14] for instance) that under assumption (S1), for any choice of the control u, there exists a unique strong solution of (9). We still denote this solution $X^u_{t,x}(\cdot)$.
In the stochastic case the concept of reachability can be defined in different ways: it can be required the existence of a control $u \in \mathcal{U}$ such that almost every path $X^u_{t,x}(\cdot)(\omega)$ reaches the target satisfying the state constraints or one could be interested to the set of initial points from which $\mathcal{T}$ is reached and the constraints satisfied which a sufficiently big probability.
In this context we deal with the first definition of reachability (clearly the strongest one), i.e.

$$\mathcal{R}^{\mathcal{T},\mathcal{K}}_{t,S} := \left\{ x \in \mathbb{R}^d : \exists u \in \mathcal{U} \text{ s.t. } \left(X^u_{t,x}(T) \in \mathcal{T} \text{ and } X^u_{t,x}(s) \in \mathcal{K}, \forall s \in [t,T] \right) \text{ a.s.} \right\}$$

where a.s. stands for "almost surely", that is for almost every $\omega \in \Omega$.
Following the ideas developed in the deterministic case let us consider two functions $f_{\mathcal{T}}$ and $f_{\mathcal{K}}$ such that:

(S3) $f_{\mathcal{T}}, f_{\mathcal{K}} : \mathbb{R}^d \to \mathbb{R}$ are Lipschitz continuous functions and

$$f_{\mathcal{T}}, f_{\mathcal{K}} \geq 0 \qquad f_{\mathcal{T}}(x) = 0 \Leftrightarrow x \in \mathcal{T}, \quad f_{\mathcal{K}}(x) = 0 \Leftrightarrow x \in \mathcal{K}$$

and let us define the following stochastic optimal control problem:

$$\vartheta(t,x) = \inf_{u \in \mathcal{U}} \mathbb{E}\Big[f_{\mathcal{T}}(X^u_{t,x}(T)) \vee \max_{s \in [0,T]} f_{\mathcal{K}}(X^u_{t,x}(s)) \Big]. \tag{10}$$

Thanks to the choice of the functions $f_{\mathcal{T}}$ and $f_{\mathcal{K}}$ one has:

Proposition 3.1 *Let assumptions (S1)-(S3) be satisfied. Then the following characterization of the state-constrained backward reachable set $\mathcal{R}_t^{\mathcal{T},\mathcal{K}}$ holds:*

$$\mathcal{R}_{t,S}^{\mathcal{T},\mathcal{K}} = \Big\{ x \in \mathbb{R}^d : \vartheta(t,x) = 0 \Big\}.$$

If in the deterministic case dealing with a cost functional in a maximum form did not add any particular technical difficulty, this is not true in the stochastic context. This kind of problems have been studied by several authors in [5, 6] also because of some interesting application in finance (look-back options). The main difficulties arise since, as a consequence of the non commutativity between expectation and maximum, ϑ does not satisfy a DPP. To avoid this difficulty it is classical to reformulate the problem by adding a new variable $y \in \mathbb{R}$ that, roughly speaking, keeps the information of the running maximum. For this reason, we introduce an auxiliary value function $\tilde{\vartheta}$ defined on $[0,T] \times \mathbb{R}^d \times \mathbb{R}$ by:

$$\tilde{\vartheta}(t,x,y) := \inf_{u \in \mathcal{U}} \mathbb{E}\Big[f_{\mathcal{T}}(X^u_{t,x}(T)) \vee \max_{s \in [0,T]} f_{\mathcal{K}}(X^u_{t,x}(s)) \vee y \Big]. \tag{11}$$

Thanks to the positivity of the functions involved one can easily verify that $\tilde{\vartheta}(t,x,0) = \vartheta(t,x)$ and then, by Proposition 3.1 one has

$$\mathcal{R}_{t,S}^{\mathcal{T},\mathcal{K}} = \Big\{ x \in \mathbb{R}^d : \tilde{\vartheta}(t,x,0) = 0 \Big\}. \tag{12}$$

In virtue of the equality above, that expresses a link between $\tilde{\vartheta}$ and the state-constrained backward reachable set, from now we focus on the characterization of $\tilde{\vartheta}$ as a solution of a suitable PDE. For $\tilde{\vartheta}$ a DPP holds as the following proposition states:

Proposition 3.2 *Assume (S1)-(S3). Then there exists $L \geq 0$ such that*

$$|\tilde{\vartheta}(t,x,y) - \tilde{\vartheta}(t',x',y')| \leq L(|x-x'| + |y-y'| + (1+|x|)|t-t'|^{1/2}),$$

for all $t, t' \in [0,T]$, $x, x' \in \mathbb{R}^d$ and $y, y' \in \mathbb{R}$.
Moreover for all $(t,x,y) \in [0,T) \times \mathbb{R}^d \times \mathbb{R}$ and all family of stopping times $\{\tau^u, u \in \mathcal{U}\}$ independent of $\mathcal{F}_t$ with values on $[t,T]$:

$$\tilde{\vartheta}(t,x,y) = \inf_{u \in \mathcal{U}} \mathbb{E}\left[\tilde{\vartheta}(\tau^u, X^u_{t,x}(\tau^u), \max_{s \in [t,\tau^u]} f_{\mathcal{K}}(X^u_{t,x}(s)) \vee y) \right].$$

Thanks to this result we can finally characterize $\tilde{\vartheta}$ as the solution of an HJB equation:

Theorem 3.3 *Under assumptions (S1)-(S3), the value function $\tilde{\vartheta}$ is the unique continuous, Lipschitz in (x,y), viscosity solution of the following HJB equation*

$$(13)\quad \begin{cases} -\partial_t\tilde{\vartheta} + \sup\limits_{u\in U}\left\{-b(t,x,u)D\tilde{\vartheta} - \frac{1}{2}Tr[\sigma\sigma^T](t,x,u)D^2\tilde{\vartheta}\right\} = 0 & \text{in } [0,T)\times D \\ -\partial_y\tilde{\vartheta} = 0 & \text{on } [0,T)\times\partial D \\ \tilde{\vartheta}(T,x,y) = f_{\mathcal{T}}(x)\vee y & \text{in } \overline{D} \end{cases}$$

where $D \subset \mathbb{R}^{d+1}$ is the domain defined by

$$\overline{D} := \left\{(x,y)\in\mathbb{R}^{d+1} : y \geq f_{\mathcal{K}}(x)\right\} = \text{Epigraph}(f_{\mathcal{K}}).$$

Equation (13) is a second order HJB with oblique derivative boundary condition in the direction $-e_y$. For the definition of solutions in the viscosity sense to this kind of boundary problems we refer to [11] and the references therein. We also refer to [10] for a complete proof of this result and for the consequent discussion on the numerical approximation.
We conclude presenting a numerical example that, even if in a simplified setting, well shows the potentiality of the level set method applied to the stochastic context.

Example 3.4 Let us consider the following dynamics:

$$(14)\quad \begin{cases} dX(s) = u(s)\begin{pmatrix}1\\0\end{pmatrix}ds + u(s)\sigma(X(s))d\mathcal{B}(s),\ s\geq t, \\ X(t) = x \end{cases}$$

where $\mathcal{B}$ is a one-dimensional Brownian motion, $U=[0,1]\subset\mathbb{R}$ and the function $\sigma(x)\in\mathbb{R}^2$ will vary depending on the example. The target set is

$$\mathcal{T} = \left\{x\equiv(x_1,x_2)\in\mathbb{R}^2,\ 0\leq x_1\leq 0.4,\ |x_2|\leq 0.5\right\}$$

and the set of state-constraints is

$$\mathcal{K} = \mathbb{R}^2 \setminus \left\{x\equiv(x_1,x_2)\in\mathbb{R}^2,\ -0.4<x_1<-0.2,\ |x_2|<0.1\right\}.$$

Given a final time $T=0.5$, we represent in Figure 3 (green colored region) the approximation of the backward reachable set $\mathcal{R}_{0,S}^{\mathcal{T},\mathcal{K}}$ obtained by our method (that is approximating the solution $\tilde{\vartheta}$ of equation (13) and then using (12)) corresponding to different choices of the function σ:

$$\text{(a)}:\sigma(x)\equiv\begin{pmatrix}0\\0\end{pmatrix},\quad \text{(b)}:\sigma(x):=\begin{pmatrix}0\\5\end{pmatrix}\quad \text{(c)}:\sigma(x):=5\,d_\Theta(x)\begin{pmatrix}0\\1\end{pmatrix}$$

where $\Theta := \left\{(x_1,x_2),\ |x_2|\geq 0.3\right\}$.

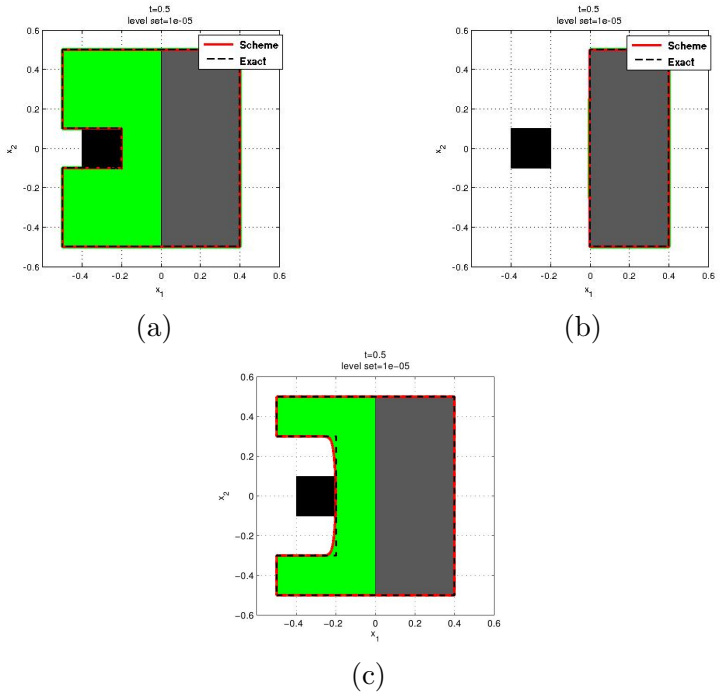


Figure 3: (a): no diffusion, (b): with diffusion and (c): degenerate diffusion.

References

[1] Altarovici A., Bokanowski O., and Zidani H., *A general Hamilton-Jacobi framework for non-linear state-constrained control problems.* ESAIM: COCV, Vol. 19 (2013), 337–357.

[2] Aubin J.-P., *Viability kernels and capture basins of sets under differential inclusions.* SIAM J. Control Optim., Vol. 40/3 (2001), 853–881.

[3] Baier R., Gerdts M., and Xausa I., *Approximation of Reachable Sets using Optimal Control Algorithms.* Numerical algebra, control and Optimization, Vol. 3/3 (2013), 519–548.

[4] Bardi M., and Capuzzo Dolcetta I., "Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman equations". Systems Control Found. Appl., Birkhäuser Boston, Inc., Boston, MA, 1997.

[5] Barles G., Daher C., and Romano M., *Optimal control on the L^∞ norm of a diffusion process.* SIAM J. Control and Optim., Vol. 32/3 (1994), 612–634.

[6] Barron E.N., *The Bellman equation for control of the running max of a diffusion and applications to lookback options.* Applicable Analysis, Vol. 48 (1993), 205–222.

[7] Barron E.N., and Ishii H., *The Bellman equation for minimizing the maximum cost.* Nonlinear Analysis, TMA, Vol. 13/9 (1989), 1067–1090.

[8] Bokanowski O., Cheng Y., and Shu C.-W., *A discontinuous Galerkin scheme for front propagation with obstacles*. Numerische Math. Vol. 126 (2014), 1-31.

[9] Bokanowski O., Forcadel N., and Zidani H., *Reachability and minimal times for state constrained nonlinear problems without any controllability assumption*. SIAM J. Control Optim., Vol. 48/7 (2010), 4292–4316.

[10] Bokanowski O., Picarelli A., and Zidani H., *Dynamic programming and error estimates for stochastic control problems with maximum cost*. Appl. Math. Optim. (2014), to appear.

[11] Crandall M.G., Ishii H., and Lions P.L., *User's guide to viscosity solutions of second order partial differential equations*. Bull. Amer. Math. Soc. (N. S.), Vol. 27/1 (1992), 1–67.

[12] Falcone M., Giorgi T., and Loreti P., *Level sets of viscosity solutions: Some applications to fronts and rendez-vous problems*. SIAM J. Appl. Math., Vol. 54 (1994), 1335–1354.

[13] Frankowska H., and Vinter R., *Existence of neighboring feasible trajectories: applications to dynamic programming for state-constrained optimal control problems*. J. Optim. Theory Appl., Vol. 104 (2000), 21–40.

[14] Karatzas I., and Shreve S., "Brownian motion and stochastic calculus". Springer, New York, 1991.

[15] Osher S., and Sethian A.J., *Fronts propagating with curvature dependent speed: algorithms on Hamilton-Jacobi formulations*. J. Comp. Phys., Vol. 79 (1988), 12-49.

[16] Saint-Pierre P., *Approximation of the viability kernel*. Appl. Math. Optim., Vol. 29 (1994), 187–209.

[17] Soner H.M., *Optimal control with state-space constraint. I*. SIAM Journal on Control and Optimization, Vol. 24 (1986), 552-561.

[18] Soner H.M., *Optimal control with state-space constraint. II*. SIAM Journal on Control and Optimization, Vol. 24 (1986), 1110-1122.

[19] Soravia P., *Optimality principles and representation formulas for viscosity solutions of Hamilton-Jacobi equations. II. Equations of control problems with state constraints*. Differential Integral Equations, Vol. 12 (1999), 275-293.

Minimum time function for linear control systems

Luong V. Nguyen (*)

Abstract. We consider linear time-optimal control problems. We first introduce some basic notions of linear control systems and then the minimum time function to reach the origin and its properties. We will focus on the set $\mathcal{S}$ of points where the minimum time function fails to be Lipschitz. More precisely, we will compute the set $\mathcal{S}$ and study its structure.

1 Linear control systems and controllability

In this section, we introduce some basic notions of linear control systems which can find in any book about optimal control theory e.g. [6, 7].
We consider a linear control process which is described by a linear differential system

$$\dot{y}(t) = Ay(t) + Bu(t) \tag{1.1}$$

The coefficients A and B are known matrices of order $N \times N$ and $N \times M$ respectively with $1 \leq M \leq N$. We choose the control $u(t)$ which takes values in the control set $U = [-1, 1]^M$ to steer or control the response $x(t)$ from an initial state x_0 to the origin. Denote by $\mathcal{U}_{ad}$ the set of all admissible control i.e., the set of measurable functions $u : [0, \infty) \to U$ and call $\mathcal{U}_{ad}$ the admissible control set. It is known that for each admissible control $u(\cdot)$, the system (1.1) with the initial condition

$$y(0) = x_0 \in \mathbb{R}^N. \tag{1.2}$$

has a unique solution $y^{x_0,u}(\cdot)$. We call $y^{x_0,u}(\cdot)$ the *trajectory* starting at x_0 corresponding to the control $u(\cdot)$. One has

$$y^{x_0,u}(t) = e^{At}x_0 + \int_0^t e^{A(t-s)}Bu(s)ds. \tag{1.3}$$

with

$$e^{At} = \sum_{k=0}^{\infty} t^k . \frac{A^k}{k!}$$

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: luonghdu@gmail.com . Seminar held on January 26th, 2014.

We are now going to discuss the possibility of steering an initial state x_0 precisely to the origin in finite time.

Definition 1.1 For a given $t > 0$. The reachable set at time t is the set of initial points x_0 for which there exists an admissible control $u(\cdot)$ such that $y^{x_0,u}(t) = 0$. We denote that set by $\mathcal{R}(t)$.

The rechable set is the set of initial points x_0 for which there is some admissible control $u(\cdot)$ such that $y^{x_0,u}(t) = 0$ for some $t > 0$ and we denote this set by $\mathcal{R}$.

From the Definition 1.1, we have

$$\mathcal{R}(t) = \{x_0 \in \mathbb{R}^N : \exists u(\cdot) \in \mathcal{U}_{ad} \text{ such that } y^{x_0,u}(t) = 0\},$$

and

$$\mathcal{R} = \bigcup_{t>0} \mathcal{R}(t).$$

Now for a given $t > 0$. By Definition 1.1, $x_0 \in \mathcal{R}(t)$ if and only if there exists $u(\cdot) \in \mathcal{U}_{ad}$ such that

$$e^{At}x_0 + \int_0^t e^{A(t-s)}Bu(s)ds = 0.$$

It follows that

$$x_0 = -\int_0^t e^{-As}Bu(s)ds.$$

Therefore

$$\mathcal{R}(t) = \left\{x_0 = -\int_0^t e^{-As}Bu(s)ds : u(\cdot) \in \mathcal{U}_{ad}\right\}. \tag{1.4}$$

From the above representation of the reachable set, one can easily prove the following properties

Proposition 1.2 *For a fixed $t > 0$, one has*

(i) $\mathcal{R}(t)$ *is convex, symmetric and compact.*

(ii) $\mathcal{R}(t) \subseteq \mathcal{R}(s)$ *for* $0 \leq t \leq s$.

Definition 1.3 The linear control system (1.1) is small-time controllable if 0 is in the interior of $\mathcal{R}$. Moreover, if $\mathcal{R} = \mathbb{R}^N$, we say that (1.1) is fully controllable.

Example 1.4 Consider a control system with

$$A = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

For $t > 0$, we can compute

$$\mathcal{R}(t) = \left\{ \begin{pmatrix} 0 \\ x_2 \end{pmatrix} : -t \leq x_2 \leq t \right\},$$

and

$$\mathcal{R} = \left\{ \begin{pmatrix} 0 \\ x_2 \end{pmatrix} : x_2 \in \mathbb{R} \right\}.$$

Since $0 \notin \operatorname{Int} \mathcal{R}$, this system is not small-time controllable.
We now give some necessary and sufficient conditions for controllability.

Theorem 1.5 *The linear control system (1.1) is small time controllable if and only if*

$$\operatorname{rank} [B, AB, \cdots, A^{N-1}B] = N. \tag{1.5}$$

Theorem 1.6 *Assume that (1.5) holds and $Re(\lambda) \leq 0$ for each eigenvale λ of A. Then the system (1.1) is fully controllable.*

Example 1.7 Consider a control system with

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

We have

$$Ab = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

then $\operatorname{rank} [b, Ab] = 2$. Moreover, A has one eigenvalue $\lambda = 0$. The control system is fully controllable.

Example 1.8 Consider a control system with

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

We have

$$Ab = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

then $\operatorname{rank} [b, Ab] = 2$. Moreover, A has one eigenvalue $\lambda = 1$. The control system is small-time controlable but not fully controllable.

We now consider a special class of linear control systems: normal linear systems.

Definition 1.9 The linear control system (1.1) is said to be normal if

$$\operatorname{rank} [b_i, Ab_i, \cdots, A^{N-1}b_i] = N \qquad \text{for all } i = 1, \cdots, M. \tag{1.6}$$

We state here a classical results for normal linear systems

Theorem 1.10 *The linear control system (1.1) is normal if and only if the reachable set $\mathcal{R}(t)$ is strictly convex for any $t > 0$.*

Proposition 1.11 *If the control system (1.1) is normal then $\mathcal{R}$ is open and consequently (1.1) is small-time controllable.*

2 Minimum time function and regularity properties

For $x \in \mathbb{R}^n$ and $u \in \mathcal{U}_{ad}$, let $y^{x,u}(\cdot)$ be the solution of (1.1) with the initial condition $y(0) = x$. We define

$$\theta(x,u) = \min\{t \geq 0 : y^{x,u}(t) = 0\}. \tag{2.1}$$

Then $\theta(x,u)$ is the time taken for $y^{x,u}(\cdot)$ to reach the origin provided $\theta(x,u) < \infty$. The minimum time to reach the origin from x is defined by

$$T(x) = \inf\{\theta(x,u) : u \in \mathcal{U}_{ad}\}. \tag{2.2}$$

Set $T(0) = 0$ and $T(x) = +\infty$ for $x \notin \mathcal{R}$, we say that $T : \mathbb{R}^N \to [0, +\infty]$ is the *minimum time function* to reach the origin for the linear control system (1.1).

If the minimum in (2.2) is attained for some $u^*(\cdot) \in \mathcal{U}_{ad}$, then we call $u^*(\cdot)$ an *optimal control* and the corresponding trajectory $y^{x,u^*}(\cdot)$ an *optimal trajectory* for x.

Theorem 2.1 *For every $x \in \mathcal{R}$, one has*

$$T(x) = \min\{\theta(x,u) : u \in \mathcal{U}_{ad}\},$$

i.e., there is an optimal control $u^*(\cdot) \in \mathcal{U}_{ad}$ such that $y^{x,u^*}(T(x)) = 0$.

Proof. See e.g. [6, 7]. □

The following is an important tool in optimal control theory which can be proved easily.

Theorem 2.2 (Dynamic programming principle) *For every $x \in \mathcal{R}$, it holds*

$$T(x) \leq t + T(y^{x,u}(t)),$$

for all $0 < t < T(x)$ and $u(\cdot) \in \mathcal{U}_{ad}$. The equality holds if $u(\cdot)$ is optimal.

From now on, for simplicity, we consider the minimum time function for normal linear control systems.
The following theorem is a version of Pontryagin's Principle for linear control systems.

Theorem 2.3 *Let $T > 0$ and let $x \in \mathcal{R}$. The following statements are equivalent*

(i) $x \in \partial\mathcal{R}(T)$.

(ii) *there exists a unique optimal control u^* steering x to the origin in time T, in particular $T(x) = T$.*

(iii) *for every $0 \neq \zeta \in N_{\mathcal{R}(T)}(x)$, we have*

$$u_i^*(t) = -\text{sign}\,\langle \zeta, e^{-At} b_i \rangle, \;\; a.e.\; t \in [0,T] \qquad for\; i = 1, \cdots, M.$$

A well known reference for this result is [6, Sections 13–15].

Theorem 2.4 *The minimum time function is everywhere finite and continuous. Moreover, T is Hölder continuous with exponent $1/N$.*

Proof. See, e.g., Theorem 17.3 in [6] and Theorem 1.9, Chapter IV, in [1] and references therein. □

Example 2.5 Consider the minimum time function to reach the origin of the system (1.1) with

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

By using Theorem 2.3, one can compute explicitly the minimum time function. For $(x, y) \in \mathbb{R}^2$, we have

$$T(x,y) = \begin{cases} y + 2\sqrt{y^2/2 + x} & \text{if } x \geq -y|y|/2 \\ -y + 2\sqrt{y^2/2 - x} & \text{if } x < -y|y|/2 \end{cases}$$

It can be shown that in this case the function T is not Lipschitz on $\mathcal{R}$. Moreover, one can compute the set $\mathcal{S}$ of points where T fails to be Lipschitz

$$\mathcal{S} = \left\{ (x,y) \in \mathbb{R}^2 : x = \frac{-y|y|}{2} \right\}.$$

In the following we will study the Lipschitz continuity of the minimum time function T. More precisely, we will study the set of points where T is not Lipschitz. By an abstract result, we can compute implicitly the non-Lipschitz set of T. We then somehow estimate the size of the non-Lipschitz set.
We first introduce the *minimized Hamiltonian* associated to the control system (1.1). Define, for $x, p \in \mathbb{R}^N$, the function

$$h(x,p) := \min_{u \in [-1,1]} \langle Ax + Bu, p \rangle. \tag{2.3}$$

We call h the minimized Hamiltonian.
The following theorem give an abstract result on the non-Lipschitz set of the minimum time function.

Theorem 2.6 [5] *Let $\mathcal{S}$ be the set of non-Lipschitz points of T. Then*

$$\mathcal{S} = \{x : \textit{there exists } 0 \neq \zeta \in N_{\mathcal{R}(T(x))}(x) \textit{ such that } h(x,\zeta) = 0\}.$$

The following is a technical lemma concerning an explicit computation of the minimized Hamiltonian which is useful to compute the set $\mathcal{S}$ explicitly.

Lemma 2.7 [5] *Let $r > 0$, $\bar{x} \in \mathbb{R}^N$ and $\bar{\zeta} \in \mathbb{S}^{N-1}$ be such that*

$$\bar{x} = \sum_{i=1}^{M} \int_0^r e^{-At} b_i \operatorname{sign}\left(\langle \bar{\zeta}, e^{-At} b_i \rangle\right) dt.$$

Then

$$h(\bar{x}, \bar{\zeta}) = -\sum_{i=1}^{M} \left| \langle \bar{\zeta}, e^{-Ar} b_i \rangle \right|. \tag{2.4}$$

Then we can characterize the set $\mathcal{S}$ of non-Lipschitz points of T as

$$\begin{aligned} \mathcal{S} = \Big\{ & x \in \mathbb{R}^N : \text{there exist } \ r > 0 \ \text{ and } \ \zeta \in \mathbb{S}^{N-1} \ \text{ such that} \\ & x = \sum_{i=1}^{M} \int_0^r e^{-At} b_i \operatorname{sign}\left(\langle \zeta, e^{-At} b_i \rangle\right) dt, \\ & \qquad \zeta \in N_{\mathcal{R}(r)}(x) \ \text{ and } \ \langle \zeta, e^{-Ar} b_i \rangle = 0 \ \ \forall i = 1, \dots, M \Big\}. \end{aligned} \tag{2.5}$$

More precisely, we can prove the following theorem:

Theorem 2.8 [5] *Let $x \in \mathbb{R}^N \setminus \{0\}$. Then T is not Lipschitz at x if and only if $x \in \mathcal{S}$.*

By changing variables and using the fact that the minimized Hamiltonian is constant along optimal trajectories, we have another representation of $\mathcal{S}$ which is useful in studying its structure.

$$\begin{aligned} \mathcal{S} = \Big\{ & x \in \mathbb{R}^N : \text{ there exist } r > 0 \text{ and } \zeta \in \mathbb{S}^{N-1} \text{ such that} \\ & x = \sum_{i=1}^{M} \int_0^r e^{A(t-r)} b_i \operatorname{sign}\left(\langle \zeta, e^{At} b_i \rangle\right) dt \ \text{ and } \ \langle \zeta, b_i \rangle = 0 \ \forall i = 1, \dots, M \Big\}. \end{aligned} \tag{2.6}$$

Set

$$k = \operatorname{rank} B. \tag{2.7}$$

Of course, $1 \leq k \leq M$. If $k = N$, then $\mathcal{S}$ is empty. If $k < N$, then $\mathcal{S}$ is nonempty.

Example 2.9 Consider the minimum time function T for a normal linear control system with

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

We apply above results to compute the non-Lipschitz set $\mathcal{S}$ of T. Consider

$$\begin{cases} \zeta \in \mathbb{S}^1 \\ \langle \zeta, b \rangle = 0 \end{cases}$$

We have

$$\zeta = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{or} \quad \zeta = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$

Moreover,

$$e^{At} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}$$

Then $x \in \mathcal{S}$ if and only if there is $r > 0$ such that

$$x = \int_0^r e^{A(t-r)} b \, \mathrm{sign}(\langle \zeta, e^{At} b \rangle) dt.$$

Thus

$$x = \int_0^r \begin{pmatrix} \sin(t-r) \\ \cos(t-r) \end{pmatrix} \mathrm{sign}(\sin t) dt \qquad \text{or} \qquad x = -\int_0^r \begin{pmatrix} \sin(t-r) \\ \cos(t-r) \end{pmatrix} \mathrm{sign}(\sin t) dt$$

Therefore we have

$$\mathcal{S} = \left\{ -\int_0^r \begin{pmatrix} \sin(t-r) \\ \cos(t-r) \end{pmatrix} \mathrm{sign}(\sin t) dt : r \geq 0 \right\}$$

In Figure 1, the non-Lipschitz set $\mathcal{S}$ of T consists of two curves: the red and blue curves.

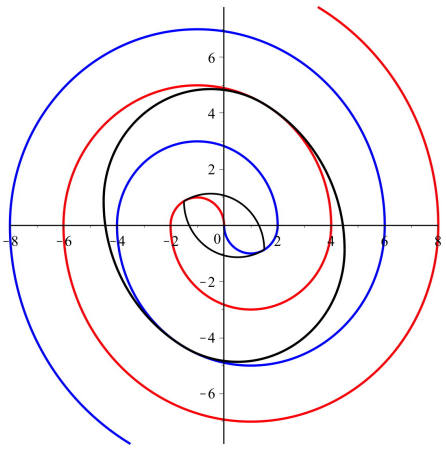


Figure 1: The non-Lipschitz set $\mathcal{S}$ of T

In the case when the matrix A has only real eigenvalues, we can compute the non-Lipschitz set of the minimum time function based on number of switchings of optimal controls. We state here a result for the case of single control i.e. $M = 1$.

Theorem 2.10 *Assume that $M = 1$ and that A has all real eigenvalues. $x \in \mathcal{S}$ if and only if x can be steered to the origin by the optimal control with $k \leq N - 2$ switchings.*

Example 2.11 Consider the minimum time function T for a normal linear control system with

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

The matrix A has only real eigenvalues. $x \in \mathcal{S}$ if and only if x can be steered to the origin by the optimal control with no switching or one switching.
Fix $T > 0$, then $x \in \mathcal{S} \cap \text{bdry}\mathcal{R}(T)$ can be steered to the origin by the optimal control of one of the following forms

- $u(s) = 1$ for $0 \leq s \leq T$.
- $u(s) = -1$ for $0 \leq s \leq T$.
- $u(s) = \begin{cases} 1 & \text{if } 0 \leq s < \alpha T \\ -1 & \text{if } \alpha T \leq s \leq T, \end{cases}$, $\alpha \in [0, 1]$.
- $u(s) = \begin{cases} -1 & \text{if } 0 \leq s < \alpha T \\ 1 & \text{if } \alpha T \leq s \leq T \end{cases}$, $\alpha \in [0, 1]$.

Then we can compute $\mathcal{S} \cap \mathcal{R}(T)$.

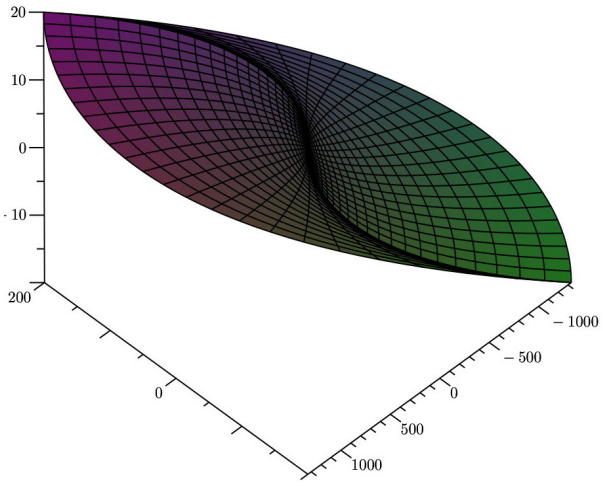


Figure 2: The set of non-lipschitz points of T within $\mathcal{R}(20)$

The structure of the non-Lipschitz set of the minimum time function is described in the following theorem.

Theorem 2.12 [5] *Let $\mathcal{S}$ be defined according to (2.6) and let k be given by (2.7). Then $\mathcal{S}$ is closed and is covered by countably many graphs of Lipschitz functions of $(N-k)$ variables.*

References

[1] M. Bardi, I. Capuzzo Dolcetta, "Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations". Birkhauser, Boston, 1997.

[2] P. Cannarsa and C. Sinestrari, "Semiconcave Functions, Hamilton-Jacobi Equations and Optimal Control". Birkhauser, Boston, 2004.

[3] G. Colombo and Khai T. Nguyen, *On the minimum time function around the origin.* Math. Control Relat. Fields 3 (2013), 51–82.

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

[5] G. Colombo, Khai T. Nguyen, Luong V. Nguyen, *Non-Lipschitz points and the SBV regularity of the minimum time function.* Calculus of Variations and Partial Differential Equations, 10.1007/s00526-013-0682-9.

[6] H. Hermes & J. P. LaSalle, "Functional analysis and time optimal control". Academic Press, New York-London, 1969.

[7] E. B. Lee and L. Markus, "Foundations of Optimal Control Theory". John Wiley, New York, 1968.

Market models with optimal arbitrage

Huy Ngoc Chau (*)

Abstract. We construct and study market models admitting optimal arbitrage. We say that a model admits optimal arbitrage if it is possible, in a zero-interest rate setting, starting with an initial wealth of 1 and using only positive portfolios, to superreplicate a constant $c > 1$. The optimal arbitrage strategy is the strategy for which this constant has the highest possible value.

1 General setting

For the theory of stochastic process and stochastic integration, we refer to Jacod and Shiryaev [4] and Protter [7].
On the stochastic basic $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, we consider a financial market with an $\mathbb{R}^d$-valued nonnegative semimartingale process $S = (S^1, ..., S^d)$ whose components model the prices of d risky assets. The riskless asset is denoted by S^0 and we assume that $S^0 \equiv 1$, that is, all price processes are already discounted. We suppose that the financial market is frictionless, meaning that there are no trading restrictions, transaction costs, or other market imperfections.

Let $L(S)$ be the set of all $\mathbb{R}^d$-valued S-integrable predictable processes. It is the most reasonable class of strategies that investors can choose, but another constraint, which is described below, is needed in order to rule out doubling strategies.

