Università di Padova – Dipartimento di Matematica

Scuole di Dottorato in Matematica Pura e Computazionale

Seminario Dottorato 2016/17

Preface

This document offers a large overview of the eight months' schedule of Seminario Dottorato 2016/17. Our "Seminario Dottorato" (Graduate Seminar) is a double-aimed activity. At one hand, the speakers (usually Ph.D. students or post-docs, but sometimes also senior researchers) are invited to think how to communicate their researches to a public of mathematically well-educated but not specialist people, by preserving both understandability and the flavour of a research report. At the same time, people in the audience enjoy a rare opportunity to get an accessible but also precise idea of what's going on in some mathematical research area that they might not know very well.
Let us take this opportunity to warmly thank the speakers once again, in particular for their nice agreement to write down these notes to leave a concrete footstep of their participation. We are also grateful to the collegues who helped us, through their advices and suggestions, in building an interesting and culturally complete program.

Padova, June 27th, 2017

Corrado Marastoni, Tiziano Vargiolu

Abstracts (from Seminario Dottorato's web page)

Wednesday 5 October 2016

Integer Linear Programming to solve Large-Scale problems

Veronica DAL SASSO (Padova, Dip. Mat.)

Integer linear programming is widely used to find optimal solutions to problems that arouse in the real world and are related to logistics, planning, management, biology and so on. However, if from a theoretical point of view it is easy to give a formulation for these problems, from a computational point of view their implementation can be impractical due to the high number of constraints and variables involved.
During this seminar I will present classical results for dealing with large-scale integer linear programs and their application to a particular bioinformatic problem, related to the study of the human genome, that helps recovering information useful to study diseases and populations' behaviours.

Wednesday 16 November 2016

Products of elementary and idempotent matrices and non-Euclidean PID's

Laura COSSU (Padova, Dip. Mat.)

It is well known that Gauss Elimination produces a factorization into elementary matrices of any invertible matrix over a field. Is it possible to characterize integral domains different from fields that satisfy the same property? As a partial answer, in 1993, Ruitenburg proved that in the class of Bézout domains, any invertible matrix can be written as a product of elementary matrices if and only if any singular matrix can be written as a product of idempotents.
In this seminar we present some classical results on these factorization properties and we focus, in particular, on their connection with the notion of weak-Euclidean algorithm. We then conclude with a conjecture on non-Euclidean principal ideal domains, rare and interesting objects in commutative algebra, and some related results.
In order to make the talk understandable to a general audience, we will recall basic definitions of Commutative Ring theory and provide easy examples of the objects involved.

Wednesday 30 November 2016

Biologically inspired deduction of Optimal Transport Problems

Enrico FACCA (Padova, Dip. Mat.)

In this talk, after a brief introduction on the Optimal Transport Problems and some PDEs based formulation, we will present a recently developed approach, based on an extension of a model proposed by Tero et al (2007), for the simulation of the dynamics of Physarum Polycephalum, a unicellular slime mold showing surprising optimization ability, like finding the shortest path connecting two food sources in a maze. We conjecture that this model is an original formulation of the PDE-based OT problems. We show some theoretical and numerical evidences supporting our thesis.

Wednesday 14 December 2016

Extension fields, and classes in the genus of a lattice

Frances ODUMODU (Université de Bordeaux, France)

In this talk, which will be accessible to a large audience, a first part will be devoted to a basic reminder on extension fields with examples, and a second part to the more specific framework of number fields, i.e. finite degree extensions of rational numbers. Concerning the latter part, the Hasse-Minkowski local-global theorem for quadratic forms fails in general at the integral level, hence there are two levels of classification, the genus (local) and the integral class (global): we shall focus on some results concerning the classes in the genus of a lattice and in particular the trace form.

Wednesday 18 January 2017

Secure And Scalable Management of Internet of Things Deployments

Moreno AMBROSIN (Padova, Dip. Mat.)

In recent years, the advent of Internet of Things (IoT) is populating the world with billions of low cost heterogeneous interconnected devices. IoT devices are quickly penetrating in many aspects of our daily lives, and enabling new innovative services, ranging from fitness tracking, to factory automation. Unfortunately, their wide use, as well as their low-cost nature, makes IoT devices also an attractive target for cyber attackers, which may exploit them to perform various type of attacks, such as Denial of Service (DoS) attacks or privacy violation of end users. Furthermore, the potentially very large scale of IoT systems and deployments, makes the use of existing security solutions practically unfeasible.
In this talk I will give an overview of the problem of secure management, and present our research

effort in defining secure and scalable solutions for managing large IoT deployments. Moreover, I will focus in particular on two important parts of the device management process: (1) software updates distribution; and (2) device's software integrity check.

Wednesday 1 February 2017

Zeta functions associated to profinite groups

Leone CIMETTA (Padova, Dip. Mat.)

In this seminar we will discuss the properties of some Dirichlet series associated to a group G satisfying specific topological properties. These series deal with two important problems arisen in the last century, which both had a great development over the last decades.
The first problem (the subgroup growth of a profinite group G) involves the behaviour of the function $a_n(G)$, that is the number of subgroups of G of index n.
The second problem consists in determining the probability that, randomly choosing n elements of a group, we get a generating set for the whole group.
The second problem, in particular, arises from a famous work by P. Hall, which solved it in 1936 in the finite case.
After recalling some basic definitions, we will present the motivations for the problems; then, starting from some examples and classical results for finite groups, we will give some ideas to develop both problems in the profinite case and show some relations between the series involved.

Wednesday 15 February 2017

Collective periodic behavior in interacting particle systems

Daniele TOVAZZI (Padova, Dip. Mat.)

Interacting particle systems constitute a wide class of models, originally motivated by Statistical Mechanics, which in the last decades have become more and more popular, extending their applications to various fields of research such as Biology and Social Sciences. These models are important tools that may be used to study macroscopic behaviors observed in complex systems. Among these phenomena, a very interesting one is collective periodic behavior, in which the system exhibits the emergence of macroscopic rhythmic oscillations even though single components have no natural tendency to behave periodically.
This talk aims to introduce to a general audience some basic tools in the theory of interacting particle systems and some of the mechanisms which can enhance the appearance of self-sustained macroscopic rhythm. After recalling some notions of Probability, we present the classical Curie-Weiss model, which doesn't exhibit periodic behavior, and we show how we can modify it in order to create macroscopic oscillations. This is also the starting point for some recent developments that will be sketched in the last part of the talk.

Wednesday 1 March 2017

Topology, analysis and the Riemann-Hilbert correspondence

Christopher LAZDA (Padova, INdAM Marie Curie Fellow)

The Riemann-Hilbert correspondence gives a way of passing back and forth between topology and differential geometry, describing the behaviour of differential equations in terms of the monodromy of their local solutions. Starting with the example of the logarithm, I will give an introduction to the ideas behind this correspondence in a concrete and down to earth manner, concentrating on the case of Riemann surfaces. If there is time I will also explain how this gives a completely algebraic way to study topological invariants.

Wednesday 15 March 2017

Quantized option pricing in Mathematical Finance

Lucio FIORIN (Padova, Dip. Mat.)

Quantization is a widely used tool in Signal Processing and Numerical Probability, and it has been recently applied to Mathematical Finance. The quantization of a continuous random variable consists in finding the "best" discrete version of it, i.e. minimizing the L^2 distance. It is possible, using this technique, to create new algorithms for the pricing of European options under different models of the underlying asset.
In this seminar we introduce the basic tools used in mathematical finance and we will present the most common results in the theory of option pricing. After a brief discussion on the existing models of the price of a financial asset, we will give the audience some ideas on how quantization can be a powerful tool able to overcome existing problems.

Wednesday 29 March 2017

An introduction to domain perturbation theory for elliptic eigenvalue problems

Francesco FERRARESSO (Padova, Dip. Mat.)

How does the sound of a drum depend on its shape? This weak variant of the classical question "Can one hear the shape of a drum?"can be considered in the framework of domain perturbation theory for elliptic differential operators. Starting with easy examples we will see that the answer to this apparently harmless question is rather different in the case of regular perturbations and in the case of singular perturbations. We will focus on the singular case, where the geometry of the

problem is deeply mixed with the differential structure, in particular with the boundary conditions. Finally, we will give an account of recent advances in the study of a specific singular perturbation (the dumbbell domain) for the Laplace operator and for the biharmonic operator.
The seminar is intended for a general audience and it aims to introduce basic concepts from spectral theory as well as more advanced research results.

Wednesday 3 May 2017

Variational Approaches in Shape Partitioning

Martin HUSKA (Padova, Dip. Mat.)

The rapid development of 3D scanning technology has incredibly increased the availability of digital models exploited for a wide range of applications varying from computer graphics and medical imaging up to industrial production. One fundamental procedure that processes the raw acquired data for further manipulation, e.g. in product design, animation, deformation and reverse engineering, is the shape partitioning. This process consists in the decomposition of an object into non-overlapping salient sub-parts determined by a shape attribute.
In this seminar, we will introduce the concept of Shape Partitioning together with the wide range of partitioning methods. Next, we will observe a few partitioning/segmentation models in the field providing some results. At last, if the time allows, we will introduce the concept of Convex-Nonconvex segmentation over surfaces.
The seminar will be held at introductory level, thus, general audience is welcome to participate.

Wednesday 31 May 2017

The influence of network structure in neuronal information transmission

Giacomo BAGGIO (Padova, DEI)

Understanding how neurons communicate is one of the most challenging open problems in neuroscience. In this talk, I will present some recent results aiming at formulating this problem from a mathematical and information-theoretic viewpoint. After an overview on neuronal network dynamical models, I will introduce a digital communication framework for studying the information transmission problem in a neuronal network driven by linear dynamics. Within this framework, a novel metric for measuring the information capacity of a neuronal network based on Shannon?s capacity and the notion of inter-symbol interference will be discussed. Finally, I will illustrate how the structure of the network matrix and, in particular, its departure from normality, affects the information capacity of a network.
The talk will be introductory in nature and it is intended for a general audience.

Wednesday 14 June 2017

Biodiversity: Mathematical Modelling and Statistics

Anna TOVO (Padova, Dip. Mat.)

Ecological systems are characterized by the emergence of universal patterns that are deemed to be insensitive to the details of the system. Such universality motivates the understanding of ecological patterns through mathematical models able to grasp basic mechanisms at work. With this talk, we will try to describe and analyze the elements that underlie these patterns as well as the patterns themselves from a mathematical point of view. In particular we will focus on biodiversity. Identifying and understanding the relationships between all the life on Earth are some of the greatest challenges in science. After a brief introduction aiming to define the basic concepts of biodiversity and its related patterns, we will see different models developed to predict and measure them. We will then tackle the problem of upscaling biodiversity through spatial scales and we will discuss some still open problems that interest the scientific community.
The seminar is intended for a general audience and it will thus be held at an introductory level.

Integer Linear Programming to solve Large-Scale problems

Veronica Dal Sasso (*)

Abstract. Integer linear programming is widely used to find optimal solutions to problems that arise in the real world and are related to logistics, planning, management, biology and so on. However, if from a theoretical point of view it is easy to give a formulation for these problems, from a computational point of view their implementation can be impractical due to the high number of constraints and variables involved.
In this note I will present classical results for dealing with large-scale integer linear programs and their application to a particular bioinformatic problem, related to the study of the human genome, that helps recovering information useful to study diseases and populations' behaviours.

1 Introduction

Operations research is a branch of Mathematics that deals with a wide range of problems, arising in different contexts such as logistics, supply chain management, vehicle routing, facility location, scheduling, biology. In particular, it deals with those problems that can be formulated as the minimization or maximization of a function (the objective function) subject to some constraints. Moreover, if both the objective function and the constraints are linear on the variables involved, these problems can be formulated by means of *linear* or *mixed integer linear programs*.

Different procedures, shown in Section 2, are used to exactly solve these programs. However, for some applications it can happen that we need to deal with a large amount of data, thus finding the optimal solution can be impractical as the standard approaches do not provide a good performance. It is necessary to combine them with extra procedures that allow us to deal with large amount of data. This will be shown in Section 3. In Section 4 we will then present how these procedures to find a solution of large-scale problems can be applied to a computational biology problem, called the Haplotype Inference by Pure Parsimony problem.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: veronica.dalsasso@gmail.com . Seminar held on October 5th, 2016.

2 Solving Linear and Mixed-Integer Linear programs

Given vectors $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ and a matrix $A \in \mathbb{R}^{m \times n}$, we introduce a set of continuous variables x_i, $i \in N = \{1, \dots, m\}$. A linear program takes the following form:

$$\max \quad cx \tag{1}$$
$$\text{s.t.} \quad Ax \leq b \tag{2}$$
$$x \geq 0 \tag{3}$$

where the set identified by $P := \{x \geq 0 \,|\, Ax \leq b\}$ describes the polyhedron of all possible solutions, called the feasible region.

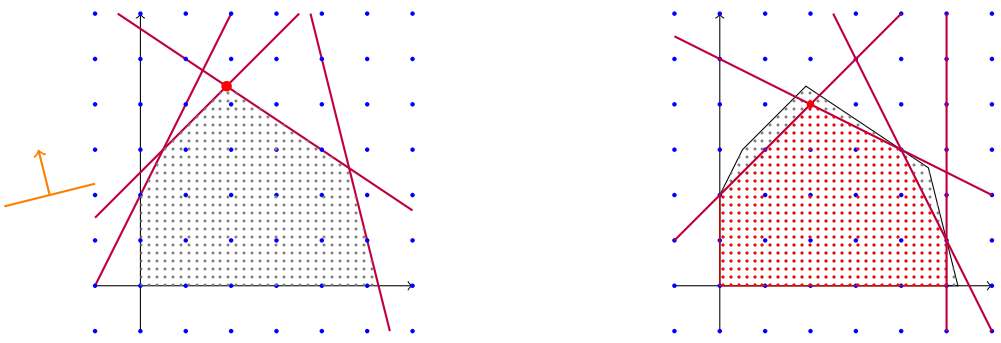
There are different algorithms that are used to solve a linear program, as can be seen for example in [4, 9]. Among these, the *Simplex Method* is widely used. It is a systematic algorithm that proceeds by moving from a vertex to another one of the feasible region, improving the value of the current solution, until it reaches optimality.

If some variables $x_i, i \in I \subseteq N$ are also required to be integer, we have the following mixed-integer linear program:

$$\max \quad cx \tag{4}$$
$$\text{s.t.} \quad Ax \leq b \tag{5}$$
$$x \geq 0 \tag{6}$$
$$x_i \in \mathbb{Z} \quad \forall\, i \in I \tag{7}$$

that becomes a pure integer program when $I = N$. We define $Q := \{x \geq 0 \,|\, Ax \leq b, x_i \in \mathbb{Z} \forall\, i \in I\}$.

In the presence of integer variables, the simplex method can still be applied to the linear relaxation of the problem, which results in having P as feasible region. However, the solution we get can be infeasible for the original problem, as there is no guarantee that x_i is integer for each $i \in I$. This is due to the fact that the simplex method gives as solution the optimal vertex of P, that is not guaranteed to be in Q, as can be seen in Figure 1(a).



(a) Linear relaxation (b) Convex hull

Figure 1. Linear relaxation and convex hull.

There are particular cases in which the integer optimal solution can be obtained as solution of a linear program: when the feasible region has integer vertices. Thus, the integer optimal solution can be found as a vertex of a different polyhedron, let us say $\tilde{P} := \{x \geq 0 \,|\, \tilde{A}x \leq \tilde{b}\}$, as seen in Figure 1(b), that is in fact the set of points obtained as convex combination of two any feasible points of Q.

Definition 1 (Convex hull) This polyhedron $\tilde{P}$ is called the *convex hull* ($\text{conv}(Q)$) of the feasible points.

Notice that $\tilde{P}$ has all vertices in Q. However, the convex hull is not always easy to be identified and for these reasons, alternative approaches are needed.

In order to tighten the feasible region of the program's linear relaxation, we can add *cutting planes* to the linear relaxation P. We want to find a hyperplane $cx \leq d$ that is able to divide the optimal solution $\tilde{x}$ given by the simplex method from the feasible region Q and hence from $\text{conv}(Q)$. In this way, we can update the feasible region P to $P \cap \{cx \leq d\}$ and the simplex method will give back a solution closer to the integer optimal one, as seen in Figure 2.

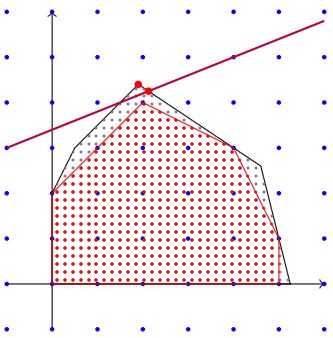
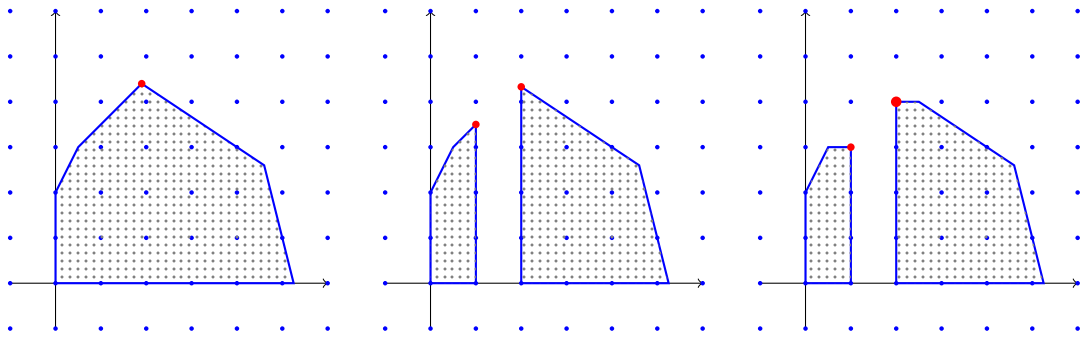


Figure 2. Example of a cutting plane.

As it is not always easy to reach the integer optimal solution adding cutting planes, the optimal solution is seeked applying a *branching scheme*, that is based on the following idea: suppose x_i should be an integer variable that takes a fractional value α in the solution of the linear relaxation. As this means that the solution is infeasible for our original problem, we can force x_i to take either a value less or equal to $\lfloor \alpha \rfloor$ or greater or equal to $\lfloor \alpha \rfloor + 1$. In this way we are splitting the feasible region into two polyhedra, whose union contains the optimal integer solution but leaves out the solution of the linear relaxation. At this point, we reoptimize over the two different polyhedra. Note that while the solution found is integer-infeasible, we keep on branching on different variables. An example of this procedure, called *branch-and-bound*, is shown in Figure 3. This recursive method can be represented using a tree structure, where each node branches into two child nodes, each of which is associated with a portion of the feasible region of the parent node. In Figure 4 we represent the branching tree associated to the branching procedure shown in Figure 3. Branching at a node is prevented if one of the following scenario happens: the solution is integer feasible, or the problem at the node is infeasible, or the best possible solution in the subtree is worse than a feasible solution already found.



(a) Original feasible region (b) First branching (c) Second branching

Figure 3. Example of branching scheme

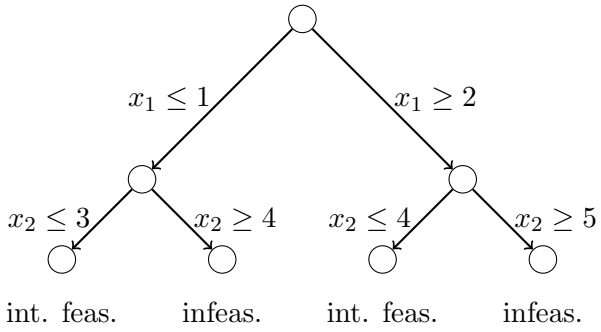


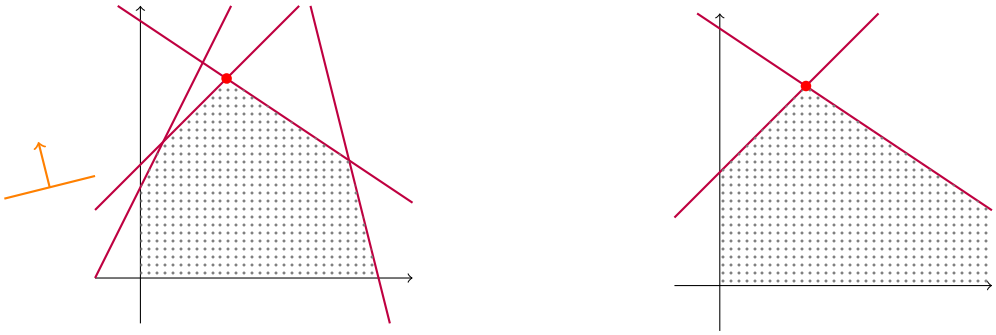
Figure 4

3 Methods for Large-Scale optimization problems

The methods presented so far are effective when the total data to be considered is limited. When dealing with a large-scale problem, instead, extra strategies are necessary to identify only a subset of variables and constraints that are sufficient to determine the optimal solution.

3.1 Linear programs

Consider the case in which a linear program has a large number of constraints. As seen in Figure 5, the optimal solution can be found by optimizing over a different feasible region, where only a subset of the original constraints is taken into account: given the direction in which the objective function improves, we see in Figure 5(b) how, by leaving out of the description of the feasible region two halfspaces, we do not affect the final solution of the problem.



(a) Original feasible region.

(b) Alternative feasible region.

Figure 5. Row generation procedure.

The *row generation* procedure, that allows us to identify a subset of constraints sufficient to find the optimal solution, works as follows. We define a subset of constraints $A'x \leq b'$, where A' is a subset of rows of A and b' is a vector made of the entries of b corresponding to the rows in A'. Then, we solve the reduced model and get the solution $\tilde{x}$. We check if this solution is feasible for the original formulation or if there is a violated constraint. In the first case, we can say that $\tilde{x}$ is optimal and stop the algorithm, otherwise we add the found constraint and solve again the resulting linear program. The core point of this method is the detection of a procedure that gives back a violated constraint, that strongly depends on the structure of the program and on the meaning underlying the involved constraints.

If, instead, the original program has a huge number of variables, we start with a reduced set of variables x_i, $i \in S \subseteq N$ that form the *Restricted Master Problem RMP* and we proceed with a *column generation* procedure. This procedure needs to consider the dual problem associated to the original (also called primal) problem, that in particular accociates a dual value to each constraint of the primal program. We optimize the RMP and get the solution $\tilde{z}$ and the dual values $\tilde{\pi}$. If $\tilde{\pi}$ is feasible for the dual problem, then the solution found is optimal and we stop the procedure. Otherwise, we need to solve a subproblem, called *pricing problem*, in order to find a dual constraint violated by $\tilde{\pi}$, that identifies a suitable variable to be added to the RMP. It can be easily seen that the column generation procedure has been translated in a row generation procedure for the dual program.

3.2 Mixed integer linear programs

The same procedures for large-scale problems seen for the linear case can be embedded into the procedures used to solved mixed integer linear programs. In particular, if we need to deal with a large amount of constraints we apply a *branch-and-cut* procedure: the root node is solved as seen for the linear programs, then, each time we branch and explore a new node, we have at first to look for violated constraints that, due to the partitioning of the feasible region, can now be necessary.

When dealing with a large number of variables, instead, we apply a *branch-and-price*

procedure. Given a node of the branching tree, its feasible region is given by the feasible region of the parent node intersected with the branching constraint. Thus, we do not know if the set of variables considered so far is enough to ensure we get the optimal solution and we need to look for variables that can improve the current objective value.

4 Application to the Haplotype Inference by Pure Parsimony problem

The methods shown so far are now applied to find the optimal solution to a problem arising in computational biology.

