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

Scuole di Dottorato in Matematica Pura e Computazionale

Seminario Dottorato 2019/20

Preface

This document offers a large overview of the eight months' schedule of Seminario Dottorato 2019/20. Our "Seminario Dottorato" (Graduate Seminar) is a double-aimed activity. At one hand, the speakers (usually Ph.D. students or post-docs, but sometimes also senior researchers) are invited to think how to communicate their researches to a public of mathematically well-educated but not specialist people, by preserving both understandability and the flavour of a research report. At the same time, people in the audience enjoy a rare opportunity to get an accessible but also precise idea of what's going on in some mathematical research area that they might not know very well.

Of course, this academic year was very peculiar, and possibly unique in this aspect: while in the first half activities were happening as usual, from February on we had to deal with COVID-19 and its consequences with respect to activities in presence, as this seminar traditionally was held. After a period of reassessing, we opted to continue this seminar online, as already done in many other series of seminars around the world. For this, it has been crucial to have had three brave PhD students who accepted to hold their seminars online: their contribution has been crucial for this series, and in a broader sense it helped a lot the other fellow PhD students to make them feel that activities, and life in general, were somehow continuing despite SARS-COV-2.

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

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

Padova, June 20th, 2020

Corrado Marastoni, Tiziano Vargiolu

Abstracts (from Seminario Dottorato's webpage)

Wednesday 2 October 2019

An overview on non-unique factorization

Federico CAMPANINI (Padova, Dip. Mat.)

In this seminar we present some basic facts about non-unique factorizations we hope will be interesting also for the non-algebraic audience. We start with the Fundamental Theorem of Arithmetic and we recall some elementary notions and examples on unique and non-unique factorization monoids. Then we move to the classical Krull-Schmidt Theorem for modules, which states that any module of finite length decomposes as a direct sum of indecomposable modules in an essentially unique way. We briefly discuss about some classes of modules in which direct-sum decompositions are not unique. Finally, we present some properties on factorizations in the monoid of polynomials with non-zero integral coefficients.

Wednesday 20 November 2019

Potential Theory and Boundary Element Method for the Laplace equation. An introduction.

Andrei-Florin ALBISORU (Babes-Bolyai University, Cluj-Napoca, Romania)

We aim to give an overview of Potential theory for Laplace's equation. We introduce the fundamental solution of this equation. Next, we define the layer potentials and we state their properties. Using the layer potentials we will construct a solution of the interior Dirichlet problem for the Laplacian. We also describe a numerical method of solving Laplace's equation, namely the Boundary Element Method. Finally, we present some numerical results.

Wednesday 11 December 2019

A smooth introduction to the semi-classical problem in Quantum Mechanics

Enrico PICARI (Padova, Dip. Mat.)

It is very well-known today that Quantum Mechanics plays a crucial role in describing and understanding nature, at least (but not only) at the smallest scales, i.e. the ones of atoms and subatomic particles. The problem of connecting physical and mathematical features of Classical and Quantum Mechanics dates back to the early days of the theory and although reduction of one to the other is understood heuristically in terms of a limit process in which the Planck constant goes to zero, the mathematical ground of such procedure, known as Semi-classical Analysis, is still an active field of research which includes techniques of apparently distant topics like Harmonic Analysis and Deformations of Poisson Algebras. The aim of this seminar is to introduce the main features of elementary Quantum Mechanics, with a brief historical note, and to give an insight into the state-of-the-art of the semi-classical limit problem.

Wednesday 18 December 2019

An introduction to sheets of conjugacy classes in reductive groups

Filippo AMBROSIO (Padova, Dip. Mat.)

Linear algebraic groups arose as a generalization of Lie groups, introduced in the late 1800s to study continuous symmetries of differential equations. The development of the modern theory of algebraic groups with the use of algebraic geometry is mostly due to Borel: in the 1950s, his work led to the definition of Chevalley groups, an important family of finite simple groups. This suggests that algebraic groups can be approached from different perspectives (Group Theory, Algebraic Geometry, Combinatorics) and have applications in several directions (Invariant Theory, Physics). In the first part of the talk we will introduce basic notions and examples of linear algebraic groups. The last part of the seminar aims at describing some of the geometric structure of these groups.

Wednesday 15 January 2020

Stable hypersurfaces in the complex projective space

Alberto RIGHINI (Padova, Dip. Mat.)

The classification of complete oriented stable hypersurfaces in the complex projective space could be an important step for the classification of isoperimetric sets. Indeed, the boundary of an isoperimetric set, if smooth, is a hypersurface with constant mean curvature which is stable for variations fixing the volume. In this talk we give an introductory overview of the problem and present some new results, in particular we will characterize the geodesic spheres as the unique stable connected and complete hypersurfaces subject to a certain bound on the curvatures.

Wednesday 5 February 2020

A random journey among stochastic processes

Samuele STIVANELLO (Padova, Dip. Mat.)

This seminar will cover a broad selection of topics, ranging from the basic definition of a probability space, passing through some famous results like the Law of Large Numbers and the Central Limit Theorem, and ending with the notion of convergence of stochastic processes. In order to fulfill this intent, and in the spirit to be accessible to a mathematical audience of non experts, in this introductory talk to the field of stochastic processes I will make extensive use of examples and intuitive definitions. In the very last part of the talk I will mention some results of my research, regarding the convergence of a random walk in random environment.

Wednesday 12 February 2020

Some features of finite simple groups

Daniele GARZONI (Padova, Dip. Mat.)

Finite simple groups are the building blocks of finite groups. For this reason, since the early days of group theory lots of efforts were devoted to understanding this class of groups. These culminated in the 1980's in an enormous theorem — known as Classification Theorem — which exhibits a very precise list of these groups. In this seminar we will state the theorem, and briefly describe the objects involved. We will then focus on some features of finite simple groups. For instance, we will see that it is amazingly easy to generate these groups by few elements. Along the way, we will try to explain the impact of the Classification in the field of group theory.

Wednesday 6 May 2020

Permutation group theory

Mariapia MOSCATIELLO (Padova, Dip. Mat.)

The study of permutation groups is an old subject with a rich history, stretching all the way back to the origins of group theory in the early 19th century. Of course, the modern notion of a permutation group is extremely flexible, and they arise naturally throughout mathematics, with important applications across the sciences. In this seminar, we will focus on finite permutation groups, which continue to be a very active area of current research. After introducing some very basic concepts, we will see, with some examples, how the Classification of Finite Simple Groups has revolutionized the study of finite permutation groups. Some of the topics we will discuss have interesting connections to other areas of mathematics, such as combinatorics, representation theory, graph theory. We will highlight these applications.

Wednesday 20 May 2020

Computational problems in mathematical physical modelling with DLTI systems

Marta GATTO (Padova, Dip. Mat.)

Mathematical physical models are often used for the description of physical phenomena and are essential in industrial applications for various aims, such as control and estimation of unmeasurable variables and physical parameters. In this talk, some numerical methods at the basis of experimental modelling, i.e. modelling through experimental data, will be described for different kind of model classes, in particular for DLTI (Dynamic Linear Time Invariant) systems. The reasons for their importance will be explained and the computational problems of parameter estimation and data denoising subject to the DLTI model constraint will be introduced with examples.

Wednesday 10 June 2020

Efficient representation of supply and demand curves on day-ahead electricity markets

Mariia SOLOVIOVA (Padova, Dip. Mat.)

Accurate modeling and forecasting electricity demand and prices are very important issues for decision making in deregulated electricity markets. In this seminar I will explain some basic facts about price formation process in day-ahead electricity market, then I will speak mainly about the problem of approximation of supply and demand curves, with a special attention to Italian case. Finally, I will show how supply and demand curves evolve as stochastic processes in functional spaces.

An overview on non-unique factorization

Federico Campanini [*]

Preface

These notes would be a gentle approach to (some aspects of) the theory of non-unique factorizations. In the first section, we introduce the very basics of this theory in the framework of commutative cancellative monoids, starting from the Fundamental Theorem of Arithmetics and going ahead through some of the key notions and facts about unique and non-unique factorizations. In Section 2 we focus our attention on the multiplicative monoid $\mathbb{N}_0[x]^*$ of polynomials with non-negative integral coefficients. This is a very easy example that hides some interesting and unexpected phenomena. Finally, in Section 3 we move to Module Theory, trying to say something about non-unique factorizations in this context. We recall the classical Krull-Schmidt theorem for modules and we describe some situations in which "weak versions" of this theorem can be stated.

1 Basics on non-unique factorizations

Definition 1.1 A monoid is a set M endowed with a binary operation $\cdot : M \times M \to M$ such that:

(a) $x \cdot (y \cdot z) = (x \cdot y) \cdot z$ for every $x, y, z \in M$ (associative);

(b) there exists an identity element $1 \in M$ such that $x \cdot 1 = x$ for every $x \in M$;

(c) $x \cdot y = y \cdot x$ for every $x, y \in M$ (commutative);

(d) $x \cdot y = x \cdot z \Rightarrow y = z$ for every $x, y, z \in M$ (cancellative).

Notice that we always assume that our monoids are commutative and cancellative by definition. This means that sometimes we are forced to remove the zero element in our multiplicative monoids. For instance, if we want to deal with the multiplicative monoid

[*] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: campanin@math.unipd.it . Seminar held on 2 October 2019.

of the integers, we will consider the monoid $(\mathbb{Z}^*, \cdot)$, where $\mathbb{Z}^* := \{\pm 1, \pm 2, \dots\}$ (0 is not included). For the natural numbers, we adopt the notations $\mathbb{N} := \{1, 2, 3, \dots\}$ and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. For our monoids, we may also use the additive notation. In this case we have:

Definition 1.2 A monoid is a set M endowed with a binary operation $+ : M \times M \to M$ such that:

(a) $x + (y + z) = (x + y) + z$ for every $x, y, z \in M$ (associative);

(b) there exists a zero element $0 \in M$ such that $x + 0 = x$ for every $x \in M$.

(c) $x + y = y + x$ for every $x, y \in M$ (commutative);

(d) $x + y = x + z \Rightarrow y = z$ for every $x, y, z \in M$ (cancellative).

We will continue to give definitions and general results in multiplicative notations, even if sometimes we shall use the additive notation, which is more natural in some contexts. The main and easiest example about unique factorization monoids is described by the Fundamental Theorem of Arithmetic:

Theorem 1.3 *Every natural number ≥ 2 can be written as a product of prime elements, and this representation is unique up to the order of the factors.*

For example, in $(\mathbb{N}, \cdot)$ we can write $6 = 2 \cdot 3 = 3 \cdot 2$ and there are no other ways to factor 6 as a product of prime elements. But if we move to the integers, we have a slight different situation. In fact, in $(\mathbb{Z}^*, \cdot)$ we have $6 = 2 \cdot 3 = 3 \cdot 2 = (-2) \cdot (-3) = (-3) \cdot (-2)$. Here we have two possible factorizations up to the order. This lack of uniqueness is caused by the invertible elements (and it is actually not a big deal). Recall that given a monoid $(M, \cdot)$, an element $u \in M$ is **invertible** if there exists $v \in M$ such that $uv = 1$, and two elements $x, y \in M$ are **associated** (we write $x \sim y$) if there exists an invertible element $u \in M$ such that $x = uy$. In $(\mathbb{Z}^*, \cdot)$ we have two invertible elements, namely ± 1, then $2 \sim -2$ and $3 \sim -3$. Therefore we can say that 6 can be written as a product of prime elements in a unique way up to the order of the factors and up to associated elements, and of course it holds for every integer. Now look at the following example.

Example 1.4 Consider $\mathbb{Z}[\sqrt{-5}]^* := \{a + b\sqrt{-5} \mid a, b \in \mathbb{Z}\} \setminus \{0\}$ as a multiplicative monoid. In $(\mathbb{Z}[\sqrt{-5}]^*, \cdot)$ we can write

$$6 = 2 \cdot 3 = (-2) \cdot (-3) = (1 + \sqrt{-5}) \cdot (1 - \sqrt{-5}) = (-1 - \sqrt{-5}) \cdot (-1 + \sqrt{-5})$$

and the other four factorizations in reverse order. The only invertible elements of $\mathbb{Z}[\sqrt{-5}]^*$ are ± 1 and 2 is not associated neither to $1 + \sqrt{-5}$ nor $1 - \sqrt{-5}$ (and similarly for 3). Therefore we have two (essentially) different factorizations up to the order and associated elements.

The previous example raises other questions. Is it correct to consider prime elements when dealing with factorizations? For instance, are 2 or $1 + \sqrt{-5}$ prime elements in $\mathbb{Z}[\sqrt{-5}]^*$? And what is the “right” definition of prime elements? Here the answers.

Definition 1.5 Let $(M, \cdot)$ be a monoid.

(a) A non-invertible element $x \in M$ is **irreducible** (or an **atom**, or **indecomposable**) if, whenever $x = yz$ for some $y, z \in M$, then y or z is invertible.
We denote by $\mathcal{A}(M)$ the set of all atoms of M.

(b) An non-invertible element $x \in M$ is **prime** if for every $y, z \in M$, $x \mid yz$ implies $x \mid y$ or $x \mid z$.

Definition 1.6 A monoid $(M, \cdot)$ is a **unique factorization monoid** (UF for short) if:

- M is **atomic**, that is, every non-invertible element can be written as a product of finitely many atoms;
- for every non-invertible element $x \in M$, if

$$x = a_1 \cdots a_n = b_1 \cdots b_m, \qquad a_i, b_j \in \mathcal{A}(M)$$

are two factorizations into irreducible elements, then $m = n$ and there exists a permutation σ of $\{1, 2, \dots, n\}$ such that $a_i \sim b_{\sigma(i)}$.

Unique factorization monoids are also called **factorial monoids**.

In a UF monoid an element is irreducible if and only if it is a prime element, but in general only the implication “prime $\Rightarrow$ irreducible” hold. In Example 1.4, the elements $2, 3, 1 \pm \sqrt{-5} \in \mathbb{Z}[\sqrt{-5}]^*$ are all irredubile but not prime.

Example 1.7 Examples of unique factorization monoids are:

(a) $(\mathbb{N}, \cdot)$ natural numbers;

(b) $(\mathbb{Z}^*, \cdot)$ integers;

(c) $(\mathbb{Z}[X]^*, \cdot)$ polynomials with integral coefficients;

(d) $(K[X_1, \dots, X_n]^*, \cdot)$ polynomials over a field;

(e) $(K[[X_1, \dots, X_n]]^*, \cdot)$ formal power series over a field;

(f) $(\mathbb{Z}[i]^*, \cdot)$ Gaussian integers.

A monoid may fail to be factorial in several ways. For instance, the monoid $(\mathbb{Z}[\sqrt{-5}]^*, \cdot)$ we have seen in Example 1.4 has the property that if $x = a_1 \cdots a_n = b_1 \cdots b_m$ are two factorizations of an element $x \in \mathbb{Z}[\sqrt{-5}]^*$ into atoms, then $n = m$ (monoids with this property are called half-factorial). In the last part of this section, we briefly present some tools that allow us to “measure” how much a monoid may fail to be factorial. What we are going to present is definitely not comprehensive and we refer the readers to [8, Chapter

1] for more details about these topics.

Let $(M, \cdot)$ be a commutative cancellative multiplicative monoid and $U(M)$ be the group of its units (= invertible elements). If $U(M) = \{1\}$, we will say that M is **reduced**. Notice that a monoid M is atomic if and only if it is generated by the set $\mathcal{A}(M) \cup U(M)$. Assume that M is reduced and atomic. For every $x \in M \setminus U(M)$, the **set of lengths of** x **in** M is defined as the set

$$L(x) := \{n \in \mathbb{N} \mid \text{ there are } a_1, \dots, a_n \in \mathcal{A}(M) \text{ with } x = a_1 \cdots a_n\}.$$

A monoid M is called a **bounded factorization monoid** (BF for short) if $L(x)$ is finite for every $x \in M \setminus U(M)$. If $|L(x)| = 1$ for every $x \in M \setminus U(M)$, the monoid is called **half-factorial**.

Example 1.8 Consider the ring $\mathbb{Z}[x_1, x_2, \dots]$ of polynomials over $\mathbb{Z}$ with countably many indeterminates and let I be the ideal generated by the elements

$$x_1 - x_2x_3, \quad x_1 - x_4x_5x_6, \quad x_1 - x_7x_8x_9x_{10}, \quad x_1 - x_{11}x_{12}x_{13}x_{14}x_{15}, \quad \dots$$

Take the quotient ring $R := \mathbb{Z}[x_1, x_2, \dots]/I$. In the monoid $(R^*, \cdot)$ we have that all the elements $x_2 + I, x_3 + I, x_4 + I, \dots$ are irreducible and

$$x_1+I = (x_2+I)(x_3+I) = (x_4+I)(x_5+I)(x_6+I) = (x_7+I)(x_8+I)(x_9+I)(x_{10}+I) = \dots$$

are all factorizations of the element $x_1 + I$ into atoms. It follows that $L(x_1 + I) = \mathbb{N}_{\geq 2}$ and therefore $(R^*, \cdot)$ is not a bounded factorization monoid.

The **elasticity of** x is given by the ratio

$$\rho(x) := \sup L(x) / \min L(x).$$

Notice that if M is not a BF-monoid, there exist elements whose elasticity is infinite. The **elasticity of** M is defined as

$$\rho(M) := \sup\{\rho(x) \mid x \in M\} \in \mathbb{Q}_{\geq 1} \cup \{\infty\}.$$

A monoid M is said to have **full-infinite elasticity** if for every rational number $q \in \mathbb{Q}_{\geq 1}$ there exists an element $x \in M$ such that $\rho(x) = q$.

Another important tool for investigating the structure of sets of lengths of a monoid M is the **set of distances of** M (or the Δ-set of M). It is the subset $\Delta(M)$ of $\mathbb{N}$ consisting of all $d \in \mathbb{N}$ for which there exist $x \in M \setminus U(M)$ and $\ell \in L(x)$ such that $L(x) \cap [\ell, \ell + d] = \{\ell, \ell + d\}$. In other words, if $L(x) = \{n_1, n_2, \dots, \}$ is the (possibly infinite) set of lengths of x, where $n_i < n_{i+1}$ for $1 \leq i < |L(x)|$, then the Δ-set of x is defined by $\Delta(x) := \{n_{i+1} - n_i \colon 1 \leq i < |L(x)|\}$. The Δ-set of a monoid M (or the **set of distances of** M) is $\Delta(M) := \bigcup_{x \in M \setminus U(M)} \Delta(x) \subseteq \mathbb{N}$. It is immediately seen that a monoid M is half-factorial if and only if $\Delta(M) = \emptyset$.

2 Factorizations of polynomials with non-negative integer coefficients

Let $\mathbb{N}_0$ be the set of all non-negative integers $0, 1, 2, \dots$, let $\mathbb{N}_0[x]$ be the set of all polynomials in the indeterminate x with coefficients in $\mathbb{N}_0$, and $\mathbb{N}_0[x]^* := \mathbb{N}_0[x] \setminus \{0\}$ be the set of all non-zero elements of $\mathbb{N}_0[x]$.

Of course, $\mathbb{N}_0[x]^*$ is a commutative, cancellative and reduced monoid, whose unique invertible element is the identity 1. It is a submonoid of the monoid $\mathbb{Z}[x]^*$ of polynomials of integral coefficients, which is a UF monoid. So, it is natural to ask if also $\mathbb{N}_0[x]^*$ is a UF monoid itself. First of all, notice that there exist polynomials $f(x), g(x) \in \mathbb{N}_0[x]^*$ such that $g(x)$ divides $f(x)$ in $\mathbb{Z}[x]$ but not in $\mathbb{N}_0[x]^*$. This implies that we can consider two different notions of divisibility in $\mathbb{N}_0[x]^*$. Given $f, g \in \mathbb{N}_0[x]^*$, we write $f|_{\mathbb{N}}g$ if there exists $h \in \mathbb{N}_0[x]^*$ with $g = fh$, and we write $f|_{\mathbb{Z}}g$ if there exists $h \in \mathbb{Z}[x]$ with $g = fh$. Both $|_{\mathbb{N}}$ and $|_{\mathbb{Z}}$ are partial orders on $\mathbb{N}_0[x]^*$ with least element 1 and with no maximal elements. Clearly, $f|_{\mathbb{N}}g$ implies $f|_{\mathbb{Z}}g$, but not conversely, as the factorization $x^3+1 = (x+1)(x^2-x+1)$ shows. Now, looking at the factorization of the cyclotomic polynomial $f(x) := x^5 + x^4 + x^3 + x^2 +^x +1$ in $\mathbb{Z}[x]$, we have:

$$x^5 + x^4 + x^3 + x^2 +^x +1 = (x + 1)(x^2 - x + 1)(x^2 + x + 1),$$

where, of course, $(x^2 + x + 1) \notin \mathbb{N}_0[x]^*$. This factorization leads to two different factorizations of $f(x)$ in $\mathbb{N}_0[x]^*$, namely

$$x^5 + x^4 + x^3 + x^2 +^x +1 = (x^3 + 1)(x^2 + x + 1) = (x + 1)(x^4 + x^2 + 1).$$

This shows that $\mathbb{N}_0[x]^*$ is not a UF monoid. In particular, not all the irreducible elements of $\mathbb{N}_0[x]^*$ are prime elements (for example, it is immediate that $x + 1$ is an atom of $\mathbb{N}_0[x]^*$ which is not prime). The following result shows that actually we have "very few" prime elements in $\mathbb{N}_0[x]^*$.

Proposition 2.1 [5, Proposition 2.4] *The only prime elements of* $\mathbb{N}_0[x]^*$ *are the prime numbers and the polynomial* x*.*

This fact implies that every irreducible polynomial in $\mathbb{N}_0[x]^*$, except for the prime numbers and the polynomial x, divides some element in $\mathbb{N}_0[x]^*$ which has a non-unique factorization. In particular, it means that examples of non-unique factorizations in $\mathbb{N}_0[x]^*$ appear very frequentely. So, we can ask if $\mathbb{N}_0[x]^*$ is a half-factorial monoid, and in case is not, try to understand how far it is from being half-factorial, using the notions of elasticity and Δ-set.

Notice that the multiplicative monoid $\mathbb{N}_0[x]^*$ is a BF-monoid, since $\mathbb{Z}[x]^*$ is, but it is possible to find several examples demonstrating the non-half-factoriality of $\mathbb{N}_0[x]^*$ (see [11] and [1]). Moreover, the following example shows that $\mathbb{N}_0[x]^*$ is surprisingly very far from being half-factorial.

Example 2.2 In the proof of [6, Theorem 2.3], the authors consider polynomials of the form $g_{n,k}(x) = (x + n)^n(x^2 - x + 1)(x + k) \in \mathbb{N}_0[x]^*$, where $n, k \in \mathbb{N}$, and they observe that there are only two factorizations of $g_{n,k}(x)$ into atoms of $\mathbb{N}_0[x]^*$, given by

$$g_{n,k}(x) = [(x + n)^n(x^2 - x - 1)] \cdot [x + 1]^k$$

and

$$g_{n,k}(x) = [x+n]^n \cdot [(x^2-x+1)(x+1)] \cdot [x+1]^{k-1},$$

which have lengths $1+k$ and $n+k$ respectively. Thus $L(g_{n,k}) = \{1+k, n+k\}$ and the elasticity of $g_{n,k}$ is $\rho(g_{n,k}) = (n+k)/(1+k)$. From the arbitrarity of n and k it follows that given any rational number $q \geq 1$, there exists some $f(x) := g_{n_q,k_q}(x) \in \mathbb{N}_0[x]^*$ such that the elasticity $\rho(f)$ of $f(x)$ is equal to q and hence $\rho(\mathbb{N}_0[x]^*) = \infty$ (therefore $\mathbb{N}_0[x]^*$ has full infinite elasticity). Moreover, $\Delta(\mathbb{N}_0[x]^*) = \mathbb{N}$, because $\Delta(g_{n,k}) = \{n+1\}$.

3 Direct-sum decompositions of modules

In this section we want to talk about (non-)unique factorizations in Module Theory. Here "factorization of a module" means "direct-sum decomposition", that is, given a module M over some ring R, we are interested in studying if M admits a decomposition $M = N_1 \oplus N_2 \oplus \cdots$ as a direct sum of its indecomposable submodules. Since we are dealing with direct sums, the additive notation is more natural in this context.

Let us start by considering the following situation, coming from linear algebra. Let k be a field and consider the set

$$M(k) := \{0, k, k \oplus k, k \oplus k \oplus k, \dots\}$$

which can be viewed as the set of all finite dimensional k-vector spaces up to isomorphism. Then $(M(k), \oplus)$ is a UF monoid with a unique atom k. It is worth noting that if we consider the set of all vector spaces over k (not only the finite-dimensional ones) we do not get an atomic monoid, because, for instance, $k^{(\mathbb{N})}$ can not be written as a finite direct sum of atoms.

That of fields is a particular situation. If we want to deal with arbitrary (unitary) rings we have to be careful, because there exist rings R such that $R \cong R \oplus R$, as the following example shows.

Example 3.1 Let V be an infinite-dimensional vector space over a field k and consider its endomorphism ring $R := \operatorname{End}(V)$. Since $V \cong V \oplus V$ we have

$$R = \operatorname{End}(V) \cong \operatorname{End}(V \oplus V) \cong \operatorname{End}(V) \oplus \operatorname{End}(V) = R \oplus R.$$

In this case, the set of all free (right) R-modules up to isomorphism is

$$M(R) := \{0, R, R \oplus R, \dots\} = \{0, R\}.$$

The monoid $(M(R), \oplus)$ consists of just two elements, the identity 0 and the idempotent element R. In particular, there are no atoms, hence $(M(R), \oplus)$ is not atomic.

Definition 3.2 A ring R is called **IBN** (invariant basis number) if for every pair of positive integers $n, m \in \mathbb{N}$, $R^n \cong R^m$ implies $n = m$.

As we have seen, there are rings that are not IBN. Anyway, any commutative ring is necessarily an IBN ring. Other classes of IBN rings are (left/right) Noetherian rings and semilocal rings. Free modules over an IBN ring satisfy the analogue of the dimension theorem for vector spaces: any two bases for a free module over an IBN ring have the same cardinality. Therefore, the rank of a free module R^n over an IBN ring is well-defined.

The most important theorem about unique factorization (= direct-sum decomposition) in Module Theory is given by the Krull-Schmidt Theorem. It states that any module of finite composition length decomposes as a direct sum of indecomposable modules in an essentially unique way up to isomorphism.

Theorem 3.3 (Classical Krull-Schmidt Theorem for modules) *Let R be a ring and M be a module of finite length. Then there exists a decomposition*

$$M = M_1 \oplus \cdots \oplus M_r$$

into indecomposable submodules. Moreover, if $M = N_1 \oplus \cdots \oplus N_t$ is another decomposition of M into indecomposable submodules, then $r = t$ and there exists a permutation σ of $\{1, 2, \ldots, r\}$ such that $M_i \cong N_{\sigma(i)}$ for every $i = 1, 2, \ldots, r$.

Notice that any module with a local endomorphism ring is necessarily indecomposable and by Fitting's Lemma, the converse holds as well for modules of finite composition length. Theorem 3.3 was extended by G. Azumaya to the case of possibly infinite direct sums of modules with a local endomorphism ring.

Theorem 3.4 (Krull-Schmidt-Remak-Azumaya Theorem) *Let R be a ring and let M be a module that is a direct sum of modules with local endomorphism rings. Then M is a direct sum of indecomposable modules in an essentially unique way in the following sense. If*

$$M = \bigoplus_{i \in I} M_i \cong \bigoplus_{j \in J} N_j$$

where all the submodules M_i, $i \in I$ and N_j, $j \in J$ are indecomposable, then there exists a bijection $\sigma : I \to J$ such that $M_i \cong N_{\sigma(i)}$ for all $i \in I$.

Now, fix a ring R and denote by $\mathcal{L}$ the class of all right R-modules with a local endomorphism ring. Then, we can consider the set $M(\mathcal{L})$ of all finite direct-sums of modules in $\mathcal{L}$ up to isomorpshism. We have that $(M(\mathcal{L}), \oplus)$ is a UF monoid, whose atoms are the modules in $\mathcal{L}$.

Nowadays the name "Krull-Schmidt" is given to any theorem concerning uniqueness of direct-sum decompositions into indecomposable direct summands. This is a very classical topic that has a crucial relevance in the study of algebraic structures. Of course, there are cases in which direct-sum decompositions are not essentially unique. It is worth mentioning that Krull already knew that the Krull-Schmidt Theorem does not hold for arbitrary Noetherian modules, which means that the ascending chain condition does not suffices to have uniqueness of direct-sum decompositions. In light of this fact, a question that naturally arises is if the Krull-Schmidt Theorem holds for the class of Artinian modules. This problem was originally posed by Krull himself in 1932 but the answer was found only sixty years later, in 1995, when Facchini, Herbera, Levy and Vámos showed that the Krull-Schmidt Theorem fails for Artinian modules. They prove that for any integer $n \geq 2$ there exists an artinian module M over a suitable ring R such that M can be written as a direct sum of k indecomposable modules for $k = 2, \dots, n$.

In the last three decades, new interesting examples in which direct-sum decompositions are not unique made their appearance. Even though the lack of uniqueness, these situations display some kind of regularity: direct-sum decompositions can be classified via two invariants and a weak version of the Krull-Schmidt Theorem can be proved. Here, we briefly outline the case of uniserial modules, but such behaviour can be found in several classes of modules, including biuniform modules, cyclically presented modules over local rings, couniformly presented modules, kernels of morphisms between indecomposable injective modules.

A right R-module U is **uniserial** if the lattice of its submodules is linearly ordered under inclusion, that is, for every $V, W \leq U$, either $V \subseteq W$ or $W \subseteq V$. The endomorphism ring $\mathrm{End}(U)$ of a non-zero uniserial module U has at most two maximal right ideals: the two-sided completely prime ideals $I_U := \{ f \in \mathrm{End}(U) \mid f \text{ is not injective} \}$ and $K := \{ f \in \mathrm{End}(U_R) \mid f \text{ is not surjective} \}$, or only one of them [7].

In order to discuss direct-sum decompositions of uniserial modules, we need to introduce the notions of monogeny class and epigeny class. These notions will turn out to be the "invariants" needed to classify direct-sum decompositions of uniserial modules (see Theorem 3.6 below).

Definition 3.5 [7] Two right R-modules M and N are said to have the same *monogeny class*, denoted by $[M]_m = [N]_m$, if there exist two right R-module monomorphisms $f : M \to N$ and $g : N \to M$. Similarly, M and N are said to have the same *epigeny class*, denoted by $[M]_e = [N]_e$, if there exist two right R-module epimorphisms $f : M \to N$ and $g : N \to M$.

For uniserial modules, we have the following weak form of the Krull-Schmidt Theorem.

Theorem 3.6 [7, Theorem 1.9] *Let $U_1, \dots, U_r, V_1, \dots, V_t$ be uniserial modules over an arbitrary ring R. Then*

$$U_1 \oplus \cdots \oplus U_r \cong V_1 \oplus \cdots \oplus V_t$$

if and only if $r = t$ and there are two permutations σ, τ of $\{1, 2, \dots, r\}$ such that $[U_i]_m = [V_{\sigma(i)}]_m$ and $[U_i]_e = [V_{\tau(i)}]_e$ for every $i = 1, 2, \dots, r$.

Therefore, if we consider the set $M(\mathcal{U})$ of all finite direct-sums of uniserial modules up to isomorphism, we have that $(M(\mathcal{U}), \oplus)$ is a half-factorial monoid, whose atoms are the non-zero uniserial modules.

As we have said before, this behaviour can be found in several classes of modules. But there are cases in which direct-sum decompositions are described by a higher number of invariants. Examples of these situations were studied in [2, 3, 4, 5]. It is worth noting that the invariants needed to describe direct-sum decompositions are closely related to the maximal ideals of the endomorphism rings of the modules.

We conclude with another example about non-unique factorizations (with respect to direct product) in the category of finite partially ordered sets. It is based on the non-unique factorization of the cyclotomic polynomial $x^5 + x^4 + x^3 + x^2 +^x +1$ in $\mathbb{N}_0[x]^*$ we have seen before.

Example 3.7 This example is due to Hashimoto and Nakayama [9, 10]. They showed that the Krull-Schmidt Theorem does not hold for finite partially ordered sets. The category of partially ordered sets has coproducts (disjoint unions $\dot\cup$) and products (direct products with the component-wise order $\times$). Let $L = \{0, 1\}$ be the partially ordered set with two elements $0 < 1$. For every $n \geq 0$, the direct product L^n is a connected partially ordered set with 2^n elements. Looking at the two factorizations $(x^3+1)(x^2+x+1) = (x+1)(x^4+x^2+1)$ of $x^5 + x^4 + x^3 + x^2 + x + 1$ into irreducible polynomials in $\mathbb{N}_0[x]$ (atoms of the commutative monoid $\mathbb{N}_0[x]^*$), we get two essentially different direct-product decompositions of the partially ordered set $1\dot\cup L\dot\cup L^2\dot\cup L^3\dot\cup L^4\dot\cup L^5$ into indecomposable partially ordered sets, given by

$$(L^3\dot\cup 1) \times (L^2\dot\cup L\dot\cup 1) \cong (L\dot\cup 1) \times (L^4\dot\cup L^2\dot\cup 1).$$

It is worth noting that this technique can be applied in studying the Krull-Schmidt property in any distributive category.

References

[1] H. Brunotte, *On some classes of polynomials with nonnegative coefficients and a given factor.* Period. Math. Hungar. 67 (2013), no. 1, 15–32.

[2] F. Campanini, *On a category of chain of modules whose endomorphisms rings have at most $2n$ maximal ideals.* Communications in Algebra 49 (2018), 1971–1982.

[3] F. Campanini, S.F. El-Deken and A. Facchini, *Homomorphisms with semilocal endomorphism rings between modules.* Algebr. Represent. Th., accepted (2019).

[4] F. Campanini and A. Facchini, *On a category of extensions whose endomorphisms rings have at most four maximal ideals.* In "Advances in Rings and Modules" S. López-Permouth, J.K. Park, C. Roman and S.T. Rizvi eds, Contemp. Math. 715 (2018), 107–126.

[5] F. Campanini and A. Facchini, *Factorization of polynomials with integral non-negative coefficients.* Semigroup Forum 99 (2018), 317–322.

[6] P. Cesarz, S.T. Chapman, S. McAdam and G.J. Schaeffer, *Elastic properties of some semirings defined by positive systems.* Commutative Algebra and its Applications (M. Fontana, S.-E. Kabbaj, B. Olberding and I. Swanson eds., deGruyter, 2009, pp. 89–101.

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

[8] A. Geroldinger and F. Halter-Koch, "Non-Unique Factorizations. Algebraic, Combinatorial and Analytic Theory". Pure and Applied Mathematics, vol. 278, Chapman & Hall/CRC, 2006.

[9] J. Hashimoto, *On the product decomposition of partially ordered sets.* Math. Japonicae 1 (1948), 120–123.

[10] J. Hashimoto and T. Nakayama, *On a problem of G. Birkhoff.* Proc. Amer. Math. Soc. 1 (1950), 141–142.

[11] C.E. van de Woestijne, *Factors of disconnected graphs and polynomials with nonnegative integer coefficients.* Ars Math. Contemp. 5 (2012), no. 2, 303–319.

Potential Theory and Boundary Element Method for the Laplace equation. An introduction.

Andrei-Florin Albisoru [*]

Abstract. We aim to give an overview of Potential theory for Laplace's equation. We introduce the fundamental solution of this equation. Next, we define the layer potentials and we state their properties. Using the layer potentials we will construct a solution of the interior Dirichlet problem for the Laplacian. We also describe a numerical method of solving Laplace's equation, namely the Boundary Element Method. Finally, we present some numerical results.

1 Introduction

Potential theory and the layer potential method have been used heavily in the study of elliptic boundary value problems (see, e.g., [4], [8]). This method is very useful for researchers in order to establish existence results. In the latter, we indicate some works in which the layer potential method plays a very important role, not only in the case of Laplace's equation, but in a more advanced setting such as the Stokes equations or the Brinkman equations. For additional information regarding Laplace's equation, see, e.g., [2, Chapter 2], [12, Chapter 3].

Going beyond the Laplace equation, we mention some particular works in which the layer potential method plays a very important role. Mitrea and Wright [9] have treated the main boundary value problems for the Stokes system in arbitrary Lipschitz domains in $\mathbb{R}^n$, for $n \geq 2$, using boundary integral methods. Kohr, Lanza de Cristoforis and Wendland [7] have used the methods of layer potential theory in order to establish existence results for boundary value problems for the the semilinear Brinkman system in Lipschitz domains in $\mathbb{R}^n$. Kohr, Lanza de Cristoforis, Mikhailov and Wendland [6] have used a layer potential method in order to prove the existence properties of linear and nonlinear transmission problems in Lipschitz domains in $\mathbb{R}^3$.

As Partridge, Brebbia and Wrobel described in [11], the Boundary Element Method is

[*] Faculty of Mathematics and Computer Science, Babes-Bolyai University, Cluj-Napoca, Romania. E-mail: `florin.albisoru@math.ubbcluj.ro` . Seminar held on 20 November 2019.

“an efficient alternative to the prevailing finite difference and finite element methods” and its attractiveness is its ”unique ability to provide a complete problem solution in terms of boundary values only”.

Although this method could be applied only to linear partial differential equations, the Dual Reciprocity Method has been developed to take into account non-homogeneous terms and keep the ”boundary-only” characteristic of the method (see e.g., [11, Chapters 2, 3, 4]).

We mention in the latter some very useful works for someone who wishes to study this numerical method. The book of Katsikadelis [5] provides an extensive look on the Boundary Element Method (as well as the DRM and another method called the Analog Equation Method) and diverse examples ranging from torsion of noncircular bars to fluid flow problems. Nishad, Chandra and Raja Sekhar [10] have used the Boundary Element Method to solve the Stokes equation in order to study the streamline profiles in a microchannel. Bozkaya [3] has provided an extensive overview of the Boundary Element Method and its application to diverse problems such as the MHD flow in rectangular ducts and in infinite regions, as well as the lid-driven cavity flow and natural convection cavity flow.

The structure of this work is as follows. In Section 2, we describe the interior Dirichlet problem for the Laplace operator. We introduce the notion of fundamental solution (see Definition 2.1). We introduce the Newtonian potential (see Definition 2.2) and we give a very important property (see Theorem 2.2). We define the single-layer and the double-layer potentials (see Definition 2.3) and state their properties (see Theorem 2.3 and Theorem 2.4). By exploiting auxiliary lemmas (see Lemma 2.5 - Lemma 2.8), one may show that indeed, as desired, the existence of a solution for the Dirichlet problem for the Laplacian holds (see Theorem 2.9). In Section 3, we solve numerically two boundary value problems using the Boundary Element Method. First, we state the problems (see (4) and (5)). Next, we give the main steps of the Boundary Element Method (see Steps 1-4). Finally, we apply those steps in order to obtain our desired numerical results. We conclude our study with two appendices that concern Fredholm operators and adjoint operators. In addition to their definition, we include several examples and properties are provided.

2 Potential Theory for Laplace’s equation

Unless otherwise specified, we assume that Ω is a bounded open subset of $\mathbb{R}^n$ with $n \geq 2$. We now consider the Poisson equation:

$$\Delta u = f \text{ in } \Omega, \tag{1}$$

where f plays the role of a given datum, u that of an unknown. Note that, in the particular case $f \equiv 0$, we obtain the Laplace equation

$$\Delta u = 0 \text{ in } \Omega. \tag{2}$$

Let $f \in C(\Omega)$ and $\eta \in C(\partial\Omega)$ be given functions. We concern ourselves with the following boundary value problems for the Poisson equation:

(a) **Dirichlet problem**: Find $u \in C^2(\Omega) \cap C(\overline{\Omega})$ such that

$$\begin{cases} \Delta u = f \text{ in } \Omega, \\ \quad u = \eta \text{ on } \partial\Omega. \end{cases}$$

(b) **Neumann problem**: Assume also that Ω is also of class C^1. Find $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ such that

$$\begin{cases} \Delta u = f \text{ in } \Omega, \\ \dfrac{\partial u}{\partial \mathbf{n}} = \eta \text{ on } \partial\Omega. \end{cases}$$

We emphasize that we have introduced above the basic problems for the Poisson equation. Later on we will exploit the volume potential to convert such problems into problems for the Laplace equation.