Definition 1.1 Let $x \in \mathbb{R}_+$. An element $H \in L(S)$ is said to be an x-admissible strategy if $H_0 = 0$ and $(H \cdot S)_t \geq -x$ for all $t \in [0, T]$ $\mathbb{P}$-a.s. An element $H \in L(S)$ is said to be an admissible strategy if it is an x-admissible strategy for some $x \in \mathbb{R}_+$.

For $x \in \mathbb{R}_+$, we denote by $\mathcal{A}_x$ the set of all x-admissible strategies and by $\mathcal{A}$ the set of all admissible strategies. As usual, H_t is assumed to represent the number of units of risky asset that we hold at time t. For $(x, H) \in \mathbb{R}_+ \times \mathcal{A}$, we define the portfolio value process $V_t^{x,H} = x + (H \cdot S)_t$. This is equivalent to requiring that portfolios are only generated by self-financing admissible strategies.

Given the semimartingale S, we denote by $\mathcal{K}_x$ the set of all outcomes that one can

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: cnhuy@math.unipd.it . Seminar held on March 12th, 2014.

realize by x-admissible strategies starting with zero initial cost:

$$\mathcal{K}_x = \{(H \cdot S)_T | H \text{ is } x\text{-admissible}\}$$

and by $\mathcal{X}_x$ the set of outcomes of strategies with initial cost x:

$$\mathcal{X}_x = \{x + (H \cdot S)_T | H \text{ is } x\text{-admissible}\}.$$

Remark that all elements in $\mathcal{X}_x$ are nonnegative. The unions of all $\mathcal{K}_x$ and all $\mathcal{X}_x$ are denoted by $\mathcal{K}$ and $\mathcal{X}$, respectively. All bounded claims which can be superreplicated by admissible strategies are contained in

$$\mathcal{C} = \left(\mathcal{K} - L^0_+\right) \cap L^\infty.$$

Now, we recall some no-free-lunch conditions, which are studied in the works of Delbaen and Schachermayer [2], Karatzas and Kardaras [5] and Kardaras [6].

Definition 1.2

- We say that the market satisfies the No Arbitrage (NA) condition with respect to general admissible integrands if

$$\mathcal{C} \cap L^\infty_+ = \{0\}.$$

- We say that the market satisfies the No Free Lunch with Vanishing Risk (NFLVR) property, with respect to general admissible integrands, if

$$\overline{\mathcal{C}} \cap L^\infty_+ = \{0\},$$

  where the bar denotes the closure in the supnorm topology of L^∞.

- There is No Unbounded Profit With Bounded Risk (NUPBR) if the set $\mathcal{K}_1$ is bounded in L^0, that is, if

$$\lim_{c\to\infty} \sup_{H\in\mathcal{A}_1} \mathbb{P}\left[V^{0,H} > c\right] = 0$$

  holds.

We are interested in financial markets satisfying the following assumption.

Assumption 1.3 S is locally bounded, the market satisfies NUPBR but the condition NFLVR fails under the physical measure $\mathbb{P}$.

Under the local boundedness assumption, by the Fundamental Theorem of Asset Pricing, the NFLVR condition is equivalent to the existence of a ELMM (see Corollary 1.2 in Delbaen and Schachermayer [2]). When the NFLVR condition fails but the NUBPR condition holds, the ELMM is replaced with a weaker notion of "deflator".

Definition 1.4 An equivalent local martingale deflator (ELMD) is a nonnegative process Z with $Z_0 = 1$ and $Z_T > 0$ such that ZV is a local martingale for all $V \in \mathcal{X}$.

In particular, an ELMD is a nonnegative local martingale. Fatou's Lemma implies that it is also a supermartingale and its expectation is less or equal to one. Hence, a local martingale density is an ELMD with expectation one. In general, we cannot use an ELMD to define a new probability measure, since the new measure loses mass.

The following result has recently been proven in Kardaras [6] in the one dimensional case.

Theorem 1.5 *The NUPBR condition is equivalent to the existence of at least one ELMD.*

2 Optimal arbitrage

It is well known that NFLVR holds if and only if both NUPBR and NA hold, see Corollary 3.4 and 3.8 of Delbaen and Schachermayer [2] or Proposition 4.2 of Karatzas and Kardaras [5]. Moreover, Lemma 3.1 of Delbaen and Schachermayer [3] shows that if NA fails then either the market admits an immediate arbitrage or an arbitrage that is created by a strategy in $\mathcal{A}_1$. Furthermore, if there exists an immediate arbitrage, the associated strategy is in $\mathcal{A}_0$ and can be freely scaled to produce an unbounded arbitrage. But this situation is not allowed in our market due to Assumption 1.3. Therefore, it is only possible to exploit arbitrages by using strategies in the set $\mathcal{A}_x$ where $x > 0$. For these reasons, arbitrages in our market are limited and the question of optimal arbitrage profit arises naturally.

Definition 2.1 For a fixed time horizon T, we define

$$U(T) := \sup\left\{c > 0 : \exists H \in \mathcal{A}_1, V_T^{1,H} \geq c, \mathbb{P} - a.s\right\}.$$

If $U(T) > 1$, we call $U(T)$ optimal arbitrage profit.

The quantity $U(T)$ is the maximum deterministic amount that one can realize at time T starting from unit initial capital. Obviously, this value is bounded from below by 1.

2.1 Optimal arbitrage and superhedging price

Definition 2.2 Given a claim $f \geq 0$, we define

$$SP_+(f) := \inf\left\{x \geq 0 : \exists H \in \mathcal{A}_x, V_T^{x,H} \geq f, \mathbb{P} - a.s\right\},$$

that is the minimal amount starting from which one can superhedge f by a nonnegative wealth process.

The following lemma is simple but useful to our problem.

Lemma 2.3 $U(T) = 1/SP_+(1)$.

Proof. ($\leq$) Take any $c > 0$ such that there exists a strategy $H \in \mathcal{A}_1$ which satisfies

- $V_T^{1,H} = 1 + (H \cdot S)_T \geq c, \mathbb{P} - a.s.$

- $(H \cdot S)_t \geq -1.$ for all $0 \leq t \leq T$.

Then a simple scaling argument gives us a strategy to hedge 1

- (superheging) $1/c + 1/c(H \cdot S)_T \geq 1, \mathbb{P} - a.s.$
- (admissibility) $1/c(H \cdot S)_t \geq -1/c$ for all $0 \leq t \leq T$.

By Definition 2.2, one can superhedge 1 at cost $1/c$. Therefore, we get an upper bound for optimal arbitrage profit

$$U(T) \leq \frac{1}{SP_+(1)}.$$

$(\geq)$ The converse inequality can be proved by the same argument. □

The above lemma has two consequences. First, the knowledge of $SP_+(1)$ is enough to find optimal arbitrage profit. Second, one should find the strategy to superhedge 1 in order to realize optimal arbitrage. Obviously, $SP_+(1) \leq 1$. If $SP_+(1) < 1$, there is optimal arbitrage. If $SP_+(1) = 1$, optimal arbitrage does not exist, but arbitrages may still exist.

3 A construction based on a predictable stopping time

We consider a measure $\mathbb{Q}$ on the space $(\Omega, \mathcal{F}, (\mathcal{F})_{t\geq 0})$. Let σ be a stopping time such that $\mathbb{Q}(\sigma > T) > 0$. We define a new probability measure, absolutely continuous with respect to $\mathbb{Q}$, by

$$\left.\frac{d\mathbb{P}}{d\mathbb{Q}}\right|_{\mathcal{F}_t} = \frac{\mathbb{Q}[\sigma > T|\mathcal{F}_t]}{\mathbb{Q}[\sigma > T]} := M_t. \tag{1}$$

Under the new measure, $\mathbb{P}(\sigma > T) = \mathbb{E}^{\mathbb{Q}}(M_T 1_{\sigma>T}) = 1$.

Theorem 3.1 *Assume that the following conditions hold*

- *The risky asset process S is a locally bounded semimartingale which satisfies NFLVR under $\mathbb{Q}$.*
- *The filtration $\mathbb{F}$ is quasi-left continuous.*
- *σ is a predictable stopping time such that for any stopping time θ,*

$$\mathbb{E}^{\mathbb{Q}}[1_{\sigma>T}|\mathcal{F}_\theta] > 0, \mathbb{Q} - a.s. \qquad on\ \{\sigma > \theta\}.$$

Then the $(\mathbb{P}, S)$-market satisfies NUPBR. Given a $\mathcal{F}_T$-measurable claim $f \geq 0$, we have

$$SP_+^{\mathbb{P}}(f) = SP_+^{\mathbb{Q}}(f1_{\sigma>T}).$$

In addition, if

$$SP_+^{\mathbb{Q}}(1_{\sigma>T}) = \sup_{\bar{\mathbb{Q}}\in ELMM(\mathbb{Q},S)} \mathbb{E}^{\bar{\mathbb{Q}}}[1_{\sigma>T}] < 1,$$

then the $(\mathbb{P}, S)$-market admits optimal arbitrage.

Proof. See in Chau and Tankov [1]. □

This construction has the following economic interpretation. Consider an event (E), such as the default of a company or a sovereign state, whose occurence is characterized by a stopping time σ. Given a planning horizon T, we are interested in the occurence of this event (E) before the planning horizon. Suppose that the market agents have common anticipations of the probability of future scenarios, which correspond to the arbitrage-free probability measure $\mathbb{Q}$, and that under this probability, the event (E) has nonzero probabilities of occuring both before and after the planning horizon. Consider now an informed economic agent who believes that the event (E) will not happen before the planning horizon T. For instance, the agent may believe that the company or the state in question will be bailed out in case of potential default. Our informed agent may then want to construct an alternative model $\mathbb{P}$, in which the arbitrage opportunity due to mispricing may be exploited and the arbitrage strategy may be constructed in a rigorous manner.

4 A complete market example

Let $W^{\mathbb{Q}}$ be a Brownian motion and let $\mathbb{F}$ be its completed natural filtration. We assume that the price of a risky asset evolves as follows

$$S_t = 1 + W_t^{\mathbb{Q}}$$

and define a predictable stopping time by $\sigma = \inf\{t > 0 : S_t \leq 0\}$. Using the law of infimum of Brownian motion, we get

$$\mathbb{Q}[\sigma > T] = \mathbb{Q}[(W^{\mathbb{Q}})_T^* > -1] = 1 - 2\mathcal{N}\left(-\frac{1}{\sqrt{T}}\right) > 0,$$

where $\mathcal{N}$ denotes the standard normal distribution function. Next, by Markov property we compute

$$\mathbb{E}^{\mathbb{Q}}[1_{\sigma>T}|\mathcal{F}_t] = \mathbb{Q}[(W^{\mathbb{Q}})_T^* > -1|\mathcal{F}_t] = \begin{cases} 0 & \text{on } \sigma \leq t \\ 1 - 2\mathcal{N}\left(-\frac{S_t}{\sqrt{T-t}}\right) > 0 & \text{on } \sigma > t. \end{cases} \tag{2}$$

Hence, on $\{\tau > t\}$, we obtain $\mathbb{E}^{\mathbb{Q}}[1_{\sigma>T}|\mathcal{F}_t] > 0$. This means that the construction of Section 3 applies and we may define a new measure $\mathbb{P}$ via (1). Since the $(\mathbb{Q}, S)$-market is complete and $ELMM(\mathbb{Q}, S) = \{\mathbb{Q}\}$, the superhedging price of the claim $1_{\sigma>T}$ is

$$\mathbb{Q}[\sigma > T] = 1 - 2\mathcal{N}\left(-\frac{1}{\sqrt{T}}\right) < 1,$$

which means that the $\mathbb{P}$-market admits optimal arbitrage. Applying the Itô formula to (2), we get the martingale representation:

$$\mathbb{E}^{\mathbb{Q}}[1_{\sigma>T}|\mathcal{F}_t] = \mathbb{Q}[\sigma > T] + \sqrt{\frac{2}{\pi}} \int_0^{\sigma\wedge t} \frac{1}{\sqrt{T-s}} e^{-\frac{S_s^2}{2(T-s)}} dW_s^{\mathbb{Q}}. \tag{3}$$

Therefore,

$$H_t = \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{T-t}} e^{-\frac{S_t^2}{2(T-t)}} \mathbf{1}_{t \leq \sigma}$$

is the optimal arbitrage strategy, that is, the hedging strategy for $1_{\sigma > T}$ in the $(\mathbb{Q}, S)$-market as well as the hedging strategy for 1 in the $(\mathbb{P}, S)$-market. Let us now compute the dynamics of S under $\mathbb{P}$. By Girsanov's Theorem (see, e.g., Theorem 41 on page 136 of Protter [7]),

$$W_t^{\mathbb{P}} = W_t^{\mathbb{Q}} - \frac{2}{\mathbb{Q}[\sigma > T]\sqrt{2\pi}} \int_0^{\sigma \wedge t} \frac{1}{M_s} e^{-\frac{S_s^2}{2(T-s)}} \frac{1}{\sqrt{T-s}} ds$$

is a $\mathbb{P}$-Brownian motion. The dynamics of S under $\mathbb{P}$ are therefore given by

$$S_t = 1 + W_t^{\mathbb{P}} + \frac{2}{\mathbb{Q}[\sigma > T]\sqrt{2\pi}} \int_0^{\sigma \wedge t} \frac{e^{-\frac{S_s^2}{2(T-s)}}}{M_s \sqrt{T-s}} ds \tag{4}$$

$$= 1 + W_t^{\mathbb{P}} + \sqrt{\frac{2}{\pi}} \int_0^{\sigma \wedge t} \frac{1}{1 - 2\mathcal{N}\left(-\frac{S_s}{\sqrt{T-s}}\right)} \frac{e^{-\frac{S_s^2}{2(T-s)}}}{\sqrt{T-s}} ds. \tag{5}$$

References

[1] Chau, H. N. and Tankov, P., *Market models with optimal arbitrage.* In `arXiv:1312.4979` (2013).

[2] Delbaen, F. and Schachermayer, W., *A general version of the fundamental theorem of asset pricing.* Mathematische Annalen 300/1 (1994), 463–520.

[3] Delbaen, F. and Schachermayer, W., *The existence of absolutely continuous local martingale measures.* The Annals of Applied Probability 5/4 (1995), 926–945.

[4] Jacod, J. and Shiryaev, A. N., "Limit Theorems for Stochastic Processes". Springer, 2nd edition, 2002.

[5] Karatzas, I. and Kardaras, C., *The numéraire portfolio in semimartingale financial models.* Finance and Stochastics 11/4 (2007), 447–493.

[6] Kardaras, C., *Market viability via absence of arbitrage of the first kind.* Finance and Stochastics 16/4 (2012), 651–667.

[7] Protter, P., "Stochastic Integration and Differential Equations". Springer, 2nd edition, 2003.

Shape optimization and polyharmonic operators

Davide Buoso (*)

Shape optimization has become a very important research area in the last decades. The "fundamental problem" of shape optimization is

$$\min_{\Omega\in\mathcal{C}} F(\Omega),$$

where F is an appropriate functional depending on the domain, and $\mathcal{C}$ is a suitable class of domains. Typically, F represents a cost which we want to minimize. As an example, we may think of the construction of an airplane or of a train. In this case, F will take into account the production costs together with the resistance of the air, while Ω represents the shape of the airplane or of the train. Clearly, the choice of Ω cannot be completely free, otherwise we will end up with a segment, which is not an acceptable answer. For this reason, we add the constraint $\Omega \in \mathcal{C}$, where with $\mathcal{C}$ we will represent the family of all domains containing at least a certain compact K, which will be thought of as the allowance space.

A more specific problem is the so called eigenvalue shape optimization problem, namely

$$\min_{\Omega\in\mathcal{C}} F(\lambda_1(\Omega),\dots,\lambda_k(\Omega)),$$

where $\lambda_1(\Omega),\dots,\lambda_k(\Omega)$ are the first k eigenvalues of a (differential) operator $\mathcal{L}$ acting on $V \subseteq W^{l,p}(\Omega)$. In the literature, the most studied operator is the Laplacian $\mathcal{L} = -\Delta$ acting on $H_0^1(\Omega)$,

$$\begin{cases} -\Delta u = \lambda u, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases}$$

with $\mathcal{C}$ the family of all open sets Ω in $\mathbb{R}^N$ with fixed (finite) measure, namely $|\Omega| = c$, and $F(\lambda_1(\Omega),\dots,\lambda_k(\Omega)) = \lambda_k(\Omega)$. Note that this problem (also called Helmholtz problem) is related to the vibrations of a fixed membrane.

Regarding λ_1, we have the following result, which was proved indipendently by Faber [8] and Krahn [11].

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: dbuoso@math.unipd.it. Seminar held on March 26th, 2014.

Theorem 1 *Let Ω be an open set in $\mathbb{R}^N$, and let B be a ball such that $|B| = |\Omega|$. Then*

$$\lambda_1(\Omega) \geq \lambda_1(B).$$

As for λ_2, we have the following Theorem, which has been proved independently by Szegö [21] and Hong [10]. The proof was however hidden also inside a paper by Krahn [12].

Theorem 2 *Let Ω be a bounded open set in $\mathbb{R}^N$, and let $\tilde{\Omega}$ be the union of two disjoint identical balls such that $|\Omega| = |\tilde{\Omega}|$. Then*

$$\lambda_2(\Omega) \geq \lambda_2(\tilde{\Omega}).$$

Regarding higher order eigenvalues (namely λ_k with $k > 2$), Wolf and Keller [25] (see also [19]) proved that balls are no longer related to minimizers. Anyway, if we turn to different boundary conditions, we can see how they are still interesting. Regarding the Neumann problem,

$$\begin{cases} -\Delta u = \lambda u, & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = 0, & \text{on } \partial\Omega, \end{cases}$$

which is related to the vibrations of a free membrane, we have the following result, proved by Szegö in dimension $N = 2$ [21] and soon generalized by Weinberger [23].

Theorem 3 *Let Ω be a Lipschitz open set in $\mathbb{R}^N$ of finite measure, and let B be a ball in $\mathbb{R}^N$ such that $|B| = |\Omega|$. Then*

$$\lambda_2(\Omega) \leq \lambda_2(B).$$

For the Robin problem,

$$\begin{cases} -\Delta u = \lambda u, & \text{in } \Omega, \\ \alpha u + (1-\alpha)\frac{\partial u}{\partial \nu} = 0, & \text{on } \partial\Omega, \end{cases}$$

which is related to the vibrations of an elastically supported membrane, we have the following result, proved by Bossel [2] in dimension $N = 2$, and recently generalized by Daners [7].

Theorem 4 *Let Ω be an open set in $\mathbb{R}^N$ of finite measure, and let B be a ball in $\mathbb{R}^N$ such that $|B| = |\Omega|$. Then, for any $\alpha \in]0, 1]$,*

$$\lambda_1(\Omega) \geq \lambda_1(B).$$

Finally, let us consider a slightly different problem, i.e. the Laplace operator with Steklov boundary conditions,

$$\begin{cases} -\Delta u = 0, & \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = \lambda u, & \text{on } \partial\Omega. \end{cases}$$

This problem is related to the vibrations of a free membrane whose mass is dislocated on the boundary only. The following theorem was proved by Weinstock [24] in dimension $N = 2$, and later generalized by Brock [3].

Theorem 5 *Let Ω be an open set in $\mathbb{R}^N$ of finite measure, and let B be a ball in $\mathbb{R}^N$ such that $|B| = |\Omega|$. Then*

$$\lambda_2(\Omega) \leq \lambda_2(B).$$

Next we present the following result, proved by Buttazzo and Dal Maso [6]. This is a very general result, and it can be easily adapted to the case $F(\Omega) = \lambda_k(\Omega)$.

Theorem 6 *Let D be a bounded open set in $\mathbb{R}^N$, and $\mathcal{C}(D)$ be the family of all quasi-open subsets of D. Let $F : \mathcal{C}(D) \to \overline{\mathbb{R}}$ be a l.s.c. map (w.r.t. γ-convergence) and decreasing (i.e. $F(A) \geq F(B)$ if $A \subset B$). Then, for every $c \in [0, |D|]$, the minimum*

$$\min\{F(\Omega) : \Omega \in \mathcal{C}(D), |\Omega| = c\}$$

is achieved.

This Theorem essentially says that the problem

$$\min \lambda_k(\Omega)$$

has a solution for any $k \in \mathbb{N}$, but the minimum is taken in the class of quasi-open sets contained in a fixed "box" D. The notion of quasi-open set is a generalization of that of open set, and in this sense it is bad because we would expect minimizers to be "nice" sets. Regarding the box D, there is a recent work by Mazzoleni and Pratelli [16] where they show that the minimizers of $\lambda_k(\Omega)$ can be taken without the restriction $\Omega \in D$.

A related, although different problem is the criticallity of domains under volume constraint. The aim here is to compute Hadamard-type formulas (i.e. derivatives of the eigenvalues) and apply Lagrange Multipliers Theorem. Clearly, the family of open sets in $\mathbb{R}^N$ does not enjoy a linear structure, therefore we introduce the following class

$$\Phi(\Omega) = \{\phi \in (C^1(\overline{\Omega}))^N : \min_{\overline{\Omega}} |\det \nabla\phi| > 0\},$$

where Ω will be thought of as a fixed domain. We will consider the eigenvalues λ_j on $\phi(\Omega)$ as functions of ϕ:

$$\phi \mapsto \lambda_j[\phi].$$

With this definition, Lamberti and Lanza de Cristoforis proved the following, valid both for the Dirichlet problem [13] and for the Neumann problem [14].

Theorem 7 *Let Ω be a bounded domain in $\mathbb{R}^N$ of class C^1, F a finite subset of $\mathbb{N}$. Fix $t > 0$. Let*

$$\mathcal{A}_\Omega[F] = \{\phi \in \Phi(\Omega) : \lambda_l[\phi] \notin \{\lambda_j[\phi] : j \in F\}\ \forall l \notin F\}.$$

Let $s \in \{1, \dots, |F|\}$. Then the function $\Lambda_{F,s}[\cdot]$ from $\mathcal{A}_\Omega[F]$ to $\mathbb{R}$ defined by

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

is real analytic.

The previous Theorem deals with elementary symmetric functions of the eigenvalues: this is because multiple eigenvalues are continuous but not differentiable. This is due to the fact that, around the multiplicity, bifurcation phenomena typically arise. Symmetric functions then allow to avoid non-differentiability situations.

Now that we know that the functions $\Lambda_{F,s}$ are analytic, we can compute their derivatives. We define

$$\Theta_\Omega[F] = \{\phi \in \mathcal{A}_\Omega[F] : \lambda_{j_1}[\phi] = \lambda_{j_2}[\phi] \ \forall j_1, j_2 \in F\}.$$

We have the following result (see [13, 14]).

Theorem 8 *Let Ω be a bounded domain in $\mathbb{R}^N$ of class C^1, F a finite subset of $\mathbb{N}$. Fix $t > 0$. Moreover, let $\tilde{\phi} \in \Theta_\Omega[F]$ be such that $\partial\tilde{\phi}(\Omega) \in C^2$. Then*

$$d|_{\phi=\tilde{\phi}} \Lambda_{F,s}[\psi] = \lambda_F^s[\tilde{\phi}] \binom{|F|-1}{s-1} \sum_{l=1}^{|F|} \int_{\partial\tilde{\phi}(\Omega)} H(v_l)(\psi \circ \tilde{\phi}^{-1}) \cdot \nu d\sigma,$$

for all $\psi \in (C^1(\Omega))^N$, where $H(v_l) = \left(\frac{\partial v_l}{\partial \nu}\right)^2$ for the Dirichlet problem, $H(v_l) = |\nabla v_l|^2 - \lambda_F v_l^2$ for the Neumann problem.

We recall that we are interested in problems like

$$\min_{V[\phi]=\text{const}} \Lambda_{F,s}[\phi] \quad \text{or} \quad \max_{V[\phi]=\text{const}} \Lambda_{F,s}[\phi],$$

where $V[\phi]$ is the N-dimensional Lebesgue measure of $\phi(\Omega)$, and that if $\tilde{\phi}$ is a minimizer or a maximizer for $\Lambda_{F,s}$, then it is a critical point for the map

$$\phi \mapsto \Lambda_{F,s}[\phi]$$

under volume constraint. Using Lagrange Multiplier Theorem it is easy to prove the following (see [15])

Theorem 9 *Let Ω be a bounded domain in $\mathbb{R}^N$ of class C^1, F a finite subset of $\mathbb{N}$. Let $\tilde{\phi} \in \Theta_\Omega[F]$ be such that $\partial\tilde{\phi}(\Omega) \in C^2$. Moreover let $\{v_l\}_{l \in F}$ be an orthonormal basis in $H_0^1(\Omega)$ (for the Dirichlet problem) or $H^1(\Omega)$ (for the Neumann problem) of the eigenspace related to $\lambda_F[\tilde{\phi}]$. Then $\tilde{\phi}$ is a critical point for $\Lambda_{F,s}$, for any $s = 1, \dots, |F|$, with volume costraint if and only if there exists a constant $c \in \mathbb{R}$ such that*

$$\sum_{l=1}^{|F|} H(v_l) = c \qquad \text{on } \partial\tilde{\phi}(\Omega). \tag{1}$$

As we said, balls are no longer minimizers when we consider eigenvalues of high order. Anyway, the following result shows that they still play a fundamental role (see [15]).

Theorem 10 *Let the same assumptions of the previous theorem hold. If $\tilde{\phi}(\Omega)$ is a ball, then condition (1) is satisfied. Thus, $\tilde{\phi}$ is a critical point for $\Lambda_{F,s}$, for any $s = 1, \dots, |F|$, with volume costraint.*

Now we want to consider a similar problem. When modelling the vibration of a clamped plate, we get to the problem

$$\begin{cases} \Delta^2 u = \lambda u, & \text{in } \Omega, \\ u = \frac{\partial u}{\partial \nu} = 0, & \text{in } \partial\Omega. \end{cases}$$

This is very close to the Dirichlet problem for the Laplace operator. There also is a conjecture, named after Lord Rayleigh (see [20]), claiming that the Faber-Krahn inequality should hold for this problem as well. The conjecture has been proved in dimension $N = 2$ by Nadirashvili [18] and soon generalized in dimension $N = 2, 3$ by Ashbaugh and Benguria [1], while the general case remains an open problem.

Theorem 11 *Let $N = 2, 3$, and let Ω be an open set in $\mathbb{R}^N$ of finite measure. If B is a ball such that $|B| = |\Omega|$, then*

$$\lambda_1(\Omega) \geq \lambda_1(B).$$

A result concerning regularity of the solution was given by Mohr [17].

Theorem 12 *If there exists a minimizer Ω of class C^2 for λ_1, then Ω is a ball.*

A similar problem turns out when studing the buckling of a plate

$$\begin{cases} \Delta^2 u = -\lambda \Delta u, & \text{in } \Omega, \\ u = \frac{\partial u}{\partial \nu} = 0, & \text{in } \partial\Omega. \end{cases}$$

Eigenvalue optimization results for this problem are very little. We cite the following from [22], which is valid also for the clamped plate.

Theorem 13 *Let Ω be an open set in $\mathbb{R}^N$ of finite measure such that the first eigenfunction is positive. Then*

$$\lambda_1(\Omega) \geq \lambda_1(B),$$

where B is a ball such that $|B| = |\Omega|$.

The following Theorem is due to Weinberger and Willms, but it has never been published (see [9]).

Theorem 14 *Let $\Omega \subset \mathbb{R}^2$ be a minimizer for λ_1 of class $C^{2,\alpha}$. Then Ω is a disk.*

Now, we want to use the same arguments of Lamberti and Lanza de Cristoforis for the clamped plate and the buckled plate. For the sake of generality, we consider the problem

$$\begin{cases} (-\Delta)^n u = \lambda(-\Delta)^m u, & \text{in } \Omega, \\ u = \frac{\partial u}{\partial \nu} = \cdots = \frac{\partial^{n-1} u}{\partial \nu^{n-1}} = 0, & \text{in } \partial\Omega, \end{cases}$$

where $n, m \in \mathbb{N}$, with $m < n$, and we introduce the following class of perturbations

$$\Phi^n(\Omega) = \{\phi \in (C^n(\overline{\Omega}))^N : \min_{\overline{\Omega}} |\det \nabla\phi| > 0\}.$$

We have the following results (see [4]).

Theorem 15 *Let Ω be a bounded domain in $\mathbb{R}^N$ of class C^1, F a finite subset of $\mathbb{N}$. Fix $t > 0$. Let*

$$\mathcal{A}_\Omega[F] = \{\phi \in \Phi^n(\Omega) : \lambda_l[\phi] \notin \{\lambda_j[\phi] : j \in F\} \ \forall l \notin F\}.$$

Let $s \in \{1, \dots, |F|\}$. Then the function $\Lambda_{F,s}[\cdot]$ from $\mathcal{A}_\Omega[F]$ to $\mathbb{R}$ defined by

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

is real analytic.

Theorem 16 *Let Ω be a bounded domain in $\mathbb{R}^N$ of class C^1, F a finite subset of $\mathbb{N}$. Fix $t > 0$. Moreover, let $\tilde{\phi} \in \Theta_\Omega[F]$ be such that $\partial\tilde{\phi}(\Omega) \in C^{2n}$. Then*

$$d|_{\phi=\tilde{\phi}} \Lambda_{F,s}[\psi] = \lambda_F^s[\tilde{\phi}] \binom{|F|-1}{s-1} \sum_{l=1}^{|F|} \int_{\partial\tilde{\phi}(\Omega)} \left(\frac{\partial^n v_l}{\partial \nu^n}\right)^2 (\psi \circ \tilde{\phi}^{-1}) \cdot \nu d\sigma,$$

for all $\psi \in (C^1(\Omega))^N$.

Theorem 17 *Let Ω be a bounded domain in $\mathbb{R}^N$ of class C^1, F a finite subset of $\mathbb{N}$. Let $\phi \in \Theta_\Omega[F]$ be such that $\partial\tilde{\phi}(\Omega) \in C^{2n}$. Moreover let $\{v_l\}_{l \in F}$ be an orthonormal basis in $H_0^n(\Omega)$ of the eigenspace related to $\lambda_F[\tilde{\phi}]$. Then $\tilde{\phi}$ is a critical point for $\Lambda_{F,s}$, for any $s = 1, \dots, |F|$, with volume costraint if and only if there exists a constant $c \in \mathbb{R}$ such that*

$$\sum_{l=1}^{|F|} \left(\frac{\partial^n v_l}{\partial \nu^n}\right)^2 = c \text{ on } \partial\tilde{\phi}(\Omega). \tag{2}$$

Theorem 18 *Let the same assumptions of the previous theorem hold. If $\tilde{\phi}(\Omega)$ is a ball, then condition (2) is satisfied. Thus, $\tilde{\phi}$ is a critical point for $\Lambda_{F,s}$, for any $s = 1, \dots, |F|$, with volume costraint.*

As we have already said, the equation

$$\Delta^2 u = \lambda u$$

is related to the study of the vibrations of a plate. Depending on the boundary conditions, we get different problems.

- Clamped plate (Dirichlet bc., i.e. $u \in H^2_0$):

$$u = 0,\ \frac{\partial u}{\partial \nu} = 0 \text{ on } \partial\Omega.$$

- Free plate (Neumann bc., i.e. $u \in H^2$):

$$\frac{\partial^2 u}{\partial \nu^2} = 0,\ \frac{\partial \Delta u}{\partial \nu} + \mathrm{div}_{\partial\Omega}\left(\nu^t D^2 u\right) = 0 \text{ on } \partial\Omega.$$

- Hinged plate (intermediate bc., i.e. $u \in H^2 \cap H^1_0$):

$$u = 0,\ \frac{\partial^2 u}{\partial \nu^2} = 0 \text{ on } \partial\Omega.$$

A similar discussion can be made for all those problems. We give here just the main result (see [4] for the clamped plate, [5] for the hinged plate; the free plate case will appear soon).

Theorem 19 *For all the previous problems, if $\tilde{\phi}(\Omega)$ is a ball, then $\tilde{\phi}$ is a critical point for $\Lambda_{F,s}$, for any $s = 1, \dots, |F|$, with volume costraint.*

References

[1] M.S. Ashbaugh, R.D. Benguria, *On Rayleigh's conjecture for the clamped plate and its generalization to three dimensions.* Duke Math. J. 78 (1995), no. 1, 1–17.

[2] M.H. Bossel, *Membranes élastiquement liées: extension du théorème de Rayleigh-Faber-Krahn et de l'inégalité de Cheeger.* C. R. Acad. Sci. Paris Série I Math., 302 (1986), no. 1, 47–50.

[3] F. Brock, *An isoperimetric inequality for eigenvalues of the Stekloff problem.* Z. Angew. Math. Mech., 81 (2001), no.1, 69–71.

[4] D. Buoso, P.D. Lamberti, *Eigenvalues of polyharmonic operators on variable domains.* ESAIM: COCV, 19 (2013), 1225–1235.

[5] D. Buoso, P.D. Lamberti, *Shape deformation for vibrating hinged plates.* Mathematical Methods in the Applied Sciences, 37 (2014), 237–244.