One of the most important achievements of the latest years in biology has been the human genome sequencing, completed in 2001, that has shown how all humans share the 99% of the information contained in the DNA, while all the significant differences are contained in the remaining information. Each site of this 1% portion of the human genome, that presents a significant variability among the individuals, is called a *Single Nucleotide Polymorphism (SNP)*.
Humans are diploid organisms, meaning that the DNA is organized in pairs of chromosomes, each copy coming from one of the two parents. Every single chain in the DNA is made of a sequence of nucleotides, each of which is made of a phosfate group, a five-sided sugar and a nitrogenous base. The nucleotide is fully characterized by the base, that can be chosen among the four: Adenine (A), Thymine (T), Cytosine (C), Guanine (G). It is known that, regarding human beings, the DNA sites and so also the SNP sites are almost always biallelic, meaning that at each site only two of the four nucleotides can be found. Thus, we can possibly encode the information for each SNP using only two symbols, 0 and 1. If the nucleotide is equal for both chains, then the SNP is *homozygous*, otherwise it is *heterozygous*. We give the following definitions.

Definition 2 (Haplotype) A *haplotype* is a sequence of values 0 and 1 that represents the single chain of SNP values for a specific portion of a chromosome copy.

Definition 3 (Genotype) A *genotype* is the chain providing information regarding the union of the two chromosome copies, that tells us if each SNP in the chain is homozygous, if it takes value within the alphabet $\{0, 1\}$, or heterozygous, that is thus denoted with a value 2.

Moreover, we say that two haplotypes resolve a certain genotype if, when paired, the information regarding homozygous and heterozygous sites they give is the same provided by that genotype. In particular:

Definition 4 (Resolving haplotypes) Two haplotypes h^1 and h^2 resolve a genotype g if, for every position p such that $g_p = 2$, we have $h^1_p \neq h^2_p$ and $h^1_p = h^2_p$ otherwise.

Haplotypes have an important role in medical and pharmacologic studies, for example to detect diseases or to study the different behaviour of various individuals to the same therapy. Sequencing them is not practical, as it is very expensive and time consuming, while it is easier to experimentally obtain the information stored in genotypes. We are

then facing the *haplotyping* problem, that consists in determining the two haplotypes that resolve a given genotype. Several approaches have been used in order to solve this problem, its difficulty consisting in the fact that, once we have k heterozygous SNPs in the same genotype, we have 2^{k-1} possible pairs of haplotypes that can represent it and we need some criteria to choose the right pair. A classical approach is to apply the *Pure Parsimony* criterion, according to which, given a set of genotypes obtained by a family of individuals, we want to select the minimum number of haplotypes that can resolve all the genotypes.

Thus, we define the problem as follows:

Definition 5 (Haplotype Inference by Pure Parsimony) The Haplotype Inference by Pure Parsimony (HIPP) problem cosists in, given a set of genotypes G, finding a set of haplotypes H such that each genotype in G is resolved by a pair of its haplotypes and H has minimum cardinality.

HIPP is well known to be NP-hard [7] and different mathematical programming approaches have been investigated. Among the others, we cite [1, 2, 8]. We present here a new formulation [6] that involves an exponential number of variables and a polynomial number of constraints. Our goal is to show how this problem can be solved in a competitive way, compared to another state-of-the-art polynomial formulation [3].

Let G be a set of m genotypes of length n, $K \subseteq G$ the subset of genotypes with at least one heterozygous site. We define a set of objects Q whose elements are $q = (h^q, G^q)$, that we call Q-pairs and are made of a haplotype h^q and a subset of genotypes of K that are compatible with h^q. Notice that a solution to the HIPP problem can be seen as a collection of Q-pairs, where if a Q-pair q is used it means that the solution uses that particular haplotype h^q to partially solve each genotype in G^q. We introduce a binary variable λ^q for each of these Q-pairs, that will take value 1 if q is used in the solution, 0 otherwise. The HIPP problem can be formulated as follows:

$$(8) \qquad (HI) \ \min \sum_{q \in Q} c^q \lambda^q \qquad + (m - |K|)$$

$$(9) \qquad s.t. \sum_{q : g^k \in G^q} \lambda^q = 2 \qquad \forall\, k = 1, \dots, |K|$$

$$(10) \qquad \sum_{\substack{q : g^k \in G^q \\ h^q_p = 1}} \lambda^q = 1 \qquad \forall\, k = 1, \dots, |K|,\ p = 1, \dots, n : g^k_p = 2$$

$$(11) \qquad \lambda^q \in \{0, 1\} \qquad \forall\, q \in Q$$

where constraints (9) ensure that each genotype is contained in exactly two selected Q-pairs, while constraints (10) say that, among the two haplotypes chosen for each genotype, only one has value 1 at each heterozygous position p (the other has necessarily value 0).

As this formulation has a polynomial number of constraints but an exponential number of variables, column generation and branch-and-price procedures are required to solve it more efficiently.

Starting from a reduced subset of variables, we need to solve the pricing subproblem

to see if there exists another variable that can possibly improve the value of the current solution. In our case, the subproblem takes the following shape:

$$(12) \quad (\text{PP})\ z = \min \quad c_q - \sum_{k=1}^{m'} \bar{\pi}^k \chi^k + \sum_{k=1}^{m'} \sum_{p=1,\dots,n: g_p^k = 2} \bar{\mu}_p^k \, \zeta_p \chi^k$$

$$(13) \quad \text{s.t.} \quad \zeta_p \leq 1 - \chi^k \qquad \forall\, k = 1, \dots, |\tilde{K}|,\ p = 1, \dots, n : g_p^k = 0$$

$$(14) \quad \zeta_p \geq \chi^k \qquad \forall\, k = 1, \dots, |\tilde{K}|,\ p = 1, \dots, n : g_p^k = 1$$

$$(15) \quad \chi^k, \zeta_p, \in \{0, 1\} \qquad \forall\, k = 1, \dots, |\tilde{K}|,\ p = 1, \dots, n$$

where a solution (ζ, χ) represents a Q-pair, in the sense that the entries of ζ determine a haplotype h^q and χ is the characteristic vector of a subset G^q of genotypes. Constraints (14) and (15) ensure compatibility between the selected haplotype and the subset of genotypes. The found Q-pair leads to add a new variable to the RMP if the optimal value z is negative. Notice that the objective function is not linear, as there is a quadratic term. In order to apply the methods seen for integer linear programs, we substitute the quadratic terms with a new set of variables w_p^k and linearize the pricing problem as follows:

$$(16) \quad (\text{LPP})\ z = \min \quad c_q - \sum_{k=1}^{m'} \bar{\pi}^k \chi^k + \sum_{k=1}^{m'} \sum_{p=1,\dots,n: g_p^k = 2} \bar{\mu}_p^k \, w_p^k$$

$$(17) \quad \text{s.t.} \quad \zeta_p \leq 1 - \chi^k \qquad \forall\, k = 1, \dots, |\tilde{K}|,\ p = 1, \dots, n : g_p^k = 0$$

$$(18) \quad \zeta_p \geq \chi^k \qquad \forall\, k = 1, \dots, |\tilde{K}|,\ p = 1, \dots, n : g_p^k = 1$$

$$(19) \quad w_p^k \leq \zeta_p \qquad \forall\, k = 1, \dots, |\tilde{K}|,\ p = 1, \dots, n : g_p^k = 2$$

$$(20) \quad w_p^k \leq \chi^k \qquad \forall\, k = 1, \dots, |\tilde{K}|,\ p = 1, \dots, n : g_p^k = 2$$

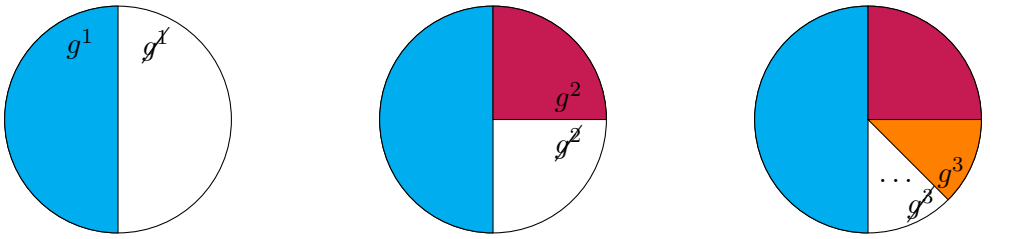
$$(21) \quad w_p^k \geq \chi^k + \zeta_p - 1 \qquad \forall\, k = 1, \dots, |\tilde{K}|,\ p = 1, \dots, n : g_p^k = 2$$

$$(22) \quad \chi^k, \zeta_p, \in \{0, 1\} \qquad \forall\, k = 1, \dots, |\tilde{K}|,\ p = 1, \dots, n$$

This program can be solved using a standard branch-and-bound approach, but it can possibly have a great number of variables and constraints, thus finding the optimal solution in this way can be computationally expensive. In [5] we propose a *Smart Enumeration* approach, i.e. an alternative way of solving (LPP) that exploits the structure of the involved variables. The main idea underlying this approach is the observation that, given a genotype $g^1 \in K$, the solution (h^q, G^q) of the pricing problem either will include g^1, or not. For compatibility reasons, if $g^1 \in G^q$ we can derive extra information on the shape of the solution, such as the values of some components of h^q, and already exclude from G^q those genotypes that are not compatible with g^1. In this way, the size of (LPP) can be significantly reduced, provided we suppose g^1 belongs to the optimal solution. We then need to look also to what happens when g^1 is not in the solution. Consider Figure 6, where the circle represents all the possible solutions (h^q, G^q) of the pricing problem. All the solutions in which $g^1 \in G^q$ are colored in blue in Figure 6(a). In Figure 6(b), we can see how the solutions not involving g^1 are divided into solutions in which $g^2 \in G^q$ (in purple) and those in which $g^2 \notin G^q$, that are again divided into solutions such that

$g^3 \in G^q$ (in orange in Figure 6(c)) or not. Thus, exploring recursively all the genotypes in K and solving the associated pricing problems with additional information, we are able to recover the optimal solution of the original pricing problem.

Proposition 1 *The Smart Enumeration procedure exactly solve (LPP).*



(a) $g^1 \in G^q$ (b) $g^1 \notin G^q$,$g^2 \in G^q$ (c) $g^1 \notin G^q$,$g^2 \notin G^q$, $g^3 \in G^q$

Figure 6. Idea of how Smart Enumeration proceeds.

In this way, we end up solving $|K|$ easy pricing problems per iteration of the column generation procedure. We can possibly further speed up the process if, instead of solving exactly (LPP), we stop as soon as we find a Q-pair such that the objective value of (LPP) is negative. In the worst case, this *Early terminated Smart Enumeration* needs to solve all the $|K|$ subproblems in order to find a variable to be added to the RMP or to be able to say that no such a variable exists.

Once we have a more efficient way of solving the linear relaxation of (HI), we proceed by applying a branching procedure to obtain the optimal integer solution. The easiest way of applying a branching scheme for our problem consists in selecting a variable that takes a fractional value $\alpha \in]0,1[$ and forcing it to be either less than or equal to 0 or greater than or equal to 1, as briefly explained in the previous section. In our case, however, having a large number of variables this scheme is sure to give back a highly unbalanced tree, as fixing a variable to 1 highly affects the structure of the solution, while the same does not happen fixing a variable to 0. We then apply another, more general, branching strategy that consists in finding a hyperplane $y = \eta x$ with integer coefficients η that, associated to the current linear solution, takes a fractional value δ. We then force this hyperplane to take either value less than or equal to $\lfloor \delta \rfloor$ or greater than or equal to $\lfloor \delta \rfloor + 1$. The scheme we present is based on the structure of a family of constraints of (HI) (23) and a family of redundant constraints (24):

$$\sum_{\substack{q \in Q : g^k \in G^q \\ h^q = 1}} \lambda^q = 1 \qquad \forall\, k = 1, \dots, |K|,\ p = 1, \dots, n : g^k_p = 2 \tag{23}$$

$$\sum_{\substack{q \in Q : g^k \in G^q \\ h^q = 0}} \lambda^q = 1 \qquad \forall\, k = 1, \dots, |K|,\ p = 1, \dots, n : g^k_p = 2 \tag{24}$$

Notice that, given two constraints of set (23) or (24) related to the same position p and two different genotypes g^s and g^t, for each feasible integer solution either the constraints are

fulfilled by the same variable, i.e. there is only one variable λ^q involved in both constraints that takes value 1, or by different ones. Starting from these considerations, we define a branching scheme based on the following conditions:

$$\sum_{\substack{q\in Q:\, g^s,g^t\in G^q \\ h^q_p=1}} \lambda^q \in \{0,1\} \qquad \forall\, k=1,\dots,|K|,\ p=1,\dots,n : g^s_p = 2 \tag{25}$$

$$\sum_{\substack{q\in Q:\, g^s,g^t\in G^q \\ h^q_p=0}} \lambda^q \in \{0,1\} \qquad \forall\, k=1,\dots,|K|,\ p=1,\dots,n : g^s_p = 2 \tag{26}$$

We branch identifying any violated condition of the kind (25) or (26) and forcing the associated sum to take either value 0 or 1. We distinguish the branching strategy according to whether we also have g^t_p equal to 2 or if $g^t_p \in \{0,1\}$. In fact, if $g^t_p = \beta$ with $\beta \in \{0,1\}$, for compatibility reasons only one of the sets $\{q \in Q : g^s, g^t \in G^q \wedge h^q_p = 1\}$ and $\{q \in Q : g^s, g^t \in G^q \wedge h^q_p = 0\}$ is not empty, thus it makes sense to check only one between conditions (25) and (26). In this case, the branching at the node produces two child nodes. We enumerate here the types of child nodes created and the corresponding inequalities that has to be added to the feasible region of the parent node.

DIFFER g^s and g^t do not belong to the same Q-pair:

$$\sum_{\substack{q:\, g^s,g^t\in G^q \\ h^q_p=\beta}} \lambda^q \le 0$$

SAME g^s and g^t belong to the same Q-pair:

$$\sum_{\substack{q:\, g^s,g^t\in G^q \\ h^q_p=\beta}} \lambda^q \ge 1$$

If $g^t_p = 2$, both genotypes can be contained in a Q-pair q having $h^q_p = 0$ or $h^q_p = 1$, so that a single child node is not enough to describe correctly the case in which both genotypes are resolved by a common haplotype. Thus, we need to generate three different child nodes to properly give a partition of the feasible region:

BI-DIFFER g^s and g^t do not belong to the same Q-pair:

$$\sum_{\substack{q:\, g^s,g^t\in G^q \\ h^q_p=1}} \lambda^q \le 0 \ \wedge \ \sum_{\substack{q:\, g^s,g^t\in G^q \\ h^q_p=0}} \lambda^q \le 0$$

SAME0 g^s and g^t belong to the same Q-pair and the associated haplotype has value 0 in position p:

$$\sum_{\substack{q:\, g^s,g^t\in G^q \\ h^q_p=0}} \lambda^q \ge 1 \ \wedge \ \sum_{\substack{q:\, g^s,g^t\in G^q \\ h^q_p=1}} \lambda^q \le 0$$

SAME1 g^s and g^t belong to the same Q-pair and the associated haplotype has value 1 in position p:

$$\sum_{\substack{q:\, g^s, g^t \in G^q \\ h^q_p = 0}} \lambda^q \leq 0 \;\wedge \sum_{\substack{q:\, g^s, g^t \in G^q \\ h^q_p = 1}} \lambda^q \geq 1$$

Proposition 2 *The proposed branching scheme is feasible, that is it cuts off the optimal solution of the parent node, it partitions the feasible region and it ensures integrality of the solution at the end of the procedure.*

The application of these procedures, the Smart Enumeration for solving the pricing problem and the branching strategy just presented, together with other refinements that reduce computational issues due to numerical instability, make the solution process of (HI) competitive compared with a state-of-the-art polynomial-sized model. Some results on this fact are presented in Table 1, for which we tested 100 instances made of a number of genotypes equal to 80, 90, 100, 110 or 120 and 20 or 30 SNPs.

	%solved	%faster	time	#nodes	#vars
PIP	97.00	-	2448.53	5.24	30509.68
SM	93.00	88.00	816.17	5.08	1956.32
ESM	100.00	100.00	119.50	4.42	2092.53

Table 1

In this table we compare the performance of the model presented in [3], denoted by PIP, with our branch-and-price approach, where we solve the pricing problem either with the Smart Enumeration (SM) or with the Early terminated Smart Enumeration (ESM) procedures. It is easily seen that the computational times are greatly reduced.

References

[1] Bertolazzi P., Godi A., Labbé M., Tininini L., *Solving Haplotyping Inference Parsimony Problem using a new basic polynomial formulation.* Computers and Mathematics with Applications 55 (2008), 900–911.

[2] Brown D., Harrower I., "Integer Programming Approaches to Haplotype Inference by Pure Parsimony", 2006.

[3] Catanzaro D., Godi A., Labbé M., *A class representative model for pure parsimony haplotyping.* INFORMS Journal on Computing 22(2) (2010), 195–209.

[4] Conforti M., Cornuéols, G., Zambelli, G., "Integer programming", 2014.

[5] Dal Sasso V., De Giovanni L., Labbé M., *A column generation approach for Pure Parsimony Haplotyping*. Proceedings 5th Student Conference on Operations Research (SCOR2016), OpenAccess Series in Informatics (OASIcs) 50 (2016), 1–11.

[6] De Giovanni L., Labbé M., *A column generation approach for pure parsimony haplotyping*. CBBM 2014 Conference, Poznan, 26-28/06/2014.

[7] Lancia G., Pinotti M.C., Rizzi, R., *Haplotyping populations by pure parsimony: Complexity of exact and approximation algorithms*. INFORMS Journal on computing 16/4 (2004), 348–359.

[8] Lancia G., Serafini P., *A set-covering approach with column generation for parsimony haplotyping*. INFORMS Journal on Computing 21/1 (2009), 151–166.

[9] Korte B., Vygen J.,, "Combinatorial optimization". Springer, 2012.

Products of elementary and idempotent matrices and non-Euclidean PID's

Laura Cossu (*)

Abstract. It is well known that Gauss Elimination produces a factorization into elementary matrices of any invertible matrix over a field. Is it possible to characterize integral domains different from fields that satisfy the same property? As a partial answer, in 1993, Ruitenburg proved that in the class of Bézout domains, any invertible matrix can be written as a product of elementary matrices if and only if any singular matrix can be written as a product of idempotents. In this article we give an overview of the classical results on these two factorization properties and we focus, in particular, on their connection with the notion of *weak-Euclidean algorithm*. We conclude presenting an open conjecture on non-Euclidean principal ideal domains, rare and interesting objects in commutative algebra, and some related results. The dissertation has been thought for a general reading audience, so we will recall basic definitions of commutative ring theory and provide easy examples of the objects involved.

1 Notation and basic definitions

The aim of this section is to fix the notation and to recall some basic definitions and results useful for the sequel. For further details we refer to Kaplansky's book [10].

In what follows R will always denote an integral domain, i.e. a nonzero commutative ring with no nonzero zero divisors, and $U(R)$ its multiplicative group of units.

Definition 1.1 An *ideal* I of R is an additive subgroup of R such that $rI \subseteq I$ for every $r \in R$.
An ideal I of R is said to be *finitely generated* if there exist $a_1, \dots, a_n \in R$ such that $I = \{r_1a_1 + \cdots + r_na_n \,|\, r_1, \dots, r_n \in R\} = \sum_{i=1}^{n} a_iR$. We will denote the ideal generated by $a_1, \dots, a_n \in R$ as $\langle a_1, \dots, a_n \rangle$.
An ideal I is said to be *principal* if it is generated by a single element $a \in R$. In this case we will write $I = aR$.
A proper ideal P of R is said to be *prime* if it satisfies the following property: if a and b are two elements of R such that $ab \in P$, then $a \in P$ or $b \in P$.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: lcossu@math.unipd.it . Seminar held on November 16th, 2016.

A *maximal* ideal $\mathfrak{M}$ of R is an ideal of R such that, if $I \subseteq R$ is an ideal of R that contains $\mathfrak{M}$, then $I = R$ or $I = \mathfrak{M}$. Every maximal ideal is a prime ideal.

Definition 1.2 An integral domain R is said to be a *principal ideal domain* (PID for short) if every ideal of R is principal. An integral domain R such that every finitely generated ideal of R is principal is called a *Bézout domain*.

Clearly, every principal ideal domain is also a Bézout domain.

Example 1.3 Any field K, the ring of integers $\mathbb{Z}$, the ring $K[X]$ of the polynomials over a field K in one indeterminate X, are examples of PID's. The ring $R = \mathbb{Z} + X\mathbb{Q}[X]$ is an example of Bézout domain that is not a PID. In fact, the ideal $X\mathbb{Q}[X]$ is not finitely generated.

We will need also the following

Definition 1.4 An integral domain R is said to be *local* if it has a unique maximal ideal $\mathfrak{M}$. A *valuation domain* is an integral domain R such that, for every $a, b \in R$, either $a|b$ or $b|a$. Every valuation domain is a local domain.

It is well known that a local domain is a Bézout domain if and only if it is also a valuation domain.

As usual, $M_n(R)$ denotes the R-algebra of the $n \times n$ matrices with entries in R. A matrix $\mathbf{M} \in M_n(R)$ is said to be *invertible* if $\det(\mathbf{M}) \in U(R)$; *singular* if $\det(\mathbf{M}) = 0$.
We will be particularly interested in *elementary* and *idempotent* matrices, a special kind of invertible and singular matrices respectively.

Definition 1.5 An *idempotent* matrix is a square matrix $\mathbf{M}$ such that $\mathbf{M}^2 = \mathbf{M}$.

It can be checked by a direct computation that every 2×2 non-identity idempotent matrix over R has the following standard form:

$$\begin{pmatrix} a & b \\ c & 1-a \end{pmatrix}, \text{ with } a(1-a) = bc.$$

Definition 1.6 An *elementary matrix* of dimension n is a square matrix obtained by applying elementary transformations to the identity matrix $\mathbf{I}_n$. There exist three different types of elementary matrices, corresponding respectively to three different types of elementary transformations:

- *transpositions* $\mathbf{P}_{ij}$, with $i \neq j$, obtained from the identity matrix $\mathbf{I}_n$ by exchanging row i and row j;
- *dilations* $\mathbf{D}_i(u)$, obtained from $\mathbf{I}_n$ by multiplying row i by the unit $u \in U(R)$;
- *transvections* $\mathbf{T}_{ij}(r)$, with $i \neq j$ and $r \in R$, obtained from $\mathbf{I}_n$ by adding to row i, r times row j. It turns out that $\mathbf{T}_{ij}(r)$ is nothing but the identity matrix with r in the ij position.

Example 1.7 The following matrices are respectively an example of transposition, dilation and transvection:

$$\mathbf{P}_{12} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \mathbf{D}_1(u) = \begin{pmatrix} u & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \mathbf{T}_{23}(r) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & r \\ 0 & 0 & 1 \end{pmatrix},$$

where $u \in U(R)$ and $r \in R$.

2 Introduction and motivations

2.1 Factorization properties (GE_n) and (ID_n)

Our discussion starts from two major properties concerning the factorization of square matrices over an integral domain R.
We say that an integral domain R satisfies the property

- (GE_n) if any *invertible* $n \times n$ matrix over R is a product of *elementary* matrices;
- (ID_n) if any *singular* $n \times n$ matrix over R is a product of *idempotent* matrices.

The problem of characterizing integral domains satisfying properties (GE_n) and (ID_n), for all $n > 0$, has been considered since the middle of the 1960's.