For convenience, we also state the analogous boundary value problems for the Laplace equation:

(a) **Dirichlet problem**: Find $u \in C^2(\Omega) \cap C(\overline{\Omega})$ such that

$$\begin{cases} \Delta u = 0 \text{ in } \Omega, \\ \quad u = \eta \text{ on } \partial\Omega. \end{cases}$$

(b) **Neumann problem**: Assume also that Ω is also of class C^1. Find $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ such that

$$\begin{cases} \Delta u = 0 \text{ in } \Omega, \\ \dfrac{\partial u}{\partial \mathbf{n}} = \eta \text{ on } \partial\Omega. \end{cases}$$

2.1 Fundamental Solution

We first look for radial solutions of the Laplace operator in $\mathbb{R}^n \setminus \{0\}$, i.e., functions of the form

$$u(\mathbf{x}) = v(\mathbf{r}),$$

where $\mathbf{r} = |\mathbf{x}|$ and v is a function from $(0, \infty)$ to $\mathbb{R}$.

Using the above ansatz one can obtain the fundamental solution of the Laplace equation (up to an additive constant; see also [12, Definition 3.1]).

Definition 2.1 The fundamental solution of the Laplace's equation is the function $\mathbf{G} : \mathbb{R}^n \setminus \{0\} \to \mathbb{R}$,

$$\mathbf{G}(\mathbf{x}) = \begin{cases} -\dfrac{1}{(n-2)\omega_n |\mathbf{x}|^{n-2}}, \text{ for } n \geq 3, \\ \dfrac{1}{2\pi} \log |\mathbf{x}|, \text{ for } n = 2, \end{cases}$$

where $\omega_n = 2\pi^{n-2}\Gamma(\frac{n}{2})$ and Γ denotes Euler's Gamma function.

We have the following theorem (see, e.g., [12, Proposition 3.1]).

Theorem 2.1 $\Delta \mathbf{G}(\mathbf{x}) = 0$ *for all* $\mathbf{x} \in \mathbb{R}^n \setminus \{0\}$.

2.2 Newtonian Potential

We introduce the Newtonian(volume) potential by the following definition.

Definition 2.2 The Newtonian potential of density f is defined by

$$(\mathcal{N}f)(\mathbf{x}) := \int_{\Omega} \mathbf{G}(\mathbf{x}-\mathbf{y})f(\mathbf{y})\mathrm{d}\mathbf{y},$$

whenever such integral converges.

We have the following useful result (see, e.g., [12, Theorem 3.23]).

Theorem 2.2 *Let $\Omega \subset \mathbb{R}^n$ be a bounded open set of class C^2. If $f \in C^1(\overline{\Omega})$, then $\mathcal{N}f \in C^2(\overline{\Omega})$ and*

$$\Delta \mathcal{N} f = f \ \textit{in}\ \Omega.$$

The following remark allows us to convert a Poisson equation to a Laplace equation.

Remark 1 If u solves the equation

$$\Delta u = f \text{ in } \Omega,$$

where $f \in C^1(\Omega) \cap C(\overline{\Omega})$, then $v := u - \mathcal{N}f$ solves

$$\Delta v = 0 \text{ in } \Omega.$$

Hence, theoretically, we are justified to study only the Laplace equation.

2.3 Layer Potentials

In this section we define the single layer and double layer potential and give their properties.

Let $\Omega \subset \mathbb{R}^n$ be a bounded and open set of class C^2. Let $g, h \in C(\partial\Omega)$.

Definition 2.3 Introduce the:

- **Single-layer potential** of density g, by the following relation:

$$(Vg)(\mathbf{x}) := \int_{\partial\Omega} \mathbf{G}(\mathbf{x}-\mathbf{y})g(\mathbf{y})\mathrm{d}\sigma_{\mathbf{y}}.$$

- **Double-layer potential** of density h, by the following relation:

$$(Wh)(\mathbf{x}) := -\int_{\partial\Omega} \frac{\partial \mathbf{G}}{\partial \mathbf{n}_y}(\mathbf{x}-\mathbf{y})h(\mathbf{y})\mathrm{d}\sigma_{\mathbf{y}}.$$

We give the properties of the double layer potential (see, e.g, [12, Theorem 3.30]).

Theorem 2.3

(i) $\Delta Wh = 0$ *in* $\mathbb{R}^n \setminus \partial\Omega$ *and* $Wh|_{\partial\Omega}$ *is continuous.*

(ii) *For all* $\mathbf{x} \in \partial\Omega$*, we have*

$$\begin{aligned} Wh(\mathbf{z}) &\to Wh(\mathbf{x}) + \frac{1}{2}h(\mathbf{x}) \ \textit{as}\ \mathbf{z} \to \mathbf{x}, \mathbf{z} \in \Omega, \\ Wh(\mathbf{z}) &\to Wh(\mathbf{x}) - \frac{1}{2}h(\mathbf{x}) \ \textit{as}\ \mathbf{z} \to \mathbf{x}, \mathbf{z} \in \mathbb{R}^n \setminus \overline{\Omega}. \end{aligned}$$

We also have the properties of the single layer potential (see, e.g, [12, Theorem 3.31]).

Theorem 2.4

(i) $\Delta Vg = 0$ *in* $\mathbb{R}^n \setminus \partial\Omega$ *and* $Vg|_{\partial\Omega}$ *is continuous.*

(ii) *For all* $\mathbf{x} \in \partial\Omega$*, we have*

$$\begin{aligned} \frac{\partial Vg}{\partial \mathbf{n}}^{-}(\mathbf{x}) &= \int_{\partial\Omega} \frac{\partial \mathbf{G}}{\partial \mathbf{n}_{\mathrm{x}}}(\mathbf{x}-\mathbf{y})g(\mathbf{y})\mathrm{d}\sigma_y + \frac{1}{2}g(\mathbf{x}), \\ \frac{\partial Vg}{\partial \mathbf{n}}^{+}(\mathbf{x}) &= \int_{\partial\Omega} \frac{\partial \mathbf{G}}{\partial \mathbf{n}_{\mathrm{x}}}(\mathbf{x}-\mathbf{y})g(\mathbf{y})\mathrm{d}\sigma_y - \frac{1}{2}g(\mathbf{x}). \end{aligned}$$

2.4 Method of Integral Equations

Remark 2 Layer potentials and the Newtonian potential can be used to prove the existence of solutions to boundary value problems for the Laplace (Poisson) equation.

We consider the Laplace equation.

(i) We seek the solution of the Dirichlet problem for the Laplacian in the form:

$$u(\mathbf{x}) = (Wh)(\mathbf{x}).$$

(ii) As for the solution of the Neumann problem for the Laplacian:

$$u(\mathbf{x}) = (Vg)(\mathbf{x}).$$

Taking into account the properties of the layer potentials, the interior Dirichlet and Neumann problems are equivalent to the following linear integral equations, respectively (see, e.g., [12, Subsection 3.14.6]):

$$\begin{aligned} &\frac{1}{2}h(\mathbf{x}) - \int_{\partial\Omega} \frac{\partial \mathbf{G}}{\partial \mathbf{n}_y}(\mathbf{x}-\mathbf{y})h(\mathbf{y})\mathrm{d}\sigma_y = \eta(\mathbf{x}),\ \mathbf{x} \in \partial\Omega, \\ &-\frac{1}{2}g(\mathbf{x}) + \int_{\partial\Omega} \frac{\partial \mathbf{G}}{\partial \mathbf{n}_x}(\mathbf{x}-\mathbf{y})h(\mathbf{y})\mathrm{d}\sigma_y = \eta(\mathbf{x}),\ \mathbf{x} \in \partial\Omega, \end{aligned}$$

where $\mathbf{G}$ is the fundamental solution.

We can rewrite, say the first integral equation (corresponding to the Dirichlet problem), in the following manner:

$$\left(\frac{1}{2}\mathbb{I} + \mathbf{K}\right) h = \eta. \tag{3}$$

In order to show that the equation (3) has a (unique) solution, one can show that the following results hold (see, e.g., [2, Lemma 2.8.1, Lemma 2.8.2, Lemma 2.8.3, Lemma 2.8.4]). Another way to tackle the existence problem is illustrated in Appendices of this study.

Lemma 2.5 *The operator* $\mathbf{K}$ *is continuous on* $C(\partial\Omega)$.

Lemma 2.6 *The operator* $\mathbf{K}$ *is completely continuous from* $L^2(\partial\Omega)$ *to itself.*

Lemma 2.7 *If* $\eta \in C(\partial\Omega)$ *and* $h \in L^2(\partial\Omega)$ *is a solution of (3), then* $h \in C(\partial\Omega)$.

Lemma 2.8 *Equation (3) has a unique solution* $h \in C(\partial\Omega)$.

The above lemmas allow us to state the existence result for the interior Dirichlet problem for the Laplacian (see, e.g., [2, Theorem 2.8.1]).

Theorem 2.9 *Assume that* Ω *is a bounded domain of class* C^2 *and* $\eta \in C(\partial\Omega)$. *Then, the Dirichlet problem for the Laplacian has a solution* $u \in C^2(\Omega) \cap C(\overline{\Omega})$.

We mention that in order to show the uniqueness of the solution, one would employ other the maximum principle or the Green formulas (see, e.g., [12, Theorem 3.6]).

3 The Boundary Element Method for Laplace's equation

In this section, we concern ourselves with the numerical study of two boundary value problems which will be described further on. We also present the necessary steps that one must follow in order to implement the Boundary Element Method and for convenience, we indicate how certain coefficients can be computed numerically. We use the approach described in [11, Chapter 2] (see also [5]).

Consider $\Omega \subset \mathbb{R}^2$. Our goal is to solve numerically the following two boundary value problems:

$$\begin{cases} \Delta u = 0 \text{ in } \Omega, \\ u = xy \text{ on } \partial\Omega, \end{cases} \tag{4}$$

and

$$\begin{cases} \Delta u = -1 \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega, \end{cases} \tag{5}$$

where $\Omega = [0,1] \times [0,1]$.

Using the fundamental solution of the Laplace equation (see Definition 2.1) and the second Green formula, one may show that (see, e.g., [5, (3.31)]):

$$\eta(\mathbf{p})u(\mathbf{p}) = -\int_{\partial\Omega} \left[\mathbf{G}(\mathbf{p}-\mathbf{p}')\frac{\partial u}{\partial \mathbf{n}}(\mathbf{p}') - u(\mathbf{p}')\frac{\partial \mathbf{G}}{\partial \mathbf{n}}(\mathbf{p}-\mathbf{p}') \right] \mathrm{d}\sigma, \ \mathbf{p}' \in \overline{\Omega}, \tag{6}$$

where $\eta(\mathbf{p}) = 1$ if $\mathbf{p} \in \Omega$ and $\eta(\mathbf{p}) = 1/2$ if $\mathbf{p} \in \partial\Omega$.

We need to discretize the boundary integral equation (6). We show how this can be done using the steps described below.

Step 1. We divide $\partial\Omega$ into a series of N small boundary elements Γ_j, $j = \overline{1, N}$.
In our case, we have the following:

- Γ_j - straight-line segment,
- $\mathbf{p}_{j-1} = (x_{j-1}, y_{j-1})$ and $\mathbf{p}_j = (x_j, y_j)$ - endpoints of Γ_j,
- $\tilde{\mathbf{p}}_j = (\tilde{x}_j, \tilde{y}_j)$ - midpoint of Γ_j,
- $\partial\Omega$ is approximated by $\cup_{j=1}^{N}\Gamma_j$.

Step 2. Assume that u and $\frac{\partial u}{\partial \mathbf{n}}$ are constant over each element Γ_j, $j = \overline{1, N}$ and equal to their values in $\tilde{\mathbf{p}}_j$, $j = \overline{1, N}$. We use the following notations

$$u(\mathbf{p}) = u(\tilde{\mathbf{p}}_j) =: u_j \text{ for } \mathbf{p} \in \Gamma_j.$$
$$\frac{\partial u}{\partial \mathbf{n}}(\mathbf{p}) = \frac{\partial u}{\partial \mathbf{n}}(\tilde{\mathbf{p}}_j) =: q_j \text{ for } \mathbf{p} \in \Gamma_j.$$

Remark 3 To increase the accuracy of the numerical solution, one may consider different boundary elements (i.e., linear or quadratic).

Step 3. Using the constant element assumption, we discretize our BIE as follows:

$$\frac{1}{2}u_i + \sum_{j=1}^{N} u_j \int_{\Gamma_j} \frac{\partial \mathbf{G}}{\partial \mathbf{n}_j}(\tilde{\mathbf{p}}_i - \mathbf{y})\mathrm{d}\Gamma_j = \sum_{j=1}^{N} q_j \int_{\Gamma_j} \mathbf{G}(\tilde{\mathbf{p}}_i - \mathbf{y})\mathrm{d}\Gamma_j. \tag{7}$$

In the latter, denote:

$$\overline{H}_{ij} = \int_{\Gamma_j} \frac{\partial \mathbf{G}}{\partial \mathbf{n}_j}(\tilde{\mathbf{p}}_i - \mathbf{y})\mathrm{d}\Gamma_j, \ G_{ij} = \int_{\Gamma_j} \mathbf{G}(\tilde{\mathbf{p}}_i - \mathbf{y})\mathrm{d}\Gamma_j,$$

and obtain:

$$\frac{1}{2}u_i + \sum_{j=1}^{N} \overline{H}_{ij}u_j = \sum_{j=1}^{N} G_{ij}q_j, \tag{8}$$

for every $i = \overline{1, N}$.

We now introduce

$$H_{ij} = \overline{H}_{ij} + \frac{1}{2}\delta_{ij},$$

that allows us to rewrite the equation (8) in the following manner

$$\sum_{j=1}^{N} H_{ij}u_j = \sum_{j=1}^{N} G_{ij}q_j.$$

Step 4. Solve the resulting system of N linear algebraic equations, after applying the boundary conditions. The system is

$$\mathbf{Hu} = \mathbf{Gq}. \tag{9}$$

After the vectors $\mathbf{u}$ and $\mathbf{q}$ are determined, the values of $\mathbf{u}$ can be computed at any internal point i, by the formula

$$u_i = \sum_{j=1}^{N} G_{ij}q_j - \sum_{j=1}^{N} \overline{H}_{ij}u_j, \tag{10}$$

but the coefficients $G_{ij}, \overline{H}_{ij}$ must be computed again for each different internal point.

It can be shown that (using Gaussian quadrature)

$$G_{ij} = \begin{cases} \frac{l_j}{2\pi}\left[\log(\frac{2}{l_j}) + 1\right], \text{ if } i = j \\ \sim \frac{lj}{4\pi}\sum_{k=1}^{4} \log(\frac{1}{r(\xi_k)}) \cdot w_k, \text{ if } i \neq j, \end{cases}$$

and

$$\overline{H}_{ij} = \begin{cases} 0, \text{ if } i = j \\ \sim -\frac{lj}{4\pi}\sum_{k=1}^{4} \frac{1}{r^2(\xi_k)}\left[r(\xi_k)_x n_x + r(\xi_k)_y n_y\right] \cdot w_k, \text{ if } i \neq j, \end{cases}$$

where

$$r = \sqrt{(x(\xi_k) - x_i)^2 + (y(\xi_k) - y_i)^2},$$
$$x(\xi_k) = \frac{x_{j+1} - x_j}{2}\xi_k + \frac{x_{j+1} + x_j}{2},$$
$$y(\xi_k) = \frac{y_{j+1} - y_j}{2}\xi_k + \frac{y_{j+1} + y_j}{2}.$$

We now return to (4) and (5).

Recall that we wish to solve numerically the following boundary value problem

$$\begin{cases} \Delta u = 0 \text{ in } \Omega, \\ u = xy \text{ on } \partial\Omega. \end{cases}$$

More precisely, our aim is to see how the number of boundary elements N influences the value of $u(0.5, 0.5)$.

Consequently, the program has been written in Python, and the system has been solved with the use of the Python routine np.linalg.solve. and we have obtained the following results:

N	$u(0.5, 0.5)$
8	0.2500179812626709
16	0.25000420387321753
20	0.25000266385349235
40	0.2500006588893487

Regarding the problem (5), recall that we would like to solve numerically:

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

More precisely, our goal is as follows. We want to see the distribution of u on the line $y = 0.5$.

To attain our goal, we can choose one of the following approaches:

- Transform the Dirichlet problem for the Poisson equation into a Dirichlet problem for the Laplacian.
- Apply an advanced version of BEM, that is the Dual Reciprocity BEM.

We have chosen the Dual Reciprocity BEM and we have also solved the problem using the Finite Difference Method and a comparison (which indicates a good agreement) of both methods is given in Figure 1.

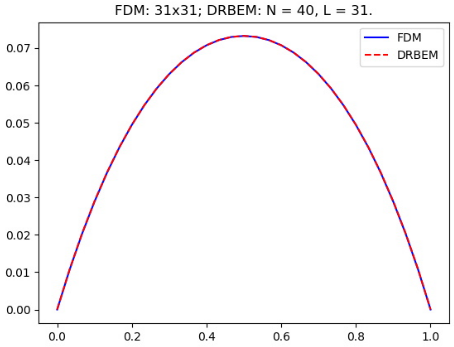


Figure 1. Distribution of u at $y = 0.5$.

Acknowledgement

The author is grateful to the members of this department and especially to Prof. Massimo Lanza de Cristoforis for their hospitality. The author is grateful to Prof. Mirela Kohr for her continuous support and encouragement.

Appendices

We describe in the latter the notions of Fredholm operator and adjoint operator and we state some of their properties.

Our motivations is as follows. We return to (3) having in mind the following question : If $\left(\frac{1}{2}\mathbb{I}+\mathbf{K}\right)$ is an isomorphism, then $h=\left(\frac{1}{2}\mathbb{I}+\mathbf{K}\right)^{-1}\eta$ and we have constructed a solution of our problem, in the form:

$$u = W\left(\left(\frac{1}{2}\mathbb{I}+\mathbf{K}\right)^{-1}\eta\right).$$

In order to study the properties of $\left(\frac{1}{2}\mathbb{I}+\mathbf{K}\right)$, we view this operator through the lenses of Fredholm operator theory, which will be described in the latter and the objective would be to show that this operator is Fredholm of index zero. For additional information we refer the reader to the work of Adams [1] and Wloka et al [13].

A Fredholm Operators

In this section, we give the definition of what means for an operator to be Fredholm. We also define the notion of a compact operator. Useful properties about Fredholm operators are illustrated, also. All notions are accompanied by examples.

Let X, Y be Banach spaces.

Definition A.1 Let $A \in LC(X,Y)$. Then A is a Fredholm operator if the following hold

(i) $\dim \operatorname{Ker}(A) = n_0 < +\infty$,

(ii) $\dim(Y/\operatorname{Im}(A)) = n_1 < +\infty$.

We introduce also the index of A which is given as

$$\operatorname{ind}(A) := n_0 - n_1 < +\infty.$$

Example A.2 Let $\mathbf{H}$ be a Hilbert space with an orthonormal basis $(e_k)_{k\in\mathbb{N}}$. The operator $S:\mathbf{H}\to\mathbf{H}$, given by

$$S(e_k) = e_{k+1}, k \geq 0,$$

is Fredholm with $\operatorname{ind}(S) = -1$.

Definition A.3 Let X, Y be normed spaces and let $A \in L(X,Y)$. A is compact if it maps bounded sets from X into relatively compact sets in Y.

Example A.4 The operator $T_n : \ell^2 \to \ell^2$ given by

$$T_n((x_k)_{k\in\mathbb{N}}) := (x1, x2, ..., x_n, 0, 0, ...)$$

is compact.

Example A.5 The operator $T : L^2([0,1]) \to L^2([0,1])$, defined by

$$T[f](x) := \int_0^1 K(x,y)f(y)dy,$$

where $K \in C([0,1]^2)$, is a compact operator.

Lemma A.1 *Let X, Y be Banach spaces. Let A be a Fredholm operator. If $K \in L(X,Y)$ such that K is compact, then $A + K : X \to Y$ is a Fredholm operator, and*

$$\operatorname{ind}(A + K) = \operatorname{ind}(A).$$

Example A.6 Define the compact operator $T : \ell^2 \to \ell^2$ by

$$T((x_n)_{n\in\mathbb{N}}) = (t_n x_n)_{n\in\mathbb{N}},$$

where $(t_n)_{n\in\mathbb{N}} \to 0$.

One can show that, indeed,

$$\operatorname{ind}(I + T) = \operatorname{ind}(I),$$

where $I : \ell^2 \to \ell^2$ is the identity operator.

Corollary A.1.1 *Let X, Y be Banach spaces and $A : X \to Y$ be a Fredholm operator of index 0. Then A is an isomorphism if one of the following (equivalent) condition holds:*

(i) *A is one-to-one.*

(ii) *A is onto.*

B Adjoint Operators

In this section, we give a short overview of adjoint operators. Such operators are very useful in the situation in which the properties of the operator under study cannot be easily ascertained. To this end, one defines the adjoint operator and studies its properties. We end this section with a lemma that links a Fredholm operator with its adjoint.

Definition B.1 Let X be a Banach space. The anti-dual of X is defined by

$$X^* := \{F : X \to \mathbb{C} : F \text{ is antilinear and continuous}\}.$$

Definition B.2 Let $B \in LC(X, Y)$. Then, the adjoint of B, $B^* : Y^* \to X^*$ is defined by

$$\langle Bx, y^* \rangle = \langle x, B^* y^* \rangle,$$

for all $x \in X$, $y^* \in Y^*$.

Example B.3 Let $K : L^2([0, 1]) \to L^2([0, 1])$, given by

$$(Kf)(x) := \int_0^1 k(x, y) f(y) \mathrm{d}y,$$

where $k : [0, 1]^2 \to \mathbb{C}$. Then, the adjoint is given by

$$(K^* f)(x) = \int_0^1 \overline{k(x, y)} f(y) \mathrm{d}y.$$

Indeed, we have in $L^2([0, 1])$ the following relations:

$$\begin{aligned}\langle f, K^* g \rangle &= \int_0^1 f(y) \overline{(K^* g)(y)} \mathrm{d}y = \int_0^1 \int_0^1 k(x, y) f(y) \mathrm{d}y \overline{g(x)} \mathrm{d}x \\ &= \langle Kf, g \rangle.\end{aligned}$$

Lemma B.4 *If $B \in LC(X, Y)$ is a Fredholm operator, then $B^* \in LC(Y^*, X^*)$ is also a Fredholm operator and*

$$\mathrm{ind}(B^*) = -\mathrm{ind}(B).$$

References

[1] R.A. Adams, "Sobolev spaces". Academic Press, 2003.

[2] V. Barbu, "Partial Differential Equations and Boundary Value Problems". Mathematics and Its Applications, Vol. 441, Springer-Science+Business Media, Dordrecht, 1998.

[3] C. Bozkaya, "Boundary Element Method Solution of Initial and Boundary Value Problems in Fluid Dynamics and Magnetohydrodynamics". PhD Thesis, 2008.

[4] M. Costabel, *Boundary Integral Operators on Lipschitz Domains: Elementary Results.* SIAM J. Math. Anal. 3/19 (1988), 613–626.

[5] J.T. Katsikadelis, "The Boundary Element Method for Engineers and Scientists. Theory and Applications". Second Edition, Academic Press, Elsevier, London, 2016.

[6] M. Kohr, M. Lanza de Cristoforis, S.E. Mikhailov, W.L. Wendland, *Integral potential method for a transmission problem with Lipschitz interface in* $\mathbb{R}^3$ *for the Stokes and Darcy-Forchheimer-Brinkman PDE Systems.* Z. Angew. Math. Phys. 67:116, 5 (2016), 1–30.

[7] M. Kohr, M. Lanza de Cristoforis, W.L. Wendland, *Poisson problems for semilinear Brinkman systems on Lipschitz domains in* $\mathbb{R}^n$. Z. Angew. Math. Phys. 66 (2015), 833–864.

[8] M. Kohr, I. Pop, "Viscous Incompressible Flow for Low Reynolds Numbers". WIT Press, Southampton, 2004.

[9] M. Mitrea, M. Wright, "Boundary value problems for the Stokes system in arbitrary Lipschitz domains". Astérisque 344, viii+241 pp., 2012.

[10] C.S. Nishad, A. Chandra, G.P. Raja Sekhar, *Stokes Flow Inside Topographically Patterned Microchannel Using Boundary Element Method.* International Journal of Chemical Reactor Engineering 15:5 (2017), 1–17.

[11] P.W. Partridge, C.A. Brebbia, L.C. Wrobel, "The Dual Reciprocity Boundary Element Method". Computational Mechanics Publications, Southampton, 1992.

[12] R. Precup, "Linear and Semilinear Partial Differential Equations. An Introduction". De Gruyter, Berlin-Boston, 2013.

[13] J.T. Wloka, B. Rowley, B. Lawruk, "Boundary Value Problems for Elliptic Systems". Cambridge University Press, 1995.

A smooth introduction to the semi-classical problem in Quantum Mechanics

Enrico Picari [(*)]

Abstract. It is very well-known today that Quantum Mechanics plays a crucial role in describing and understanding nature, at least (but not only) at the smallest scales, i.e. the ones of atoms and subatomic particles. The problem of connecting physical and mathematical features of Classical and Quantum Mechanics dates back to the early days of the theory and although reduction of one to the other is understood heuristically in terms of a limit process in which the Planck constant goes to zero, the mathematical ground of such procedure, known as Semi-classical Analysis, is still an active field of research which includes techniques of apparently distant topics like Harmonic Analysis and Deformations of Poisson Algebras. The aim of this note is to introduce the main features of elementary Quantum Mechanics, with a brief historical note, and to give an insight into the state-of-the-art of the semi-classical limit problem.

1 A brief review of Classical mechanics

It is very well known that *Hamiltonian mechanics* is one possible way to describe the motion of classical particles, equivalent to Newton's laws but rather more abstract. For a single particle moving inside some region $M \in \mathbb{R}^3$, the time evolution of its position x and momentum p obeys the following differential equation

$$\begin{cases} \dot{x}_j = \dfrac{\partial H}{\partial p_j} \\ \dot{p}_j = -\dfrac{\partial H}{\partial q_j} \end{cases} \qquad j = 1, 2, 3$$

for a prescribed *Hamiltonian function* H depending on x and p, which represents particle's *energy*. Generally speaking, if M is sufficiently regular manifold, the pair (x, p) is an element of the *cotangent bundle* $T^*M =: \Gamma$, called *phase space* in this context, and thus $H\colon \Gamma \to \mathbb{R}$.[(1)] This formalism can be extended as well to systems of many interacting

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: `picari@math.unipd.it` . Seminar held on 11 December 2019.

[(1)] H can depend on t also, but for simplicity we won't consider this case here.

particles: the *state* of the system is a point in Γ and it evolves along the integral curves of

$$(1) \qquad \begin{cases} \dot{x}_j = \dfrac{\partial H}{\partial p_j} \\ \dot{p}_j = -\dfrac{\partial H}{\partial q_j} \end{cases} \qquad j = 1, \dots, n$$

where n is the dimension of M. Some instances of H for n particles moving without contraints in $\mathbb{R}^3$ are, for example

(i) free particles: their energy is purely kinetic and $H = \sum_{j=1}^{3n} \frac{p_j^2}{2m_j}$, where m_j is the mass of the j-th particle;

(ii) particles moving in a gravitational field in near-Earth approximation: $H = \sum_{j=1}^{3n} \frac{p_j^2}{2m_j} + m_j g z_j$, where z_i is the coordinate along the direction of gravity acceleration $\boldsymbol{g}$;

(iii) charged particles interacting via Coulomb potential: $H = \sum_{j=1}^{3n} \frac{p_j^2}{2m_j} - \sum_{j<k} \frac{q_j q_k}{|x_j - x_k|}$, where q_j is the j-th particle's charge;

and so on. We introduce now some concepts that will be extremely useful to understand basic features of semi-classical analysis. It is an elementary fact in ordinary differential equations theory that the solutions of systems of the form (1) define a *flow* on phase space, that is a map $\Phi^t \colon \Gamma \to \Gamma$ which maps the initial data (x, p) of the Cauchy problem associated with Eq. (1) to the solution at time t

$$\Phi^t(x, p) = (x(t), p(t))$$

which has some nice group properties

(i) $\Phi^0 = \mathrm{Id}_\Gamma$;

(ii) $\Phi^{(t+s)} = \Phi^t \circ \Phi^s$;

(iii) $\Phi^{-t} = (\Phi^t)^{-1}$;

and will be central in the following discussion.[2]

A concept of capital importance in the study of physical systems is the one of *observable*, that is some quantity which can me measured in an experiment, such as particles' positions, their angular momentum, energy and so on and are usually modeled as smooth functions $f \in \mathcal{C}^\infty(\Gamma)$, or some other functional space depending on the application one has in mind. Since we're often interested in measuring how observables evolve along the Hamiltonian flow, it is desirable to write down some evolutionary equation for them and

[2] In general property (iii) holds only if the vector field at right-hand side of (1) is *complete*, but this fact is not so important in these notes.

this can be made by defining *Poisson brackets* between smooth functions, that is a binary operation $\{\cdot,\cdot\}\colon C^\infty(\Gamma)\times C^\infty(\Gamma)\to C^\infty(\Gamma)$

$$\{f,g\}=\sum_j\left(\frac{\partial f}{\partial x_j}\frac{\partial g}{\partial p_j}-\frac{\partial f}{\partial p_j}\frac{\partial g}{\partial x_j}\right)$$

so that the variation of f along the solutions of Eq. (1) can be written as

$$\frac{d}{dt}f\circ\Phi^t=\{f\circ\Phi^t,H\}.$$

It is worth noting that Hamilton's equations themselves can be written in this form by taking $f=x_i$ and $f=p_i$. Poisson brackets satisfy the following properties

(i) bilinearity;

(ii) skew-symmetry: $\{f,g\}=-\{g,f\}$;

(iii) Jacobi identity$\{\{f,g\},h\}+\{\{g,h\},f\}+\{\{h,f\},g\}=0$;

(iv) Leibniz rule: $\{fg,h\}=f\{g,h\}+\{f,h\}g$,

giving $C^\infty(\Gamma)$ the structure of a *Poisson algebra*, that is a Lie algebra with a derivation. To end these quick reviews of Hamiltonian mechanics we establish the so-called *canonical commutation relation*, that is we compute Poisson bracket between coordinates and momenta

$$\begin{aligned}&\{q_j,q_k\}=\{p_j,p_k\}=0\\&\{q_j,p_k\}=\delta_{jk}\end{aligned}\tag{2}$$

which, despite their simplicity, play a central role in quantum-classical correspondence.

2 The crisis of classical physics

Hamilton's equations are a powerful tool to describe how nature works: for a given system, a physicist can make a guess about the correct Hamiltonian he should choose and then he could try to integrate them, analytically or numerically, or at least to understand the qualitative behaviour of the solutions using techniques of the theory of dynamical systems. However at the turn of the 19th and 20th centuries it became clear that the Hamiltonian picture gives wrong predictions for certain systems, that is the ones at very short length scales such as molecula and atoms. For example, at the time the most popular model for the Hydrogen atom was the *planetary model*: an electron of negative charge moves in closed orbits around a proton of positive charge, like planets do around the Sun in the Solar System. While this model was confirmed by Rutherford's experiments around 1911, every attempt to use classical electromagnetism and Hamiltonian mechanics fails to predict the stable behavior of electron's motion. Indeed, according to *Larmor formula* a charged particle which moves along trajectory of non-zero curvature (like a closed curve), should loose energy due to interaction with the electromagnetic field of some other charge in proportion to the squared norm of its acceleration

$$\frac{dE}{dt} \propto -||\boldsymbol{a}(t)||^2 \qquad \text{(Larmor formula).}$$

This loss forces the electron to fall towards the nucleus and the atom should collapse, while we know that matter is indeed stable!

Moreover, it was observed in experiments in spectroscopy that energy exchanges between atoms and radiation are not continuous, as suggested by classical theory, but they are *quantized*, namely they come as *integer* multiples of some *energy packet*

$$E_0 = \hbar\omega,$$

where ω is the frequency of an electromagnetic wave interacting with the atom and $\hbar$ is a proportionality constant found by interpolation with experimental data. That was a completely new phenomenon at the time and it was completely inexplicable by the sole use of any classical theory. The constant $\hbar$ seemed to pop up every time physicists tried to fit experimental results regarding atomic particles and was soon regarded as a *fundamental constant of nature*. Nowadays we call it *Planck constant*, its magnitude is

$$\hbar \approx 1.054 \times 10^{-34} \text{ Joule} \cdot \text{seconds}$$

and has a prominent role in every known *quantum theory*.

3 De Broglie hypothesis and Schrödinger equation

During 1924, quantization of energy was interpreted as a *wave condition* by the french physicist L. De Broglie: he supposed that to each electron in the atom there exists an associated wave whose wavelength λ and frequency ω are related to particle's momentum and energy by

$$p = \frac{\hbar}{\lambda}, \qquad E = \hbar\omega, \qquad \text{(De Broglie relations).}$$

The enigmatic character of such statement was quite well highlighted by De Broglie himself in his Nobel Prize speech:

> ”*Determination of the stable motion of electrons in the atom introduces integers, and up to this point the only phenomena involving integers in physics were those of interference and of normal modes of vibration. This fact suggested to me the idea that electrons too could not be considered simply as particles, but that frequency (wave properties) must be assigned to them also.*”

However, De Broglie's hypothetis was exploited by austrian physicist E. Schrödinger in a series of paper in 1926. Following the same reasoning which leads from wave theory of light to optical geometry and proceeding by analogy using Hamilton-Jacobi theory (see e.g. [4]), he argued that if a particle of mass m interacts with some potential V depending on position, then its associated wave function ψ should evolve in time according to the equation

$$i\hbar\frac{\partial}{\partial t}\psi(t,x) = -\frac{\hbar^2}{2m}\Delta\psi(t,x) + V(x)\psi(t,x). \tag{3}$$

The first example of usage of Schrödinger equation is the time-independent case. As in ordinary wave theory, a solution ψ of Eq. (3) is said to be *stationary* if its time dependence can be factored out

$$\psi(t,x) = e^{-it\omega}\phi(x).$$

By identifying particle's energy with its frequency $E = \hbar\omega$, it's easy to see that then ϕ satisfies

$$-\frac{\hbar^2}{2m}\Delta\phi(x) + V(x)\phi(x) = E\phi(x)$$

which has the form of an eigenvalue equation for the operator $H = -\frac{\hbar^2}{2m}\Delta + V$, called *Hamiltonian operator*. One can consider, for example, the Hydrogen atom case: ϕ represents electron's wave function, while V is the Coulomb interaction between the electron and the proton

$$-\frac{\hbar^2}{2m}\Delta\phi(x) - \frac{e^2}{|x|}\phi(x) = E\phi(x). \tag{4}$$

Eq. (4) admits indeed solutions under suitable boundary conditions on ϕ, namely fast decreasing for $|x| \to \infty$, for negative values of E given by

$$E_n = -\frac{me^4}{2n^2\hbar}. \tag{5}$$

The corresponding functions ϕ_n then represent electron's *state* and are given in terms of Laguerre polynomials and spherical harmonics (see e.g. [7]). Formula (5) called *Rydberg's formula*, is in perfect accordance with experimental data on the energy levels of Hydrogen atom.

4 Hilbert space approach and canonical operators

Let us consider an initial value problem for the Schrödinger equation for a function $\psi\colon \mathbb{R} \times \mathbb{R}^3 \to \mathbb{C}$

$$\begin{cases} i\hbar\partial_t\psi(t,x) = (H\psi)(t,x) \\ \psi(0,x) = \psi_0(x), \end{cases}$$

where the Hamiltonian operator H is defined on a suitable space of functions. One immediate consequence of the equation form is that if the initial value ψ_0 is square integrable, that is $\psi_0 \in L^2(\mathbb{R}^3)$, then the solution $\psi_t := \psi(t,\cdot)$ is again in $L^2(\mathbb{R}^3)$ *for all times.* In particular, for a normalized initial condition, one has

$$||\psi_t||_{L^2} = ||\psi_0||_{L^2} = 1. \tag{6}$$

This particular feature allowed early workers on quantum theory to interpret the wave function of a particle *statistically*: ψ represents the probability amplitude for particles position, that is the probability of finding a particle in a certain region $B \subset \mathbb{R}^3$ at time t is

$$\int_B |\psi_t(x)|^2 \mathrm{d}x.$$

It follows from elementary probability theory that the mean value of a certain coordinate, say the j-th one, can be computed by

$$\int_{\mathbb{R}^3} x_j |\psi_t(x)|^2.$$

Then by defining the j-th position operator in $L^2(\mathbb{R}^3)$

$$(X_j\varphi)(x) := x_j \varphi(x),$$

we see that it encodes entirely the mean value by means of inner product in L^2

$$\int_{\mathbb{R}^3} x_j |\psi_t(x)|^2 = \int_{\mathbb{R}^3} \overline{\psi_t(x)} x_j \psi_t(x) = \langle \psi_t, X_j \psi_t \rangle.$$

By analogy with the classical case one can define a derivation operator

$$(P_j\varphi)(x) = -i\hbar \frac{\partial \varphi}{\partial x_j}(x)$$

on a certain domain in $L^2(\mathbb{R}^3)$ [(3)], so that the Hamiltonian operator can be written as

$$H = \frac{1}{2m} \sum_{j=1}^{3} P_j^2 + V(x).$$

So we have three operators resembling classical observables such as position, momentum and energy which are self-adjoint operators on L^2 once a correct domain is given (see the footnote below):

$$\begin{aligned} \langle \phi, X_j \chi \rangle &= \langle X_j^* \phi, \chi \rangle = \langle X_j \phi, \chi \rangle, \\ \langle \phi, P_j \chi \rangle &= \langle P_j^* \phi, \chi \rangle = \langle P_j \phi, \chi \rangle, \\ \langle \phi, H \chi \rangle &= \langle H^* \phi, \chi \rangle = \langle H \phi, \chi \rangle, \qquad \text{(requires V real)}, \end{aligned}$$

but what happens if one would like to consider more general quantum observables?

[(3)] usually the minimal domain for derivations is the space of compactly supported smooth functions C_0^∞, but for many purposes it is preferable to take the Schwartz space $\mathcal{S}$.

5 Axioms of Quantum Mechanics and time evolution

The first attempt to arrange some minimal hypotheses on the mathematical structure of a Quantum Theory was made in 1930 by P.A.M. Dirac. His axioms are now well known and were refined during successive decades in order to include the role of relativity (see [1] or [8] for the algebraic aspects), but they are sufficient for our discussion on the semi-classical problem:

A1. A state ψ of a quantum system is a unit vector of some Hilbert space $\mathfrak{h}$, up to scalar multiples.

A2. A quantum observable is represented by some self-adjoint (possibly unbounded) operator A on $\mathfrak{h}$.

A3. The possible outcomes of an experiment measuring the observable A are contained in the spectrum of A.

A4. The expectation value of the observable A for a system in a state ψ is given by the inner product $\langle \psi, A\psi \rangle$.

Even in this abstract setting, time evolution of a quantum system is still carried by a Schrödinger equation, which should now be regarded as an abstract differential equation on $\mathfrak{h}$. An equivalent way of defining quantum evolution is the following: consider the equation

$$i\hbar\partial_t \psi_t = H\psi_t, \tag{7}$$

for ψ_t in the domain of H, and define a t-dependent operator U_t which sends the initial condition ψ_0 to the solution of (7) at time t: $U_t\psi_0 = \psi_t$. Then U_t has the following properties:

- $U_0 = \mathbb{I}$
- group property $\rightsquigarrow$ $U_{t+s} = U_t U_s$ and $U_{-t} = U_t^{-1}$;
- unitarity $\rightsquigarrow$ $U_t^* U_t = U_t U_t^* = \mathbb{I}$;
- satisfies (in the strong sense) the operator differential equation

$$i\hbar\partial_t U_t = HU_t \qquad \Longrightarrow \qquad U_t = e^{-itH/\hbar}.$$

Time evolution on states can be as well transferred to observables. Indeed, given a self-adjoint operator A on $\mathfrak{h}$, define the time dependent operator

$$A(t) = U_t^* A U_t.$$

Then by formal derivation with respect to t, $A(t)$ satisfies the *Heisenberg equation of motion*

$$i\hbar\partial_t A(t) = -HU_t^* AU_t + U_t^* AU_t H = [A(t), H],$$

where the commutator $[C, D] = CD - DC$ between operators is introduced. The motivation is simple: since one is interested in the possible values attained by the observable A during and experiment on a system whose state is ψ_t, then, by Dirac's axioms, its expectation value along ψ_t is given by

$$\langle \psi_t, A\psi_t \rangle = \langle U_t\psi_0, AU_t\psi_0 \rangle = \langle \psi_0, U_t^* AU_t\psi_0 \rangle = \langle \psi_0, A(t)\psi_0 \rangle.$$

In other words, the values attained by A can be determined both by evolution of states and evolution of its associated operator.