[6] G. Buttazzo, G. Dal Maso, *An Existence Result for a Class of Shape Optimization Problems.* Arch. Rational Mech. Anal., 122 (1993), 183–195.

[7] D. Daners, *A Faber-Krahn inequality for Robin problems in any space dimension.* Math. Ann. 335 (2006), no. 4, 767–785.

[8] G. Faber, *Beweis, dass unter allen homogenen Membranen von gleicher Fläche und gleicher Spannung die kreisförmige den tiefsten Grundton gibt.* Sitz. Ber. Bayer. Akad. Wiss. (1923), 169-172.

[9] A. Henrot, "Extremum problems for eigenvalues of elliptic operator". Frontiers in Mathematics, Birkhäuser Verlag, Basel, 2006.

[10] I. Hong, *On an inequality concerning the eigenvalue problem of membrane.* Kodai Math. Sem. Rep. (1954), 113–114.

[11] E. Krahn, *Über eine von Rayleigh formulierte Minimaleigenschaft des Kreises.* Math. Ann. 94 (1924), 97–100.

[12] E. Krahn, *Über Minimaleigenschaften der Kugel in drei un mehr Dimensionen.* Acta Comm. Univ. Dorpat., A9 (1926), 1–44.

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

[14] P.D. Lamberti, M. Lanza de Cristoforis, *A real analyticity result for symmetric functions of the eigenvalues of a domain-dependent Neumann problem for the Laplace operator.* Mediterr. J. Math. 4 (2007), no. 4, 435–449.

[15] P.D. Lamberti, M. Lanza de Cristoforis, *Critical points of the symmetric functions of the eigenvalues of the Laplace operator and overdetermined problems.* J. Math. Soc. Japan 58 (2006), no. 1, 231–245.

[16] D. Mazzoleni, *Existence of minimizers for spectral problems.* J. Math. Pures Appl. (9) 100 (2013), no. 3, 433–453.

[17] E. Mohr, *Über die Rayleighsche Vermutung: unter allen Platten von gegebener Fläche und konstanter Dichte und Elastizität hat die kreisförmige den tiefsten Grundton.* Ann. Mat. Pura Appl. (4) 104 (1975), 85–122.

[18] N.S. Nadirashvili, *Rayleigh's conjecture on the principal frequency of the clamped plate.* Arch. Rational Mech. Anal. 129 (1995), no. 1, 1–10.

[19] E. Oudet, *Numerical minimization of eigenmodes of a membrane with respect to the domain.* ESAIM: COCV, 10 (2004), 315–330.

[20] J.W.S. Rayleigh, "The theory of sound". Dover Pub. New York, 1945 (republication of the 1894/96 edition).

[21] G. Szegö, *Inequalities for certain eigenvalues of a membrane of given area.* J. Rational Mech. Anal. 3 (1954), 343–356.

[22] G. Szegö, *On membranes and plates.* Proc. Nat. Acad. Sci. 36 (1950), 210–216.

[23] H.F. Weinberger, *An isoperimetric inequality for the $N-$dimensional free membrane problem.* J. Rat. Mech. Anal., no 5 (1956), 633–636.

[24] R. Weinstock, *Inequalities for a classical eigenvalue problem.* J. Rational Mech. Anal., 3 (1954), 745–753.

[25] S.A. Wolf, J.B. Keller, *Range of the first two eigenvalues of the Laplacian.* Proc. Roy. Soc. London Ser. A 447, no. 1930 (1994), 397–412.

Geometric modeling and splines: state of the art and outlook

Michele Antonelli (*)

Abstract. We will give an introductory presentation of the research field of geometric modeling and its applications, with specific attention to the use of splines for the representation of curves and surfaces. In particular, we will start by introducing basic notions of geometric modeling leading up to the definition of splines, which are piecewise functions with prescribed smoothness at the locations where the pieces join. Splines will be exploited for the representation of parametric curves and surfaces, and we will present their application in the context of computer-aided geometric design for shape description by means of approximation and interpolation methods. Finally, we will discuss some open problems in this topic and we will sketch some recent approaches for addressing them.

1 Introduction

Geometric modeling is the branch of applied mathematics that studies mathematical and computational methods for the description of geometric shapes and objects for use in computer graphics, manufacturing, or analysis. Since today most geometric modeling is done with computers and for computer-based applications, we also refer to this discipline as *computer-aided geometric design* (CAGD, for short).

CAGD is inherently interdisciplinary: in fact, it draws upon the fields of geometry, numerical analysis, approximation theory, data structures and computer algebra. Moreover, many applied fields have emerged because of the advancements in geometric design, and in turn the motivation for further development of geometric design techniques has originated from several sources during the years. The earliest influences came from mechanical engineering and the emerging field of computer-aided design (CAD), as well as from shipbuilding, aeronautics, and numerical control of milling machines for computer-aided manufacturing. The solution to the new problems that arose in these contexts involved results from approximation theory, differential and computational geometry, but also from computer science and graphics.

For these reasons, it is not surprising that geometric modeling has a wide range of

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: antonelm@math.unipd.it . Seminar held on April 9th, 2014.

applications. The main one is in the context of computer-aided design and manufacturing for the automotive, aerospace and naval industry, as well as for prosthetic design and image processing in the biomedical sector, and for the conservation and restoration of cultural heritage. Also other applied technical fields, such as civil and mechanical engineering, architecture, and geology, make a wide use of models realized with CAD systems. Another flourishing application is in computer graphics for gaming and animation. All these contexts require handling complicated geometric entities, and geometric modeling gives the mathematical tools for the representation of these objects.

It is worth pointing out that the shapes considered in the context of CAGD are mostly two- or three-dimensional (plane and space curves, surfaces, solids), but many of the tools and principles can be applied to geometric objects of any finite dimension.

We suggest [11, 7, 8] as valuable references for a detailed presentation of the most relevant topics in CAGD and their application.

1.1 A brief historical retrospective

Before presenting some of the mathematics involved in the context of geometric modeling, we briefly review the history of this field (see [8, Chapter 1] for an extended overview).

The earliest recorded use of curves in a manufacturing context dates back to Ancient Romans for the purpose of shipbuilding. At that time, the ribs of a ship (which are wooden planks emanating from the keel) were produced based on templates that could be reused many times. Thus the basic geometry of a vessel could be stored and did not have to be recreated every time.

Similar techniques were used in the following centuries, but during the Renaissance more rigorous constructions started to appear in the context of drafting. In particular, Italian naval architects adopted design methods that involved conic sections in a systematic way, such as tangent-continuous circular arcs.

These design techniques were refined through the centuries and culminated in the invention of some tools for technical drawing, such as French curves and the so-called mechanical splines. In particular, French curves are templates composed of different pieces of conics and spirals. Instead, mechanical splines are thin flexible strips (originally made of wood, metal or even plastic in more recent times) that can be bent into optimal shapes to draw smooth curves. The splines are held in place by metal weights, which are called "ducks" because of their shape. The elasticity of the spline material, combined with the constraint given by the weights, causes the strip to take the shape that minimizes the strain energy required for bending between the fixed locations. In this way, the spline provides an interpolation of the given points by means of the smoothest possible curve, which is both physically optimal and aesthetically pleasing.

Near the middle of the twentieth century, further progress was made in the context of drafting techniques, in particular in the field of aircraft design. Liming proposed that the conic sections that were exploited to streamline the overall design of the outside fuselage of airplanes should have been stored in terms of numbers, parameters, and numerical algorithms for their construction.
In the same years, Schoenberg [15] gave the first mathematical definition of basis splines

defined over uniform knots, and shortly after Curry [5] extended this definition over nonuniform knots. It is interesting to note that many quantities that started to appear in this new theory were named after terms used when referring to the mechanical spline device.

Moreover, in the Fifties the advent of numerical control of milling machines drew attention to the need for new blueprint-to-computer concepts in design, and the development of suitable hardware and software dates back to this period.

In the next decade, computer-aided design started spreading in the automotive and aeronautic industry around the world as the first CAD software appeared. In addition, scientific investigations in the field went on.

However, all the research efforts in those years were undertaken mainly in isolation and gave rise to a quite fragmented scenery. This situation changed in the Seventies, where the confluence of the different approaches culminated in the creation of a new coherent scientific discipline, CAGD. In particular, the CAGD Conference at the University of Utah in 1974 may be regarded as the founding event of the field. In the same year, Gordon and Riesenfeld [10] introduced parametric spline curves, which in the context of geometric design are much more important than spline functions, which instead have a relevant role in approximation theory.

Naturally, since then research has progressed and we conclude this brief overview by just mentioning some application and new trends of geometric modeling that appeared in the last twenty years. In particular, in the Nineties CAGD methods started to be applied in an extensive way for computer graphics and animation for entertainment. Moreover, the very last decade has seen the birth of isogeometric analysis [12], a new computational approach that aims at integrating the power of finite element analysis into conventional spline-based CAD systems. In this way, it is possible to design, test and adjust a model in one go, bypassing the time-consuming step of conversion between the different forms of representation used by the different software that in the usual workflow takes care of each phase of the design and production process.

1.2 Splines in geometric modeling

In the above overview we have mentioned the fact that in the middle of the last century a mathematical counterpart to a mechanical spline was introduced: a spline curve, which is a composite curve made of different polynomial pieces that join with a prescribed continuity and minimize certain functionals.
Apart from their many properties, these curves can be easily exploited for the interpolation of a set of points, and therefore they were largely adopted for design in the automotive and aerospace industry, where computer software was replacing the old drafting techniques on paper. The name "spline" is due to the fact that the functions proposed in the first works minimized a functional similar to the physical properties of mechanical splines. However, the meaning of the term "spline curve" has undergone a subtle change in the years. In fact, instead of referring to curves that minimize certain functionals, spline curves are now mostly thought of just as piecewise curves with certain smoothness properties.

Splines quickly gained popularity among the design community thanks to their flexibility, computational efficiency and because they provide an intuitive way for defining a

shape, as we will see. In particular, a generalization of the Schoenberg's polynomial B-splines, called Non-Uniform Rational B-Splines (NURBS) has become the standard form of representation for curves and surfaces in the CAD industry. Their success is due to the fact that they offer a unified representation of spline and classical conic geometries: in fact, every conic admits a piecewise rational polynomial representation.

Regarding references, the monograph by Schumaker [16] is recognized as a fundamental resource for learning the basic theory of splines and some early generalizations.

In the following sections we will introduce some basic notions of geometric modeling leading up to the definition of splines, and we will present some of their properties and applications. Finally, we will discuss some open problems in CAGD and we will sketch some recent approaches for addressing them.

2 Towards the definition of splines

In this section we present some mathematical tools that are essential for the representation of shapes in geometric design.

2.1 Bernstein polynomials

As a first step, we recall the definition of the Bernstein polynomial basis, which is a basic tool in geometric design because of its many properties and the computational benefits that it offers. However, this basis originally appeared in a different context, that of approximation theory. In fact, it was originally introduced by Bernstein [4] 100 years ago to provide a constructive proof of Weierstrass approximation theorem, which states that polynomials can uniformly approximate any continuous function over a closed interval. More precisely:

Theorem 1 *Given any continuous function f on an interval $[a, b]$ and a tolerance $\epsilon > 0$, there exists a polynomial p_n of sufficiently high degree n such that $|f(x) - p_n(x)| \leq \epsilon$ for all $x \in [a, b]$.*

This result represented a significant advance over using the Taylor expansion to generate polynomial approximations of a function, but the first proofs developed by Weierstrass and others were existential, rather than constructive, and relied on analytic limit arguments, rather than concrete algebraic processes. On the contrary, the main feature of Bernstein's proof was the explicit construction of a sequence of polynomials approaching the given function more closely at every point of the interval as the degree increases.

In particular, the approximating Bernstein polynomial of degree n associated with the function f is defined as follows. Note that it is not restrictive to work in the unit interval $[0, 1]$ instead of $[a, b]$, since we can exploit a linear change of variable to go back to the considered interval.

Definition 1 Let f be a continuous function on $[0, 1]$. The *Bernstein polynomial associ-*

ated with f is given by

$$p_n(t) := \sum_{i=0}^{n} f\left(\frac{i}{n}\right) B_{i,n}(t), \qquad t \in [0,1],$$

where $B_{i,n}$, $i = 0, \ldots, n$, are the *Bernstein polynomial basis functions* of degree n on $t \in [0,1]$:

$$B_{i,n}(t) := \binom{n}{i}(1-t)^{n-i}\, t^i, \qquad i = 0, \ldots, n.$$

Figure 1 shows the basis functions for some values of n.

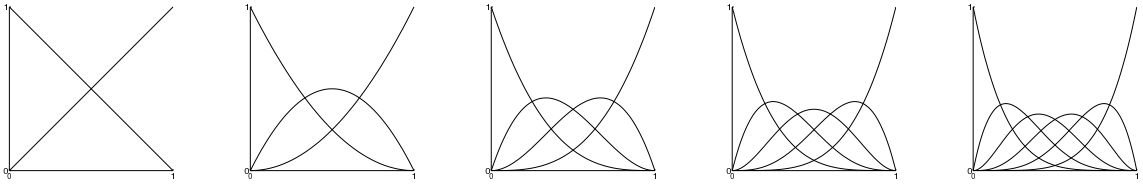


Figure 1: Bernstein basis functions for some degrees n (here $n = 1, 2, 3, 4, 5$)

One can prove that the Bernstein functions enjoy many properties:

(a) $B_{i,n}$, $i = 0, \ldots, n$, form a basis of the space of polynomials of degree at most n;

(b) $B_{i,n}(0) = \delta_{i,0}$, $B_{i,n}(1) = \delta_{i,n}$, $B_{i,n}^{(k)}(0) = 0$, $k = 1, \ldots, i-1$, and $B_{i,n}^{(k)}(1) = 0$, $k = 1, \ldots, n-i-1$;

(c) symmetry: $B_{n-i,n}(1-t) = B_{i,n}(t)$;

(d) non-negativity: $B_{i,n}(t) \in [0,1]$ for $t \in [0,1]$;

(e) partition of unity: $\sum_{i=0}^{n} B_{i,n}(t) \equiv 1$ for $t \in [0,1]$;

(f) unimodality: $B_{i,n}$ has a single maximum on $[0,1]$, at $t = \frac{i}{n}$;

(g) recursion: $B_{i,n+1}(t) = tB_{i-1,n}(t) + (1-t)B_{i,n}$.

In particular, the uniform convergence of the Bernstein approximating polynomials to a given function relies on the properties of non-negativity and partition of unity.
From a computational point of view, the existence of a recurrence relation that allows generating the basis of degree $n+1$ from the basis of degree n gives an efficient way for evaluating the Bernstein functions of a certain degree at a given value. Such a procedure consists in evaluating and combining basis functions of increasing degree up to the desired degree. Moreover, since only convex combinations of values between 0 and 1 are involved, this algorithm is numerically stable.

Unfortunately, the leisurely convergence rate of Bernstein polynomial approximations to continuous functions (see Figure 2(a)) caused them to find no practical application in approximation theory, and thus to languish in obscurity, until the advent of computers. In fact, with the desire to exploit the power of computers for geometric design applications, the Bernstein form began to enjoy a widespread use as a versatile way of constructing and manipulating geometric shapes, leading to further development of basic theory and stable computational methods.

2.2 Bézier curves

In particular, Bézier and de Casteljau[(1)], who were working in the automotive industry during the Sixties, were not concerned with the approximation of given functions, but rather with formulating novel mathematical tools that would allow designers to construct and manipulate in an intuitive way complex shapes using digital computers. This problem was especially critical for the so-called free-form shapes, which did not have a description in terms of a few simple geometric parameters like other classical entities. The motivation was to avoid the expensive process of sculpting clay models to specify the desired shape.

Although a parametric curve or surface is an infinite collection of points, its computer representation must employ just a finite data set. The mapping from the finite set of input values to a continuous locus is achieved by interpreting those values as coefficients for certain basis functions in the parametric variables. Moreover, it is advisable that the coefficients furnish natural "shape handles" that permit the intuitive creation or modification of the geometry in order to satisfy prescribed aesthetic or functional requirements. For this reason, the choice of the basis is fundamental to a successful design scheme. Ultimately, the work of Bézier and de Casteljau lead to the adoption of the Bernstein form, giving rise to what is now called a Bézier curve.

Definition 2 Let a set of *control points* $\boldsymbol{p}_i \in \mathbb{R}^d$, $i = 0, \ldots, n$, be given. The *Bézier curve* described by these points is the parametric curve with polynomial components expressed in terms of the Bernstein basis functions of degree n that has the form

$$\boldsymbol{c}(t) = \sum_{i=0}^{n} \boldsymbol{p}_i B_{i,n}(t), \qquad t \in [0,1]. \tag{1}$$

The control points form the so-called *control polygon*, and it can be used to analyze and manipulate the curve shape in a simple and natural manner.

Figure 2(b) shows a Bézier curve of degree 3 and its four control points.

Bézier curves enjoy many properties that are intimately related to the properties of the underlying Bernstein basis. We briefly describe some of the most relevant ones in the context of geometric modeling, which make the use of Bézier curves very intuitive for design and shape representation:

[(1)] See [14] and [6] respectively for some biographic and scientific sketch, and refer to [9, Section 4] for a concise overview of their work.

(a) endpoints interpolation: the curve interpolates the first and the last control point, and the tangents at the endpoints are given by the first and the last edge of the control polygon;

(b) variation diminution: no straight line may intersect the curve more often than it intersects the control polygon;

(c) convex hull: a Bézier curve is confined within the convex region described by its control polygon;

(d) shape preservation: the control polygon of a Bézier curve expressed in term of the Bernstein basis is optimal, in the sense that it mimics in the best way the overall shape of the curve.

Bézier curves enjoy also some good computational properties, thanks to the numerical stability of the Bernstein form with respect to perturbations of initial data or rounding errors that occur during floating-point calculations. This attribute is very important for the robustness of the geometrical computations performed in CAD systems, where the output consists in geometric models and it is not an end in itself, but it is the starting point for other applications, such as meshing for finite-element analysis or path planning for manufacturing tools, which cannot succeed without accurate and consistent geometrical representations.

Another important feature from the computational point of view is the existence of an efficient algorithm for the evaluation of a Bézier curve $\boldsymbol{c}(t)$ of the form (1) at a given parameter $\bar{t} \in [0, 1]$. It is named after de Casteljau and it is based on the recurrence relation that we have recalled in the previous section for the Bernstein basis functions. The algorithm consists in multiple steps of linear interpolation between pairs of control points, weighted by the given evaluation parameter $\bar{t}$, and each resulting new point is placed on the edge between two old ones. After each step, a certain number of new points is obtained, and the procedure goes on by combining these points, until we get only one new point, inserted at the last step. This point is exactly the sought value $\boldsymbol{c}(\bar{t})$.
Moreover, as a by-product of the algorithm we also get the control points of the two arcs in which the curve is split by the evaluated point as if they were two separate Bézier curves.

In the context of design, it is essential to be able to represent complex shapes in a flexible way. However, if we were to describe this kind of shapes by using only one Bézier curve, we would need a very large number of control points, and thus a very high polynomial degree. Therefore, a better strategy consists in gluing together different Bézier curves of low degree, usually quadratic or cubic, with the desired continuity, thus increasing the flexibility for shape representation without increasing the polynomial degree and the computational cost.
Because of the shape properties recalled above, the continuity conditions required when joining the different curves translate into conditions on the location of control points. As an example, let us consider the case of joining two Bézier curves $\boldsymbol{c}(t) = \sum_{i=0}^{n} \boldsymbol{p}_i B_{i,n}(t)$ and $\boldsymbol{d}(t) = \sum_{i=0}^{n} \boldsymbol{q}_i B_{i,n}(t)$. In particular, to achieve C^0 *continuity* we let the last and the

first control points of the curves coincide:

$$\boldsymbol{p}_n = \boldsymbol{q}_0.$$

Furthermore, G^1 *continuity*, that means tangent continuity, is obtained by aligning the last and the first edge of the control polygons:

$$\boldsymbol{p}_n - \boldsymbol{p}_{n-1} = a\,(\boldsymbol{q}_1 - \boldsymbol{q}_0)\,, \quad a > 0,$$

and C^1 *continuity*, that is the real parametric continuity of the first derivatives of the two curves, additionally requires that the two edges have the same length:

$$\boldsymbol{p}_n - \boldsymbol{p}_{n-1} = \boldsymbol{q}_1 - \boldsymbol{q}_0.$$

Higher orders of continuity require similar conditions which involve a larger number of control points.

An application of connected Bézier curves is the description of the outline of a digital font. In particular, in the PostScript format and in the usual TEX typesetting systems the text font is specified in terms of straight lines and quadratic or cubic Bézier curves.

However, joining pieces and assuring that the desired continuity conditions are satisfied may be a tedious task, even with the support of a computer. In addition, these conditions must be checked again every time that one or more control points are moved by designer to adjust the desired shape. Therefore, a ground-breaking idea was to exploit piecewise-defined functions in the definition of the underlying basis used for the representation, and to embed the smoothness properties directly into the basis. This lead to the definition of the B-spline basis for the space of polynomial splines.

Figure 2: (a) Bernstein polynomials of degrees $n = 10, 30, 100, 300, 1000$ approximating a piecewise linear continuous function; (b) Cubic Bézier curve.

3 Spline functions

Let us consider a closed bounded interval $[a, b] \subset \mathbb{R}$ and define a set $\boldsymbol{\Delta} := \{x_i\}_{i=1,\dots,k}$ of distinct points, called *knots*, within this interval. $\boldsymbol{\Delta}$ induces a partition of $[a, b]$ into subintervals $I_i := [x_i, x_{i+1})$, $i = 0, \dots, k-1$, and $I_k := [x_k, x_{k+1}]$. Moreover, let m be a positive integer, which is called *order* of the spline, and a vector $\boldsymbol{M} := (m_1, \dots, m_k)$ of *knot multiplicities* such that $m_i \in \mathbb{N}$ and $1 \leq m_i \leq m$ for all $i = 1, \dots, k$.

Definition 3 We define the *space of polynomial splines* of order m with knots $\boldsymbol{\Delta}$ of multiplicities $\boldsymbol{M}$ as:

$$
\begin{aligned}
S(\mathcal{P}_{m-1}, \boldsymbol{M}, \boldsymbol{\Delta}) := \big\{ s \,\big|\, & \exists\, s_i \in \mathcal{P}_{m-1}, i = 0, \dots, k, \text{ such that:} \\
& \text{i) } s(x) = s_i(x) \text{ for } x \in I_i, i = 0, \dots, k; \\
& \text{ii) continuity conditions at knots:} \\
& \quad s_{i-1}^{(r)}(x_i) = s_i^{(r)}(x_i) \text{ for } r = 0, \dots, m - m_i - 1, \text{ and } i = 1, \dots, k \\
\big\}\,. &
\end{aligned}
$$

This means that $S(\mathcal{P}_{m-1}, \boldsymbol{M}, \boldsymbol{\Delta})$ is the space of piecewise functions such that each piece is defined in one of the intervals I_i induced by the knot partition $\boldsymbol{\Delta}$, and it belongs to the space $\mathcal{P}_{m-1}$ of univariate polynomials of degree at most $m-1$. Moreover, consecutive pieces s_{i-1} and s_i have to join with the specified order of continuity, related to the order m and the multiplicity m_i of the common knot x_i.
One can show that the spline space $S(\mathcal{P}_{m-1}, \boldsymbol{M}, \boldsymbol{\Delta})$ has dimension $m + K$, where $K := \sum_{i=1}^{k} m_i$.

3.1 Normalized B-splines

A crucial task is to find a basis of the spline space $S(\mathcal{P}_{m-1}, \boldsymbol{M}, \boldsymbol{\Delta})$ with good properties from the point of view of geometric modeling, just like the Bernstein basis is for the space of polynomials. To this aim, we introduce an *extended partition* $\boldsymbol{\Delta}^* := \{t_i\}_{i=1,\dots,2m+K}$ obtained from the previous partition $\boldsymbol{\Delta}$ in the following way:

i) $t_1 \le t_2 \le \ldots \le t_{2m+K}$;

ii) $t_m \equiv a$ and $t_{m+K+1} \equiv b$;

iii) $(t_{m+1} \le \ldots \le t_{m+K}) \equiv (\underbrace{x_1 = \ldots = x_1}_{m_1 \text{ times}} < \ldots < \underbrace{x_k = \ldots = x_k}_{m_k \text{ times}})$.

Definition 4 The *normalized B-spline functions* $N_{i,m}$, $i = 1, \dots, m + K$, are defined by the following recurrence relation:

$$
N_{i,1}(x) = \begin{cases} 1, & \text{if } t_i \le x < t_{i+1}, \\ 0, & \text{otherwise,} \end{cases}
$$

and for $h = 2, \dots, m$:

$$
N_{i,h}(x) = \begin{cases} \dfrac{x - t_i}{t_{i+h-1} - t_i} N_{i,h-1}(x) + \dfrac{t_{i+h} - x}{t_{i+h} - t_{i+1}} N_{i+1,h-1}(x), & \text{if } t_i \ne t_{i+h}, \\ 0, & \text{otherwise,} \end{cases}
$$

where we set $\frac{0}{0} := 0$.

Figure 3 shows some single B-spline functions defined over a uniform knot partition of various orders.

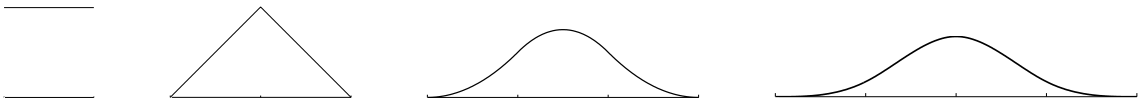


Figure 3: Single B-spline functions over a uniform knot partition ($t_i = i$) for some orders m (here $m = 1, 2, 3, 4$).

One can show that the normalized B-spline functions $\{N_{i,m}\}_{i=1,\dots,m+K}$ form a basis of the space of polynomial splines $S(\mathcal{P}_{m-1}, \boldsymbol{M}, \boldsymbol{\Delta})$. In particular, the B-spline basis may be regarded as an extension of the Bernstein basis, generalizing the description of a single polynomial to piecewise-polynomial functions. In fact, the B-spline basis retains the non-negativity and partition-of-unity properties of the Bernstein basis. Another property, pertinent to the spline context, characterizes the B-spline functions: they have compact support, which means that they are non-zero over a finite number of intervals of the extended partition.

Regarding continuity as piecewise functions, if $m_i = 1$ for all i, each B-spline basis functions of order m is C^{m-2} at the knot locations. Instead, in presence of multiple knots the order of continuity and the width of the support of the basis functions influenced by such knots are reduced.

As an example, let us consider the situation illustrated in Figure 4, which shows the B-spline basis of the space of cubic splines ($m = 4$) defined over the non-uniform knot partition $\boldsymbol{\Delta} = \{0.25, 0.5, 0.6, 0.8\}$ with multiplicities $\boldsymbol{M} = (3, 1, 1, 1)$. In particular, the function highlighted in red is only C^0 at 0.25 since it is a knot of multiplicity 3. On the contrary, from the point of view of the function highlighted in blue 0.25 has multiplicity 1, as well as all the other knots contained in its support; thus this function has the maximal continuity at knots, that is C^2 in this case.

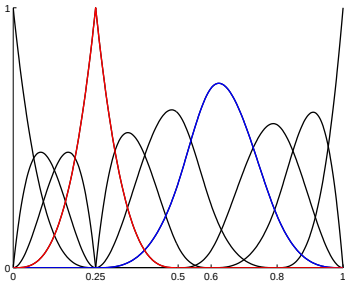


Figure 4: Example of the B-spline basis functions for a spline space with non-uniform spacing of the knots.

3.2 B-spline curves

Definition 5 Given a set of control points $\boldsymbol{p}_i \in \mathbb{R}^d$, $i = 0, \ldots, m+K$, a *B-spline curve* may be defined (similarly to Bézier curves) by associating a control point with each B-spline basis function in the space $S(\mathcal{P}_{m-1}, \boldsymbol{M}, \boldsymbol{\Delta})$:

$$\boldsymbol{c}(x) = \sum_{i=1}^{m+K} \boldsymbol{p}_i N_{i,m}(x), \qquad x \in [a, b].$$

B-spline curves enjoy the same shape properties of Bézier curves, and many computational methods exist for working with them. For instance, an efficient algorithm for the evaluation of a B-spline curve is due to de Boor and it is the generalization of the de Casteljau's algorithm.

As we can see from Figure 5, one of the advantages of the spline representation over that based on Bézier pieces is the lower number of control points that are exposed to the designer and that have to be stored by a computer, leading to a simpler and more efficient way of describing shapes.
Moreover, the flexibility of B-spline curves in geometric design is evident if we consider the possibility of exploiting multiple knots and the resulting smoothness reduction in order to achieve particular shape effects, such as sharp corners.

Apart from the good shape-approximation properties that can be used to facilitate free-form design, B-spline curves can be exploited also to solve more rigorous problems, such as interpolation or approximation of a given set of distinct points $\boldsymbol{q}_i \in \mathbb{R}^d$, $i = 1, \ldots, n$. In these contexts, we are interested in finding a B-spline curve

$$\boldsymbol{c}(x) = \sum_{i=1}^{m+K} \boldsymbol{p}_i N_{i,m}(x),$$

whose components belong to a spline space $S(\mathcal{P}_{m-1}, \boldsymbol{M}, \boldsymbol{\Delta})$, and a suitable knot sequence $\boldsymbol{\Delta}^* = \{t_i\}$ for the curve that solve one of the following problems:

- *interpolation* (when $m + K = n$): $\boldsymbol{c}$ is such that $\boldsymbol{c}(t_i) = \boldsymbol{q}_i$;
- *approximation* (when $m + K \ll n$): $\boldsymbol{c}$ is such that minimizes $\sum_{i=1}^{n} \|\boldsymbol{q}_i - \boldsymbol{c}(t_i)\|^2$ (least-squares problem).

In particular, spline interpolation entails the solution of a linear system for each component and, in this case, the problem of determining a suitable parameterization for the spline curve consists in prescribing the extended knot partition such that the existence and the uniqueness of the solution is guaranteed (Schoenberg-Whitney conditions). Also, the resulting curve shall not exhibit any undesired behavior, such as oscillations in between

the constrained points. For this reason, the knots shall depend on the location of the interpolation points:

$$t_{i+1} = t_i + \left\| \boldsymbol{q}_{i+1} - \boldsymbol{q}_i \right\|_2^{\alpha}, \qquad \alpha \in [0, 1].$$

A good compromise for automatic setting is to choose $\alpha = \frac{1}{2}$ (centripetal parameterization), while for $\alpha = 0$ and $\alpha = 1$ we have the uniform and the chordal parameterization, respectively.

As a side note, in the context of approximation theory the use of splines would allow us to attain a much faster convergence rate than Bernstein polynomials.

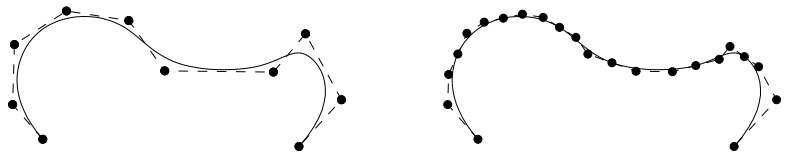


Figure 5: Spline curve of order $m = 4$ represented (a) in the B-spline basis and (b) as a collection of cubic Bézier pieces: in the latter case, the number of control points that are needed is larger.

3.3 NURBS

An extension of B-splines is given by NURBS, which are rational functions defined as the ratio of two polynomials expressed in the B-spline basis.

Definition 6 Let m, $\boldsymbol{\Delta}$ and $\boldsymbol{M}$ be specified as in the previous section. Moreover, let $\boldsymbol{W} := (w_1, \dots, w_k)$ be a vector of positive *weights*. Then, we define the space $R(\mathcal{P}_{m-1}, \boldsymbol{M}, \boldsymbol{\Delta}, \boldsymbol{W})$ of *Non-Uniform Rational B-Splines* (NURBS) as the space generated by the rational basis functions

$$R_{i,m}(x) = \frac{w_i \, N_{i,m}(x)}{\sum_j^{m+K} w_j \, N_{j,m}(x)}, \qquad x \in [a, b], \qquad i = 1, \dots, m + K,$$

where $N_{i,m}$ are the normalized B-splines.

In analogy with a B-spline curve, we can define a *NURBS curve* with given control points $\{\boldsymbol{p}_i\}$ as a curve whose components are functions in $R(\mathcal{P}_{m-1}, \boldsymbol{M}, \boldsymbol{\Delta}, \boldsymbol{W})$:

$$\boldsymbol{c}(x) = \sum_{i=1}^{m+K} \boldsymbol{p}_i \, R_{i,m}(x), \qquad x \in [a, b].$$

Among the properties of NURBS, they reduce to B-splines when all the weights w_i are set to 1, and they can reproduce conic sections by properly choosing the weights and the control points. For these reasons, they are widely used in geometric design and became

the standard form of representation for curves and surfaces in CAD systems.

We refer the interested reader to the monograph [13] that Piegl and Tiller devoted to NURBS and the related computational methods: it is regarded as the fundamental textbook on this subject, complete with a C library for the implementation of NURBS-based algorithms.

3.4 Spline surfaces

It is possible to extend the definition of spline curves to objects of higher dimension, leading for instance to spline surfaces.