The main impulse to investigate the property (GE_n) comes from the fundamental 1966 paper by Cohn, *On the structure of the GL_2 of a ring*, [6]. Cohn observed that the study of the general linear group $GL_n(R)$ of a ring R is related to the study of the property (GE_n) on R, and he called the rings satisfying (GE_n), for all $n > 0$, *generalized-Euclidean* (*GE-rings* for short). In fact Euclidean domains are the second instance, after the fields, of integral domains in which every invertible matrix is product of elementary matrices.
In 1967, J.A. Erdos initiated in [8] the study of integral domains satisfying property (ID_n) for all $n > 0$. Generalizing a 1966 Howie's result, *every transformation of a finite set that is not a permutation can be written as a product of idempotents*, Erdos proved that fields satisfy property (ID_n) for any $n > 0$.

2.2 Classical results

In this section we give an overview of the first and main results on the study of the properties (GE_n) and (ID_n) over an integral domain R, starting from the easiest cases.
When R is a field, it is well known from Gauss Elimination algorithm, that R satisfies property (GE_n), for every $n > 0$. Moreover, as we said in the previous section, Erdos proved that R also satisfies property (ID_n) for all $n > 0$.
The same is true for Euclidean domains (we refer to Section 3 for a precise definition). In fact, when R is Euclidean, it was proved by van der Waerden in 1937 that every invertible matrix over R is a product of elementary matrices; while Laffey , in 1983, proved that R satisfies also property (ID_n) for all $n > 0$, i.e. every singular matrix over an Euclidean domain can be written as a product of idempotents.

The situation is different when R is a PID. Principal ideal domains do not satisfy in general the property (GE_n) for any n. We can find in [6] and [2] some examples of PID's that are not generalized Euclidean. Since every Euclidean domain satisfies property (GE_n) for all $n > 0$, these particular principal ideal domains can not be Euclidean. The property (ID_n) has been studied in the class of PID's by Fountain in 1991 (cf. [9]). He found some properties equivalent to (ID_n) in this environment and, using these new characterizations, he proved that discrete valuation rings (i.e. local PID's that are not fields) and $\mathbb{Z}$ satisfy (ID_n) for all $n > 0$.

When Laffey proved that Euclidean domains satisfy property (ID_n) for all $n > 0$, a crucial part of his proof was a *reduction argument* from any dimension $n > 0$ to dimension 2:

when R is an Euclidean domain, all singular matrices are products of idempotent matrices if and only if all 2×2 singular matrices are products of idempotent matrices.

Laffey's reduction argument was extended by Bhaskara Rao (cf. [3]) to principal ideal domains, and by Salce and Zanardo (cf. [15]) to Bézout domains. Therefore,

if R is a Bézout domain, R satisfies property (ID_2) if and only if it satisfies property (ID_n), far all $n > 0$.

A sort of reduction argument holds also for the property (GE_n) in the class of Bézout domains. In fact, a celebrated result by Kaplansky, Theorem 7.1 in [11], says that

if R is a Bézout domain, all invertible matrices are products of elementary matrices if and only if all 2×2 invertible matrices are products of elementary matrices.

Summing up the above results, it follows that a Bézout domain satisfies property (ID_2) if and only if it satisfies (ID_n) for all $n > 0$, and it satisfies (GE_2) if and only if it satisfies (GE_n) for all $n > 0$. Hence, to study the factorization of square matrices with entries in a Bézout domain into elementary or idempotent matrices, it is enough to consider the 2×2 case.

A natural question that arises at this point is: are properties (GE_n) and (ID_n) related to each other? The answer is yes, at least for Bézout domains.
The main result in Ruitenburg's 1993 paper [14], given in a simplified version according to our purposes, says that:

Theorem 2.1 (Ruitenburg - 1993) *For a Bézout domain R the following conditions are equivalent:*

(1) (ID_n) *holds for every integer $n > 0$;*

(2) (GE_n) *holds for every integer $n > 0$.*

By reduction arguments, if R is a Bézout domain R satisfies (ID_2) if and only if it satisfies (ID_n) for all $n > 0$, if and only if it satisfies (GE_n) for all $n > 0$, if and only if it satisfies (GE_2).

It is worth noting that the equivalence of the properties (GE_n) and (ID_n) in Ruitenburg's theorem is not valid outside Bézout domains. In fact, Cohn proved in [6] that local domains always satisfy property (GE_2), but Salce and Zanardo proved (cf. [15]) that a local domain satisfies property (ID_2) if and only if it is also a valuation domain. Therefore, local domains that are not valuation domains (equivalently that are not Bézout domains) satisfy (GE_2) but not (ID_2).

Example 2.2 $R = k[\![X, Y]\!]$ is a local domain ($\mathfrak{M} = \langle X, Y\rangle$) but it is not a valuation domain (both X/Y and Y/X are not in R); thus R satisfies property (GE_2) but not property (ID_2).

3 Weak Euclidean algorithm and the factorization properties

The aim of this section is to present the relation between the notion of *weak Euclidean algorithm* and our factorization properties (GE_n) and (ID_n). Let us start recalling the definition of *Euclidean algorithm*.

Definition 3.1 (Euclidean algorithm) An integral domain R admits an *Euclidean algorithm* if there exists a map

$$\varphi : R \to W,$$

called *algorithm*, with W a well-ordered set, such that for any pair of elements $a, b \in R$ with $b \neq 0$, there exists a finite sequence of relations

$$r_i = q_{i+1}r_{i+1} + r_{i+2}, \quad r_i, q_i \in R,\ -1 \leq i \leq n-2,$$

such that $a = r_{-1}$, $b = r_0$, $r_{n-1} \neq 0$, $r_n = 0$ and $\varphi(r_{i+1}) > \varphi(r_{i+2})$ for every i.
If R admits an Euclidean algorithm it is said to be an *Euclidean domain*.

It is well known and it can be easily proved that Euclidean domains are principal ideal domains.

Now, we are ready to give the following

Definition 3.2 (Weak Euclidean algorithm) An integral domain R admits a *weak (Euclidean) algorithm* if, for any pair of elements $a, b \in R$, there exists a finite sequence of relations

$$r_i = q_{i+1}r_{i+1} + r_{i+2}, \quad r_i, q_i \in R,\ -1 \leq i \leq n-2,$$

such that $a = r_{-1}$, $b = r_0$, $r_{n-1} \neq 0$ and $r_n = 0$. If R admits a weak algorithm it is said to be a weakly Euclidean domain.

The difference between a weak algorithm and an Euclidean algorithm is that, in the weak case, we don't have any order relation between the elements of R, so we can not assume a condition on the rests r_i analogous to the condition $\varphi(r_{i+1}) > \varphi(r_{i+2})$ of the Euclidean algorithm. Clearly, an Euclidean domain is also weakly Euclidean.

It follows from the definition that, if R is weakly Euclidean, then $\langle a, b\rangle = r_{n-1}R$ and R is also a Bézout domain.

Example 3.3 Valuation domains admit a weak algorithm. In fact, if:

– $a|b$, say $b = ca$ with $c \neq 0$, then

$$a = b + (a - b)$$
$$b = -(a - b) + a$$
$$(a - b) = a - ca = a(1 - c)$$

– $b|a$, then $a = bd$ with $c \neq 0$.

The link between the notion of weak Euclidean algorithm and the properties (GE_n) and (ID_n) comes from the following nice result by O'Meara.

Theorem 3.4 (O'Meara [13], Theorem 14.3) *Let R be a Bézout domain. Then TFAE:*

(i) *R admits a weak Euclidean algorithm;*

(ii) *R satisfies property (GE_2).*

Using Kaplansky's reduction argument and Ruitenburg's theorem, we can state the previous theorem in a more general form.

Theorem 3.5 (O'Meara generalized) *Let R be a Bézout domain. Then TFAE:*

(i) *R admits a weak Euclidean algorithm;*

(ii) *R satisfies property (GE_n) for all $n > 0$;*

(iii) *R satisfies property (ID_n) for all $n > 0$.*

4 A conjecture on non-Euclidean PID's

In this section we present a conjecture, proposed by Salce and Zanardo in [15], that concerns non-Euclidean PID's. We introduce this particular kind of principal ideal domains starting from Samuel's characterization of Euclidean domains (Proposition 9 of [16]).

4.1 Samuel's characterization for Euclidean domains

In Samuel's 1970 paper [16] it is shown that, among all Euclidean algorithms on a given domain R, there exists a unique *smallest* algorithm $\theta : R \to W$ that associates to any element of R the smallest value it can take among the values associated to the same element by each Euclidean algorithm of R. Therefore θ is called "smallest" because it is point-wise the minimum of all Euclidean algorithms. The smallest algorithm θ can be constructed by transfinite induction.
The transfinite construction of the smallest algorithm can be applied to any domain R and it is the following: let R be a domain and W an ordinal such that $\mathrm{card}(R) < \mathrm{card}(W)$, set

- $R_0 = \{0\}$ and,
- for $\alpha > 0$, set $R'_\alpha = \bigcup_{\beta<\alpha} R_\beta$ and

$$R_\alpha = \{0\} \cup \left\{ b \in R \,|\, R'_\alpha \to R/bR \,\text{is surjective}\right\}.$$

The sequence $(R_\alpha)_{\alpha\in W}$ is clearly increasing and we have that R is Euclidean if and only if $R = \bigcup_{\alpha\in W} R_\alpha$. In this case the smallest algorithm θ is defined as:

$$\theta(x) = \alpha \in W \Leftrightarrow x \in R_\alpha \setminus R'_\alpha$$

Note that at any rate we have:

- $R_0 = \{0\}$;
- $R_1 \setminus R_0 = R_1 \setminus R'_1 = U(R)$;
- $R_2 \setminus R_1 = R_2 \setminus R'_2 = \{b \in R \,|\, R/bR$ admits a system of representatives made of 0 and units $\}$.

It follows that if $R_2 \setminus R_1 = \emptyset$, then R is not Euclidean unless it is a field, and this condition gives an easy criterion to understand if a domain fails to be Euclidean.

4.2 The conjecture

We already said that every Euclidean domain is a PID. However there exist principal ideal domains that are not Euclidean.

The classical examples of non-Euclidean PID's are the ring of integers in $\mathbb{Q}\sqrt{-d}$, with $d = 19, 43, 67, 163$ and a Bass'technical example constructed with algebraic K-theory tools (cf. [6], [2]).
These examples were exhibited by Cohn and Bass in order to show that not every PID satisfies property (GE_2), so, in view of O'Meara's result, they don't even admit a weak algorithm. Somehow surprisingly, an example of non-Euclidean PID satisfying (GE_2) was not found up to now.
This fact suggested the following conjecture

(C) If a principal ideal domain R is not Euclidean, then R does not satisfy (GE_n), for some $n > 0$.

By the results of O'Meara [13] and Ruitenburg [14] it follows that **(C)** is equivalent to the following

(C_1) If a principal ideal domain R is not Euclidean, then R does not satisfy (GE_2).

(C_2) If a principal ideal domain R is not Euclidean, then R does not satisfy (ID_2).

(C_3) If a principal ideal domain R satisfies (ID_n) for every $n > 0$, then R is Euclidean.

(C_4) If a principal ideal domain R admits a weak algorithm, then R is Euclidean.

In the last part of this note we summarize some new results that support the validity of the conjecture. For more details we refer to [7].
We proved that the conjecture is valid when R is either an Anderson's PID or the coordinate ring of a special algebraic curve.

4.3 Anderson's PIDs

In this section D will denote an assigned UFD (unique factorization domain) and f any prime element of D. We set $D_f := D[1/f]$.
The following theorem, due to Anderson, is a slightly less general statement of the theorem on page 1222 of [1], and it provides an easy way to construct principal ideal domains that are not Euclidean.

Theorem 4.1 (Anderson - 1988) *Let D be a two-dimensional UFD, and let f be a prime element contained in the Jacobson radical $J(D)$. Then $D_f = D[1/f]$ is always a PID, and it is Euclidean if and only if D/Df is Euclidean. Moreover, if D_f is Euclidean, then D is regular and $f \notin \mathfrak{M}^2$ for any maximal ideal $\mathfrak{M}$ of D. Otherwise, D_f is a non-Euclidean PID.*

Definition 4.2 A non-Euclidean PID of the form $D_f = D[1/f]$ as in the theorem above, is called *Anderson's PID*.

It is important to remark that, in the above theorem, the case when D_f is non-Euclidean was solved using Samuel's characterization (see Section 4.1), which does not exclude the possible existence of weak algorithms.

Example 4.3 Let K be a field, X, Y two indeterminates, and let $D = K[\![X, Y]\!]$. The domain D is a two-dimensional regular local ring with maximal ideal $\mathfrak{M} = \langle X, Y \rangle$. Take any principal prime element f of D that lies in $\mathfrak{M}^2$ (for instance, we can take $f = X^2 + Y^3$) and define $D_f = D[1/f]$. Then D_f is an Anderson's PID.

We proved the following

Theorem 4.4 *Any Anderson's PID does not satisfy property (ID_2).*

Therefore we proved the validity of (C_2), equivalent to **(C)**, for the class of Anderson's PID's.

4.4 Special coordinate rings

Dealing with coordinate rings of algebraic curves that are non-Euclidean PID's, we will distinguish two cases: the case of elliptic curves having the point at infinity as unique rational point, and the case of conics without rational points.

Let k be a perfect field, $\bar{k}$ its algebraic closure, and $f \in k[x]$ a cubic polynomial without multiple factors.

The equation $y^2 = f(x)$ defines an affine smooth curve E_0 over k. Its projective completion in $\mathbb{P}_2$ has a unique (smooth) point at infinity $\mathcal{O} = (0, 1, 0)$ and thus, it defines an elliptic curve E.
We consider the coordinate ring (over k) of our affine curve E_0,

$$R := \frac{k[x, y]}{(y^2 - f(x))} = k[x, y] = k[E_0].$$

Then we have the following theorem, whose proof is due to U. Zannier:

Theorem 4.5 *The ring R is a PID if and only if $\mathcal{O} = (0, 1, 0)$ is the unique k-rational point of E. The ring is never Euclidean.*

We remark that, also in this theorem, the last statement was proved using Samuel's characterization for Euclidean domains.

Example 4.6 There are several known examples of elliptic curves over $\mathbb{Q}$ having a unique rational point. For instance, in the book by Cassels [5], Lemma 2, page 86, one may find the example of the affine curve E_0 with equation

$$y^2 = x^3 - 2^8 3^5 5^2.$$

The proof that E has no rational points other than $O = (0, 1, 0)$ is far from being easy.

Example 4.7 An easy example of elliptic curve defined over the field of rational functions $\mathbb{C}(t)$, having a unique rational point is

$$\mathcal{C} \equiv y^2 = x^3 - t.$$

An easy computation shows that the point at infinity $O = (0, 1, 0)$ is the unique rational point of $\mathcal{C}$.

Our aim was to prove that the non-Euclidean PID $R := k[x, y] = k[E_0]$ as in Theorem 4.5 does not satisfy property (ID_2), thus verifying the conjecture **(C)**.
Actually, we proved a more general result:

Theorem 4.8 *If $\mathcal{C} \subseteq \mathbb{P}^n$ is a smooth curve of genus ≥ 1, with a unique point at infinity, whenever the coordinate ring $R = k[\mathcal{C}_0]$ of the affine curve $\mathcal{C}_0 = \mathcal{C} \cap \mathbb{A}^n$ is a PID, then R does not satisfy property (ID_2).*

Therefore we get as a corollary that

Corollary 4.9 *In the above notation, if $R = k[\mathcal{C}_0]$ is a PID, then it doesn't satisfy property (GE_2), it doesn't admit a weak algorithm and, in particular, it is never Euclidean.*

The above corollary shows that, in the case of the affine elliptic curve E_0 without rational points, the PID $k[E_0]$ cannot be Euclidean. So the last part of the proof of Theorem 4.5 should not be necessary, *a priori*. Moreover, the coordinate rings $R = k[\mathcal{C}_0]$ as in Theorem 4.8 are non-Euclidean PID's verifying the conjecture.

Remark 1 In Brown's paper [4], statement of Theorem 1.1, one finds a list of four principal ideal domains that are not Euclidean, namely

$$R_1 = \mathbb{F}_2[X,Y]/(Y^2+Y+X^3+X+1)\ ;\ R_2 = \mathbb{F}_3[X,Y]/(Y^2-X^3+X+1);$$

$$R_3 = \mathbb{F}_4[X,Y]/(Y^2+Y+X^3+\eta)\ ;\ R_4 = \mathbb{F}_2[X,Y]/(Y^2+Y+X^5+X^3+1),$$

where η is a generator of $\mathbb{F}_4^*$. Since R_1–R_3 are the rings of smooth curves with genus 1 and a unique point at infinity, they do not satisfy property (ID_2), by Theorem 4.8. On the other hand, the curve $\mathcal{C}_4$ of equation $y^2+y+x^5+x^3+1=0$ on $\mathbb{F}_2$ has a singular point at infinity, so we cannot directly apply the above arguments to the coordinate ring R_4. However, a slight generalization of the techniques used to prove Theorem 4.8, allowed us to show that also R_4 does not satisfy property (ID_2). It follows that R_i satisfies (C_2), and then **(C)**, for $i = 1,\dots,4$.

The bodies of the proofs of Theorem 4.4 and of Theorem 4.8 are basically the same: in both cases we assume by contradiction that (ID_2) holds and fix a minimal length factorization into idempotents for the matrix $\begin{pmatrix}\eta & \xi\\ 0 & 0\end{pmatrix}$, where η and ξ satisfy some *technical properties*. It follows that

$$\begin{pmatrix}\eta & \xi\\ 0 & 0\end{pmatrix} = \begin{pmatrix}\eta' & \xi'\\ 0 & 0\end{pmatrix}\begin{pmatrix}a & b\\ c & 1-a\end{pmatrix},$$

and we reach the absurd by proving that η' and ξ' satisfy the same properties of η and ξ, contradicting the minimality of the length of the factorization.
However, it is important to remark the two theorems are based on very different lemmas, concerning the *technical properties* of η and ξ mentioned above; in particular, the case of an Anderson's PID is harder to establish than the case of the coordinate rings.

We conclude with the case of the coordinate rings of conics without rational points.
From Samuel's Proposition 19 in [16], we get the following Corollary:

Corollary 4.10 *Let $\mathcal{C} \subseteq \mathbb{P}^2$ be a plane smooth curve of genus zero over k, $\mathcal{C}_0 = \mathcal{C} \cap \mathbb{A}^2$ the affine part of $\mathcal{C}$, and $R = k[\mathcal{C}_0]$ the affine coordinate ring of $\mathcal{C}_0$. Then R is a non-Euclidean PID if and only if $\mathcal{C}$ has no rational points.*

Example 4.11 The coordinate ring of the conic over $\mathbb{R}$ with equation $x^2+y^2+1=0$ is a non-Euclidean PID. The coordinate ring of the conic over $\mathbb{Q}$ with equation $x^2-3y^2+1=0$ is a non-Euclidean PID over $\mathbb{Q}$ but not over $\mathbb{R}$.

Using a Cohn's result on what he calls k-rings with a *degree-function* (cf. Proposition 7.3 of [6]), we proved the following theorem.

Theorem 4.12 *Let $\mathcal{C} \subseteq \mathbb{P}^2$ be a plane smooth curve of genus zero over k, such that its coordinate ring $R = k[\mathcal{C}_0]$ is a non-Euclidean PID. Then R does not satisfy property (GE_2).*

Therefore, the conjecture **(C)** is verified also for this family of non-Euclidean PID's.

We point out that the classical examples of non-Euclidean PID's in number fields, Bass'e-xample and the two classes of PID's mentioned above, namely Anderson's PID's and the particular coordinate rings, seem to be the only examples of non-Euclidean principal ideal domains that may be found in the literature. Thus, with our results we proved that the conjecture **(C)** is verified for all non-Euclidean PID's appeared in the literature up to now.

References

[1] D.D. Anderson, *An existence theorem for non-Euclidean PID's.* Comm. Algebra 16/6 (1988), 1221–1229.

[2] H. Bass, "Introduction to some methods of algebraic K-theory". Expository lectures from the CBMS Regional Conference held at Colorado State University (1973). AMS 20, 1974.

[3] K.P.S. Bhaskara Rao, *Products of idempotent matrices over integral domains.* Linear Algebra Appl. 430 (2009), 2690–2695.

[4] M.L. Brown, *Euclidean rings of affine curves.* Math. Z. 208 (1991), 467–488.

[5] J.W.S. Cassels, "Lectures on elliptic curves". London Mathematical Society Student Texts, 24. Cambridge University Press, Cambridge, 1991.

[6] P.M. Cohn, *On the structure of the GL_2 of a ring.* Inst. Hautes Études Sci. Publ. Math. 30 (1966), 5–53.

[7] L. Cossu, P. Zanardo, U. Zannier, *Products of elementary matrices and non-Euclidean principal ideal domains.* Submitted (2016).

[8] J.A. Erdos, *On products of idempotent matrices.* Glasgow Math. J. 8 (1967), 118–122.

[9] J. Fountain, *Products of idempotents integer matrices.* Math. Camb. Phil. Soc. 110 (1991), 431–441.

[10] I. Kaplansky, "Commutative Rings". Revised edition. The University of Chicago Press, Chicago, Ill.-London, 1974.

[11] I. Kaplansky, *Elementary divisors and modules.* Trans. Amer. Math. Soc. 66 (1949), 464–491.

[12] T.J. Laffey, *Products of idempotent matrices.* Linear and Multilinear Algebra 14 (4) (1983) 309–314.

[13] O.T. O'Meara, *On the finite generation of linear groups over Hasse domains.* J. Reine Angew. Math. 217 (1965), 79-128.

[14] W. Ruitenburg, *Products of idempotents matrices over Hermite domains.* Semigroup Forum, 46 (1993), no. 3, 371–378.

[15] L. Salce and P. Zanardo, *Products of elementary and idempotent matrices over integral domains.* Linear Algebra Appl. 452 (2014), 130–152.

[16] P. Samuel, *About Euclidean rings.* J. Algebra 19 (1971), 282–301.

Biologically inspired deduction of Optimal Transport Problems

Enrico Facca (*)

In this report we report a mathematical model that we introduced in [7], which is a generalization of a model born to describe the behavior of slime mold dynamics called *Physarum Polycephalum* (PP). Recent experimental evidence [10] shows that PP grows following the most efficient path between food sources. This evidence is exemplified in the picture shown in Figure 1 that shows the experimental setup developed by [10] suggesting that PP, after colonizing the entire maze paths, concentrates growing along the network shortest path connecting the two food sources.

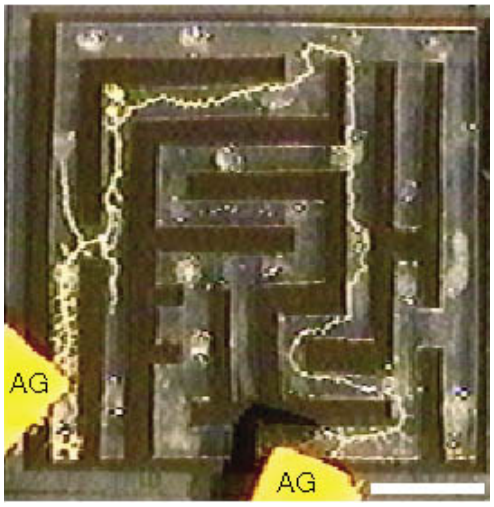


Figure 1. *Physarum Polycephalum* solving a maze (photo from [10])

This behavior has been used for the experimental analysis of transportation networks, with many researchers suggesting that this slime mold is capable of identifying the optimal many-site connecting transportation network, such as railroad systems of Tokyo and Spain [13, 1]. Many further surprising properties of Physarum Polycephalum have been experimentally identified, but in this work we are interesting in studying and possibly extending mathematical models that mimic the slime mold behavior.