6 Correspondence principle and quantization

From the form of commutator, it's easy to see that $[\cdot, \cdot]$ is a bilinear and skew-symmetric application on operators and it satisfies Jacobi identity and Leibniz property. Recalling the Hamiltonian formulation of classical mechanics, we have the following scenario:

Classical mechanics

- States: points in phase space Γ
- Observables: sufficiently regular functions on Γ
- Flows: $\Phi^t\colon \Gamma \to \Gamma$
- Poisson algebra of observables: $\{f, g\}$
- Canonical commutations: $\{x_j, p_k\} = \delta_{jk}$
- Evolution of observables: $\dot{f} = \{f, H\}$

Quantum mechanics

- States: elements in $\mathfrak{h}$
- Observables: self-adjoint operators on $\mathfrak{h}$
- Flows: $U_t\colon \mathfrak{h} \to \mathfrak{h}$
- Poisson algebra of observables: $[A, B]$
- Canonical commutations: $[X_j, P_k] = i\hbar\delta_{jk}$
- Evolution of observables: $i\hbar\dot{A}(t) = [A(t), H]$

Both theories work well in their natural environment: Classical Mechanics is an excellent description of systems at ordinary length scales, while Quantum Mechanics is the only available theory at atomic length scales and becomes over-complicated at ordinary ones. So one would like to have a criterion for which one theory *reduces* to the other. This is often understood at heuristic level as a *limit process*: Quantum Mechanics should reduce, in some sense, to Classical Mechanics if Planck's constant is small compared to classical actions, in other terms the reduction should take place if we perform the formal limit $\hbar \to 0^+$. The task of *semi-classical* analysis is precisely to make rigorous a statement of this kind.

The first problem one has to address is the following: considering the one-dimensional case for simplicity, the *quantization* of the canonical variables (x, p) is well understood by setting

$$x \to X, \qquad p \to P = -i\hbar\frac{d}{dx}, \tag{8}$$

that is, given a classical Hamiltonian $H = H(x,p)$ its quantum version is obtained by taking the same functional form a performing the substitution in Eq. (8). This, however cause some ordering ambiguities, since if $H(x,p)$ contains a term like xp, then one should choose between the two non-equivalent quantizations XP and PX, which moreover are not self-adjoint and thus cannot represent some observable. However, the symmetrized product $(XP+PX)/2$ has no ordering ambiguities and is self-adjoint. Generalizing a bit, there is indeed a classic result in analysis regarding quantization of functions.

Proposition 1 *For any function $a\colon \mathbb{R}^{2n} \to \mathbb{C}$ with suitable growing property at infinity*[4], *there exist an operator-valued map $Q\colon a \to Q(a)$, whose action on Schwartz functions is given by*

$$(Q(a)u)(x) = \frac{1}{(2\hbar\pi)^2} \int_{\mathbb{R}^{2n}} a\left(\frac{x+y}{2}, p\right) e^{\frac{i}{\hbar}(x-y)\cdot p} u(y) dy d\xi \tag{9}$$

with the properties

(i) $a \to Q(a)$ *is linear;*

(ii) $Q(a)$ *is self-adjoint if a is real;*

(iii) $Q(a) = Id$*;*

(iv) *for monomials of the form $x^n p^m$, $Q(x^n p^m)$ is the sum of all possible permutations of n times the X operator and m times P.*

The operator $Q(a)$ is called Weyl quantization *of a and if a certain operator A is the quantization of some function a, a is called* Weyl symbol *of A.*

The integral in Eq. (9) can be defined in general defined for smooth functions which have at most polynomial growth at infinity together with their derivatives, that is for all multi-indices $\gamma \in \mathbb{N}^n$

$$|\partial^\gamma_{x,p} a(x,p)| \leq C_\gamma \langle (x,p) \rangle^m, \qquad \langle (x,p) \rangle := (1 + |x|^2 + |p|^2)^{1/2} \tag{10}$$

for some constant C_γ and some real number m. This space of functions is denoted as $S(m)$, its elements are referred to as *classical symbols of order m* and has the structure of a Fréchet space with seminorms

$$\sup_{(x,p)\in\mathbb{R}^{2n}} \sup_{|\gamma|\geq k} \langle (x,p) \rangle^{-m} |\partial^\gamma_{x,p} a(x,p)| =: |a_{m,k}|.$$

One of the most important features of definition in Eq. (9) is the *composition property.*

Proposition 2 *Let $a \in \mathcal{S}(m_1)$ and $b \in \mathcal{S}(m_2)$ two classical symbols. Then there exist a classical symbol $c \in \mathcal{S}(m_1+m_2)$ such that $Q(c) = Q(a) \circ Q(b)$. Moreover c has the explicit representation*

$$c(x,p;\hbar) = \frac{1}{(\pi\hbar)^{2n}} \int_{\mathbb{R}^{4n}} e^{-\frac{2i}{\hbar}(y\cdot\theta - z\cdot\eta)} a(x+y, p+\eta) b(x+z, p+\theta) dy dz d\eta d\theta. \tag{11}$$

[4] see the discussion below.

It is worth to notice that the correspondence $(a, b) \to c$ is bilinear and c has an explicit dependence on $\hbar$ but more importantly it has an asymptotic expansion in terms of $\hbar$

$$c(x, p; \hbar) = \sum_{k=0}^{N} \frac{(i\hbar/2)^k}{k!} (\partial_x \partial_\eta - \partial_y \partial_p)^k a(x, p) b(y, \eta) \Big|_{\substack{y=x \\ \eta=p}} + \hbar^{N+1} R_N(x, p),$$

where the remainder term R_N is in $\mathcal{S}(m_1 + m_2)$. Recalling the definition of Poisson brackets, the first terms in the expansion can be expressed as[(5)]

$$c(x, p; \hbar) = a(x, p) b(x, p) + \frac{i\hbar}{2}\{a, b\} + \mathcal{O}_{S(m_1+m_2)}(\hbar^2),$$

so c can be regarded as a non-commutative *deformation* of the usual product of function and is denoted as $a \star b$. As a corollary we have that the symbol of the commutator of $Q(a)$ and $Q(b)$ can be written as an asymptotic series which turns out to be a deformation of Poisson brackets, usually called *Moyal brackets* of a and b:

$$\{a, b\}_\star = i\hbar\{a, b\} + \mathcal{O}_{S(m_1+m_2)}(\hbar^2).$$

7 Egorov theorem

From the preceding discussion, we see that it is useful to consider a somewhat larger class of symbols $\mathcal{S}_{sc}(m)$, that is the ones which admit some asymptotic expansion of the form

$$c \sim \sum_k \hbar^k c_k,$$

where all the c_k's and all the possible remainders are $\mathcal{S}(m)$ for some $m \in \mathbb{R}$. Symbols with this property are called *semi-classical symbols* of order m. With this in mind, we are ready to state a form of the *Egorov theorem*, that is a result in semi-classical analysis which connects the Heisenberg evolution of quantum observables emerging from the quantization of some symbols and the classical hamiltonian flow of their symbols (for the original statement and proof, see [3]).

Proposition 3 *Let H be a semiclassical observable with sub-quadratic growth at infinity*

$$|\partial^\gamma_{x,p} H_j(x, p)| < C_{\gamma,j}, \qquad \textit{with } |\gamma| + j \geq 2,$$

and let $a \in S(m)$ for some $m \in \mathbb{R}$.
Then there exist a $\hbar_0 > 0$ such that for all $\hbar \in [0, \hbar_0)$ we have

(i) *$Q(H)$ is self-adjoint $\mathcal{S}(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ so that U_t is unitary for all $t \in \mathbb{R}$;*

(ii) *$A(t) = U_t^* Q(a) U_t$ is the quantization of a symbol $a_t \in S_{sc}(m)$, $A(t) = Q(a_t)$;*

[(5)] the "$\mathcal{O}$ notation" refers to the semi-norms $|\cdot|_{m,k}$ introduced before.

(iii) *there exist a $T(\hbar) > 0$ such that, uniformly in $[-T(\hbar), T(\hbar)]$, a_t has an asymptotic expansion*

$$a_t \sim \sum_k \hbar^k a_k(t), \quad a_k(t) \in S(m).$$

(iv) *All the $a_k(t)$'s can be explicitly computed. In particular the principal and sub-principal symbols are given by*

$$\begin{aligned} a_0(t; x, p) &= a(\Phi^t(x, p)) \\ a_1(t; x, p) &= \int_0^t \{a \circ \Phi^s, H_1\} \circ \Phi^{(t-s)}(x, p) ds, \end{aligned}$$

where Φ^t is the Hamiltonian flow of the principal symbol H_0 of H.

(v) *$T(\hbar)$ is of order $-log(\hbar)$.*

Although the statement of the theorem is quite involved, it gives a precise meaning to the Correspondence Principle. Indeed, according to this result quantum and classical evolution *stay close* in terms of symbols

$$a_t(x, \xi) = a(\Phi^t(x, \xi)) + \mathcal{O}_{S(m)}(\hbar), \qquad |t| \leq T(\hbar),$$

at least for a certain time in an observation window given approximatively as $\sim [0, -\log \hbar)$. The technical assumptions on the semi-classical hamiltonian H seem to be quite restrictive, since many hamiltonian functions of interest (like the one in FPU model [5], the DNLS hamiltonian [6] and so on) grow at least as a polynomial of degree strictly greater than 2. However, in some cases the presence of *first integrals* for the flow of H (see [2]) allows to reduce the dynamics from $\mathbb{R}^{2n}$ to some compact manifold M and in this case symbols can be taken to be some reduction on M if cut-offs outside the manifold are introduced (see [9]). On the other hand, Proposition 3 deals only with *uniform* estimates on symbols in (x, p), so one can ask himself if the result of Egorov theorem can be enlarged to include some weaker type of convergence, e.g. introducing some probability density in (x, p) and studying the closedness *in measure* of quantum and classical evolution. This is precisely the content of my research with prof. A. Ponno and L. Zanelli, but it is another story.

References

[1] A. Arai, “Analysis on Fock spaces and mathematical theory of Quantum Fields”. World Scientific, 2017.

[2] O. Babelon, D. Bernard, and M. Talon, “Introduction to Classical Integrable Systems”. Cambridge University Press, 2003.

[3] A. Bouzouina and D. Robert, *Uniform Semiclassical Estimates for the Propagation of Quantum Observables.* Duke Mathematical Journal 111/2 (2002), 223–252.

[4] E. Fermi, “Notes on Quantum Mechanics”. University of Chicago Press, 1961.

[5] E. Fermi, J. Pasta, and S. Ulam, “Studies on Non Linear Problems”. In: Los-Alamos International Report Document LA-1940, 1955.

[6] P.G. Kevrekidis, “The Discrete Non-linear Schrödinger Equation”. Springer, 2009.

[7] L.D. Landau and E.M. Lifshitz, “Quantum Mechanics: non relativistic theory”. Pergamon Press, 1965.

[8] V. Moretti, “Spectral Theory and Quantum Mechanics”. Springer, 2017.

[9] M. Zworski, “Semiclassical Analysis”. AMS, 2012.

An introduction to sheets of conjugacy classes in reductive groups

Filippo Ambrosio [(*)]

Abstract. Linear algebraic groups arose as a generalization of Lie groups, introduced in the late 1800s to study continuous symmetries of differential equations. The development of the modern theory of algebraic groups with the use of algebraic geometry is mostly due to Borel: in the 1950s, his work led to the definition of Chevalley groups, an important family of finite simple groups. This suggests that algebraic groups can be approached from different perspectives (Group Theory, Algebraic Geometry, Combinatorics) and have applications in several directions (Invariant Theory, Physics). In the first part of the talk we will introduce basic notions and examples of linear algebraic groups. The last part of the seminar aims at describing some of the geometric structure of these groups.

1 Essential facts from Algebraic Geometry

In this preliminary section we recollect some facts to be found for example in [Spr98, Chapters 1,5].

Let $\mathbb{C}[X_1, \dots, X_n]$ be the ring of polynomials with complex coefficients in n indeterminates. Denote by $\mathbb{A}^n(\mathbb{C})$ the n-dimensional complex affine space.

Definition 1.1 Let $S \subset \mathbb{C}[X_1, \dots, X_n]$. The *vanishing locus* of S is $V(S) := \{x \in \mathbb{A}^n(\mathbb{C}) \mid f(x) = 0 \text{ for all } f \in S\}$.

It is easy to show that if $S \subset \mathbb{C}[X_1, \dots, X_n]$, then $V(S) = V(\langle S \rangle)$ where $\langle S \rangle = \{\sum_{j=1}^m s_j f_j \mid m \in \mathbb{N}, s_j \in S, f_j \in \mathbb{C}[X_1, \dots, X_n]\}$ denotes the ideal in $\mathbb{C}[X_1, \dots, X_n]$ generated by S.

Definition 1.2 $X \subset \mathbb{A}^n(\mathbb{C})$ is an *affine algebraic set* if it is the vanishing locus of some ideal $I \triangleleft \mathbb{C}[X_1, \dots, X_n]$. If $I_X \triangleleft \mathbb{C}[X_1, \dots, X_n]$ is such that $V(I_X) = X$, we call I_X the *defining ideal* of X.

Remark that $\varnothing = V(\{1\})$ and $\mathbb{A}^n(\mathbb{C}) = V(\{0\})$. The class of affine algebraic sets is

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: `ambrosio@math.unipd.it` . Seminar held on 18 December 2019.

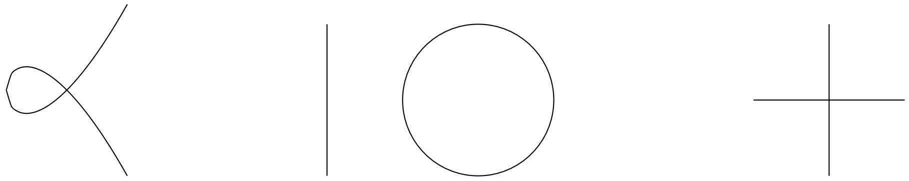
closed by taking arbitrary intersections: if $\{X_j\}_{j\in J}$ is a family of affine algebraic sets with defining ideals $\{I_j\}_{j\in J}$, then $\bigcap_{j\in J} X_j = V(\sum_{j\in J} I_j)$, where $\sum_{j\in J} I_j := \{\sum_{k=1}^{m} f_k \mid m \in \mathbb{N}, f_k \in I_j \text{ for some } j \in J\} \lhd \mathbb{C}[X_1,\dots,X_n]$. The class of affine algebraic sets is closed by taking finite unions: if X, Y are affine algebraic sets with defining ideals I_X and I_Y, then $X \cap Y = V(I_X I_Y)$, where $I_X I_Y := \{fg \mid f \in I_X, g \in I_Y\} \lhd \mathbb{C}[X_1,\dots,X_n]$. Hence, affine algebraic sets fulfill the axioms of the closed subsets of a topology on the space $\mathbb{A}^n(\mathbb{C})$.

Definition 1.3 The *Zariski topology* on $\mathbb{A}^n(\mathbb{C})$ is the topology in which the closed subsets are affine varieties.

Example 1.4

(i) Any point $P \in \mathbb{A}^1(\mathbb{C})$ on the affine complex line is completely determined by its coordinate x_P. It is an affine algebraic set, as $\{P\} = V(X - x_P)$. Similarly, any finite collection of points on the affine complex line is an algebraic set.

(ii) The following are algebraic sets in the complex affine plane $\mathbb{A}^2(\mathbb{C})$ with coordinates X_1, X_2.



(a) $V(X_2^2 - X_1^3 - X_1^2)$ (b) $V((X_2^2 - X_1^2 - 1)(X_1 + 2))$ (c) $V(X_1 X_2)$

Figure 1. Examples of affine algebraic sets in the plane.

Definition 1.5 If $X \subset \mathbb{A}^m(\mathbb{C}), Y \subset \mathbb{A}^n(\mathbb{C})$ are affine algebraic sets, an *(algebraic) morphism* $\phi \colon X \to Y$ is the restriction of a map

$$\begin{aligned} \tilde{\phi} \colon \mathbb{A}^m(\mathbb{C}) &\to \mathbb{A}^n(\mathbb{C}) \\ (x_1,\dots,x_m) &\mapsto \begin{pmatrix} \phi_1(x_1,\dots,x_m) \\ \vdots \\ \phi_n(x_1,\dots,x_m) \end{pmatrix} \end{aligned}$$

with $\phi_i \in \mathbb{C}[X_1,\dots,X_m]$ for all $i = 1,\dots,n$.

Example 1.6 We give two easy examples of algebraic morphisms.

(i) The maps $\alpha, \mu : \mathbb{A}^n(\mathbb{C}) \to \mathbb{A}^1(\mathbb{C})$ defined by $\alpha(x_1,\dots,x_n) = \sum_{i=1}^{n} x_i$ and $\mu(x_1,\dots,x_n) = \prod_{i=1}^{n} x_i$ are morphisms.

(ii) The projection on the i-th component $\pi_i : \mathbb{A}^n(\mathbb{C}) \to \mathbb{A}^1(\mathbb{C})$ defined by $(x_1,\dots,x_n) \mapsto x_i$ is a morphism, for $i \in \{1,\dots,n\}$.

Definition 1.7 Let Z be a topological space. A subset $Z' \subset Z$ is said to be *irreducible* if for all closed subset $C_1, C_2 \subset Z$ satisfying $Z' = C_1 \cup C_2$, then $C_1 = Z'$ or $C_2 = Z'$.

Notice that irreducibility implies connectedness, since a disconnection of a set is the possibility of expressing such set as the (disjoint) union of two closed subsets.

Example 1.8 The affine space $\mathbb{A}^n(\mathbb{C})$ is irreducible. In Example 1.4 we can see an irreducible curve (the node in Figure 1a), a disconnected (hence reducible) curve (Figure 1b) and a connected but reducible curve (the union of the axes in Figure 1c).

Proposition 1.9 *Let $\phi\colon X \to Y$ be a continuous map between two topological spaces.*

(i) *If $Z \subset X$ is irreducible, then so is $\phi(Z)$.*

(ii) *If $\varnothing \neq A$ is open in the irreducible set Z, then $\overline{A} = Z$ and A is irreducible (hence connected).*

Example 1.10 The set $\mathbb{A}^1(\mathbb{C}) \setminus \{0\}$ is open, since it is the complementary of a point. By Proposition 1.9 it is irreducible (hence connected) and dense in $\mathbb{A}^1(\mathbb{C})$ with respect to the Zariski topology.

Definition 1.11 The *dimension* of an affine algebraic set X is the largest $n \geq 0$ such that there exists a chain $\varnothing \subsetneq X_0 \subsetneq X_1 \subsetneq \cdots \subsetneq X_n \subset X$ of irreducible closed subsets $X_j \subset X$.

Example 1.12 The affine space $\mathbb{A}^n(\mathbb{C})$ has dimension n; points have dimension zero while all algebraic sets in Example 1.4 are 1-dimensional.

Proposition 1.13 *Every affine algebraic set X can be written uniquely as a finite union of maximal irreducible closed subsets $X = X_1 \cup \cdots \cup X_r$ with $X_i \not\subset X_j$ for all $i \neq j$, called* irreducible components *of X.*

Example 1.14 Consider the reducible curve in Figure 1c. It consists of two irreducible components:

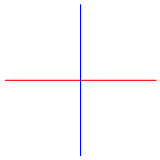


Figure 2. $V(X_1X_2) = V(X_2) \cup V(X_1)$.

For our purposes, we will call *variety* a set which admits a covering by finitely many affine open subsets (in the Zariski topology). For a rigorous definition, see [Spr98, §1.6].

Definition 1.15 A morphism of varieties $\phi\colon X \to Y$ is said to be *dominant* if $\overline{\phi(X)} = Y$.

Proposition 1.16 *Let $\phi\colon X \to Y$ be a dominant morphism of irreducible varieties. Then there exists a non-empty open subset $U \subseteq Y$ such that $\dim \phi^{-1}(y) = \dim X - \dim Y$ for all $y \in U$.*

Remark 1.17 The product of two varieties is endowed with the structure of a variety. We alert the reader that the Zariski topology on the product is not given by the product of the Zariski topologies on the varieties, see [Spr98, §1.5].

2 Linear algebraic groups

Definition 2.1 A *linear algebraic group* is an affine algebraic set G with a group structure such that:

$$\mu\colon G \times G \to G \quad (x, y) \mapsto xy \qquad \text{and} \qquad \iota\colon G \to G \quad x \mapsto x^{-1}$$

are algebraic morphisms.

Let $n \in \mathbb{N}\backslash\{0\}$ and denote with $(X_{ij})_{i,j}$ the n^2 indeterminates parametrized by ordered pairs (i, j) with $1 \leq i, j, \leq n$. With these coordinates, we can identify the vector space $M_n(\mathbb{C})$ of $n \times n$ square matrices with complex coefficients with $\mathbb{A}^{n^2}(\mathbb{C})$.

Definition 2.2 The *general linear group* of order n over $\mathbb{C}$ is $\mathrm{GL}_n(\mathbb{C})$, consisting of all invertible $n \times n$ square matrices with complex coefficients.

We briefly explain the reason why $\mathrm{GL}_n(\mathbb{C})$ is a linear algebraic group. As an affine algebraic set, we have:

$$\mathrm{GL}_n(\mathbb{C}) = \{x = ((x_{ij})_{i,j}, t) \in \mathbb{A}^{n^2+1}(\mathbb{C}) \mid \det((x_{ij})_{i,j})t - 1 = 0\}.$$

If $A, B \in \mathrm{GL}_n(\mathbb{C})$, then each entry of the product AB can be expressed as a polynomial functions in the entries of A and B. Similarly, the entries of A^{-1} can be expressed by means of Laplace's rule as polynomial functions in the entries of A and of $(\det A)^{-1}$.

The group $G = \mathrm{GL}_n(\mathbb{C})$ is connected and irreducible. Observe that $\mathcal{N} = \{A \in M_n(\mathbb{C}) \mid \det A = 0\}$ is a closed subset of $M_n(\mathbb{C})$ (called *nilpotent cone*). Hence $G = M_n(\mathbb{C}) \setminus \mathcal{N}$ is a non-empty open subset of the irreducible set $M_n(\mathbb{C}) \simeq \mathbb{A}^{n^2}(\mathbb{C})$, so it is dense and irreducible by Proposition 1.9.

Actually, the general linear group is a "prototype object", in the following sense:

Theorem 2.3 *G is a linear algebraic group if and only if it is isomorphic to a (Zariski) closed subgroup of $\mathrm{GL}_n(\mathbb{C})$ for some $n \in \mathbb{N}$.*

Proof. If G is a closed subset of $\mathrm{GL}_n(\mathbb{C})$ then it is an affine algebraic set and its group structure is inherited by the one of $\mathrm{GL}_n(\mathbb{C})$, hence it is a linear algebraic group. The other implication is not trivial, see [Spr98, Theorem 2.3.7]. □

Example 2.4

(i) The multiplicative group $\mathrm{GL}_1(\mathbb{C})$ of invertible complex numbers $\mathbb{C}^*$.

(ii) Consider the additive group $G = (\mathbb{C}, +)$. This is a linear algebraic group as the maps:

$$\begin{aligned} \mu\colon G \times G &\to G & \iota : G &\to G \\ (x_1, x_2) &\mapsto x_1 + x_2 & x &\mapsto -x \end{aligned}$$

are clearly morphisms of algebraic sets.

This suggests that G can be realized also as a group of matrices. This is done as follows:

$$G \to \mathrm{GL}_2(\mathbb{C}) \qquad\qquad z \mapsto \begin{pmatrix} 1 & z \\ 0 & 1 \end{pmatrix}.$$

(iii) Consider the algebraic group homomorphism $\det\colon \mathrm{GL}_n(\mathbb{C}) \twoheadrightarrow \mathbb{C}^*$. Then $\det^{-1}(1) = \mathrm{SL}_n(\mathbb{C}) \leq \mathrm{GL}_n(\mathbb{C})$ is the *special linear group of order n*. By Proposition 1.16, $\mathrm{SL}_n(\mathbb{C})$ has dimension $n^2 - 1$ and it is connected (see [HH03, Proposition 1.10]).

Proposition 2.5 *Let G be a linear algebraic group. Then its connected components coincide with its irreducible component. The connected component containing the unity is a closed normal subgroup of G called the identity component of G and denoted with G°.*

Proof. See [Spr98, Proposition 2.2.1]. □

Example 2.6 We give examples of disconnected algebraic groups.

(i) The *orthogonal group* of order n over $\mathbb{C}$ is $G = \mathrm{O}_n(\mathbb{C}) = \{A \in \mathrm{GL}_n(\mathbb{C}) \mid A^T A = 1\}$. It has two connected (i.e. irreducible) components, $G^\circ = \mathrm{SO}_n(\mathbb{C}) = \{A \in G \mid \det A = 1\}$ and $(-1)G^\circ$.

(ii) Any finite group G is a linear algebraic group. Let $n = |G|$ be the order of G. Fix a base of $\mathbb{C}^n$ of vectors $(e_h)_{h \in G}$. For all $g \in G$, define the linear map σ_g via $\sigma_g(e_h) = e_{gh}$ for all $h \in G$. Then we have the group morphism: $G \to \mathrm{GL}_n(\mathbb{C})$ defined by $g \mapsto \sigma_g$. If $n \geq 1$, then G is totally disconnected with respect to the Zariski topology.

2.1 Jordan decomposition

This part is devoted to one of the founding instruments to study linear algebraic groups.

Definition 2.7 Let $G \leq \mathrm{GL}_n(\mathbb{C})$ and $A \in G$.

(i) A is *semisimple* if it is diagonalizable, i.e. if there exists a basis $\{v_1, \dots, v_n\}$ of $\mathbb{C}^n$ and $\lambda_i \in \mathbb{C}^\times$ such that $Av_i = \lambda_i v_i$ for all $i = 1, \dots, n$;

(ii) A is *unipotent* if $A - 1$ is nilpotent, i.e. if there exists $k \in \mathbb{N}$ such that $(A-1)^k = 0$.

Proposition 2.8 *$A \in G$ admits a unique decomposition such that $A = A_s A_u = A_u A_s$ with $A_s \in G$ semisimple and $A_u \in G$ unipotent.*

Proof. See [Spr98, Corollary 2.4.5]. □

Remark 2.9 1 is the only element which is both semisimple and unipotent.

Example 2.10 In $G = \mathrm{GL}_2(\mathbb{C})$ consider $A = \begin{pmatrix} \alpha & 1 \\ 0 & \alpha \end{pmatrix}$. Then $A = A_s A_u$, with $A_s = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}$ and $A_u = \begin{pmatrix} 1 & \alpha^{-1} \\ 0 & 1 \end{pmatrix}$. Hence A is unipotent if and only if $\alpha = 1$, otherwise it is nor semisimple nor unipotent.

3 Sheets

3.1 Generalities on group actions

Definition 3.1 Let G be a linear algebraic group and X be an affine algebraic set. We say that G *acts (morphically)* on X if there is a morphism of affine algebraic sets:

$$\phi \colon G \times X \to X \qquad\qquad (g, x) \mapsto g \cdot x$$

satisfying:

(i) $g_1 \cdot (g_2 \cdot x) = (g_1 g_2) \cdot x$ for all $g_1, g_2 \in G, x \in X$;

(ii) $1 \cdot x = x$ for all $x \in X$.

We introduce some terminologies regarding group actions.

Definition 3.2 Let G be a group acting on set X, let $x \in X$. The *stabilizer* of x in G is $G_x = \{g \in G \mid g \cdot x = x\}$. The G-*orbit* of x is $G \cdot x = \{g \cdot x \mid g \in G\} \subset X$.

It is well-known that if a group G acts on a set X then X is partitioned into its G-orbits. From now on we deal with a linear algebraic group G acting on an affine algebraic set X.

Definition 3.3 For all $x \in X$ we define the *orbit map*:

$$\phi_x \colon G \to X \qquad\qquad g \mapsto g \cdot x.$$

Remark 3.4 In this case $G_x = \phi_x^{-1}(x)$ is a *closed* (hence linear algebraic) subgroup of G and $\phi_x(G) = G \cdot x$ is a variety. Moreover, if G is connected (i.e. irreducible), the orbit $G \cdot x$ is an irreducible subset of X by Proposition 1.9.

Example 3.5 Let G be a linear algebraic group. The *conjugacy action* of G on itself is: $(g, x) \to g \cdot x = gxg^{-1}$ for all $g, x \in G$. When $G = \mathrm{GL}_n(\mathbb{C})$ and $A \in G$, then $G_A = C_G(A) = \{P \in G \mid PA = AP\}$ is called the *centralizer* of A in G. The G-orbit of a matrix $A \in G$ is $\{PAP^{-1} \mid P \in G\}$ i.e. all matrices in G which are *similar* to A.

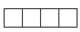
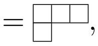
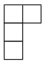
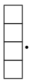
Definition 3.6 Let G act on X and $n \in \mathbb{N}$. The n-th *level set* of X is $X_{(n)} = \{x \in X \mid \dim G \cdot x = n\} = \{x \in X \mid \dim G_x = \dim G - n\}$. A *sheet* of X for the action of G is an irreducible component of $X_{(n)}$ for some $n \in \mathbb{N}$.

3.1.1 Conjugacy classes of unipotent elements in $\mathrm{GL}_n(\mathbb{C})$

Definition 3.7 Let $n \in \mathbb{N} \setminus \{0\}$. A *partition* $[d_1, d_2, \ldots, d_r]$ of n is a sequence of non-decreasing integers $d_1 \geq d_2 \geq \ldots d_r > 0$ such that $n = \sum_{j=1}^{r} d_j$.

Partitions of n are usually denoted with Young diagrams: to the partition $[d_1, d_2, \ldots, d_r]$ of n we associate a diagram where the j-th row consists of d_j squares.

Example 3.8 The partitions of 4 are:

$[4] =$ , $[3,1] =$ , $[2,2] =$, $[2,1,1] =$ , $[1,1,1,1] =$ .

Proposition 3.9 *The unipotent conjugacy classes of* $\mathrm{GL}_n(\mathbb{C})$ *are in one-to-one correspondence with the partitions of* n.

Proof. This follows from the fact that $A \in \mathrm{GL}_n(\mathbb{C})$ is unipotent if and only if its only eigenvalue is 1 and elementary Jordan Theory. □

3.2 Sheets of $\mathrm{GL}_n(\mathbb{C})$ for the conjugacy action

We are interesting in the following problem: consider $X = G = \mathrm{GL}_n(\mathbb{C})$ and describe sheets of G for the conjugation action of G on itself, i.e. irreducible components of all sets $G_{(n)} = \{A \in G \mid \dim C_G(A) = \dim G - n\}$ for $n \in \mathbb{N}$.

Why are we interested in studying sheets rather than conjugacy classes?

(i) Sheets are more treatable, as they finitely many. This is due to the fact that the level sets are finitely many ($G_{(n)} = \varnothing$ for all $n \geq \dim G$) and irreducible components of a variety are finitely many (Proposition 1.13).

(ii) Sheets are varieties, so they can be studied from a geometric point of view.

(iii) Sheets are strongly related to Representation Theory.

We have not introduced enough instruments to give a complete parametrization of sheets in G. We will limit ourselves to give some heuristics on the structure of sheets and then state a theorem describing the behaviour of sheets of G.

We would like to use this method to compute sheets: start from a semisimple conjugacy class in G, see if this is a "deformation" of a unipotent class and compute all others classes which can be obtained as its deformations.

3.2.1 An example: $\mathrm{GL}_4(\mathbb{C})$

We make an example in $n = 4$ to explain the above idea. The empty entries in the matrix conventionally substitute entries which are equal to zero.

Let $A = \begin{pmatrix} \alpha_1 &&& \\ & \alpha_2 && \\ && \alpha_3 & \\ &&& \alpha_4 \end{pmatrix} \in G$ with $\alpha_i \neq \alpha_j$ for $i \neq j$. Then the centralizer is: $C_G(A) = \left\{ \begin{pmatrix} a &&& \\ & b && \\ && c & \\ &&& d \end{pmatrix} \mid a, b, c, d \in \mathbb{C}^* \right\}$ and $\dim C_G(A) = 4$. In particular, $\dim G \cdot A = \dim G - \dim C_G(A) = 16 - 4 = 12$, hence $A \in G_{(12)}$

Now let $\alpha_1 = \alpha_2$, then we obtain $A' = \begin{pmatrix} \alpha_1 &&& \\ & \alpha_1 && \\ && \alpha_3 & \\ &&& \alpha_4 \end{pmatrix}$ with $\alpha_i \neq \alpha_j$ for $i \neq j$. The centralizer is $C_G(A') = \left\{ \begin{pmatrix} a_1 & a_2 && \\ a_3 & a_4 && \\ && c & \\ &&& d \end{pmatrix} \mid c, d \in \mathbb{C}^*, \left(\begin{smallmatrix} a_1 & a_2 \\ a_3 & a_4 \end{smallmatrix}\right) \in \mathrm{GL}_2(\mathbb{C}) \right\}$, with dimension $\dim C_G(A') = 6$, which exceeds 4, but if we allow a unipotent part on the upper-left block, we get $A'' = \begin{pmatrix} \alpha_1 & 1 && \\ & \alpha_1 && \\ && \alpha_3 & \\ &&& \alpha_4 \end{pmatrix}$. One simply computes $C_G(A'') = \left\{ \begin{pmatrix} a & b && \\ & a && \\ && c & \\ &&& d \end{pmatrix} \mid a, c, d \in \mathbb{C}^*, b \in \mathbb{C} \right\}$, which satisfies $\dim C_G(A'') = 4$. Actually, one can prove that A'' and A are contained in the same sheet.

The next natural question is: are A'' and A obtainable as "deformations" of the same unipotent element? The answer is positive, as one can check in the following Table: taking $\alpha_1 = 1$ in the last row, we get the unipotent class corresponding to the partition [4].

degeneration	deformation	centralizer
$\alpha_i \neq \alpha_j$ for $i \neq j$	$\begin{pmatrix} \alpha_1 &&& \\ & \alpha_2 && \\ && \alpha_3 & \\ &&& \alpha_4 \end{pmatrix}$	$\left\{ \begin{pmatrix} a_1 &&& \\ & a_2 && \\ && a_3 & \\ &&& a_4 \end{pmatrix} \mid a_i \in \mathbb{C} \right\} \cap G$
$\alpha_4 \neq \alpha_1 = \alpha_2 \neq \alpha_3 \neq \alpha_4$	$\begin{pmatrix} \alpha_1 & 1 && \\ & \alpha_1 && \\ && \alpha_3 & \\ &&& \alpha_4 \end{pmatrix}$	$\left\{ \begin{pmatrix} a_1 & a_2 && \\ & a_1 && \\ && a_3 & \\ &&& a_4 \end{pmatrix} \mid a_i \in \mathbb{C} \right\} \cap G$
$\alpha_1 = \alpha_2 = \alpha_3$	$\begin{pmatrix} \alpha_1 & 1 && \\ & \alpha_1 & 1 & \\ && \alpha_1 & \\ &&& \alpha_4 \end{pmatrix}$	$\left\{ \begin{pmatrix} a_1 & a_2 & a_3 & \\ & a_1 & a_2 & \\ && a_1 & \\ &&& a_4 \end{pmatrix} \mid a_i \in \mathbb{C} \right\} \cap G$
$\alpha_1 = \alpha_2 \neq \alpha_3 = \alpha_4$	$\begin{pmatrix} \alpha_1 & 1 && \\ & \alpha_1 && \\ && \alpha_3 & 1 \\ &&& \alpha_3 \end{pmatrix}$	$\left\{ \begin{pmatrix} a_1 & a_2 && \\ & a_1 && \\ && a_3 & a_4 \\ &&& a_3 \end{pmatrix} \mid a_i \in \mathbb{C} \right\} \cap G$
$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$	$\begin{pmatrix} \alpha_1 & 1 && \\ & \alpha_1 & 1 & \\ && \alpha_1 & 1 \\ &&& \alpha_1 \end{pmatrix}$	$\left\{ \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ & a_1 & a_2 & a_3 \\ && a_1 & a_2 \\ &&& a_1 \end{pmatrix} \mid a_i \in \mathbb{C} \right\} \cap G$

Table 1. Deformations of the unipotent class of $\mathrm{GL}_4(\mathbb{C})$ corresponding to the partition [4].

One can check that in Table 1 we have listed all conjugacy classes in $G_{(12)}$, we deduce that $G_{(12)}$ has a unique irreducible component, called *regular sheet*$^{(\mathbf{6})}$ (it consists of all conjugacy classes of maximal dimension).

One can carry out the computations for any semisimple class in G: we collect the relevant information in the following Table.

semisimple el.	unipotent el.	partition	$\dim C_G(A)$	$G_{(n)}$
$\begin{pmatrix} \alpha_1 & & & \\ & \alpha_2 & & \\ & & \alpha_3 & \\ & & & \alpha_4 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & & \\ & 1 & 1 & \\ & & 1 & 1 \\ & & & 1 \end{pmatrix}$	$[4]$	4	$G_{(12)}$
$\begin{pmatrix} \alpha_1 & & & \\ & \alpha_1 & & \\ & & \alpha_3 & \\ & & & \alpha_4 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & & \\ & 1 & 1 & \\ & & 1 & \\ & & & 1 \end{pmatrix}$	$[3,1]$	6	$G_{(10)}$
$\begin{pmatrix} \alpha_1 & & & \\ & \alpha_1 & & \\ & & \alpha_3 & \\ & & & \alpha_3 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & & \\ & 1 & & \\ & & 1 & 1 \\ & & & 1 \end{pmatrix}$	$[2,2]$	8	$G_{(8)}$
$\begin{pmatrix} \alpha_1 & & & \\ & \alpha_1 & & \\ & & \alpha_1 & \\ & & & \alpha_4 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}$	$[2,1,1]$	10	$G_{(6)}$
$\begin{pmatrix} \alpha_1 & & & \\ & \alpha_1 & & \\ & & \alpha_1 & \\ & & & \alpha_1 \end{pmatrix}$	$\begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}$	$[1,1,1,1]$	16	$G_{(0)}$

Table 2. Sheets of $\mathrm{GL}_4(\mathbb{C})$.

In general, the following result holds.

Theorem 3.10 (Classification of sheets in $\mathrm{GL}_n(\mathbb{C})$) *Let S be a sheet of $G = \mathrm{GL}_n(\mathbb{C})$. Then:*

(i) *S contains a unique unipotent conjugacy class, in particular sheets of G are parametrized by partitions of n;*

(ii) *S is a smooth variety;*

(iii) *G is the disjoint union of its sheets.*

3.2.2 Historical remarks and generalizations

The problem of determining sheets of $\mathrm{GL}_n(\mathbb{C})$ is rather old: main results were collected and formalized by Dixmier, Kostant, Kraft, Lusztig, starting from the 1960s.

The situation in $\mathrm{GL}_n(\mathbb{C})$ is particularly well-behaved, but what is still true for a connected reductive$^{(\mathbf{7})}$ algebraic group G?

The approach in the general case needs quite more work. When working with algebraic groups, a standard technique is to study first an analogue problem in a "linearized context" and then try to adapt the obtained results back to the group. The tangent space to the

[6] This fact generalizes to the action of any algebraic group G on an irreducible variety X: the union of orbits of maximal dimension is a sheet and is denoted X^{reg}.

[7] Reducing to this class of algebraic groups is standard for people working in the field. An algebraic group is *reductive* if its *unipotent radical* (the maximal connected normal subgroup consisting of unipotent elements) is trivial.

group G at the identity 1 is a Lie algebra $\mathfrak{g} = \mathrm{T}_1 G$. There exists a natural action of G on $\mathfrak{g}$, called *adjoint action.* Sheets of $\mathfrak{g}$ for the adjoint action of G were completely understood in [BK79, Bor82].

We notice different behaviours for G connected reductive:

1. There are sheets which do not contain unipotent classes, hence they lose their special role in the parametrization, i.e. these sheets cannot be seen as "deformations" of a particular conjugacy class.

2. In many simple algebraic groups (e.g. $\mathrm{SO}_n(\mathbb{C}), \mathrm{Sp}_{2n}(\mathbb{C}), \dots$) sheets may intersect and may present singularities.

It would be of little use to state the Theorem classifying sheets in any connected reductive group, as it requires the knowledge of more definitions and constructions: we address the interested reader to [CE12].

References

[BK79] W. Borho and H. Kraft, *Über Bahnen und deren Deformationen bei linearen Aktionen reduktiver Gruppen.* Comm. Math. Helv. 54/1 (Dec 1979), 61–104.

[Bor82] W. Borho, *Über Schichten halbeinfacher Lie-Algebren.* Invent. Math. 65 (1981/82), 283–318.