One way to define spline surfaces is to consider the space obtained by tensor product of two spline spaces of the type considered in the previous sections.

Definition 7 Let $S_x := S(\mathcal{P}_{m-1}, \boldsymbol{M}, \boldsymbol{\Delta}_x)$ be a spline space defined on $[a, b]$ with knots $\boldsymbol{\Delta}_x$ and multiplicities $\boldsymbol{M} = (m_1, \dots, m_k)$, and $S_y := S(\mathcal{P}_{n-1}, \boldsymbol{N}, \boldsymbol{\Delta}_y)$ be another spline space defined on $[c, d]$ with knots $\boldsymbol{\Delta}_y$ and multiplicities $\boldsymbol{N} = (n_1, \dots, n_h)$. In addition, let a grid of control points $\boldsymbol{p}_{i,j} \in \mathbb{R}^d$, $i = 1, \dots, m + K$, $j = 1, \dots, n + H$, be assigned. Then, a *tensor-product spline surface* in $S_x \otimes S_y$ has the form

$$\boldsymbol{s}(x, y) = \sum_{i=1}^{m+K} \sum_{j=1}^{n+H} \boldsymbol{p}_{i,j} N_{i,m}(x)\, N_{j,n}(y), \qquad (x, y) \in [a, b] \times [c, d],$$

where $K := \sum_i^k m_i$, $H := \sum_j^h n_j$.

These surfaces have a good flexibility for design purposes, therefore they are widespread in CAD systems and geometric modeling in general. However, since tensor-product surfaces have an intrinsic rectangular, grid-like topology, they can represent only objects with genus 0 or 1. Thus, their use for the description of complicated objects with arbitrary topology is not straightforward, and it often requires some kind of workaround to achieve the desired surface continuity or shape quality.

4 Open problems

In this final section we only touch on some open problems in the context of geometric modeling with splines and sketch some recent approaches for addressing them.

4.1 Generalized splines

A generalization of polynomial splines consists in allowing the different pieces to belong to different function spaces, other than polynomials. In particular, let $\boldsymbol{\mathcal{U}}_m = \{\mathcal{U}_{i,m}\}_{i=1,\dots,k}$ be a sequence of (Quasi) Extended Chebyshev spaces of dimension m, and $\boldsymbol{\Delta}$ and $\boldsymbol{M}$ be respectively a knot partition and a vector of multiplicities as in Definition 3. Then we

define the *space of generalized splines* as follows:

$$
\begin{aligned}
S(\boldsymbol{\mathcal{U}}_m, \boldsymbol{M}, \boldsymbol{\Delta}) := \big\{ s \,\big|\, \exists\, s_i \in \mathcal{U}_{i,m}, i = 0, \dots, k, \text{ such that} \\
\text{(a) } & s(x) = s_i(x) \text{ for } x \in I_i, i = 0, \dots, k \\
\text{(b) } & \text{continuity conditions at knots:} \\
& s_{i-1}^{(r)}(x_i) = s_i^{(r)}(x_i) \text{ for } r = 0, \dots, m - m_i - 1, \text{ and } i = 1, \dots, k \\
\big\}\,.
\end{aligned}
$$

The motivation for this kind of spaces is the increased flexibility that they offer with respect to purely polynomial splines. In fact, according to our needs, we are allowed to consider spaces $\mathcal{U}_{i,m}$ spanned by functions of trigonometric, hyperbolic, polynomial or other type, and exploit them to achieve particular shape and tension effects. In particular, a mixed polynomial-trigonometric space would allow us to represent both polynomials and circular arcs, without resorting to rational splines that imply more complicated computations and cannot recover the arc-length parameterization.

The theory of generalized spline spaces dates back to the Eighties [16, Chapter 11], but some problems are still open. For example, finding necessary and sufficient conditions for establishing, in a general setting, whether a given spline space admits a basis that is the analogous of the B-spline basis for polynomial splines, in the sense that it enjoys all the good properties that we have recalled. Moreover, many aspects that are crucial from the point of view of application have been neglected, so that there is a lack of efficient algorithms for the computation with generalized splines.

For these reasons, we have recently devised [3] a general and effective procedure to construct a basis for all generalized spline spaces useful in a design context, and this basis is precisely the analogous of the B-spline basis (thus, it is optimal), if the space admits it. Our construction and the related proofs are simpler and more general than the ones proposed so far in the literature. In fact, the other existing approaches mainly rely on integral recurrence relations or require the burdensome task of determining particular weight functions (related to the Extended Chebyshev spaces) such that they can be joined and form piecewise functions satisfying a certain set of continuity conditions. Thus, these approaches are of little use in practice and often tailored to specific cases. Moreover, our formulation can be exploited to devise a practical criterion for establishing whether a spline space admits the optimal basis. Finally, this approach simply translates into numerical methods useful in the context of geometric design, and, for example, it provides a simple way of deriving the analogous of the de Casteljau and de Boor's algorithms.

4.2 Local spline interpolation

Another important issue concerns the definition of *local spline interpolants*, which, unlike the interpolants that solve the problem mentioned in Section 3.2 and which are called global spline interpolants, are spline where each piece is influenced only by a small subset of interpolation points.

In particular, in this context we proposed [1] the construction of piecewise-polynomial local interpolants of minimal degree with given smoothness and approximation order, defined on non-uniform knot partitions. Such interpolants are defined as the combination between some proper blending (compactly supported, "bell-shaped") functions $\{N_i\}$ and polynomials $\{\boldsymbol{P}_i\}$ of low degree that interpolate only a small number of the given points $\{\boldsymbol{q}_i\}$:

$$\boldsymbol{c}(x) = \sum_i \boldsymbol{P}_i(x)\, N_i(x) = \sum_i \boldsymbol{q}_i\, \psi_i(x), \qquad x \in [a, b].$$

Formally, the polynomials take the place of the control points in the definitions of Bézier and B-spline curves that we have recalled in the previous sections. The blending functions need not be the B-spline basis, but are still characterized by non-negativity, partition-of-unity, local support and non-maximal continuity with respect to global interpolation (where the spline has maximal continuity).
An alternative formulation is in terms of the given interpolation points $\{\boldsymbol{q}_i\}$ and fundamental functions $\{\psi_i\}$ deduced from the interpolating polynomials and blending functions in the previous expression.

4.3 Local interpolatory surfaces of arbitrary topology

Finally, when dealing with surfaces, the problem of interpolation gets harder, in particular when surfaces of arbitrary topology are sought. In fact, while in the context of approximating surfaces many approaches have been proposed and are now finding their way in applications, the definition of interpolating surfaces that are smooth and aesthetically pleasant is still an open problem.

In this regard, we have devised [2] an innovative approach that exploits local spline interpolants of the type that we have mentioned above in combination with local patching to produce surfaces that interpolate assigned quadrilateral meshes or curve networks and exhibit a good quality, besides satisfying the required continuity properties.

References

[1] M. Antonelli, C. V. Beccari, and G. Casciola, *A general framework for the construction of piecewise-polynomial local interpolants of minimum degree.* Advances in Computational Mathematics (2013), to appear.

[2] M. Antonelli, C. V. Beccari, and G. Casciola, *High-quality local interpolation of arbitrary-topology meshes and curve networks by composite parametric surfaces.* In preparation (2014).

[3] M. Antonelli, C. V. Beccari, G. Casciola, and L. Romani, *A constructive approach for the B-spline basis of generalized spline spaces.* In preparation (2014).

[4] S. N. Bernstein, *Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités*. Communications de la Société Mathématique de Kharkov 2/13 (1912), 1–2.

[5] H. B. Curry, *Review*. Mathematical Tables and Other Aids to Computation 2 (1947): 167–169, 211–213.

[6] P. d. F. de Casteljau, *De Casteljau's autobiography: My time at Citroën*. Computer Aided Geometric Design 16/7 (1999), 583–586.

[7] G. Farin, "Curves and surfaces for CAGD: a practical guide". Morgan Kaufmann Publishers, 5th edition, 2002.

[8] G. Farin, J. Hoschek, and M.-S. Kim, "Handbook of Computer Aided Geometric Design". Elsevier Science Publishers, 2002.

[9] R. T. Farouki, *The Bernstein polynomial basis: A centennial retrospective*. Computer Aided Geometric Design 29/6 (2012), 379–419.

[10] W. J. Gordon and R. F. Riesenfeld, *B-spline curves and surfaces*. In R. E. Barnhill and R. F. Riesenfeld, editors, Computer Aided Geometric Design, Academic Press (1974) 95–126.

[11] J. Hoschek and D. Lasser, "Fundamentals of Computer Aided Geometric Design". A. K. Peters, Ltd., 1993.

[12] T. J. R. Hughes, J. A. Cottrell, and Y. Bazilevs, *Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement*. Computer Methods in Applied Mechanics and Engineering, 194/39-41 (2005), 4135–4195.

[13] L. Piegl and W. Tiller, "The NURBS book". Springer-Verlag, 2nd edition, 1997.

[14] C. Rabut, *On Pierre Bézier's life and motivations*. Computer-Aided Design 34/7 (2002), 493–510.

[15] I. J. Schoenberg, *Contributions to the problem of approximation of equidistant data by analytic functions*. Quarterly of Applied Mathematics, 2/1-2 (1946): 45-99, 112–141.

[16] L. L. Schumaker, "Spline Functions: Basic Theory". Cambridge University Press, 3rd edition, 2007.

An introduction to Representation Theory of groups

Martino Garonzi (*)

Abstract. Label the faces of a cube with the numbers from 1 to 6 in some order, then perform the following operation: replace the number labeling each given face with the arithmetic mean of the numbers labeling the adjacent faces. What numbers will appear on the faces of the cube after this operation is iterated many times? This is a sample problem whose solution is a model of the application of the theory of representations of groups to diverse problems of mathematics, mechanics, and physics that possess symmetry of one kind or another. In this introductory talk I will present the tools from representation theory needed to solve this problem. I will also point out the connection with harmonic analysis by expressing Fourier analysis as an instance of representation theory of the circle group (the multiplicative group of complex numbers with absolute value 1) and by stating a version of Heisenberg's uncertainty principle for finite cyclic groups.

1 A sample problem

I will start by stating a sample problem which I found in [2].

Take a cube and write the numbers $1, 2, 3, 4, 5, 6$ on its faces, in any way you like. Then perform the following operation on the cube: substitute to the number on each face the arithmetic mean of the numbers written on the (four) adjacent faces. Iterate this. The question is: what do the numbers on the faces of the cube look like after n iterations, where n is a large number? For example in the case of a die, from the first iteration onward the value on each face is constantly 3.5, because in a die the sum of the numbers labeling two opposite faces is always 7.

The idea to solve this problem is the following. Let F be the set of faces of the cube, and let W_F be the set of functions $F \to \mathbb{C}$. W_F is a $\mathbb{C}$-vector space of dimension $|F| = 6$ spanned by $\{\delta_x \ : \ x \in F\}$ where $\delta_x(y) = 1$ if $x = y$ and $\delta_x(y) = 0$ if $x \neq y$.

Call L the operator $W_F \to W_F$ that takes a face label to the arithmetic mean of the four adjacent face labels: $L(f)(x) := \frac{1}{4} \sum_{y \in A_x} f(y)$.

(*)Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: mgaronzi@gmail.com. Seminar held on April 30th, 2014.

L is a linear operator whose matrix in the base $\{\delta_x \ : \ x \in F\}$ is

$$L = \begin{pmatrix} 0 & 1/4 & 0 & 1/4 & 1/4 & 1/4 \\ 1/4 & 0 & 1/4 & 0 & 1/4 & 1/4 \\ 0 & 1/4 & 0 & 1/4 & 1/4 & 1/4 \\ 1/4 & 0 & 1/4 & 0 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \end{pmatrix}$$

It obviously has rank 3 (opposite faces take the same value).

We are interested in the powers L^n. So our aim is to **diagonalize** L (obviously the powers of a diagonal matrix are easy to compute). Of course, this can be done computationally, but what we want to do is to look for some geometrical way to do it, i.e. by means of some **group action on the space**.

The solution of this problem is in Section 7.

2 Representations

Suppose we want to understand a set X which has some symmetries. The idea is to consider the vector space

$$V_X := \{\text{functions } X \to \mathbb{C}\} = \{\text{vectors } (c_x)_{x\in X} \ x \in X, c_x \in \mathbb{C}\}.$$

This is a $\mathbb{C}$-vector space of dimension $|X|$. Consider the group G of the symmetries of X you are interested in. In other words, G is some subgroup of the group $\mathrm{Sym}(X) = \{\text{bijections } X \to X\}$. Denote by $GL(V_X)$ the group of the linear isomorphisms $V_X \to V_X$ (it is a group with respect to composition of functions). Then we have a group homomorphism

$$\pi : G \to GL(V_X), \ g \mapsto \pi_g : \ v = (c_x)_{x\in X} \mapsto \pi_g(v) = (c_{g^{-1}(x)})_{x\in X}.$$

In other words, π sends $g \in G$ to π_g, which is the linear isomorphism $V_X \to V_X$ that sends a vector $v = (c_x)_{x\in X}$ to the vector obtained by permuting the coordinates according to g^{-1}, $\pi_g(v) = (c_{g^{-1}(x)})_{x\in X}$. This is the object we want to study. The reason why it is a homomorphism is the following.

$$\pi_{gh}((c_x)_{x\in X}) = (c_{(gh)^{-1}(x)})_{x\in X} = (c_{h^{-1}(g^{-1}(x))})_{x\in X},$$

$$\pi_g(\pi_h((c_x)_{x\in X})) = \pi_g((c_{h^{-1}(x)})_{x\in X}) = (c_{h^{-1}(g^{-1}(x))})_{x\in X}.$$

Hence $\pi_{gh} = \pi_g \circ \pi_h$.

Definition 1 A (complex, linear) representation of the group G is a $\mathbb{C}$-vector space V endowed with a group homomorphism

$$\pi : G \to GL(V), \qquad g \mapsto \pi_g$$

where $GL(V)$ is the group of the linear isomorphisms $V \to V$ with the operation given by composition of functions. This is the data of V and π, so we will also write (V, π) to denote this representation. The dimension of V is called the dimension of the representation (V, π).

Let us give some examples.

- If V is a $\mathbb{C}$-vector space, the group $GL(V)$ itself admits a representation given by the identity $GL(V) \to GL(V)$.
- If G is a group of bijections $X \to X$ where X is a finite set of cardinality n then G admits the n-dimensional representation $G \to GL(V_X)$ described above.
- The group $(\mathbb{Z}, +)$ (i.e. the set $\mathbb{Z}$ endowed with the operation $+$) admits a representation $\mathbb{Z} \to GL(\mathbb{C}^2)$ given by $n \mapsto \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}$. This is indeed a group homomorphism:

$$\begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & n+m \\ 0 & 1 \end{pmatrix},$$

$$\begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -n \\ 0 & 1 \end{pmatrix}.$$

  The same is true if we replace $\mathbb{Z}$ with any additive subgroup of $\mathbb{C}$, for example $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$ itself.

2.1 Permutation matrices

We give now an explicit instance of the second example given above, $G \to GL(V_X)$. Consider the group S_4 of bijections $X \to X$ where $X = \{1, 2, 3, 4\}$. The representation $\pi : S_4 \to GL(\mathbb{C}^4)$ described above sends a permutation σ to the corresponding "**permutation matrix**", that is, the 1-0 matrix whose 1-entries are in the $(\sigma(i), i)$ positions, for $i \in \{1, 2, 3, 4\}$.

So for example for the permutation $(123) \in S_4$ (i.e. the permutation $1 \mapsto 2 \mapsto 3 \mapsto 1$, $4 \mapsto 4$) we have

$$\pi_{(123)} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \in GL(\mathbb{C}^4).$$

Note that this is precisely the linear operator that permutes the four canonical basis vectors e_1, e_2, e_3, e_4 the same way (123) moves $1, 2, 3, 4$: it takes e_1 to e_2, e_2 to e_3, e_3 to e_1 and e_4 to e_4.

It is worth noting that the **trace** of π_σ (i.e. the sum of its diagonal entries) is the number of **fixed points** of σ.

The map $\sigma \mapsto \mathrm{Tr}(\pi_\sigma)$ will be called the "**character**" of π.

2.2 Invariant subspaces

Fix a representation (V, π) of G. The group G "acts" on the vector space V, in the sense that $g \in G$ "moves the vectors" by sending $v \in V$ to $\pi_g(v)$. In this setting, the notion of "subspace" is weak: we are much more interested in "G-invariant" subspaces! What does this mean?

Definition 2 A subspace W of V is called G-invariant (or simply, "invariant") if whenever $w \in W$ and $g \in G$, $\pi_g(w) \in W$.

2.3 A 2-dimensional representation of $(\mathbb{R}, +)$

For example the additive group $G = \mathbb{R}$ has a 2-dimensional representation given by

$$\mathbb{R} \to GL(\mathbb{C}^2), \qquad a \mapsto \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix},$$

and the subspace W of $V = \mathbb{C}^2$ given by $W = \{\begin{pmatrix} 0 \\ y \end{pmatrix} \ : \ y \in \mathbb{C}\}$, is NOT G-invariant. Indeed for example

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \notin W.$$

Instead, the subspace $L = \{\begin{pmatrix} x \\ 0 \end{pmatrix} \ : \ x \in \mathbb{C}\}$ is G-invariant, being the eigenspace of 1 for π_a, for all $a \neq 0$.

2.4 Invariant homomorphisms

Also the familiar notion of linear homomorphism between to spaces is too weak for us. We need the notion of G-invariant homomorphism! What does this mean?

Suppose (V, π) and (W, ν) are two representations of G and $f : V \to W$ is a linear map. f is called G**-invariant** if it satisfies

$$f(\pi_g(v)) = \nu_g(f(v)) \qquad \forall g \in G, v \in V, w \in W.$$

Such map is also called "**intertwining operator**", since it "intertwines" the two representations π and ν.

We give an example which constitutes the archetipe of intertwining operator. Consider the case $V = W$, choose a basis of V and think of π_g and ν_g as matrices. Suppose there is an invertible matrix A such that $\nu_g = A\pi_g A^{-1}$ for all $g \in G$, in other words ν_g is obtained by π_g via a **change of basis**. Then the map

$$f : V \to V \qquad v \mapsto Av$$

is G-invariant (i.e. it is an intertwining operator), indeed $\nu_g = A\pi_g A^{-1}$ means that $A\pi_g = \nu_g A$.

2.5 Irreducible subspaces

Usually when dealing with a big space what we want to do is try to decompose it in smaller pieces that cannot be further decomposed.

Definition 3 A representation (V, π) of G is called irreducible if the only G-invariant subspaces of V are $\{0\}$ and V.

For example the representation of the additive group $G = \mathbb{R}$ considered above,

$$\mathbb{R} \to GL(\mathbb{C}^2), \qquad a \mapsto \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix},$$

is NOT irreducible, having the invariant subspace $L = \{\begin{pmatrix} x \\ 0 \end{pmatrix} \ : \ x \in \mathbb{C}\}$. Indeed we have, for $a \in \mathbb{R}$,

$$\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{pmatrix} x \\ 0 \end{pmatrix} \in L.$$

3 Looking for decompositions

Our aim in life is now the following. Let (V, π) be a representation of a group G. If possible, we want to write V as a direct sum of irreducible G-invariant subspaces (irreducible representations), i.e. we want an expression of the form

$$V = \bigoplus_{i=1}^{n} W_i = W_1 \oplus W_2 \oplus \cdots \oplus W_n$$

where W_i is an irreducible G-invariant subspace of V for $i \in \{1, \ldots, n\}$. In order to do this, we need first of all to understand better irreducible subspaces.

For (V, π) and (W, ν) two representations of G, denote by $\mathrm{Hom}_G(V, W)$ the set of G-invariant homomorphisms $V \to W$.

Lemma 1 *If $f \in Hom_G(V, W)$ then both its kernel and its image are G-invariant subspaces, of V and W respectively.*

Proof. Let $f \in \mathrm{Hom}_G(V, W)$.

We prove that $\ker(f)$ is a G-invariant subspace of V. Let $v \in \ker(f)$, $g \in G$. We need to prove that $\pi_g(v) \in \ker(f)$, i.e. that $f(\pi_g(v)) = 0$. Since f is G-invariant, $f(\pi_g(v)) = \nu_g(f(v)) = \nu_g(0) = 0$, where $f(v) = 0$ being $v \in \ker(f)$.

We prove that $\mathrm{Im}(f) = f(V)$ is a G-invariant subspace of W. Let $w = f(v) \in \mathrm{Im}(f)$, with $v \in V$, and let $g \in G$. We need to prove that $\nu_g(w) \in \mathrm{Im}(f)$, i.e. that there is some $v' \in V$ with $f(v') = \nu_g(w)$. Since f is G-invariant, $\nu_g(w) = \nu_g(f(v)) = f(\pi_g(v))$. Choose $v' = \pi_g(v)$. □

This easily implies a fundamental fact, Schur's lemma, which is the starting point of representation theory.

Theorem 1 (Schur's lemma) *Let (V, π) and (W, ν) be two irreducible representations of the group G. Then any nonzero G-invariant map $V \to W$ is an isomorphism.*

Proof. Let $f : V \to W$ be a nonzero G-invariant map. Since f is nonzero, $\ker(f) \neq V$ and $\mathrm{Im}(f) \neq \{0\}$. On the other hand, $\ker(f)$ and $\mathrm{Im}(f)$ are G-invariant subspaces of V and W respectively (by Lemma 1), and V, W do not have nontrivial G-invariant subspaces (they are irreducible!), hence it must be $\ker(f) = \{0\}$ and $\mathrm{Im}(f) = W$, in other words f is an isomorphism. □

Observe that $\mathrm{End}_G(V) := \mathrm{Hom}_G(V, V)$ has the structure of ring with respect to (SUM) pointwise sum and (PRODUCT) composition of functions. This allows to restate Schur's Lemma in the following form:

Theorem 2 (Schur's lemma) *Let (V, π) be an irreducible representation of G. Then in the ring $End_G(V)$ every nonzero element is invertible. In other words, $End_G(V)$ is a* ***skew field****.*

Now suppose V is irreducible and finite dimensional, say $\dim_{\mathbb{C}}(V) = n$. Then also $\mathrm{End}_G(V)$ is a finite dimensional $\mathbb{C}$-vector space (it is a vector subspace of $\mathrm{End}(V)$, which has dimension n^2: it is isomorphic to the space of $n \times n$ matrices). Also, by Schur's lemma it is a skew field. Moreover, it contains a copy of $\mathbb{C}$ given by the scalar operators

$$V \to V,\ v \mapsto \lambda v, \qquad \lambda \in \mathbb{C}.$$

From the fact that $\mathbb{C}$ is **algebraically closed** (equivalently, it does not admit finite dimensional field extensions) and finite dimensionality it follows that $\mathrm{End}_G(V) \cong \mathbb{C}$. In other words, every G-invariant map $V \to V$ is scalar! Let us re-state Schur's Lemma accordingly.

Theorem 3 (Schur's lemma) *Let (V, π) be a finite dimensional irreducible representation of the group G, and let $f : V \to V$ be a G-invariant homomorphism. Then there exists $\lambda \in \mathbb{C}$ such that $f(v) = \lambda v$ for all $v \in V$.*

Let us see what this means in the case G is **abelian**. In this case for $g, h \in G$ we have $gh = hg$, so that

$$\pi_g \pi_h = \pi_{gh} = \pi_{hg} = \pi_h \pi_g$$

($\pi : G \to GL(V)$ is a homomorphism!). This, by the very definition of intertwining operator, implies that π_h is G-invariant (i.e. it is an intertwining operator) for all $h \in G$! So Schur's Lemma implies that π_h is a scalar operator, for all $h \in G$. Hence irreducibility forces the dimension of V to be 1: an irreducible representation that sends every group element to a scalar operator must be one-dimensional, because the scalar operators stabilize all subspaces (actually they are the only operators that stabilize all subspaces). We conclude that:

Corollary 1 *Let (V, π) be an irreducible finite dimensional representation of the* ***abelian*** *group G. Then* $\dim(V) = 1$.

Let us go back to the 2-dimensional representation of the additive group $G = \mathbb{R}$

$$\pi : \mathbb{R} \to GL(\mathbb{C}^2), \qquad a \mapsto \pi_a = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}.$$

We know that $L = \{\begin{pmatrix} x \\ 0 \end{pmatrix} \ : \ x \in \mathbb{C}\}$ is a G-invariant subspace of $V = \mathbb{C}^2$. Now we ask, **is V the direct sum of two 1-dimensional G-invariant subspaces?** This would be great, because the way of understanding a space is by writing it as a direct sum of irreducible subspaces.

But what is a subspace invariant under π_a? It is just an **eigenspace** of π_a. Hence being able to write V as direct sum of two 1-dimensional G-invariant subspaces would mean, in particular, being able to **diagonalize** π_a (simultaneously, i.e. uniformly with respect to a !).

The problem is that π_a is not diagonalizable if $a \neq 0$. Hence V is NOT the direct sum of two 1-dimensional G-invariant subspaces.

3.1 Unitarisability

So, we have a problem. We might find a G-invariant subspace W of the finite dimensional space V without a G-invariant complement (a complement of W is a subspace U of V such that $V = U \oplus W$). Note that decomposing into direct sums is indeed equivalent to finding invariant subspaces complementing each other.

UNITARISABILITY (Weyl's unitary trick). Suppose that G is a **finite** group. The formula

$$B(u, v) := \frac{1}{|G|} \sum_{g \in G} \pi_g(u) \cdot \overline{\pi_g(v)}$$

defines a hermitian inner product on V, which has the property of being ***G*-invariant** :

$$B(\pi_g(u), \pi_g(v)) = B(u, v) \qquad \forall u, v \in V.$$

If U is a G-invariant subspace of V then $U^{\perp}$, the space of vectors v such that $B(u, v) = 0$ for all $u \in U$, is a G-invariant complement of U.

We deduce that if G is finite then we can indeed decompose the space as direct sum of G-invariant irreducible subspaces:

Theorem 4 (Maschke Theorem) *Suppose that G is finite. Then any finite dimensional representation of G is completely reducible, i.e. it is a direct sum of irreducible G-invariant subspaces.*

4 Characters

Now for $g \in G$ define $\chi_\pi(g) := Tr(\pi_g)$, the **trace** of the (matrix!) operator π_g (i.e. the sum of its diagonal entries).

For example, in the case of permutation matrix representations $\chi_\pi(g)$ is the number of fixed points of the permutation g.

The map $\chi_\pi : G \to \mathbb{C}$, $g \mapsto \chi_\pi(g)$ is called the "**character**" of the representation π.

Theorem 5 (Frobenius) *Let π_1, π_2 be two representations of the finite group G and let χ_1, χ_2 be their characters. Then $\pi_1 \cong \pi_2$ if and only if $\chi_1 = \chi_2$.*

In other words, it is enough to know the traces (!) of the matrices π_g to recover the whole representation π.

The idea to prove this is the following. Let χ_1, χ_2 be two functions $G \to \mathbb{C}$ (for example, two characters of G), i.e. elements of $\mathbb{C}^G$. Set

$$B(\chi_1, \chi_2) := \frac{1}{|G|} \sum_{g \in G} \chi_1(g)\overline{\chi_2(g)}.$$

This defines a hermitian inner product on $\mathbb{C}^G = \{functions\ G \to \mathbb{C}\}$. Now suppose χ_i is the character of the representation π_i for $i = 1, 2$, and suppose π_1 is irreducible. Then $B(\chi_1, \chi_2)$ equals the multiplicity of π_1 in the decomposition of π_2 into irreducibles. Let us state this explicitly.

Proposition 1 (Orthogonality of Characters) *Let χ be the character of an irreducible representation V of the finite group G, and let θ be a representation of G. Write $\theta = \bigoplus_{i=1}^n \theta_i^{\oplus m_i}$ where $\theta_1, \ldots, \theta_n$ are irreducible representations of G and $\theta_i^{\oplus m_i}$ means $\theta_i \oplus \ldots \oplus \theta_i$, m_i times, where m_i (the multiplicity of θ_i in θ) is a positive integer. Denote by χ_π and χ_θ the character of π, χ respectively. Then $B(\chi_\theta, \chi_\pi) = 0$ unless $\pi \cong \theta_i$ for some $i \in \{1, \ldots, n\}$ and in this case $B(\chi_\theta, \chi_\pi) = m_i$.*

The proof of this is a bit technical and we omit it. It can be found in any textbook of representation theory, cf. [1]. For example, suppose $\pi = \alpha \oplus \beta \oplus \beta$ with α, β irreducible. Then $B(\chi_\pi, \chi_\alpha) = 1$ and $B(\chi_\pi, \chi_\beta) = 2$. If γ is an irreducible representation not isomorphic to α or β then $B(\chi_\pi, \chi_\gamma) = 0$.

Corollary 2 *Let χ be a character of G. Then χ is irreducible if and only if $B(\chi, \chi) = 1$. Moreover if χ is irreducible and ψ is an irreducible character of G then $B(\chi, \psi) = 1$ if $\chi = \psi$ and $B(\chi, \psi) = 0$ if $\chi \neq \psi$. In other words, distinct irreducible characters are orthogonal to each other.*

We include the following corollary because it is beautiful.

Corollary 3 (n-th Burnside Theorem) *Let $n_1, \ldots, n_t$ be the degrees of the irreducible characters of the finite group G. Then $n_1^2 + \ldots + n_t^2 = |G|$.*

Proof. For every $g \in G$ consider the function $\gamma_g : G \to G$ given by $\gamma_g(x) = xg$. This is a bijective map whose inverse is $\gamma_{g^{-1}}$. The map $G \to \mathrm{Sym}(G)$ that sends g to γ_g is a group homomorphism, hence by considering permutation matrices we can associate to it a representation $\pi : G \to GL(V_G)$ whose character χ takes $g \in G$ to the number of fixed points of γ_g. But γ_g is fixed-point-free for all $g \neq 1$, indeed if $xg = x$ then multiplying by x^{-1} on the left we find $g = 1$. Clearly $\gamma_1 = 1$ has $|G|$ fixed points. It follows that $\chi(1) = |G|$ and $\chi(g) = 0$ if $g \neq 1$.

With this information we can now determine the decomposition of π as direct sum of irreducible representations. Let $\pi_1, \ldots, \pi_t$ be the irreducible representations of G and let $\chi_1, \ldots, \chi_t$ be their characters. Using Maschke's Theorem we can write $\pi = \bigoplus_{i=1}^{t} \pi_i^{\oplus m_i}$, so that $\chi = m_1\chi_1 + \ldots + m_t\chi_t$. We want to compute the multiplicities m_i. For $i = 1, \ldots, t$ we have

$$m_i = B(\chi, \chi_i) = \frac{1}{|G|} \sum_{x \in G} \chi(x)\overline{\chi_i(x)} = \frac{1}{|G|}|G|\chi_i(1) = \chi_i(1) = n_i.$$

It follows by bilinearity of B and orthogonality of irreducible characters that

$$\begin{aligned} n_1^2 + \ldots + n_t^2 &= B(\sum_{i=1}^{t} n_i\chi_i, \sum_{i=1}^{t} n_i\chi_i) = B(\chi, \chi) \\ &= \frac{1}{|G|} \sum_{x \in G} \chi(x)\overline{\chi(x)} = \frac{1}{|G|}|G|^2 = |G|. \end{aligned}$$

This concludes the proof. □

4.1 The character table of S_3

Consider S_3, the group of bijections $\{1, 2, 3\} \to \{1, 2, 3\}$. S_3 permutes naturally the three basis vectors e_1, e_2, e_3 of $\mathbb{C}^3$. Thinking of the elements of S_3 as permutation matrices we can imagine that $S_3 \leq GL(\mathbb{C}^3)$. It turns out that every element is conjugated to one of the matrices displayed below.

S_3	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$
$\chi_1 = 1$	1	1	1
$\chi_2 = \det$	1	-1	1
χ_3	2	0	-1

χ_3 is the character of the following representation: S_3 acts on $W := \{(x_1, x_2, x_3) \in \mathbb{C}^3 \ : \ x_1 + x_2 + x_3 = 0\}$ (2-dimensional) by permuting the canonical basis vectors (indeed, the equation $x_1 + x_2 + x_3 = 0$ is invariant under any permutation of the three indices $1, 2, 3$).

How does $V = \mathbb{C}^3$ decompose via this representation? Let U be the space of constant vectors: $U := \{(a, a, a) \ : \ a \in \mathbb{C}\}$. Then U and W are irreducible S_3-invariant subspaces of V and

$$V = U \oplus W.$$

Also, note that the column of the character table with the identity matrix on top gives precisely the degrees of the irreducible representations of S_3 (indeed, the trace of the identity matrix is just the dimension of the space!) and this fits with the n-th Burnside Theorem (Corollary 3) because

$$1^2 + 1^2 + 2^2 = 6 = 3! = |S_3|.$$

5 Fourier analysis

Let us relax the condition of finiteness of G and consider **compactness**. Consider the **circle** of center the origin and radius 1:

$$G := S^1 = \{e^{i\theta} \ : \ 0 \leq \theta \leq 2\pi\}.$$

Why did I call it G? Because it is a group with respect to multiplication:

$$e^{i\theta_1}e^{i\theta_2} = e^{i(\theta_1+\theta_2)}, \qquad (e^{i\theta})^{-1} = e^{-i\theta} = e^{i(2\pi-\theta)}.$$

Algebraically it can be viewed as the quotient $\mathbb{R}/2\pi\mathbb{Z}$.