Physarum Polycephalum is a slime mold that grows on humid and cool environments. They are typically multi-nuclei unicellular organisms containing "a network of tubes by means of which nutrients and chemical signals circulate throughout the organism" [12].

It is perhaps this peculiar tubular structure that allows this slime mold to explore the entire space and identify the optimal paths. Little is known about the biological mecha-

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: facca@math.unipd.it. Seminar held on November 30th, 2016.

nisms governing these processes. On the other hand, the observation that optimal paths are singled out suggests that there must exists a global functional that drives the mold in its exploration. Recently, [12] have proposed a general mathematical model for the description of PP behavior. By numerical experimentations the authors have shown that, depending on the chosen model configuration, the simulated mold density correctly identifies optimal (shortest) paths in a network graph. The proposed model was subsequently applied successfully in a number of network optimization problems, such as the railway transportation problem mentioned above [13, 1].

The model proposed in [12] considers a connected planar graph on which an equation describing the behavior of the mold density is defined. This equation is a typical conservation law coupled to a nonlinear dynamic equation for the flow conductivity defined on each graph edge. The mathematical model reads as follows. Consider a simple graph $G = (V, E)$, with V being the set of n vertices and E the set of m edges. The flux on each edge $e = (u, v) \in E$ connecting nodes $u \in V$ and $v \in V$ is denoted by Q_e. Given a function p defined on the nodes of G the flux is given by:

$$Q_e = \frac{D_e}{L_e}(p_u - p_v), \tag{1}$$

where D_e is the conductivity (inverse of the resistance to flow) and L_e is the length of edge e. Note that this equation can be interpreted from the physical point of view as the Poiseuille equation governing laminar flow of a fluid in a pipe if we let $D_e = \pi r_e^4/8\rho$, where r_e is the radius of the pipe and ρ is the dynamic viscosity of the fluid. For all nodes except the source v_0 and the sink v_1 nodes, mass conservation is imposed (i.e., the total flux leaving the node must equal the total flux entering the node), while on the source and sink nodes unit inflow (+1) and outflow (-1) is assumed. In summary, the following equations are imposed on each vertex $v \in V$:

$$\sum_{e \in \sigma(v)} Q_e = f(v) := \begin{cases} 0 & v \neq v_0, v_1, \\ 1 & v = v_0, \\ -1 & v = v_1, \end{cases} \tag{2}$$

where $\sigma(v)$ is the set of edges having v as one of the vertices. Finally, the conductivity D_e of each edge e varies in time according to:

$$\frac{d}{dt} D_e(t) = g(|Q_e(t)|) - D_e(t) \tag{3}$$

where g is an increasing function from $\mathbb{R}^+$ into itself with $g(0) = 0$. In [12] numerical solutions of the above model under different configurations have shown that, for particular forms of the function g, the conductivity degenerates (tends to zero) in every edge except the edges forming the shortest path along connecting the graph vertices v_0 and v_1, and, obviously, the mass density of the slime mold p accumulates on these paths.

More recently, [3] have proved that this model naturally identifies the shortest paths of a simple undirected planar graph, in the sense that the dynamics of D_e reaches an equilibrium state characterized by conductivities achieving nonzero values only on the edges of the shortest paths of G connecting the source and sink vertices. The authors

showed that a the same model can be applied to solve a generalization of the problem that considers general forcing terms $f(v)$ satisfying $\sum_{v\in V} f(v) = 0$. They show that the system $(D(t), Q(t) = (D_e(t), Q_e(t))_{e\in E(G)}$ converges to a steady state configuration $(D^*, Q^*) = (D^*_e, Q^*_e)_{e\in E(G)}$ with Q^* solving the following transportation problem on a graph: find Q that solves

$$\begin{aligned} \min \qquad & \sum_e Q_e L_e \\ \sum_{e\in\delta(v)} Q_e = f_v \qquad & \forall v \in V \end{aligned}$$

In [12] we extend the model considering a general continuum (dense) domain $\Omega \subset \mathbb{R}^n$. A straight forward generalization of the above model to such a continuous domain yields the following nonlinear system of equations, with $(\mu, u) : ([0, +\infty) \times \Omega) \mapsto (\mathbb{R}^+, \mathbb{R})$ solving

$$-\nabla\cdot(\mu(t,x)\nabla u(t,x)) = f(x) = f^+(x) - f^-(x) \tag{4a}$$

$$\mu'(t,x) = \mu(t,x)(|\nabla u(t,x)| - 1) \tag{4b}$$

$$\mu(0,x) = \mu_0(x) \tag{4c}$$

completed with zero Neumann boundary condition and having f forcing term with zero mean. Hence the gradient and the divergence are computed with respect to the spatial variable x, while $'$ indicates the time-derivative. We can try to justify the heuristics that is behind the idea that the above model is a natural extension of the model proposed by [12]. Assuming the existence of a solution of system (4), we will try to determine if the pair $(\mu(t,x), u(t,x))$ has asymptotic convergence toward an "equilibrium" configuration, i.e.

$$\begin{aligned} (\mu'(t,x), u'(t,x)) &\overset{t\to+\infty}{\longrightarrow} (0,0) \\ (\mu(t,x), u(t,x)) &\overset{t\to+\infty}{\longrightarrow} (\mu^*, u^*) \end{aligned} \tag{5}$$

This assumptions allows us to look at the system of equations "solved" by the pair (μ^*, u^*) and given by:

$$\begin{cases} -\nabla\cdot(\mu^*(x)\nabla u^*(x)) = f(x) \\ 0 = \mu^*(|\nabla u^*(x)| - 1) \end{cases} \tag{6}$$

This system can be rewritten as:

$$\begin{cases} -\nabla\cdot(\mu^*(x)\ \nabla u^*(x)) = f(x), \\ |\nabla u^*(x)| = 1 \text{ where } \mu^*(x) > 0, \\ \text{ no assumption on } (|\nabla u^*(x)| - 1) \text{ where } \mu^*(x) = 0. \end{cases} \tag{7}$$

These equations resemble the so called the Monge-Kantorovich equations [4]:

$$\begin{cases} -\nabla\cdot(a(x)\nabla u^*(x)) = f(x) = f^+(x) - f^-(x) \\ |\nabla u(x)| \le 1 \quad \text{on } \Omega \\ |\nabla u(x)| = 1 \quad \text{on } \{a(x) > 0\} \end{cases} \tag{8}$$

These equation, introduced in [6], are a PDE formulation of the *Optimal Transport Problem* (OTP), first proposed by Gaspard Monge in 1791 and reformulated in 1942 by Leonid Kantorovich, which studies the optimal way to move f^+ into f^-. We refer the reader to [14] for a complete treatment of this problem and its variants. The similarity between equations (7) and (8) gave as the heuristic justification to suppose that, if the system (4) tends to an equilibrium state, then *the pair* (μ^*, u^*) *is the solution* (a, u^*) *of the Monge-Kantorovich equation.* In order to study the long time behavior of system (4), we first need to prove existence and uniqueness in time. We obtained local existence under the assumption that $\mu_0 \in \mathcal{D} := \left\{\mu > 0 \in \mathcal{C}^\delta\Omega\right\}$ and $f \in L^\infty(\Omega)$.

Under such hypothesis we can introduced two operators $\mathcal{U}$ defined as

$$\mathcal{U} : \mathcal{D} \mapsto \mathcal{C}^{1,\delta}(\Omega)$$
$$\bar{\mu} \mapsto \mathcal{U}(\bar{\mu}) := \bar{u} \text{ solution of}$$
$$\int_\Omega \bar{\mu}\nabla\bar{u}\nabla\varphi\, dx = \int_\Omega f\varphi\, dx \qquad \forall\varphi \in H^1(\Omega)$$

With the above definitions we can recast the problem (4) in ODE-form:

$$\left\{ \begin{array}{c} \mu'(t) = \mu(t)|\mathcal{U}(\mu(t))| - \mu(t) \\ \mu(0) = \mu_0 \in \mathcal{D} \end{array} \right. \tag{9}$$

In [7] we use well known tools of the regularity theory of elliptic equation (see [9]), properly adapted to our problem, to prove a local Lipschitz continuity of the right hand side of ODE (9), which provides sufficient condition for existence and uniqueness, at least in a small time interval $[0, \tau(\mu_0)[$ which depends on the initial data μ_0. Besides this partial result we prove in [8] that the functional

$$\mathcal{S}(\mu) := \frac{1}{2}\int_\Omega \mu(x)|\nabla\mathcal{U}(\mu(x))|^2\, dx + \frac{1}{2}\int_\Omega \mu(x)\, dx \tag{10}$$

is strictly decreasing in time. In fact, its Lie derivative along the $\mu(t)$-trajectory is

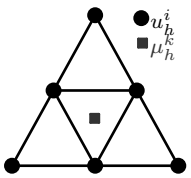
$$\frac{d}{dt}\mathcal{S}(\mu(t)) = -\frac{1}{2}\int_\Omega \mu(t,x)\left(|\nabla\mathcal{U}(\mu(t,x))| - 1\right)^2\left(|\nabla\mathcal{U}(\mu(t,x))| + 1\right) dx \tag{11}$$

which is strictly negative $\forall t \geq 0$ and is equal to zero only if $|\nabla\mathcal{U}(\mu)| = 1$ within the support of μ. This is one of the constraint of the MK-equations , while the uniform bound on the whole domain can not be deduced yet.

In order to obtain more evidence supporting the convergence of system (4) toward the solution of the MK-equations, we have developed a numerical algorithm based on the separation of the time and spatial variable. Starting from a triangulation $\mathcal{T}_h(\Omega)$ (where h is the typical length of the triangulation) we approximate the function μ and u as follows

$$\mu_h(t,x) = \sum_{r=1}^{M} \mu_r(t)\chi_r(x) \quad \chi_r(x) = \left\{ \begin{array}{l} 1 \text{ if } x \in T_r \\ 0 \text{ if } x \notin T_r \end{array} \right. \qquad T_r \in \mathcal{T}_h(\Omega)$$

$$u_h(t,x) = \sum_{i=1}^{N} u_i(t)\varphi_i(x) \quad \varphi_i \in V_h = P1(\mathcal{T}_{h/2})$$



where $\mathcal{T}_{h/2}$ indicates the mesh generated uniformly refining the triangulation $\mathcal{T}_h$. On the figure on the right we show the relation between the mesh used for the discretization of μ and u on one triangle. The projection into the finite-dimensional system leads to the following set of equations:

$$\begin{cases} \int_\Omega \mu_h \nabla u_h \cdot \nabla \varphi_i = (f, \varphi_i) = \int_\Omega f\varphi_i \qquad \forall \varphi_i \in V_h \\ \int_\Omega \mu_h' \chi_r = \int_\Omega \mu_h(|\nabla u_h| - 1)\chi_r \qquad \forall \chi_r \in W_h \\ \mu_h(0, x) = \mu_h^0(x) > 0 \ \in W_h \end{cases}$$

which represents a Differential Algebraic Equation (DAE) that can be rewritten in a more compact form:

$$\begin{cases} \boldsymbol{A}(\boldsymbol{\mu}(t))\ \boldsymbol{u}(t) = \boldsymbol{b} \\ \boldsymbol{\mu}'(t) = \boldsymbol{B}(\boldsymbol{u}(t))\ \boldsymbol{\mu}(t) \\ \boldsymbol{\mu}(0) = \boldsymbol{\mu}^0 > 0 \end{cases} \tag{12}$$

where $\boldsymbol{\mu}(t), \boldsymbol{u}(t)$ represents the vector describing the time evolution of the projected system, $\boldsymbol{A}$ is the stiffness matrix, $\boldsymbol{b}$ is right hand side of the linear system defined by the forcing term f and $\boldsymbol{B}$ is a diagonal matrix.

In order to solve the DAE (12) we operate a discretization in time using either a forward or a backward Euler scheme. Denoting with Δt_k the time-step size so that $t_{k+1} = t_k + \Delta t_k$ we obtain an approximating sequence $(\boldsymbol{u}^k, \boldsymbol{\mu}^k) = (\boldsymbol{u}(t_k), \boldsymbol{\mu}(t_k))$. Using the explicit Euler Scheme the approximation sequence is described by the following equation

$$\text{EE} \begin{cases} \boldsymbol{A}[\boldsymbol{\mu}^k] \cdot \boldsymbol{u}^k = \boldsymbol{b} \\ \boldsymbol{\mu}^{k+1} = (\boldsymbol{I} + \Delta t^k \boldsymbol{B}[\boldsymbol{u}^k]) \cdot \boldsymbol{\mu}^k \\ \boldsymbol{\mu}^0 = \boldsymbol{\mu}_0 \end{cases}$$

Using the Implicit Euler Scheme $(\boldsymbol{u}^k, \boldsymbol{\mu}^k)$ solves

$$\text{EI} \begin{cases} \boldsymbol{\mu}^{k+1} = \boldsymbol{\mu}^k + \Delta t_k (\boldsymbol{B}[\boldsymbol{u}^{k+1}] \cdot \boldsymbol{\mu}^{k+1}) \\ \boldsymbol{\mu^0} = \boldsymbol{\mu}_0 \end{cases}$$

where the non-linear equation is solved by Picard iterations method, which reads as:

$$\text{PIC} \begin{cases} \boldsymbol{A}[\boldsymbol{\mu}^{m,k}] \cdot \boldsymbol{u}^{m,k} = \boldsymbol{b} \\ \boldsymbol{\mu}^{m+1,k+1} = \left(\boldsymbol{I} - \Delta t^k \boldsymbol{B}[\boldsymbol{u}^{m,k}]\right)^{-1} \cdot \boldsymbol{\mu}^k \\ \boldsymbol{\mu}^{0,k} = \boldsymbol{\mu}^k \end{cases}$$

until $\frac{\|\mu_h^{m+1,k+1} - \mu_h^{m,k+1}\|_{L^2(\Omega)}}{\|\mu_h^{m,k+1}\|_{L^2(\Omega)}} < 10^{-11}$. The results obtained with the procedure described above show that (μ_h, u_h) converges towards an equilibrium state as time progresses, and the flux constraint is satisfied. Moreover equilibrium configurations does not depend on

the initial data μ_0. We show in Figure 2 the results obtained with our numerical implementation for μ_h^*, which compares well with published numerical solutions of [2]. Moreover we compare the solution obtained with our approach with exacts solution of the MK-equations (see [5]) obtaining an experimental convergence of when the mesh is align with the support of the optimal solution and when the mesh is not aligned. This suggests that our method can be used to solve efficiently the MK-equations.

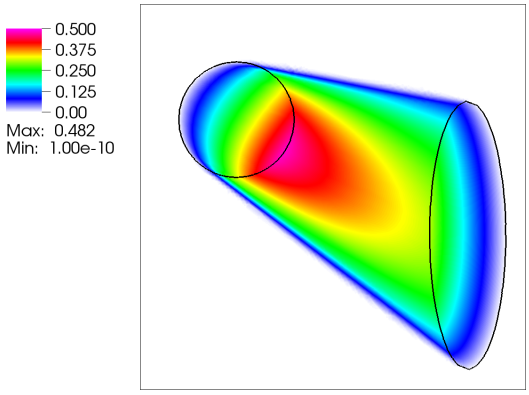


Figure 2. Spatial distribution of the numerical approximation of μ_h^* obtained using a triangulation $\mathcal{T}_{h/2}$ with 49421 nodes. The circle and the ellipse indicate respectively the supports of the positive and negative part of the forcing term f.

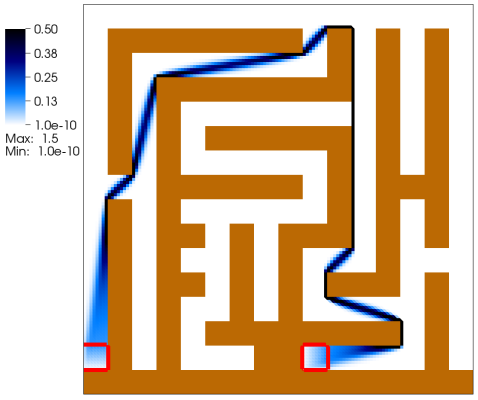


Figure 3. Equilibrium distribution of Physarum Polycephalum mass on the maze. The red squares indicates the support of f^+ and f^- representing the food sources. The brown areas indicates where $k(x) = 1000$, corresponding to the walls of the maze. The blue path is the equilibrium μ^* of equation (13).

To address the dynamics of PP in the maze, we need to reconcile the model with the fact that some portions of the domain (the maze barriers in this case) may hinder through-flow. This can be obtained by imposing the gradient to be large where the flux must be small, thus forcing the conductivity μ to become small. Thus, equation (4b) is replaced by:

$$\begin{aligned} -\nabla\cdot(\mu(t,x)\nabla u(t,x)) &= f(x) \\ \mu'(t,x) &= \mu(t,x)\Big(|\nabla u(t,x)| - k(x)\Big) \\ \mu(0,x) &= \mu_0(x) \end{aligned} \tag{13}$$

where $k(x)$ is a positive function describing the spatial pattern of the resistance to flow, whereby large values of k imply large energy losses and hence large gradients of the potential u. As shown in Figure 3 the equilibrium configuration of the solution of (13) with $k(x)$ and f adapted by the maze problem, shows that the PP mass concentrates along the the shortest path between the food sources.

A further modification of dynamic part of the model, inspired by the original model

on the graph, is to consider

$$\begin{aligned} -\nabla\cdot(\mu(t,x)\nabla u(t,x)) &= f(x) \\ \mu'(t,x) &= (\mu(t,x)|\nabla u(t,x)|)^{\beta} - \mu(t,x) \\ \mu(0,x) &= \mu_0(x) \end{aligned} \tag{14}$$

with $\beta > 1$. Unlike the case $\beta = 1$ there is no candidate equation describing the steady state, thus we use the numeric scheme described before to study the behavior of the system in (14). Also in this case the above system presents a steady state configuration, result confirmed by considering different forcing terms f and powers β. As shown in Figure 4 the equilibrium configuration μ^* obtained with different exponent β and with two forcing terms, one piecewise constant, and the other atomic. The path described by the support of μ^* resembles a network structure, which through the flux going from f^+ to f^- tends first to concentrate and then to split. This effect increases with the power β.

We note that, for $\beta > 1$, the asymptotic state μ^* of equation (14) depends on the initial data μ_0. We report in Figure 5 the comparison between two equilibrium μ^* starting from two different initial data. The results obtained using $\beta > 1$ in equation (14) strongly suggest a relation between our model and the, so called, *Branched Transport Problem* (BTP) (see [11] for a complete overview). This sub-area of OTP studies the problem of reallocating f^+ into f^- favoring mass concentration, which is a more realist description of real-life transport problem. One formulation of the BTP can be given as follows: consider $\boldsymbol{a} = \sum_{i=1}^{n} a_i \delta_{x_i}$, $\boldsymbol{b} = \sum_{j=1}^{m} b_j \delta_{y_j}$ (δ_p indicates the Dirac function located at $p \in \mathbb{R}^N$) and satisfying $\sum_{i=1}^{n} a_i = \sum_{j=1}^{m} b_j$. Find a graph $G = (V, E)$ and a weight function defined on the graph edges $q : E(G) \mapsto R^+$ such that

$$\sum_{e\in\delta(v)} q_e = \begin{cases} a_i & \text{if } v = x_i \text{ for some } i \\ -b_j & \text{if } v = y_j \text{ for some } j \\ 0 & \text{otherwise} \end{cases}$$

$$\min \mathcal{E}_\alpha(G, q) := \sum_{e\in E} q(e)^\alpha L_e \quad 0 \le \alpha \le 1$$

The spatial distributions of μ^* on the right panels of Figure 4 resemble the solutions of the BTP using $\boldsymbol{a} = f^+$ and $\boldsymbol{b} = f^-$ (f^+, f^- are forcing terms described in Figure 4), for different values of $0 < \alpha \le 1$, even if it is no clear yet the relation between the exponent β in our model and the exponent α of the BTP.

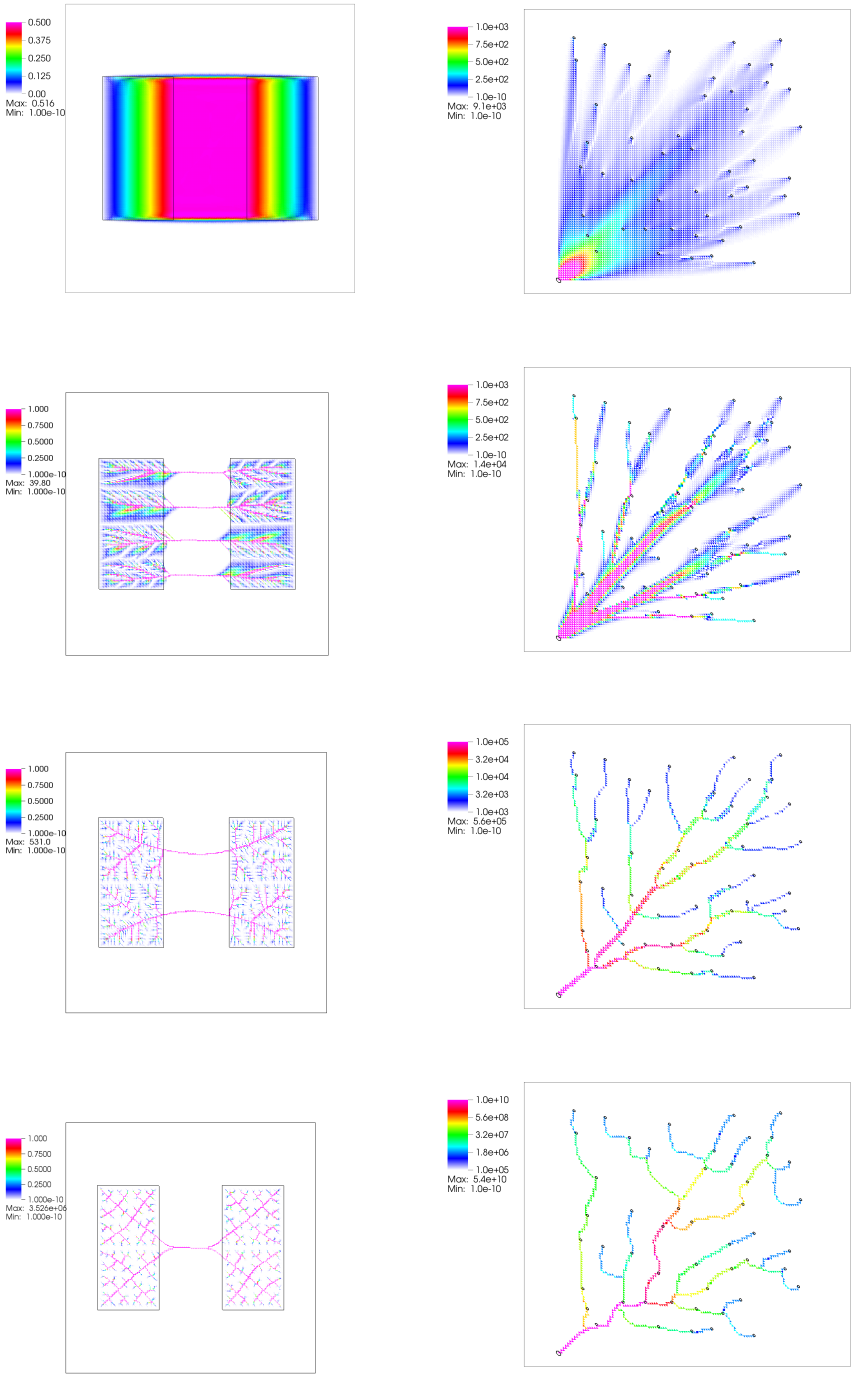


Figure 4. Spatial distribution of μ^* obtained with two different forcing terms. On the left column f is a piecewise constant forcing term f, where the black rectangles indices the supports of f^+ (left) and f^- (right). On the right column f^+ is the sum of 30 Dirac sources randomly distributed in the square $[0.1, 0.9] \times [0.1, 0.9]$, and f^- is the concentrated in the point $(0.05, 0.05)$. The powers β used in (14) are from top to bottom 1.0 (which corresponds to the MK equations), 1.05, 1.4 and 3.0.