[CE12] G. Carnovale and F. Esposito, *On sheets of conjugacy classes in good characteristic.* Int. Math. Res. Not. IMRN 2012/4 (2012), 810–828.

[HH03] B. Hall and B.C. Hall, "Lie Groups, Lie Algebras, and Representations: An Elementary Introduction". Graduate Texts in Mathematics, Springer, 2003.

[Spr98] T.A. Springer, "Linear algebraic groups". Progress in mathematics. Birkhäuser, 1998.

Stable hypersurfaces in the complex projective space

Alberto Righini [*]

Abstract. The classification of complete oriented stable hypersurfaces in the complex projective space could be an important step for the classification of isoperimetric sets. Indeed, the boundary of an isoperimetric set, if smooth, is a hypersurface with constant mean curvature which is stable for variations fixing the volume. In this talk we give an introductory overview of the problem and present some new results, in particular we will characterize the geodesic spheres as the unique stable connected and complete hypersurfaces subject to a certain bound on the curvatures.

Introduction

The origin of the isoperimetric problem goes back to the legend of Dido and the foundation of Carthage. According to the legend, when Dido landed on the north coast of Africa, she managed to buy a small piece of land, as much as she could enclose with an oxhide. Then she cut the hide to obtain a long string and she faced the problem of enclosing the largest area with the given perimeter (the length of the string), which is exactly what we call the isoperimetric problem. The guessed solution is the circle of circumference L (if L is the given perimeter).

A first formal solution to this problem, although a partial one, is due to Steiner in 1841. He gave five proofs of the isoperimetric theorem, namely that the circle is actually the set with a given perimeter L, which encloses the greatest area, but all his proofs assumed the existence of a solution.

The question in the Euclidean space was solved in 1958 by Ennio De Giorgi, who proved the isoperimetric property of the hypersphere in a space of arbitrary dimension, among the class of sets of finite perimeter.

A classical approach to the problem is by means of the calculus of variations: the idea is to take a curve and wiggle it a little bit, while keeping fixed its length. If the curve is the one with maximal area then we are at an optimum, so an infinitesimal wiggle will cause zero change in the area. In order to find the optimal figure, therefore, we calculate the change in area caused by an infinitesimal wiggle and set this equal to zero.

[*] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: `righini@math.unipd.it`. Seminar held on 15 January 2020.

Now, if we consider a hypersurface M in the Euclidean space of any dimension (say $n+1$) and we denote by $\mathcal{A}(t)$ the area of all compactly supported variations M_t that leave constant the volume enclosed by M, then we are considering a variational isoperimetric problem and it is well known that being a critical point to this problem is equivalent to having constant mean curvature. But when it comes to the second variation of the area, the two problems are no longer equivalent. Here comes into play the definition of stability and the works by Barbosa and Do Carmo (1984) and Barbosa, Do Carmo and Eschenburg (1988), where they studied stability in the Euclidean space (giving a complete classification of stable hypersurfaces) and they approached the study in the more general setting of Riemannian Manifold.

The Isoperimetric Problem

In the plane, we can formulate the isoperimetric problem in the following ways.

1. Consider all *isoperimetric* bounded domains in $\mathbb{R}^2$ (i.e. all open connected sets with fixed given perimeter). Find the domain with the greatest area.
2. Consider all bounded domains with a fixed given area. Find the domain with minimal perimeter.
3. Express the isoperimetric problem as proving the following analytic inequality:

   $$L^2 \geq 4\pi A, \tag{1}$$

   where A is the area of the domain and L is its perimeter.

The answer to this problem will be the disk. Inequality (1) is referred as *isoperimetric inequality* and one can prove that it holds for every bounded domain of $\mathbb{R}^2$, with equality if and only if the domain is a disk.

The isoperimetric inequality in the plane can be proved in many ways. In [9] Hurwitz made use of Fourier Series, in [16] Topping did it using complex variables and, in [7], Howards, Hutchings and Morgan proved a weaker version with a symmetry-and-convexity argument. All these proofs are presented in Chavel [5].

If we consider the problem in $\mathbb{R}^n$ for any $n \geq 2$, the isoperimetric inequality becomes

$$\frac{A(\partial\Omega)}{V(\Omega)^{1-1/n}} \geq \frac{A(\mathbf{S}^{n-1})}{V(\mathbf{B}^n)^{1-1/n}}, \tag{2}$$

where Ω is any bounded domain in $\mathbb{R}^n$, V is the n-measure and A the $(n-1)$-measure in $\mathbb{R}^n$, $\mathbf{B}^n$ and $\mathbf{S}^{n-1}$ are the unit disk and the unit sphere respectively.

As in the case of the plane, one wants to prove that inequality (2) holds for every Ω, with equality if and only if $\Omega = \mathbf{B}^n$.

Remark 1.1 (Spaces with constant sectional curvature) If we want to extend the investigation to model spaces with consant sectional curvature, say κ, things change. Still we have an isoperimetric inequality, meaning that all domains with the same volume have the area of their boudnaries minimized by disks. For $n = 2$ the isoperimetric inequality is

$$L^2 \geq 4\pi A - \kappa A^2, \tag{3}$$

with equality if and only if the domain in question is a disk.

Stability

A useful way to approach the isoperimetric problem can be by taking variations of the boundary of a domain (in the appropriate class of domains) and trying to get some necessary condition for it to be isoperimetric, by imposing the minimality.

Now we will consider the Euclidean space $\mathbb{R}^{n+1}$. Let M be an orientable, n-dimensional differentiable manifold and let $x : M \to \mathbb{R}^{n+1}$ be an immersion. Let $D \subset M$ be a relatively compact domain with smooth boundary. We will denote by $\mathcal{A}_D(x)$ the n-area of D.

Let $x_t : \bar{D} \to \mathbb{R}^{n+1}$, $t \in (-\varepsilon, \varepsilon)$, $x_0 = x$ be a variation of D and let $\mathcal{A}_D(t) := \mathcal{A}_D(x_t)$, $\mathcal{V}_D(t) := \mathcal{V}_D(x_t)$. We are interested in *volume-preserving*, *normal* variations, namely variations for which $\mathcal{V}_D(t) = \mathcal{V}_D(0)$ for all $t \in (-\varepsilon, \varepsilon)$ and for which the variation vector is proportional to the outward normal N along x. We also want variations that fix the boundary, namely such that $x_t(p) = x_0(p)$ for all $p \in \partial D$ and $t \in (-\varepsilon, \varepsilon)$.

The first important result is obtained by studying the first variation of the area $\mathcal{A}'$ and it's the following.

Proposition 1.2 *Let $x : M \to \mathbb{R}^{n+1}$. The following statements are equivalent:*

1. *x has constant mean curvature H_0.*
2. *For each relatively compact domain $D \subset M$ with smooth boundary, and each volume-preserving variation $x_t : \bar{D} \to \mathbb{R}^{n+1}$ that fixes the boundary, $\mathcal{A}'_D(0) = 0$.*

The mean curvature of a domain is defined as the trace *of the second fundamental form of the domain.*

This result can be found in [3] and it characterizes the extremals for the variational isoperimetric problem as the domains with constant mean curvature.

If we want to extend the study up to the second order, it is useful to give the following definition of *stability.*

Definition 1.3 Let $x : M \to \mathbb{R}^{n+1}$ have constant mean curvcature and let $D \subset M$ be a relatively compact domain with smooth boundary. We say that D is *stable* if $\mathcal{A}''_D(0) \geq 0$ for every volume-preserving variation that fixes the ∂D. If every such D is stable, we say that the immersion is stable.

If we consider the variational isoperimetric problemd described in 2. of Proposition 1.2, clearly the boundary of an *isoperimetric set E* is a critical point of such problem, hence it's a hypersurface with constant mean curvature.

Morover, it is a minimum for the area $\mathcal{A}(t)$ for all compactly supported volume-preserving variations x_t.Thus, by the definition of stability, ∂E must be *stable* under such variations.

This points out the importance that a complete characterization of stable, constant mean curvature hypersurfaces would have for the study of the isoperimetric problem in a suitable class of domains. For the Euclidean space, this characterization is done by Barbosa and Do Carmo in [3]. The result is the following.

Theorem 1.4 (Barbosa, do Carmo) *Let M be a compact, orientable, n-dimensional manifold and let $x : M \to \mathbb{R}^{n+1}$ be an immersion with constant mean curvature. Then x is stable if and only if $x(M) \subset \mathbb{R}^{n+1}$ is a round sphere $S^n \subset \mathbb{R}^{n+1}$.*

A consequence of this theorem is that in $\mathbb{R}^{n+1}$, even if we request our manifold M to be immersed, from the stability hypothesis follows that the immersion $x : M \to \mathbb{R}^{n+1}$ is actually an embedding (it cannot self-intersect). This is interesting in view of the fact that there are many examples of compact nonspherical hypersurfaces with constant mean curvature in $\mathbb{R}^{n+1}$. A famous example due to Wente [17] for $n = 2$, but others can be found in the works of Abresch [1] also for $n = 2$, and Hsiang, Teng and Yu [8] for $n > 2$. In view of this last result, all these hypersurfaces are not stable and, in fact, they cannot be the boundary of isoperimetric sets.

These results can be extended to the more general case of Riemannian Manifolds with *constant sectional curvature c*. In [4], Barbosa, do Carmo and Eschenburg proved the following.

Theorem 1.5 (Barbosa, Do Carmo, Eschenburg) *Assume that M is a compact manifold of dimension n without boundary and that $x : M \to \bar{M}^{n+1}(c)$ is an immersion with constant mean curvature, with $\bar{M}^{n+1}(c)$ a Riemannian manifold with constant sectional curvature c. Then x is stable if and only if $x(M) \subset \bar{M}^{n+1}(c)$ is a geodesic sphere.*

For a general Riemannian manifold with no assuption on its sectional curvarure, the situation is different. Not all the (geodesic) spheres are stable and not all stable hypersurfaces are spheres. For example, in [4] Barbosa, Do Carmo and Eschenburg prove the following result about stability of tubes and spheres in projective spaces.

Let $\bar{M} = P^{r-1}\mathbb{K}$ be the projective space over the field $\mathbb{K}$, with metric of diameter $\frac{\pi}{2}$ and curvature between 1 and 4, where $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$.

For $q < r$ let $U_\rho(P^{q-1}\mathbb{K})$ be the tubular neighborhood of radius ρ around the totally geodesic subspace $P^{q-1}\mathbb{K}$ of $P^{r-1}\mathbb{K}$, and put $T_\rho(q) = \partial U_\rho(P^{q-1}\mathbb{K})$. Note that $T_\rho(q)$ is congruent to $T_{\frac{\pi}{2}-\rho}(p)$ if $p = r - q$ and that $T_\rho(1)$ is the geodesic sphere of radius ρ. Set $d = \dim_{\mathbb{R}} \mathbb{K}$ and assume that r is even if $\mathbb{K} = \mathbb{R}$ (for orientability reasons).

Theorem 1.6 (Barbosa, Do Carmo, Eschenburg) *For $2 \leq q \leq r - 2$, $T_\rho(q)$ is stable in $P^{q-1}\mathbb{K}$ if and only if*

$$\frac{pd-1}{qd+1} \leq \tan^2 \rho \leq \frac{pd+1}{qd-1}.$$

For $q = 1$ ($q = r - 1$), the lower (upper) bound is not present: a sphere of radius ρ is *stable if and only if* $\tan^2\rho \leq \frac{(r-1)d+1}{d-1}$.

Regarding the isoperimetric problem, we have seen, with Theorem 1.4 and Theorem 1.5, that in the Euclidean case and for a Riemannian manifold with constant sectional curvaure, isoperimetric variational problems are equivalent to stability of a hypersurface.

Remark 1.7 In general, if we are not in the Euclidean case and we are not taking specific geometric assumptions, stability is only a necessary condition for a surface to be the boundary of an isoperimetric set.

This fact can be seen, for example, by considering the classical Schwarz minimal P-surface of genus three in the cubic three-torus, which is a *stable constant mean curvature hypersurface*, but the domain enclosed by it is *not* a solution of the isoperimetric problem (see [13] for details).

Another remarkable result is due to Ritoré and Ros. In [11] they give a complete solution of the stability problem in the 3-dimensional projective space. Their result reads as follows.

Theorem 1.8 (Ritoré, Ros) *Let* $x : M \to \mathbf{R}P^3$ *be a complete orientable CMC surface immersed into the real projective space. If the immersion is stable, then either*

(i) *M is a compact surface with* $\text{genus}(M) = 0$ *and x is an embedded geodesic sphere or a twofold covering of a totally geodesic projective plane, or*

(ii) *M is a compact surface of genus* 1 *and x is an embedded flat tube of radius r about a geodesic, with* $\pi/6 \leq r \leq \pi/3$.

Stable Hypersurfaces in the Complex Projective Space

Now we look at the Complex Projective Space $\mathbf{C}P^n$ and we characterize the geodesic sphere with radius $\tan^2 r = 2n+1$ as the unique stable connected and complete hypersurface subject to a bound either on the characteristic curvature or on the restriction of the second fundamental form to the complex tangent space.

The n-dimensional complex projective space is the quotient of the unit sphere $\mathbf{S}^{2n+1} = \{z \in \mathbf{C}^{n+1} : |z| = 1\}$ by the Hopf-action of $\mathbf{S}^1$, $(\mathrm{e}^{i\vartheta}, z) \mapsto \mathrm{e}^{i\vartheta} z$. We denote by $[z]$ the equivalence class of $z \in \mathbf{S}^{2n+1}$. The tangent space of $\mathbf{C}P^n$ at the point $[z]$ is

$$T_{[z]}\mathbf{C}P^n = \{w \in \mathbf{C}^{n+1} : z \cdot \bar{w} = 0\},$$

where $z \cdot \bar{w} = z_1\bar{w}_1 + \ldots + z_{n+1}\bar{w}_{n+1}$ is the standard Hermitian product of $\mathbf{C}^{n+1}$. The complex structure on $T_{[z]}\mathbf{C}P^n$ is given by $Jw = iw$, the standard multiplication by i of $w \in T_{[z]}\mathbf{C}P^n \subset \mathbf{C}^{n+1}$.

The metric $\langle \zeta, w\rangle_{FS} = \mathrm{Re}(\zeta \cdot \bar{w})$, with $\zeta, w \in T_{[z]}\mathbf{C}P^n$, is the *Fubini-Study* metric of $\mathbf{C}P^n$, that makes the complex projective space a Riemannian manifold. The induced distance function $d : \mathbf{C}P^n \times \mathbf{C}P^n \to [0, \pi/2]$ is $d([z],[w]) = \arccos|z \cdot \bar{w}|$.

The characteristic curvature κ of a hypersurface $\Sigma \subset \mathbf{C}P^n$ is the curvature in direction JN, where N is the normal to Σ and J is the complex structure of $\mathbf{C}P^n$, i.e., $\kappa = h(JN, JN)$ where h is the second fundamental form of Σ.

For any fixed $[w] \in \mathbf{C}P^n$ and $0 < r < \pi/2$, the geodesic sphere centered at $[w]$ with radius r is

$$\Sigma_r = \big\{[z] \in \mathbf{C}P^n : |z \cdot \bar{w}| = \cos r\big\}.$$

The curvatures of Σ_r are well-known, see e.g. [6, Example 1 page 493]. Letting $t = \tan r$, they are

$$\begin{aligned} &\lambda = \cot r = \frac{1}{t}, \quad \text{with multiplicity } 2(n-1), \\ &\kappa = 2\cot(2r) = \frac{1}{t} - t, \quad \text{the characteristic curvature.} \end{aligned} \tag{4}$$

These two curvatures are constant and distinct for each value of $t > 0$. In [15], Takagi proved that this property characterizes the sphere.

We now discuss tubes around $\mathbf{C}P^k$. For $k = 1, \dots, n-1$, the natural inclusion $\mathbf{S}^{2k+1} = \{z \in \mathbf{S}^{2n+1} : z_{k+2} = \ldots = z_{n+1} = 0\} \subset \mathbf{S}^{2n+1}$ induces the inclusion $\mathbf{C}P^k \subset \mathbf{C}P^n$. For $0 < r < \pi/2$, we define the tube

$$\begin{aligned} T_r^k &= \{[z] \in \mathbf{C}P^n : \operatorname{dist}([z], \mathbf{C}P^k) = r\} \\ &= \{[z] = [(z', z'')] \in \mathbf{C}P^n : |z| = 1, z' \in \mathbf{C}^{k+1}, |z'| = \cos r\}. \end{aligned}$$

The curvatures of T_r^k are computed in [6]. Letting $t = \tan r$, they are

$$\begin{aligned} &\lambda_1 = \cot\Big(r - \frac{\pi}{2}\Big) = -t, \quad \text{with multiplicity } 2k, \\ &\lambda_2 = \cot r = \frac{1}{t}, \quad \text{with multiplicity } 2\ell = 2(n-1-k), \\ &\kappa = 2\cot(2r) = \frac{1}{t} - t, \quad \text{the characteristic curvature.} \end{aligned} \tag{5}$$

These three curvatures are constant and distinct for each value of $t > 0$. In particular, T_r^k has constant mean curvature. For $r + s = \pi/2$ and $k + \ell = n - 1$ the hypersurfaces T_r^k and T_s^ℓ are congruent.

Theorem 1.6 gives us some stability intervals for the radius of a geodesic sphere and a geodesic tube around $\mathbf{C}P^k \subset \mathbf{C}P^n$.

With the next result we move forward to a characterization of stable hypersurfaces in $\mathbf{C}P^n$, by studing the converse problem: what geometric properties can be deduced by a stability assumption? It turns out that, if we add a constraint on the characterisctic curvature, we get that a stable hypersurface must be a geodesic sphere, precisely with the radius that bounds the stability interval of Theorem 1.6.

For fixed $H \in \mathbb{R}$ and $n \in \mathbb{N}$, let $p(\cdot; H, n)$ be the quadratic polynomial of the real variable $t \in \mathbb{R}$

$$p(t; H, n) = (2n+1)t^2 - 2Ht - H^2 - 4(n^2 - 1). \tag{6}$$

The result is the following.

Theorem 1.9 (Battaglia, Monti, R.) *Let $\Sigma \subset \mathbf{C}P^n$, $n \geq 2$, be a complete connected stable hypersurface with constant $H = \mathrm{tr}(h)$. If the characteristic curvature κ of Σ satisfies $p(\kappa; H, n) \geq 0$ then Σ is a geodesic sphere of radius $r > 0$ with $\tan^2 r = 2n + 1$.*

The details of every step of the proof can be found in [2]. Here we will give just an outline.

The main step is to prove the following inequality, wchich is implied by stability. Let $\nabla^\top$ be the Levi-Civita connection of $\mathbf{C}P^n$ and consider the covariant derivative $\nabla^\top_{JN} N \in T\Sigma$ of the normal N to Σ. We denote by $h_N \in \mathbf{C}T\Sigma$ the projection of $\nabla^\top_{JN} N$ onto $\mathbf{C}T\Sigma$, where $\mathbf{C}T\Sigma$ is the complex tangent space to Σ. By $|h|^2$ we denote the squared norm of h.

Theorem 1.10 *Let $\Sigma \subset \mathbf{C}P^n$, $n \geq 2$, be a complete stable hypersurface with constant $H = \mathrm{tr}(h)$. Then we have*

$$\int_\Sigma \Big\{ |h|^2 + \frac{(H+\kappa)^2 + |h_N|^2}{2(n^2-1)} - \frac{H^2}{n-1} - 2n \Big\} d\mu \leq 0, \tag{7}$$

where μ is the Riemannian hypersurface measure.

The method for obtaining formula (7) starts from an idea contained in the proof of Theorem 1.4. The first step is the isometric embedding of $\mathbf{C}P^n$ into H^{n+1}, the space of $(n+1) \times (n+1)$ Hermitian matrices, see [12] and [14]. Once the hypersurface Σ is embedded in H^{n+1}, we can consider the *position matrix* $A \in \Sigma$ and compute its tangential Laplacian ΔA.

For any fixed $V \in H^{n+1}$, the function $u_V = \langle \Delta A, V \rangle$ has zero mean and we can compute the trace of the quadratic form Q_Σ on H^{n+1} defined by $Q_\Sigma(V) = \mathcal{A}''(u_V)$. If Σ is stable, this trace is nonnegative and this fact is precisely inequality (7).

The trace of the quadratic form Q_Σ is

$$\mathrm{tr}(Q_\Sigma) = 4 \int_\Sigma \Big\{ 2(n+1)H^2 + 2(n^2-1)\big(2n - |h|^2\big) - (H+\kappa)^2 - |h_N|^2 \Big\} d\mu. \tag{8}$$

By testing this formula on geodesic spheres, with a simple computation we obtain

Lemma 1.11 *For the sphere $\Sigma_r \subset \mathbf{C}P^n$ we have $\mathrm{tr}(Q_{\Sigma_r}) \geq 0$ if and only if $\tan^2 r \leq 2n+1$. The trace is zero if and only if $\tan^2 r = 2n + 1$.*

Proof of Theorem 1.9. We denote by $\widehat{h}$ the restriction of the second fundamental form h of Σ to the complex tangent space $\mathbf{C}T\Sigma$ and by $\widehat{H}$ the trace of $\widehat{h}$. At any point of Σ, we have the identities

$$H = \widehat{H} + \kappa \quad \text{and} \quad |h|^2 = |\widehat{h}|^2 + 2|h_N|^2 + \kappa^2,$$

and the inequalities

$$|h|^2 \geq |\widehat{h}|^2 + \kappa^2 \quad \text{and} \quad |\widehat{h}|^2 \geq \frac{\widehat{H}^2}{2(n-1)} = \frac{(H-\kappa)^2}{2(n-1)}. \tag{9}$$

Inserting these inequalities and $|h_N| \geq 0$ into (8) we obtain

$$\begin{aligned} \operatorname{tr}(Q_\Sigma) &\leq 4\int_\Sigma \Big\{2(n+1)H^2 + 2(n^2-1)\Big(2n - \kappa^2 - \frac{(H-\kappa)^2}{2(n-1)}\Big) - (H+\kappa)^2\Big\}d\mu \\ &= -4n\int_\Sigma p(\kappa; H, n)\, d\mu, \end{aligned} \tag{10}$$

where $p(\cdot; H, n)$ is the polynomial in (6). By our assumption $p(\kappa; H, n) \geq 0$ on Σ, we deduce that $\operatorname{tr}(Q_\Sigma) \leq 0$. On the other hand, the stability of Σ implies that $\operatorname{tr}(Q_\Sigma) \geq 0$. We deduce that $\operatorname{tr}(Q_\Sigma) = 0$ and that we have equality in (10). In turn, the equality in (10) implies that $p(\kappa; H, n) = 0$, that

$$|h|^2 = |\widehat{h}|^2 + \kappa^2 \quad \text{and} \quad |\widehat{h}|^2 = \frac{\widehat{H}^2}{2(n-1)}, \tag{11}$$

and also that $h_N = 0$ on Σ.

The equation $h_N = 0$ means that JN is an eigenvector of h. By Maeda's theorem [10], this implies that the characteristic curvature κ is constant. This also simply follows from the fact that κ is one of the roots of $p(\kappa; H, n) = 0$. Here we use the fact that Σ is connected.

The identity in the right-hand side of (11) implies that Σ is umbilical in $\mathbf{C}T\Sigma$, i.e., each unit vector in $\mathbf{C}T\Sigma$ is an eigenvector of h with eigenvalue $\lambda = \widehat{H}/2(n-1)$. Moreover, λ is constant on Σ, because $\widehat{H} = H - \kappa$ is constant.

The two constants κ and λ are different, because in $\mathbf{C}P^n$ there are no totally umbilical hypersurfaces. By Takagi's theorem (see [15]) Σ is a geodesic sphere: up to a suitable choice of the center of the sphere, we have $\Sigma = \Sigma_r$ for some $r \in (0, \pi/2)$. By Lemma 1.11 the equation $\operatorname{tr}(Q_{\Sigma_r}) = 0$ implies that $\tan^2 r = 2n + 1$.

□

References

[1] Abresch, U., *Constant mean curvature tori in terms of elliptic functions.* Journal für die Reine und Angewandte Mathematik 374 (1987), 169–192.

[2] Battaglia E., Monti R. and Righini A., *Stable hypersurfaces in the complex projective space.* Annali di Matematica 199, 231–251 (2020). https://doi.org/10.1007/s10231-019-00875-4.

[3] J. Lucas Barbosa and Manfredo do Carmo, *Stability of hypersurfaces with constant mean curvature.* Math. Z. 185 (1984), no. 3, 339–353. MR 731682.

[4] J. Lucas Barbosa, Manfredo do Carmo, and Jost Eschenburg, *Stability of hypersurfaces of constant mean curvature in Riemannian manifolds.* Math. Z. 197 (1988), no. 1, 123–138. MR 917854.

[5] I. Chavel,, "Isoperimetric inequalities. Differential geometric and analytic perspectives". Cambridge Tracts in Mathematics 145. Cambridge: Cambridge University Press. xii, 268 p. , 2001.

[6] Thomas E. Cecil and Patrick J. Ryan, *Focal sets and real hypersurfaces in complex pro- jective space.* Trans. Amer. Math. Soc. 269 (1982), no. 2, 481–499. MR 637703.

[7] Howards H., Hutchings M., Morgan F., *The isoperimetric problem on surfaces.* Amer. Math. Monthly 106 (1999), 430–439.

[8] Hsiang, W.Y., Teng, Z.H., Yu, W., *New examples of constant mean curvature immersions of* $(2k-1)$*-spheres into Euclidean* $2k$*-space.* Ann. Math. 117 (1983), 609–625.

[9] Hurwitz A., *Sur le problème des isopérimètres.* C. R. Acad. Sci. Paris 132 (1901), 401–403.

[10] Hurwitz A., *On real hypersurfaces of a complex projective space.* J. Math. Soc. Japan 28 (1976), no. 3, 529–540. MR 0407772.

[11] M. Ritoré, A. Ros, *Stable constant mean curvature tori and the isoperimetric problem in three space forms.* Commentarii Mathematici Helvetici 67 (1992), 293–305.

[12] Antonio Ros, *Spectral geometry of CR-minimal submanifolds in the complex projective space.* Kodai Math. J. 6 (1983), no. 1, 88–99. MR 698330.

[13] M. Ross, "Stability properties of complete two–dimensional minimal surfaces in Euclidean space". PhD Thesis, Berkeley, 1989.

[14] Shin-Sheng Tai,, *Minimum imbeddings of compact symmetric spaces of rank one.* J. Differential Geometry 2 (1968), 55–66. MR 0231395.

[15] Ryoichi Takagi,, *Real hypersurfaces in a complex projective space with constant principal curvatures.* J. Math. Soc. Japan 27 (1975), 43–53. MR 0355906.

[16] Topping P., *The optimal constant in Wente's* $L^{\circ\circ}$ *estimate.* Comment. Mat. Helv. 139 (1997), 316–328.

[17] Wente H., *Counter-example to the Hopf conjecture.* Pac. J. Math. 121 (1986), 193–244.

A random journey among stochastic processes

Samuele Stivanello [*]

Abstract. This seminar will cover a broad selection of topics, ranging from the basic definition of a probability space, passing through some famous results like the Law of Large Numbers and the Central Limit Theorem, and ending with the notion of convergence of stochastic processes. In order to fulfill this intent, and in the spirit to be accessible to a mathematical audience of non experts, in this introductory talk to the field of stochastic processes I will make extensive use of examples and intuitive definitions. In the very last part of the talk I will mention some results of my research, regarding the convergence of a random walk in random environment.

1 Introduction and Basic Definitions

In this short seminar we will give a basic introduction of random variables. Once given the definition and some basic examples, we want to focus on convergence of sequence of random variables. We will deeply analyze the notion of weak convergence, recalling some famous theorems and specifying the differences with the usual convergence in metric spaces. Then we will move on to the definition of stochastic processes, seen as random variables taking values on a functional space. Again, weak convergence will be our center of interest. Finally we will move to a particular case of processes whose limit has a super diffusive behavior. We will try to say some words on the origin of this anomalous rescaling, and provide the sketch of the proof of the latest results.

Figure 1. Example of Random Variables

[*] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: `samuele.stivanello@gmail.com` . Seminar held on 5 February 2020.

To start with, we give some examples of random variables we encounter in everyday's life. Rolling a dice, tossing a coin, shuffle a deck of cards: randomness is everywhere and a random variable can be seen as a quantity that depends on some randomness. Suppose you are throwing a dart on a professional board. The score you will made is a random variable. If you train yourself, you probably score higher points. In this context probability is seen as a score measure, and it depends on the player.

Figure 2. Example: Different Random Variables on the same space

Let's try to give precise definitions of the quantities that we want to deal with.

Definition 1 A Random Variable is a measurable function from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a metric space $(E, \mathcal{E})$:

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (E, \mathcal{E}) \tag{1}$$

Hence randomness can be seen as the probability measure on the underlying probability space. The underlying probability space is something abstract, and we want to see how to measure the actual distribution of the random variable.
The distribution of X is the probability measure $P^X = \mathbb{P} \circ X^{-1}$ on $(E, \mathcal{E})$, i.e. for every $A \in \mathcal{E}$ we have:

$$P^X(A) = \mathbb{P}\left(X^{-1}(A)\right) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in A\}\right) \tag{2}$$

Let try to gain some more confidence through an easy example. Suppose we are tossing 3 identical coins. The underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ can be identified with the space $\left(\{H, T\}^3 \,,\, \mathcal{P}(\{H, T\}^3) \,,\, \mathbb{P}(A) = \frac{|A|}{|\Omega|}\right)$, where $\mathbb{P}$ is the normalized counting measure, or the uniform discrete measure, on the space of possible outcomes (H =Heads, T =Tails). The random variable X represents the number of Tails after 3 Coin Tosses. Hence the space $(E, \mathcal{E})$ in which X takes values is $(\ \{0, 1, 2, 3\} \,,\, \mathcal{P}(\{0, 1, 2, 3\})\ )$. Suppose that we want to compute the probability of the following event:

$$A = \textit{"after the tossing of 3 coins, we have at least 2 Tails"}$$

The event A can be identified with the subset $\{2, 3\}$. Following (2) we have that

$$\begin{aligned} P^X(A) &= \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in A\}\right) \\ &= \mathbb{P}\left(\{\{T, T, H\}, \{T, H, T\}, \{H, T, T\}, \{T, T, T\}\}\right) = \frac{4}{2^3} = \frac{1}{2} \end{aligned} \tag{3}$$

2 Convergence of Random Variables

Suppose that $X, (X_n)_{n\in\mathbb{N}}$ are random variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P}) = (\,[0,1]\,,\, \mathcal{B}([0,1])\,,\, \lambda)$ and taking values on the metric space $(E, \mathcal{E}) = (\,\mathbb{R}\,,\, \mathcal{B}(\mathbb{R})\,)$. Here λ is the usual Lebesgue measure. The random variable are defined as follows:

$$X_n(\omega) := \omega + \omega^n \quad \forall n \in \mathbb{N} \qquad \text{and} \quad X(\omega) \equiv \omega \tag{4}$$

We ask whether it is true or not that $\lim\limits_{n\to\infty} X_n = X$. How should we interprete the limit? Suppose that we require pointwise convergence. It means that for every $\omega \in \Omega$ the difference $|X_n(\omega) - X(\omega)|$ should be small for n large enough. But pointwise convergence fails at $\omega = 1$. In facts

$$\lim_{n\to\infty} X_n(1) = 2 \neq 1 = X(1) \tag{5}$$

A first solution is passing to almost sure convergence. It means that we require pointwise convergence to hold almost everywhere, i.e. pointwise convergence can fail in a set of zero probability. Here we give some other definitions of convergence that are commonly used in probability theory.

Definition 2 We say that $X_n \xrightarrow{a.s.} X$ **almost surely** if

$$\lim_{n\to\infty} X_n(\omega) = X(\omega) \qquad \forall\, \omega \in A \subseteq \Omega, \quad \text{with} \quad \mathbb{P}(A) = 1 \tag{6}$$

Definition 3 We say that $X_n \xrightarrow{P} X$ **in Probability** if, for every $\varepsilon >= 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\omega \,:\, |\, X_n(\omega) - X(\omega)\,| > \varepsilon\right) = 0 \tag{7}$$

Definition 4 We say that $X_n \xrightarrow{L^p} X$ **in** L^p if $|X_n|, |X|$ are in L^p and

$$\lim_{n\to\infty} \int_\Omega |\, X_n(\omega) - X(\omega)\,|^p \,\mathbb{P}(\omega) d\omega = 0 \tag{8}$$

While all the previous definitions differ from types of convergence seen in elementary Calculus courses, they are nevertheless squarely in the analysis tradition, and they can be thought of as variants of standard pointwise convergence. These types of convergence are natural and useful in probability. However, there is another notion of convergence which is profoundly different from the four we have already seen, known as weak convergence. As its name implies, it is a weak type of convergence. The weaker the requirements for convergence, the easier it is for a sequence of random variables to have a limit. What is unusual about weak convergence, however, is that the actual values of the random variables themselves are not important. It is simply the probabilities that they will assume those values that matter. That is, it is the probability distributions of the random variables that will be converging, and not the values of the random variables themselves.

Let's start by defining weak convergence of probability measures.

Definition 5 Let μ_n, μ be probability measures on $\mathbb{R}^d$. We say $\mu_n \xrightarrow{w} \mu$ weakly if

$$\lim_{n\to\infty} \int g(x) d\mu_n(x) = \int g(x) d\mu(x) \tag{9}$$

for every continuous, bounded function $g : \mathbb{R}^d \to \mathbb{R}$.

Note that convergence of integrals as stated in (9) is actually a definition. If facts, if two different probability measures P, Q are such that $\int g dP = \int g dQ$ for every continuous bounded function g, then $P \equiv Q$.
Now, let $(X_n)_{n\in\mathbb{N}}, X$ be random variables taking values on a metric space $(E, \mathcal{E})$. Suppose for example the space to be $\big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\big)$ with the usual Euclidean distance.

Definition 6 We say $X_n \xrightarrow{D} X$ in distribution if $P^{X_n} \xrightarrow{w} P^X$ weakly, i.e.

$$\lim_{n\to\infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(X)] \tag{10}$$

for every continuous, bounded function $g : \mathbb{R}^d \to \mathbb{R}$.

Note that, if $= \mathbb{R}^d$ and the measure P^X is absolutely continuous with respect to the Lebesgue measure, then

$$\mathbb{E}[g(X)] = \int g(x) dP^X(x) = \int g(x) f(x) dx \tag{11}$$

and the same hold for X_n. Hence the fact that $f_n \to f$ pointwise as $n \to \infty$ is a sufficient condition to have convergence in distribution $X_n \xrightarrow{D} X$.

Remark 1 If the space E is discrete, then $\mathbb{E}[g(X)] = \sum g(x_k) p(x_k)$. Hence, a sufficient condition for $X_n \xrightarrow{D} X$ is that, for every $x_k \in E$, we have that $g_n(x_k) \to g(x_k)$ as $n \to \infty$.

Suppose now we have a sequence of random variables $(\xi_n)_{n\in\mathbb{N}}$, all independent and with the same distribution (i.i.d.) such that

$$\mathbb{E}[\xi_k] = \mu \qquad \mathbb{V}ar[\xi_k] = \sigma^2 \tag{12}$$

Then we have the following limit, that takes the name of Law of Large Numbers (LLN)

Theorem 2.1 (Law of Large Numbers)

$$\frac{\xi_1 + \xi_2 + \cdots + \xi_n}{n} \xrightarrow[n\to\infty]{a.s.} \mu \tag{13}$$

It states that the limit of the empirical mean of n random variables with the same distribution is the theoretical mean. Note that the theorem holds for any distribution. Now we would like to say something more. From the LLN we have that the difference between the sum of n independent identical distributed random variable and n-times their mean converges to 0 almost surely. We want to analyze the fluctuations of this difference, as $n \to \infty$. The following theorem says that the fluctuation behave like a standard normal, with amplitude proportional to $\sqrt{n}$.

Theorem 2.2 (Central Limit Theorem)

$$\frac{\sum_{k=1}^{n} \xi_k \; - \; n\mu}{\sqrt{n\sigma^2}} \xrightarrow[n\to\infty]{D} \mathcal{N}(0,1) \tag{14}$$

3 Stochastic Processes

Our journey moves on to stochastic processes. Let's go straight to the definition.

Definition 7 (Stochastic Process) A Stochastic Process is a family of random variables $\{X_t\}_{t\in J}$ all defined in the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in the same measurable space $(E, \mathcal{E})$.

The former definition is really simple but far from intuitive. The complexity of a stochastic process is shown in Figure 3.

Let's try to formalize.

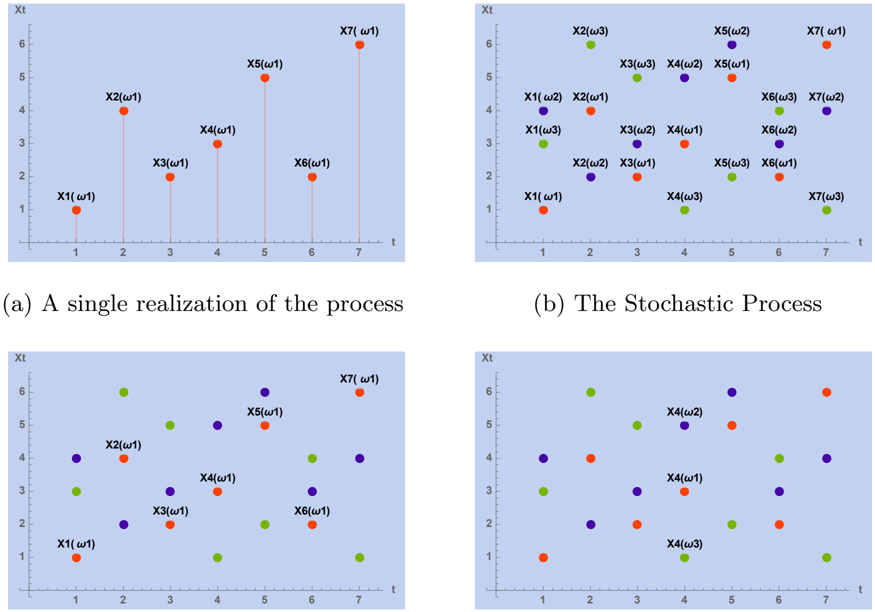
Definition 8 For any given $k \in \mathbb{N}$, $t_1, ..., t_k \in J$, the distributions of the vector $(X_{t_1}, ..., X_{t_k})$ take the name of finite-dimensional distributions of the process X.

$$X = \{X_t\}_{t\in J} \quad \text{where} \quad X_t = (X_t(\omega))_{\omega\in\Omega} \tag{15}$$

Hence a stochastic process $X = \{X_t\}_{t\in J}$ with values in $(E, \mathcal{E})$ can be seen as a unique random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \to \left(E^J, \mathcal{E}^J\right)$ with values in the product space $\left(E^J, \mathcal{E}^J\right)$, known as the space of the trajectories, i.e.

$$X = \{X_t(\omega)\}_{t\in J, \omega\in\Omega} \quad \text{where} \quad X(\omega) = (X_t(\omega))_{t\in J} \tag{16}$$

Definition 9 The law of the process μ^X is the distribution induced by X on $\left(E^J, \mathcal{E}^J\right)$.



(a) A single realization of the process

(b) The Stochastic Process

(c) A trajectory of the process

(d) Distribution of the process at a fixed time

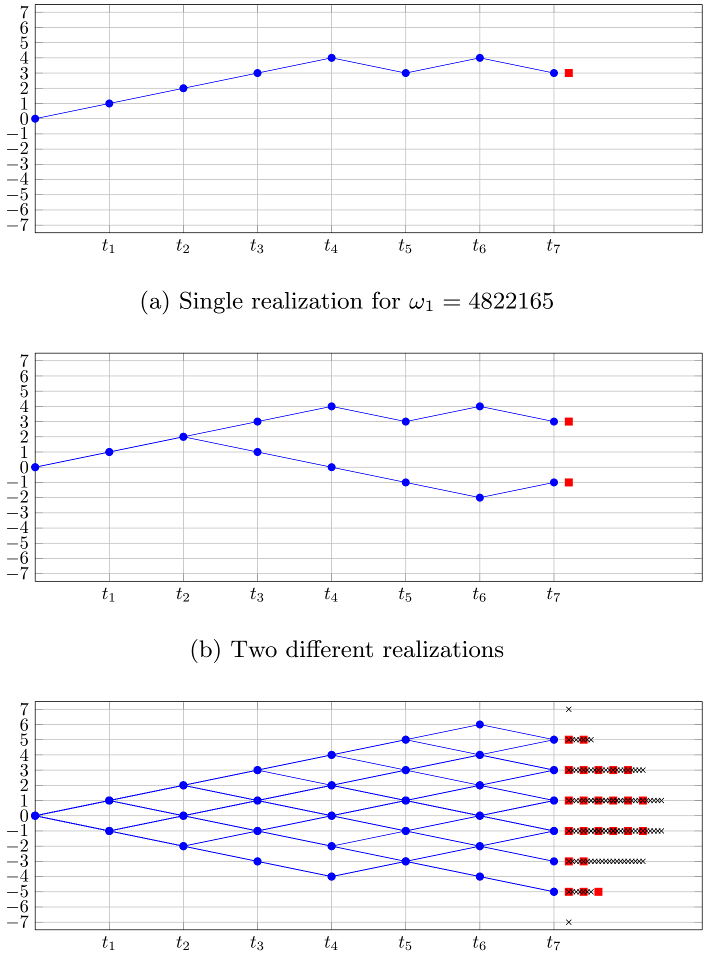
Figure 3. Stochastic Process: a visual example. If we fix ω as in Figure 3a, we can see a single realization of the process. The process itself is the random variable that take values in the space of realizations. Figure 3b shows three different realizations at the same time. A trajectory (Figure 3c) is a single realization of the process. If we fix the time index, the process at a fixed time behaves like a random variable (Figure 3d).