- To make the theory of representations meaningful in this setting **we must take the topology into account**. In other words, we will have a topology on the space, and the notion of "G-invariant subspace" will be substituted by "**closed** G-invariant subspace". Also, we require morphisms to be **continuous**. Moreover, the notion of direct sum will be substituted with the notion of **orthogonal** direct sum (cf. below).
- What are the irreducible representations of S^1? Since S^1 is abelian (and compact: cf. Section 6), they are all 1-dimensional (by Schur's lemma!), i.e. they are continuous homomorphisms $S^1 \to \mathbb{C}^\times$. This forces them to be of the form $e^{i\theta} \mapsto e^{in\theta}$ where $n \in \mathbb{Z}$.

Now we need a vector space representing our group $G = S^1$.

- Let $V := L^2(G) = \{f : G \to \mathbb{C} \ : \ \int_G |f(e^{i\theta})|^2 d\theta < \infty\}$. It is a **Hilbert space** with the hermitian inner product given by

$$B(u,v) := \frac{1}{2\pi}\int_G u(e^{i\theta})\overline{v(e^{i\theta})}d\theta.$$

  Note that $2\pi = \int_G d\theta$, hence it substitutes $|G|$, which was used in the finite case. Note that a function $S^1 \to \mathbb{C}$ can be thought of as a periodic function $\mathbb{R} \to \mathbb{C}$ of period 2π. Orthogonality of characters now amounts to the following easy computation. Suppose n, m are distinct integers. Then

$$\begin{aligned}\frac{1}{2\pi}\int_0^{2\pi} e^{in\theta}\overline{e^{im\theta}}d\theta &= \frac{1}{2\pi}\int_0^{2\pi} e^{in\theta}e^{-im\theta}d\theta = \frac{1}{2\pi}\int_0^{2\pi} e^{i(n-m)\theta}d\theta \\ &= \frac{1}{2\pi}\frac{1}{i(n-m)}[e^{i(n-m)\theta}]_0^{2\pi} \\ &= \frac{1}{2\pi i(n-m)}(e^{i(n-m)2\pi} - e^{i(n-m)0}) = 0.\end{aligned}$$

Instead, if $n = m$ then we have

$$\begin{aligned}\frac{1}{2\pi}\int_0^{2\pi} e^{in\theta}\overline{e^{in\theta}}d\theta &= \frac{1}{2\pi}\int_0^{2\pi} e^{in\theta}e^{-in\theta}d\theta = \frac{1}{2\pi}\int_0^{2\pi} e^{i(n-n)\theta}d\theta \\ &= \frac{1}{2\pi}\int_0^{2\pi} d\theta = 1.\end{aligned}$$

- Consider the following representation of G:

  $$\pi : G \to GL(V), \qquad \pi_{e^{i\theta}}(f)(e^{i\theta_0}) := f(e^{i(\theta_0+\theta)}).$$

  It is **unitary**, i.e. $B(\pi_g(f_1), \pi_g(f_2)) = B(f_1, f_2)$ (this follows easily by the change of variables $\tau = \theta_0 + \theta$).

  – So Maschke Theorem holds!

- The irreducible invariant subspaces of V are, for $n \in \mathbb{Z}$,

  $$V_n := \{f \in V \ : \ f(e^{i(\theta_0+\theta)}) = e^{in\theta}f(e^{i\theta_0})\} = \mathbb{C}\{e^{i\theta} \mapsto e^{in\theta}\}.$$

  In other words, an element of V_n has the form $e^{i\theta} \mapsto \lambda e^{in\theta}$ for some $\lambda \in \mathbb{C}$.

We want to **decompose our V as direct sum of irreducible subspaces**. Now V is a Hilbert space, and in this setting the right notion of direct sum to use is the notion of **orthogonal** direct sum. If V is a Hilbert space and $\{H_i \ : \ i \in I\}$ is a family of subspaces, the orthogonal direct sum $\widehat{\bigoplus} H_i$ is the set of elements $(h_i)_{i\in I} \in \prod_{i\in I} H_i$ such that $\sum_{i\in I} ||h_i||^2 < \infty$. It turns out that the orthogonal direct sum of the subspaces H_i is the closure in V of the algebraic direct sum $\bigoplus_{i\in I} H_i$, which is by definition the set of elements $(h_i)_{i\in I} \in \prod_{i\in I} H_i$ such that the set $\{i \in I \ : \ h_i \neq 0\}$ is finite.

The decomposition of $V = L^2(S^1)$ into irreducible G-invariant subspaces is the following:

$$\begin{aligned}V &= \widehat{\bigoplus}_{n\in\mathbb{Z}} V_n = \overline{\{\sum_{finite} v_n \ : \ v_n \in V_n\}} \\ &= \{\theta \mapsto \sum_{n\in\mathbb{Z}} c_n e^{in\theta} \ : \ c_n \in \mathbb{C} \ \forall n \in \mathbb{Z}, \ \sum_{n\in\mathbb{Z}} |c_n|^2 < \infty\}.\end{aligned}$$

This says that any function $S^1 \to \mathbb{C}$ (i.e. any periodic function of period 2π !) which is square-integrable admits an espression of the form

$$f(\theta) = \sum_{n\in\mathbb{Z}} c_n e^{in\theta}$$

for some $c_n \in \mathbb{C}$ with $\sum_{n\in\mathbb{Z}} |c_n|^2 < \infty$ (this ensures that f is square-integrable). This is the Fourier series of f.

Now we want to compute the Fourier coefficients c_m (what in the discrete case we were calling "multiplicities"!). Using orthogonality of characters we find

$$\frac{1}{2\pi}\int_0^{2\pi} f(\theta)e^{-im\theta}d\theta = B(f(\theta), e^{im\theta}) = B(\sum_{n\in\mathbb{Z}} c_n e^{in\theta}, e^{im\theta}) = c_m.$$

Thus **Fourier analysis is the representation theory of the circle group** S^1. This is related to the representation theory of $G = SL(2,\mathbb{R})$, the group of 2×2 matrices with real coefficients and determinant 1: the matrices of the form $\begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}$ form a maximal compact subgroup K of G isomorphic to S^1. Using Harish-Chandra modules, it is possible to study the representations of the Lie group G only using K and the Lie algebra of G. A good source of material about this can be found in the videos of the $SL(2,\mathbb{R})$ Summer School in Utah in June 2006 [5].

6 Finite dimensionality of irreducible representations of compact groups

Let us spend some more words on why the irreducible representations of a compact group are finite dimensional. A Hilbert space is a vector space H endowed with a positive definite Hermitian inner product $\langle\cdot,\cdot\rangle$ which makes it a complete metric space. Let H be a Hilbert space and let $U(H)$ be the group of all bounded unitary operators on H. Recall that a linear operator $U : H \to H$ is said to be bounded if there exists a constant M such that $||Uh||/||h|| \leq M$ for all $h \in H - \{0\}$. A bounded linear operator $U : H \to H$ is said to be unitary if U is surjective and $\langle Ux, Uy\rangle_H = \langle x, y\rangle_H$ for all $x, y \in H$.

Let G be a compact group, and let $\pi : G \to U(H)$ be a group homomorphism such that for all $v \in H$, the map $G \to H$, $g \mapsto \pi(g)v$ is continuous. This π is what we call a unitary representation of G. π is irreducible if H has no closed invariant subspaces except for $\{0\}$ and H. Let $\pi : G \to U(H)$ be a unitary representation of G and suppose that it is irreducible. We want to show that then H must be finite-dimensional.

Theorem 6 (Schur's Lemma) *If T is a bounded linear operator on H such that $T\pi(g) = \pi(g)T$ for all $g \in G$ then T is the multiplication by a scalar $\lambda \in \mathbb{C}$.*

Since G is compact it admits a measure, called left (normalized) Haar measure. It is characterized as follows. A Borel set in G is an element of the σ-algebra of G generated by the open subsets of G. There is, up to a multiplicative constant, a unique countably additive, nontrivial measure μ on the Borel subsets of G satisfying the following properties:

- μ is left-translation-invariant: $\mu(gE) = \mu(E)$ for every $g \in G$ and Borel set E.
- μ is finite on every compact set: $\mu(K) < \infty$ if K is compact.
- μ is outer regular on Borel sets E:

$$\mu(E) = \inf\{\mu(U) \ : \ E \subseteq U,\ U \text{ open}\}.$$

- μ is inner regular on open sets E:

$$\mu(E) = \sup\{\mu(K) \ : \ K \subseteq E, \ K \text{ compact}\}.$$

It is called a "left Haar measure". The normalized left Haar measure on G is the unique left Haar measure μ on G such that $\mu(G) = 1$.

Using the Haar measure we can compute integrals. The map $G \to \mathbb{C}$, $g \mapsto \langle \pi(g)u, v\rangle$ is continuous on G for all $u, v \in H$. For $v, v', w, w' \in H$ consider

$$I(v, v', w, w') := \int_G \langle \pi(g)v, w\rangle \cdot \overline{\langle \pi(g)v', w'\rangle} dg.$$

Think of w, w' as fixed. It follows from the Riesz representation theorem that there is a bounded linear operator $T_{w,w'} : H \to H$ such that $I(v, v', w, w') = \langle T_{w,w'}v, v'\rangle$ for all $v, v' \in H$. Now we prove that for every $g \in G$ we have $\pi(g)T_{w,w'} = T_{w,w'}\pi(g)$. We have

$$\begin{aligned} \langle T_{w,w'}\pi(g)v, v'\rangle &= \int_G \langle \pi(h)\pi(g)v, w\rangle \overline{\langle \pi(h)v', w'\rangle} dh \\ &= \int_G \langle \pi(h)v, w\rangle \overline{\langle \pi(hg^{-1})v', w'\rangle} dh \\ &= \langle T_{w,w'}v, \pi(g^{-1})v'\rangle = \langle \pi(g)T_{w,w'}v, v'\rangle. \end{aligned}$$

The second equality follows from the fact that the measure dg is G-invariant, the fourth equality follows from unitarity of π (apply $\pi(g)$ to both the arguments of $\langle \cdot, \cdot\rangle$). This proves that $\pi(g)T_{w,w'} = T_{w,w'}\pi(g)$. By Schur's Lemma we deduce that $T_{w,w'}$ is the multiplication by the scalar $\lambda(w, w') \in \mathbb{C}$. We obtain that

$$\int_G \langle \pi(g)v, w\rangle \cdot \overline{\langle \pi(g)v', w'\rangle} dg = \lambda(w, w')\langle v, v'\rangle.$$

Repeating the same argument thinking of v, v' as fixed we find that there is a function $\mu(v, v')$ such that

$$\int_G \langle \pi(g)v|w\rangle \cdot \overline{\langle \pi(g)v', w'\rangle} dg = \mu(v, v')\overline{\langle w, w'\rangle}.$$

It follows that for all $v, v', w, w' \in H$ we have $\lambda(w, w')\langle v, v'\rangle = \mu(v, v')\overline{\langle w, w'\rangle}$. Choosing $v = v' = v_0$ of norm 1 we find $\lambda(w, w') = \mu(v_0, v_0)\overline{\langle w, w'\rangle}$. Call $C := \mu(v_0, v_0)$. Then we find

$$\int_G \langle \pi(g)v, w\rangle \cdot \overline{\langle \pi(g)v', w'\rangle} dg = C\langle v, v'\rangle\overline{\langle w, w'\rangle}.$$

Suppose $v = v'$, $w = w'$ and $||v|| = ||w|| = 1$. Then

$$\int_G |\langle \pi(g)v, w\rangle|^2 dg = C > 0. \qquad (*)$$

We want to show that H is finite dimensional. Suppose by contradiction that H is infinite dimensional. Then for every positive integer n we can find $e_1, \ldots, e_n \in H$ mutually

orthogonal of length 1. Since π is unitary, $\pi(g)e_1, \ldots, \pi(g)e_n$, for $g \in G$, are also mutually orthogonal of length 1. Using Bessel inequality for $v \in H$ we find

$$\sum_{i=1}^{n} |\langle v, \pi(g)e_i\rangle|^2 \leq ||v||^2. \qquad (**)$$

Now integrating over G and using $(*)$ and $(**)$ we find

$$\begin{aligned} nC||v||^2 &= \sum_{i=1}^{n} C||v||^2 = ||v||^2 \sum_{i=1}^{n} \int_G |\langle \pi(g)e_i, \frac{v}{||v||}\rangle|^2 dg \\ &= \int_G \sum_{i=1}^{n} |\langle v, \pi(g)e_i\rangle|^2 \leq \int_G ||v||^2 dg = ||v||^2 \end{aligned}$$

hence $n \leq 1/C$. This cannot hold for every positive integer n, thus we found a contradiction. So H is finite dimensional. Let now n be its dimension. Then $\sum_{i=1}^{n} |\langle v, \pi(g)e_i\rangle|^2 = ||v||^2$ (the equality in $(**)$ becomes an equality). The above computation then shows that $C = 1/n$.

6.1 Heisenberg's Uncertainty Principle

Let $G = \mathbb{Z}/N\mathbb{Z} = \{0, 1, \ldots, N-1\}$ (cyclic group of order N) and let

$$L(G) := \mathbb{C}^G = \{\text{functions } G \to \mathbb{C}\}.$$

It is a $\mathbb{C}$-vector space of dimension $|G|$.

Let $\hat{G} := \{\textit{group homomorphisms } \chi : G \to \mathbb{C}^\times\}$.

- It is a group isomorphic to G, generated by $1 \mapsto e^{i2\pi/N}$.
- It is the set of linear (1-dimensional) characters of G.
- It is a basis of $L(G)$.

Fourier transform:

$$\mathcal{F} : L(G) \to L(\hat{G}), \qquad \mathcal{F}(f)(\chi) := B(f, \chi) = \frac{1}{|G|} \sum_{x \in G} f(x)\overline{\chi(x)},$$

the coefficient of χ in the expression of f in the base $\hat{G}$.

In other words, if $f = \sum_{\chi \in \hat{G}} \hat{f}(\chi)\chi$ then $\mathcal{F}(f)(\chi) = \hat{f}(\chi)$.

For $f \in L(G)$ let $\mathrm{Supp}(f) := \{x \in G \ : \ f(x) \neq 0\}$.

Theorem 7 (Heisenberg's Uncertainty Principle) *If $f \in L(G)$ then*

$$|\mathrm{Supp}(f)| \cdot |\mathrm{Supp}(\mathcal{F}(f))| \geq |G|.$$

Proof. For $f, h \in L(G)$ define

$$\langle f, h\rangle_{L(G)} := \frac{1}{|G|}\sum_{x\in G} f(x)\overline{h(x)}$$

and

$$||f|| := \sqrt{\langle f, f\rangle_{L(G)}}.$$

This defines a positive definite hermitian inner product in $L(G)$. For $f, h \in L(\hat{G})$ define

$$\langle f, h\rangle_{L(\hat{G})} := \sum_{\chi\in\hat{G}} f(\chi)\overline{h(\chi)}$$

and

$$||f|| := \sqrt{\langle f, f\rangle_{L(\hat{G})}}.$$

This defines a positive definite hermitian inner product in $L(\hat{G})$. We now prove that the Fourier transform $\mathcal{F}$ is unitary (this is known as **Parseval formula**): if $f, h \in L(G)$ then $\langle f, h\rangle_{L(G)} = \langle \hat{f}, \hat{h}\rangle_{L(\hat{G})}$. Indeed, by orthogonality of characters we have

$$\begin{aligned}
\langle f, h\rangle_{L^2(G)} &= \frac{1}{|G|}\sum_{x\in G} f(x)\overline{h(x)} = \frac{1}{|G|}\sum_{x\in G}\sum_{\chi,\eta\in\hat{G}} \hat{f}(\chi)\chi(x)\overline{\hat{h}(\eta)}\overline{\eta(x)} \\
&= \frac{1}{|G|}\sum_{\chi,\eta\in\hat{G}} \hat{f}(\chi)\overline{\hat{h}(\eta)}\sum_{x\in G}\chi(x)\overline{\eta(x)} = \sum_{\chi\in\hat{G}} \hat{f}(\chi)\overline{\hat{h}(\chi)} \\
&= \langle \hat{f}, \hat{h}\rangle_{L^2(\hat{G})}.
\end{aligned}$$

Recall that in both spaces $L(G)$, $L(\hat{G})$ we have the **Cauchy-Schwarz inequality**

$$|\langle f, h\rangle| \leq ||f|| \cdot ||h||$$

for any f, h in the space. With these ingredients we can proceed to our computation. For a function f defined on a set X define $||f||_\infty := \max_{x\in X} |f(x)|$. For $f \in L(G)$, $A = \mathrm{Supp}(f)$, $B = \mathrm{Supp}(\hat{f})$ we then have

$$\begin{aligned}
||\hat{f}||_\infty &= \max_{\chi\in\hat{G}} |f(\chi)| = \max_{\chi\in\hat{G}} |\frac{1}{|G|}\sum_{x\in G} f(x)\overline{\chi(x)}| \\
&= \max_{\chi\in\hat{G}} |\frac{1}{|G|}\sum_{x\in G} 1_A(x)f(x)\overline{\chi(x)}| \\
&= \max_{\chi\in\hat{G}} |\langle f, 1_A\chi\rangle| \leq \max_{\chi\in\hat{G}} ||f||_{L^2(G)}||1_A\chi||_{L^2(G)}
\end{aligned}$$

where the last inequality is a consequence of the Cauchy-Schwarz inequality. Now since the elements of $\hat{G}$ are linear characters, they are homomorphisms $G \to \mathbb{C}^\times$ hence $\chi(x)$ is a root of unity for all $\chi \in \hat{G}$, $x \in G$ hence $|\chi(x)| = 1$. We deduce that

$$||1_A\chi||_{L^2(G)} = \sqrt{|\frac{1}{|G|}\sum_{x\in G} 1_A(x)\overline{\chi(x)}|} \leq \sqrt{\frac{1}{|G|}\sum_{x\in G} 1_A(x)|\overline{\chi(x)}|} = \sqrt{|A|/|G|}.$$

We can now proceed with our estimation recalling that $\mathcal{F}$ is unitary.

$$\begin{aligned} ||\hat{f}||_\infty &\leq \max_{\chi\in\hat{G}} ||f||_{L^2(G)}||1_A\chi||_{L^2(G)} \leq \sqrt{|A|/|G|}\cdot||f||_{L^2(G)} \\ &= \sqrt{|A|/|G|}\cdot||\hat{f}||_{L^2(\hat{G})} = \sqrt{|A|/|G|}\sqrt{\sum_{\chi\in\hat{G}} 1_B(\chi)|\hat{f}(\chi)|^2} \\ &\leq \sqrt{|A|/|G|}\cdot||\hat{f}||_\infty\sqrt{\sum_{\chi\in\hat{G}} 1_B(\chi)} = \sqrt{|A||B|/|G|}\cdot||\hat{f}||_\infty. \end{aligned}$$

In conclusion $||\hat{f}||_\infty \leq \sqrt{|A||B|/|G|}\cdot||\hat{f}||_\infty$ hence $|A||B| \geq |G|$. □

7 The solution of the cube problem

Now we go back to our original problem. Let G be **the group of rotations of the cube**. Then $|G| = 24$ (if you place a cube on a table, you can put each of the 6 faces up, and rotate that face in 4 ways).

It turns out that $G \cong S_4$ (the idea is to observe that G permutes the four diagonals of the cube in any possible way!).

G permutes the six faces of the cube. Let F be the set of faces of the cube. This gives a permutation matrix representation

$$\pi : G \to GL(W_F) \quad \text{where} \quad W_F := \mathbb{C}^F = \{functions\ F \to \mathbb{C}\} \cong \mathbb{C}^6.$$

By computing the fixed points of the elements of G we can compute the character of this representation. Call it χ.

It turns out that $B(\chi,\chi) = \frac{1}{|G|}\sum_{g\in G}\chi(g)\overline{\chi(g)} = 3$.

Now we know by Maschke Theorem that π is a direct sum of irreducible representations, so we can write $\pi = \bigoplus_{i=1}^n \pi_i^{\oplus m_i}$ with $m_1,\ldots,m_n$ positive integers (m_i is the multiplicity of π_i in π, i.e. the number of times π_i appears in the decomposition of π: the notation $\pi_i^{\oplus m_i}$ means $\pi_i\oplus\ldots\oplus\pi_i$, m_i times) and we deduce $\chi = \sum_{i=1}^n m_i\chi_i$ with χ_i the character of π_i for $i = 1,\ldots,n$. On the other hand $B(\chi,\chi) = 3$ hence, by orthogonality of characters,

$$3 = B(\chi,\chi) = B(\sum_{i=1}^n m_i\chi_i, \sum_{i=1}^n m_i\chi_i) = m_1^2 + \ldots + m_n^2.$$

Since $m_1,\ldots,m_n$ are positive integers with the property that $m_1^2 + \ldots + m_n^2 = 3$, we deduce that $n = 3$ and $m_1 = m_2 = m_3 = 1$. In other words $\pi = \pi_1 \oplus \pi_2 \oplus \pi_3$ where π_i is irreducible for $i = 1,2,3$ and $\chi = \chi_1 + \chi_2 + \chi_3$.

Hence $W_F = \mathbb{C}^F = \{functions\ F \to \mathbb{C}\} \cong \mathbb{C}^6$ **is the direct sum of three G-invariant irreducible subspaces**. We are left to find them. This is where the geometry comes in: our problem is now reduced to find G-invariant subspaces.

But we know the characters of the corresponding representations! This means that we have a lot of information about them, which leads us close to determining them explicitly. This is the use of representation theory: collect as much information as possible about the decomposition.

For a face x let $-x$ denote the face opposite to x. Consider

- $W_1 := \{constant\ functions\ F \to \mathbb{C}\}$. This is clearly one-dimensional: $\dim_{\mathbb{C}}(W_1) = 1$.
- $W_2 := \{f : F \to \mathbb{C} \ : \ f(-x) = f(x)\ \forall x \in F,\ \sum_{x\in F} f(x) = 0\}$. This is given by $3 + 1 = 4$ equations so $\dim_{\mathbb{C}}(W_2) = 6 - 4 = 2$.
- $W_3 := \{f : F \to \mathbb{C} \ : \ f(-x) = -f(x)\ \forall x \in F\}$. This is given by 3 equations so $\dim_{\mathbb{C}}(W_3) = 6 - 3 = 3$.

The decomposition of W_F into irreducible subspaces is

$$W_F = W_1 \oplus W_2 \oplus W_3.$$

The operator we are concerned with is

$$L : W_F \to W_F, \qquad L(f)(x) := \frac{1}{4} \sum_{y\in A_x} f(y)$$

where A_x denotes the set of faces adjacent to the face x. It turns out that L is G-invariant! This is because *rotating after averaging is the same as averaging after rotating.*

Since W_1, W_2, W_3 are irreducible, by Schur's lemma $L|_{W_i}$ is a scalar operator. Using $W_F = W_1 \oplus W_2 \oplus W_3$ it turns out that

$$L \sim \left(\begin{array}{c|cc|ccc} 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & -1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1/2 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

The operator L^n has eigenvalues 1, $(-1/2)^n$ and 0, hence if n is large then $L^n(f)$ is approximately equal to the projection of f onto W_1. The projection of any vector f whose entries are $1, 2, 3, 4, 5, 6$ in some order onto W_1 is always $(3.5, 3.5, 3.5, 3.5, 3.5, 3.5)$. Hence no matter what is the initial configuration, i.e. the initial position of the numbers from

1 to 6 labelling the faces of the cube, after many iterations the value on each face gets arbitrarily close to 3.5.

This phenomenon can be interpreted using ergodic theory. To give an idea of this fact let us state a version of Von Neumann ergodic theorem taken from [4] (the original source is [3]). Note however that the following result does not apply right away to our case because L is not invertible.

Theorem 8 (Von Neumann Ergodic Theorem) *Let $U : H \to H$ be a unitary operator on a separable Hilbert space H. Then for every $v \in H$ we have*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} U^n v = \pi(v)$$

where $\pi : H \to H^U$ is the orthogonal projection from H to the closed subspace $H^U := \{v \in H \ : \ Uv = v\}$ consisting of the U-invariant vectors.

In our problem we actually dealt with the powers of the operator L and not the arithmetic mean of the powers. But note that if a sequence a_n converges to a in a normed vector space then also the arithmetic mean $\frac{1}{N}\sum_{n=0}^{N-1} a_n$ converges to a as $N \to \infty$.

References

[1] I. M. Isaacs, "Character Theory of Finite Groups". Corrected reprint of the 1976 original; Academic Press, New York, 2006.

[2] A. A. Kirillov, "Elements of the Theory of Representations".Translated from the Russian by Edwin Hewitt. Grundlehren der Mathematischen Wissenschaften, Band 220. Springer-Verlag, Berlin-New York, 1976

[3] J. von Neumann, *Proof of the quasi-ergodic hypothesis.* Proc. Nat. Acad. Sci. USA 18 (1932), 70–82.

[4] T. Tao, Lecture on von Neumann's Mean Ergodic Theorem, notes available at `http://terrytao.wordpress.com/2008/01/30/254a-lecture-8-the-mean-ergodic-theorem/`.

[5] University of Utah, Summer School on $SL_2(\mathbb{R})$ (May-June 2006), notes and videos available at `http://www.math.utah.edu/vigre/minicourses/sl2/`.

Extrapolation techniques and applications to row-action methods

Anna Karapiperi (*)

1 Introduction

Let (S_n) be a sequence of (real or complex) numbers which converges to S. We shall transform the sequence (S_n) into another sequence (T_n) and denote by T such a transformation.
For example, we can have

$$T_n = \frac{S_n S_{n+2} - S_{n+1}^2}{S_{n+2} - 2S_{n+1} + S_n}, \quad n = 0, 1, \ldots$$

which is the well-known Δ^2 **process** due to Aitken [1].

In order to present some practical interest, the new sequence (T_n) should satisfy, at least for some particular classes of convergent sequences (S_n), the following properties

(a) (T_n) must converge

(b) (T_n) must converge to the same limit as (S_n)

(c) (T_n) must converge to S faster than (S_n), that is $\lim_{n\to\infty} \frac{T_n - S}{S_n - S} = 0$.

In the last case, we say that the transformation T accelerates the convergence of the sequence (S_n).

The three aforementioned properties usually do not hold for all converging sequences (S_n) and, in particular, the last one. In fact, it has been proved that a universal transformation T able to accelerate all the converging sequences cannot exist [5].

In other words, every sequence transformation is only able to accelerate the convergence of certain classes of sequences. For example, Aitken's process accelerates the convergence

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: akarapi@math.unipd.it. Seminar held on May 7th, 2014.

of all the sequences for which $\exists\ \lambda \in [-1, +1[$ such that

$$\lim_{n\to\infty} \frac{S_{n+1} - S}{S_n - S} = \lambda.$$

However, properties 1 and 2 are not satisfied for all convergent sequences. There are examples of convergent sequences (S_n) for which the sequence (T_n) obtained by Aitken's process has two accumulation points (see [3, Section 2.3]). But it is also true that if such a (T_n) converges, then its limit is the same as the limit of the sequence (S_n) [12].

2 Extrapolation methods

In the study of a sequence transformation T an important notion is that of the *kernel* $\mathcal{K}_T$, that is the set of all the sequences (S_n) for which

$$\exists S \text{ such that } \forall n \geq N,\ T_n = S.$$

For instance, the kernel of Aitken's Δ^2 process is the set of sequences of the form

$$S_n = S + a\lambda^n \tag{1}$$

where $a \in \mathbb{C}$ is different from 0 and $\lambda \in \mathbb{C}$ is different from 0 and 1.

The equation (1) is the *explicit form* of the kernel since it gives explicitly the form of the sequences belonging to the kernel of Aitken's process. An equivalent expression is the so-called *implicit form* of the kernel, which for Δ^2 process is described by

$$S_{n+1} - S = \lambda(S_n - S) \tag{2}$$

The solution of the above difference equation is given by (1).

If the sequence to be accelerated belongs to the kernel of the transformation used then, by construction, $\forall n \geq N,\ T_n = S$. Usually, S is the limit of the sequence (S_n). But if (S_n) diverges, then S is called its anti-limit.

Now we are ready to give the definition of an extrapolation method.

Definition 1 A sequence transformation $T : (S_n) \longmapsto (T_n)$ is said to be an *extrapolation method* if it is such that

$$\forall n \geq N,\ T_n = S \quad \text{if and only if} \quad (S_n) \in \mathcal{K}_T.$$

Thus any sequence transformation can be viewed as an extrapolation method.

We shall now explain how a transformation T is built from its kernel. We already saw that the implicit form of the kernel consists of a relation among consecutive terms of the sequence, that is a relation of the form

$$R(S_n, ..., S_{n+q}, S) = 0 \tag{3}$$

which must be satisfied $\forall n \geq N$, if and only if (S_n) belongs to the kernel $\mathcal{K}_T$ of the transformation T. Given $S_n, S_{n+1}..., S_{n+q}$, we are looking for the sequence $(u_n) \in \mathcal{K}_T$ satisfying the interpolation conditions

$$u_i = S_i,\ i = n, ..., n+p+q.$$

Since (u_n) belongs to the kernel, then it satisfies the relation (3) that is $\forall i\ R(u_i, ..., u_{i+q}, S) = 0$, and the interpolation conditions impose

$$R(S_i, ..., S_{i+q}, S) = 0, \quad i = n, ..., n+p.$$

This system of $p+1$ equations with $p+1$ unknowns $\alpha_1, ..., \alpha_p, S$, has a solution (which we denote by T_n to recall that it depends on n, the index of the first interpolation condition) if the derivative of R with respect to S is different from zero. Under this assumption, the implicit function theorem implies the existence of a function G (depending on the unknown parameters $\alpha_1, ..., \alpha_p$) such that

$$S = G(S_i, ..., S_{i+q}),\ i = n, ..., n+p.$$

The solution $T_n := S$ of this system depends on $S_n, ..., S_{n+k}$, that is $T_n = F(S_n, ..., S_{n+p+q})$.

An example: We assume that R has the form

$$R(u_i, u_{i+1}, S) = \alpha_1(u_i - S) + \alpha_2(u_{i+1} - S) = 0,$$

with $\alpha_1\alpha_2 \neq 0$, $\alpha_1 + \alpha_2 \neq 0$. Thus we have to solve the system

$$\left\{ \begin{array}{lcl} \alpha_1(S_n - S) & + & \alpha_2(S_{n+1} - S) = 0 \\ \alpha_1(S_{n+1} - S) & + & \alpha_2(S_{n+2} - S) = 0 \end{array} \right.$$

The derivative of R with respect to the last variable is $-(\alpha_1 + \alpha_2) \neq 0$. Then G is given by $S = (\alpha_1 u_i + \alpha_2 u_{i+1})/(\alpha_1 + \alpha_2)$ and we have to solve the system

$$\left\{ \begin{array}{l} S = (\alpha_1 S_n + \alpha_2 S_{n+1})/(\alpha_1 + \alpha_2) \\ S = (\alpha_1 S_{n+1} + \alpha_2 S_{n+2})/(\alpha_1 + \alpha_2) \end{array} \right.$$

Without loss of generality, we can assume that $\alpha_1 + \alpha_2 = 1$, therefore the system to be solved becomes

$$\left\{ \begin{array}{lcl} S & = & \alpha_1 S_n + (1-\alpha_1) S_{n+1} \\ S & = & \alpha_1 S_{n+1} + (1-\alpha_1) S_{n+2} \end{array} \right. \tag{4}$$

or $0 = \alpha_1 \Delta S_n + (1-\alpha_1)\Delta S_{n+1}$ [(1)]. From the last relation we obtain $\alpha_1 = \dfrac{\Delta S_{n+1}}{\Delta^2 S_n}$ ($\Delta^2 S_n \neq 0$ since $\alpha_1 + \alpha_2 \neq 0$). Hence, $T_n = S = \dfrac{\Delta S_{n+1}}{\Delta^2 S_n} S_n + (1 - \dfrac{\Delta S_{n+1}}{\Delta^2 S_n}) S_{n+1}$, that is

$$T_n = \frac{S_n S_{n+2} - S_{n+1}^2}{\Delta^2 S_n} \tag{5}$$

[(1)] The difference operator Δ is defined by $\Delta u_n = u_{n+1} - u_n$ and $\Delta^{k+1} u_n = \Delta^k u_{n+1} - \Delta^k u_n$.

which is Δ^2 process, whose name is due to the Δ^2 appearing in the denominator.

Formula (5) is highly numerically unstable since, if S_n,S_{n+1} and S_{n+2} are almost equal, cancellation errors arise both in the numerator and in the denominator and T_n is badly computed [3]. For that reason, we use one of the following equivalent formulas

$$T_n = S_n - \frac{(\Delta S_n)^2}{\Delta^2 S_n} \tag{6}$$

$$= S_{n+1} - \frac{\Delta S_n \Delta S_{n+1}}{\Delta^2 S_n} \tag{7}$$

$$= S_{n+2} - \frac{(\Delta S_{n+1})^2}{\Delta^2 S_n}, \quad n = 0, 1, \ldots \tag{8}$$

Of course, cancellation errors again arise but in the correcting term to S_n,S_{n+1}, S_{n+2} respectively, thus formulas (6)–(8) are much more stable than formula (5).