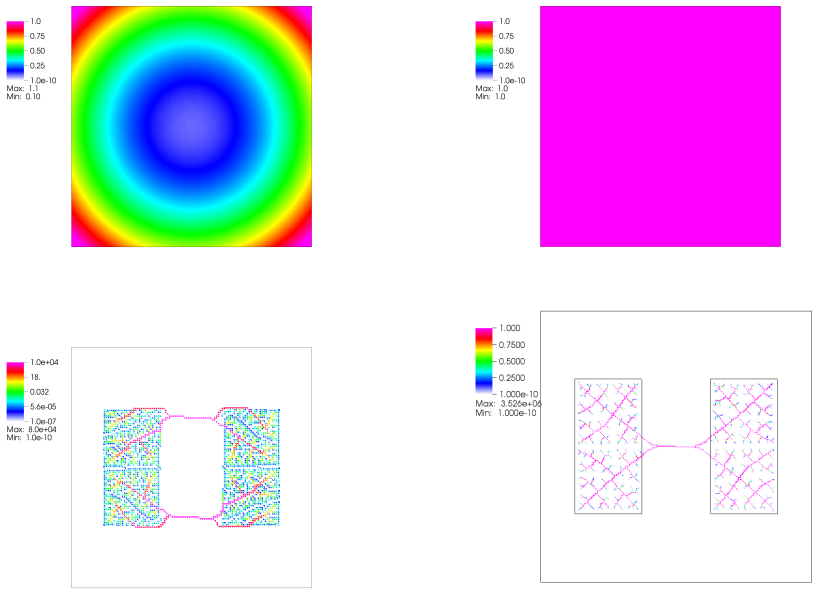


Figure 5. Spatial distribution of μ^* obtained with a piecewise constant forcing term f, where the black rectangles denote the supports of f^+ (left) and f^- (right). The upper panels show the initial data μ_0 and the lower panels the correspondence equilibrium μ^*.

References

[1] A. Adamatzky,, "Physarum Machines, Computers from Slime Mould". World Scientific, 2010.

[2] J.W. Barrett and L. Prigozhin, *A mixed formulation of the Monge-Kantorovich equations.* Math. Model. Num. Anal., 41 (2007), 1041–1060.

[3] V. Bonifaci, K. Mehlhorn, and G. Varma, *Physarum can compute shortest paths.* J. Theor. Biol., 309 (2012), 121–133.

[4] G. Bouchitté, G. Buttazzo, and P. Seppecher, *Shape optimization solutions via Monge-Kantorovich equation.* C. R. Acad. Sci. Paris Sr. I Math, 324 (1997), 1185–1191.

[5] G. Buttazzo and E. Stepanov, *On regularity of transport density in the Monge?Kantorovich problem.* SIAM J. Control Optim, 42 (2003), 1044–1055.

[6] L.C. Evans and W. Gangbo, *Differential equations methods for the Monge-Kantorovich mass transfer problem.* Mem. Am. Math. Soc., 137 (1999), 1–66.

[7] E. Facca, F. Cardin, and M. Putti, *Toward?s a stationary Monge- Kantorocich dynamics: the Physarum Polycephalum experience.* Journal of Applied Mathematics, submitted (2017).

[8] E. Facca, S. Daneri, F. Cardin, and M. Putti, *Numerical solution of Monge-Kantorovich equations via a dynamic formulation.* To appear (2017).

[9] M. Giaquinta and L. Martinazzi, "An Introduction to the Regular- ity Theory for Elliptic Systems, Harmonic Maps and Minimal Graphs". Springer Science and Business Media, Pisa, July 2013..

[10] T. Nakagaki, H. Yamada, and A. Toth, *Maze-solving by an amoeboid organism.* Nature, 407 (2000), 470–470.

[11] F. Santambrogio, "Optimal transport for applied mathematicians". 2015.

[12] A. Tero, R. Kobayashi, and T. Nakagaki, *A mathematical model for adaptive transport network in path finding by true slime mold.* J. Theor. Biol., 244 (2007), 553–564.

[13] A. Tero, S. Takagi, T. Saigusa, K. Ito, D.P. Bebber, M.D. Fricker, K. Yumiki, R. Kobayashi, and T. Nakagaki, *Rules for biologically inspired adaptive network design.* Science, 327 (2010), 439–442.

[14] C. Villani, "Optimal Transport". Vol. 338 of Old and New, Springer Science and Business Media, Berlin, Heidelberg, 2008.

Extension fields, and classes in the genus of a lattice

Frances Odumodu (*)

1 Introduction

To classify objects in a certain collection, one needs to determine when two objects in that collection are equivalent and give a complete set of invariants that determine an equivalence class. Thus, one can give a non-redundant enumeration of the objects by placing each object in exactly one class.

At the level of fields one has the Hasse-Minkowski local-global theorem which gives the classification of quadratic forms. This fails in general at the integral level, hence there are two levels of classification, the genus (local) and the integral class (global). The idea then is to study the classes in a given genus.

This talk which will be accessible to a large audience, will focus on some existing results concerning the classes in the genus of a lattice in a quadratic space, and in particular the trace form. We will start by recalling several facts on extension of fields which we will need for the talk.

Let E and F be fields such that $F \subseteq E$. Then, E is called an extension field of F and F a subfield of E. Example. The field of complex numbers $\mathbb{C}$ is an extension of the reals $\mathbb{R}$.

We may view E as a vector space over F; the elements of E are vectors and F is the field of scalars. Then the degree of the extension is $\dim_F E$. If this dimension is finite, then E/F is a finite extension.

Let $\alpha \in E$. The field $F(\alpha)$ is the subfield of E generated by F and α. It is the smallest (intersection) of all such fields. It is called the field obtained from F by adjoining α.

A polynomial $f(x) \in F[x]$ is irreducible if it cannot be expressed as a product $g(x)h(x)$ of polynomials both of degree less than $f(x)$. Let $p(x) \in F[x]$ be the unique polynomial such that

(*)Institut de Mathématiques de Bordeaux, Université de Bordeaux, 351 Cours de la Libération, 33400 Talence, France; E-mail: `francesodumodu@gmail.com` . Seminar held on December 14th, 2016.

- $p(x)$ is irreducible over F.
- $p(x)$ is monic. That is, the leading coefficient is 1.
- $p(\alpha) = 0$ for some α.

Then, $p(x)$ is called the minimum polynomial of α over F. Moreover, we have

$$F(\alpha) = F[x]/(p(x)).$$

where $(p(x))$ is the ideal generated by the irreducible polynomial $p(x)$. Also, $\deg F(\alpha) = \deg p(x) = \deg \alpha$ over F. Example. $\mathbb{C} = \mathbb{R}(i) = \mathbb{R}[x]/(x^2 + 1)$.

Let E/F be a finite extension of fields and $\alpha \in E$. Then, α is algebraic over F if there is a nonconstant polynomial $f(x) \in F[x]$ such that $f(\alpha) = 0$. Algebraic numbers form a subfield of $\mathbb{C}$. Every finite extension of fields is algebraic.

2 Algebraic Number Fields

A special finite extension of fields in which we are interested is the number field. An algebraic number field K is a finite extension field of the field of rationals $\mathbb{Q}$. It is a subfield of $\mathbb{C}$. Examples include:

(a) The smallest and most basic example is $\mathbb{Q}$ itself.

(b) The quadratic number field. Let d be a square free integer, different from 1. Then $K = \mathbb{Q}(\sqrt{d}) = \{a + b\sqrt{d} : a, b \in \mathbb{Q}\}$ is a number field of degree 2 over $\mathbb{Q}$.

(c) The real numbers $\mathbb{R}$ and the complex numbers $\mathbb{C}$ are not algebraic number fields since they have infinite dimension over $\mathbb{Q}$.

Let K be a number field. By the primitive element theorem, there is an element $\alpha \in K$ such that $K = \mathbb{Q}(\alpha)$. Let $f(x) \in \mathbb{Q}(x)$ be the minimum polynomial of α over $\mathbb{Q}$. The roots of f in $\mathbb{C}$ are $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_n$ and are conjugates of α. They give rise to n $\mathbb{Q}$-linear embeddings of K in $\mathbb{C}$:

$$\tau_i : \alpha \mapsto \alpha_i.$$

The images $K_i = \mathbb{Q}(\alpha_i)$ in $\mathbb{C}$ of K under the above map are called the conjugate fields of K. They are isomorphic to K.

The signature of a number field K is given by (r_1, r_2). Here, r_1 is the number of embeddings of K whose image lie in $\mathbb{R}$ and $2r_2$ is the number of non-real complex embeddings such that $r_1 + 2r_2 = n$. These come in pairs. Also, r_1(resp. $2r_2$) is the number of real (resp. complex) roots of $f(x)$ in $\mathbb{C}$. Now, if $r_2 = 0$, then K is said to be totally real, if $r_1 = 0$ then K is totally complex, otherwise it is nontotally real, that is, $1 \leq r_1, r_2 < n$.

Let $K = \mathbb{Q}(\sqrt[3]{2})$ be the number field obtained by adjoining the real cube root of 2 to $\mathbb{Q}$. The minimum polynomial of $\sqrt[3]{2}$ over $\mathbb{Q}$ is $x^3 - 2$. It has roots $\sqrt[3]{2}, \omega\sqrt[3]{2}, \omega^2\sqrt[3]{2}$ in $\mathbb{C}$ with $\omega = \frac{-1+i\sqrt{3}}{2}$, a cube root of unity. The embeddings of K into $\mathbb{C}$ are given by the mappings

$$\sqrt[3]{2} \mapsto \sqrt[3]{2}; \qquad \sqrt[3]{2} \mapsto \omega\sqrt[3]{2}; \qquad \sqrt[3]{2} \mapsto \omega\sqrt[3]{2}$$

Thus the signature of K is $(1,1)$.

The trace map. Let $x \in K$. We have the $\mathbb{Q}$-linear transformation given by multiplication by x:

$$m_x : K \to K; \quad y \mapsto xy.$$

By a choice of basis, we have a matrix $A(x) = (a_{ij}(x))$. The trace of x relative to $K/\mathbb{Q}$ is then given by $T_K(x) = \text{trace } m_x = \text{trace } A(x)$.

Example Quadratic field. Let $K = \mathbb{Q}(\sqrt{d})$. A $\mathbb{Q}$-basis of K is $\{1, \sqrt{d}\}$. If $x = a + b\sqrt{d} \in K$, then letting x act on the basis elements we have

$$x \cdot 1 = a + b\sqrt{d}$$
$$x \cdot \sqrt{d} = bd + a\sqrt{d}$$

Thus $A(x) = \begin{pmatrix} a & bd \\ b & a \end{pmatrix}$. Hence $T_K(x) = \text{trace } A(x) = 2a$.

The trace map on K defines a nondegenerate symmetric bilinear form on K.

$$t_K : K \times K \to \mathbb{Q}; \quad (x, y) \mapsto T_K(xy).$$

This is called the trace form on K. Thus, (K, t_K) is a quadratic space. The ring of integers $\mathcal{O}_K$ of K is a $\mathbb{Z}$-lattice in this quadratic space. That is, a finitely generated free abelian group.

Besides the degree, the most important invariant of a number field K is its discriminant. It is defined as follows. Let $\{\omega_1, \ldots, \omega_n\}$ be a $\mathbb{Q}$-basis of K. Then, the discriminant is given by

$$d_K = \det(t_K(\omega_i\omega_j)_{i,j}) \in \mathbb{Q}^\times/(\mathbb{Q}^\times)^2.$$

Example Let $K = \mathbb{Q}(\sqrt{d})$ a quadratic number field. Then its discriminant is given by

$$d_K = \begin{cases} d & \text{if } d \equiv 1 \bmod 4 \\ 4d & \text{if } d \equiv 2, 3 \bmod 4 \end{cases}$$

If $d \equiv 2, 3 \bmod 4$, then K has $\mathbb{Z}$-basis $\{1, \sqrt{d}\}$. Since, Trace of $a + b\sqrt{d}$ is $2a$, the discriminant of K is

$$d_K = \det\left(\begin{pmatrix} T(1) & T(\sqrt{d}) \\ T(\sqrt{d}) & T(d) \end{pmatrix}\right) = \det\left(\begin{pmatrix} 2 & 0 \\ 0 & 2d \end{pmatrix}\right) = 4d.$$

3 Classification

From the definition of the discriminant, we have for a quadratic number field a one-one correspondence

$$d \text{ square free } \leftrightarrow \mathbb{Q}(\sqrt{d}).$$

Thus, quadratic number fields are characterised by their discriminant. This is not true in general for number fields of higher degree. In fact, we have the following result of Hermite.

Theorem 3.1 (Hermite) *There are only finitely many (up to isomorphism) number fields of fixed (bounded) degree.*

Is there a more refined invariant than the discriminant? Let's look at the form that defines the discriminant - the trace form. For arithmetic purposes we consider the integral trace form. This is the trace form restricted to the ring of integers $\mathcal{O}_K$ of K. View $\mathcal{O}_K$ as a $\mathbb{Z}$-lattice in the quadratic space (K, T_K).

What can one say? Are number fields classified by their forms?

Now, Given two quadratic forms, one can classify them up to some equivalence relation. The isometries are isomorphisms of the underlying vector space which preserve the form. These form a group, the orthogonal group of the vector space.

We first classify the underlying space (K, T_K). According to the Hasse-Minkowski local-global theorem, a complete set of invariants is

(a) Its rank which is $\dim_{\mathbb{Q}} K$.

(b) Its discriminant which is the discriminant of K up to squares.

(c) Its signature which is the signature of the number field.

(d) The Hasse-Witt invariant which assigns ± 1 to the form at each prime p of K.

Before proceeding to the classification at the integral level, we recall a few things we need from p-adic valuation. Let $p \in \mathbb{Z}$ be a prime number and $x \in \mathbb{Z}$, then

$$x = p^k a \text{ with } (a, p) = 1 \text{ and } k \in \mathbb{Z}.$$

Define

$$v_p(x) = k \quad \text{and} \quad |x|_p = p^{-k}$$

Example

$$|3|_3 = \frac{1}{3}; \quad |1097|_3 = 1; \quad |54|_3 = \frac{1}{27}$$

If $x \in \mathbb{Q}$, then $x = \frac{a}{b}$ with $a, b \in \mathbb{Z}$ and $b \neq 0$ and $v_p(x) = v_p(a) - v_p(b)$. The field of p-adic numbers is the field obtained on completing the rationals $\mathbb{Q}$ with respect to the p-adic absolute value. Its ring of integers is called the p-adic integers, denoted by $\mathbb{Z}_p$.

Now at the integral level, the Hasse-Minkowski theorem fails in general and hence we have two levels of classification:

- Global case: The **class** is isometry, that is equivalence over $\mathbb{Z}$. The isometries are given by matrices with entries in $\mathbb{Z}$.
- Local case: The **genus** is local isometry at all p, that is, equivalence over $\mathbb{Z}_p$ for all p prime. The local isometries at each p are given by matrices with entries in $\mathbb{Z}_p$.

The obstruction is given by the number of classes in the genus. This is finite.

4 Classes in a genus

Conjugate number fields are isomorphic and have isometric integral trace forms. The question to ask is:

Does isometry of integral trace forms imply conjugation? NO!

Example 4.1 [1] There exists two nonconjugated number fields of degree 7 with isometric integral trace forms.

In fact, the number fields have defining polynomials

$$p_1 = x^7 - 3x^6 + 4x^4 + x^3 - 4x^2 - x + 1$$
$$p_2 = x^7 - 3x^6 + 2x^5 + 4x^4 - 3x^3 - 2x^2 - x - 1$$

In example (4.1), the two fields considered are arithmetically equivalent. That is, they have the same Dedekind zeta functions. Now, arithmetically equivalent fields have the same

- rank
- discriminant
- signature
- genus
- ...

The question becomes

$$\text{Arithmetical equivalence} \overset{?}{\Leftrightarrow} \text{Isometric trace forms}$$

To answer this we study classes in a genus. Now, there is an intermediate classification – the spinor genus. It is obtained by studying in more details elements the orthogonal group of the underlying vector space. The spinor genus breaks down the problem of finding classes in a genus in two. One then studies

- spinor genera in a genus
- classes in a spinor genus

The importance of this intermediate classification stems from the following result of Eichler.

Theorem 4.1 (Eichler) *For an indefinite form of rank at least* 3*, the spinor genus and the class coincide.*

It happens that for integral trace forms, the genus and spinor genus coincide.

Theorem 4.2 [2] *For a nonquadratic number field, the genus of the integral trace form contains only one proper spinor genus.*

Thus for an indefinite trace form (nontotally real number field), the local-global theorem holds. Hence, one can check isometry classes by looking locally. Now, the question about arithmetical equivalence and isometry of trace form is resolved by the following theorem of Mantilla-Soler.

Example 4.2 [3] Let K, L be two non-totally real, tamely ramified, arithmetically equivalent number fields. Then the integral trace forms are isometric.

Now, let $F/K/\mathbb{Q}$ be a finite extension of number fields. One of the things I am working on is the generalisation of the theorem of Mantilla-Soler to the relative case.

References

[1] Guillermo Mantilla-Soler., *On Number Fields with Equivalent Integral Trace Forms*. International Journal of Number Theory 8/7 (2012), 1569–1580.

[2] Guillermo Mantilla-Soler, *Weak Arithmetic Equivalence*. Canadian Mathematical Bulletin 58/1 (2015), 115-127. doi: `http://dx.doi.org/10.4153/CMB-2014-036-7`..

[3] Guillermo Mantilla-Soler, *The Spinor Genus of the Integral Trace*. Transactions of the American Mathematical Society 369/3 (2017), 1611–1626. doi: `http://dx.doi.org/10.1090/tran/6723`..

Secure And Scalable Management of Internet of Things Deployments

Moreno Ambrosin (*)

Abstract. Recent years have seen the advent of Internet of Things (IoT), which is populating the world with billions of low cost heterogeneous interconnected devices. IoT devices are quickly penetrating in many aspects of our daily lives, and enabling new innovative services, ranging from fitness tracking, to factory automation. Unfortunately, their wide use, as well as their low-cost nature, makes IoT devices also an attractive target for cyber attackers, which may exploit them to perform various type of attacks, such as Denial of Service (DoS) attacks or privacy violation of end users. Furthermore, the potentially very large scale of IoT systems and deployments, makes the use of existing security solutions practically unfeasible.
This document gives an overview of the problem of secure management, and presents our research effort in defining secure and scalable solutions for managing large IoT deployments. We focus in particular on two important parts of the device management process: (1) software updates distribution; and (2) device integrity check.

1 Context

Recent years have seen a growing trend in the diffusion of "smart", interconnected, and low-cost devices. Such trend gave the rise to a new paradigm known as the *Internet of Things* (IoT) [18], which is defined by the International Telecommunication Union as [13]:

> "... a global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving interoperable information and communication technologies (ICT)."

The term "Internet of Things" was coined in 1999 by Kevin Ashton executive director of the Auto-ID Center at the Massachusetts Institute of Technology (MIT), in the domain of RFID, and has later evolved and is commonly used in press, books and scientific literature [24]. IoT is a broad term. It is not the result of a single technology, but rather synthesizes a set of emerging technologies that try to "fill the gap" between the physical and the digital world.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: ambrosin@math.unipd.it . Seminar held on January 18th, 2017.

The IoT is expected to populate the world with billions of interconnected smart objects, according to forecasts of authoritative companies and research centers. For example, Cisco estimates 24 billion Internet-connected objects by 2019 [21], while Huawei forecasts 100 billion IoT connections by 2025 [14]. Smart devices are used for various purposes, ranging from simple sensing and data collection (e.g., temperature, energy, pollution measurement) to automation. Moreover, IoT devices have very heterogeneous characteristics, such as physical equipment, e.g., network card, CPU, memory, sensors, supported communication technologies and protocols (e.g., ZigBee [11] or Bluetooth Low Energy [4]), and firmware.

There are several envisioned IoT use cases, and, consequently, deployment and communication models (Device-to-Device, Device-to-Cloud, or Device-to-Gateway [26, 30] and architectures. Examples are small and medium scale scenarios, e.g., home automation [9], or large and massive scale deployments, e.g., industrial automation (or smart factories [31]), smart health infrastructures [19], buildings automation [23], or energy management (e.g., smart metering systems [12]. This plethora of objects generates massive amount of data, is naturally distributed and often organized in heterogeneous subsystems, and has different requirements, e.g., latency or processing time [20].

2 Problem Statement

The pervasiveness of IoT devices and systems, makes them a potential source of security and privacy "headaches". Indeed, on one hand, many safety critical systems rely on the correct operative state and the security of IoT devices, e.g., sensors and actuators in a smart car. Ensuring the correct operation of these devices, and quickly react to attacks is fundamental, as vulnerabilities may be a threat for the life of the end users [5]. On the other hand, IoT devices constantly collect large amount of data from the environment, or directly from end users, and often communicate them to cloud services, either directly or through gateways. This represents a potential threat users' privacy, as indicated by several studies [22, 25]. Additionally, low cost popular IoT devices, such as fitness trackers, have been shown to be easily attackable by hackers, interested in violating users' privacy or in manipulating the tracked data to fraudulently gain financial benefits or even influence a court trial [3].

3 Research Contributions

Our research work looked at the IoT from a security and privacy perspective. Our research analyzes specific emerging security requirements in various IoT scenarios, mainly targeting medium and large scale deployments, and providing novel practical solutions to solve them. In particular, this document describes our research contribution in designing *Secure solutions for scalable management of IoT devices*, targeting large scale deployments.

Secure and efficient management of IoT deployments is an important aspect considered in recent years in industrial environments [1, 6], and an open research field [26, 29]. At a high-level, management of IoT deployments must take into account two key aspects that have a direct impact on the design of secure management solutions. First, IoT devices are

often low cost, and have low computing power, and small size. As a consequence, IoT devices are easier to be violated by attackers, often unable to perform complex cryptographic operations, and have scarce storage resources. Second, every management solution must be scalable: IoT deployments, such as smart metering or healthcare infrastructures, have potentially billions of interconnected devices.

In our research work, we considered the system model in Figure 1, where a (trusted) management entity is in charge of carrying out management activities (e.g., software updates distribution, or deployment integrity verification) on a potentially large deployment of heterogeneous IoT devices. These operations are typically facilitated by a third-party distribution network, which provides (at least) two main features: data caching (to speed-up one-to-many data distribution), and data aggregation (for many-to-one data collection). Examples are Content Delivery Networks (CDN) such as Akamai [2], or generic cache-enabled network such as Named-Data Networking (NDN) [7]. We consider this network as *untrusted*, meaning that it cannot be trusted for data confidentiality or integrity preservation.