4 Convergence of processes: Random Walk and Brownian Motion

A Random Walk is the basic example of stochastic process. Suppose we have a sequence $(\xi_k)_{k\in\mathbb{N}}$ of random variables defined on the same probability space. A Random Walk $S = \{S_n(\omega)\}_{n\in\mathbb{N},\omega\in\Omega}$ is defined as follows

$$S_n(\omega) := \sum_{k=1}^{n} \xi_k(\omega) \tag{17}$$

If the $(\xi_k)_{k\in\mathbb{N}}$ are i.i.d. such that $\mathbb{P}(\xi_k = 1) = \mathbb{P}(\xi_k = -1) = 1/2$ we obtain a simple symmetric random walk (SSRW). Suppose now we define the following SSRW based on the last seven digits of a phone numbers. If the k-th digit is even, then $\xi_k = 1$, whereas if the k-th digit is odd, then $\xi_k = -1$.



(a) Single realization for $\omega_1 = 4822165$

(b) Two different realizations

(c) The PhDs phone numbers

Figure 4. Figure 4a represents the trajectory of the random walk up to time $t = 7$. This can be seen as a realization of S at $\omega_1 = 4822165$. For $\omega_2 \in \Omega$, Figure 4b depicts a different trajectory: $S(\omega_2)$. Finally, Figure 4c depicts together the n trajetories obtained from the phone numbers of the n PhD students of my office.

The red squares in Figure 4c represents the empirical distribution of the Random Walk defined above a the fixed time $t = 7$. Note that this empirical distribution is close to the theoretical distribution (black crosses). The empirical distribution can be seen as a random sample of the theoretical one.

Now we want to talk about convergence of the random walk as a stochastic process. In order to make the game easier, we define a new random walk in the following way:

$$
\begin{aligned}
&S_0 = 0, \\
&S_n(\omega) := \sum_{k=1}^{n} \xi_k(\omega) \qquad \text{with } \xi_k \sim \mathcal{N}(0,1)
\end{aligned}
\tag{18}
$$

Now the increments ξ_k are gaussian, with zero mean and variance equal to 1. In what follows it's not necessary that the increments have this normal distribution, but this make the computations easier. We want to define the following sequence of processes, for $n \in \mathbb{N}$:

$$
W_t^{(n)}(\omega) := \frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor}(\omega) + \frac{1}{\sqrt{n}}(nt - \lfloor nt \rfloor)\xi_{\lfloor nt \rfloor + 1}(\omega) \qquad t \geq 0 \tag{19}
$$

The first term is basically the random walk S rescaled by n and $\sqrt{n}$ on the x-y axes respectively; whereas the second term accounts for interpolation in order to make trajectories continuous, but it goes to 0 as $n \to \infty$, hence we can neglect it for our purposes. We have that

$$
\begin{aligned}
&\mathbb{E}[W_t^{(n)}] = 0 \\
&\mathbb{V}ar[W_t^{(n)}] = \frac{\lfloor nt \rfloor}{n} + \frac{(nt - \lfloor nt \rfloor)^2}{n} \xrightarrow[n \to \infty]{} t
\end{aligned}
\tag{20}
$$

It is actually possible to show that $W_t^{(n)} \xrightarrow{D} \mathcal{N}(0,t)$. We want to define the Brownian Motion as the natural limit process for the sequence $(W^{(n)})_{n \in \mathbb{N}}$.

Definition 10 The Brownian Motion W is a real values stochastic process that satisfies

- $W(0) = 0$ almost surely.
- W has independent increments. It means that for every $k \in \mathbb{N}$ and every $0 = t_0 \leq t_1 \leq ... \leq t_k < \infty$, we have $\{W_{t_i} - W_{t_{i-1}}\}_{1 \leq t \leq k}$ for $i = 1, .., k$.
- W has stationary gaussian increments, meaning that for every $0 \leq s < t < \infty$ we have $(W_t - W_s) \sim \mathcal{N}(0, t-s)$.
- W has continuous trajectories $t \mapsto W_t(\omega)$

We would like to say that $W^{(n)} \xrightarrow{D} W$. Is that true? The answer is more complicated than it seems. We showed that, for any fixed $t \in [0, \infty)$ $W_t^{(n)} \xrightarrow{D} W_t$. Hence we have convergence of the finite-dimensional distributions, but this is not enough to say that the entire process converge in distribution.

We have $W^{(n)} \xrightarrow{D} W$ as a process if and only if $\mu^{W^{(n)}} \xrightarrow{w} \mu^W$, i.e. the law induced by the process on the space of trajectories converges weakly. We want again to underline that convergence of the finite-dimensional distribution is not enough to guarantee this weak convergence. We need to require something more. We need the sequence of laws $(\mu^{W^{(n)}})_{n \in \mathbb{N}}$ to be relatively compact.

Definition 11 A family Π of probability measures on a metric space is relatively compact if every sequence of elements of Π contains a weakly convergent subsequence.

Proposition 4.1 *Let $(P_n)_{n\in\mathbb{N}}, P$ be probability measures on the space of continuous functions $\mathcal{C}$. Assume that the finite-dimensional distributions of $(P_n)_{n\in\mathbb{N}}$ converge weakly to those of P. Assume moreover that the sequence $(P_n)_{n\in\mathbb{N}}$ is relatively compact. Then $P_n \xrightarrow{w} P$.*

Proof. This is not a complete proof, just a sketch with some ideas. Starting from the definition of relative compactness, we have that every subsequence $(P'_n)_{n\in\mathbb{N}}$ has a further subsequence $(P''_n)_{n\in\mathbb{N}}$ converging weakly to some limit Q. Since $P''_n \xrightarrow{w} Q$, then also the finite-dimensional distributions of P''_n converge weakly to those of Q; and since the finite-dimensional distributions of P_n converge weakly to those of P, then the finite-dimensional distributions of Q must coincide to those of P. Moreover, every probability measure on $\mathcal{C}$ is completely determined by its finite-dimensional distributions, hence we conclude that we must have $P \equiv Q$. We showed that any subsequence $(P'_n)_{n\in\mathbb{N}}$ contains a further subsequence $(P''_n)_{n\in\mathbb{N}}$ that is weakly convergent to P. This is enough to ensure weak convergence $P_n \xrightarrow{w} P$. □

This idea is a powerful technique to prove weak convergence in $\mathcal{C}$ and other functional spaces. As first we have to prove that the finite-dimensional distributions converge weakly to some limit and we have to identify the limit. Next we have to prove that the sequence is relatively compact. Going back to our random walk, we can actually prove that the sequence $(\mu^{W^{(n)}})_{n\in\mathbb{N}}$ is relatively compact. Hence we have that $W^{(n)} \xrightarrow{D} W$.

To conclude this section we want to show that relatively compactness is fundamental. Suppose that P gives unit mass to the function $x \equiv 0$ and, for every $n \in \mathbb{N}$, the probability measure P_n gives unit mass to the function x_n defined as follows

$$x_n(t) = \begin{cases} nt & t \in [0, 1/n] \\ 2 - nt & t \in [1/n, 2/n] \\ 0 & t \in [2/n, \infty) \end{cases} \tag{21}$$

We can have weak convergence $P_n \xrightarrow{w} P$ if and only if we have uniform convergence of x_n to x, i.e. if we have convergence in the topology of $\mathcal{C}$. But $\sup_{t\in[0,\infty)} |x_n(t) - x(t)| = 1$ for all $n \in \mathbb{N}$. Hence weak convergence of $(P_n)_{n\in\mathbb{N}}$ is not possible. However, we have convergence of the finite-dimensional distributions.

5 Anomalous Limits

In the previous section we started from a Random Walk, describing the dynamics of the discrete step. If suitably rescaled, we showed that it converges to the Brownian Motion. Is some way we can think of the Random Walk as the infinitesimal dynamic that is responsible for the macroscopic behavior. Note that the Brownian Motion is diffusive, meaning that portion of the system "explored" by a random particle moving according to a Brownian Motion grows linearly in time, or better, that we have $\mathbb{E}[|W(t)|^2] \sim t$. There are many cases in which processes Y have ”anomaluos” rescaling, meaning that $\mathbb{E}[|Y(t)|^2] \sim t^{2\delta}$ with $\delta \neq 1/2$. Superdiffusive processes ($\delta > 1/2$) arise naturally, mainly connected to motion in disorder media, like a tracer in a turbolent flow, a light particle in an optical lattice or molecular diffusion in porous media. We are interested in the study of microscopic dynamics that give birth to superdiffusive behavior.

A first possibility to explore is represented by the Levy Flight. The typical trajectory consists of long ballistic flights and short disorder motion. The notion of stable random variable will be essential. Given $\alpha \in (0,2)$, an α-stable random variable ζ satisfies the following relations

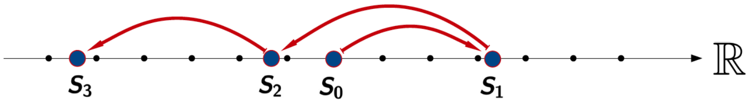
$$\mathbb{P}\left(|\zeta| > x\right) \approx x^{-\alpha} \qquad \text{for} \quad |x| \gg 1 \tag{22}$$

$$\frac{\zeta_1 + \zeta_2 + \cdots \zeta_n}{n^{1/\alpha}} \xrightarrow{\sim} \zeta \tag{23}$$

Property (22) indicates that an α-stable random variable has heavy tails. The anomalous rescaling is reflected in condition (23). We skip in this lecture notes the proper definition and properties of α-stable random variables. The two equations above should only be intended as a motivations.

The Levy Flight is a Random Walk with jump-steps given by a sequence of i.i.d. α-stable random variables, with $\alpha \in (0,2)$. We have

$$S_n = \sum_{k=1}^{n} \zeta_k \qquad \text{with} \quad \zeta_k \sim \alpha\text{-stable} \tag{24}$$

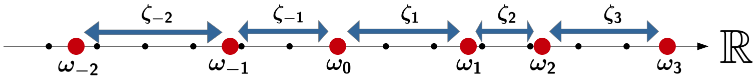


This way of modeling contains the first idea to include non-standard rescaling, but does not include the idea of moving in a medium, meaning that each jump lends with probability 1 into a new point.

We define the Random Environment $\omega = \{\omega_k\}_{k\in\mathbb{Z}}$ as a Point Process on $\mathbb{R}$ in the following way:

$$\begin{aligned} \omega_0 &= 0 \\ \omega_k - \omega_{k-1} &= \zeta_k \qquad \zeta_k \sim \beta\text{-stable} \end{aligned} \tag{25}$$

We are fixing some scattered targets on the real line, assuming the distance between them to be drawn from an β-stable distribution, as represented in the next Figure.



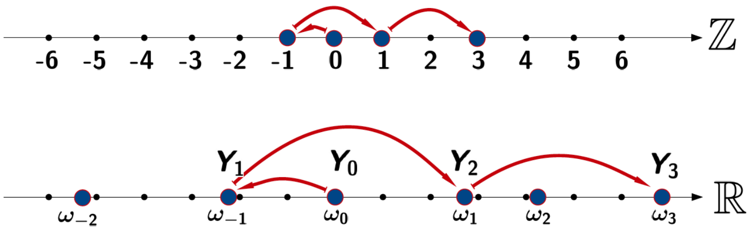
Then we consider a Random Walk S whose increments are i.i.d. random variables taking values in $\mathbb{Z}$.

$$\begin{aligned} S_0 &= 0 \\ S_n &= \sum_{j=1}^{n} \xi_j \end{aligned} \tag{26}$$

Finally we define the Levy Lorentz Gas as a random walk on the random environment

$$Y_n := \omega_{S_n} \equiv \omega \circ S(n) \tag{27}$$

The Levy Lorentz gas Y is a Random Walk on the Random Environment ω. It jumps accordingly to the Underlying Random Walk S.



6 Results

Suppose that $\beta \in (1,2)$. Hence the distance between targets in the random environment has finite mean $\mu = \mathbb{E}[\zeta]$, but infinite variance. This case has been studied in several articles, for example [5], [6],, where they proved the following result.

Theorem 6.1 *The process* $\bar{Y}^{(n)}(t) := \dfrac{Y_{\lfloor nt \rfloor}}{n^{1/2}}$ *converges weakly to* $\mu W(t)$.

Proof. We give a sketch of the proof. This only serves to motivate that $\sqrt{n}$ is the right rescaling. First note that, by the LLN, the environment ω scales linearly with n,

$$\frac{\omega_n}{n} = \frac{\sum_{j=1}^{n} \zeta_j}{n} \xrightarrow[n\to\infty]{a.s.} \mu \tag{28}$$

Using this results, we decompose Y_n and find the suitable rescaling in the following way

$$\frac{Y_n}{n^{1/2}} = \underbrace{\frac{\omega \circ S_n}{S_n}}_{\xrightarrow[LLN]{a.s.} \mu} \cdot \underbrace{\frac{S_n}{n^{1/2}}}_{\xrightarrow[CLT]{D} W(1)} \xrightarrow{D} \mu \cdot W(1) \tag{29}$$

where we used the convergence of a finite variance random walk to the Brownian Motion. □

Note that we obtained a diffusive behavior, that was not our goal.

Suppose now to take $\beta \in (0,1)$, as done in [7]. It means that also the mean distance between targets is infinite. In this case we have that

$$\bar{\omega}^{(n)} = \left(\frac{\omega_{\lfloor nx \rfloor}}{n^{1/\beta}}\right)_{x\in\mathbb{R}} = \left(\frac{\sum_{j=1}^{\lfloor nx \rfloor} \zeta_j}{n^{1/\beta}}\right)_{x\in\mathbb{R}} \xrightarrow[\beta\text{-stability}]{w} Z \tag{30}$$

$$\bar{S}^{(n)} = \left(\frac{S\lfloor nt \rfloor}{n^{1/2}}\right)_{t\in\mathbb{R}^+} = \left(\frac{\sum_{i=1}^{\lfloor nt \rfloor} \xi_i}{n^{1/2}}\right)_{t\in\mathbb{R}^+} \xrightarrow{w} W \tag{31}$$

Following the same procedure of the previous theorem, we would like to say that the process $\bar{Y}^{(n)}(t) := \frac{Y_{\lfloor nt \rfloor}}{n^{1/2\beta}}$ converges weakly to $Z \circ W$, where Z is the process defined as follows

$$Z(s) := \begin{cases} Z_+^{(\beta)}(s) & \text{if } s \geq 0, \\ -Z_-^{(\beta)}(-s) & \text{if } s < 0. \end{cases} \tag{32}$$

with $(Z_\pm^{(\beta)}(x))_{x\geq 0}$ be two i.i.d. càdlàg β-stable processes with independent and stationary increments such that $Z_\pm^{(\beta)}(0) = 0$ and $Z_\pm^{(\beta)}(1)$ is distributed as a β-stable random variable.

Theorem 6.2 *The finite-dimensional distributions of* $\left(\frac{Y_{\lfloor nt \rfloor}}{n^{1/2\beta}}\right)_{t\geq 0}$ *converge to the corresponding distributions of* $(Z \circ W(t))_{t\geq 0}$.

Here the rescaling is superdiffusive, but the theorem above only shows convergence of the finite-dimensional distributions, and can not be extended to the entire process.

To resume we analyzed 3 different cases:

- **Lévy Flights**: It converge to a process with superdiffusive behavior, but the environment changes at every step. Actually there is no environment, the jumps are independent, and this is not the situation we encounter in many physical situations we want to model.

- **Lévy Lorentz Gas** with $\alpha \in (1,2)$: We are able to show convergence of the process, but the limit has a classical diffusive behavior.

- **Lévy Lorentz Gas** with $\alpha \in (0,1)$ The limit process is superdiffusive, however we can only show convergence of the finite-dimensional distributions.

Our goal is to obtain a sequence of processes in a random (fixed) environment, that converge weakly to a superdiffusive limit. The solution we propose is a double source of randomness, coming both from an α-stable random walk (as the Levy walk) and from a β-stable random environment (as in the Levy Lorentz gas).

We define the random environment and the underlying random walk with the following parameters:

- $\omega_n = \sum_{i=1}^{n} \zeta_i \quad (\zeta_i)_{i\in\mathbb{Z}} \sim \ \beta\text{-stable} : \quad \beta \in (1,2), \quad \mathbb{E}[\zeta] = \mu$
- $S_n = \sum_{j=1}^{n} \xi_j \quad (\xi_j)_{j\in\mathbb{N}} \sim \ \alpha\text{-stable} : \quad \begin{cases} \alpha \in (0,1) \\ \\ \alpha \in (1,2), \quad \mathbb{E}[\xi] = 0 \end{cases}$

We provide in [8] a solution to the problem, as stated in this final theorem

Theorem 6.3 (Stivanello, Bet, Bianchi, Lenci, Magnanini) *The the properly rescaled Random Walk in Random Environment*

$$\bar{Y}^{(n)} := \left(\frac{Y_{\lfloor nt \rfloor}}{n^{1/\alpha}} \right)_{t\geq 0} \tag{33}$$

converges weakly to $\mu W^{(\alpha)}$, a superdiffusive α-stable process.

References

[1] P. Billingsley, "Convergence of probability measures". Wiley & Sons. New York, 1968.

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

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

[4] N. Berger, R. Rosenthal, *Random walks on discrete point processes.* Ann. Inst. Henri Poincaré Probab. Stat. (2015).

[5] A. Bianchi, G. Cristadoro, M. Lenci, M. Ligabò, *RandomWalks in a One-Dimensional Lévy Random Environment.* Journal of Statistical Physics (2016).

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

[7] A. Bianchi, M. Lenci, F. Penè, *Continuous-time random walk between Lévy-spaced targets in the real line.* Stochastic Process. Appl. (2020).

[8] S. Stivanello, G. Bet, A. Bianchi, M. Lenci, E. Magnanini, *Limit theorems for Lévy flights on a 1D Lévy random medium.* In preparation.

Some features of finite simple groups

Daniele Garzoni [(*)]

Abstract. Finite simple groups are the building blocks of finite groups. For this reason, since the early days of group theory lots of efforts were devoted to understanding this class of groups. These culminated in the 1980's in an enormous theorem — known as Classification Theorem — which exhibits a very precise list of these groups. In this seminar we will state the theorem, and briefly describe the objects involved. We will then focus on some features of finite simple groups. For instance, we will see that it is amazingly easy to generate these groups by few elements. Along the way, we will try to explain the impact of the Classification in the field of group theory.

1 Introduction

The interest in finite simple groups arose very early in the life of group theory. This culminated in the 1980's in a theorem which provides a precise (infinite) list of these groups: this theorem goes under the name of Classification of Finite Simple Groups (CFSG for short). The proof of this theorem is probably one of the most complicated in the history of mathematics. In this note we will state the theorem, and give a brief description and present some feature of the objects involved.

The study of finite simple groups is a vast topic, and we are not in position here to give a satisfactory account; hence the author has made some choices. One of the main choices is that the historical overview is reduced to a mininum. What is more, the proof of the CFSG is essentially not discussed. There are two main reasons for this. The first is that, as we already mentioned, the proof is rather complicated, and it seems impossible to introduce the main ideas in a reasonable amount of space. The second, more important, reason, is that the author lacks any serious understanding of the proof: he would not be able to discuss the main ideas even if the reader was prepared to digest a long manuscript. Well, the author would probably be able to write something — nobody can ask questions when someone is hidden behind the pen. However, it does not seem very respectful to act in this way. Fortunately, there are some places in the literature where one can find very interesting discussions on the history, the proof and the consequences of the CFSG: two are certainly [Sol01] and [Sol18].

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: daniele.garzoni@phd.unipd.it . Seminar held on 12 February 2020.

The bibliography is far from exhaustive: only the references which have been explicitly mentioned along the note are reported. With one exception: [EG19] is a very tiny contribution that the author, in joint work with Sean Eberhard, gave on some problems of the flavour as those discussed in Section 5.

The style of exposition is friendly: there are few definitions, and few concepts are treated in detail. The author has tried to avoid technical details whenever possible; he hopes that the reader will not find this approach sloppy, rather than friendly!

2 Definition, motivation and a bit of history

Assume we are given a finite group G, and assume we want to understand its structure. As always, one tries to decompose the object into simpler pieces; then one tries to understand the different pieces, and finally tries to gather together the information. For many purposes, in a finite group G the simple pieces are indeed called *simple groups*. Let us give the definition.

A subgroup H of a group G is called *normal* if it is closed under conjugation, that is, $g^{-1}Hg = H$ for every $g \in G$. Whenever H is a normal subgroup of G, one can construct a group, usually denoted G/H, whose elements are the right cosets of H in G. Groups constructed in this way are called *quotient groups*. Note that G itself and the trivial subgroup $\{1\}$ are always normal subgroups of G.

Definition 2.1 A group $G \neq \{1\}$ is called *simple* if it admits only two normal subgroups, namely $\{1\}$ and G.

Let us immediately restrict to the case in which G is finite. In this case, G can be broken down into simple pieces in essentially one way. All we need to remember is this sentence; for completeness, however, we report an exact statement.

Theorem 2.2 (Jordan–Hölder Theorem) *Let G be a finite group. There exists a chain of subgroups*

$$1 = H_0 < H_1 < \cdots < H_t = G$$

such that for every $i = 0, \ldots, t-1$,

(a) *H_i is normal in H_{i+1}, and*

(b) *H_{i+1}/H_i is simple.*

Moreover, if

$$1 = H'_0 < H'_1 < \cdots < H'_\ell = G$$

is any other chain satisfying properties (a) and (b), then $t = \ell$ and there exists a permutation ρ of the set $\{0, 1, \ldots, t-1\}$ such that for every $i = 0, \ldots, t-1$, we have

$$H_{\rho(i)+1}/H_{\rho(i)} \cong H'_{i+1}/H'_i.$$

We have now learnt that finite simple groups (FSG for short) are the building blocks of finite groups. For this reason, it is very often the case that questions about general finite groups can be reduced to FSG. In other words, a certain question has a positive answer for all finite groups, provided it has a positive answer for FSG.[8]

Therefore, people got interested in FSG since the early days of group theory. The German mathematician Otto Hölder, in 1892, writes *"It would be of the greatest interest if it were possible to give an overview of the entire collection of finite simple groups."* We begin with the easy case.

2.1 The abelian finite simple groups

It is very easy to classify the abelian (i.e., commutative) finite simple groups. Indeed, they are just the cyclic groups of prime order. This is easy to prove and was known to Hölder and his contemporaries. Therefore, Hölder quote is really about *nonabelian* FSG.

2.2 From now on: nonabelian finite simple groups

From now on we focus on nonabelian finite simple groups, since we have not much else to say about the abelian ones. In 1872 the Norwegian mathematician Ludwig Sylow proved a theorem, usually divided in three parts, which had a dramatic impact on the process of understanding the structure of finite groups. We will not report the exact statement. It is a theorem which imposes some arithmetical conditions on the number of certain subgroups of a finite group. Amazingly, these conditions put severe restrictions on the structure of finite groups, and in particular on the presence of normal subgroups. Indeed, using only Sylow theorems, at the beginning of the 20th century some people managed to make some progress towards the understanding of finite simple groups. For example, the following was proved thanks to the joint efforts of Hölder, Cole, Miller and Ling.

Theorem 2.3 (1900) *The finite simple groups of order at most* 2001 *are classified. In particular, there are only seven nonabelian among them.*

Other results of this flavour were proved. People started to get a feeling that finite simple groups might be very rare objects, suitable for some sort of classification.

There were several crucial developments towards the understanding of FSG (and of finite groups in general). A very important one happened around 1900 as well, namely, the introduction of *character theory* in the study of finite groups. This was originally done by the German mathematician Georg Frebenius, in response to a famous letter sent to him by his compatriot Richard Dedekind. The setup is as follows. We are given a finite group G, and we look at complex (finite dimensional) representations of G: these are morphisms $\rho : G \to \mathrm{GL}_n(\mathbf{C})$ for some positive integer n. To each representation ρ, we associate a mapping $\chi_\rho : G \to \mathbf{C}$, which is called *character*, and which sends $g \in G$ to the trace of the linear map $\rho(g)$. The game is to try to recover information about the structure of G by looking at the characters χ_ρ, where ρ varies among all representations.

[8] Actually, one often reduces questions to *almost simple* groups, namely, groups G such that $S \leqslant G \leqslant \mathrm{Aut}(S)$ for some simple group S. It is not necessary to go into these technicalities here, however.

(In fact, one usually only looks at *irreducible* representations, that here we do not define.) Somewhat surprisingly, this strategy works very well, and it is still nowadays object of active research.

Later, other developments occurred. Between the 1950's and the 1970's, very sophisticated methods were introduced with the purpose of penetrating the structure of finite groups. And here is where we stop our historical overview. The reasons why we do this were explained already in the introduction, so we do not insist here.

3 The Classification of Finite Simple Groups

Around 1980, it seemed that one could state the following result.

Theorem 3.1 (Classification of Finite Simple Groups) *Let G be a finite simple group. Then, G is either*

(a) *a cyclic group of prime order;*

(b) *an alternating group of degree at least* 5*;*

(c) *a finite simple group of Lie type;*

(d) *one of* 26 *sporadic finite simple groups.*

It is our purpose to give a brief description of all the groups mentioned in the statement. Before doing that, however, let us spend few words about the proof. This was immensely deep and complicated: 15000 journal pages, spread across 500 separate papers, by more than 100 mathematicians. Note that immediately before Theorem 3.1, we used the words "it *seemed* that one could state...". Indeed, around 1989 the American mathematician Michael Aschbacher noticed that there was a substantial gap in the proof. We will come back on this, with some more comments on the proof (and on some ongoing revision projects of the proof), in Section 6. It is now time to explain what the objects in Theorem 3.1 are.

4 The groups in Theorem 3.1

We already discussed item (1) in Subsection 2.1, hence we move to item (2).

4.1 Alternating groups

We begin by recalling the definition. Let Ω be a finite set, and let $\mathrm{Sym}(\Omega)$ denote the set of all bijective maps $\Omega \to \Omega$; then $\mathrm{Sym}(\Omega)$ is a group under composition, called the *symmetric group* on the set Ω. The elements of $\mathrm{Sym}(\Omega)$ are usually called *permutations*. It turns out that each permutation π can be written as a product of a certain number, say r, of *transpositions*, i.e., mappings which exchange two points and fix all the other points. This decomposition of π is not unique; however, it turns out that, given π, the parity of r is uniquely determined. Then, the set of all permutations that can be written as the product of an even number of transpositions is a normal subgroup of index 2 of $\mathrm{Sym}(\Omega)$, called the *alternating group* on the set Ω, and usually denoted by $\mathrm{Alt}(n)$. Note that, if Ω and Ω' are two sets of the same size, say n, then $\mathrm{Sym}(\Omega)$ and $\mathrm{Sym}(\Omega')$ are naturally isomorphic. Then, it is common to fix one Ω, and just denote by S_n and A_n the symmetric group and the alternating group on Ω, respectively.

Theorem 4.1 *Assume $n \geqslant 5$. Then, A_n is simple nonabelian.*

The cases $n \leqslant 4$ are genuine exceptions, as it is not hard to show. Historically, there is not much we have to say. Symmetric and alternating groups are probably the most natural families of finite groups, and were known since the very beginning of group theory.

The reader who is unfamiliar with these definitions can safely forget about A_n, and just think of S_n (which is more natural to define). Indeed, for most purposes — certainly for those which concern us in this note — A_n and S_n can be thought of as essentially the same thing: everything which is true for one group applies with easy changes to the other.

4.2 Groups of Lie type: classical groups

We begin here with an example, which should be thought of as the "key" example of finite simple group of Lie type.

Let q be a prime power, let $\mathbf{F}_q$ be a finite field with q elements (recall all such fields are isomorphic over the base field), and let n be a positive integer. We consider the set $\mathrm{GL}_n(q)$ of all invertible $n \times n$ matrices over $\mathbf{F}_q$. This set is a group under usual matrix multiplication, called the *general linear group* (of degree n, over the field with q elements). We can easily detect two normal subgroups inside $\mathrm{GL}_n(q)$. The first consists of all matrices having determinant 1. This subgroup is usually denoted by $\mathrm{SL}_n(q)$, and called the *special linear group.* The second, that we denote by $Z_n(q)$, consists of all scalar matrices, i.e., matrices of the form $\mathrm{diag}(\lambda, \dots, \lambda)$ for some nonzero $\lambda \in \mathbf{F}_q$. It turns out that, if we remove these two basic obstructions, we get a simple group. Specifically, define $\mathrm{PSL}_n(q) := \mathrm{SL}_n(q)/(\mathrm{SL}_n(q) \cap Z_n(q))$, and we have

Theorem 4.2 $\mathrm{PSL}_n(q)$ *is simple nonabelian provided $n \geqslant 2$ and $(n, q) \neq (2, 2), (2, 3)$.*

The group $\mathrm{PSL}_n(q)$ is called *projective special linear group.* This terminology comes from the fact that $\mathrm{PSL}_n(q)$ acts faithfully on the projective space $\mathbf{P}^{n-1}(\mathbf{F}_q)$, i.e., the set of all 1-dimensional subspaces of $\mathbf{F}_q^n$. Again, the exceptions in the theorem are genuine. Indeed, if $n = 1$ then $\mathrm{PSL}_n(q)$ is the trivial group, and moreover $\mathrm{PSL}_2(2) \cong S_3$ and $\mathrm{PSL}_2(3) \cong A_4$.

Other finite simple groups can be constructed in a similar fashion, starting from subgroups of $\mathrm{GL}_n(q)$ preserving suitable nondegenerate forms over $\mathbf{F}_q$. In this way one gets symplectic groups, unitary groups, and orthogonal groups. The simple versions are usually denote by $\mathrm{PSp}_{2n}(q)$ in the symplectic case (these occur only in even degree), $\mathrm{PSU}_n(q)$ in the unitary case, $\mathrm{P\Omega}_{2n+1}(q)$, $\mathrm{P\Omega}^+_{2n}(q)$ and $\mathrm{P\Omega}^-_{2n}(q)$ in the orthogonal case. Over finite fields, simple orthogonal groups are the trickiest to define. For instance, we see that in even degree we have two type of groups ("plus" type, and "minus" type).

These groups — linear, symplectic, unitary, orthogonal — are called the finite simple *classical groups*. The reader who is unfamiliar with these can safely keep in mind only the example $\mathrm{PSL}_n(q)$. For most purposes — although not quite for all! — what is true for $\mathrm{PSL}_n(q)$ is true also for the other families, with changes which can be more or less painful depending on the situation.

4.3 Groups of Lie type: exceptional groups

Groups of Lie type are not finished here, however. Indeed, there are also the finite simple *exceptional groups* of Lie type. It would be possible to define some of them in a reasonably quick way. However the mere definition might sound a bit artificial, so we prefer to avoid doing this.

4.4 Groups of Lie type: history and connections

Finite classical groups have been known for quite a long time. Indeed, the American mathematician Leonard Dickson proved the simplicity of these groups in 1901. Around the same time, Wilhelm Killing end Élie Cartan classified the complex simple Lie groups. Roughly, the strategy was to associate to each simple Lie group a certain linear strucure, called Lie algebra, and to provide a classification of the possible Lie algebras occurring in terms of certain combinatorial invariants. The classification of Lie algebras can then be pulled back to offer a classification of Lie groups. In the end, one gets two types of groups: the simple classical Lie groups, and the simple exceptional Lie groups. It was very clear to Dickson that his study of finite simple classical groups had to be connected to the Killing–Cartan classification. In particular, it should have been possible to analyze the finite classical groups using the same combinatorial techniques used by Killing and Cartan. What is more, Dickson could grasp the fact that finite analogues of the exceptional Lie groups could exist.

Dickson's intuition was correct. However, it took quite a long time to make everything precise. In particular, it was only in the 1960's that Robert Steinberg [Ste68] (following fundamental work of other people, notably Claude Chevalley and Jacques Tits) provided a uniform construction of finite simple groups of Lie type — both classical and exceptional — using similar ideas to those of Killing and Cartan. In particular, one starts with a linear algebraic group $\mathbf{G}$ over an algebraically closed field of positive characteristic. For these groups, the classification of Killing and Cartan applies. Then one constructs the finite groups of Lie type as groups of fixed points of suitable morphisms $\sigma : \mathbf{G} \to \mathbf{G}$, nowadays called Steinberg morphisms. It is amazing that one can provide such an elegant, uniform, description of groups that at first glance might look very different. It should be pointed out that in the finite case many complications occur — indeed, as already observed, it took several more decades to understand things properly. One of the essential reasons is that, in the finite case, one has *twisted* versions of groups of Lie type, which do not exist when one works over an algebraically closed field (for instance, $\mathrm{P}\Omega_{2n}^{-}(q)$ is a twisted simple group of Lie type).

We come now to the last family of finite simple groups.

4.5 Sporadic groups

This is with no doubt the most mysterious family of finite simple groups. It consists of 26 groups. One may argue why 26, and not 25. But this is a sort of philosophical question that the author is not able to address. In fact, similarly one could ask why there are exactly 5 simple complex exceptional Lie groups.

The history of these groups is strange. The first that we have to mention are the five *Mathieu groups*. These five groups were discovered by the French mathematician Émile Mathieu between 1860 and 1870. In order to see the other sporadic groups, we have to jump a century ahead in time. Three of them, the *Conway groups*, were discovered around 1970 by John Conway as group of automorphism of certain lattices [Con69]. The other 18 sporadic groups emerged during the proof of the CFSG. Roughly, when trying to prove a certain classification theorem (i.e., a theorem stating that a group with certain properties must belong to a certain, possibly infinite, list) it was noted that some exceptions occurred provided a group with specific properties existed. It was then proved, by construction, that indeed such exception occurred. (This description of the genesis of sporadic groups is true only to some extent, and not exactly satisfying.)

The author knows very little about sporadic groups; the next subsection somehow reiterates this point.

4.6 Sporadic is difficult

Let us consider the sporadic group with largest order. This is called *the Monster*, and usually denoted by M. It has order

$$808017424794512875886459904961710757005754368000000000 \approx 8 \times 10^{53}$$

This is just a finite number, and perhaps we should not be scared of it. This group can be embedded in $\mathrm{GL}_N(\mathbf{C})$ with $N = 196883$, and this N is minimal: we cannot view M as a subgroup of $\mathrm{GL}_n(\mathbf{C})$ with $n < N$. If we prefer to work with matrices over a finite field, we can embed M inside $\mathrm{GL}_N(2)$. Again, N is minimal.

On the opposite side, let us consider the projective special linear group $\mathrm{PSL}_2(q)$. For several reasons, the family $\{\mathrm{PSL}_2(q)\}_q$ is the easiest family of FSG one can consider. However the order of $\mathrm{PSL}_2(q)$ is

$$\frac{q(q-1)(q+1)}{(2, q-1)}$$

Up to a constant this number is roughly q^3 and, as q grow, gets much larger than $|M|$: the Monster looks very tiny compared to $\mathrm{PSL}_2(q)$ for large q. Therefore, the order is not necessarily a good way to measure the complexity of a structure.

Indeed, what makes $\mathrm{PSL}_2(q)$ easy to understand is that we can just write down 2×2 matrices and work with them. We can see many things just by elementary linear algebra. On the other hand, it is more difficult to work with $N \times N$ matrices. However this is not completely satisfactory. Indeed, we can play a similar game with the family

$\{\mathrm{PSL}_n(2)\}_n$: for large n, we will need to work with tremendously big matrices. Nonetheless it is fair to say that we understand much better $\mathrm{PSL}_n(2)$ than M. The explanation is that $\{\mathrm{PSL}_n(2)\}_n$ comes as a *family*, while M is just an individual entity. And it is much easier to understand things that somehow we can group, or order, together. This is the biggest difference between the sporadic groups and the other finite simple groups. Sporadics looks as strange exceptions, compared to the other families, which admit a satisfying and unified explanation. We feel it necessary, here, to quote John Thompson — one of the main characters of the proof of the CFSG — who wrote in 1982:

"(...) the classification of finite simple groups is an exercise in taxonomy. This is obvious to the expert and to the uninitiated alike. To be sure, the exercise is of colossal length, but length is a concomitant of taxonomy. Those of us who have been engaged in this work are the intellectual confreres of Linnaeus. Not surprisingly, I wonder if a future Darwin will conceptualize and unify our hard won theorems. The great sticking point, though there are several, concerns the sporadic groups. I find it aesthetically repugnant to accept that these groups are mere anomalies... Possibly... The Origin of Groups remains to be written, along lines foreign to those of Linnean outlook."

Indeed, still nowadays many of the misteries regarding the *true* structure of sporadic groups remain to be solved.[9]

It is now time to change topic. So far we have tried to define and introduce the finite simple groups. Let us move to describe some of their properties.

5 Generation properties of finite simple groups

Finite simple groups enjoy amazing properties. We have chosen to discuss some properties concerning generation. Recall that, given a group G and a subset X of G, we say that X generates G if every element of G can be written as a product of elements in $X \cup X^{-1}$ (where X^{-1} denotes the set of inverses of elements of X). Equivalently, the smallest subgroup of G containing X is G itself. We write in this case $G = \langle X \rangle$. The starting point is the following theorem, proved by Steinberg [Ste62] for groups of Lie type, and by Aschbacher–Guralnick [AG84] for sporadic groups (the case of alternating groups being folklore).

Theorem 5.1 *Let G be a finite simple group. Then, G is generated by a subset of size* 2.

We say alternatively that G is 2-generated. Note that a group generated by one element is, by definition, cyclic, hence abelian; therefore one cannot hope to improve the statement above.

For the proof of the theorem, one inspects the list of groups appearing in the CFSG, and proves the statement for each of them. It is worth remarking that, without the

[9] The reader may want to watch on YouTube a short interview to John Conway, an extraordinary mathematician who worked on groups and on many other areas. To be honest, the interview is not exactly entertaining, and one can probably find better material in which Conway is involved; however the topic suits very well our discussion here. The name of the video is "The Monster Group – John Conway".

Classification, so far nobody has been able to prove the following statement: There exists an absolute constant C such that every FSG is generated by a subset of size C. At first glance this might seem strange, but it happens very often in the study of finite simple groups: with the CFSG, one is able to prove spectacular results that one would hardly imagine to be true without it.

5.1 A probabilistic approach

If G is a finite group, denote by $\mathbf{P}_G(t)$ the probability that t randomly chosen elements of G generate G. Here, and in the following, every probabilistic statement is with respect to the uniform distribution; therefore

$$\mathbf{P}_G(t) = \frac{|\{(x_1, \dots, x_t) \in G^t : \langle x_1, \dots, x_t \rangle = G\}|}{|G^t|}$$

In general, since we are dealing with finite sets, this is particularly interesting if we consider families, or sequences, of groups. In this case we can discuss uniform lower or upper bounds to $\mathbf{P}_G(t)$, limits, etc.

Of course, in this report we have a favourite family of groups to consider: the family of all FSG. Indeed we have the following beautiful result, proved by Dixon [Dix69] for alternating groups, by Kantor–Lubotzky [KL90] for classical groups, and by Liebeck–Shalev [LS95] for exceptional groups.

Theorem 5.2 *Let G be a finite simple group. Then, $\mathbf{P}_G(2) \to 1$ as $|G| \to \infty$.*

The statement must be understood as follows: for every $\epsilon > 0$, there exists N such that if G is a finite simple group and $|G| \geqslant N$, then $\mathbf{P}_G(2) \geqslant 1 - \epsilon$.