2.1 Extrapolation algorithms

Since $T_n = S$, the system (4) can be written as follows

$$\begin{cases} T_n = \alpha_1 S_n + (1-\alpha_1) S_{n+1} \\ T_n = \alpha_1 S_{n+1} + (1-\alpha_1) S_{n+2} \end{cases} \tag{9}$$

We add and subtract S_n to the first equation and S_{n+1} to the second one. Making use of the difference operator Δ and setting $b = \alpha_1 - 1$ we obtain the following equivalent system

$$\begin{cases} S_n = T_n + b\Delta S_n \\ S_{n+1} = T_n + b\Delta S_{n+1} \end{cases}$$

The classical determinantal formulae for T_n gives the following ratio of determinants

$$T_n = \frac{\begin{vmatrix} S_n & S_{n+1} \\ \Delta S_n & \Delta S_{n+1} \end{vmatrix}}{\begin{vmatrix} 1 & 1 \\ \Delta S_n & \Delta S_{n+1} \end{vmatrix}}$$

which is equivalent to formula (5), that is Aitken's Δ^2 process.

2.1.1 Shanks' transformation

We now assume that R has the form

$$\alpha_1(\mathbf{u}_i - \mathbf{S}) + \alpha_2(\mathbf{u}_{i+1} - \mathbf{S}) + \ldots + \alpha_{k+1}(\mathbf{u}_{i+k} - \mathbf{S}) = 0$$

with $\alpha_1 \alpha_{k+1} \neq 0$, $\alpha_1 + \ldots + \alpha_{k+1} \neq 0$.
Repeating the previous steps (assuming that $\alpha_1 + \ldots + \alpha_{k+1} = 1$), we obtain the system

$$\begin{cases} \mathbf{S}_n = \mathbf{T}_n + b_1\Delta\mathbf{S}_n + \cdots + b_k\Delta\mathbf{S}_{n+k-1} \\ \mathbf{S}_{n+1} = \mathbf{T}_n + b_1\Delta\mathbf{S}_{n+1} + \cdots + b_k\Delta\mathbf{S}_{n+k} \\ \vdots = \vdots \\ \mathbf{S}_{n+k} = \mathbf{T}_n + b_1\Delta\mathbf{S}_{n+k} + \cdots + b_k\Delta\mathbf{S}_{n+2k-1} \end{cases} \tag{10}$$

The classical determinantal formulae gives

$$\mathbf{T}_n = \frac{\begin{vmatrix} \mathbf{S}_n & \mathbf{S}_{n+1} & \cdots & \mathbf{S}_{n+k} \\ \Delta\mathbf{S}_n & \Delta\mathbf{S}_{n+1} & \cdots & \Delta\mathbf{S}_{n+k} \\ \vdots & \vdots & & \vdots \\ \Delta\mathbf{S}_{n+k-1} & \Delta\mathbf{S}_{n+k} & \cdots & \Delta\mathbf{S}_{n+2k-1} \end{vmatrix}}{\begin{vmatrix} 1 & 1 & \cdots & 1 \\ \Delta\mathbf{S}_n & \Delta\mathbf{S}_{n+1} & \cdots & \Delta\mathbf{S}_{n+k} \\ \vdots & \vdots & & \vdots \\ \Delta\mathbf{S}_{n+k-1} & \Delta\mathbf{S}_{n+k} & \cdots & \Delta\mathbf{S}_{n+2k-1} \end{vmatrix}} \tag{11}$$

which is denoted by $e_k(S_n)$. This is the well-known **Shanks' transformation** [10].
Of course the computation of the above two determinants is prohibitive because of the time needed and mostly from a numerical point of view. However, every time that we have to compute determinants of a special structure, like in (11), it is possible to obtain some rules for computing recursively the ratio of determinants. Such an algorithm is called an *extrapolation algorithm*. For example, for Shanks' transformation we can use the **vector ε-algorithm** of Wynn, which is described by the following rules

$$\varepsilon_{-1}^{(n)} = 0 \in \mathbb{R}^N, \quad \varepsilon_0^{(n)} = \mathbf{S}_n \in \mathbb{R}^N, \quad n = 0, 1, ...$$

$$\varepsilon_{k+1}^{(n)} = \varepsilon_{k-1}^{(n+1)} + (\varepsilon_k^{(n+1)} - \varepsilon_k^{(n)})^{-1}, \quad k, n = 0, 1, ...$$

We have

$$\varepsilon_{2k}^{(n)} = e_k(\mathbf{S}_n).$$

2.1.2 Shanks' generalizations

The system (10) is equivalent to the system

$$\begin{cases} \mathbf{S}_n &= \mathbf{T}_n + b_1\Delta\mathbf{S}_n + \cdots + b_k\Delta\mathbf{S}_{n+k-1} \\ \Delta\mathbf{S}_n &= 0 + b_1\Delta^2\mathbf{S}_n + \cdots + b_k\Delta^2\mathbf{S}_{n+k-1} \\ \vdots &= \vdots \\ \Delta\mathbf{S}_{n+k-1} &= 0 + b_1\Delta^2\mathbf{S}_{n+k-1} + \cdots + b_k\Delta^2\mathbf{S}_{n+2k-2} \end{cases} \tag{12}$$

which results from (10) if we replace each equation, starting from the second one, by its difference with the preceding one. Always using the classical determinantal formulae for T_n, we obtain

$$e_k(\mathbf{S}_n) = \frac{\begin{vmatrix} \mathbf{S}_n & \Delta\mathbf{S}_n & \cdots & \Delta\mathbf{S}_{n+k-1} \\ \Delta\mathbf{S}_n & \Delta^2\mathbf{S}_n & \cdots & \Delta^2\mathbf{S}_{n+k-1} \\ \vdots & \vdots & & \vdots \\ \Delta\mathbf{S}_{n+k-1} & \Delta^2\mathbf{S}_{n+k-1} & \cdots & \Delta^2\mathbf{S}_{n+2k-2} \end{vmatrix}}{\begin{vmatrix} 1 & 0 & \cdots & 0 \\ \Delta\mathbf{S}_n & \Delta^2\mathbf{S}_n & \cdots & \Delta^2\mathbf{S}_{n+k-1} \\ \vdots & \vdots & & \vdots \\ \Delta\mathbf{S}_{n+k-1} & \Delta^2\mathbf{S}_{n+k-1} & \cdots & \Delta^2\mathbf{S}_{n+2k-2} \end{vmatrix}} \tag{13}$$

We observe that (11) and (13) can be unified under the following formula

$$R_k = \frac{\begin{vmatrix} \mathbf{e}_0 & \cdots & \mathbf{e}_k \\ a_1^{(0)} & \cdots & a_1^{(k)} \\ \vdots & & \vdots \\ a_k^{(0)} & \cdots & a_k^{(k)} \end{vmatrix}}{\begin{vmatrix} c_0 & \cdots & c_k \\ a_1^{(0)} & \cdots & a_1^{(k)} \\ \vdots & & \vdots \\ a_k^{(0)} & \cdots & a_k^{(k)} \end{vmatrix}}$$

Indeed, if we set $\mathbf{e}_i = \mathbf{S}_{n+i}$, $a_j^{(i)} = \Delta\mathbf{S}_{n+i+j-1}$, $c_i = 1$, $i = 0, ...,$, $j = 1, ..., k$, we recover the ratio (11), while the options $\mathbf{e}_0 = \mathbf{S}_n$, $a_j^{(0)} = \Delta\mathbf{S}_{n+i-1}$, $c_0 = 1$ and $\mathbf{e}_i = \mathbf{S}_{n+i-1}$, $a_j^{(i)} = \Delta^2\mathbf{S}_{n+i+j-2}$, $c_i = 0$ for $i \geq 1$ give formula (13).

When $\mathbf{e}_i = \mathbf{S}_{n+i}$, $a_j^{(i)} = \Delta\mathbf{S}_{n+i+j-1}$, $c_i = 1$, if we choose $a_i^{(j)} = (\Delta\mathbf{S}_{n+i-1}, \Delta\mathbf{S}_{n+j})$ then we obtain the Minimal Polynomial Extrapolation method (**MPE**), whereas for $a_i^{(j)} = (\Delta^2\mathbf{S}_{n+i-1}, \Delta\mathbf{S}_{n+j})$ we have the Reduced Rank Extrapolation method (**RRE**). For more information on these methods see [3] and the references therein.

3 Row-action methods

Let us consider the system of linear equations $A\mathbf{x} = \mathbf{b}$, where A is a real $M \times N$ matrix, initially assumed to be nonsingular. If we denote by $\mathbf{a}_i = A^T\mathbf{e}_i$ the column vector formed by the i-th row of A, and by b_i the i-th component of the right hand side $\mathbf{b}$, then the solution $\mathbf{x}_* = A^{-1}\mathbf{b}$ is the unique intersection point of the M hyperplanes in $\mathbb{R}^N$

$$\langle \mathbf{a}_i, \mathbf{x} \rangle = b_i, \ i = 1, 2, ..., N\,. \tag{14}$$

These hyperplanes play a fundamental role in row-action methods which can be used for solving a (possibly large) system of linear equations. Their great interest arise from the fact that they use no matrix-vector product, that is they don't work with the whole matrix A, but only with its rows (this explains the name of these methods).

Row-action methods can be categorized in two classes, Algebraic Reconstruction Techniques (ART) and Simultaneous Iterative Reconstruction Techniques (SIRT). One representative of each class is Kaczmarz [8] and Cimmino [4] method, respectively. Let us see how these two methods work.

Kaczmarz method starts from an arbitrary initial vector $\mathbf{x}_0$, and projects it to the first hyperplane (the one defined by the first row of the matrix). The new point, call it $\mathbf{p}_1$, is projected to the next hyperplane (the one defined by the second row of the matrix) so we obtain $\mathbf{p}_2$, and so on. Once we have passed from all the hyperplanes, one cycle of the method has been completed and we obtain the first approximation $\mathbf{x}_1$, which is simply the

last projected point $\mathbf{p}_M$. Afterwards, we take $\mathbf{x}_1$ and we repeat the previous steps, so that at the end of the second cycle we obtain $\mathbf{x}_2$ and so on. In other words, one iteration of Kaczmarzs method consists in a complete cycle of consecutive projections in their natural order, that is,

$$\begin{cases} \mathbf{p}_0 &= \mathbf{x}_n \\ \mathbf{p}_i &= \mathbf{p}_{i-1} + \dfrac{b_i - \langle \mathbf{p}_{i-1}, \mathbf{a}_i \rangle}{\|\mathbf{a}_i\|^2} \mathbf{a}_i \,, \quad i = 1, ..., M \\ \mathbf{x}_{n+1} &= \mathbf{p}_M \end{cases}$$

Cimmino followed another approach using simultaneous projections[(2)]. Starting from $\mathbf{x}_0$, he projects it to all the hyperplanes and takes as $\mathbf{x}_1$ the centroid of the projections $\mathbf{p}_1, ..., \mathbf{p}_M$. Then, he takes $\mathbf{x}_1$, he projects again to all the hyperplanes and obtains $\mathbf{x}_2$ as their centroid. So, one iteration of Cimmino's method can be described by the following scheme

$$\begin{cases} \mathbf{p}_0 &= \mathbf{x}_n \\ \mathbf{p}_i &= \mathbf{p}_0 + \dfrac{b_i - \langle \mathbf{p}_0, \mathbf{a}_i \rangle}{\|\mathbf{a}_i\|^2} \mathbf{a}_i \,, \quad i = 1, ..., M \\ \mathbf{x}_{n+1} &= \dfrac{1}{M} \displaystyle\sum_{i=1}^{M} \mathbf{p}_i \end{cases} .$$

In Figure 1 we see the first three iterations (cycles) of Kaczmarz and Cimmino method applied to the system $A\mathbf{x} = \mathbf{b}$, with

$$A = \begin{bmatrix} 1 & 2 \\ -1 & 3 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, \text{ and exact solution } \mathbf{x}_* = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

The two hyperplanes (lines in this case) involved are

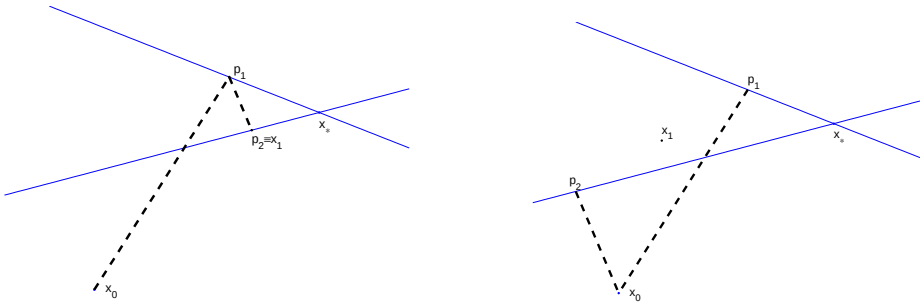
$$\begin{aligned} x_1 + 2x_2 &= 3 \\ -x_1 + 3x_2 &= 2 \end{aligned}$$

Starting form $\mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ we see how the two methods approach the solution. It is evident that in this example Kaczmarz method converges to the exact solution $\mathbf{x}_*$ faster than Cimmino method. However, they are both slow (since we deal with a 2×2 system, 3 iterations are already many). The good news is that the methods of Kaczmarz and Cimmino always converge to the solution[(3)] of the system (see [6] and [4] respectively).

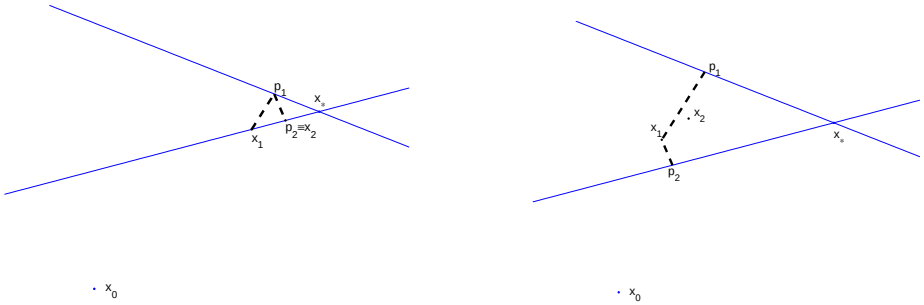
For that reason it worths a try to accelerate their convergence and we will do it using the extrapolation methods introduced in the previous section.

[(2)] Cimmino in [4] considers reflections, instead of projections, and also a weighted centroid. Here for simplicity we present the variant with the projections, which are also used in Kaczmarz method, and all the weights are set equal to 1.

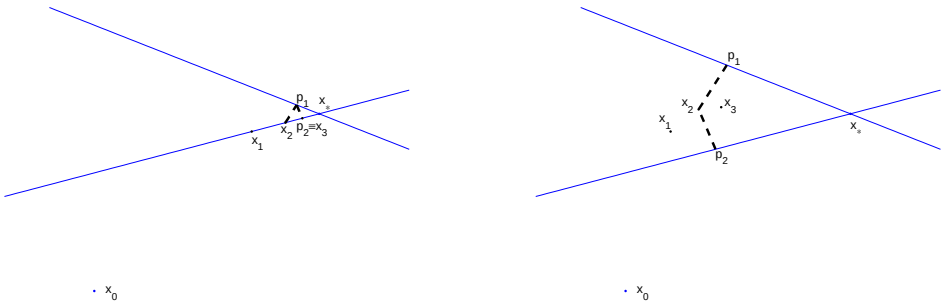
[(3)] In the case of more than one solutions the methods convergence to the least-square solution.



(a) 1st cycle



(b) 2nd cycle



(c) 3rd cycle

Figure 1: The first three iterations (cycles) of the methods of Kaczmarz (left) and Cimmino (right) applied to a 2-dimensional problem.

3.1 Convergence acceleration

The sequence of vectors $(\mathbf{x}_n)$ obtained by Kaczmarz or Cimmino method satisfies the following relation

$$\mathbf{x} - \mathbf{x}_n = Q^n(\mathbf{x} - \mathbf{x}_0)\,, \quad n = 0, 1, ... \tag{15}$$

with

$$Q = \begin{cases} Q_M \cdots Q_1, & Q_i = I - \dfrac{\mathbf{a}_i \mathbf{a}_i^T}{\|\mathbf{a}_i\|^2}, & \text{for Kaczmarz method} \\ I - \dfrac{1}{M} A^T D A, & D = \operatorname{diag}\left(\dfrac{1}{\|\mathbf{a}_i\|^2}\right), , & \text{for Cimmino method.} \end{cases}$$

We consider the minimal polynomial of the matrix Q for the vector $\mathbf{x} - \mathbf{x}_0$, that is the polynomial Π_ν of smallest degree $\nu \leq N$ such that $\Pi_\nu(Q)(\mathbf{x} - \mathbf{x}_0) = 0$. Let $\Pi_\nu(\xi) = c_0 + c_1\xi + ... + c_\nu\xi^\nu$, thus from (15), it follows

$$Q^n\Pi_\nu(Q)(\mathbf{x}-\mathbf{x}_0) = c_0(\mathbf{x}-\mathbf{x}_n)+c_1(\mathbf{x}-\mathbf{x}_{n+1})+\ldots+c_\nu(\mathbf{x}-\mathbf{x}_{n+\nu}) = 0, \quad n = 0, 1, ... \tag{16}$$

The vector $\mathbf{x}$ can be exactly computed if we are able to build a sequence transformation whose kernel consists of sequences of the form (16). However, usually in practical applications, ν is quite large. For that reason, we build a sequence transformation whose kernel contains sequences of the form

$$a_0^{(n)}(\mathbf{x} - \mathbf{x}_n) + a_1^{(n)}(\mathbf{x} - \mathbf{x}_{n+1}) + \ldots + a_k^{(n)}(\mathbf{x} - \mathbf{x}_{n+k}) = 0 \tag{17}$$

for some $k \leq \nu$. Solving this equation for the vector $\mathbf{x}$, we will obtain an approximation $\mathbf{y}_k^{(n)}$ depending on k and n.

First of all, we have to compute the unknown coefficients $a_0^{(n)}, \ldots, a_k^{(n)}$. The vector sequence transformations introduced in Section 2.1.2, namely MPE and RRE, include in their kernel sequences of the form (17). Therefore, we will use them in order to obtain a system of linear equations whose solution is $a_0^{(n)}, \ldots, a_k^{(n)}$ and then compute $\mathbf{y}_k^{(n)} = a_0^{(n)}\mathbf{x}_n + ... + a_k^{(n)}\mathbf{x}_{n+k}$. Here is how we proceed.

Step 1: We write (17) for the indices n and $n + 1$, subtract and multiply scalarly by a vector $\mathbf{y}$. Hence, we obtain

$$a_0^{(n)}(\mathbf{y}, \Delta\mathbf{x}_n) + a_1^{(n)}(\mathbf{y}, \Delta\mathbf{x}_{n+1}) + ... + a_k^{(n)}(\mathbf{y}, \Delta\mathbf{x}_{n+k}) = 0\,. \tag{18}$$

Step 2: We construct the system

$$\begin{cases} a_0^{(n)} + ... + a_k^{(n)} = 1 & \text{(normalization condition)} \\ d_{i,0}^{(n)} a_0^{(n)} + ... + d_{i,k}^{(n)} a_k^{(n)} = 0, & i = 1, \ldots, k\,. \end{cases}$$

with the k equations formed by one of the following:

- one vector $\mathbf{y}$ and the equation (18) written for the indices $n, ...n + k - 1$.
- k linearly independent vectors $\mathbf{y}$ and the equation (18) written only for the index n.

- several linear independent vectors $\mathbf{y}$ and the equation(18) written for several indices .

The coefficients $d_{i,j}^{(n)}$, $i = 1, \ldots, k$, $j = 0, \ldots, k$ are given according to the chosen strategy, as follows

$$d_{i,j}^{(n)} = \begin{cases} (\Delta\mathbf{x}_{n+i-1}, \Delta\mathbf{x}_{n+j}), & \text{if used the MPE} \\ (\Delta^2\mathbf{x}_{n+i-1}, \Delta\mathbf{x}_{n+j}), & \text{if used the RRE.} \end{cases}$$

Then, for a fixed value of k,

$$\mathbf{y}_k^{(n)} = a_0^{(n)}\mathbf{x}_n + ... + a_k^{(n)}\mathbf{x}_{n+k} = \frac{\begin{vmatrix} \mathbf{x}_n & \cdots & \mathbf{x}_{n+k} \\ d_{1,0}^{(n)} & \cdots & d_{1,k}^{(n)} \\ \vdots & & \vdots \\ d_{k,0}^{(n)} & \cdots & d_{k,k}^{(n)} \end{vmatrix}}{\begin{vmatrix} 1 & \cdots & 1 \\ d_{1,0}^{(n)} & \cdots & d_{1,k}^{(n)} \\ \vdots & & \vdots \\ d_{k,0}^{(n)} & \cdots & d_{k,k}^{(n)} \end{vmatrix}}$$

Of course, the vector ε-algorithm of Wynn can be also used, since the kernel of Shanks' transformation contains sequences of the form (17). However, in this case an underlying system of linear equations for $\mathbf{y}_k^{(n)}$ does not exist and the algorithm is applied directly to the sequence $(\mathbf{x}_n)$ after defining the inverse $\mathbf{u}^{-1}$ of a vector $\mathbf{u}$ as $\mathbf{u}^{-1} = \mathbf{u}/(\mathbf{u}, \mathbf{u})$.

3.2 Accelerated and Restarted algorithms

The aforementioned extrapolation techniques can be used in two different ways in order to accelerate the convergence of the sequence $(\mathbf{x}_n)$ produced by the original method, that is Kaczmarz or Cimmino. Since the number of the vectors $\mathbf{x}_i$ required by the various algorithms for the computation of $\mathbf{y}_k^{(n)}$ is not the same (the ε-algorithm needs $2k + 1$ terms, whereas MPE and RRE only $k + 2$), we denote by $\ell + 1$ this number, thus $\ell = 2k$ for the vector ε-algorithm, and $\ell = k + 1$ for MPE and RRE.

Now we can present the two strategies used by Brezinski, Redivo-Zaglia in [2] for the acceleration of Kaczmarz method. Here we apply them also to Cimmino method.

Restarted Algorithm: In the first strategy, we start with $\mathbf{x}_0$ and we compute $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$ by using the original method. Then we apply one of the extrapolation algorithms implementing a sequence transformation on the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$, and we obtain $\mathbf{y}_k^{(0)}$, which we consider the new initial vector let us call it $\mathbf{z}_0$. Now we apply again the original method to the new $\mathbf{x}_0$, that is to the vector $\mathbf{z}_0 = \mathbf{y}_k^{(0)}$, we compute the new sequence of ℓ vectors and we apply to them the extrapolation algorithm in order to obtain $\mathbf{z}_1 = \mathbf{y}_k^{(0)}$. We restart from $\mathbf{x}_0$ that now is the vector $\mathbf{z}_1$, and so on.

Accelerated Algorithm: In this strategy we apply one of the extrapolation algorithms on the sequence $\mathbf{x}_0, \mathbf{x}_1, \ldots$ given by the original method and, after fixing the index k, we build simultaneously the sequence $\mathbf{z}_0 = \mathbf{y}_k^{(0)}, \mathbf{z}_1 = \mathbf{y}_k^{(1)}, \ldots$. Therefore, the computation of $\mathbf{y}_k^{(0)}$ can only begin after the iterate $\mathbf{x}_\ell$ has been computed, but the computation of each new transformed vector needs only one new iterate of the original method.

Table 1 displays the scheme of the Restarted and Accelerated algorithms as described above.

$$
\begin{array}{lllll}
\mathbf{x}_0 & & & & \mathbf{x}_0 \\
\mathbf{x}_1 & & & & \mathbf{x}_1 \\
\vdots & & & & \vdots \\
\mathbf{x}_\ell \longrightarrow \mathbf{z}_0 \equiv \mathbf{x}_0 & & & & \mathbf{x}_\ell \longrightarrow \mathbf{z}_0 \\
\quad \mathbf{x}_1 & & & & \mathbf{x}_{\ell+1} \longrightarrow \mathbf{z}_1 \\
\quad \vdots & & & & \mathbf{x}_{\ell+2} \longrightarrow \mathbf{z}_2 \\
\quad \mathbf{x}_\ell \longrightarrow \mathbf{z}_1 \equiv \mathbf{x}_0 & & & & \vdots \quad \vdots \\
\qquad \mathbf{x}_1 & & & & \\
\qquad \vdots & & & & \\
\qquad \mathbf{x}_\ell \longrightarrow \mathbf{z}_2 \equiv \mathbf{x}_0 & & & & \\
\qquad \vdots \quad \ddots & & & &
\end{array}
$$

Table 1: Explanation scheme that describes Restarted (left) and Accelerated (right) algorithms.

4 Numerical experiments

The numerical experiments presented in this section were performed using `Matlab 7.12.0`. Except for the Euclidean norms of the errors (in logarithmic scale), for the comparison of the acceleration brought by each procedure, we give also the ratios of the norms of the errors between the iterate $\mathbf{z}_n$ obtained by the Accelerated or Restarted algorithm and the iterate of the original method with the highest index used in its construction (for the Accelerated algorithm) or in the case of the Restarted algorithm, the iterate with the highest index that would have been used if we had let the method continue without restarting it, that is

$$\frac{\|\mathbf{z}_n - \mathbf{x}\|}{\|\mathbf{x}_{n+\ell} - \mathbf{x}\|} \quad \text{(Accelerated)} \qquad \frac{\|\mathbf{z}_n - \mathbf{x}\|}{\|\mathbf{x}_{(n+1)(\ell+1)} - \mathbf{x}\|} \quad \text{(Restarted)}$$

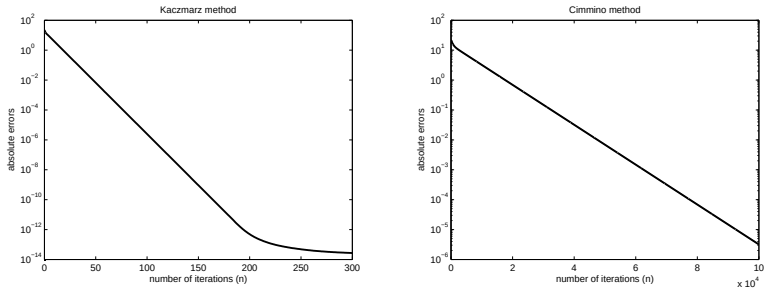
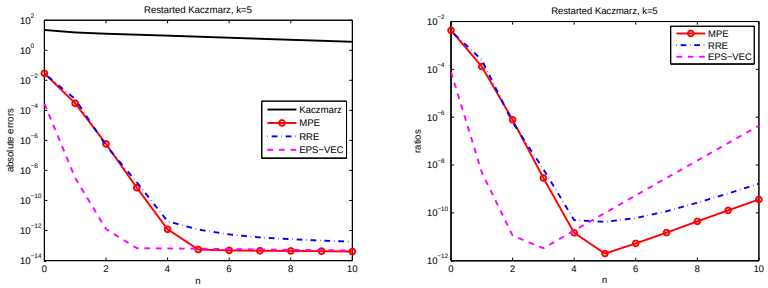


Figure 2: Convergence of the methods Kaczmarz (left) and Cimmino (right) for *parter* matrix, $M = N = 500$.

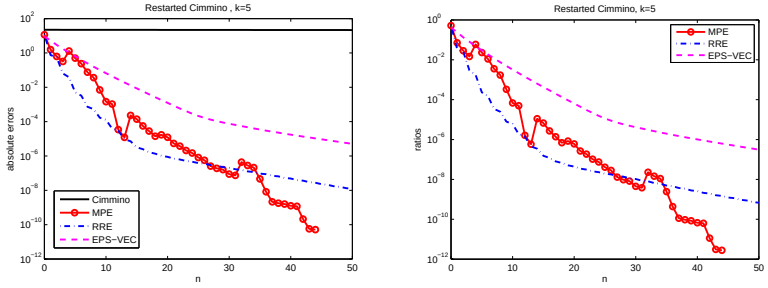
The first test matrix, *parter*, is taken from the gallery set of `Matlab`. It is a 500×500 Toeplitz matrix with singular values near π and condition number 4.21943. The solution of

the system was set $\mathbf{x} = (1, \ldots, 1)^T$, and the right-hand side $\mathbf{b}$ was computed accordingly. In Figure 2 we see how slow is the convergence of the original methods (Kaczmarz and Cimmino).

In Figure 3(a) we observe that Restarted algorithm makes all the extrapolation methods reach a good precision. In fact, the acceleration of Kaczmarz method is impressive. For instance, ε-algorithm gives an error of 10^{-13} only in three iterations. Note that when the method has attained a good precision, then the ratios increase. This happens because the error of Kaczmarzs method continues to decrease slowly, while the error of the extrapolation methods almost stagnate.



(a) Restarted Kaczmarzs errors (left) and ratios (right).



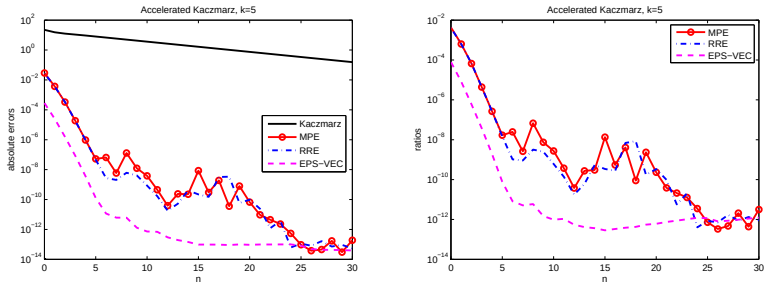
(b) Restarted Cimmino errors (left) and ratios (right).

Figure 3: Restarted algorithm: *parter* matrix, $M = N = 500$, $k = 5$.

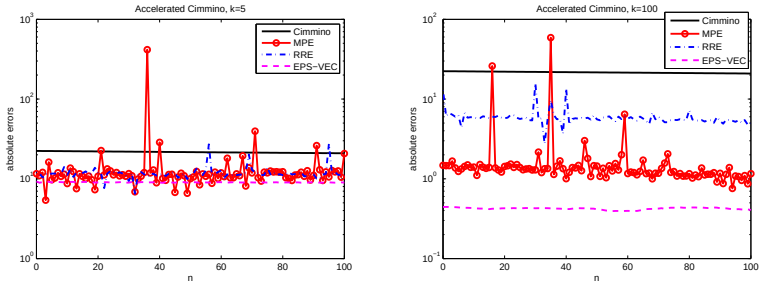
In the case of Cimmino method, MPE performs better than the other extrapolation methods (see Figure 3(b)). However, in order to reach a good precision all methods need much more iterations than when used the Restarted Kaczmarz algorithm. This comes as no surprise, since the original Cimmino method converges extremely slow, which means that the ℓ vectors used for the extrapolation procedure are quite close to each other. Nevertheless, if we do not compare with Kaczmarz method which for this example is by its own faster than Cimmino, the Restarted Cimmino algorithm gives very satisfactory results, given the small convergence rate of the original method.

More or less the same observations one can make for the Accelerated Kaczmarz algorithm in Figure 4(a). This time more iterations are needed, but we still reach a good

precision quite fast. On the other hand, Accelerated Cimmino seems to fail for all the extrapolation methods (see Figure 4(b)). The reason is the slow convergence of the original method, which implies that the vectors used for the extrapolation procedure are almost equal. This results to the slow convergence of the Accelerated Cimmino algorithm, which simply recycles the "bad information". Even if we try a large value of k like $k = 100$, the results do not change significantly. Obviously, using an even larger k could give a better approximation but at the same time it would be computationally expensive. Therefore, we may conclude that when the original method has an extremely slow rate of convergence, then Accelerated algorithm cannot save the situation.



(a) Accelerated Kaczmarzs errors (left) and ratios (right).



(b) Accelerated Cimmino errors with $k = 5$ (left) and $k = 100$ (right).

Figure 4: Accelerated algorithm: *parter* matrix, $M = N = 500$.

An important point to investigate is which stopping criterion to use. Usually iterative methods are stopped by using the residual but in our case the computation of a matrix-vector product would cancel one of the advantages of the row-action methods. Instead, we make use of our observation on the ratio figures. Since, the results given by the acceleration procedures stagnate when some precision is attained while those of the original methods continue to decrease, we decide to stop the iterations as soon as the following ratios increase significantly

$$\frac{\|\Delta \mathbf{z}_n\|}{\|\Delta \mathbf{x}_{n+\ell}\|} \quad \text{(Accelerated)} \qquad \frac{\|\Delta \mathbf{z}_n\|}{\|\Delta \mathbf{x}_{(n+1)(\ell+1)}\|} \quad \text{(Restarted)}$$

In Figure 5 we see how reliable is this stopping criterion when used for the vector ε-algorithm both with Restarted and Accelerated algorithm.