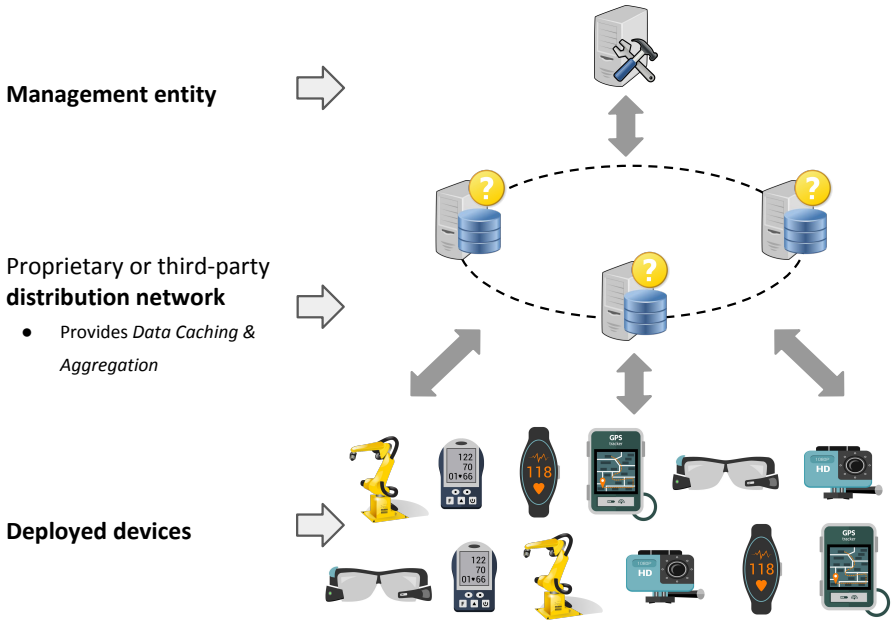


Figure 1. System Model.

Our work provided contributions on two main building blocks of a secure management service: Secure and timely delivery of software updates patches (Section 3.1), and device software integrity verification (Section 3.2).

3.1 Scalable and Secure Software Updates Delivery

Software or patch updates delivery is of paramount importance to guarantee both security and correct operation of smart devices.

From a pure security perspective, most of the new software vulnerabilities can be resolved by applying software updates; additionally, the Open Web Application Security Project (OWASP) has recently indicated insecure software update delivery as one of the top 10 security concerns in IoT [8]. Hence, fast and secure delivery of software updates plays a key role in securing software systems. In particular, once a vulnerability is published, the system becomes exposed to a large base of potential adversaries. A fast update is therefore fundamental (e.g., see the case of the recent SSL "Heartbleed" vulnerability [28]). Additionally, there are many cases in which software updates are required to be confidential. Examples include protection of embedded software against reverse-engineering, or the distribution of valuable map updates in automotive systems and portable devices. However, most of the existing remote update protocols focus on ensuring integrity and authenticity of the transmitted updates, i.e., they guarantee that only untampered updates from a legitimate source will be installed on the device. Furthermore, the use of end-to-end encryption between the software updates distributor and each device, e.g., using SSL [27], is hard to scale, or requires undesirable trust in the intermediate distribution networks, which should cache and replicate unencrypted contents.

To overcome the above issues, we designed Updaticator [15], a protocol for the distribution of confidential software updates that does not require the management entity to trust the third-party untrusted intermediate distribution infrastructure. Updaticator adopts a data-centric communication model between the end devices and the update distribution service, and guarantees confidentiality for the distributed software using Ciphertext-Policy Attribute-Based Encryption (CP-ABE), a novel cryptographic tool that allows to cryptographically enforce high-level access control policies on data. We tested Updaticator through simulation, proving its scalability, and discussing its resiliency against attacks on data confidentiality and integrity that may be carried out by the untrusted distribution network.

3.2 Scalable Collective Attestation

Unlike traditional computing devices, smart devices that are deployed in massive numbers are often limited in cost, computing power, and size. As a consequence, they often lack the security capabilities of general purpose computers: a skilled adversary can easily attack such devices, and compromise both their privacy and safety. One common attack is to modify or replace a device's firmware, as part of a larger attack scenario [5, 10]. In order to react to such attacks and ensure the safe and secure operation of a device, it is important to guarantee its *software* integrity, e.g., via *remote software attestation*. Typical remote software attestation protocols involve a *prover*, which needs to prove its software integrity to a *remote verifier*, i.e., that its firmware is in a known "good" state. However, while remote attestation is a well established research area, existing schemes are meant for one-to-one attestation, and are thus hard to scale to large IoT deployments. The existing state-of-the-art for scalable smart device attestation is represented by SEDA [17], which allows hop-by-hop aggregation of attestation responses along an attestation tree. Unfortunately, SEDA presents important limitations, and in particular: (1) it requires all the devices to be equipped with a trusted execution environment, and to participate in the attestation process; and (2) it considers a software only adversary.

We took a step forward from SEDA [17], and designed SANA [16] a collective attestation protocol that works in more realistic IoT settings. SANA's output is publicly verifiable, meaning that any third party can verify the outcome of the attestation process; furthermore, it does not require trust in the intermediate aggregation network, which is in charge of collecting and aggregating the attestation proofs from devices: the nodes of this network are not required to take part to the remote attestation process. SANA makes use of a novel signature scheme that we called Optimistic Aggregate Signature (OAS), to securely collect and aggregate attestation response from provers. Our protocol is proved to be both secure and scalable.

4 Conclusions

The ever increasing demand of connectivity and services that rely on distributed sensing is populating the world with millions of tiny devices, expected to reach 100 billions by 2025. This phenomenon is commonly referred to as the Internet of Things (IoT), and is constantly creating new marketing opportunities, and new services for users. However, despite its clear advantages, the IoT can be source of new security and privacy threats. Indeed, the penetration of IoT devices in many different scenarios, ranging from home automation, to safety critical environments, and their limited power and simple nature, makes them particularly attractive targets for attackers. For this reason, there is a growing effort both in research and industry in designing novel, efficient and effective security and privacy solutions.

This document briefly summarized our contributions in the research area of secure and scalable device management for large IoT deployments. Our work assumed a realistic scenario where devices are managed by a management entity, via an untrusted intermediate network, that provides data aggregation and caching. We proposed two protocols: Updaticator [15] and SANA [16]. Updaticator is an updates distribution protocol, that guarantees both confidentiality and integrity of the updates, while maintaining their distribution efficient; SANA is a scalable protocol for collective devices integrity verification (a.k.a., attestation), which overcomes several limitations of previous works in this area, e.g., the lack of flexibility and vulnerability to hardware-attacks.

References

[1] 5 Key Elements of IoT Device Management. `http://proximetry.com/`. [Last A.: 2016-05-22].

[2] Akamai. `http://www.akamai.com`. [Last Accessed: 2013-12-10].

[3] Are fitness trackers fit for security?. `https://www.tu-darmstadt.de/vorbeischauen/aktuell/news_details_157888.en.jsp`. [Last Accessed: 2016-09-29].

[4] Bluetooth low energy. `https://www.bluetooth.com/what-is-bluetooth-technology/bluetooth-technology-basics/low-energy`. [Last Accessed: 2016-05-22].

[5] Jeep Hacking 101. http://spectrum.ieee.org/cars-that-think/transportation/systems/jeep-hacking-101. [Last Accessed: 2015-12-10].

[6] Machine-to-machine device management. http://internet-of-things-innovation.com/products/m2m-device-management/. [Last Accessed: 2016-05-22].

[7] Named Data Networking project (NDN). http://named-data.org. [Last A.: 2013-12-10].

[8] Owasp internet of things project. https://www.owasp.org/index.php/OWASP_Internet_of_Things_Top_Ten_Project. [Last Accessed: 2016-08-05].

[9] Samsung smart home. http://www.samsung.com/us/. [Last Accessed: 2016-09-10].

[10] Target attack shows danger of remotely accessible HVAC systems. http://www.computerworld.com/article/2487452/cybercrime-hacking/target-attack-shows-danger-of-remotely-accessible-hvac-systems.html. [Last Accessed: 2015-12-10].

[11] The ZigBee Alliance. www.zigbee.com. [Last Accessed: 2016-05-22].

[12] Smart Meters and Smart Meter Systems: A Metering Industry Perspective. White paper eei-aeic-utc, 2011. Edison Electric Institute, http://www.eei.org/issuesandpolicy/grid-enhancements/documents/smartmeters.pdf.

[13] Overview of the Internet of Things. Recommendation itu-t y.2060, 2012. International Telecommunication Union (ITU), https://www.itu.int/rec/T-REC-Y.2060-201206-I.

[14] Global Connectivity Index 2015. White paper, 2015. Huawei Technologies, http://www.digitaleschweiz.ch/wp-content/uploads/2016/05/Huawei-global-connectivity-index-2015-whitepaper-en-0507.pdf.

[15] M. Ambrosin, C. Busold, M. Conti, A.-R. Sadeghi, and M. Schunter., *Updaticator: Updating billions of devices by an efficient, scalable and secure software update distribution over untrusted cache-enabled networks.* In Proceedings of the 2014 European Symposium on Research in Computer Security, ESORICS '14, 76–93. Springer, 2014.

[16] M. Ambrosin, M. Conti, A. Ibrahim, G. Neven, A.-R. Sadeghi, and M. Schunter, *SANA: Secure and Scalable Aggregate Network Attestation.* In Proceedings of the 23rd ACM Conference on Computer and Communications Security, CCS '16, 731–742. ACM, 2016.

[17] N. Asokan, F. Brasser, A. Ibrahim, A.-R. Sadeghi, M. Schunter, G. Tsudik, and C. Wachsmann, *Seda: Scalable embedded device attestation.* In Proceedings of the 2015 ACM Conference on Computer and Communications Security, CCS '15, 964–975. ACM, 2015.

[18] L. Atzori, A. Iera, and G. Morabito, *The Internet of Things: A survey.* Computer Networks, 54(15):2787–2805, Elsevier, 2010.

[19] M. M. Baig and H. Gholamhosseini, *Smart health monitoring systems: an overview of design and modeling.* Journal of medical systems, 37(2):1–14, Springer, 2013.

[20] F. Bonomi, R. Milito, P. Natarajan, and J. Zhu, *Fog computing: A platform for internet of things and analytics.* In Big Data and Internet of Things: A Roadmap for Smart Environments, 169–186. Springer, 2014.

[21] D. Evans, *The Internet of Things - How the Next Evolution of the Internet Is Changing Everything.* White paper, 2011. Cisco IBSG, https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf.

[22] M. Lisovich, D. K. Mulligan, S. B. Wicker, et al., *Inferring personal information from demand-response systems.* IEEE Security & Privacy Magazine, 8(1):11–20, IEEE, 2010.

[23] L. Mainetti, L. Patrono, and A. Vilei, *Evolution of wireless sensor networks towards the internet of things: A survey.* In Proceedings of the 2011 IEEE International Conference on Software, Telecommunications and Computer Networks, SoftCOM '11, 1–6. IEEE, 2011.

[24] F. Mattern and C. Floerkemeier., *From the internet of computers to the internet of things.* In From active data management to event-based systems and more, 242–259. Springer, 2010.

[25] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin, *Private memoirs of a smart meter. In 2010 ACM workshop on embedded sensing systems for energy-efficiency in building.* ACM BuildSys '10, 61–66. ACM, 2010.

[26] K. Rose, S. Eldridge, and L. Chapin, *The Internet of Things: An Overview.* The Internet Society, 1–50, 2015.

[27] J. Samuel, N. Mathewson, J. Cappos, and R. Dingledine, *Survivable key compromise in software update systems.* In Proceedings of the 2010 ACM Conference on Computer and Communications Security, CCS'10, 61–72. ACM, 2010.

[28] B. Schneier., *Heartbleed ssl protocol vulnerability.* `https://www.schneier.com/blog/archives/2014/04/heartbleed.html`. [Last Accessed: 2014-05-28].

[29] A. Shipley, *Security in the Internet of Things - Lessons from the Past for the Connected Future.* Technical report, 2015. Wind River, `http://www.windriver.com/whitepapers/security-in-the-internet-of-things/wr_security-in-the-internet-of-things.pdf`.

[30] H. Tschofenig, J. Arkko, D. Thaler, and D. McPherson, *Architectural Considerations in Smart Object Networking.* RFC 7452, 2015. RFC Editor, `https://tools.ietf.org/html/rfc7452`.

[31] D. Zuehlke, *Towards a factory-of-things.* Annual Reviews in Control, 34(1):129–138, Elsevier, 2010.

Zeta functions associated to profinite groups

Leone Cimetta (*)

1 Some preliminaries

Group theory is a quite recent branch of mathematics: we can consider its origin in the late 18th century, with some work of Lagrange and Ruffini, while (as it is well known) the first explicit mention of groups as an algebraic structure relies in the 1830s, with the work of Galois. Now the applications of group theory involve many branches of science, such as analysis, geometry, physics, chemistry. One of the main issues with groups is that, although they have a very simple definition, there are many pathological cases, expecially in the infinite case. In order to work with infinite groups it is often useful to be able to define a topological structure on them, with basic properties.

Definition 1 A profinite group is a Hausdorff, compact, and totally disconnected topological group.

We recall that:

Recall 2 A topological space is:

- Hausdorff if any two distinct points admit disjoint neighbourhoods;
- compact if each of its open covers has a finite subcover;
- totally disconnected if points are its only connected components.

An equivalent definition, from a categorical point of view, is the following:

Definition 3 A profinite group is the inverse limit of an inverse system of finite discrete groups.

Example 4 Finite groups are profinite.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: leone.cimetta@gmail.com . Seminar held on February 1st, 2017.

Example 5 p-adic integers (for any prime p) are profinite.

Profinite groups represent a large class of groups with some useful properties:

Proposition 6 *Products of (arbitrarily many) profinite groups are profinite.*

Proposition 7 *Closed subgroups and quotients of profinite groups are profinite.*

Proposition 8 *Given a profinite group G, it is possible to define a finite, non-trivial, left-translation-invariant measure on its open subsets (Haar's Theorem).*
It is always possible to normalize it, so that $\nu(G) = 1$, thus providing a way to compute the probability of certain events.

We recall now some basic definitions and properties of groups:

Recall 9 If $H \leq G$, then the index of H in G ($|G : H|$) is the number of cosets of H in G. Intuitively, it measures "'how many copies of H"' are required to fill G. If G is finite, then $|G : H| = |G|/|H|$.

Recall 10 $H \leq G$ is normal in G ($H \trianglelefteq G$) if $ghg^{-1} \in H$ for any $h \in H, g \in G$.
Let $S \subset G$, then S generates G ($\langle S \rangle = G$) if the smallest subgroup of G containing S is G itself.
Let $S \subset G$, then S normally generates G if the smallest normal subgroup of G containing S is G itself.
G is called finitely generated if it has a finite generating subset.

Finally, we state one last property of profinite groups.

Proposition 11 *A subgroup of a profinite group is open if and only if it is closed and has finite index.*
A subgroup of a finitely generated profinite group is open if and only if it has finite index.

2 Two problems on profinite groups

We will deal with the following problems, which has many computational implications:

Problem 12 *Given a group G, is there an easy way to compute the probability that k randomly chosen elements of G generate the whole group?*

This problem, which was solved by P. Hall in 1936 in the finite case, has unexpected relations with the problem of the subgroup growth.

Problem 13 *Given a group G, for any positive integer n consider the number $a_n(G)$ of subgroups of G of index n. Whats is the behaviour of $a_n(G)$ as a function in the variable n (it is called subgroup growth)?*

In order to face these questions, let us consider the lattice $\mathcal{L}$ of all subgroups of G: then we can define a Möbius function on it in the following way.

Definition 14 $\mu(G,G) = 1$;
$\mu(H,G) = -\sum_{H<K\leq G} \mu(K,G)$ for $H < G$.

Using $\mu(\cdot, G)$ we are able to define new coefficients and hence two Dirichlet series associated to G:

Definition 15 $b_n(G) = \sum_{|G:H|=n} \mu(H,G)$

Definition 16 Let G be a finite group, then consider the Dirichlet polynomials associated to the coefficients $a_n(G)$ and $b_n(G)$: $\zeta_G(s) := \sum_n \frac{a_n(G)}{n^s}$
is the subgroup zeta function associated to G; $p_G(s) := \sum_n \frac{b_n(G)}{n^s}$
is the inverse of the probabilistic zeta function associated to G.

In the finite case, the second Dirichlet series solves the problem of generation we stated:

Proposition 17 (Hall, 1936) *Let G be a finite group and $t \in \mathbb{N}$: then $p_G(t)$ is the probability that t randomly chosen elements of G generate G.*

Problem 18 *How does $p_G(x)$ behave, for a finite G, as a function on real numbers?*

Proposition 19 *Let M be the largest positive integer such that $b_M(G) \neq 0$. Then*

$$\lim_{x\to+\infty} p_G(x) = 1 + \lim_{x\to+\infty} \left(\sum_{n=2}^{M} \frac{b_n(G)}{n^x} \right) = 1,$$

$$\lim_{x\to-\infty} p_G(x) = \lim_{x\to-\infty} \left(\sum_{n=1}^{M} \frac{b_n(G)}{n^x} \right) = \mathrm{sgn}(b_M(G))\infty.$$

If $d(G) = 1$, then $p_G(x)$ is strictly increasing; if $d(G) > 1$, it can be shown that there exist a minimal $x_+ \in \mathbb{R}$, $x_+ > d(G) - 2$ and a maximal $x_- \in \mathbb{R}$ such that $p_G(x)$ is monotone on $]-\infty, x_-]$ and monotone increasing on $[x_+, +\infty[$.
[Shareshian]If G is a non-abelian simple group, $p'_G(1) = 0$.

It seems natural to ask if there is an easy way to extend Hall's result to the infinite case.

(a) We need a notion of probability, thus we need to focus on profinite groups.

(b) In order to use $a_n(G)$, $b_n(G)$, we need to ask that G has a finite number of subgroups of order n for any $n \in \mathbb{N}$: it is possible to prove that finite generation implies this property.

(c) We need to give conditions to ensure the absolute convergence of $\zeta_G(s)$, $p_G(s)$.

Let G be a finitely generated profinite group with a normalized Haar measure ν. Let $\Phi_G(t)$ be the set of all ordered t-uples generating G, then $\Phi_G(t)$ is the complement in G^t of $\bigcup_{H<_O G} H^t$, which is open. In particular, $\Phi_G(t)$ is closed and hence measurable, thus we can define

$$\underset{G}{Prob}(t) := \nu(\Phi_G(t)).$$

We can now define the two series

$$\zeta_G(s) = \sum_n \frac{a_n(G)}{n^s}$$

and

$$p_G(s) = \sum_n \frac{b_n(G)}{n^s}.$$

It is reasonable to ask about the convergence of these serie:

Proposition 20 (Mann, 2005) *The series $p_G(s)$ is absolutely convergent in a right half-plane of the complex plane if and only if there exist $c_1, c_2 \in \mathbb{N}$ such that:*

(a) $|\mu(H, G)| \leq n^{c_1}$ *for any $H \leq_O G$ such that $|G : H| = n$;*

(b) *the number of open subgroups of G of index n with non-zero Möbius function is bounded by n^{c_2}.*

However, it is impotant to notice that there are properties of G encoded in the coefficients of $p_G(s)$ which do not depend on its convergence. One can ask now if there is a relation between $p_G(s)$ and $\zeta_G(s)$. A first look to some easy examples gives us evidence that there can be such a relation.

Example 21 Consider the (profinite complexion of the) group $\mathbb{Z}$.Then

$$\zeta_{\hat{\mathbb{Z}}}(s) = \sum_n \frac{1}{n^s} = \zeta(s),$$

$$p_{\hat{\mathbb{Z}}}(s) = \sum_n \frac{\mu(n)}{n^s}$$

and it is easy to prove that

$$\zeta_{\hat{\mathbb{Z}}}(s) p_{\hat{\mathbb{Z}}}(s) = 1.$$

In order to find out if this behaviour is common to a wider class of groups, Damian and Lucchini introduced the following definition.

Definition 22 A finitely generated profinite group G is ζ-reversible if $\zeta_G(s)p_G(s) = 1$.

Example 23 $\hat{\mathbb{Z}}$ and $\mathbb{Z}_p$ for any prime p are ζ-reversible.

Notice that ζ-reversibility is a condition on the coefficients $a_n(G)$, $b_n(G)$ and can be introduced and studied independently of the convergence of $\zeta_G(s)$, $p_G(s)$. Hence ζ-reversible only means that $\sum_{rs=n} a_r(G)b_s(G) = 0$ for any $n > 0$.
The interest in ζ-reversible groups is motivated by computational difficulties in the finding the coefficients $a_n(G)$ for most groups. In fact, while the series $p_G(s)$ apparently has a more complicated definition than $\zeta_G(s)$, several progresses have been achieved in the last decades in its factorization and thus in its computation. In particular, Detomi and Lucchini showed in 2006 an explicit factorization of $p_G(s)$ using the notion of crowns.

Proposition 24 (Damian and Lucchini, 2014) *Let G be a finitely generated profinite group. Then G is ζ-reversible if and only if*

$$\sum_{m|n} \left(\sum_{|G:H|=m} b_{n/m}(G) - b_{n/m}(H) \right) = 0$$

for all $n \in \mathbb{N}$.

Corollary 25 (Damian and Lucchini, 2014) *If $p_G(s) = p_H(s)$ for all $H \leq_O G$, then G is ζ-reversible.*

As it is apparent from these results, ζ-reversibility is a strong property: a ζ-reversible group must have a sort of uniform subgroup structure, in the sense that the open subgroups must have a comparable structure. Motivated by many results, in particular for some classes of pro-p groups, they conjectured that the converse of the corollary holds:

Conjecture 26 (Damian and Lucchini, 2014) *A profinite group G is ζ-reversible if and only if $p_G(s) = p_H(s)$ for all $H \leq_O G$.*

3 A normal generalization of the two problems

As the coefficients of both Dirichlet series are defined from the lattice of all open subgroups of G, it is reasonable to ask what is the result if we consider a sublattice. In response to this question, in 2007 Detomi and Lucchini introduced some generalizations of the subgroup and the probabilistic zeta functions. Let $\mathcal{L}^{\triangleleft}$ be the lattice of normal subgroups of G. Then it is possible to define:

- a Möbius function $\mu^{\triangleleft}(\cdot, G)$;
- coefficients $a_n^{\triangleleft}(G)$ (the number of open normal subgroups of G of index n);
- coefficients $b_n^{\triangleleft}(G) = \sum_{H \triangleleft G, |G:H|=n} \mu^{\triangleleft}(H, G)$;

- two Dirichlet series $\zeta_G^{\triangleleft}(s)$, $p_G^{\triangleleft}(s)$ associated to $a_n^{\triangleleft}(G)$, $b_n^{\triangleleft}(G)$ respectively.

It is not difficult to prove that the probabilistic meaning still holds in the finite case (i.e., $p_G^{\triangleleft}(t)$ is the probability that t randomly chosen elements of G normally generate G). Moreover, Detomi and Lucchini found in 2007 a factorization of $p_G^{\triangleleft}(s)$ for a finite group G. We recall once again some results on finite groups:

Definition 27 A group S is simple if it has no non-trivial normal subgroups.

Lemma 28 *Let G be a finite group and $\mathcal{N}(G)$ the intersection of all maximal normal subgroups of G. Then $G/\mathcal{N}(G)$ is a direct product of finite simple groups.*

Proposition 29 (Detomi and Lucchini, 2007) *Let G be a finite group and let*

$$G/\mathcal{N}(G) \cong \prod_{i=1}^{m} S_i^{n_i}$$

where S_i are non-isomorphic simple groups. Then

$$p_G^{\triangleleft}(t) = \prod_{i=1}^{m} p_{S_i^{n_i}}^{\triangleleft}(t).$$

Moreover, for a simple group S,

$$p_{S^n}^{\triangleleft}(t) = \begin{cases} (1 - 1/|S|^t)^n & \text{if } S \text{ is not abelian;} \\ \prod_{j=1}^{n}(1 - p^{j-1}/p^t) & \text{if } S \text{ is abelian of order } p. \end{cases}$$

In the profinite case, we can define $Prob_G^{\triangleleft}(t) = \nu(\Phi_G^{\triangleleft}(t))$. It is also possible, with some tecnical instruments, to extend the factorization we have presented in the finite case.