In words, if we pick two random elements from a finite simple group G, then these elements are very likely to generate G provided G is large. Note that, being finitely many, the sporadic finite simple groups are not involved in such a statement. It is a good moment for the reader to recall that, without the Classification, we are not even able to show that a finite simple group is generated by boundedly many elements. We are now going to say something about the proof of Theorem 5.2.

5.2 Maximal subgroups

There is an intimate connection between $\mathbf{P}_G(t)$ and the maximal subgroups of G. Recall that, given a group G, a proper subgroup H of G is called maximal if there are no subgroups between H and G. More precisely, whenever $H \leqslant K \leqslant G$, it must be either $H = K$ or $K = G$. A finite group with at least 2 elements contains maximal subgroups, and what is more, every proper subgroup is contained in a maximal one.[10] Therefore, we deduce that, if a subset X of G does not generate G, then X is contained in a maximal subgroup

[10] This is not always the case: there are infinite groups which contain no maximal subgroups. In other words, every proper subgroup is properly contained in another proper subgroup.

of G. Hence, if we denote by $\mathcal{M}(G)$ the set of maximal subgroups of G, we have

$$1 - \mathbf{P}_G(t) = \frac{|\cup_{M\in\mathcal{M}(G)} M^t|}{|G^t|} \tag{1}$$

Therefore we see that giving a lower bound to $\mathbf{P}_G(t)$ is equivalent to giving an upper bound to the right-hand side of (1). A first idea could be to give a trivial union bound, namely

$$\frac{|\cup_{M\in\mathcal{M}(G)} M^t|}{|G^t|} \leqslant \sum_{M\in\mathcal{M}(G)} \frac{|M|^t}{|G|^t}. \tag{2}$$

For a general finite group, this estimate is not accurate. Indeed, it does not take into account intersections between maximal subgroups, which instead might be relevant. However, amazingly, for finite simple groups the estimate turns out to be effective. Indeed, in order to prove Theorem 5.2, the authors proved that the right-hand side of (2) for $t = 2$ goes to zero as $|G| \to \infty$. Note that this amounts to proving that maximal subgroups of finite simple groups are "few" and "small" in some sense.

5.3 The probabilistic method

Theorem 5.2 implies that, if G is a sufficiently large finite simple group, then G is generated by two elements. The reader might be not particularly enthusiastic about this corollary: we already said in Theorem 5.1 that this holds for *all* FSG. However, the method of proof is interesting, in the following sense. Theorem 5.1 s proved by a constructive argument: pick a finite simple group G, and find explicitly two elements which generate G. On the other hand, Theorem 5.2 is proved by estimating the number and the size of maximal subgroups of G. Therefore we are not directly addressing a problem of generation. Nonetheless, as a corollary, we get the groups are generated by two elements. This is an instance of the so called "probabilistic method", pioneered by the Hungarian mathematician Paul Erdős, which can be publicized as follows: "If you want to show that something exists, show it has positive probability."

It is often the case that showing that a set has positive probability — by counting arguments, or by other means — is easier than showing directly that the set is nonempty. In questions regarding generation of FSG, this is a very common approach. Indeed, there are some problems for which only a probabilistic proof is known: we know that the groups are generated by elements with certain properties, but we are not actually able to construct these elements.

5.4 Asymptotic statements in finite simple groups

We spend few words about general asymptotic questions in finite simple groups — not necessarily regarding generation. With "asymptotic", we mean that we are interested in G simple with $|G| \to \infty$. We immediately note that, in these questions, sporadic simple groups can be ignored, since they are finitely many.

It would be a breakthrough to give a direct proof of the fact that there exist only finitely many sporadic finite simple groups.(11) Indeed, for many applications only large finite simple groups are important; and these applications would not rely anymore on the CFSG. As an example, related to our previous discussion, it would imply that finite simple groups are generated by an absolute number of elements. Unfortunately, currently it seems absolutely hopeless to succeed in proving "directly" that there are finitely many sporadics.

In any case, in asymptotic statements we can ignore the sporadics, and therefore we have to deal with the remaining groups. These can be divided into three families:

(a) Alternating groups A_n with $n \to \infty$.

(b) Groups of Lie type of bounded rank (e.g. $\mathrm{PSL}_n(q)$ with n fixed and $q \to \infty$.)(12)

(c) Groups of Lie type of large rank (e.g. $\mathrm{PSL}_n(q)$ with $n \to \infty$.)

We can guess that the methods involved in (1) are usually different from the methods in (2) and (3). This is not particularly surprising, as the relevant groups have different nature. However, it is often the case that the methods in (2) are different from the methods in (3). For instance, the family $\mathrm{PSL}_2(q)$ with $q \to \infty$ (small matrices over large fields) can be quite different from the family $\mathrm{PSL}_n(2)$ with $n \to \infty$ (big matrices over small fields).

In the last decades, asymptotic questions about finite simple groups have become very popular, and are object of vibrant research. It is a realm where the finer details disappear, and the rough lines become the main focus.(13)

6 Revision projects

As we mentioned already, the proof of the CFSG was extremely long and difficult. Moreover, in 1989 Michael Aschbacher found a gap in the proof. In 1992, at a conference, he announced a new proof of the relevant parts. In the end, the task turned out to be rather complex, and the final proof of the so called "quasithin case" was published in 2004, in joint work with Stephen Smith ([AS04b] and [AS04a]; in total, more than 1000 pages!).

In addition to this, since the original announcement of the CFSG in the 80's, it was clear to the community of group theorists that a revision of the proof was necessary. In particular, it was necessary to try to collect the proof in a unique place (for instance in a series of books), at the same time trying to improve and shorten the exposition. Needless to say, this was en enormous task to undertake.

It began in 1994 with a book called "The Classification of Finite Simple Groups" [GLS94] by Daniel Gorenstein, Richard Lyons and Ronald Solomon. This was the first

(11) The reader may decide autonomously which meaning to give to the word "direct". At least, it should be less than 15000 pages!

(12) The *rank* can be defined for every group of Lie type.

(13) This quote is taken from [Sha01]. This short survey is very much recommended!

book in a series, in which the authors planned to carry out the project described above.[14]

The series is not finished yet. The eighth volume [GLS18] has been published in 2018. There should be twelve volumes in total, for a total of 5000 pages or so (to which one should sum [AS04b] and [AS04a]). There is a hope that the last volume will be pubslished within 2023. Fortunately, other mathematicians are contributing in this fundamental work.

It is fair to say that without the CFSG most of the fundamental results in the field of group theory would not have been proved (or imagined), let alone the numerous applications to other parts of mathematics. Therefore, the community of group theorists should be (and indeed, as far as I can tell, is) very grateful to the people who are taking part to this revision project. These people are spending their mathematical lives to pursue a difficult and fundamental goal, which most group theorists just ignore: it is much easier to apply the CFSG, than to read, understand and improve its proof.

References

[AG84] M. Aschbacher and R. Guralnick, *Some applications of the first cohomology group.* J. Algebra 90/2 (1984), 446–460.

[AS04a] Michael Aschbacher and Stephen D. Smith, "The classification of quasithin groups. I". Volume 111 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2004. Structure of strongly quasithin K-groups.

[AS04b] Michael Aschbacher and Stephen D. Smith, "The classification of quasithin groups. II". Volume 112 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2004. Main theorems: the classification of simple QTKE-groups.

[Con69] J.H. Conway., *A group of order 8; 315; 553; 613; 086; 720; 000.* Bull. London Math. Soc. 1 (1969), 79–88.

[Dix69] John D. Dixon, *The probability of generating the symmetric group.* Math. Z. 110 (1969), 199–205.

[EG19] Sean Eberhard and Daniele Garzoni, *Random generation with cycle type restrictions.* arXiv preprint, arXiv:1904.12180 (2019).

[GLS94] Daniel Gorenstein, Richard Lyons, and Ronald Solomon, "The classification of the finite simple groups". Volume 40 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 1994.

[GLS18] Daniel Gorenstein, Richard Lyons, and Ronald Solomon, "The classification of the finite simple groups". Number 8. Part III. Chapters 12–17. The generic case, completed. Volume 40 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2018.

[14] Daniel Gorenstein passed away in 1992. He contributed in an essential way to the CFSG, and he always appreciated the urgency of a revision of the proof. Therefore, he was included as an author of this project as a tribute.

[KL90] William M. Kantor and Alexander Lubotzky, *The probability of generating a finite classical group.* Geom. Dedicata 36/1 (1990), 67–87.

[LS95] Martin W. Liebeck and Aner Shalev, *The probability of generating a finite simple group.* Geom. Dedicata 56/1 (1995), 103–113.

[Sha01] Aner Shalev, *Asymptotic group theory.* Notices Amer. Math. Soc. 48/4 (2001), 383–389.

[Sol01] Ronald Solomon, *A brief history of the classification of the finite simple groups.* Bull. Amer. Math. Soc. (N.S.), 38/3 (2001), 315–352.

[Sol18] Ronald Solomon, *Afterword to the article "A brief history of the classification of the finite simple groups".* Bull. Amer. Math. Soc. (N.S.) 55/4 (2018), 453–457.

[Ste62] Robert Steinberg, *Generators for simple groups.* Canadian J. Math. 14 (1962), 277–283.

[Ste68] Robert Steinberg, "Endomorphisms of linear algebraic groups". Memoirs of the American Mathematical Society, No. 80. American Mathematical Society, Providence, R.I., 1968.

Permutation group theory

Mariapia Moscatiello [(*)]

Abstract. The modern notion of a permutation group arises naturally throughout mathematics, with important applications across the sciences. In this note we will focus on finite permutation groups. After introducing some basic concepts, we will see, with some examples, how the Classification of Finite Simple Groups has revolutionized the study of finite permutation groups. We will then highlight some connections to other areas of mathematics.

The study of permutation groups is an old subject with a rich history, stretching all the way back to the origins of group theory in the early 19th century. The modern notion of a permutation group is extremely flexible, and they arise naturally throughout mathematics, with important applications across the sciences. For instance, given any mathematical object or structure Σ (e.g. vector space, group, graph, topological space, etc.) based on a set of points Ω (e.g. vectors, group elements, vertices, points, etc.) then the set $\mathrm{Aut}(\Sigma)$ of automorphisms (or symmetries) of Σ (i.e. the bijective maps $f : \Omega \to \Omega$ such that f and f^{-1} preserve the structure of Σ) is a permutation group on Ω. That is, $\mathrm{Aut}(\Sigma)$ is a group of bijections from Ω to itself.

In this short note, we will focus on finite permutation groups, which continue to be a very active area of current research. In particular, we will states very recent results obtained in this contex. After introducing some very basic concepts and key tools, we will see, how the Classification of Finite Simple Groups (CFSG) has revolutionized the study of finite permutation groups. Finally, we will introduce the concept of base size, which has interesting connections to other areas of mathematics, such as computational group theory, representation theory, graph theory.

1 Notation, terminology and basic facts

Throughout this section, it is possible to refer to [4] as the main reference. From now on, Ω is a finite set. Let $\mathrm{Sym}(\Omega)$ be the set of bijections $\Omega \to \Omega$ (also called permutations of Ω). The operation of the usual composition of functions endow $\mathrm{Sym}(\Omega)$ with the structure of a group, this is called the Symmetric Group on Ω. If Ω_1 and Ω_2 are equipotent sets, then $\mathrm{Sym}(\Omega_1)$ and $\mathrm{Sym}(\Omega_2)$ are isomorphic groups. When $\Omega = \{1, \dots, n\}$, we denote denote

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: `mariapia.moscatiello@math.unipd.it` . Seminar held on 6 May 2020.

$\mathrm{Sym}(\Omega)$ by $\mathrm{Sym}(n)$, and we call this the symmetric group of degree n. The cardinality of $\mathrm{Sym}(n)$ is $n!$. To familiaze the reader with usual notations we consider as an example a couple of non trivial partitions in $\mathrm{Sym}(3)$: the partition that maps $1 \to 2 \to 3 \to 1$ is denoted by $(1, 2.3)$, and the partition that maps $1 \to 2 \to 1$ is is denoted by $(1, 2)$. Composition works as follows:

$$(1,2)(1,2,3) = (1,3), \text{ and } (1,2,3)(1,2) = (2,3). \tag{1}$$

In general, an element of the form $(i, j) \in \mathrm{Sym}(n)$ is called transposition, and an element of the form $(i_1, \dots, i_k) \in \mathrm{Sym}(n)$ is called k-cycle. Hence $(1, 2, 3)$ is a 3-cycle. Note that disjoint cycles always commute, and from (1) we deduce that $\mathrm{Sym}(n)$ is not abelian, for every $n \geq 3$.

Fact 1 Every permutation can be written (up to reordering) in a unique way as a product of disjoint cycles.

Further, every permutation can be written (not in a unique way) as product of transpositions. A permutation is called even (and it has $+$ sign) if it can be written as the product of an even number of transpositions, and odd (and it has $-$ sign) otherwise. In particular, all transposition are odd, and all the 3-cycles are even. Note that, a k-cycle is even if and only if k is odd, and a product of disjoint cycles is an even permutation if and only if the number of cycles of even length is even. Hence the product of even permutations is an even permutation. Let us denote by C_2 the set $\{-1, 1\}$ with the operation given by multiplication. Then C_2 is an abelian group, it is a cyclic group (a group generated by one element, in this case the generator is -1) of order 2. The following map

$$\mathrm{sgn} : \sigma \in \mathrm{Sym}(n) \mapsto \mathrm{sign}(\sigma) \in C_2$$

is a surjective group homomorphism, and the kernel of sgn is the normal subgorup of $\mathrm{Sym}(n)$ consisting of the even permutations. This subgroup is called the alternating group of degree n, and we denote this by $\mathrm{Alt}(n)$. By the first isomomorphism theorem, $\mathrm{Sym}(n)/\mathrm{Alt}(n) \cong \mathrm{Sym}(n)^\sigma = C_2$, hence $\mathrm{Alt}(n)$ has order $n!/2$.
A *permutation group* on Ω is a subgroup of $\mathrm{Sym}(\Omega)$; that is, a permutation group G on Ω is a set of permutations of Ω which is closed under composition, contains the identity permutation, and contains the inverse of each of its elements. The *degree* of the permutation group is the order of Ω. There is an intimately related concept, that of a group action.

Let G be a group (in the abstract sense of group theory, a set with a binary operation). Then an *action* of G on Ω is a (group) homomorphism φ from G to $\mathrm{Sym}(\Omega)$, we say that Ω is a G-set. For $x \in G$, $\alpha \in \Omega$ we write α^x to denote the element $\alpha(x\varphi) \in \Omega$. The image of φ, denoted by G^Ω is a subgroup $\mathrm{Sym}(\Omega)$. That is, G^Ω is a permutation group on Ω. The action is *faithful* if $Ker(\varphi)$ is trivial; in this case $G \cong G^\Omega$, hence G is a permutation group on Ω. Conversely, if G is a permutation group on Ω, taking φ to be the natural inclusion of G in $\mathrm{Sym}(\Omega)$, then we have an action of G on Ω.

For $x \in G$, $\alpha \in \Omega$ we define

$$\begin{array}{ll} \alpha^G := \{\alpha^x \mid x \in G\} & \text{the orbit of } \alpha \text{ in } G; \\ G_\alpha := \{g \in G \mid \alpha^g = \alpha\} & \text{the stabilizer of } \alpha \text{ in } G. \end{array}$$

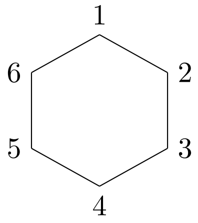
It is an easy exercise to show that G_α is a subgroup of G, and that $G_{\alpha^x} = G_\alpha^x := x^{-1}G_\alpha x$.

Let G be a finite group, let $g \in G$, and let $\lambda_g : x \in G \mapsto xg \in G$. Note that λ_g is a bijection, and that

$$\Lambda : g \in G \mapsto \lambda_g \in \mathrm{Sym}(G)$$

is a (group) homomorphism, that is G acts on itself. Note that Λ is a faithful action, hence G is a prmutation group on G. So, every finite group is a permuation group (Cayley's theorem).

Example 2 Let D_{12} be the group of the symmetries of a regular hexagon. The elements of D_{12} are the rotations of an angle of degree $60°i$, with $1 \leq i \leq 6$ and the reflections around the six axes of the regular hexagon. In particular, D_{12} consists in 12 elements. By Cayley's theorem D_{12} is a permutation group of degree 12, but it is possible to do better. Indeed, we will show that D_{12} is a permutation group of degree 6. To do this, let consider the regular hexagon with edges labeled as follows:



Let ρ be the rotation of an angle of degree 60°, and let σ be the reflexion around the axis passing through the points 1 and 4. With these two elements we are able to recostruct all the group: let $1 \leq i \leq 6$, then ρ^i is the rotation of an angle of degree $60°i$, and $\sigma\rho^i$ is a reflexion around one of the axis of the regular exagon. These elements are two generators of the group, and we used to write $D_{12} = \langle \rho, \sigma \rangle$. Using the labelling of the edges as above, we can identify the rotation ρ with the 6-cycle (123456), and the reflexion σ with the product of two transpositions (26)(35). Precisely,

$$\begin{aligned} \Lambda : D_{12} &\to \mathrm{Sym}(6) \\ \rho &\mapsto (123456) \\ \sigma &\mapsto (26)(35) \end{aligned}$$

is an injective group homomorphism. Hence D_{12} is a permutation group of degree 6, as required.

The previous example, even if quite easy, is crucial because it reveals which are the advantages to work with generators and permutation groups. Indeed, working with generators give the possibility to work (hopefully) with few elements of the group, and working with permutation groups allows working with permutations that are concrete objects.

Definition 3 Let G be a permutation group on Ω. The group G is transitive if for every $i, j \in \Omega$, there exists $g \in G$ such that $i^g = j$.

Hence if G is transitive there is only one orbit, namely Ω.

Example 4 Evidently $\mathrm{Sym}(n)$, and $\mathrm{Alt}(n)$ are transitive groups on $\{1, \dots, n\}$, and it is easy to show that D_{12} is transitive on $\{1, \dots, 6\}$.

Let G be an abstract group, and let H be a subgroup of G. Let $H \setminus G := \{Hg \mid g \in G\}$ be the set of right coset of H in G. Then

$$\begin{aligned} \Lambda : G &\longrightarrow \mathrm{Sym}(H \setminus G) \\ x &\longmapsto [\lambda_x : Hg \to Hgx] \end{aligned}$$

is a transitive action.

These kinds of maps are the link between the permutation group and abstract group structures. Let us explain this better. Let G be a transitive permutation group on Ω. By the Orbit-Stabiliser theorem, there is a bijection between α^G and the set of right cosets $G_\alpha \setminus G := \{G_\alpha g \mid g \in G\}$ of G_α in G.

Two G-sets Ω and Γ are isomorphic, denoted $\Omega \cong \Gamma$, if there exists a bijection $\varphi : \Omega \to \Gamma$ such that $(\alpha^x)\varphi = (\alpha\varphi)^x$ for all $\alpha \in \Omega$ and $x \in G$. For example, $\alpha^G \cong G \setminus G_\alpha$ (in terms of the natural action of G on the set of cosets $G \setminus G_\alpha$ described above). In particular, if G is transitive then $\Omega \cong G \setminus G_\alpha$. Further, for every $x \in G$, we have that $\Omega \cong G \setminus G_{\alpha^x} = G \setminus G_\alpha^x$. Hence, any transitive action of a group G is isomorphic to an action of G on the right cosets of a point stabilizer in G.

The connection is transparent: the transitive actions a group G correspond to the conjugacy classes of subgroups of G.

Let G be a permutation group on a set Ω, with orbits $\Delta_i = \alpha_i^G$, for some $\alpha_i \in \Omega$, $i \in I$. Then G acts transitively on Δ_i, and the permutation group G^{Δ_i}, induced by the action of G on Δ_i, is called *transitive constituent* of G. In some sense, G is built from its transitive constituents. Indeed, G is a subdirect product of the G^{Δ_i}'s (that is, G is a subgroup of the direct product of the G^{Δ_i}'s and the corresponding projection maps $G \to G^{\Delta_i}$ are surjective). The transitive constituents, in turn, may be built from smaller permutation groups. Here we need the notion of primitivity. Primitivity is a natural "irreducibility" condition that leads us to the basic building blocks of all permutation groups: the primitive groups. We introduce this notion in the following section.

2 Primitivity, Primitive componets, and System of Imprimitivity

In this section, it is possible to refer to [5] as the main reference.

Definition 5 Let G be a transitive permutation group on Ω. A non-empty subset B of Ω is a block of imprimitivity if, for every $g \in G$, either $B \cap B^g = \emptyset$ or $B = B^g$. Each translate B^g is also a block, and we say that $\{B^g \mid g \in G\}$ is a *block system* or a *system of imprimitivity* (this is a partition of Ω).

Let $\omega \in \Omega$. The singleton $\{\omega\} \subseteq \Omega$, and the whole Ω are blocks of imprimitivity; these are called *trivial blocks*, and any other block is *nontrivial*.

Example 6 Let $G = D_{12} = \langle \rho, \sigma \rangle$, and let $B = \{1, 4\} \subseteq \{1, \dots, 6\}$. Note that $B \cap B^\rho = \emptyset$, and $B^\sigma = B$, hence B is a nontrivial block of imprimitity.

Definition 7 Let G be a transitive permutation group on Ω. The group G is imprimitive if there exists a nontrivial block of imprimitivity. Accordingly, G is primitive if it admits only the trivial blocks.

It follows immediatly from Example 6 that D_{12} is an imprimitive group on $\{1, \dots, 6\}$.

Proposition 8 *Let $n \geq 3$. The symmetric group and the alternating group are primitive groups in their natural action on $\{1, \dots, n\}$.*

Proof. Let $B \subseteq \{1, \dots, n\}$ be a block, and let assumte that B containing at least 2 points i and j. Let $k \in \{1, \dots, n\} \setminus \{i, j\}$ (we are allowed to do so because $n \geq 3$). Let consider the 3-cycle $g = (i, j, k) \in \mathrm{Alt}(n) \leq \mathrm{Sym}(n)$. Since B is a block and $j \in B^g \cap B$, then $B^g = B$, and consequently $k \in B$. By the arbitrariness of k, it follows that $B = \{1, \dots, n\}$. We have shown that, in their natural action on $\{1, \dots, n\}$, $\mathrm{Alt}(n)$, and $\mathrm{Sym}(n)$ admit only the trivial blocks, that is $\mathrm{Alt}(n)$, and $\mathrm{Sym}(n)$ are primitive. □

The notion of primitivity, as the notion of transitivity, in permutation group theory has a correspondence with abstract group theory. The relation arises from the following easy result.

Proposition 9 *Let G be a transitive group on Ω. Then, G is primitive if, and only if, G_α is a maximal subgroup for some $\alpha \in \Omega$.*

Therefore the primitive actions correspond to conjugacy classes of maximal subgroups. We are going to define wreath products that are an important source of examples. Let $H \leq \mathrm{Sym}(\Gamma)$ and $K \leq \mathrm{Sym}(n)$ be permutation groups where $|\Gamma|, n \geq 2$. Let H^n be the direct product of n copies of H. The group K acts on H^n by permuting the coordinates. More specifically $\pi \in K$ acts on H^n by

$$(x_1, \dots, x_n)^\pi = (x_{1^{\pi^{-1}}}, \dots, x_{n^{\pi^{-1}}}).$$

The *wreath product* between H and K, denoted by $H \wr K$, is the semidirect product $H^n \rtimes K$, where the group operation is defined as follows:

$$(a_1, \dots, a_n)k \cdot (b_1, \dots, b_n)k' = (a_1, \dots, a_n)(b_1, \dots, b_n)^{k^{-1}} kk' = (a_1 b_{1^k}, ..., a_n b_{n^k})kk'.$$

The direct product H^n is called the *base group* of $H \wr K$, and K is the *top group*. There is a faithful action of $H \wr K$ on $\Omega = \Gamma \times \{1, \dots, n\}$ defined by

$$(\gamma, i)^{(h_1, \dots, h_n)k} = (\gamma^{h_i}, i^k).$$

We call this the *standard action* of G. Note that G is transitive if and only if H and K are both transitive. Also note that the partition $\{\Gamma \times \{i\} \mid 1 \leq i \leq n\}$ is a system of imprimitivity for $H \wr K$ on $\Omega = \Gamma \times \{1, \ldots, n\}$.

There is also a natural faithful action of G on the Cartesian product $\Omega = \Gamma^n$ defined by

$$(\gamma_1, \ldots, \gamma_n)^{(h_1,\ldots,h_n)k^{-1}} = ((\gamma_{1^k})^{h_{1^k}}, \ldots, (\gamma_{n^k})^{h_{n^k}}).$$

This is called the *product action* of G. Note that this is simply a combination of the coordinatewise action of H^n on Ω, together with the natural permuting action of K on coordinates.

Let $G \leq \mathrm{Sym}(\Omega)$ be an imprimitive permutation group. Let $\Sigma = \{B^g \mid g \in G\}$ be a system of imprimitivity. Note that G acts transitively on Σ, that is, the induced permutation group $G^\Sigma \leq \mathrm{Sym}(\Sigma)$ is transitive. The system of imprimitivity Σ is *maximal* if G^Σ is primitive.

Example 10 Let $G = D_{12}$. The set $\Sigma := \{\{1,4\},\{2,5\},\{3,6\}\}$ is a system of imprimitivity (see Example 6). Since $D_{12}^\Sigma \cong \mathrm{Sym}(3)$, by Proposition 8, we deduce that Σ is a maximal block system.
It is easy to show that $\Sigma' := \{\{1,3,5\},\{2,4,6\}\}$ is another maximal system of imprimitivity for G.

Let $G \leq \mathrm{Sym}(\Omega)$ be an imprimitive permutation group, and let $\Sigma_1 = \{B^g \mid g \in G\}$ be a maximal system of imprimitivity. Let $H = (G_B)^B$ be the permutation group induced on B by the setwise stabiliser G_B of B in G, and let $K_1 = G^{\Sigma_1}$. Note that $H \leq \mathrm{Sym}(B)$ is transitive and that $K_1 \leq \mathrm{Sym}(\Sigma_1)$ is primitive. There exists a bijection between Ω and $B \times \Sigma_1$ that embeds G into $H \wr K_1$ (see [5, Theorem 2.6A]). If H is imprimitive we can repeat the process, taking a maximal block system $\Sigma_2 = \{B_1^x \mid x \in H\}$ with respect to the action of H on B. Then H is isomorphic to a subgroup of $L \wr K_2$, where $L = (H_{B_1})^{B_1}$ is transitive and $K_2 = H^{\Sigma_2} \leq \mathrm{Sym}(\Sigma_2)$ is primitive, so we can embed G in $(L \wr K_2) \wr K_1$. Being Ω finite, this process terminates, at which point G is isomorphic to a subgroup of an iterated wreath product

$$((\ldots(K_t \wr K_{t-1}) \wr \ldots) \wr K_2) \wr K_1 \leq \mathrm{Sym}(((\ldots(\Sigma_t \times \Sigma_{t-1}) \times \ldots) \times \Sigma_2) \times \Sigma_1)$$

of primitive groups $K_i \leq \mathrm{Sym}(\Sigma_i)$ called *primitive components* of G (these depend form the different systems of imprimitivity considered each step, hence these are not uniquely determined by G).

From Example 10, it is clear that there might exist different systems of imprimitivity. So arise naturally to the following question.

Question 11 *Is the number of maximal blocks of imprimitivity through a point for a transitive group G of degree n bounded above polynomially in terms of degree n?*

This question was firstly asked by Cameron [3] (see [3] also for the motivation for this question.) The inspection of the problem shows how strong is the relation between permutation group theory and abstract group theory. To see this we fix some notation. Given a finite group G and a subgroup H of G, we denote by $\max(H, G)$, the number of maximal subgroups of G containing H. From what we have seen so far we can deduce that, if $G \leq \mathrm{Sym}(\Omega)$ and $\omega \in \Omega$, then there exists a one-to-one correspondence between the maximal systems of imprimitivity of G through the point ω and the maximal subgroups of G containing the point stabilizer G_ω. Hence Question 11 asks for a polynomial upper bound for $\max(G_\omega, G)$ as a function of the degree $n = |G : G_\omega|$. In [10], using this correspondece, we gave a positive solution to Question 11, showing the following result.

Theorem 12 (A. Lucchini, M. Moscatiello, P. Spiga, 2019) *There exists a constant a such that, for every finite group G and for every subgroup H of G, we have* $\max(H, G) \leq a|G : H|^{3/2}$. *In particular, a transitive permutation group of degree n has at most $an^{3/2}$ maximal systems of imprimitivity.*

3 The O'Nan-Scott theorem

In this section, we state one of the most important results in permutation group theory: the O'Nan-Scott Theorem. This theorem is a very powerful tool for studying finite primitive permutation groups, describing their structure and theri action. The theorem was stated independently by O'Nan and Scott in the preliminary proceedings of the Santa Cruz Conference on Finite Groups in 1979. Only Scott's version made it into the final Proceedings. Later, Aschbacher pointed out the existence of another class of groups erroneously excluded in the original version of the O'Nan-Scott Theorem. In [9], Liebeck, Praeger and Saxl give a self-contained proof.

To state the theorem, we introduce some classes of groups easy to describe; while, to avoid technicality, we only give a rough description or an example for remaining classes.

(I) Almost simple group. Recall that finite group T is *simple* if only the trivial subgroups (that is, the identity subgroup $\{1\}$ and whole group T) are normal in T. A permutation group $G \leq \mathrm{Sym}(\Omega)$ is *almost simple* if there exists a nonabelian simple group T such that $T \leq G \leq \mathrm{Aut}(T)$ (where $\mathrm{Aut}(T)$ is the group of the automorphisms of T).

(II) Affine-type. Let p be a prime and let V be a d-dimensional vector space over the field with p elements (which we will view as an additive group). Let

$$\mathrm{AGL}(V) = \{\varphi_{(A,v)} : u \in V \mapsto uA + v \in V \mid A \in \mathrm{GL}(V), v \in V\}$$

be the group of affine transformations of V. Note that $\mathrm{AGL}(V) \leq \mathrm{Sym}(V)$, hence it is a permutation group on V. A permutation group $G \leq \mathrm{Sym}(V)$ is of *affine-type* if $V \leq G \leq \mathrm{AGL}(V)$. Observe that the action of G is explicitly dermined by the action of $\mathrm{AGL}(V)$ on V. Here, G is primitive if and only if the stabiliser of the trivial vector $G_0 = \mathrm{GL}(V)$ acts irreducible on V (that is there are no nontrivial subspaces of V on which G_0 acts).

(III) Diagonal-type. Here we introduce only a special case with a nice description. Let T be a nonabelian simple group, and let $G = T^2$ be the direct product of two copies of T. Let consider the action of G on Ω the set of the cosets of the diagonal subgroup $H := \{(t,t) \mid t \in T\}$. Then $G \leq \mathrm{Sym}(\Omega)$ provided an of example permutation group diagonal-type.

(IV) Product-type & (V) Twisted wreath products. Just to give an idea about these classes, we can say that the permutation groups $G \leq \mathrm{Sym}(\Omega)$ of these types preserve a Cartesian structure on Ω (that is, an identification of Ω with Γ^n, the cartesian product of $n \in \mathbb{N} \setminus \{0,1\}$ copies of a set Γ).

Theorem 13 (O'Nan & Scott, 1979) *Let G be a nontrivial finite primitive permutation group on Ω. Then G is permutation isomorphic to a group that is either an almost simple, an affine-type, a diagonal-type, a product-type, or a twisted wreath product.*

Note that for the classes **(II)**,**(III)**, **(IV)**, **(V)** the action is specified precisely; but nothing is said about the action in the almost simple case. Maybe this explains why the class **(I)** is where most of the mystery resides.

In many situations, this theorem can be used to reduce a general problem to a much more specific problem concerning almost simple groups. At that point, one can appeal to the Classification of the finite simple group (CFSG) and the vast literature on simple groups and their subgroups, conjugacy classes, and representations. Therefore, being an essential tool to study the building blocks of the permutation group, the CFSG has revolutionized the study of finite permutation groups.

4 Base size

Let $G \leq \mathrm{Sym}(\Omega)$ be a permutation group of degree n. A subset $\mathcal{B}$ of Ω is a *base* for G if the pointwise stabilizer of $\mathcal{B}$ in G

$$G_{(\mathcal{B})} := \{g \in G \mid \omega^g = \omega \,, \forall \omega \in \Omega\}$$

is trivial. Clearly Ω is a base; so every permutation group has a base. The *base size* for G, denoted by $b(G,\Omega)$, or just $b(G)$ when the meaning is clear, is the minimal size of a base for G. It follows some examples.

Example 14

- Let $G = \mathrm{Sym}(n)$ acts naturally on $\Omega = \{1,\dots,n\}$. We claim that $b(\mathrm{Sym}(n)) = n-1$.

  Let $g \in G$. Since g is a permutation, if g stabilizes $n-1$ points $i_1,\dots,i_{n-1}$ in Ω, then g stabilizes $\Omega \setminus \{i_1,\dots,i_{n-1}\}$. That is, $G_{(i_1,\dots,i_{n-1})} = 1$. Hence $b(G) \leq n-1$. Let $\Gamma \subseteq \Omega$, with $|\Gamma| \leq n-2$. Hence there exist $i,j \in \Omega \setminus \Gamma$, with $i \neq j$, consequently $(i,j) \in G_{(\Gamma)} \neq \{1\}$. That is, $b(G) \geq n-1$. Definitely $b(G) = n-1$.
- Let $G = \mathrm{Alt}(n)$ acts naturally on $\Omega = \{1,\dots,n\}$. Arguing similarly to the previous case it is possible to show that $b(\mathrm{Alt}(n)) = n-2$.

Let $G \leq \mathrm{Sym}(\Omega)$ be a permutation group, and $\omega \in \Omega$. The orbit ω^G is *regular* if $|\omega^G| = |G|$.

By the Orbit-Stabilizer theorem we have $|G : G_\omega| = |\omega^G|$. Hence, ω^G is regular if and only if $G_\omega = \{1\}$. That is,

$$G \text{ has a regular orbit if and only if } b(G) = 1. \tag{2}$$

Example 15

(1) Let $G = D_{2n}$ be the group of the symmetries of a regulan n-gon. As we saw for the particular case $n = 6$, G acts faithfully on $\{1, \ldots, n\}$ the set of the vertices of the regulan n-gon. We claim that $b(G) = 2$. Let $i \in \{1, \ldots, n\}$, and let $\mathcal{B} = \{i, i + 1\}$. A reflexion with axis a diagonal passing through the point i and its diagonally opposit moves the point $i + 1$ (since i and $i + 1$ are not diagonally opposit), and a nontrivial rotation has no fixed points. Hence $G_{(i,i+1)} = \{1\}$. Being $|D_{2n}| = 2n$, and $G \leq \mathrm{Sym}(n)$, there are no regular orbits, and so $b(D_{2n}) = 2$.

(2) Let V be a d-dimesional vector space, and let $G = \mathrm{GL}(V)$. Note that G acts in a natuar way on V. We claim that $b(G) = d$.

First we are going to show that for any A subset of V with size $d-1$, then $G_{(A)} \neq \{1\}$. Let $u_1, \ldots, u_{d-1} \in V$, and let $g \in G_{(\{u_1,\ldots,u_{d-1}\})}$. Observe that since $u_i g = u_i$, then $ug = u$ for every $u \in \langle u_i \rangle$ (where $\langle u_i \rangle$ is the subspaces of V generated by u_i). Thus, we can assume, without loss of generality, that $u_1, \ldots, u_{d-1}$ are indepentent vectors. Let $w \in V \setminus \langle u_1, \ldots, u_{d-1} \rangle$. Then $\{u_1, \ldots, u_{d-1}, w\}$, and $\{u_1, \ldots, u_{d-1}, w + u_1\}$ are bases of the vector space V. Consequently the bijective map f that fixes the vectors $u_1, \ldots, u_{d-1}$ and that swaps the vectors w and $w + u_1$ is a nonidentity element of $G_{(\{u_1,\ldots,u_{d-1}\})}$. That is, we proved that $b(G) \geq d$. Let $\{v_1, \ldots, v_d\}$ a basis of the vector space V. It is not difficult to prove that $G_{(\{v_1,\ldots,v_d\})} = \{1\}$. Hence $b(G) \leq d$, and the claim follows.

(3) Let $G = \mathrm{Sym}(n)$ acts on Ω the set of the 2-elements subsets of $\{1, \ldots, n\}$. We claim that $b(G, \Omega) \leq \frac{2}{3}n + 1$.

Let $n = 3s + t$, with $0 \leq t \leq 2$. Let

$$\mathcal{B} = \begin{cases} \{\{1,2\}, \{2,3\}, \ldots, \{3s-1, 3s\}\}, & \text{if } 0 \leq t \leq 1 \\ \{\{1,2\}, \{2,3\}, \ldots, \{3s-1, 3s\}, \{3s, 3s+1\}\}, & \text{if } t = 2. \end{cases}$$

It is easy to deduce that $G_{\mathcal{B}} = \{1\}$, and so the claim follows.

From the following result (proved in [6]), we deduce that the base described in (3) of Example 15 gives a good estimation for the base size for the action of $G = \mathrm{Sym}(n)$ on the 2-elements subsets of $\{1, \ldots, n\}$, provided that $n \geq 4$.

Theorem 16 (Z. Halasi, 2012) *Let $G = \mathrm{Sym}(n)$, and let Ω be the set of k-elements subsets of $\{1, \ldots, n\}$. If $n \geq k^2$, then $b(G, \Omega) \leq \lceil \frac{2n-2}{k+1} \rceil$*

Definition 17 Let $G \leq \mathrm{Sym}(n)$ be a permutation group of degree n. The group G is *large base* if there exist some ℓ and $r \geq 1$ such that $(\mathrm{Alt}(\ell))^r \leq G \leq \mathrm{Sym}(\ell) \wr \mathrm{Sym}(r)$, where the action of $\mathrm{Sym}(\ell)$ is on k-sets elements of $\{1, \dots, \ell\}$, and the wreath product acts with product action.

We can say that the large base groups arise as "blows-up" of almost simple groups with socle $\mathrm{Alt}(\ell)$ acting on k-sets elements of $\{1, \dots, \ell\}$. Note that, taking $r = 1$ and $\ell = n$, the groups considered in Theorem 16 are examples of large base.

The concept of base arise naturally in others contexts of mathematics. For example, in the graph-theoretic literature, if Γ is a graph with automorphism group $G = \mathrm{Aut}(\Gamma)$, then $b(G)$ is called the fixing number (also determining number or rigidity index) of Γ and this is a well-studied graph invariant (see [1] and the references therein).

Further, some classical problems in the representation theory of groups can also be stated in terms of bases. For instance, if H is a group and V is a faithful H-module, then H has a regular orbit on V if and only if the corresponding affine group $G = VH \leq \mathrm{AGL}(V)$ admits a base of size 2. Indeed, from the simple observation written in (2), and from the fact that $b(G) = b(H) + 1$, we deduce that G has a regular orbit if and only if $b(G) = 2$.

In a different direction, bases have been used extensively in the computational study of finite permutation groups. We briefly described this. Let $G \leq \mathrm{Sym}(\Omega)$ be a permutation group, let $\mathcal{B}$ be a base, and let $x, y \in G$. Note that $\alpha^x = \alpha^y$, for every $\alpha \in \mathcal{B}$, if and only if $xy^{-1} \in G_{(\mathcal{B})} = 1$. That is, the elements of G are completely determined by their action on $\mathcal{B}$. Consequently $|G| \leq |\Omega|^{|\mathcal{B}|}$, and in particular,

$$|G| \leq |\Omega|^{b(G)}. \tag{3}$$

In computational group theory, it is used to store the elements of G as $|\Omega|$-tuples. Now, it follows from Example 10 that the element of G can be store as $b(G)$-tuples. Clearly, it is more convenient to store these elements as $b(G)$-tuples, rather than $|\Omega|$-tuples; whence the problem of calculating base sizes has important practical application (see [14, Chapter 4] for further details).

Moreover, Example 10 reveals that it is possible to find an upper bound on the order of a permutation group by bounding the minimal base size. The problem of bounding the order of a finite primitive permutation group attracted a lot of attention in the 19th century. One of the earliest results in this direction is a theorem of Bochert [2] from 1889, which states that if G is a primitive permutation group of degree n not containing the alternating group $\mathrm{Alt}(n)$, then $b(G) \leq n/2$.

Using the Classification of Finite Simple Groups (CFSG) Liebeck proved the following remarkable result.