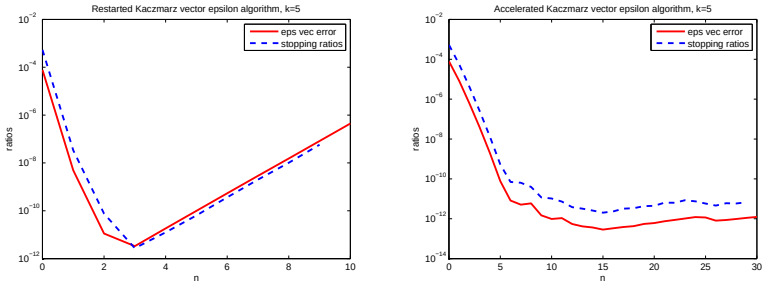


Figure 5: Restarted (left) and Accelerated (right) algorithm for Kaczmarz method: errors and stopping ratios for the vector ε-algorithm: *parter* matrix, $M = N = 500$.

4.1 Image reconstruction

An interesting application of row-action methods is image reconstruction. The next test matrix is taken from the AIR toolbox by P. C. Hansen and M. Saxild-Hansen [7]. It is called *paralleltomo* and it is a two-dimensional parallel-beam tomography test problem[(4)]. The matrix A has dimensions 2700×400. At the right-hand side we have added noise $\mathbf{e} = 10^{-3}$. The exact solution is the modified Shepp-Logan phantom head from [11].

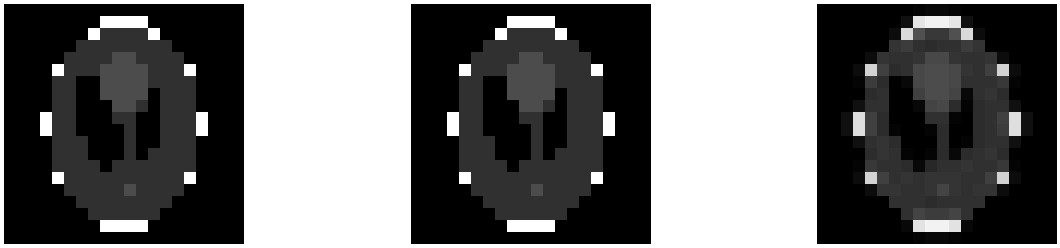


Figure 6: Shepp-Logan phantom, N=20. (a) Exact solution. (b) Kaczmarz, 15 iterations. (c) Cimmino, 50 iterations

Figure 6 displays the exact solution of the problem and the reconstruction brought by the methods of Kaczmarz and Cimmino. For Kaczmarz method we see that after only 15 iterations gives a quite faithful reconstruction of the Shepp-Logan phantom. On the other hand, Cimmino method even after 50 iterations does not succeed to approximate well the exact solution.

Regarding the acceleration techniques, if we see Figure 7 we conclude that Restarted algorithm offers an important acceleration of the convergence of Cimmino method, where

[(4)] Here we used $N = 20$, $\theta = 0 : 5 : 179$, $p = 75$. For more information on this problem refer to [9].

the original method has a slow rate of convergence. In the case of Kaczmarz method the gain is not so much, since the method by its own converges quite fast. It is important to stress that the error 10^{-2} attained by all the methods is a good precision in this case, since our problem has a noise at the level of 10^{-3}. Also here we observe a slight advantage of the vector ε-algorithm, but the difference is minor if we take into consideration the information that each iteration of the ε-algorithm uses more vectors than MPE and RRE. The last ones use exactly the same number of vectors, hence they always perform quite similar.

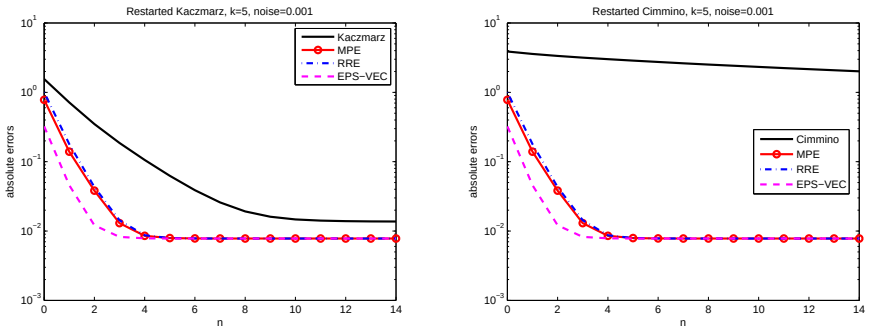


Figure 7: Restarted algorithm errors: *paralleltomo*, 2700×400, $k = 5$, noise=10^{-3}.

Figure 8 displays the reconstruction brought by the vector ε-algorithm, but exactly the same picture we obtain from MPE and RRE since, as shown in Figure 7, all the methods at the end reach the same accuracy.

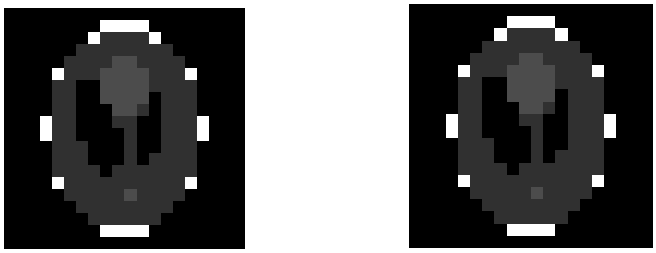
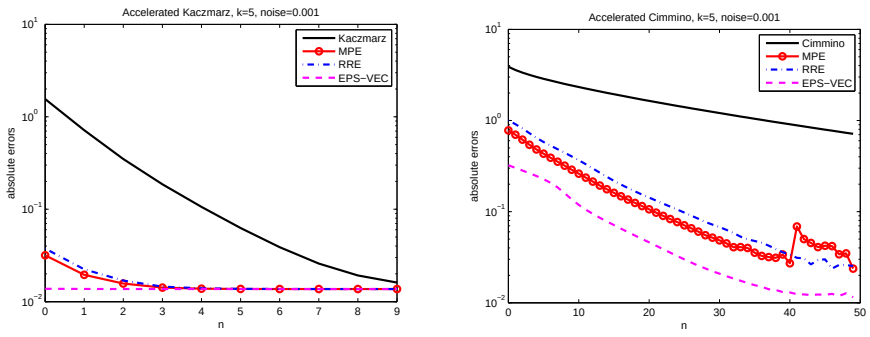


Figure 8: Restarted Kaczmarz (left, 15 iterations) and Restarted Cimmino (right, 15 iterations) algorithm, vector ε-algorithm reconstruction of the *paralleltomo* phantom.

Quite interesting are the results when used the Accelerated algorithm (see Figure 9). In the case of Kaczmarz method, all the extrapolation methods converge immediately, something that does not happen with Cimmino method. Since the original method of Cimmino is too slow, the extrapolation techniques have a small convergence rate compared to Accelerated Kaczmarz, but they still reach a good precision.

Figure 9: Accelerated algorithm errors: *paralleltomo*, 2700×400, $k = 5$, noise=10^{-3}.



Indeed, Figure 10 shows that the reconstruction brought by the vector ε-algorithm with Accelerated Kaczmarz and Accelerated Cimmino algorithms is the same, but after a different number of iterations.

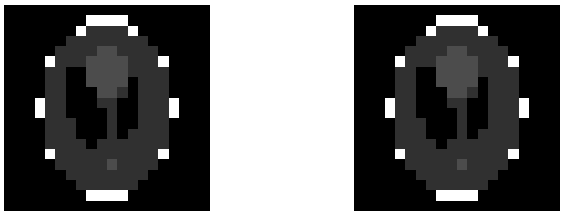


Figure 10: Accelerated Kaczmarz (left, 4 iterations) and Restarted Cimmino (right, 50 iterations) algorithm, Vector ε-algorithm reconstruction of the *paralleltomo* phantom.

References

[1] A. Aitken, *On Bernoulli's numerical solution of algebraic equations.* Proc. Royal Society of Edinburgh 46 (1926), 289–305.

[2] C. Brezinski, M. Redivo-Zaglia, *Convergence acceleration of Kaczmarz's method.* Journal of Engineering Mathematics, DOI 10.1007/s10665-013-9656-3, Article in Press.

[3] C. Brezinski, M. Redivo-Zaglia, "Extrapolation Methods. Theory and Practice". North-Holland, Amsterdam, 1991.

[4] C. Cimmino, *Calcolo approssimato per le soluzioni dei sistemi di equazioni lineari.* La Ricerca Scientifica, II, 9 (1938), 326–333.

[5] J. P. Delahaye, B. Germain-Bonne, *Résultats négatifs en accélération de la convergence.* Numerische Mathematik 35 (1980) 443–457.

[6] N. Gastinel, "Numerical linear algebra". Academic Press, New York, 1970 (English translation of *Analyse Numérique Linéaire*, Hermann, Paris, 1966).

[7] P. C. Hansen, M. Saxild-Hansen, *AIR Tools - A MATLAB package of algebraic iterative reconstruction methods.* Journal of Computational and Applied Mathematics 236 (2012), 2167–2178.

[8] S. Kaczmarz, *Angenäherte Auflösung von Systemen linearer Gleichungen.* Bull. Acad. Pol. Sci. A35 (1937), 355–357 (English translation: *Approximate solution of systems of linear equations*, Int. J. Control 57 (1993), 1269–1271.

[9] A. C. Kak, M. Slaney, "Principles of Computerized Tomographic Imaging". SIAM, 2001.

[10] D. Shanks, *Non linear transformations of divergent and slowly convergent sequences.* J. Math. Phys. 34 (1955), 1-42.

[11] P. Toft, *The Radon Transform, Theory and Implementation.* Ph.D. Thesis, DTU Informatics, Technical University of Denmark (1996), 199–201.

[12] R. R. Tucker, *The Δ^2-process and related topics.* Pacific J. Math. 22 (1967), 349–359.

A visual introduction to Tilting

Jorge Vitória (*)

Abstract. The representation theory of a quiver (i.e., an oriented graph) can sometimes be understood by... another quiver! Such pictures of complex concepts (such as categories of modules or derived categories) are a source of intuition for many phenomena, among which lie the tools for the classification and comparison of representations: Tilting Theory. The aim of this text is to give an heuristic view (example driven) of some ideas in this area of algebra.

Representation theory is an area of mathematics that studies the properties of *actions* of certain abstract objects (such as groups or rings) on other objects (such as abelian groups or vector spaces). One of the recurrent problems in the area (and in many other areas of mathematics) is the problem of classifying these actions, which are called **representations**. Tilting theory is a set of tools and techniques that allows to compare representations and, in a sense that we aim to make slightly more precise in this text, classify them. We will focus in particular on the representation theory of a quiver - which corresponds to studying the action of the associated path algebra on a vector space (see [2] for details on this correspondence).

1 Quivers and Representations

A **quiver** Q is an oriented graph. We denote by Q_0 its **vertices** and by Q_1 its **edges**. The $\mathbb{C}$-vector space whose basis elements are all **paths** in Q is denoted by $\mathbb{C}Q$.

Example 1.1 $Q = \ 1 \xrightarrow{\alpha} 2 \xrightarrow{\beta} 3\,, \quad Q_0 = \{1,2,3\}$ and $Q_1 = \{\alpha, \beta\}$;
$\mathbb{C}Q$ is a six-dimensional $\mathbb{C}$-vector space with basis $\mathbb{P} = \{e_1, e_2, e_3, \alpha, \beta, \beta\alpha\}$, where e_1, e_2 and e_3 are **lazy paths** (paths *starting at a vertex and ending at the same vertex without going through any arrow*) and $\beta\alpha$ is the long path going from vertex 1 to vertex 3. Note that for long paths we use a notation similar to that of composition of maps: $\beta\alpha$ can be read β after α.

Example 1.2 $Q = \ 1 \circlearrowleft \gamma\,, \quad Q_0 = \{1\}$ and $Q_1 = \{\gamma\}$;
$\mathbb{C}Q$ is an infinite-dimensional $\mathbb{C}$-vector space with basis $\mathbb{P} = \{e_1, \gamma^n : n \in \mathbb{N}\}$.

(*)Dipartimento di Informatica - Settore di Matematica, Università degli Studi di Verona, Strada le Grazie 15 - Ca' Vignal, I-37134 Verona, Italy; E-mail: jorge.vitoria@univr.it. Seminar held on May 21st, 2014.

The examples suggest a further operation on the vector space of paths: **concatenation of paths**. In the first example, for instance, it is evident that the path given by going first from 1 to 2 by α and then from 2 to 3 by β yields the long path $\beta\alpha$. When concatenation is not possible, we set it to be zero. These rules yield a *multiplication* on the vector space $\mathbb{C}Q$. The sum of all the lazy paths acts as a **multiplicative identity** (check!) on any path. Thus, $\mathbb{C}Q$ has a ring structure and we call it **the path algebra of** Q.

Example 1.3 For the quiver $Q = \ 1 \xrightarrow{\alpha} 2 \xrightarrow{\beta} 3$, $\mathbb{C}Q$ is a finite-dimensional $\mathbb{C}$-vector space with basis $\mathbb{P} = \{e_1, e_2, e_3, \alpha, \beta, \beta\alpha\}$. Given two elements:

$$\Phi = \lambda_1 e_1 + \lambda_2 e_2 + \lambda_3 e_3 + \lambda_4 \alpha + \lambda_5 \beta + \lambda_6 \beta\alpha, \quad \Psi = \mu_1 e_1 + \mu_2 e_2 + \mu_3 e_3 + \mu_4 \alpha + \mu_5 \beta + \mu_6 \beta\alpha$$

with λ_i, μ_i in $\mathbb{C}$, the multiplication $\Phi\Psi$ is defined **distributively**, multiplying the scalars and using the concatenation rules (such as, for example, $e_1\alpha = 0,\ \beta e_2 = \beta,\ e_2 e_1 = 0 = e_1 e_2,\ \beta\alpha = \beta\alpha$). As before suggested, our notation for concatenation is analogous to the composition of functions. Therefore, $e_1\alpha$ should read e_1 *after* α, which is visibly an impossible path and, thus, we set its concatenation to be zero.

The following easy facts (which we leave as exercises) relate the combinatorics of the quiver and the algebraic structure of the path algebra. Note that these path algebras are certainly not unfamiliar - the reader will have come across versions of them in any introductory course of algebra, as the statements below show.

Exercise 1.4 (1) The path algebra $\mathbb{C}Q$ of the quiver $Q = \ 1 \xrightarrow{\alpha} 2 \xrightarrow{\beta} 3$ is isomorphic, as a ring, to the ring of lower triangular matrices $\begin{pmatrix} \mathbb{C} & 0 & 0 \\ \mathbb{C} & \mathbb{C} & 0 \\ \mathbb{C} & \mathbb{C} & \mathbb{C} \end{pmatrix}$.

(2) The path algebra $\mathbb{C}Q$ of the quiver $Q = \ 1 \circlearrowright \gamma$ is isomorphic, as a ring, to the polynomial ring $\mathbb{C}[X]$.

(3) The vector space $\mathbb{C}Q$ of a quiver Q is finite dimensional if and only if Q has no oriented cycles.

Definition 1.5 A **representation** of a quiver Q is a pair $((V_i)_{i\in Q_0}, (f_\alpha)_{\alpha\in Q_1})$ where each V_i is a $\mathbb{C}$-vector space and for any arrow $\alpha : i \to j$, f_α is a linear map $V_i \to V_j$. A **morphism between representations** of Q

$$\phi : ((V_i)_{i\in Q_0}, (f_\alpha)_{\alpha\in Q_1}) \longrightarrow ((W_i)_{i\in Q_0}, (g_\alpha)_{\alpha\in Q_1})$$

is a family $(\phi_i)_{i\in Q_0}$ of linear maps $\phi_i : V_i \to W_i$ such that, for any arrow $\alpha : i \to j$ in Q_1, the diagram commutes

$$\begin{array}{ccc} V_i & \xrightarrow{f_\alpha} & V_j \\ \downarrow{\scriptstyle \phi_i} & & \downarrow{\scriptstyle \phi_j} \\ W_i & \xrightarrow{g_\alpha} & W_j \end{array}$$

The morphism ϕ is said to be an **isomorphism** if all the ϕ_i's are isomorphisms of vector spaces.

Example 1.6 The following are examples of representations of $Q = \ 1 \xrightarrow{\alpha} 2 \xrightarrow{\beta} 3$:

$$M := \ \mathbb{C}^2 \xrightarrow{\begin{pmatrix} 1 & 0 \end{pmatrix}} \mathbb{C} \xrightarrow{0} 0, \qquad N := \ \mathbb{C} \xrightarrow[\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}]{} \mathbb{C}^3 \xrightarrow[\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}]{} \mathbb{C}^2$$

A morphism between M and N can be given, for example, as follows.

$$\begin{array}{ccccc} \mathbb{C}^2 & \xrightarrow{\begin{pmatrix} 1 & 0 \end{pmatrix}} & \mathbb{C} & \xrightarrow{0} & 0 \\ {\scriptstyle \begin{pmatrix} 1 & 0 \end{pmatrix}} \downarrow & & \downarrow {\scriptstyle \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}} & & \downarrow {\scriptstyle 0} \\ \mathbb{C} & \xrightarrow[\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}]{} & \mathbb{C}^3 & \xrightarrow[\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}]{} & \mathbb{C}^2 \end{array}$$

2 The theorem of Gabriel

How can we understand and classify (up to isomorphism) *all the representations* (and their morphisms) of a quiver Q? This can be a difficult, if not *impossible*, task. Instead, one tries to classify some representations of Q from which we can build all other.

Definition 2.1 A representation M of a quiver Q is said to be **indecomposable** if it is not isomorphic to the direct sum of two other representations.

Example 2.2 Consider the quiver $Q = \ 1 \xrightarrow{\alpha} 2 \xrightarrow{\beta} 3$. The representation

$$M := \ \mathbb{C}^2 \xrightarrow{\begin{pmatrix} 1 & 0 \end{pmatrix}} \mathbb{C} \xrightarrow{0} 0$$

is a **decomposable** representation as it can be written $S_1 \oplus I_2$, where

$$S_1 := \ \mathbb{C} \xrightarrow{0} 0 \xrightarrow{0} 0\,, \qquad I_2 := \ \mathbb{C} \xrightarrow{1} \mathbb{C} \xrightarrow{0} 0\,.$$

Theorem 2.3 (Krull-Remak-Schmidt) *Every finite dimensional representation of a quiver decomposes uniquely as a direct sum of indecomposable representations.*

We can, therefore, think of indecomposable representations as the *atoms of the category of finite dimensional representations*. There are also **irreducible morphisms of representations** such that every other morphism can be *built from them* by forming compositions, linear combinations and matrices. Still, it may occur that there are *too many* indecomposable representations to be classified...

Definition 2.4 We say that a quiver Q is of **finite representation type** if Q has only finitely many indecomposable representations (up to isomorphism).

The theorem of Gabriel will say precisely which quivers have finite representation type. Among quivers of **infinite representation type**, there are two *subtypes*:

- **Quivers of tame type**: Infinitely many indecomposable finite dimensional representations (up to isomorphism) but which are *possible to parametrise*;
- **Quivers of wild type**: Infinitely many indecomposable finite dimensional representations (up to isomorphism) which *cannot be parametrised*.

Theorem 2.5 *A quiver Q is of finite representation type if and only if the underlying graph belongs to one of the following families of graphs:*

$A_n =$ $1 - 2 - 3 - \dots - n-1 - n$, $n \geq 1$

$D_n =$ $1 - 2 - 3 - \dots - n-2 - n-1$, $n \geq 4$ (with n attached to $n-2$)

$E_6 =$ $1 - 2 - 3 - 4 - 5$ (with 6 attached to 3)

$E_7 =$ $1 - 2 - 3 - 4 - 5 - 6$ (with 7 attached to 3)

$E_8 =$ $1 - 2 - 3 - 4 - 5 - 6 - 7$ (with 8 attached to 3)

Quivers of tame representation type can be classified in a similar way (see [2] for details). For quivers of finite representation type, the number of indecomposable representations up to isomorphism depends on its underlying graph:

- Type A_n, $n \geq 1$: $n(n+1)/2$ indecomposable representations;
- Type D_n, $n \geq 4$: $n(n-1)$ indecomposable representations;
- Type E_6, 36 indecomposable representations;
- Type E_7, 63 indecomposable representations;
- Type E_8, 120 indecomposable representations.

Example 2.6 Consider the quivers

$$Q_1 = \ 1 \longrightarrow 2, \quad Q_2 = \ 1 \rightrightarrows 2, \quad Q_3 = \ 1 \ \substack{\longrightarrow \\ \longrightarrow \\ \longrightarrow} \ 2.$$

Then Q_1 is of finite type, Q_2 of tame type and Q_3 of wild type.

3 The Auslander-Reiten quiver of linearly oriented A_3

A useful way of condensing the information of the indecomposable representations and the irreducible morphisms between them for a quiver Q is through its Auslander-Reiten quiver.

Definition 3.1 The **Auslander-Reiten quiver** of a quiver Q is a quiver defined by:

- The vertices are the finite dimensional indecomposable representations of Q;
- The arrows are the irreducible morphisms between the indecomposable finite dimensional representations.

We will construct the Auslander-Reiten quiver of the following quiver of type A_3:

$$Q = \ 1 \xrightarrow{\alpha} 2 \xrightarrow{\beta} 3.$$

It is of finite representation type, by Gabriel's theorem, and it has 6 indecomposable representations. It can be checked that the indecomposable representations of Q are given as follows.

- $P_1 := \ \mathbb{C} \xrightarrow{1} \mathbb{C} \xrightarrow{1} \mathbb{C}$, sometimes denoted by (1 1 1);
- $P_2 := \ 0 \longrightarrow \mathbb{C} \xrightarrow{1} \mathbb{C}$, sometimes denoted by (0 1 1);
- $P_3 := \ 0 \longrightarrow 0 \longrightarrow \mathbb{C}$, sometimes denoted by (0 0 1);
- $I_2 := \ \mathbb{C} \xrightarrow{1} \mathbb{C} \longrightarrow 0$, sometimes denoted by (1 1 0);
- $S_1 := \ \mathbb{C} \longrightarrow 0 \longrightarrow 0$, sometimes denoted by (1 0 0);
- $S_2 := \ 0 \longrightarrow \mathbb{C} \longrightarrow 0$, sometimes denoted by (0 1 0).

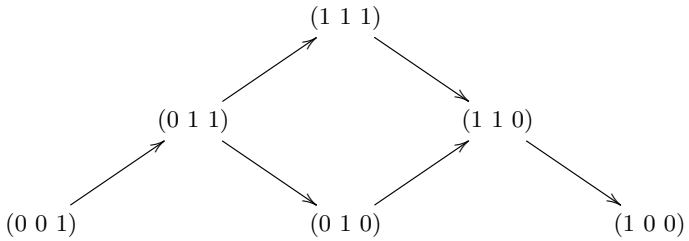
The irreducible morphisms between representations of Q can also be explicitly computed. For example, there is an injective morphism from $P_3 = (0\ 0\ 1)$ to $P_2 = (0\ 1\ 1)$, defined by:

$$\begin{array}{ccccc} 0 & \longrightarrow & 0 & \longrightarrow & \mathbb{C} \\ \downarrow & & \downarrow & & \downarrow{\scriptstyle 1} \\ 0 & \longrightarrow & \mathbb{C} & \xrightarrow{1} & \mathbb{C}. \end{array}$$

Similar considerations give the following morphisms:

- An injective morphism from $P_2 = (0\ 1\ 1)$ to $P_1 = (1\ 1\ 1)$;
- A surjective morphism from $P_2 = (0\ 1\ 1)$ to $S_2 = (0\ 1\ 0)$;
- An injective morphism from $S_2 = (0\ 1\ 0)$ to $I_2 = (1\ 1\ 0)$;
- A surjective morphism from $P_1 = (1\ 1\ 1)$ to $I_2 = (1\ 1\ 0)$;
- A surjective morphism from $I_2 = (1\ 1\ 0)$ to $S_1 = (1\ 0\ 0)$.

We are now ready to build the Auslander-Reiten quiver of A_3.



As said before, due to Krull-Remak-Schmidt's theorem, this quiver contains *all the information* about the *category of finite dimensional representations of* Q. The triples identifying the representations are called **dimension vectors** and they help us to keep in mind what the morphisms are.

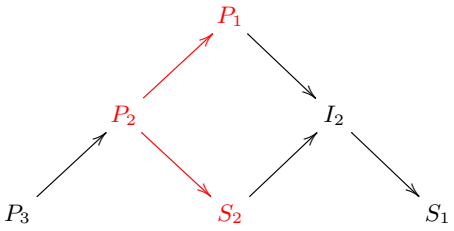
4 Examples of tilting representations and their endomorphism rings

In this section we will show some examples of tilting representations and *compute their endomorphism rings*. This serves as motivation for the next section, where we will see how the representations of a path algebra and the representations of the endomorphism ring of a tilting representation are related.

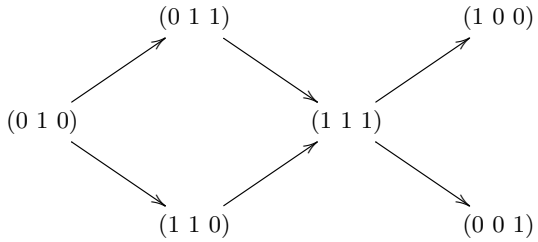
Recall that, given finite dimensional representations M and N of a quiver Q, we denote by $Hom_Q(M, N)$ the set of morphisms of representations between M and N. It is clear that $Hom_Q(M, N)$ is a $\mathbb{C}$-vector space and it is, moreover, finite dimensional. If $M = N$, we write $End_Q(M)$ for this space and $End_Q(M)$ has an additional operation, **composition of morphisms**, which is distributive with respect to addition - i.e., $End_Q(M)$ has a ring structure. It is called **the endomorphism ring of** M.

Example 4.1 As before, consider $Q = 1 \xrightarrow{\alpha} 2 \xrightarrow{\beta} 3$. With the help of the Auslander-Reiten quiver, we can compute endomorphism rings of representations. Let $T = P_2 \oplus P_1 \oplus S_2$. To compute $End_Q(T)$ we signal in red the indecomposable summands of T and we

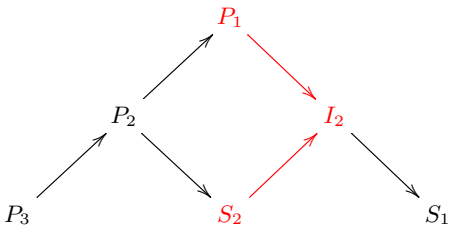
look at irreducible morphisms between them.



It turns out that $End_Q(T)$ can be computed explicitly as the path algebra of a quiver obtained as *the opposite quiver* (i.e., the quiver with the same vertices and arrows in the opposite direction) of the one suggested by the irreducible morphisms. Hence, we have that $End_Q(T) \cong \mathbb{C}(\ 1 \longrightarrow 2 \longleftarrow 3\ )$, where we *identify* the vertex 2 with the representation P_2, the vertices 1 and 3 with the representations P_1 and S_2 and the arrows are in the opposite directions of the morphisms. One also can compute the Auslander-Reiten quiver of $End_Q(T) \cong \mathbb{C}(\ 1 \longrightarrow 2 \longleftarrow 3\ )$ as follows (see the next section for the relation with the Auslander-Reiten quiver of Q).

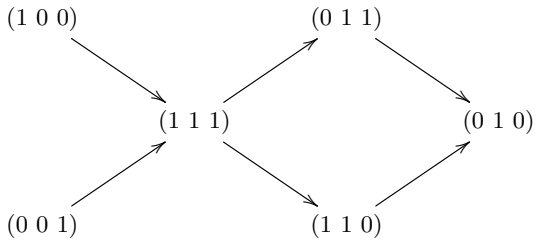


Example 4.2 Consider again the quiver $Q = \ 1 \xrightarrow{\alpha} 2 \xrightarrow{\beta} 3$. As before, we signal in red the indecomposable summands of $V = I_2 \oplus P_1 \oplus S_2$ and we look at the irreducible morphisms between them.



Again, the endomorphism ring of V can be computed as the path algebra of the opposite quiver of the one suggested by the morphisms, i.e., $End_Q(V) \cong \mathbb{C}(\ 1 \longleftarrow 2 \longrightarrow 3\ )$, where we *identify* the vertex 2 with the representation I_2, the vertices 1 and 3 with the representations P_1 and S_2 and the arrows are in the opposite directions of the morphisms. The Auslander-Reiten quiver of $End_Q(V) \cong \mathbb{C}(\ 1 \longleftarrow 2 \longrightarrow 3\ )$ is the following (see the

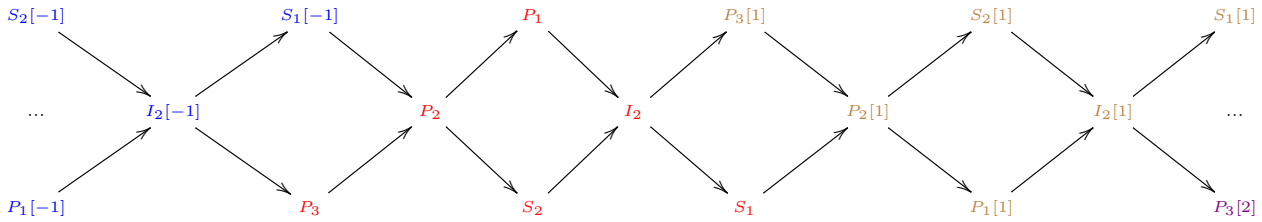
next section for the relation with the Auslander-Reiten quiver of Q).



5 Tilting representations and Happel's theorem

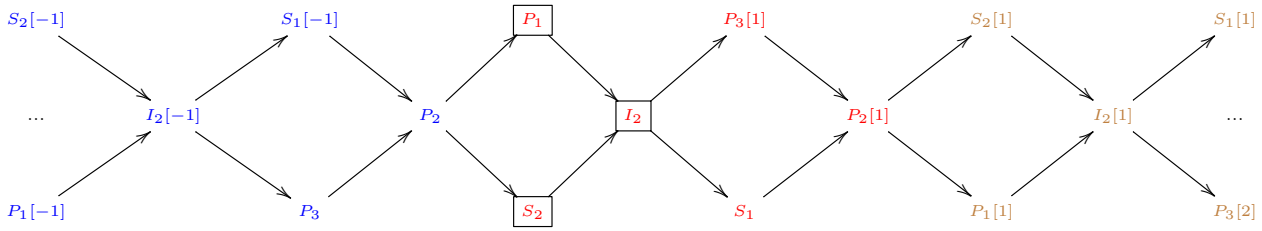
The two representations T and V considered in the above examples are **tilting representations**. A tilting representation M has good properties that allow to *compare* representations of Q and representations of $End_Q(M)$. More precisely, it allows to compare the **derived categories of representations** of Q and $End_Q(M)$ - denoted by $\mathcal{D}^b(Q)$ and $\mathcal{D}^b(End_Q(M))$, respectively. The Auslander-Reiten quiver of the derived category of a quiver Q can be drawn by *repetition* of the Auslander-Reiten quiver of Q.

Example 5.1 The Auslander-Reiten quiver of $\mathcal{D}^b(\mathbb{C}Q)$, $Q = 1 \longrightarrow 2 \longrightarrow 3$, is as follows



The different colours signal the copies of the Auslander-Reiten quiver of Q in different *positions* or *degrees* (as indicated in the square brackets in front of the corresponding representations).

Let us consider, as before, the tilting representation $V = I_2 \oplus P_1 \oplus S_2$ over $Q = 1 \longrightarrow 2 \longrightarrow 3$. We signal its indecomposable summands in the Auslander-Reiten quiver of $\mathcal{D}^b(Q)$ with squares.



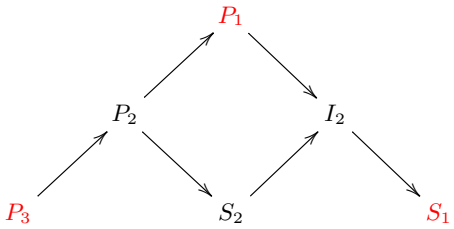
This time, we identify the different colours with repeated copies of the Auslander-Reiten

quiver of $End_Q(V) \cong \mathbb{C}(\ 1 \longleftarrow 2 \longrightarrow 3\ )$. This suggests that the derived categories $\mathcal{D}^b(Q)$ and $D^b(End_Q(V))$ are **equivalent**. Similarly, once can expect that $\mathcal{D}^b(Q) \cong \mathcal{D}^b(End_Q(T))$. These observations can be formalised as follows.

Definition 5.2 A finite dimensional representation T of a quiver Q is said to be **tilting** if $Ext^1_Q(T,T) = 0$ (i.e., every short exact sequence of representations of the form $0 \to T \to M \to T \to 0$ splits) and the number of indecomposable summands of T equals the number of vertices of Q.

Theorem 5.3 (Happel, 1987) *If T is a tilting representation of a quiver Q, then $\mathcal{D}^b(Q) \cong \mathcal{D}^b(End_Q(T))$.*

Example 5.4 The endomorphism ring of a tilting representation is not always of the form $\mathbb{C}Q'$ for some quiver Q'. Consider the representation $W = P_3 \oplus P_1 \oplus S_1$ over $Q = 1 \longrightarrow 2 \longrightarrow 3$.