Definition 30 Let G be a profinite group and S a finite simple group. Then $\gamma_G(S)$ is the maximum $t \in \mathbb{N} \cup \{\infty\}$ such that $S^{\gamma_G(S)}$ is a continuous epimorphic image of G.

Definition 31 A profinite group G is positively finitely normally generated (PFNG) if there exists $t \in \mathbb{N}$ such that $Prob_G^{\triangleleft}(t) > 0$.

Remark 32 Clearly, PFNG implies normally finitely generated. The converse does not hold.

Proposition 33 (Detomi and Lucchini, 2007) *Let G be a PFNG profinite group. Then $\gamma_G(S)$ is finite for any simple group S, the infinite products*

$$A(G, s) = \prod_{p \in \mathcal{P}} \prod_{i=1}^{\gamma_G(C_p)} \left(1 - \frac{p^{i-1}}{p^s}\right)$$

$$B(G,s) = \prod_{S \in \Sigma} \left(1 - \frac{1}{|S|^s}\right)^{\gamma_G(S)}$$

(where $\mathcal{P}$ is the set of all prime numbers and Σ the sets of all non-abelian finite simple groups) are absolutely convergent in a right half-plane of the complex plane. Moreover, $p_G^{\triangleleft}(s) = A(G,s)B(G,s)$ in the same half-plane and

$$\overset{\triangleleft}{Prob}_G(t) = p_G^{\triangleleft}(t)$$

for t large enough.

It can be proved that the condition for G to be PFNG is crucial for the convergence of $p_G^{\triangleleft}(s)$ and can be verified looking at $G/\mathcal{N}(G)$.

Proposition 34 (Cimetta and Lucchini, 2016) *Let G be a profinite group. Then the following are equivalent:*

(1) *The infinite sum $\displaystyle\sum_{H \trianglelefteq_O G} \frac{\mu^{\triangleleft}(H,G)}{|G:H|^s}$ absolutely converges in a right half-plane of the complex plane;*

(2) *G is PFNG;*

(3) *$G/\mathcal{N}(G)$ is finitely generated.*

As it is apparent, we are able (also in this generalization) to provide a factorization for the series $p_G^{\triangleleft}(s)$, while it is general very difficult to compute $\zeta_G^{\triangleleft}(s)$; furthermore, groups like $\hat{\mathbb{Z}}$, $\mathbb{Z}_p$ satisfy the property $\zeta_G^{\triangleleft}(s)p_G^{\triangleleft}(s) = 1$, so it is reasonable to define normally ζ-reversible groups.

Definition 35 A profinite group G is normally ζ-reversible if $\zeta_G^{\triangleleft}(s)p_G^{\triangleleft}(s) = 1$.

Normal ζ-reversibility seems to be a very rare property, that only few groups satisfy: it is possible to extend to the normal case some results in the non-normal case. Our main conjecture is the following:

Conjecture 36 *A profinite group G is normally ζ-reversible if and only if it is abelian and torsion-free (i.e., $g^n \neq 1$ for any $g \in G$, $n \in \mathbb{N}$).*

Some partial results supporting this conjecture have been proved.

References

[1] B. Benesh, *The probabilistic zeta function.* Computational group theory and the theory of groups, II, 1–9, Contemp. Math., 511, Amer. Math. Soc., Providence, RI, 2010.

[2] N. Boston, *A probabilistic generalization of the Riemann zeta function.* Analytic number theory, Vol. 1 (Allerton Park, IL, 1995) 138 (1996), 155–162.

[3] K. Brown, *The coset poset and probabilistic zeta function of a finite group.* J. Algebra 225 (2000), no. 2, 989–1012.

[4] H.M. Crapo, *Möbius inversion in lattices.* Arch. Math. (Basel) 19 1968, 595–607 (1969).

[5] E. Damian and A. Lucchini, *Profinite groups in which the probabilistic zeta function coincides with the subgroup zeta function.* J. Algebra 402, 92–119 (2014).

[6] E. Detomi and A. Lucchini, *Some generalizations of the probabilistic zeta function.* Ischia group theory 2006, 56–72, World Sci. Publ., Hackensack, NJ, 2007.

[7] M.P.F. du Sautoy and L. Woodward, "Zeta functions of groups and rings". Lecture Notes in Mathematics, vol. 1925, Springer, Heidelberg (2008).

[8] P. Hall, *The eulerian functions of a group.* Quart. J. Math. (1936), no. 7, 134–151.

[9] A. Lubotzky and D. Segal, "Subgroup growth". Progress in Mathematics, 212. Birkhauser Verlag, Basel, 2003.

[10] A. Mann, *Positively finitely generated groups.* Forum Math. 8 (1996), no. 4, 429–459.

[11] A. Mann, *A probabilistic zeta function for arithmetic groups.* Internat. J. Algebra Comput. 15 (2005), 1053–1059.

[12] J. Shareshian, *On the probabilistic zeta function for finite groups.* J. Algebra, 210 (1998), 703–770.

Collective periodic behavior in interacting particle systems

Daniele Tovazzi (*)

Abstract. Interacting particle systems constitute a wide class of models, originally motivated by Statistical Mechanics, which in the last decades have become more and more popular, extending their applications to various fields of research such as Biology and Social Sciences. These models are important tools that may be used to study macroscopic behaviors observed in complex systems. Among these phenomena, a very interesting one is collective periodic behavior, in which the system exhibits the emergence of macroscopic rhythmic oscillations even though single components have no natural tendency to behave periodically.
This talk aims to introduce to a general audience some basic tools in the theory of interacting particle systems and some of the mechanisms which can enhance the appearance of self-sustained macroscopic rhythm. After recalling some notions of Probability, we present the classical Curie-Weiss model, which doesn't exhibit periodic behavior, and we show how we can modify it in order to create macroscopic oscillations. This is also the starting point for some recent developments that will be sketched in the last part of the talk.

1 Introduction

Living systems are characterized by the emergence of self-organized collective behaviors in large communities of interacting components: prey-predator equilibria, flocks of birds, fireflies glowing in a synchronized manner constitute very examples. A fundamental problem in complex systems is to understand how many interacting individuals organize to produce such coherent behaviors at a macroscopic level. A phenomenon whose enhancing mechanisms are not well-understood yet consists in the emergence of macroscopic stable periodic oscillation in systems whose units have no natural tendency to evolve periodically: real examples in this sense come from Biology, Ecology and Socio-Economics [1, 7, 11, 12]. The attempt of modeling such complex systems leads naturally to consider large families of N microscopical identical units and then, following a typical approach of Statisical Mechanics, to study the *infinite volume limit* dynamics of the system by letting $N \uparrow +\infty$.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: tovazzi@math.unipd.it. Seminar held on February 15th, 2017.

When working with this type of models, it is often assumed to have a mean-field interaction between particles: this means that each single agent interacts with all the others in the same way, and the strength of interaction is of order $O(\frac{1}{N})$. This implies that there is no spatial geometry in the system and the graph describing interactions between particles is the complete graph with N nodes. Mean-field interaction dramatically simplifies the analysis of the system and its macroscopic limit and, even though is a very strong assumption, it is somehow justified for modelling systems where there exists a large number of connections (e.g. neuronal networks) or where information is is shared by all the agents (e.g. financial markets).
Several papers in literature deals with the mechanisms enhancing self-sustained periodic behavior in mean-field models (see [2-6, 8-10]) but little is known whether those mechanisms induce collective periodic behavior in system where the interaction in local (i.e. not of mean-field type).
In this notes, we firstly introduce the Curie-Weiss model, a classical toy model which explains polarization in a ferromagnetic system. Then, we briefly sketch a modification of this model analysed in [4]: by adding a dissipative term in the interaction energy the macroscopic dynamics will present a stable limit cycle at sufficiently low temperature. Finally, we will present some recent results (due to a joint work with Raphael Cerf, Paolo Dai Pra and Marco Formentin) concerning an Ising model with dissipation, which show that the dissipation mechanism enhances the appearance of collective oscillation even if the model is not of mean-field type.

2 Preliminary notions

Definition 1 A **probability space** is a triple $(\Omega, \mathcal{F}, P)$ where Ω is a sample space, $\mathcal{F}$ is a σ-algebra of events and P is probability measure on $(\Omega, \mathcal{F})$.
Given a measurable space $(E, \mathcal{E})$, a **random variable** is a measurable function $X : \Omega \to E$.

A random variable X induces a probability measure μ_X on $(E, \mathcal{E})$, usually called the law of X, defined by

$$\mu_X(A) = P(X \in A), \quad \forall A \in \mathcal{E}.$$

Definition 2 A **stochastic process** is an object

$$\underline{X} = (\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathcal{T}}, (X_t)_{t\in\mathcal{T}}, P)$$

where

- $(\Omega, \mathcal{F}, P)$ is a probability space;
- $\mathcal{T} \subseteq \mathbb{R}^+$ is the time set;
- $(\mathcal{F}_t)_{t\in\mathcal{T}}$ is a **filtration**, i.e. an increasing family of sub-σ-algebras of $\mathcal{F}$: for any $s, t \in \mathcal{T}$ with $s \leq t$

$$\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F};$$

- $(X_t)_{t\in\mathcal{T}}$ is a family of random variables on $(\Omega,\mathcal{F})$ taking values in a measurable space $(E,\mathcal{E})$ and such that it is **adapted** to the filtration $(\mathcal{F}_t)_{t\in\mathcal{T}}$, i.e.

$$X_t \text{ is } \mathcal{F}_t\text{-measurable} \quad \forall\, t\in\mathcal{T}.$$

Definition 3 Let $(E,\mathcal{E})$ be a metric space provided with the Borel σ-algebra, $S\subset\mathbb{R}$. The **Skorohod space**, indicated by $\mathcal{D}(S;E)$, is the set of functions $f: S\to E$ which in any point are right-continuous and have finite limit from the left (often called *cadlag*). This space can be endowed with the **Skorohod topology**, which provides a metric and a Borel σ-algebra $\mathcal{B}$.

Note that if $(X_t)_{t\in\mathcal{T}}$ is a stochastic process with values in E and *cadlag* trajectories, then it can be interpreted as a random variable X taking values in $\mathcal{D}(\mathcal{T};E)$:

$$\begin{array}{rccc} X: & (\Omega,\mathcal{F},P) & \longrightarrow & (\mathcal{D}(\mathcal{T};E),\mathcal{B}) \\ & \omega & \longmapsto & [t\mapsto X_t(\omega)]. \end{array}$$

In particular, being a random variable, $X=(X_t)_{t\in\mathcal{T}}$ induces a probability measure μ_X on $(\mathcal{D}(\mathcal{T};E),\mathcal{B})$.

Definition 4 A sequence of random variables $(X_n)_{n\in\mathbb{N}}$ with values in $(E,\mathcal{E})$ is said to **weakly converge** to a random variable X with values in $(E,\mathcal{E})$ if, for every continuous and bounded $f: E\to\mathbb{R}$,

$$\int_E f(x)\mu_{X_n}(dx) \underset{n\to+\infty}{\longrightarrow} \int_E f(x)\mu_X(dx).$$

Then, the concept of weak convergence for stochastic processes with *cadlag* trajectories is obtained by interpreting it as weak convergence of random variables on the Skorohod space:

$$\begin{array}{c} (X_t^n)_{t\in\mathcal{T}} \underset{n\to+\infty}{\overset{w}{\longrightarrow}} (X_t)_{t\in\mathcal{T}} \\ \Updownarrow \\ \displaystyle\int_{\mathcal{D}(\mathcal{T};E)} f(x)\mu_{X^n}(dx) \underset{n\to+\infty}{\longrightarrow} \int_{\mathcal{D}(\mathcal{T};E)} f(x)\mu_X(dx) \quad \forall\, f\in\mathcal{C}_b(\mathcal{D}(\mathcal{T};E)) \end{array}$$

where μ_{X^n} is law of $(X_t^n)_{t\in\mathcal{T}}$ on $\mathcal{D}(\mathcal{T};E)$ and μ_X is law of $(X_t)_{t\in\mathcal{T}}$ on $\mathcal{D}(\mathcal{T};E)$.

3 The classical Curie-Weiss Model

We consider N sites and we associate with each of them a spin value, which is a random variable taking values in the set $\{-1,+1\}$. We call a configuration of the system

the collection of the spin values $\underline{\sigma} = (\sigma_j)_{j=1}^N \in \{-1,+1\}^N$. Given a configuration, the corresponding **magnetization** is the quantity

$$m_N = \frac{1}{N}\sum_{j=1}^{N}\sigma_j.$$

At time $t = 0$, we construct a configuration $\underline{\sigma}(0) = (\sigma_j(0))_{j=1}^N \in \{-1,+1\}^N$, choosing N independent and identically distributed spins, with common law μ. Our aim is to define a stochastic process $(\underline{\sigma}(t))_{t\in[0,T]}$ with values in $\{-1,+1\}^N$ in which the spins interact with each other by trying to align/cooperate: this means that we have to assign a flipping rate for each spin. We can interpret flipping rates in the following way: at time $t \in [0,T]$, if the j-th spin has flipping rate $\lambda_j(t)$, then for $h > 0$,

$$P(\sigma_j \text{ flips in }]t, t+h[) = \lambda_j(t)h + o(h),$$

hence the higher the flipping rate of a spin at time t, the higher the probability to observe it flip on the interval $]t, t+h[$, for any $h > 0$. The classical Curie-Weiss model is obtained by choosing the rates of transitions in the following form:

$$\sigma_j \longrightarrow -\sigma_j \qquad \text{at rate} \qquad e^{-\beta\sigma_j m_N},$$

where β, positive parameter, measure the strength of interaction (it actually the inverse of the temperature of the system). At this point, it is worth to stress some aspects:

- the interaction depends entirely on the value of the magnetization of the system $m_N(t)$: in particular, this interaction is of mean-field type;
- we obtained a cooperative type of interaction, in the following sense: if $m_N(t) > 0$ (so there is a majority of positive spins in the configuration $\underline{\sigma}(t)$), then positive spins have flipping rates less than 1 and negative spins have flipping rates grater than one (and vice versa with $m_N(t) < 0$), meaning that spins are trying to align with each other. Note also that the strength of this alignment mechanism depend on the parameter β.

Actually, one can take a step further: thanks to the mean-field hypothesis, the magnetization $m_N(t)$ encodes all the information of the state of the system (how many spins are positive and how many are negative) and also on the flipping rates of the system. Hence, the stochastic process $(m_N(t))_{t\in[0,T]}$ is an **order parameter** of the system, i.e. it is a sufficient statistic to describe the dynamics of the N-particle process. The magnetization process takes values in $\{-1, -\frac{N-2}{N}, \dots, \frac{N-2}{N}, +1\}$ and it jumps with amplitude $+\frac{2}{N}$ when a negative spin flips to positive and with amplitude $-\frac{2}{N}$ when a positive spin flips to negative. Hence, it is easy to see that, at time $t \in [0,T]$, the process jumps "up" with a rate given by the flipping rate of a single negative spin at time t multiplied by the total number of negative spins present in the configuration $\underline{\sigma}(t)$, namely

$$\frac{N(1-m_N(t))}{2}\, e^{\beta m_N(t)}.$$

Analogously, it jumps "down" with a rate given by the flipping rate of a single positive spin at time t multiplied by the total number of positive spins present in the configuration $\underline{\sigma}(t)$, namely

$$\frac{N(1+m_N(t))}{2}\, e^{-\beta m_N(t)}.$$

Hence, for any $t \in [0,T]$ and $h > 0$,

$$P\left(m_N(t+h) = m_N(t) + \frac{2}{N}\right) = h\frac{N(1-m_N(t))}{2}\, e^{\beta m_N(t)} + o(h);$$

$$P\left(m_N(t+h) = m_N(t) - \frac{2}{N}\right) = h\frac{N(1+m_N(t))}{2}\, e^{-\beta m_N(t)} + o(h).$$

Now we are interested in describing the dynamics of the system (hence, of the process $m_N(t)_{t\in[0,T]}$), in the limit as $N \uparrow +\infty$.

Theorem 1 *For $t \in [0,T]$, T fixed, in the limit as $N \uparrow +\infty$, the magnetization process $(m_N(t))_{t\in[0,T]}$ converges, in sense of weak convergence of stochastic processes, to $(m(t))_{t\in[0,T]}$, which is the solution of the following ordinary differential equation*

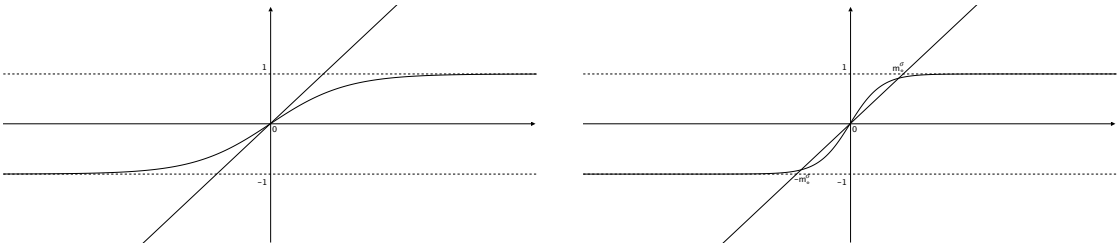
$$\begin{cases} \dot{m}(t) = -2m(t)\cosh(\beta m(t)) + 2\sinh(\beta m(t)), \\ m(0) = m_0 \end{cases} \tag{1}$$

where $m_0 \in [-1,1]$ is such that

$$m_N(0) \xrightarrow[N\uparrow+\infty]{w} m_0.$$

Notice that the macroscopic dynamics is deterministic: the equation (1) drives the behavior of the classical Curie-Weiss model in the infinite volume limit. Now we want to study the long-time behavior of the macroscopic system.

Lemma 1 *Any equilibrium solution of $\dot{m}(t) = -2m(t)\cosh(\beta m(t)) + 2\sinh(\beta m(t))$ is of the form $m^* = \tanh(\beta m^*)$.*



(a) Case $\beta < 1$ (b) Case $\beta > 1$

Figure 1. Solution(s) of $m_* = \tanh(\beta m_*)$.

Theorem 2 *Consider the equation (1):*

- *for $\beta \leq 1$, it has 0 as a unique equilibrium solution and it is globally asymptotically stable, i.e. for every initial condition $m_0 \in [-1, 1]$*

$$\lim_{t \to +\infty} m(t) = 0;$$

- *for $\beta > 1$, the point 0 is still an equilibrium and two further equilibria arise:*

$$m^*_\beta \qquad \text{and} \quad -m^*_\beta\,,$$

where m^*_β is the unique positive solution of $x = \tanh(\beta x)$. In this case, the phase space $[-1, 1]$ is bi-partitioned by the origin in two domains of attraction: given an initial condition m_0,

$$\lim_{t \to +\infty} m(t) = \begin{cases} m^*_\beta & \text{if } m_0 \in (0, 1] \\ -m^*_\beta & \text{if } m_0 \in [-1, 0) \\ 0 & \text{if } m_0 = 0\,. \end{cases}$$

So, the macroscopic system exhibits a phase transition, which is the appearance of multiple stable equilibria, at a critical value $\beta_c = 1$, so its behavior is deeply influenced by the choice of β. If $\beta \leq \beta_c$, the interaction is not strong enough to maintain any form of self-organization and, regardless of any initial condition, we will only observe disorder (i.e. magnetization close to zero) in the long-time behavior. On the other hand, if $\beta > \beta_c$ the interaction is sufficiently strong to allow a polarization phenomenon, where the magnetization converges to a value $m^*(\beta)$ or $-m^*(\beta)$ depending on the initial condition (with the exception of $m_0 = 0$).

4 A Curie-Weiss model with dissipation

In the previous section, we have seen that the classical Curie-Weiss model is toy model which explains self-organized polarization. Since our goal is to study the emergence of macroscopic periodic behavior, we may ask whether it is possible to modify the classical CW model in order to obtain a dynamics which is still simple to analyse but that will exhibit collective periodic behavior. An answer to this question is given by the Curie-Weiss model with dissipation introduced in [4] that will be presented in this section.

We consider again the same framework: we have N sites and at time $t = 0$, we construct a configuration $\underline{\sigma}(0) = (\sigma_j(0))_{j=1}^N \in \{-1, +1\}^N$, choosing N independent and identically distributed spins, with common law μ. The only change lies in flipping rates: now we consider them in the form

$$\sigma_j \longrightarrow -\sigma_j \quad \text{at rate} \quad e^{-\sigma_j(t)\lambda_N(t)},$$

where $(\lambda_N(t))_{t\in[0,T]}$ is a stochastic process that evolves according to

$$d\lambda_N(t) = -\alpha\lambda_N(t)dt + \beta dm_N(t),$$

with $\alpha > 0$, $\lambda(0) = \lambda_0 \in \mathbb{R}$. The term $\lambda_N(t)$ describes *interaction energy*: it jumps with amplitude $\pm\frac{2\beta}{N}$ whenever the magnetization $m_N(t)$ jumps with amplitude $\pm\frac{2}{N}$ and, between two consecutive jumps, it decays exponentially to 0. In this sense, dissipation dumps the influence of interaction when no spin-flip occurs for a long time. Note that:

- if $\alpha = 0$ we recover the classical CW model;
- we are still dealing with a mean-field interaction.

In this case we need a two-dimensional process to completely describe the system: $m_N(t)$ contains the information on the configuration at time $t \in [0,T]$ while $\lambda_N(t)$ is needed to describe the jump rates at time $t \in [0,T]$. In this sense, $(m_N(t), \lambda_N(t))_{t\in[0,T]}$ is an order parameter for the system, hence we want to study its infinite volume dynamics.

Theorem 3 *In the limit as $N \uparrow +\infty$, the process $(m_N(t), \lambda_N(t))_{t\in[0,T]}$ converges, in sense of weak convergence of stochastic processes, to $(m(t), \lambda(t))_{t\in[0,T]}$, solution of the ordinary differential equation*

$$\begin{cases} \dot{m}(t) = -2m(t)\cosh(\lambda(t)) + 2\sinh(\lambda(t)), \\ \dot{\lambda}(t) = -\alpha\lambda(t) + \beta\dot{m}(t), \\ m(0) = m_0, \ \ \lambda(0) = \lambda_0 \end{cases} \tag{2}$$

where $m_0 \in [-1,1]$, $\lambda_0 \in \mathbb{R}$ are such that

$$(m_N(0), \lambda_N(0)) \xrightarrow[N\uparrow+\infty]{w} (m_0, \lambda_0).$$

Lemma 2 *Fix $\alpha > 0$ and consider (2). Then,*

- *for any $\beta > 0$, $(0,0)$ is the only stationary solution;*
- *for $\beta = 1 + \frac{\alpha}{2}$ a Hopf bifurcation is present.*

Recall that, roughly speaking, a Hopf bifurcation occurs when an equilibrium point looses stability and a periodic solution arises. This type of bifurcation can be detected by studying the eigenvalues of the Jacobian matrix in the equilibrium point: there must exist a pair of conjugate eigenvalues crossing the imaginary axis with positive velocity.

Theorem 4 *Fix $\alpha > 0$ and consider (2).*

- *for $\beta \leq 1 + \frac{\alpha}{2}$, $(0,0)$ is globally asymptotically stable, i.e. for every initial condition $(m_0, \lambda_0) \in [-1,1] \times \mathbb{R}$,*

$$\lim_{t\to+\infty} (m(t), \lambda(t)) = (0,0); \ .$$

- *for $\beta > 1 + \frac{\alpha}{2}$, the system has a unique periodic orbit which attracts all other trajectories except the one starting at the equilibrium point, i.e. $(m_0, \lambda_0) = (0, 0)$.*

So, in this case, the macroscopic system exhibits a phase transition, which is the appearance of a stable limit cycle, at a critical value $\beta_c = 1 + \frac{\alpha}{2}$, so its behavior is deeply influenced by the choice of β. If $\beta \leq \beta_c$, the interaction is not strong enough to maintain any form of self-organization and, regardless of any initial condition, we will only observe disorder in the long-time behavior. On the other hand, if $\beta > \beta_c$ the interaction is sufficiently strong to allow for a collective phenomenon, which is no longer polarization as in the classical model but rather a stable oscillation between magnetized states.