Theorem 18 (Liebeck [8]) *Let G be a primitive permutation group of degree n. If G is not large base, then $b(G) \leq 9 \log n$.*

The previous theorem provided a strong example of how much it is useful the tools of CFSG in the study of permutation groups.

More recently, Liebeck, Halasi and Maroti showed in [7] that for almost all non-large base primitive groups, $b(G) \leq 2\lceil \log n \rceil + 26$; Roney-Dougal and Siccha then noted in [12] that this bound applies to all primitive groups that are not large base. A conclusive result (to apper in [11]) in this direction is the following.

Theorem 19 (M. Moscatiello, C.M. Roney-Dougal, 2020^+) *Let G be a primitive permutation group of degree n. If G is not large base then $b(G) \leq \max\{7, \lceil \log n \rceil + 1\}$. Furthermore, there are infinitely many such groups G for which $b(G) > \log n + 1$.*

Notice that if G is the largest Mathieu group M_{24} in its 5-transitive action of degree 24 then $b(G) = 7 > \lceil \log n \rceil + 1$. If $G = \mathrm{Sp}_{2m}(2)$, acting on the cosets of the maximal subgroup $\mathrm{O}^-_{2m}(2)$, then $b(G) = \lceil \log n \rceil + 1 > \log n + 1$. Hence, Theorem 18 is the best possible in this context.

References

[1] R.F. Bailey and P.J. Cameron, *Base size, metric dimension and other invariants of groups and graphs.* Bull. London Math. Soc. 43 (2011), 209–242.

[2] A. Bochert, *Uber die Zahl verschiedener Werte, die eine Funktion gegebener Buchstaben durch Vertauschung derselben erlangen kann.* Math. Ann. 33 (1889), 584–590.

[3] P.J. Cameron, `https://cameroncounts.wordpress.com`.

[4] I.M. Isaacs, "Finite Group Theory". American Mathematical Soc., 2008.

[5] J.D. Dixon and B. Mortimer, "Permutation Groups". Springer-Verlag, New York, 1996.

[6] Z. Halasi, *On the base size for the symmetric group acting on subsets.* Article in Studia Scientiarum Mathematicarum Hungarica 49/4 (2012), 492–500.

[7] Z. Halasi, M.W. Liebeck, A. Maroti, *Base sizes of primitive groups: bounds with explicit constants.* Preprint, `https://arxiv.org/abs/1802.06972`.

[8] M.W. Liebeck, *On minimal degrees and base sizes of primitive permutation groups.* Arch. Math. (Basel) 43/1 (1984), 11–15.

[9] M.W. Liebeck, C.E. Praeger, J. Saxl, *On the O'Nan-Scott theorem for finite primitive permutation groups.* J. Australian Math. Soc. (A) 44 (1988), 389–396.

[10] A. Lucchini, M. Moscatiello and P. Spiga, "Bounding the maximal size of independent generating sets of finite group". Proc. A Royal Soc. Edinburgh (2019).

[11] M. Moscatiello and C.M. Roney-Dougal, *Base size of primitive permutation groups.* In preparation.

[12] C.M. Roney-Dougal and S. Siccha, *Normalisers of primitive permutation groups in quasipolynomial time.* Bull. London Math. Soc., to appear.

[13] L.L. Scott, *Representations in characteristic p.* The Santa Cruz Conference on Finite Groups, Proceedings of Symposia in Pure Mathematics, vol. 37 (1980), 319–331.

[14] A. Seress, "Permutation group algorithms". Cambridge Tracts in Mathematics 152, Cambridge University Press, 2003.

Computational problems in mathematical physical modelling with DLTI systems

Marta Gatto [*]

Abstract. Mathematical physical models are often used for the description of physical phenomena and are essential in industrial applications for various aims, such as control and estimation of unmeasurable variables and physical parameters. In this talk, some numerical methods at the basis of experimental modelling, i.e. modelling through experimental data, will be described for different kind of model classes, in particular for DLTI (Dynamic Linear Time Invariant) systems. The reasons for their importance will be explained and the computational problems of parameter estimation and data denoising subject to the DLTI model constraint will be introduced with examples.

1 Introduction

Mathematical physical models are often used for the description of physical phenomena and are essential in industrial applications for various aims, such as control and estimation of unmeasurable variables and physical parameters.

1.1 Some examples of industrial applications

The uses of models in industrial applications are very different and are referred to with various expressions: predictive and adaptive control, model-based design and shape optimization, indirect measures and virtual sensors, fault detection and predictive maintenance, virtual testing, parameter estimation, digital twins and prototypes and lot of others. We describe for example some of these cases:

(a) Shape Optimization and Model-Based Design, are terms referred to the case in which an analytical and numerical study on the shape of a certain object is done through the optimization of mathematical equations to obtain a certain aim before building an object or a machine;

(b) Indirect measures and Virtual Sensors are terms that indicate the calculation of a certain quantity through physical equations, when it is not possible to measure it

[*] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: mgatto@math.unipd.it . Seminar held on 20 May 2020.

with true sensors, for example when the place has a too high temperature or the sensor is expensive;

(c) The monitoring of the good functioning of a machine can be checked through models and this is referred to as Fault Detection and Predictive Maintenance;

(d) In general the simulation of a model that describes a machine is called Virtual Testing, and it may be useful when the tests on the machine are expensive, dangerous, or when the machine is not available (suppose for example that workers are forced to smart-working at home for a certain time).

1.2 Some basics of Experimental Physical Modelling

As various the applications in which models are used, as various are the mathematical models that can be chosen. The activity of getting a model from data is called with different terms, depending on the approach and the techniques used. Some examples are *System Identification*, *Signal Estimation or enhancement* [2, 1], and *Experimental Modelling*, and indicate the techniques used to extract the useful information of the system from the available measurement corrupted with noise. The structure of these procedures has some basic common concepts that can be pointed out. The process starts with the collection of some data of the system under study, that comes with a measurement error. Sometimes it is also possible to choose how to take the data of the system (and this step is called Experiment Design). Then, some physical information about the system and relations among the measured (and also unmeasured) quantities can be collected. Among these information only the ones useful to the aim of the problem must be considered.

The procedure can in general be divided in three parts:

(a) in each case we can define a *criterion function* to be minimized, that represents our aim. For example deterministic (square error, absolute error, ...) or probabilistic (maximum likelihood error, ...).

(b) secondly, a *model* structure must be chosen to describe the system: the range goes from simpler to more complex. We can divide them in the following characterizations: linear or nonlinear, static or dynamic, with lumped or distributed parameters (Differential Equations or Partial Differential Equations), dynamic with continuous or discrete time. Another distinction is the one between White Box models (i.e. governed by physical equations, in which parameters have physical meaning) and Black Box ones (or data-driven, in which parameters do not have physical meaning).

(c) finally we have to choose the numerical *algorithm* to solve the minimization of the criterion function (for example batch or recursive methods).

The choice of the model structure can be done following the main properties of a good model:

- generality: is the model still able to describe the system for little modifications of the setting? For example the Hook's law that describe a spring is valid for springs with different stiffness factors;

- predictability: is the model able to describe the phenomena also in situations that were not used in the creation of the model?
- simplicity: this is a really important point and can be summarized with the Occam's Razor principle, and described by the quotation: "Everything should be made as simple as possible, but not simpler", Albert Einstein. This means that in the description of the model, only useful information must be taken into consideration, while non-useful details of the phenomena must be neglected.

The last one is a very important principle in modelling, and is the reason why simple model equations are the most widespread and we will focus on them. We will show two examples, Least Squares for Linear Systems and the Nonlinear case, following the three-step structure described above. Consequently, we will introduce the State-Space form and the Discrete Linear Time Invariant systems (DLTI).

2 Least-Squares and Nonlinear Least-Squares, static and dynamic

2.1 Least-Squares

The easiest, and hoped-for, *model* is the linear one, of the form

$$Ax = y$$

where x is the unknown variable, A is the matrix built from data and the right hand side y is a measured quantity. When the matrix A is full-rank, the solution is easily given by $x = A^{-1}b$, since the inverse can be computed. However, in general the system is built from lots of noisy data that generate a matrix A with more rows than columns, and y does not belong to $Im(A)$. Hence, what we look for is the vector x that gives the minimum norm of the error, i.e. such that Ax is the projection of b on the subspace generated by the columns of A. In this way we obtain the least-squares formulation (*criterion function*)

$$\min_x \|Ax - b\|_2^2.$$

We may want to solve this problem on a batch of data (offline mode), or in a recursive way (online), i.e. updating the system and solving it whenever a new data is available. These two problems lead to different numerical methods or *algorithms*.

Moreover, the numerical problem gets more complicated when the matrix A is nearly singular: these problems are called *ill-posed* and regularization techniques must be considered. For example, the Tikhonov regularization consists in solving the problem

$$\min_x \|Ax - b\|_2^2 + \lambda\|Lx\|_2^2$$

for a certain value of the weight λ and for a certain matrix L, that is usually the identity or the discretization of a derivative operator, usually of order 2.

2.2 Examples of Least-Squares

Although the linear model is very simple, it is sufficient to describe lots of different situations.

Example of a static model: One of the basic examples is the *polynomial regression or fitting*, in which the aim is to describe the quantity y as a polynomial in the variables u. In Figure 1 an example of a second order polynomial fitting is shown. Given measurements at samples of time $t_0, \dots, t_N$, we can build the linear system

$$\begin{bmatrix} 1 & u(t_0) & u(t_0)^2 \\ 1 & u(t_1) & u(t_1)^2 \\ \vdots & \vdots & \vdots \\ 1 & u(t_N) & u(t_N)^2 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} y(t_0) \\ y(t_1) \\ \vdots \\ y(t_N) \end{bmatrix}.$$

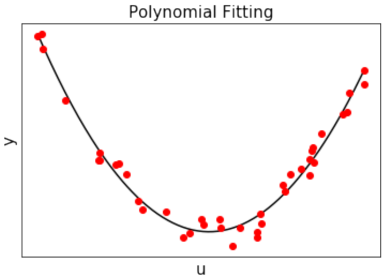


Figure 1. Polynomial fitting

This problem can be written as a linear least squares problem, in which the unknown x are the weights of each term of the polynomial.

Example with a dynamic model: Given the mechanical equation of the motor

$$J_M \frac{d\omega}{dt}(t) + B_M \omega(t) = T_M(t) - T_L(t),$$

suppose to know the measurements of $T_M, T_L, \omega, \frac{d\omega}{dt}$ and to need the estimate of the parameters J_M and B_M. We can write the problem as a linear system as follows, and solve it as a least-squares problem

$$\begin{bmatrix} \frac{d\omega}{dt}(t_0) & \omega(t_0) \\ \frac{d\omega}{dt}(t_1) & \omega(t_1) \\ \vdots & \vdots \\ \frac{d\omega}{dt}(t_N) & \omega(t_N) \end{bmatrix} \begin{bmatrix} J_M \\ B_M \end{bmatrix} = \begin{bmatrix} T_M(t_0) - T_L(t_0) \\ T_M(t_1) - T_L(t_1) \\ \vdots \\ T_M(t_N) - T_L(t_N) \end{bmatrix}.$$

2.3 Nonlinear Least-Squares

When a linear model is not sufficient to describe the system at hand, we must consider a nonlinear *model* of the kind

$$y = f(u, x)$$

where y and u are vectors of measured quantities of which we know the physical relation described by the nonlinear function f and we want to estimate a set of parameters x. The *criterion* to find the solution is the generalization of the linear case, called *nonlinear least squares*. It consists in minimizing the norm of the error between the measured vector

y^{meas} and the vector y estimated through the model. For this reason it is also called the *prediction error method* (PEM), and consists in the following problem

$$\min_x \|f(u,x) - y^{meas}\|_2^2 \quad = \quad \begin{array}{ll} \displaystyle\min_x & \|y - y^{meas}\|_2^2 \\ \text{s.t.} & y = f(u,x) \end{array} \quad = \quad \begin{array}{ll} \displaystyle\min_x & \displaystyle\sum_{i=0}^{N} (y_i - y_i^{meas})^2. \\ \text{s.t.} & y_i = f(u, x_i) \quad \forall i \end{array}$$

The optimization methods that can be used to solve this problem are various, for example Gauss-Newton and Levenberg-Marquardt methods. Note that the resolution of a nonlinear optimization is iterative and computationally more difficult than the solution of the linear case. Moreover the problem of local minima may require the use of global optimization algorithms.

2.4 Examples of Nonlinear Least-Squares

We give two examples of static and dynamic models, as in the linear case.

Example of a static model: In the static case we can consider a simple example of Nonlinear exponential fitting, of which an example is shown in Figure 2.

Given two quantities y and u, we want to describe the variable y with the exponential function

$$y = f(u) = x_1 e^{x_2 u}$$

estimating the value of the parameter vector $x = [x_1, x_2]$. Given measurements u_i^{meas}, y_i^{meas} at samples of time $t_0, \ldots, t_N$, we can build the nonlinear least squares problem

$$\min_{x=[x_1,x_2]} \sum_{i=0}^{N} (y(t_i) - y_i^{meas})^2$$
$$\text{s.t. } y(t_i) = x_1 e^{x_2 u_i^{meas}} \qquad \text{for } t_i = t_0, \ldots, t_N.$$

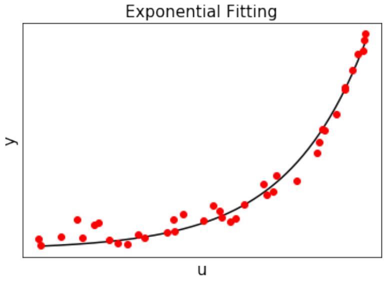


Figure 2. Exponential fitting

Example of a dynamic model: The Lorentz model is nonlinear with respect to the parameters p, r (Prandtl and Rayleigh numbers). Calling $y = [y(0)^T y(1)^T \ldots]$ with $y(k) = [y_1(k), y_2(k), y_3(k)]$, we solve

$$\underset{p,r}{\text{minimize}} \ \|y - y^{meas}\|_2^2$$
$$s.t. \ \begin{cases} \frac{dy_1}{dt}(t) & = -py_1(t) + py_2(t), \\ \frac{dy_2}{dt}(t) & = (r - y_3(t))y_1(t) - y_2(t), \\ \frac{dy_3}{dt}(t) & = y_1(t)y_2(t) - by_3(t). \end{cases}$$

This system of equations has a peculiarity, in fact for high values of the Rayleigh number, the system is near to chaotic, i.e. for little perturbations of the initial condition, there are big variations in the dynamic. Numerical methods that take into account this aspect must be used for the estimation of the initial condition in this case.

3 State-Space modelling

Until now, we considered models in which there were two sets of variables, an independent variable vector u and a dependent vector y. These kinds of models are called *Input-Output*.

A State-Space model is characterized by the presence of three main variables, not only the input vector u and the output y, but also the *state vector* x. Usually, the state is also called internal or hidden variable, because some of its components may be not measurable. More precisely, the *state* of a system at time t is the minimum set of variables that, with the input, is sufficient to uniquely specify the dynamic system behaviour for all t over the interval $t \in [t, \infty)$.

The continuous case has the following structures for the nonlinear and linear cases:
Continuous Nonlinear/Linear Time State-Space Models

$$\text{Nonlinear:} \quad \begin{cases} \dot{x}(t) = A_c(t)x(t) + B_c(t)u(t) \\ y(t) = C_c(t)x(t) + D_c(t)u(t) \end{cases} \qquad \text{Linear:} \quad \begin{cases} \dot{x}(t) = a(t, x(t)) + b(t, u(t)) \\ y(t) = c(t, x(t)) + d(t, u(t)) \end{cases}$$

with $x \in \mathbb{R}^{n_x}$, $y \in \mathbb{R}^{n_y}$, $u \in \mathbb{R}^{n_u}$, $A_c \in \mathbb{R}^{n_x \times n_x}$, $B_c \in \mathbb{R}^{n_x \times n_u}$, $C_c \in \mathbb{R}^{n_y \times n_x}$ and $D_c \in \mathbb{R}^{n_y \times n_u}$.

Some systems may arise directly in a discrete form, or the continuous ones can be discretized to obtain the following discrete equations:
Discrete Linear State-Space Model

$$\begin{cases} x(k+1) &= A(k)x(k) + B(k)u(k) \\ y(k) &= C(k)x(k) + D(k)u(k) \end{cases} \tag{1}$$

and the simpler case, that we are interested in:
Discrete Linear Time-Invariant State-Space (DLTI) Models

$$\begin{cases} x(k+1) &= Ax(k) + Bu(k) \\ y(k) &= Cx(k) + Du(k). \end{cases} \tag{2}$$

3.1 Properties of DLTI Models

The importance and extensive use of DLTI models is due to the following aspects:

- DLTI models are a common structure for lots of physical phenomena, i.e. very different systems (from mechanical to electrical, thermal, ...) can be described with the same mathematical structure of equations, with different meaning of the states and variables. This concept is called system analogy. The consequence is that, from the mathematical point of view, the theory and the dynamics of these systems can be studied independently on the application.

- The other characterization is that system theory is well developed, a lot of properties can be characterized and, in the deterministic case, the solution is known explicitly.
- In the linear case, linear algebra problems can be solved online, with small computational effort.

3.1.1 Properties: Controllability and Observability

Controllability, Reachability and Observability are properties of a system important in Control Theory. When the model is used to control a certain system it is important to know how the input can influence the dynamic. We recall here the definitions [6].

Definition 1 (Controllability) The DLTI system (2) is controllable if, given any initial state $x(k_a)$, there exists an input signal $u(k)$ for $k_a \leq k \leq k_b$ such that $x(k_b) = 0$ for some k_b.

Definition 2 (Reachability) The DLTI system (2) is *reachable* if for any two states x_a and x_b there exists an input signal $u(k)$ for $k_a \leq k \leq k_b$ that will transfer the system from the state $x(k_a) = x_a$ to $x(k_b) = x_b$.

In few words, controllability means that the system, through a certain input, can always be brought to the origin, and reachability that the state can always be moved from one point to another. Observability is the possibility to deduce univocally the state from the output measurement:

Definition 3 (Observability) The DLTI system (2) is *observable* if any initial state $x(k_a)$ is uniquely determined by the corresponding zero-input response $y(k)$ for $k_a \leq k \leq k_b$ with k_b finite.

It is easy to check these properties on DLTI systems, since two theorems hold:

- reachability is equivalent to the matrix $C_n = \begin{bmatrix} B & AB & \dots & A^{n-1}B \end{bmatrix}$ to be full rank,
- observability is equivalent to the matrix $O_n = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}$ to be full rank.

3.1.2 Explicit formula of the Solutions

The solution equations for the State-Space problems, given the initial conditions on the state, have analytic expressions

- in the *continuous* deterministic LTI case is:

$$
\begin{cases} x_t = \Phi_{t,t_0} x_{t_0} + \int_{t_0}^{t} \Phi_{t,\alpha} B_c u_\alpha d\alpha \\ y_t = C_c \Phi_{t,t_0} x_{t_0} + \int_{t_0}^{t} C_c \Phi_{t,\alpha} B_c u_\alpha d\alpha + D_c u_t \end{cases} \tag{3}
$$

  with $\Phi_{t,t_0} = e^{A_c(t-t_0)}$.

- in the *discrete* deterministic linear case is:

$$
\begin{cases} x(t) = \Phi(t,0)x(0) + \sum_{k=0}^{t-1} \Phi(t,k)B(k)u(k) \\ y(t) = C(t)\Phi(t,0)x(0) + \sum_{k=0}^{t-1} C(t)\Phi(t,k)B(k)u(k) + D(t)u(t) \end{cases} \tag{4}
$$

  with $\Phi(t,k) = A(t-1)\,A(t-2)\,A(t-3)\cdots A(k)$ for $t > k$, for a general discrete linear time variant state-space; while $\Phi(t,k) = A^{t-k}$ for $t > k$ for a DLTI state-space.

4 Computational Problems of DLTI systems with noise

All the measures we obtain from real systems are corrupted with noise. The term "model-based" processing was used in literature [1, 2] to represent the introduction of the description of noise inside the description of the system.

We introduce some basic probability concepts to add some noise terms in the state space models considered until now.

4.1 Probability Preliminaries

Consider a *probability space* $(\Omega, \mathcal{S}, P)$, i.e. the triplet with Ω the sample space, $\mathcal{S}$ a σ-algebra on it, is the set of events, and P a probability measure, i.e. a measure with $P(\Omega) = 1$. Then a *random variable* with values in $\mathbb{R}^n$ is a measurable mapping $X : \Omega \to \mathbb{R}^n$, $X(\omega) = x$ for $\omega \in \Omega$. This generates a probability measure $P_X(B) = P(X^{-1}(B))$ with $B \in \mathcal{B}$ borel sets of $\mathbb{R}$. Moreover, the *cumulative probability distribution function* is defined by

$$F_X(x_i) = Pr(X(\omega_i) \leq x_i)$$

called *probability distribution of* X or *probability mass function.*

Definition 4 (Cumulative distribution function) The cumulative distribution function (CDF) $F_X(\alpha)$ of a random variable X yields the probability of the event $\{X \leq \alpha\}$, which is denoted by

$$F_X(\alpha) = P[X \leq \alpha], \quad \text{for} \quad -\infty < \alpha < \infty.$$

Definition 5 (Probability density function) The probability density function (PDF) $f_X(\alpha)$ of a random variable X, if it exists, is equal to the derivative of the cumulative distribution function $F_X(\alpha)$, which is denoted by

$$f_X(\alpha) = \frac{dF_X(\alpha)}{d\alpha}.$$

We recall the following definitions:

$$\begin{aligned}\text{Expected value (mean):}\quad & m_X \quad = \mathbb{E}[X] = \int_{-\infty}^{\infty} \alpha f_X(\alpha) d\alpha, \\ \text{Cross-Correlation:}\quad & C_{XY} = \mathbb{E}[X\,Y], \\ \text{Covariance:}\quad & R_{XY} = \mathbb{E}[(X - m_X)(Y - m_Y)].\end{aligned}$$

We will need the following particular case of random signal:

Definition 6 (A Gaussian (or normal) random variable) is defined by its probability density function:

$$f_{x(t)}(\alpha) = \frac{1}{\sqrt{2\pi R_{xx}}} exp\left\{-\frac{1}{2}\frac{(\alpha - m_x)^2}{R_{xx}}\right\} \quad \text{i.e.} \quad x \sim \mathcal{N}(m_x, R_{xx})$$

with $m_x \in \mathbb{R}$ and $R_{xx} \in \mathbb{R}^+$. **White noise** is a Gaussian process with zero mean.

Vectorial case The definitions of mean and covariance for the case of two random variables can be extended to the vector case. Let X be a vector with entries X_i for $i = 1, 2, \ldots, n$ that jointly have a Gaussian distribution with mean equal to:

$$m_X = \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{bmatrix} \quad \text{and covariance matrix } C_X \text{ equal to} \quad C_X = \begin{bmatrix} C_{X_1,X_1} & \ldots & C_{X_1,X_n} \\ C_{X_2,X_1} & & C_{X_2,X_n} \\ \vdots & \ddots & \vdots \\ C_{X_n,X_1} & \ldots & C_{X_n,X_n} \end{bmatrix}$$

then the joint probability density function is given by

$$f_X(\alpha) = f_{X_1,X_2,\ldots,X_n}(\alpha_1, \alpha_2, \ldots, \alpha_n) = \frac{1}{(2\pi)^{n/2}\det(C_X)^{1/2}} exp\{\frac{1}{2}(\alpha - m_X)^T C_X^{-1}(\alpha - m_X)\}$$

where α is a vector with entries α_i, $i = 1, 2, \ldots, n$.

Time dependence A *Random Signal* or *stochastic process* can be seen as a sequence of ordered in time random variables, in fact if we add the dependence on time (continuous or discrete) we obtain the function $X : \mathcal{I} \times \Omega \to \mathbb{R}$ with $(k, \omega) \mapsto X(k, \omega)$. Fixed each distinct value of time we obtain again a random variable $X(k) : \Omega \to \mathbb{R}$. Each entry of the discrete-time vector random signal $x(k, \omega)$ for a fixed k, is a random variable.

We recall some useful definitions we will use from now on: given x, y two random signals, we define

$$
\begin{aligned}
\text{Expected value (mean):}\quad & m_x(t) \quad = \mathbb{E}[x(t)] = \int_{-\infty}^{\infty} \alpha f_X(\alpha) d\alpha, \\
\text{Auto-Correlation:}\quad & \psi_{xx}(t,k) = \mathbb{E}[x(t)x(k)], \\
\text{Cross-Correlation:}\quad & C_{xy}(t,k) = \mathbb{E}[x(t)y(k)], \\
\text{Variance:}\quad & R_{xx}(t,k) = \mathbb{E}[(x(t) - m_x(t))(x(k) - m_x(k))], \\
\text{Covariance:}\quad & R_{xy}(t,k) = \mathbb{E}[(x(t) - m_x(t))(y(k) - m_y(k))].
\end{aligned}
$$

4.2 DLTI models with process and measure noise

We can now introduce noise in the DLTI model already described. The most common formulation is the one in which some error in the model and in the output measure y is added, while the input measure u is supposed exact, with no noise.

Definition 7 (DLTI State-Space Models with process and measure noise)

$$
\begin{cases} x(k+1) = Ax(k) + Bu(k) + v(k) \\ y(k) = Cx(k) + Du(k) + w(k) \end{cases} \quad \text{with} \quad \begin{cases} v(k) \sim \mathcal{N}(0, R_{vv}) \\ w(k) \sim \mathcal{N}(0, R_{ww}) \end{cases} \tag{5}
$$

where $v(k)$ is the process or model noise and $w(k)$ it the measurement noise, both gaussian white.

4.2.1 Computational Problems with DLTI models

We can consider different computational problems, depending on what are the known and unknown variables of the system:

- estimation of the unmeasurable variable x, solved through the Kalman Filter,
- estimation of physical parameters (i.e. estimation of the model) and unmeasurable variable x, for which two approaches can be considered, the Nonlinear least-squares method and Subspace Methods,
- estimation of Input/Output data, i.e. denoising or filtering of Input/Output data, that brings to the modified Kalman Filter.

4.3 State Estimation: the Kalman Filter

The problem we want to solve is to recover a state estimate of (5) using both the information of the model and the measurements. Since models are inaccurate and sensors have errors, we must take into consideration the noise values. The Kalman Filter gives us the right way (in a sense we will define) to weight both the information, i.e. gives us the weights for the combination of the two information. It is a very famous algorithm since its successful use in the navigation systems for the Apollo mission. From its discover, the applications for which it has been used cover different fields like signal processing, voice recognition, video stabilization, and automotive, control, global positioning, computer vision and lots more. Moreover, various generalizations of this method have been studied and are still open problems.

The term *filtering* is referred to the fact that it is an online algorithm, i.e. it works one discrete instant at a time, in contrast to the offline procedures, which are called *smoothing*. More precisely, at each time instant we have the knowledge of an estimate of the state at the actual moment k, calculated with the model, and the measurements up to the previous time instant.

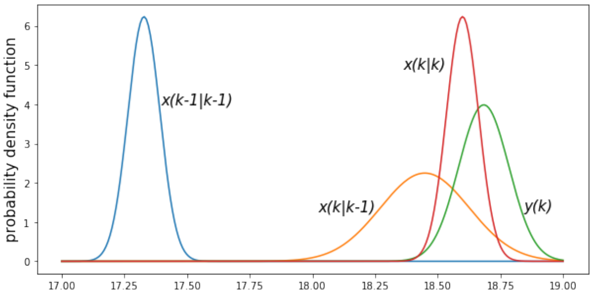


Figure 3. Kalman Filter iteration of predictor-corrector form

In Figure 3 the prediction-correction procedure is shown.

Let us define the Kalman Problem more precisely.

Problem 1 (Kalman Filter Problem) *We are given the signal-generation model*

$$\begin{cases} x(k+1) = A(k)x(k) + B(k)u(k) + w(k) \\ y(k) = C(k)x(k) + D(k)u(k) + v(k) \end{cases} \tag{6}$$

with the process noise $w(k)$ and measurement noise $v(k)$ assumed to be zero mean white-noise sequences with joint covariance matrix

$$E\left[\begin{bmatrix} v(k) \\ w(k) \end{bmatrix} \begin{bmatrix} v(j)^T & w(j)^T \end{bmatrix}^T\right] = \begin{bmatrix} R_{vv}(k) & R_{wv}(k)^T \\ R_{wv}(k) & R_{ww}(k) \end{bmatrix} \Delta(k-j) \geq 0$$

with $R_{vv}(k) > 0$ and where $\Delta(k)$ is the unit pulse. At time instant $k-1$, we have an estimate of $x(k)$, which is denoted by $\hat{x}(k|k-1)$ with properties

$$E[x(k)] = E[\hat{x}(k|k-1)],$$

$$E[(x(k) - \hat{x}(k|k-1))(x(k) - \hat{x}(k|k-1))^T] = P(k|k-1) \geq 0.$$

This estimate is uncorrelated with the noise $w(k)$ and $v(k)$. The problem is to determine a linear estimate of $x(k)$ and $x(k+1)$ based on the given data $u(k)$, $y(k)$, and $\hat{x}(k|k-1)$, such that both estimates are minimum variance unbiased estimates; that is, estimates with the properties

$$E[\hat{x}(k|k)] = E[x(k)], \quad E[\hat{x}(k+1|k)] = E[x(k+1)], \tag{7}$$

and the expressions below are minimal:

$$(8)\quad E[(\hat{x}(k)-\hat{x}(k|k))(x(k)-\hat{x}(k|k))^T], \quad E[(x(k+1)-\hat{x}(k+1|k))(x(k+1)-\hat{x}(k+1|k))^T].$$

Note that, with the condition of the mean, for the unbiased estimate, asking for the minimum variance of the error is equivalent to asking the minimum cross-correlation of the error, and the minimum variance of the solution. The conditions on means and variances of the errors give the right Kalman weights for the combination of the measurements.

There are lots of formulations and derivation of this result, we will present here the "conventional Kalman filter".

Predictor-Corrector form:

$$\begin{cases} \hat{x}(k|k-1) & = A(k-1)\hat{x}(k-1|k-1) + B(k-1)u(k-1) \\ \tilde{P}(k|k-1) & = A(k-1)\tilde{P}P(k-1|k-1)A'(k-1) + R_{ww}(k-1) \end{cases} \qquad \textbf{predictor}$$

$$\begin{cases} e(k) & = y(k) - \hat{y}(k|k-1) = y(k) - C(k)\hat{x}(k|k-1) \\ R_{ee}(k) & = C(k)\tilde{P}(k|k-1)C'(k) + Rvv(k) \end{cases}$$

$$K(k) = \tilde{P}(k|k-1)C'(k)R_{ee}^{-1}(k)$$

$$\begin{cases} \hat{x}(k|k) & = \hat{x}(k|k-1) + \mathbf{K}(k)e(k) \\ \tilde{P}(k|k) & = [I - K(k)C(k)]\tilde{P}(k|k-1) \end{cases} \qquad \textbf{corrector}$$

$$\hat{x}(0|0), \tilde{P}(0|0)$$

where $K(k)$ is called *Kalman gain or weight*, that gives us exactly that correct combination of the two information we started with. The equations of the algorithm descend directly from the conditions on means and covariances (7), (8).

4.4 State and Parameters Estimation

The problem of identifying state-space models on the basis of measured data can be solved with two approaches. The first one is using *Prediction Error Methods (PEM)*, in which we obtain a constrained optimization problem

$$\begin{array}{ll} \text{given} & u_{meas} \text{ and } y_{meas} \\ \underset{A,B,C,D}{\text{minimize}} & ||y - y_{meas}||_2^2 \\ \text{subject to} & \begin{cases} x(k+1) = A\,x(k) + B\,u_{meas}(k) \\ y(k) = C\,x(k) + D\,u_{meas}(k) \end{cases} \end{array}$$

where

- the variables are the unknown parameters of the model,
- the objective function is the difference between the measured data and the predictions obtained from the model, i.e. the *Prediction Error*,

- the constraints are the model equations.

Another kind of methods are the Subspace Methods [6]: first they estimate the states of the system using a projection of certain subspaces generated from the data, then determine the state-space model by a linear least squares method. They are a class of methods that look for the estimated state on particular subspaces generated by the data, from which comes the name. They are preferable with respect to the first method since they are faster, no iterative methods are needed and they exploit linear algebra structure. Their result however is suboptimal, hence they are used as an initial point for the PEM method.

4.5 Input/Output Denoising

The last problem we consider is the estimation of noisy Input/Output measurements through the knowledge of the model and the covariance matrices.

Proposition 2 (Problem of I/O estimation: DLTI Noisy I/O problem) *Given the DLTI system with noisy input/output data*

$$\begin{cases} x(k+1) = Ax(k) + Bu(k) \\ y(k) = Cx(k) + Du(k) \end{cases} \quad \textit{with} \begin{cases} u_d(k) = u(k) + e_u(k) \\ y_d(k) = y(k) + e_y(k) \end{cases}$$

the DLTI Noisy I/O problem is the following

$$\begin{aligned} &\min_{\hat{u},\hat{y},\hat{x}} \left\| \begin{bmatrix} V_{\tilde{u}} & \\ & V_{\tilde{y}} \end{bmatrix}^{-1/2} \begin{bmatrix} \hat{u} - u_d \\ \hat{y} - y_d \end{bmatrix} \right\|_2^2 \\ &s.t.\ \ \hat{x}(k+1) = A\hat{x}(k) + B\hat{u}(k) \\ &\qquad\ \ \hat{y}(k) = C\hat{x}(k) + D\hat{u}(k) \qquad \textit{for}\ \ k = 1, \dots, t_f - 1. \end{aligned} \tag{9}$$

The optimal smoothed state estimate $\hat{x}(\cdot, t_f)$ *is the solution of (9).*

Even if this problem starts with a different model structure, it can be simply reduced to a Kalman Filter problem, as shown in [5], in the following way:

Modified Kalman Filter problem:

$$\begin{cases} x(t+1) &= Ax(t) + Bu_d(t) + w(t) \\ y_d(t) &= Cx(t) + Du_d(t) + v(t) \end{cases} \quad \text{with} \quad \begin{cases} w := -Be_u \\ v := -De_u + e_y. \end{cases} \tag{10}$$

4.5.1 Denoising of Input/Output data without the knowledge of covariances

In the problem above, the covariance values of Input and Output signals are supposed known. In few words, this information gives us the precision and confidence of our measurements. However if the knowledge of covariance values is not available, the Kalman filter does not give an optimal solution. In this case the following approach can be considered [3]: rewrite the problem as a linear system $G\tilde{z} = d$, with $\tilde{z} = [\tilde{e}_u, \tilde{e}_y, \bar{e}_u, \bar{e}_y]$ and minimize the following least squares problem with four additive terms

$$\begin{aligned}\min_{\tilde{z}} \big(\|G\tilde{z} - d\|_2^2 + \|\Lambda_{eu}^{min}\tilde{e}_u\|_2^2 + \|\Lambda_{ey}^{min}\tilde{e}_y\|_2^2 + \\ \|\Lambda_{eu}^{curv} L_{n_u}(u_e - \hat{e}_u)\|_2^2 + \|\Lambda_{ey}^{curv} L_{n_y}(y_e - \hat{e}_y)\|_2^2\big).\end{aligned}$$

The minimization problem comes from the need to satisfy the following points: the Input and Output must be related through the model (least squares problem), the error should have small magnitude, and the signals should have small curvature, i.e. they should not vary too much (Tikhonov regularization terms). The choice of the weights Λ_* can be based on statistical properties of the Gaussian noise, i.e. all frequencies have the same probability (Normalized Cumulative Periodogram must be near to a line).

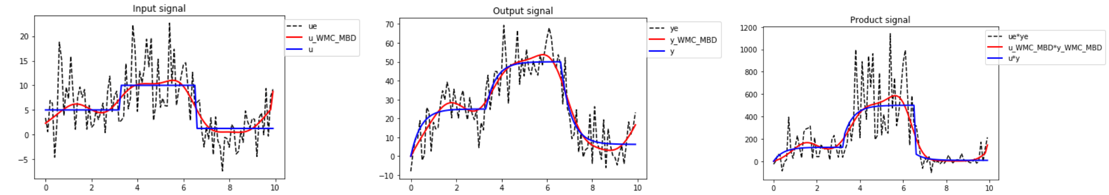


Figure 4. The true (blue), noisy (dotted black) and denoised (red) signals of input (left), output (center) and product of input and output (right) are shown.

The importance of this problem in applications is that the errors on Input/Output may be intensified by the application of a function that computes a derivative quantity (e.g. product of input/output, such as the mechanical/electrical power). In Figure 4 an example with the true, noisy and denoised signals of input, output and product of input and output is shown.

4.6 Unbiased Least-Squares modelling

Finally, we summarize a computational problem, described in [4], for linear models in which the importance of the physical meaning of the parameters is central. We suppose that a physical system is described by the real linear model

$$f = A\bar{x} = [A_a \, A_u] \begin{bmatrix} \bar{x}_a \\ \bar{x}_u \end{bmatrix} = A_a\bar{x}_a + A_u\bar{x}_u$$

with $A \in \mathbb{R}^{n\times n}$ full rank, $A_a \in \mathbb{R}^{n\times n_a}$ full column rank approximated known model $A_a x_a = f$, x_a parameters to be estimated, A_u unknown un-modelled component, with A_u non-orthogonal to A_a. The non-orthogonal residual implies that the least squares solution is not the real physical one

$$\hat{x}^{\|} = \operatorname*{argmin}_{x' \in \mathbb{R}^n} \|A_a x' - f\|^2 \qquad \neq \bar{x}_a.$$

This is a case in which, even if the model describes in a good way the system, the estimated parameter vector x_a does not have the true physical meaning.

A solution can be found supposing a-priori conditions on the norm of the modelled part and ratio between the modelled and unmodelled parts:

$$\begin{cases} N_f = \|A_a x_a\| \\ I_f = \frac{\|A_a x_a\|}{\|A_u x_u\|} \end{cases} \iff \begin{cases} x_a \in A_a^{\dagger}(\partial B_n(0, N_f)) \\ x_a \in A_a^{\dagger}(\partial B_n(f^{\parallel}, T_f)) \end{cases} \qquad \text{with } \; T_f := \sqrt{\left(\frac{N_f}{I_f}\right)^2 - \|f^{\perp}\|^2}.$$

These two conditions constraint the solution to lie on two hyper-ellipsoids which are the boundaries of two n-dimensional balls in $\mathbb{R}^n$.

If the linear system is built from a physical system, we can build more linear systems considering different time instants, or different input signal values. Intersecting the solutions sets of each test an considering a sufficient number of tests a unique solution is determined.

References

[1] J. Candy, “Bayesian Signal Processing: Classical, Modern, and Particle Filtering Methods”. Wiley IEEE Press, July 2016, ISBN 978-1-119-12545-7.

[2] J. Candy, “Model-Based Signal Processing”. Wiley IEEE Press, 2005, ISBN 0471236322.

[3] M. Gatto, F. Marcuzzi, *An algorithm for model-based denoising of input-output data.* Dolomites Research Notes on Approximation, 2019, 12(1), 73–85.

[4] M. Gatto, F. Marcuzzi, *Unbiased Least-Squares Modelling.* Mathematics 2020, 8(6), 982.

[5] I. Markovsky, B. De Moor, *Linear Dynamic Filtering with Noisy Input and Output.* IFAC Proceedings Volumes, 36/16 (2003), 1711–1716.

[6] M. Verhaegen, V. Verdult, “Filtering and System Identification: A Least Squares Approach”. Cambridge University Press, 2007, doi:10.1017/CBO9780511618888.

Modeling of Supply and Demand Curves for Day-Ahead Electricity Market

Mariia Soloviova [(*)]

Abstract. Accurate modeling and forecasting electricity demand and prices are very important issues for decision making in deregulated electricity markets. In this seminar I will explain some basic facts about price formation process in day-ahead electricity market, then I will speak mainly about the problem of approximation of supply and demand curves, with a special attention to Italian case. Finally, I will show how supply and demand curves evolve as stochastic processes in functional spaces.

1 Introduction

Accurate modeling and forecasting electricity demand and prices are very important issues for decision making in deregulated electricity markets. Different techniques were developed to describe and forecast the dynamics of electricity load. Short term forecast proved to be very challenging task due to these specific features. Figures 1 and 2 demonstrate changing of electricity equilibrium price and quantity during one week. Functional data analysis is extensively used in other fields of science, but it has been little explored in the electricity market setting.

In the beginning of the 2000s the amount of papers focused on electricity price forecasting started to increase dramatically. A great variety of methods and models occurred during last twenty years. Weron [12] (2014) made an overview of the existing literature on electricity price forecasting and divided electricity price models into five different groups: multi-agent, fundamental, reduced-form, statistical and computational intelligence models. A review of probabilistic forecasting was done in [7] (2018) by Weron and Nowotarski. Most models have in common that they focus on the price itself or related time series. In such a way these models does not take into account the underlying mechanic which determines the price process – the intersection between the part of the electricity supply and demand.