To understand $End_Q(T)$, identify P_3, P_1 and S_1 with the vertices 1, 2 and 3, respectively, of a quiver and identify arrows with the opposite directions of the irreducible morphisms. However, we must remember that the composition $P_3 \to P_1 \to S_1$ is a morphism between the representations $P_3 = (0\ 0\ 1)$ and $S_1 = (1\ 0\ 0)$, i.e., it is the zero morphism. Thus, we can deduce that

$$End_Q(W) \cong \mathbb{C}(\ 1 \xleftarrow{\delta} 2 \xleftarrow{\gamma} 3\ )/ <\delta\gamma>,$$

where $<\delta\gamma>$ is the ideal generated by the path $\delta\gamma$. In this case, representations of $End_Q(W)$ will be representations $((M_i)_{i\in Q_0}, (f_\omega)_{\omega\in Q_1})$ of Q satisfying the relation $f_\delta f_\gamma = 0$.
Still, Happel's theorem still applies and $\mathcal{D}^b(Q) \cong \mathcal{D}^b(End_Q(W))$.

References

[1] Angeleri Hügel, L., Happel, D., Krause, H., "Handbook of tilting theory". London Mathematical Society, 2007.

[2] Assem, I., Simson, D. Skowroński, "Elements of the representation theory of associative algebras, 1: Techniques of Representation Theory". London Math. Soc. Student Texts 65, 2006.

[3] Auslander, M., Reiten, I., Smalø, S., "Representation Theory of Artin Algebras". Cambridge University Press, 1997.

[4] Happel, D., "Triangulated categories in the representation theory of finite dimensional algebras". London Mathematical Society, 1988.

Jacobi matrices, orthogonal polynomials and Gauss quadrature. An introduction and some results for the non-hermitian case.

Stefano Pozza (*)

1 Introduction

Let $A \in \mathbb{C}^{N\times N}$ a square matrix, let f a *matrix function*, i.e. a function such that $f(A) \in \mathbb{C}^{N\times N}$ (we will give a definition in Section 3). Now, let $\mathbf{v}$ a vector in $\mathbb{C}^N$, we are interested in estimating the bilinear form

$$\mathbf{v}^* f(A)\mathbf{v}, \tag{1}$$

where $\mathbf{v}^*$ is the conjugate transpose vector of $\mathbf{v}$.

In our studies this problem arises in some *Complex Network Theory* problems. A complex network is particular kind of graph. A graph is defined as follow.

Definition 1 (Graph) A *graph* G is a ordered pair of sets $(V(G), E(G))$ such that $V(G)$ is the *nodes (or vertices) set* and $E(G) \subset V(G) \times V(G)$ is the *edges set*.

The element of $E(G)$ could be ordered or unordered, we will call *directed graph* or *digraph* the graph of the first case and *undirected graph* the second one (see Figure 1).

A edge (u, v) of a directed graph id usually represented by a arrow that goes from the first node u, *tail*, to the second one v, *head*. We will say that two nodes are *adjacent* if there exists an edge between the two nodes.

Definition 2 (adjacency matrix) We call *adjacency matrix* of a graph G the matrix A such that $A_{i,j} = 1$ if i and j are nodes such that $(i, j) \in E(G)$ and $A_{i,j} = 0$ elsewhere.

We remark that the adjacency matrix is symmetric if and only if the graph is undirected.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: stefano.pozza@gmail.com. Seminar held on June 18th, 2014.

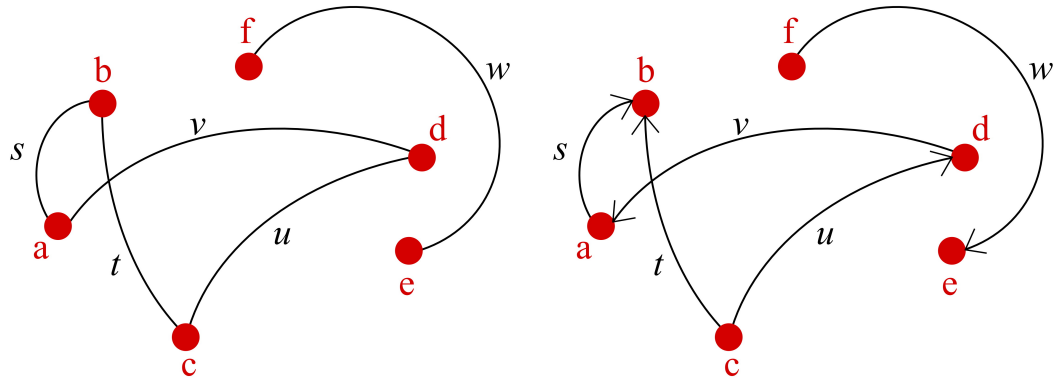


Figure 1: Graph and digraph representation relative to $G = (V(G), E(G))$, with $V(G) = \{a, b, c, d, e, f\}$ and $E(G) = \{s = (a, b), t = (c, b), u = (c, d), v = (d, a), w = (f, e)\}$

A Complex Network is a graph that for our purpose can be think as a graph obtained from the representation of relationships existent in nature. For example the relations of people in a social network or the mutual citations of the scientific articles in a database. In particular we are interested in this two important properties of the adjacency matrix A of a Complex Network:

- A is a matrix with great dimension;
- A is a sparse matrix.

A *sparse* matrix is a matrix such that the number of non-zero elements of the matrix is of the order of $\log(N)$, where N is the dimension of the matrix.

We call *path* a sequence of node $i_1, i_2, \ldots, i_n$ of a graph, then we have the following proposition.

Proposition 1 *Let A be the incidence matrix of the graph, then we have*

$$(A^k)_{i,j} = \textit{number of paths of length } k \textit{ from } i \textit{ to } j$$

Proof. By induction. $A_{i,j}$ clearly counts the number of path of length 1 between i and j. Now, if $A^{k-1}_{i,\ell}$ is the number of paths of length $k-1$ from i to any nodes ℓ, hence multiplying the i-th row of A^{k-1} by the j-th column of A means to add the number of paths of length $k-1$ from i to the nodes ℓ such that there exists an edge between ℓ and j. Hence $A^k_{i,j}$ is the number of paths of length k from i to j. □

In this talk we focus on the problem of the computation of the centrality indexes of a Complex Network. These are indexes that measure the centrality of every node in terms of connection and communicability with the other nodes. We call Subgraph Centrality an index that measures the centrality of a node counting the number of subgraph containing it [7]. For an introduction on *Subgraph Centrality* see [1, 7], for further information about *Centrality* and *Complex Networks* we refer to [5, 6].

Using Proposition 1 we can define an index SC that consists in a weighted sum of the number of closed paths (cycles) passing through a node i, that is

$$SC(i) = \left(\sum_{k=0}^{\infty} c_k A^k \right)_{i,i}$$

Hence,

- for $c_k = \frac{1}{k!}$ we have $SC(i) = (e^A)_{i,i}$, where e^A is the exponential of a matrix;
- for $c_k = c^k$ with $0 < c < \frac{1}{\lambda_N}$ we have $SC(i) = (I - cA)^{-1}_{i,i}$, the resolvent of a matrix.

Thus, we can compute the centrality index using a bilinear form

$$SC(i) = \mathbf{e_i}^T f(A) \mathbf{e_i},$$

where $\mathbf{e}_i$ is the i-th vector of the canonical basis.

To approximate this bilinear form we will follow the model reduction presented in [9]. However, this reduction is given only for the case in which A is symmetric. In Section 2 and 3 we will treat this case, while in Section 4 and 5 we will try to extend these results to the non-symmetric (or generally complex) case.

2 Orthogonal Polynomials

In this section we will give the definitions and the main properties of *orthogonal polynomials* theory. For more details on this topic we refer to [8] or [9]. Let $\mu(x)$ a non decreasing function supported by the real axis $\mathbb{R}$ having a finite limit for $x \to \pm\infty$ and let the induced positive measure $d\mu$ such that

$$\int_{\mathbb{R}} t^i d\mu(t) = m_i < \infty \text{ for all } i = 0, 1, 2, \ldots,$$

where m_i are called *moments* of the measure.

Let $\mathbb{P}$ be the space of real polynomials and $p, q \in \mathbb{P}$. We can define the following inner product

$$(p, q) = \int_{\mathbb{R}} p(t) q(t) d\mu(t), \tag{2}$$

and the norm

$$||p|| = \sqrt{(p, p)}.$$

Moreover, we say that p and q are *orthogonal* with respect to the measure $d\mu$ if

$$(p, q) = 0.$$

We say that a inner product $(\cdot,\cdot)$ is *positive definite* on $\mathbb{P}$ if $||p|| > 0$ for every $p \in \mathbb{P}$ such that $p \neq 0$. Now, if we define the Hankel determinant as

$$H_n^{(i)} = \begin{vmatrix} m_i & m_{i+1} & \cdots & m_{i+n-1} \\ m_{i+1} & m_{i+2} & \cdots & m_{i+n} \\ \vdots & \vdots & & \vdots \\ m_{i+n-1} & m_n & \cdots & m_{i+2n-2} \end{vmatrix} \tag{3}$$

we can give the following statement.

Theorem 1 *The inner product (2) is positive definite if and only if*

$$H_n^{(0)} > 0 \quad \text{for } n = 1, 2, \ldots.$$

Moreover, we define

Definition 3 (Family of Orthogonal Polynomials) We call a *family of orthogonal polynomials* with respect to a inner product $(\cdot,\cdot)$ a set of polynomials $p_0, p_1, p_2, \ldots$ such that every p_i has degree i and

$$(p_i, p_j) = 0 \text{ for every } i \neq j.$$

We first give a theorem about the existence of such families.

Theorem 2 *If the inner product (2) is positive definite then there exists a unique family of orthogonal polynomials with respect to the inner product such that every polynomial of the family is monic.*

The proof can be found in [8, Chapter 1].

We will denote this family of monic orthogonal polynomials with $\pi_0, \pi_1, \ldots$. However, we notice that since the inner product is bilinear every other family of orthogonal polynomial with respect to $d\mu$ is obtained rescaling the monic polynomials π_i.

Now we introduce an important property of orthogonal polynomials

Theorem 3 (Three-term recurrence relation) *Given a family of* orthogonal polynomials*, it satisfies a three terms recurrence relationship of the form*

$$\beta_{n+1} p_{n+1}(x) = (x - \alpha_{n+1}) p_n(x) - \gamma_n p_{n-1}(x), \quad \text{for} \quad n = 0, 1, \ldots \quad , \tag{4}$$

where $\alpha_n, \beta_n, \gamma_n \in \mathbb{C}$, $p_{-1}(x) = 0$ and $p_0(x)$ is a given scalar.

A proof can be found in in [8, Chapter 1].

Using orthogonality conditions we can give the following expressions for the coefficients of the three terms recurrence relation

$$\alpha_{n+1} = \frac{\int x p_n^2(x) d\mu}{\int p_n^2(x) d\mu}, \quad \beta_{n+1} = \frac{\int x p_n(x) p_{n+1}(x) d\mu}{\int p_{n+1}^2(x) d\mu}, \quad \gamma_n = \frac{\int x p_{n-1}(x) p_n(x) d\mu}{\int p_{n-1}^2(x) d\mu}. \tag{5}$$

Now, for every fixed λ we define the vector $\mathbf{P_n}(\lambda) = (p_0(\lambda), \dots, p_{n-1}(\lambda))^T$ and the real tridiagonal matrix

$$T_n = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & \cdots & 0 \\ \gamma_1 & \alpha_2 & \beta_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \beta_{n-1} \\ 0 & \cdots & 0 & \gamma_{n-1} & \alpha_n \end{pmatrix}. \tag{6}$$

Hence, using the three terms recurrence relation (4) we obtain the equation

$$\lambda \mathbf{P_n}(\lambda) = T_n \mathbf{P_n}(\lambda) + \beta_n p_n(\lambda) \mathbf{e}_n. \tag{7}$$

From this equality we immediately deduce the following result.

Theorem 4 *The zeros of the formal orthogonal polynomial p_n are the eigenvalues of the tridiagonal matrix T_n.*

It is always possible to obtain a symmetric T_n normalizing every polynomial so that $\mathcal{L}(\tilde{p}_n^2) = 1$. In fact, using expression (5) we have that $\gamma_k = \beta_k$ for $k = 1, \dots, n$. Hence

$$\beta_{n+1}\tilde{p}_{n+1}(x) = (x - \alpha_{n+1})\tilde{p}_n(x) - \beta_n \tilde{p}_{n-1}(x), \quad \text{for} \quad n = 0, 1, \dots. \tag{8}$$

We call this kind of family a family of *orthonormal polynomials* and we will denote them as $\tilde{p}_0, \tilde{p}_1, \dots$.

An important property of orthonormal polynomials is the Christoffel-Darboux identity, for the proof we refer to [8, 9].

Theorem 5 (Christoffel-Darboux Identity) *Let $\tilde{p}_0, \tilde{p}_1, \dots$ a family of orthonormal polynomials, then for $n \geq 0$ the following identity holds*

$$\beta_{n+1}\left(\tilde{p}_{n+1}(x)\tilde{p}_n(t) - \tilde{p}_{n+1}(t)\tilde{p}_n(x)\right) = (x - t)\sum_{i=0}^{n} \tilde{p}_i(x)\tilde{p}_i(t), \tag{9}$$

where β_n is the n-th coefficient of the relation (8).

In this case T_n is a real tridiagonal matrix with positive entries on the super or sub-diagonal. We denote such a matrix J_n and we will call it *Jacobi* matrix. Usually this is exactly the definition of Jacobi matrix (see for example [9]), however we will give another definition that will be useful in the non-symmetric case we will treat in Section 4.

Definition 4 (Jacobi matrix) A *Jacobi matrix* is a square matrix in the complex field such that it is tridiagonal, symmetric and its super or sub-diagonal has all non-zero entries.

We remark that a Jacobi matrix is Hermitian if and only if it is real.

Now, we present some spectral properties of real Jacobi matrices.

Proposition 2 *Any real $n \times n$ Jacobi matrix J can be orthogonally diagonalized, i.e.*

$$JZ = Z\,\mathrm{diag}\,(\lambda_1, \dots, \lambda_n),$$

$\lambda_1, \ldots, \lambda_n$ are the eigenvalues and $Z = [z_1, \ldots, z_n]$ are the eigenvectors such that $Z^T Z = I$.

The proof is evident since every real symmetric matrix can be orthogonally diagonalized. Moreover, we have the following two propositions.

Proposition 3 *The following properties stand for every real Jacobi matrix*

(a) it has real distinct eigenvalues;

(b) the first component of each of its eigenvectors is nonzero.

Proposition 4 (Interlacing property) *Let $\lambda_i^{(k)}$ for $i = 1, \ldots, k-1$ be the eigenvalues of J_k. Then the eigenvalues of J_{k+1} interlace the eigenvalues of J_k, i.e.*

$$\lambda_1^{(k+1)} < \lambda_1^{(k)} < \lambda_2^{(k+1)} < \lambda_2^{(k)} < \cdots < \lambda_k^{(k)} < \lambda_{k+1}^{(k+1)}.$$

For the proofs we refer to [9].

3 Gauss Quadrature and Model Reduction of a Bilinear Form

Let $\mathbf{v}^T f(A)\mathbf{v}$ a bilinear form such that A is a real and symmetric $N \times N$ matrix and $\mathbf{v}$ is a real non-zero vector. Moreover, let $A = Q\Lambda Q^T$ be the orthogonal diagonalization of A. Hence, $\Lambda = \text{diag}\,(f(\lambda_1), \ldots, f(\lambda_N))$ is the matrix of the eigenvalues and the columns of Q are the normalized eigenvectors of A.

As shown for example in [10] in this case a matrix function can be defined in the following way.

Definition 5 (Matrix Function) Let A a diagonalizable matrix and f a function defined on the spectrum of A, the *matrix function* $f(A)$ is a matrix such that

$$f(A) = Z f(\Lambda) Z^{-1}.$$

where $A = Z\Lambda Z^{-1}$, with Λ diagonal and Z invertible and $f(\Lambda) = \text{diag}\,(f(\lambda_1), \ldots, f(\lambda_N))$.

Notice that if p is a polynomial interpolating f on the spectrum of A, then $p(A) = f(A)$. Hence, we can always see $f(A)$ as a matrix polynomial depending on f and A and of degree less or equal to N.

Our purpose is to approximate the bilinear form $\mathbf{v}^T f(A)\mathbf{v}$, as we have seen in the Introduction. Setting $\mathbf{q} = Q\mathbf{v}$ we can rewritten the bilinear form as

$$\mathbf{v}^T f(A)\mathbf{v} = \mathbf{q}^T f(\Lambda)\mathbf{q} = \sum_{k=1}^{N} f(\lambda_k)\mathbf{q}_k^2.$$

And this last sum can be seen as a Riemann-Stieltjes integral with a measure $d\mu$ such defined (see [9, Chapter 7])

$$\mathbf{v}^T f(A)\mathbf{v} = \int_{\lambda_1}^{\lambda_N} f(\lambda)\mathrm{d}\mu(\lambda) \qquad \mu(\lambda) = \begin{cases} 0 & \text{if } \lambda < \lambda_1 \\ \sum_{j=1}^{i} \mathbf{q}_j^2 & \text{if } \lambda_i \leq \lambda < \lambda_{i+1} \\ \sum_{j=1}^{N} \mathbf{q}_j^2 & \text{if } \lambda \geq \lambda_N \end{cases} \tag{10}$$

Hence, to approximate this bilinear form is equivalent to approximate an integral with measure $d\mu$.

In general, given a measure $d\alpha$ and a integrable function f a quadrature rule is a relation

$$\int f(x)d\alpha(x) = \sum_{i=1}^{n} \omega_i f(t_i) + R[f], \tag{11}$$

where ω_i are the weights, t_i are the distinct nodes of the quadrature formula, whereas R is the remainder. A quadrature rule is said to be of exact degree d if $R[p] = 0$ for every polynomial p of degree d, while there exists a polynomial q of degree $d+1$ such that $R[q] \neq 0$. The optimal quadrature rule of degree $2n-1$ is called a *Gaussian quadrature rule*.

If $d\alpha$ is a measure for which there exists a family of orthogonal polynomials (p_n), then the nodes and the weights of a Gauss quadrature are linked to this family.

In fact, let p be a polynomial of degree $2n-1$, if we divide p by the formal orthogonal polynomial p_n we obtain

$$p(x) = p_n(x)q_{n-1}(x) + r_{n-1}(x), \tag{12}$$

where q_{n-1} and r_{n-1} are both of degree at most $n-1$.

Hence we have

$$\int p\, d\alpha = \int p_n q_{n-1}\, d\alpha + \int r_{n-1}\, d\alpha.$$

Since p_n is formally orthogonal to every polynomial of degree lower or equal than $n-1$ we obtain

$$\int p\, d\alpha = \int r_{n-1}\, d\alpha.$$

Let $x_1, \ldots, x_n$ be the roots of p_n. Notice that they are distinct (it enough to combine Theorem 4 and Proposition 3). Now, let

$$l_i(x) = \prod_{j=1 j\neq i}^{n} \frac{x - x_j}{x_i - x_j}$$

be the Lagrange basis polynomials. Note that from equation (12) it follows that $r_{n-1}(x_i) = p(x_i)$ for $i = 1, \ldots, n$. Thus we have $r_{n-1}(x) = \sum_{i=1}^{n} p(x_i) l_i(x)$. Hence we obtain

$$\int p\, d\alpha = \int r_{n-1}\, d\alpha = \sum_{i=1}^{n} p(x_i) \int l_i\, d\alpha.$$

If we set $\omega_i = \int l_i \, d\alpha$ for $i = 1, \dots, n$ we have an exact quadrature formula. Hence, taking as nodes the roots x_i of p_n and the weighs ω_i we obtain a Gauss quadrature formula.

Hence we have proven the following theorem.

Theorem 6 *Let x_i for $i = 1, \dots, n$ the zeros of the polynomial p_n of a family of orthogonal polynomials with respect to a measure $d\alpha$. If we set $t_i = x_i$ and $\omega_i = \int_{\mathbb{R}} l_i \, d\alpha$ as the nodes and weights of a quadrature rule (11) then it is a Gauss quadrature rule.*

3.1 Model Reduction and Moment Matching Property

Now, let J_n be the Jacobi matrix related to the measure $d\alpha$, hence such that the family $p_0, p_1, \dots$ satisfies the relation (7). Since by Theorem 4 the eigenvalues of J_n are the zeros of the polynomial p_n then the nodes of a Gauss quadrature of Theorem 6 can be seen as the eigenvalues of the Jacobi matrix. Moreover, we can give a matrix interpretation of the weights of the rule.

Theorem 7 *The eigenvalues of J_n are the nodes of the Gauss quadrature of Theorem 6 and the weights ω_i are the squares of the first element of the normalized eigenvectors of J_n.*

For a proof see [9, Chapter 6]. In particular, we remark that to prove this theorem we need the Christoffel-Darboux relation (9).

Let $\lambda_1, \dots, \lambda_n$ the eigenvalues of J_n, then from the theorem we have just stated we deduce that for every $i \leq 2n - 1$

$$\int_{\mathbb{R}} x^i d\alpha(x) = \sum_{i=1}^{n} \lambda_i^k (\mathbf{e_1}^T \mathbf{z_i})^2 = \mathbf{e_1}^T \left(\sum_{i=1}^{n} \lambda_i^k \mathbf{z_i} \mathbf{z_i}^T \right) \mathbf{e_1} = \mathbf{e_1}^T J_n^k \mathbf{e_1}.$$

From which we obtain the *moment matching property* of the Jacobi matrix

Proposition 5 (Moment Matching Property) *Let J_n a Jacobi matrix related to the measure $d\alpha$, then*

$$\int_{\mathbb{R}} x^i d\alpha(x) = \mathbf{e_1}^T J_n^i \mathbf{e_1} \qquad \text{for } i = 0, \dots, 2n - 1.$$

Moreover, let $d\alpha = d\mu$ the measure defined in (10) then by the definition of matrix function (Definition 5) we obtain

$$\mathbf{v}^T f(A) \mathbf{v} = \int_{\mathbb{R}} f \, d\alpha \approx \sum_{i=1}^{n} f(\lambda_i) (\mathbf{e_1}^T \mathbf{z_i})^2 = \mathbf{e_1}^T \left(\sum_{i=1}^{n} f(\lambda_i) \mathbf{z_i} \mathbf{z_i}^T \right) \mathbf{e_1} = \mathbf{e_1}^T f(J_n) \mathbf{e_1}.$$

Hence, since J_n is a really small matrix compare to A we have the *model reduction*

$$\mathbf{v}^T f(A) \mathbf{v} \approx \mathbf{e_1}^T f(J_n) \mathbf{e_1}.$$

In the following sections we will try to show some results of the extension of this model reduction to the case in which A is non-symmetric (or generally complex).

4 Formal Orthogonal Polynomials

First of all, we notice that if A is a square matrix then in general $\mathbf{v}^T f(A)\mathbf{v}$ cannot be seen as a integral of the function f under a certain measure. However we can still give a linear functional $\mathcal{L}$ defined on the space of complex polynomials such that

$$\mathcal{L}(x^i) = \mathbf{v}^T A^i \mathbf{v}, \quad \text{for } i = 0, 1, \dots.$$

Hence, we will try to give an approximation for this linear functional. To do this, we use some well known results about functional and *formal orthogonal polynomials* (we mainly refer to the second chapter of [2]. Moreover, we mention the results obtained by Saylor and Smolarski in [12] and [11] for the functional defined as

$$\mathcal{L}(f)_w = \int_\gamma f(x)w(x)|dx|, \tag{13}$$

where γ is an arc in the complex plane, $|dx|$ is the arc length and $w(x)$ is a weight function.

Let (m_i) be a sequence of real or complex numbers and $\mathcal{L}$ be a linear functional defined on the vector space of polynomials by

$$\mathcal{L}(x^i) = \begin{cases} m_i & i \geq 0 \\ 0 & i < 0 \end{cases}.$$

Under some assumptions on (m_i) we are going to see later, there exists a set of polynomials (p_n), named *family of Formal Orthogonal Polynomials (FOP) with respect to* $\mathcal{L}$, such that for all n

$$\mathcal{L}(x^i p_n(x)) = 0 \quad \text{for} \quad i = 0, \dots, n-1.$$

These last conditions are known as *orthogonality conditions*.

Hence, for every polynomial p of degree lower than the degree of p_n we have $c(p(x)p_n(x)) = 0$. Moreover

$$\mathcal{L}(p_n(x)p_k(x)) = 0 \quad \text{for} \quad k \neq n.$$

Note that when the linear functional $\mathcal{L}$ can be represented as the integral on the real line of a positive measure, we obtain the usual orthogonal polynomials.

From the orthogonal conditions we can derive the necessary and sufficient condition for the existence of p_n (see [2]). Let $H_n^{(k)}$ be the Hankel determinat (defined at equation (3)). If $H_n^{(0)} \neq 0$, then p_n exists and has exactly degree n. If this condition holds for all n, the linear functional $\mathcal{L}$ is called *definite*.

We can recover some properties of the orthogonal polynomials. The first one is the three-term recurrence relation.

Theorem 8 (Three-term Recurrence Relation [2]) *Given a family of* formal orthogonal polynomials*, it satisfies a three terms recurrence relationship of the form*

$$\beta_{n+1}p_{n+1}(x) = (x - \alpha_{n+1})p_n(x) - \gamma_n p_{n-1}(x), \quad \textit{for} \quad n = 0, 1, \dots \quad , \tag{14}$$

where $\alpha_n, \beta_n, \gamma_n \in \mathbb{C}$, $p_{-1}(x) = 0$ *and* $p_0(x)$ *a given constant.*

Again the coefficients can be obtained by

$$\alpha_{n+1} = \frac{\mathcal{L}(xp_n^2(x))}{\mathcal{L}(p_n^2(x))}, \quad \beta_{n+1} = \frac{\mathcal{L}(xp_n(x)p_{n+1}(x))}{\mathcal{L}(p_{n+1}^2(x))}, \quad \gamma_n = \frac{\mathcal{L}(xp_{n-1}(x)p_n(x))}{\mathcal{L}(p_{n-1}^2(x))}. \tag{15}$$

If $\mathcal{L}$ is not definite some Hankel determinants $H_n^{(0)}$ are equal to zero and the corresponding polynomials p_n do not exist. In these cases a division by zero occurs in the computation of the coefficients of the three terms relationship. We call them *true breakdowns*. If a division by zero occurs, but it is not related with the non-existence of a polynomial, then we call it *ghost breakdown*. It is possible to avoid *breakdowns* jumping over the non-existing polynomials and considering only the existing ones, that are called *regular* (see [3, 4]).

The study of *breakdowns* is not in the purpose of this talk. Hence, from now on we assume that for $k = 0, \ldots, n$, p_k exists and that $\mathcal{L}(p_k p_k) \neq 0$

Now, using the three-term recurrence relation we can obtain again the equation

$$\lambda \mathbf{P_n}(\lambda) = T_n \mathbf{P_n}(\lambda) + \beta_n p_n(\lambda) \mathbf{e}_n,$$

where T_n now is a generally complex tridiagonal matrix such that the elements in the super or sub-diagonal are non-zero.

We remark that it is always possible to obtain a symmetric T_n normalizing every polynomial so that $\mathcal{L}(\tilde{p}_n^2) = 1$. In this case T_n is Jacobi matrix J_n (Definition 4) and we denote the family of formal orthogonal polynomials $\tilde{p}_0, \ldots, \tilde{p}_n$.

Some of the differences between orthogonal polynomials and formal orthogonal polynomials arise from the spectral properties of the generally complex Jacobi matrix compared with the properties of the real case (Section 2). First of all we notice that a Jacobi matrix may not be diagonalizable since it is symmetric but not hermitian. However, at least we have the following theorem and its corollary.

Theorem 9 *Every tridiagonal matrix* $T \in \mathbb{C}^{n \times n}$ *with nonzero elements on its super or sub-diagonal is non-derogatory, i.e. each one of its eigenvalues has geometric multiplicity* 1*.*

Corollary 1 *Every generally complex tridiagonal matrix without any zero entry on super or sub-diagonal is diagonalizable if and only if it has distinct eigenvalues.*

5 Gauss Quadrature in the Complex Plane

Now, we want to give an approximation of a linear functional with the same formulation of a quadrature rule

$$\mathcal{L}(f) = \sum_{i=1}^{n} \omega_i f(t_i) + R[f], \tag{16}$$

where the nodes t_i and the weights ω_i can be complex scalars.

We will say that the rule (16) has degree d if $R[p] = 0$ for every polynomial p of degree d. Moreover, rule (16) will be said a rule of the Gauss kind if it has degree $2n-1$.

Let J_n the Jacobi matrix associated with the linear functional $\mathcal{L}$. If J_n is diagonalizable, then we can redo the same path we have already seen for the real case in Section 3 where instead of the integration we apply the linear functional and the orthogonal polynomials are replaced by formal orthogonal polynomials.

This is possible only because we can defined the Lagrange basis polynomials

$$l_i(x) = \prod_{j=1, j\neq i}^{n} \frac{x-\lambda_j}{\lambda_i - \lambda_j}, \quad \lambda_1, \ldots, \lambda_n \text{ distinct eigenvalues of } J_n.$$

since $\lambda_1, \ldots, \lambda_n$, the eigenvalues of J_n, are distinct by Corollary 1.

Thus we can state the following theorem

Theorem 10 *Let λ_i the eigenvalues of J_n and $\omega_i = \mathcal{L}(l_i(x))$, then rule (16) is a rule of th Gauss kind.*

Moreover, it is easy to deduce a moment matching property for the linear functional.

Proposition 6 (Moment Matching Property) *Let J_n a Jacobi matrix related to the linear functional $\mathcal{L}$, then*

$$\mathcal{L}(x^i) = \mathbf{e_1}^T J_n^i \mathbf{e_1} \qquad \text{for } i = 0, \ldots, 2n-1.$$

5.1 Model Reduction and Moment Matching Property

Hence, we defined a linear functional $\mathcal{L}$ such that

$$\mathcal{L}(x^i) = \mathbf{v}^T A^i \mathbf{v} \text{ for } i = 0, 1, \ldots.$$

If J_n is a diagonalizable Jacobi matrix related to $\mathcal{L}$ then, using the results we have just seen, we obtain a model reduction

$$\mathbf{v}^T f(A)\mathbf{v} \approx \mathbf{e_1}^T f(J_n)\mathbf{e_1}.$$

In particular the approximation is exact if f is a polynomial of degree lower of equal to $2n-1$.

However, the assumption on the diagonalizability of J_n seems to be quite artificial. In fact, it is possible to extend the moment matching property such that for any matrix A and vector $\mathbf{v}$

$$\mathbf{v}^T A^i \mathbf{v} = \mathbf{e_1}^T J_n^i \mathbf{e_1} \qquad \text{for } i = 0, \ldots, 2n-1,$$

see [13] for a proof.

Hence, in conclusion it seems possible to give a moment matching property for any linear functional $\mathcal{L}$. This lead us to the question: what does it means from the point of view of the Gauss quadrature?

References

[1] Michele Benzi, Ernesto Estrada, and Christine Klymko, *Ranking hubs and authorities using matrix functions*. CoRR, abs/1201.3120 (2012).

[2] C. Brezinski, "Padé-type approximation and general orthogonal polynomials". International Series of Numerical Mathematics. Birkhäuser, 1980.

[3] C. Brezinski and M. Redivo-Zaglia, *Breakdowns in the computation of orthogonal polynomials*. In Annie Cuyt, editor, "Nonlinear Numerical Methods and Rational Approximation II", volume 296 of Mathematics and Its Applications, 49–59. Springer Netherlands (1994).

[4] C. Brezinski, H. Sadok, and M. Redivo Zaglia, *Orthogonal polynomials and the Lanczos method*. n "Numerical analysis and mathematical modelling", volume 29 of Banach Center Publ., 19–33. Polish Acad. Sci., Warsaw, 1994.

[5] E. Estrada, "The Structure of Complex Networks: Theory and Applications". OUP Oxford, 2011.

[6] Ernesto Estrada, Naomichi Hatano, and Michele Benzi, *The physics of communicability in complex networks*. CoRR, abs/1109.2950 (2011).

[7] Ernesto Estrada and Desmond J. Higham, *Network properties revealed through matrix functions*. SIAM Rev. 52/4 (2010), 696–714.

[8] W. Gautschi, "Orthogonal Polynomials: Computation and Approximation". Numerical mathematics and scientific computation. Oxford University Press, 2004.

[9] G. H. Golub and G. Meurant, "Matrices, Moments and Quadrature with Applications". Princeton Series in Applied Mathematics. Princeton University Press, 2009.

[10] Nicholas J. Higham, "Functions of Matrices: Theory and Computation". Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[11] Paul E. Saylor and Dennis C. Smolarski, *Addendum to: Why gaussian quadrature in the complex plane?* Numerical Algorithms 27/2 (2001), 215–217.

[12] Paul E. Saylor and Dennis C. Smolarski, *Why gaussian quadrature in the complex plane?* Numerical Algorithms 26/3 (2001), 251–280.

[13] Zdeněk Strakŏs, *Model reduction using the Vorobyev moment problem*. Numerical Algorithms 51/3 (2009), 363–379.