[(*)] Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy. E-mail: `soloviov@math.unipd.it` . Seminar held on 10 June 2020.

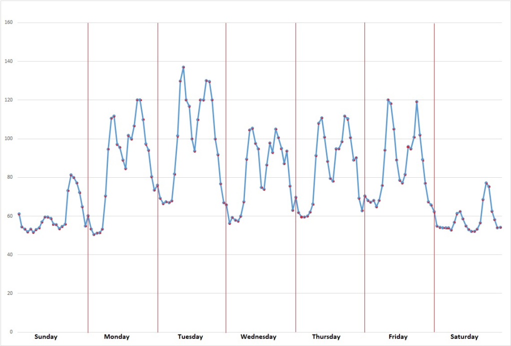


Figure 1. Electricity equilibrium prices during a week

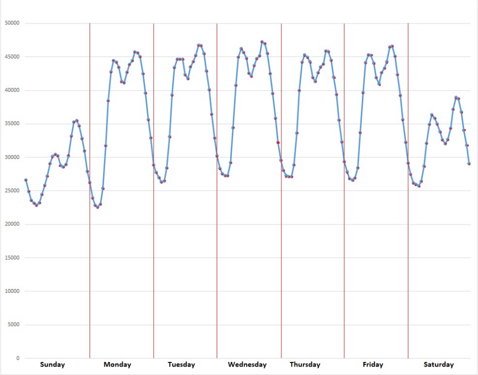


Figure 2. Electricity equilibrium quantities during a week

We consider the Italian electricity market (IPEX). IPEX consists of different markets, including a day-ahead market. The day-ahead market is managed by Gestore del Mercato Elettrico where prices and demand are determined the day before the delivery. Supply and demand curves on day-ahead electricity markets are the results of thousands of bid and ask entries in the day-ahead auction, this for all the 24 hours. In principle, it would be possible to represent, and forecast, these curves by taking into account each production and each consumption unit as a separate time series, and then joining these together to construct the final curves, and thus the resulting price. However, the huge number of these units makes this naive strategy infeasible, unless one has extremely high computing capacity with complex machine learning algorithms available.

In this paper, we are going to present one more parsimonious approach. Our idea is to represent each curve using non-parametric mesh-free interpolation techniques, so that we can obtain an approximation of the original curve with far less parameters than the original one. The original curve, in fact, in principle depends on about hundreds of parameters and is obtained as follow.

The producers submit offers where they specify the quantities and the minimum price at which they are willing to sell. The demanders submit bids where they specify the quantities and the maximum price at which they are willing to buy. They are then aggregated by an independent system operator (ISO) in order to construct the supply and demand curves. Once the offers and bids are received by the ISO, supply and demand curves are established by summing up individual supply and demand schedules. In the case of demand, the first step is to replace ”zero prices“ bids by the market maximum price (for Italian electricity market, the market maximum price is 3000 Euro) without changing the corresponding quantities. After this replacement, the bids are sorted from the highest to the lowest with respect to prices. The corresponding value of the quantities is obtained by cumulating each single demand bid. For supply curve, in contrast, the offers are sorted from the lowest to the highest with respect to prices and the corresponding value of the quantities is obtained by cumulating each single supply offer. The market equilibrium is the point where both curves intersect each other and the price balances supply and demand schedules (see Figure 3). This point determines the market clearing price and the quantity. Accepted offers and bids are those that fall to the left of the intersection of the two curves and all of them are exchanged at the resulted price.

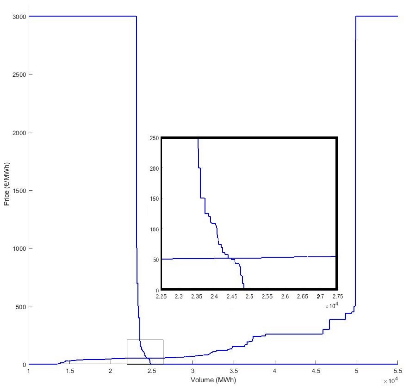


Figure 3. The market equilibrium point

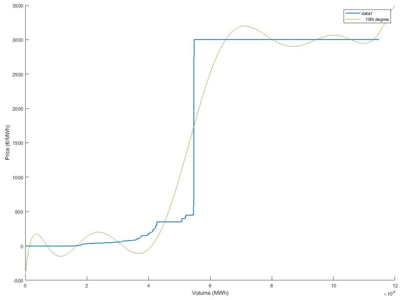


Figure 4. Approximation of supply curve with polynomials

Some of the recent approaches try to to analyse the real offered volumes for selling and purchasing electricity. This commonly leads to a problem of a large amount of data and, therefore, high complexity.

Then Ziel and Steinert in 2016 [13] proposed a model for the German European Power Exchange (EPEX) market, which considers all the supply and demand information of the system and discusses the effects of the changes in supply and demand. Their idea was to fill the gap between research done in time-series analysis, where the structure of the market is usually left out, and the research done in structural analysis, where empirical data is utilized very rarely and even less thoroughly. They provided deep insight on the bidding behavior of market participants. They also showed that incorporating the sale and purchase data yields promising results for forecasting the likelihood of extreme price events.

As far as we know, non-parametric mesh-free interpolation techniques were never considered for the problem of modeling the daily supply and demand curves.

We are going to use a relatively new modeling technique based on functional data analysis for demand and price prediction. The first task for this purpose is to make an appropriate algorithm to present the information about electricity prices and demands, in particular to approximate a monotone piecewise constant function.

We want to make an appropriate algorithm to present this information, in particular, to approximate a monotone piecewise constant function. Accuracy of the approximation and running time are very important for us. As we already said, the basic novelty of our problem is that we are going to present the information about electricity prices and demands using functional data analysis approach. The main idea behind functional data analysis is, instead of considering a collection of data points, to consider the data as a single structured object. This allows to use additional information contained in the functional structure of the data. Once the data are converted to functional form, it can be evaluated at all values over some interval.

The most promising technique to do so is the use of (integrals of) Radial Basis Functions, which are been used in several other applications (image reconstruction, medical imaging, geology, etc.) and allow a very flexible adaptation of the interpolating curves to real data. The use of radial basis functions have attracted increasing attention in recent years as an elegant scheme for high-dimensional scattered data approximation, an accepted method for machine learning, one of the foundations of meshfree methods and so on (see [8], for example). We will present different techniques for this interpolation, with their advantages and drawbacks, and with an application to the Italian day-ahead market.

We consider the Italian electricity market (IPEX). IPEX consists of different markets, including a day-ahead market. The day-ahead market is managed by Gestore del Mercato Elettrico (GME) where prices and demand are determined the day before the delivery by means of hourly concurrent auctions. In this paper, we consider supply and demand curves as stochastic processes with values in a functional space. Recall that, in the price formation process, the producers submit offers, where they specify the quantities of electricity and the minimum price at which they are willing to sell. However, the market operator allows a minimum increase, or tick, of 1 kWh for quantities and 0.01 euro/MWh for prices. Then, the dimension of our model is 60 000 000. In order to deal with the huge amount of bid data, one needs dimension reduction techniques and high-dimensional statistical methods. Using the tools of functional data analysis, it is possible to approximate the original supply and demand curves with far less parameters than the original ones. Then we continue research considering supply and demand curves as stochastic processes with values in functional space.

Linear processes on functional spaces were born about fifteen years ago. The linear processes on function spaces generalize the classical scalar or vector linear processes to random elements which are curves or functions and more generally valued in an infinite-dimensional separable Hilbert space H. Random curves designed to improve the quality of prediction. This approach offers a wide spectrum of models suited to the statistical inference and also leads to challenging theoretical and applied problems.

2 Meshless approximation of supply and demand curves

Let us briefly notice some features of supply and demand curves that are relevant for our modeling:

1. By construction, the curves are monotone.
2. The values attained by the supply curve are roughly clustered around **layers**, corresponding to different production technologies. In Italy they are non-dispatchable renewables, gas, coal, hydro, oil.
3. The fact that renewables are the first ones make the supply curve intrinsically ”meshless”.
4. Demand is much more inelastic than supply.

So, we are dealing with a scattered data interpolation problem. We have a large amount of points (each point represents price and amount of electricity) that we want to approximate. We can formalize this problem as follows.

Given a set of N distinct *data points* $X_N = \{x_i : i = 1, 2, \dots, N\}$ arbitrarily distributed on a domain $\Omega \subset \mathbb{R}$ and a set of *data values* (or function values) $Y_N = \{y_i : i = 1, 2, \dots, N\} \subset \mathbb{R}$, the data interpolation problem consists in finding a function $s_f : \Omega \to \mathbb{R}$ such that

$$s_f(x_i) = y_i,\ i = 1, \dots, N. \tag{2.1}$$

Notice that for all methods, the interpolant s_f is expressed as a linear combination of some basis functions B_i , i.e.

$$s_f(t) = \sum_{k=1}^{d} c_k B_k(t).$$

The basis functions in polynomial interpolation does not depend on the data points. Another approach is to use a basis which depends on the data points.

One simple way to solve problem (2.1) is to choose a fixed function $\phi : \mathbb{R} \to \mathbb{R}$ and to form the interpolant as

$$s_f(x) = \sum_{i=1}^{N} \alpha_i \phi(\|x - x_i\|),$$

where the coefficients α_i are determined by the interpolation conditions $s_f(x_i) = y_i$. Therefore, the scattered data interpolation problem leads to the solution of a linear system

$$A\alpha = y, \text{ where } A_{i,j} = \phi(|x_i - x_j|).$$

The solution of the system requires that the matrix A is non-singular. It is enough to know in advance that the matrix is positive definite (see [11] for more details). Let us recall the definition of strictly positive definite function.

Definition 2.1 A real-valued function $\Phi : \mathbb{R} \longrightarrow \mathbb{R}$ is called *positive semi-definite* if , for all $m \in \mathbb{N}$ and for any set of pairwise distinct points $x_1, x_2, \dots, x_m$, the $m \times m$ matrix

$$A = (\Phi(x_i - x_j))_{i,j=1}^m$$

is positive semi-definite, i.e. for every column vector z of m real numbers the scalar $z^T Az \geqslant 0$. The function $\Phi : \mathbb{R} \longrightarrow \mathbb{R}$ is called (strictly) *positive definite* if the matrix A is positive definite, i.e. for every non-zero column vector z of m real numbers the scalar $z^T Az > 0$.

The most important property of positive semi-definite matrices is that their eigenvalues are positive and so is its determinant.

A radial function is a real-valued positive semi-definite function whose value depends only on the distance from the center $\mathbf{c}$. One useful characterization for positive semi-definite univariate function was given by Schoenberg in 1938 in the terms of completely monotone functions: a continuous function $\phi : [0, \infty) \to \mathbb{R}$ is positive semi-definite if and only if $\phi \in C^\infty(0, \infty)$ and $(-1)^k \phi^{(k)}(r) \geqslant 0$ for all $r \geqslant 0$, for $k = 0, 1, \dots$.

Some standard radial basis functions are

- $\phi(r) = e^{-(\varepsilon r)^2}$ (Gaussian),
- $\phi(r) = e^{-\varepsilon r}(\varepsilon r + 1)$ (Matérn),
- $\phi(r) = (1 - \varepsilon r)_+^4 (4\varepsilon r + 1)$ (Wendland),

where $\varepsilon > 0$ denote a shape parameter, $r = \|x\|_2$.

The idea of meshless approximation with radial basis functions is to find an approximant of f in the following form:

$$s_f(x) := \sum_{i=1}^{N} \alpha_i \phi(\|x - x_i\|)$$

where:

- the coefficients α_i and the **centers** x_i are to be chosen so that the interpolant is as near as possible as the original function f;
- $\phi : \mathbb{R} \to \mathbb{R}$ is a **radial basis function** (RBF).

Notice that the radial basis function $\phi \geqslant 0$, with $\alpha_i \geqslant 0$, so

$$\sum_{i=1}^{M} \alpha_i \phi(\|x - x_i\|) \geqslant 0.$$

As we need to approximate piecewise constant monotone function from $[0, M]$ to $\mathbb{R}^+$, we decided to use the integrals of RBF. Namely, we want to find an approximant of the form

$$s_f(t) = \int_0^t \sum_{i=1}^{M} \alpha_i \phi(\lambda_i \|x - x_i\|) \, dx = \sum_{i=1}^{M} \alpha_i \int_0^t \phi(\lambda_i \|x - x_i\|) \, dx$$

where λ_i is a shape parameter for every center x_i.

Evidently, any supply curve and any demand curve can be approximated by a combination of error functions, which is the integral of a normalized Gaussian function. The standard error function is defined as:

$$\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} \, dt = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, dt.$$

Since we want to approximate monotone curves we came up with the idea to use the integral of radial basis function. In order to find unknown coefficients α_i, λ_i, x_i we need to solve global minimization problem:

$$\min_p \|s_f(x_i, p) - y_i\|_2^2,$$

where $p = (\alpha_i, \lambda_i, x_i)_{i=1,\dots,N}$ and

$$s_f(t, p) := \sum_{i=1}^{M} \alpha_i \int_0^t \phi(\lambda_i \|x - x_i\|) \, dx$$

and $\phi(t) = (\operatorname{erf}(t) + 1)/2$ is the primitive of a Gaussian kernel.

2.1 Data set

In our work we are using the data about supply bids from the Italian electricity market from the GME website www.mercatoelettrico.org. We consider time period from 01.01.2017 to 31.12.2017. These data are in aggregated form, i.e. bids coming from different agents but with the same price are aggregated in the price layer. Even in this form, we are dealing with the massive amount of data. For instance, there were observed **2 800 687** offer and **558 926** bid layers during this period.

Date	Hour	Volume (MW)	Price (Euro)
01-01-2017	1	13392.7	0
01-01-2017	1	25	0.1
01-01-2017	1	113.8	1
01-01-2017	1	11	3.5
01-01-2017	1	270.3	5
01-01-2017	1	0.5	6
....	......	....	....
31-12-2017	24	370	554.2
31-12-2017	24	352	554.3
31-12-2017	24	365	554.5
31-12-2017	24	97	700
31-12-2017	24	60000	3000

Table 1. Data

So, it means, that on average there are 324 offer and 65 bid layers for each hour of the year, which corresponds to one supply curve and one demand curve respectively.

2.2 Numerical experiments

Since the maximum market clearing price for the period under review (i.e. from 01.01.2017 to 31.12.2017) is 350 €, in all the experiments we restricted ourselves to a maximum price of 400 €. For the realization of our algorithm we are using the function `lsqcurvefit` from MATLAB Optimization Toolbox.

First, we download the data from a text file and choose the number of basis function M. After that, we need to divide our problem into M sub-problems. Then each part of the supply curve must be approximated by one error function.

Our first attempt (Method 1) was just to divide y-axis uniformly into M equal intervals (see Figure 5). However this approach is ineffective, as a huge jump concentrates on itself, keeping uselessly many components.

To resolve this problem we created a simple algorithm - Method 2 - that finds the points $p_1, \dots, p_M$ on the y-axis such that our supply curve takes the value exactly p_i on some non-trivial interval (see Figure 6).

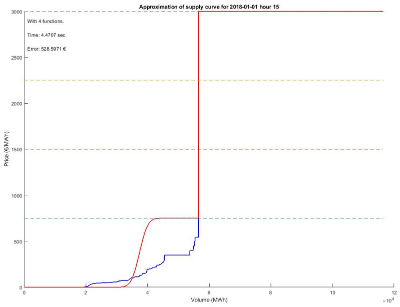


Figure 5. Method 1

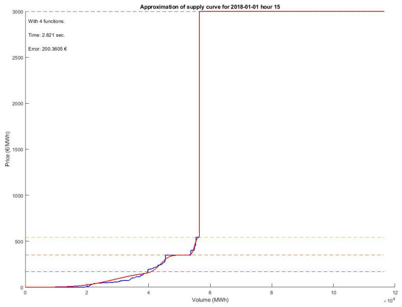


Figure 6. Method 2

Then M times we resolve the same optimization problem for the values of the supply curve between p_i and p_{i+1} using function `lsqcurvefit` (see Figure 7). On each part we need to find only 3 coefficients a_i, b_i, c_i of the function

$$G(x) = \sum_{i=1}^{k} a_i(\operatorname{erf}(c_i \cdot (x - b_i)) + 1). \tag{2.2}$$

Here, for convenience of representation we are using $\{\operatorname{erf}(c_i \cdot (x - b_i)) + 1\}$ instead of $\{\operatorname{erf}(c_i \cdot (x - b_i))\}$, as our data values are never negative.

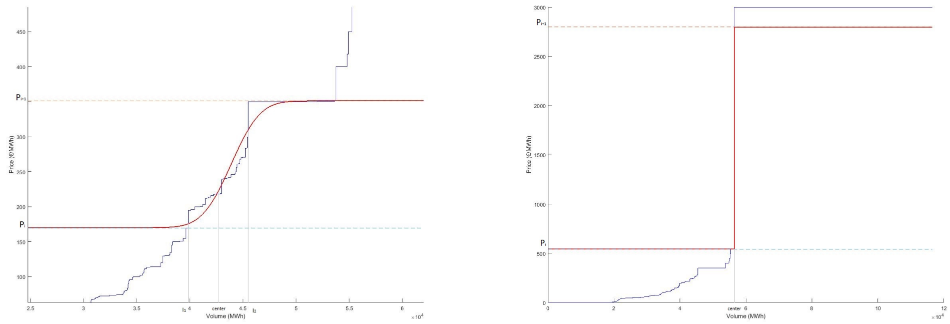


Figure 7. Local interpolation by one error function with `lsqcurvefit` function

The `lsqcurvefit` function solves nonlinear data-fitting problems in least-squares sense. Suppose that we have data points $X_N = \{x_i : i = 1, 2, \ldots, N\}$ and data values $Y_N = \{y_i : i = 1, 2, \ldots, N\} \subset \mathbb{R}$ and we want to find a function f such that $f(x_i) \approx y_i$, $i = 1, \ldots, N$. We can consider the family of functions $\{f(x, p) : p \in \mathbb{R}^k\}$, depending of some parameter $p \in \mathbb{R}^k$. Let $p_0 \in \mathbb{R}^k$ be an "initial guess" such that $f(x_i, p)$ is reasonably close to y_i. The function `lsqcurvefit` starts at p_0 and finds coefficients p from some neighborhood of p_0 to best fit the data set Y_N:

$$\min_p \|f(x_i, p) - y_i\|_2^2.$$

Notice that this function works well only if the number of parameters $(p_1, \dots, p_k)$ is not very big. That is why we are forced to divide our problem into many local problems.

For optimizing the numerical procedure we solved some parts of the optimization problem by ourselves: in fact, when the interval $[p_i, p_{i+1}]$ contains only one jump, then

$$a_i := f(p_{i+1}) - f(p_i)$$

for any kernel function ϕ with unit integral.

A summary of the results is shown in Table 2. For all experiments we proceed with the data for period from 01.01.2017 to 31.12.2017. We used different number of basis function to approximate supply and demand curves, and then compared the equilibrium price, which was received as intersection of approximants (P_{appr}), with the correct equilibrium price (P). We did this for each hour of each day, and then computed the average value of $|P - P_{appr}|$ (Error) for all 8 664 hours of the year and the maximum value of $|P - P_{appr}|$ (Max error).

This empirical results show that the accuracy of our approximation is good enough, if we use 5 basis function for the demand curve and 15 basis function for the supply curve. Then the increase in the number of functions leads to more time consumption, but the increase of the accuracy is less significant.

Number of functions		Results		
For demand	For supply	Error	Max error	Running time
5	5	3.9 €	28.6 €	69 min.
5	10	2.2 €	14.9 €	82 min.
5	15	1.5 €	11.1 €	103 min.
5	20	1.3 €	9.1 €	110 min.
5	25	1.2 €	9.3 €	135 min.
5	30	1.2 €	9.4 €	159 min.
5	35	1.2 €	9.8 €	177 min.
5	40	1.2 €	9.6 €	190 min.
5	45	1.2 €	9.6 €	199 min.
5	50	1.2 €	9.6 €	207 min.
10	5	3.9 €	39.5 €	100 min.
10	10	2.1 €	14.9 €	128 min.
10	15	1.4 €	8.9 €	146 min.
10	20	1.2 €	9.1 €	162 min.
10	25	1.1 €	9.5 €	183 min.
10	30	1.1 €	9.3 €	199 min.
10	35	1.0 €	9.4 €	223 min.
10	40	0.98 €	9.8 €	241 min.
10	45	0.98 €	9.6 €	255 min.
10	50	0.98 €	9.6 €	273 min.

Table 2. Results of numerical experiment

3 Supply and demand curves as stochastic processes

Let us give a simple example where infinite-dimensional modeling is a useful tool for applications. If one observes temperature in continuous time during N days, and wants to predict its evolution during the $(N+1)$ day, then $(X_n), n \in \mathbb{N}$ is a sequence of random variables with values in a suitable function space, say $C([0,24])$.

Another example of modeling in large dimensions is the following: consider an economic variable associated with individuals. At instant n, the variable associated with the individual i is $X_{n,i}$. In order to study global evolution of that variable for a large number of individuals, and during a long time, it is convenient to set

$$X_n = (X_{n,i}, i \geqslant 1), \quad n \in \mathbb{Z},$$

which defines a process $X = (X_n, n \in \mathbb{Z})$ with values in some sequence space F.

Ziel and Steinert in [13] analyzed the hourly day-ahead electricity price auction data of Germany and Austria provided by the EPEX Spot from 01.10.2012 to 19.04.2015, using a subtle data processing technique as well as dimension reduction and lasso-based estimation methods. Their model consists of three parts: 1. Construction of price classes in order to overcome the massive amount of data. 2. Forecasting for each price class by using time series model. 3. Reconstruction of supply and demand curves and computation of market clearing price. We reformulated the model of Ziel and Steinert in terms of linear transformation of multivariate stochastic processes. They use a time series model for the bid volume processes $X_{S,t}^{(c)}$ and $X_{D,t}^{(c)}$ for each price classes c. The original bid volume processes are $V_{S,t}(p)$ and $V_{D,t}(p)$ for each possible price $p \in P = \{-500, -499.9, \ldots, 2999.9, 3000\}$. Denote $p_1 = -500, p_2 = -499.9, \ldots, p_N = 3000$, where $N = 35001$. So, we can say that the stochastic processes

$$V_{S,t} = (V_{S,t}(p_1), V_{S,t}(p_2), \ldots, V_{S,t}(p_N)),$$
$$V_{D,t} = (V_{D,t}(p_1), V_{D,t}(p_2), \ldots, V_{D,t}(p_N)),$$

are processes with values in $\mathbb{R}^N$, which represents the information about the whole supply and demand curves. More precisely, the sale and purchase curves are characterized by

$$S_t(p) = \sum_{i:p_i \leqslant p} V_{S,t}(p_i), \text{ and } D_t(p) = \sum_{i:p_i \geqslant p} V_{D,t}(p_i).$$

In order to reduce the dimensionality of the problem, Ziel and Steinert defined price classes for supply and demand curves as $C_S = (c_1, c_2, \ldots, c_M)$ and $C_D = (\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_M)$, where $-500 = c_1 < c_2, \ldots < c_M = 3000$ and $3000 = \tilde{c}_1 > \tilde{c}_2 > \ldots > \tilde{c}_M = -500$. In such a way, M is a new dimension for the studied processes and it is much less than N.

Let us explain the dimensionality reduction, i.e the process of finding the points $c_1, c_2, \ldots, c_M$ on the example of supply curve in more detail. In order to create the price classes Ziel and Steinert considered for any $p \in P$ the mean bid volume

$$\overline{V_S}(p) = \frac{1}{T} \sum_{t=1}^{T} V_{S,t}(p)$$

as a measure of importance for the price $p \in P$, where T is the number of observations across all hours in the database. Let $P_S = \cup_{t=1}^{T} P_{S,t}$ be the sets of all bid prices for the supply side. Then the mean supply curve $\overline{S}$ is characterized by

$$\overline{S}(p) = \sum_{\substack{x \in P_S \\ x \leqslant p}} \overline{V_S}(x) \text{ for } p \in P_{S,t}.$$

This equation define the curves explicitly only on the price grid P_S, but the complete mean supply curve can be obtained, for example, by a linear interpolation of the characterized points. The function $\overline{S}$ on the price grid P is monotone, so there exist the inverse function $\overline{S}^{-1}$. After that the authors defined the upper and lower values of the price classes by

$$C_S = \{\overline{S}^{-1}(nV_*) : n \in \mathbb{N}\}$$

where V_* is an amount of volume, which will give the average amount of volume that should be represented by every price class (in [13], for example, $V_* = 1000$). For any price class upper bound $c \in C_S$ the price class is given by

$$\begin{aligned} P_S(c) = \{p \in P : p \leqslant \min\{s \in C_S : s \geqslant c\} \text{ and } p > \max\{s \in C_S : s < c\}\} \\ = \{p \in P : p \leqslant c \text{ and } p > \max\{s \in C_S : s < c\}\} \\ = [\max\{s \in C_S : s < c\}, c) \cap P. \end{aligned}$$

Then the associated volumes at time t to the price classes for $c \in C_S$ is given by

$$X_{S,t}^{(c)} = \sum_{p \in P_S(c)} V_{S,t}(p).$$

Now, turning back to modeling the supply and demand curves, we can state that

$$X_{S,t} = T_S(V_{S,t}) \text{ and } X_{D,t} = T_D(V_{D,t})$$

where $T_S, T_D : \mathbb{R}^N \to \mathbb{R}^M$ are linear continuous operators such that

$$\begin{aligned} T_S(x_1, x_2, \ldots, x_n) &= \left(\sum_{i=1}^{k_1} x_i, \sum_{i=k_1+1}^{k_2} x_i, \ldots, \sum_{i=k_{M-1}+1}^{N} x_i \right) \text{ and} \\ T_D(x_1, x_2, \ldots, x_n) &= \left(\sum_{i=1}^{m_1} x_i, \sum_{i=k_1+1}^{m_2} x_i, \ldots, \sum_{i=m_{M-1}+1}^{N} x_i \right). \end{aligned}$$

So, the processes $X_{S,t}$ and $X_{D,t}$ are linear transformations of the processes $V_{S,t}$ and $V_{D,t}$.

Linear transformations of stochastic processes are used in many ways in economic analyses. In the paper [4] it is proved that an N-dimensional linear transformation of a process possessing an MA(q) representation gives a process that also has a finite order MA representation with order not greater than q. The more general fact that a linear transformation of a vector ARMA process is again an ARMA process is also proved. These results are of importance because many temporal as well as contemporaneous aggregation procedures can be represented as linear transformations.

Proposition 3.1 [4, Lemma 1] *Let X_t be an N-dimensional $MA(q)$ process, and $T = [t_{ij}]_{i,j} \neq 0$ be a real $M \times N$ matrix. Then $Y_t = T(X_t)$ is an $MA(q^*)$ process, where $q^* \leqslant q$.*

However, there are transformations of a finite order $AR(p)$ process that do not admit a finite order AR representation, but just a mixed $ARMA$ representation.

Let us show that there is no version of this theorem for AR processes.

Example 3.2 Consider the case $N = 2, M = 1$ and define the $AR(1)$ process X_t as

$$X_t = \begin{pmatrix} x_t^{(1)} \\ x_t^{(2)} \end{pmatrix} = \begin{pmatrix} ax_{t-1}^{(1)} + w_t^{(1)} \\ bx_{t-1}^{(2)} + w_t^{(2)} \end{pmatrix}$$

Let $T = [1 \quad 1] : \mathbb{R}^2 \to \mathbb{R}$. Then $Y_t = T(X_t) = ax_{t-1}^{(1)} + bx_{t-1}^{(2)} + w_t^{(1)} + w_t^{(2)}$. Evidently, unless $a = b$, Y_t is not autoregressive.

For modeling the electricity price Ziel and Steinert follow to a simple regression approach described in [5]. So, in this case, the initial processes $X_{S,t}$ and $X_{D,t}$, and the transformed processes $V_{S,t}$ and $V_{D,t}$ are vector-valued autoregressive process. We studied the following problem: suppose that V_t is an $\mathbb{R}^N$-valued process and $V_t \in AR(p)$, i.e.

$$V_t = A_1V_{t-1} + A_2V_{t-2} + \ldots + A_pV_{t-p} + W_t,$$

where A_i are $(N \times N)$ coefficient matrices W_t is an $(N \times 1)$ unobservable zero-mean white noise vector process. Let $T : \mathbb{R}^N \to \mathbb{R}^M$ be a linear continuous operator, and $Y_t = T(V_t)$. We have explored the question under which condition $Y_t \in AR(p)$ and what is the connection between A_i, B_i, W_t and Z_t, if

$$Y_t = B_1V_{t-1} + B_2V_{t-2} + \ldots + B_pV(t-p) + Z_t.$$

In particular, we investigated for which transformations T the process $T(X_t)$ has an $AR(p)$ representation. We obtained the following result.

Theorem 3.3 *Let X_t be an N-dimensional $AR(p)$ process with the representation*

$$X_t = A_1X_{t-1} + A_2X_{t-2} + \ldots + A_pX_{t-p} + W_t,$$

and $T : \mathbb{R}^N \to \mathbb{R}^M$, $M < N$ be linear transformation. Then $Y_t = T(X_t)$ is an M-dimensional $AR(p)$ if and only if there exist $(M \times M)$ matrices B_i such that

$$B_iT = TA_i \quad for\ all\ i = 1, \ldots, p. \tag{3.1}$$

Moreover, the process (Y_t) has the representation:

$$Y_t = B_1Y_{t-1} + B_2Y_{t-2} + \ldots + B_pY_{t-p} + Z_t,$$

Remark 3.4 If rank $T = M$ (i.e. rows of T are linearly independent), the pseudoinverse of T can be expressed as follows

$$T^+ = T^T(TT^T)^{-1}.$$

Then from (3.1) we obtain

$$TA_iT^+ = B_iTT^+ = B_iTT^T(TT^T)^{-1} = B_i$$

So, $B_i = TA_iT^+$ and

$$Y_t = TA_1T^+Y_{t-1} + TA_2T^+Y_{t-2} + \ldots + TA_pT^+Y_{t-p} + TW_t.$$

We also formulated some conditions for the model of Ziel and Steinert in order to have Theorem 3.3 to hold true. For any matrix T let R_i^T denote the row i of matrix T and let C_j^T denote the column j of matrix T. The following simple lemma is valid.

Lemma 3.5 *Let A be an $(N \times N)$ matrix and T be an $(M \times N)$ matrix ($M < N$) with columns*

$$C^T_{k_{s-1}} = C^T_{k_{s-1}+1} = C^T_{k_{s-1}+2} = \ldots = C^T_{k_s} = e_s, \quad 1 \leqslant s \leqslant M$$

where $0 = k_0 < k_1 < k_2 < \ldots < k_M = N$ and $\{e_1, e_2, \ldots, e_M\}$ is the standard basis of $\mathbb{R}^M$.

Then the following two conditions are equivalent :

1. *There is $B \in \mathbb{R}^{(M\times M)}$ such that $TA = BT$;*
2. *For any $1 \leqslant i, l \leqslant M$ there exist number D_{il} such that*

$$\sum_{s=k_{l-1}+1}^{k_l} a_{sj} = D_{il} \quad for\ all\ k_{i-1} < j \leqslant k_i.$$

In particular condition (2) is satisfied if the first k_1 columns of A are the same, the second $k_2 - k_1$ are the same, and so on, until the last $k_M - k_{M-1}$ columns.

Proof. Suppose that $A = \Psi T$. From the formula

$$a_{ij} = \langle R_i^\Psi, C_j^T \rangle$$

we obtain that $C_l^A = C_d^A$ for any $k_{s-1} \leqslant l, d \leqslant k_s$.

Conversely, suppose now that

$$A = \left(\underbrace{C^A_{k_1} \ldots C^A_{k_1}}_{k_1} \underbrace{C^A_{k_2} \ldots C^A_{k_2}}_{k_2 - k_1} \ldots \underbrace{C^A_{k_M} \ldots C^A_{k_M}}_{k_M - k_{M-1}} \right).$$

Then $A = \Psi T$ holds true for $\Psi = \left(C^A_{k_1}\, C^A_{k_2}\, \ldots\, C^A_{k_M} \right)$. □

In the model of Ziel and Steinert we start from the stochastic processes with values in $\mathbb{R}^N$, which represents the information about the whole supply curve:

$$V_{S,t} = (V_{S,t}(p_1), V_{S,t}(p_2), \ldots, V_{S,t}(p_N)),$$

and then we define the modified process with values in $\mathbb{R}^M$

$$X_{S,t} = (X_{S,t}(c_1), X_{S,t}(c_2), \ldots, X_{S,t}(c_M)),$$

such that

$$X_{S,t} = T_S(V_{S,t}).$$

The next result follows directly from an application of Lemma 3.5 and Theorem 3.3.

Proposition 3.6 *Suppose that X_t is AR(p) process with the representation*

$$X_t = A_1 X_{t-1} + A_2 X_{t-2} + \ldots + A_p X_{t-p} + W_t,$$

$T : \mathbb{R}^N \to \mathbb{R}^M$ is the linear continuous operators such that

$$T(x_1, x_2, \ldots, x_n) = \left(\sum_{i=1}^{k_1} x_i, \sum_{i=k_1+1}^{k_2} x_i, \ldots, \sum_{i=k_{M-1}+1}^{k_M} x_i \right),$$

for some $0 = k_0 < k_1 < k_2 < \ldots < k_M = N$.

Then $Y_t = T(X_t)$ is AR(p) process if and only if In particular, if in all the matrices A_i, the first k_1 columns are the same, the second $k_2 - k_1$ are the same, and so on, until the last $N - k_{M-1}$ columns. Then $Y_t = T(X_t)$ is AR(p) with the representation

$$X_t = B_1 X_{t-1} + B_2 X_{t-2} + \ldots + B_p X_{t-p} + (TW_t),$$

with $B_i = T\Psi_i$ are $M \times M$ matrices, where $\Psi = \left(C_{k_1}^{A_i}\, C_{k_2}^{A_i} \, \ldots \, C_{k_M}^{A_i} \right) \in \mathbb{R}^{N \times M}$.

As we already mentioned, stochastic processes with values in infinite-dimensional space often allows to represent the initial data with relatively small number of parameters and improve the quality of prediction. So, we also obtained the version of Proposition 3.6 for the infinite-dimensional AR processes. For this matter, we recall the definition of autoregressive Hilbertian process [2].

Let H be a real separable Hilbert space with its norm $\|\cdot\|$ and its scalar product $\langle \cdot, \cdot \rangle$, $\mathcal{L}$ denote the space of continuous linear operators from H to H. We consider the space $L_H^2 := L_H^2(\Omega, \mathcal{A}, P)$ of random variables X, with values in H, and such that $\mathbb{E}\|X\|^2 < \infty$. If $\mathbb{E}\|X\|^2 < \infty$, then the mathematical expectation $\mathbb{E}X$ exists as an element of H . The mean $\mathbb{E}X$ is the unique element of H such that $\langle \mathbb{E}X, h \rangle = \mathbb{E}\langle X, h \rangle$ for all $h \in H$. The scalar product for $X, Y \in L_H^2$ is defined as follow:

$$\langle X, Y \rangle_{L_H^2} = \mathbb{E}\langle X, Y \rangle.$$

For any $X, Y \in L^2_H$ the cross-covariance operator of X and Y, which is an infinite-dimensional analogous to the covariance matrix, is defined as

$$C_{X,Y}(h) = \mathbb{E}[\langle X - \mathbb{E}X, h\rangle (Y - \mathbb{E}Y)] : H \to H.$$

The covariance operator $C_{X,X}$ of X is denoted by C_X.

One important case arises when one assumes that $X = (X_n,\, n \in \mathbb{Z})$ is a stationary zero-mean autoregressive H-valued process of order 1 (ARH(1)), which has the following characterization:

$$X_n = \varepsilon_n + \rho(X_{n-1}), \tag{3.2}$$

with $\rho \in \mathcal{L}$ and ε_n, $n \in \mathbb{Z}$ a strong H-valued white noise, (i.e. $\mathbb{E}\varepsilon_n = 0$, $C_{\varepsilon_n} = C_{\varepsilon_0} \neq 0$ for any $n \in \mathbb{Z}$; $C_{\varepsilon_n,\varepsilon_m} = 0$ for any $n \neq m$; (ε_n) and, in addition, ε_n are independent and identically distributed).

Proposition 3.7 *Consider a zero-mean ARH*(1) *process* $X = (X_n, n \in \mathbb{Z})$*, satisfying, for all* $n \in \mathbb{Z}$*, the equation*

$$X_n = \rho(X_{n-1}) + \varepsilon_n,$$

where ρ *denotes the autocorrelation operator of the process* X*, which belongs to the class* $\mathcal{L}$*.*

Let $T : H \to H$ *is the linear continuous operators and* $Y_t = T(X_t)$ *is AR*(p)*. Then* $Y = (Y_n, n \in \mathbb{Z})$ *is an ARH*(1) *if and only if there exist* $\vartheta \in \mathcal{L}$ *such that*

$$T\rho = \vartheta T. \tag{3.3}$$

Proof. The following sequence of equality shows that condition (3.3) is necessary and sufficient:

$$\begin{aligned} Y_t = T(X_t) &= T\rho X_{t-1} + T\varepsilon_t \\ &= \vartheta T X_{t-1} + T\varepsilon_t \\ &= \vartheta Y_{t-1} + \xi_t, \end{aligned}$$

where $\xi_t = T\varepsilon_t$ is a zero-mean strong white noise $\vartheta \in \mathcal{L}$. □

4 Conclusions

We presented a parsimonious way to represent supply and demand curves, using a mesh-free method based on Radial Basis Functions. Using the tools of functional data analysis, we are able to approximate the original curves with far less parameters than the original ones. Namely, in order to approximate piece-wise constant monotone functions, we are using the combination of the integral of a normalized Gaussian function.

We also consider supply and demand curves as stochastic processes with values in a functional space. In order to deal with the huge amount of bid data, we studied linear transformation of multivariate stochastic process. It is known fact that a linear transformation of a vector ARMA process is again an ARMA process. However, in general, there are transformations of a finite order AR(p) process that do not admit a finite order AR representation, but just a mixed ARMA representation. We obtained some partial results regarding the conditions that guarantee that linear transformation of a vector AR process is again an AR process.

References

[1] D. Bosq, "Linear Processes in Function Spaces". Lecture Notes in Statistics, vol. 149. Springer, New York, 2000.

[2] D. Bosq and P. Blanke, "Inference and Prediction in Large Dimensions". John Wiley & Sons Ltd. 2007.

[3] Fasshauer G.E., McCourt M.J., "Kernel-based Approximation Methods Using MATLAB". World Scientific, Singapore, 2015.

[4] Lutkepohl, H., *Linear transformations of vector ARMA processes.* Journal of Econometrics, Elsevier, vol. 26/3 (1984), 283–293.

[5] Maciejowska, K., Nowotarski, J., Weron, R., *Probabilistic forecasting of electricity spot prices using factor quantile regression averaging.* Int. J. Forecast. 32/3 (2016), 957–965.

[6] Micchelli C.A., *Interpolation of scattered data: Distance matrices and conditionally positive definite functions.* Constr. Approx. 2 (1986), 11–22.

[7] Nowotarski J., Weron R., *Recent advances in electricity price forecasting: A review of probabilistic forecasting.* Renew. Sustain. Energy Rev. 81 (2018), 1548–1568.

[8] Perracchione E., Stura I., *RBF kernel method and its applications to clinical data.* Dolomites Res. Notes Approx. 9 (2016), 13–18.

[9] Schaback R., "Native Hilbert Spaces for Radial Basis Functions I". In: Müller M.W., Buhmann M.D., Mache D.H., Felten M. (eds) New Developments in Approximation Theory. ISNM International Series of Numerical Mathematics, vol 132. Birkhäuser, Basel, 1999.

[10] Shah, I., "Modeling and Forecasting Electricity Market Variables ". PhD Thesis, 2016.

[11] Wendland H., "Scattered Data Approximation". Cambridge Monogr. Appl. Comput. Math., vol. 17, Cambridge Univ. Press, Cambridge, 2005.

[12] Weron R., *Electricity price forecasting: A review of the state-of-the-art with a look into the future.* Int. J. Forecast, 30/4 (2014), 1030–1081.

[13] Ziel F., Steinert R., *Electricity price forecasting using sale and purchase curves: the X-Model.* Energy Econ. 59 (2016), 435–454.