5 Some recent developments: an Ising model with dissipation

Now we want to analyse a spin-flip model where the mean-field hypothesis no longer holds, and we want to check if the dissipation mechanism described above is still capable to produce macroscopic rhythmic oscillation. To do so, we start with the classical 1-dimensional Ising model.

In the classical 1-dimensional Ising model with periodic boundary conditions we have N site lying on a 1-dimensional torus. At any site, we associate a spin value $\sigma_i \in \{-1, +1\}$, for $i = 1, \dots, N$. Moreover, for any site, we introduce a **local magnetization**

$$m_{i,N} = \sum_{j \sim i} \sigma_j = \sigma_{i-1} + \sigma_{i+1}, \quad i = 1, \dots, N,$$

which only depends on the spins located in the first neighbours of site i. The flipping rate now is given by

$$\sigma_i \longrightarrow -\sigma_i \quad \text{at rate} \quad e^{-\sigma_i(t) m_{i,N}(t)}.$$

Notice that interaction is no longer of mean-field type since the flipping rate is site-dependent. This also means that the total magnetization $m_N(t) = \frac{1}{N} \sum_{j=1}^{N} \sigma_j(t)$ is far from being an order parameter for the system, so there is no hope to reduce the model to a 1-dimensional process as in the Curie-Weiss framework.

We modify the classical 1-dimensional Ising model by introducing a dissipation, following the idea of the Curie-Weiss model with dissipation: we get that the flipping rates in the form

$$\sigma_i \longrightarrow -\sigma_i \quad \text{at rate} \quad e^{-\sigma_i(t) \lambda_i(t)}$$

where, for any $i = 1, \dots, N$, the process $(\lambda_i(t))_{t \in [0,T]}$ evolves according to

$$d\lambda_i(t) = -\alpha \lambda_i(t) dt + \beta dm_{i,N}(t),$$

with $\alpha, \beta > 0$.
Our aim is to show that in a suitable large volume - low temperature limit, the total

magnetization has a periodic behavior after a suitable time scaling. We assume the dynamics starts with all spins equal to -1. The analysis of the evolution is divided into two parts. We begin by studying the occurrence time of the first spin flip. After the first spin-flip occurs, the change in the local field and the low temperature favors the growth of a "droplet" (just a segment in the one-dimensionale case) of $+1$, which invades the whole state space. At this point we are back the situation of all equal spins. By assigning the initial local fields $\lambda_i(0)$ in a suitable way, the local fields at the time the droplet has invaded the space is essentially opposite to the initial one, producing the iteration of the same phenomenon. To guarantee the growth of the droplet to occur with overwhelming probability, we will assume $\beta, N \uparrow \infty$ in such a way that $\beta \gg \log N$.
In this regime, the waiting time for the first spin flip is large, but has very small fluctuations. These fluctuations, however, have impact on the growth time of the droplet (which is very short). For this reason, while the waiting time of the first spin flip, rescaled by its mean, has a deterministic limit, the rescaled growth time of the droplet keeps some randomness in the limit.

Let $\lambda_i(0) = -\gamma < 0$ and $\sigma_i(0) = -1$ for all $i = 1, \dots, N$, denote with T_1 the time at which the first spin flip is observed, namely

$$T_1 = \inf\{t \in [0,T] \mid \sigma_j(t) = +1 \text{ for some } j \in \{1, \dots, N\}\};$$

denote also with T_c the time take by the positive droplet to expand and cover all the sites, i.e.

$$T_c = \inf\{t \in [0,T] \mid \sigma_i(T_1 + t) = +1 \text{ for all } i = 1, \dots, N\}.$$

Then, the following results hold:

Proposition 1 *Suppose $N, \gamma \uparrow +\infty$ with the condition $\frac{N}{\gamma} e^{-\gamma} \to 0$. Then*

$$\alpha \log N \Big(T_1 - t(\gamma, N) \Big) \xrightarrow[\gamma, N \uparrow +\infty]{d} X,$$

where X is a random variable distributed according to

$$P(X > x) = \exp\left(-e^x\right), \qquad \forall\ x \in \mathbb{R}$$

and

$$t(\gamma, N) := \frac{1}{\alpha} \log \frac{\gamma}{L^{-1}\big(L(\gamma) + \frac{\alpha}{N}\big)}$$

where, for $x > 0$,

$$L(x) = \int_x^{+\infty} \frac{e^{-y}}{y} dy.$$

Proposition 2 *Let $\gamma, \beta, N \uparrow +\infty$ with the condition*

$$\lim_{\beta, N \uparrow +\infty} \frac{\log N}{\beta} = 0 \quad \liminf_{\beta, \gamma \uparrow +\infty} \frac{\gamma}{\beta} > 0.$$

Then

$$\frac{T_c}{\frac{N^2}{2\alpha \log N} e^{-2\beta}} \xrightarrow[\gamma,\beta,N\uparrow+\infty]{d} Z,$$

where $1/Z$ *is an exponential random variable with mean* 1 *and* $X_N := \alpha \log N \Big(T_1 - t(\gamma, N)\Big)$.

Remark 1 As stated before, in order to see a repetition of this phenomenon (long waiting for the first flip and fast covering by the droplet) one has to choose γ in such a way the local fields at the time the droplet has invaded the space is essentially opposite to the initial one, i.e. $\lambda_i(0) \approx -\lambda_i(T_1 + T_c)$. One can check that the correct choice to produce the phenomenon is $\gamma = 4\beta$. Actually, after time T_1 the local fields will no longer be homogeneous, but will have small differences due to the dissipation and the fact that the droplet takes some time to cover all the spins. Anyway, these fluctuations are so small that they don't have any effect in the limit: in the end, we will choose the initial conditions

$$\lambda_i(0) = -4\beta + \varepsilon_i(\beta, N), \qquad i = 1, \dots, N$$

where $\varepsilon_i(\beta, N)$ are allowed to be random perturbation such that

$$\lim_{\beta,N\uparrow+\infty} \frac{\log N}{\beta} \max_i |\varepsilon_i(\beta, N)| = 0$$

in probability. One can prove that with this choice, Proposition 1 and Proposition 2 are still true and, at time $T_1 + T_c$, we recover a condition comparable to the initial one, which allows repetition of the phenomenon hence, in the limit, macroscopic oscillations.

Notice that $t(4\beta, N)$ is a diverging quantity as $\beta, N \uparrow +\infty$, while $\frac{N^2}{2\alpha \log N} e^{-2\beta}$ converges to 0 as $\beta, N \uparrow +\infty$, assuming to work in the regime $\beta^{-1} \log N \to 0$. Therefore, in order to state a convergence theorem for the total magnetization process $(m_N(t))_{t\in[0,T]}$ we have to consider a time rescaled version of it. Define $(\tilde{m}_N(t))_{t\in[0,T]}$ in the following way:

$$\tilde{m}_N(t) = m_N(\theta_N(t)), \qquad t \in [0, T]$$

where

$$\theta_N(t) = \int_0^t \left(t(4\beta, N) \mathbb{1}_{\{|m_N(s)|=1\}} + \frac{N^2 e^{-2\beta}}{2\alpha \log N} \mathbb{1}_{\{|m_N(s)|<1\}} \right) ds, \qquad t \in [0, T].$$

Now, let $(Z_i)_{\{i\in\mathbb{N}\}}$ be a sequence of i.i.d. random variables such that $\frac{1}{Z_1}$ is distributed as an exponential r.v. of mean 1. Then, define the sets

$$A = [0, 1 \wedge T] \cup \left(\bigcup_{k=1}^{+\infty} \left[\left(\sum_{i=1}^{2k} Z_i + 2k \right) \wedge T, \left(\sum_{i=1}^{2k} Z_i + 2k + 1 \right) \wedge T \right] \right),$$

$$B = \bigcup_{k=1}^{+\infty} \left[\left(\sum_{i=1}^{2k-1} Z_i + 2k - 1 \right) \wedge T, \left(\sum_{i=1}^{2k-1} Z_i + 2k \right) \wedge T \right],$$

$$C =]1 \wedge T, (1 + Z_1) \wedge T[\,,$$

and, for any integer $k \geq 1$,

$$U_k = \left]\left(\sum_{i=1}^{2k} Z_i + 2k + 1\right) \wedge T, \left(\sum_{i=1}^{2k+1} Z_i + 2k + 1\right) \wedge T\right[,$$

$$V_k = \left]\left(\sum_{i=1}^{2k-1} Z_i + 2k\right) \wedge T, \left(\sum_{i=1}^{2k} Z_i + 2k\right) \wedge T\right[.$$

Note that

$$[0, T] = A \cup B \cup C \cup \left(\bigcup_{k=1}^{+\infty} U_k \cup V_k\right),$$

then define the process $(\tilde{m}(t))_{t\in[0,T]}$ in the following way:

$$\tilde{m}(t) = \begin{cases} -1 \text{ if } t \in A, \\ +1 \text{ if } t \in B, \\ \frac{2t-2}{Z_1} - 1 \text{ if } t \in C, \\ \frac{2t-2\sum_{i=1}^{2k} Z_i + 4k + 2}{Z_{2k+1}} - 1 \text{ if } t \in U_k \text{ for some } k \geq 1, \\ \frac{-2t+2\sum_{i=1}^{2k} Z_i + 4k}{Z_{2k}} - 1 \text{ if } t \in V_k \text{ for some } k \geq 1. \end{cases}$$

Theorem 5 *As $\beta, N \uparrow +\infty$, the process $(\tilde{m}_N(t))_{t\in[0,T]}$ converges, in sense of weak convergence of stochastic processes, to the process $(\tilde{m}(t))_{t\in[0,T]}$.*

The behavior of the process $(\tilde{m}(t))_{t\in[0,T]}$ is intuitively illustrated in Figure 2: it presents oscillations between the full-magnetized states -1 and +1. In this case, some randomness is still present and in particular it appears each time the magnetization moves from -1 to +1 (or vice-versa) influencing the time taken to perform this shift. Nevertheless, Theorem 5 shows that a self-sustained oscillating behavior can be induced by the mechanism of dissipation even if the interactions are not of mean-field type.

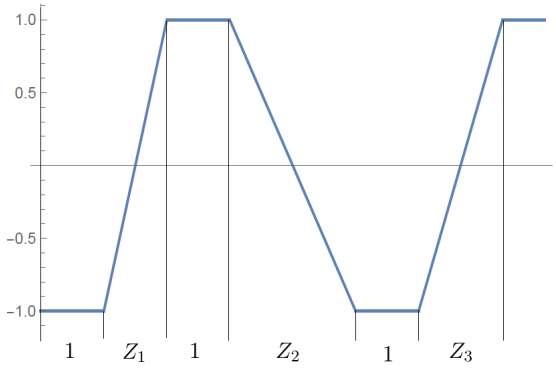


Figure 2. Limit process $(\tilde{m}(t))_{t\in[0,T]}$.

References

[1] W.C. Chen, *Nonlinear dynamics and chaos in a fractional-order financial system.* Chaos, Solitons & Fractals, 36/5 (2008), 1305–1314.

[2] F. Collet, P. Dai Pra and M. Formentin, *Collective periodicity in mean-field models of cooperative behavior.* NoDEA, 22 (2015), 1461–1482.

[3] F. Collet, M. Formentin and D. Tovazzi, *Rhythmic behavior in a two-population mean-field Ising model,.* Phys. Rev. E, 94 (2016), 042139.

[4] P. Dai Pra, M. Fischer and D. Regoli, *A Curie-Weiss model with dissipation.* J. Stat. Phys., 152 (2013), 37–53.

[5] P. Dai Pra, G. Giacomin and D. Regoli, *Noise-induced periodicity: some stochastic models for complex biological systems.* In *Mathematical Models and Methods for Planet Earth*, pages 25-35. Springer, 2014.

[6] S. Ditlevsen and E. Löcherbach, *Multi-class oscillating systems of interacting neurons. arXiv:1512.00265*.

[7] G.B. Ermentrout and D.H. Terman, "Mathematical foundations of neuroscience". Volume 35. Springer Science & Business Media, 2010.

[8] G. Giacomin and C. Poquet, *Noise, interaction, nonlinear dynamics and the origin of rhythmic behaviors.* Braz. J. Prob. Stat., 29(2) (2015), 460–493.

[9] G. Giacomin, C. Poquet and A. Shapira, *Small noise and long time phase diffusion in stochastic limit cycle oscillators. arXiv: 1512.04436*.

[10] B. Lindner, J. Garcia-Ojalvo, A. Neiman and L. Schimansky-Geier, *Effects of noise in excitable systems.* Phys. Rep., 392 (2004), 321–424.

[11] P. Turchin and A.D. Taylor, *Complex dynamics in ecological time series.* Ecology, 73(1) (1992), 289–305.

[12] W. Weidlich and G. Haag, "Concepts and models of a quantitative sociology: the dynamics of interacting population". Volume 14. Springer Science & Buisiness Media, 2012.

Quantized option pricing in Mathematical Finance

Lucio Fiorin (*)

1 Introduction on Financial Markets

What is Mathematical Finance? We take the definition from Wikipedia:

Mathematical finance, also known as quantitative finance, is a field of applied mathematics, concerned with financial markets. Generally, mathematical finance will derive and extend the mathematical or numerical models without necessarily establishing a link to financial theory, taking observed market prices as input. Mathematical consistency is required, not compatibility with economic theory.

The main issues in Mathematical Finance are the following:

- How and why does the price of an asset change?
- Is it possible to make money using mathematical tools?
- Can we understand something about Mathematical Finance without knowing anything about Finance (and / or Mathematics)?

A perfect example of how markets work, and how Mathematics can deal with financial markets, is the movie Trading Places. Trading Places is a 1983 American comedy film, broadcast on the Italian TV every Christmas eve. It tells the story of an upper-class commodities broker and a homeless street hustler, and the end of the movie explains how financial markets work, or at least in a toy model. The scene tells the story of how the Duke brothers, the bad people in the movie, and Valentine and Winthorpe, the main characters, behave on the trading floor.

- The Duke brothers buy information on the harvest of oranges (insider trading: it is a crime!)

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: `fiorin@math.unipd.it` . Seminar held on March 15th, 2017.

- Valentine and Winthorpe replace the report with fake information. In the fake report the harvest is bad, while in reality it is good.
- Trading floor scene: Market of Orange Juice futures. The opening is before the announcement of the minister.
- The Duke brothers buy at lowest possible price (the opening) and sell at the highest possible price (after the annuncement): they buy something they do not own and they sell something the just bought.
- Valentine and Winthorpe sell at the highest possible price (before the announcement) and buy at the lowest possible price (after the announcement): they sell something they do not own and then buy what they just sold.
- At the end nobody has bought any Orange Juice, they just made a profit (Valentine and Winthorpe) or a loss (The Duke brothers).
- **They are both committing a crime!**

We can learn different things from this scene of the movie. The first one is the presence in the market of **Futures** (3 months in the movie): the buyer and the seller agree now on the price to pay in 3 months for a given quantity of OJ. Then we see the dynamic of the formation of the price: when people buy, the price raises, while when people sell, the price falls. Before the announcement the price is fluctuating, due to the expectations of the traders, a phenomenon called high volatility. After the announcement the price is falling, since the harvest has been good, so there will be a lot of OJ in the market. Futures are used in the markets for different reasons, but mainly for speculation and hedging purposes. Speculation is exactly the one seen in the movie, as people are betting on the price in the future in order to have a profit. Hedging is a technique used to reduce the risk of price changes in the future. For example, if we are a juice producer company, we prefer to fix now the price for Orange Juice in 3 months, instead of buying/selling it at the spot price in 3 months.
In terms of risk management, futures are not really useful in covering risk: we can face great losses (as well as great profits). This is why there has been the introduction in financial markets of European Put/Call option. At time $t = 0$ I pay an amount of money so I have the opportunity to buy at time $t = 3$ months OJ at a fixed price. We give here an example of a European Call option:

- The OJ price now is 100. We buy a Call option (paying an amount C) so we have the right, but not the obligation, to buy OJ in 3 months at 100.
- If at maturity the price of OJ is below 100, we do not exercise the option: we lose C.
- If at maturity the price of OJ is above 100, we exercise, so we buy OJ at 100.
- The profit is less than with Futures (we subtract the cost of C) but the loss is bounded by C.

2 Mathematical modeling

In this section we introduce some of the mathematical tools that are needed to understand the basic aspects of Mathematical Finance. The price of a stock has (usually) the following behavior:

Figure 1. Stock price of the FTSE MIB from June 2015 to March 2017. Data from Yahoo! Finance.

It is impossible to model something like this with deterministic functions, so we need to use stochastic processes!
We assume then that the increment of the price process from time t to time $t + dt$ is not given only by the position in t but also by a source of randomness. We can then write what is called a Stochastic Differential Equation (SDE):

$$dX_t = a(X_t, t)dt + b(X_t, t)dW_t, \quad X_0 = x, \tag{1}$$

where $dW_t \sim \mathcal{N}(0, dt)$, i.e. dW_t is a Gaussian random variable with zero mean and variance equal to dt. This is a generalization of the most celebrated Brownian motion, which can be obtained taking the coefficients $a = 0$ and $b = 1$.
Usual assumptions in Mathematical Finance are that the price process S_t follows a Geometric Brownian motion: we consider $dS_t = \mu S_t dt + \sigma S_t dW_t$, where the drift term μ and the volatility component σ are considered to be constant. This is called the Black - Scholes dynamic for the asset price.
Another important tool is the Ito's Lemma, which permits to write the dynamics of a function of a stochastic process. It is a stochastic generalization of the chain rule for derivation.
Let X_t satisfy the SDE $dX_t = a(X_t, t)dt + b(X_t, t)dW_t$ and let $f \in C^{2,1}$. Then

$$df(X_t, t) = f_t(X_t, t)dt + f_x(X_t, t)dX_t + \frac{1}{2} f_{xx}(X_t, t)(b(X_t, t))^2 dt. \tag{2}$$

Using this formula, it is possible after trivial computations to write the log price process of an asset, that we will call X_t:

$$dX_t = d\log(S_t) = \left(\mu - \frac{1}{2}\sigma^2\right) dt + \sigma dW_t, \quad X_0 = x = \log(x). \tag{3}$$

This characterization of the log price is very useful because it gives us the possibility to write the solution of the SDE in a closed form, and to analyze separately the deterministic and the stochastic part. In fact we have that

$$\log(S_T) = X_T \sim \mathcal{N}\left(\left(\mu - \frac{1}{2}\sigma^2\right) T, \sigma^2 T\right)$$

This means that under our assumptions, the price is lognormally distributed. with a closed form expression for the mean and the variance.
The lognormal assumption is a nice feature, but empirical evidence shows that we are far from this (really nice) formulation, so both academia and financial industries have introduced more complicated models to deal with peculiar behaviors of asset prices and of their derivatives.
Another very important tool is the theory of self financing portfolios. We create a portfolio V_t with (a convex combination of) two assets: our risky asset S_t and a risk-free asset B_t, i.e. $B_t = B_0 e^{rt}$. The risk-free asset is a mathematical idealization. In fact there is no possibility to have zero risk, but the usual assumption is to consider American Treasury bonds as the safest in the world. The portfolio is composed as

$$V_t = \delta S_t + (1 - \delta) B_t.$$

We introduce the concept of self financing: the purchase of a new asset must be financed by the sale of an old one. Assuming that we can buy and sell at every instant of time, we get the (stochastic) differential equation for the portfolio:

$$dV_t = \delta dS_t + (1 - \delta) r B_t dt.$$

If we collect the two conditions we have the SDE satisfied by a self financing portfolio with one risky asset:

$$dV_t = \delta dS_t + (V_t - \delta S_t)\, r dt. \tag{4}$$

Self financing portfolios have been created to mimic the price of a Call option, in order to give correct prices. This is due to the non Arbitrage theory, which tells then if two portfolios (the self financing and the Call option) have the same value at maturity, then the must have the same value at the time when the contract is signed. Since a Call option is a function of the price of the asset, it can be considered as a portfolio, but self financing portfolios are easier to price.
In fact, using together the Ito's lemma and the self financing portfolio dynamic, we get that, considering $V_t = f(S_t, t)$:

- Ito's Lemma $\Rightarrow dV_t = f_t dt + f_s dS_t + \frac{1}{2} f_{ss} \sigma^2 S_t dt$.

- Self financing portfolio $\Rightarrow dV_t = \delta dS_t + (V_t - \delta S_t)\, rdt$.

Putting them together we have that $\delta = f_s$ and

$$f_t + \frac{1}{2} f_{ss}\sigma^2 s^2 + f_s rs = rf, \tag{5}$$

with final condition: $f(s,T) = \max(s-K, 0)$, because we exercise the option only if the price is above a fixed threshold K.
This equation is the most celebrated Black Scholes equation, and gives the opportunity to compute the price of a Call option using PDE techniques. The important result is not only the way to compute the price, but also the fact that we have an explicit formula for the δ part. In fact, if we want to cover the risk of a call option, we know exactly how to construct the self financing portfolio able to mimic the behavior of the call price.

3 Option pricing and Numerical Probability

Another important tool in Mathematical Finance is the Feynman - Kac formula. This formula establishes a link between parabolic partial differential equations and stochastic processes. It offers a method of solving certain PDEs using probability approaches. Consider the PDE with terminal condition

$$f_t + \frac{1}{2} f_{ss}\sigma^2 s^2 + f_s rs = rf, \quad f(s,T) = \max(s-K, 0),$$

then we can write f as an expected value:

$$f(s,t) = \mathbb{E}\left[e^{-r(T-t)} \max(S_T - K, 0) \middle| S_t = s \right], \tag{6}$$

where the stochastic process S_t has the following dynamic:

$$dS_t = rS_t dt + \sigma S_t dW_t.$$

This result is an incredibly powerful tool, since it gives the opportunity to compute the prize of an option with two completely different approaches. Sometimes the pricing via the PDE is easier, while sometimes the expected value is easier to solve and / or compute. Remember the definition expected value $\mathbb{E}[g(X)] = \int g(z)\mathbb{P}(X \in dz)$, where $\mathbb{P}(X \in dz)$ is the density of X, which can be a random variable or a stochastic process at a given time.
The Feynman - Kac formula gives also some really interesting results. First of all the drift term in the price process must be r if we want to price using probabilistic tools. This also means that, after some really basic computations, the discounted price process, i.e. the process $e^{-rt}S_t$ is a martingale, i.e. $\mathbb{E}[e^{-rt}S_t] = S_0 \quad \forall t \geq 0$. Furthermore, the price at time t of a Call option can be seen as the discounted expectation of the final payoff at maturity.
A first attempt to tackle the issue of computing the expected value in (6) using Numerical Probability is with Monte Carlo methods. In Monte Carlo algorithms, the expectation is

approximated using an empirical mean. In the case of a Geometric Brownian motion, we need to be able to simulate a standard Gaussian random variable W. In fact, using Ito's Lemma

$$S_T = S_0 e^{(r-\frac{1}{2}\sigma^2)T+\sigma\sqrt{T}W} =: g(W),$$

then, having N simulation of W, that we call W_i,

$$\mathbb{E}\left[\max(S_T - K, 0)\right] \approx \frac{1}{N}\sum_{i=1}^{N} \max\left(g(W_i) - K, 0\right)$$

The Law of Large Numbers tells us that the empirical mean converges towards the expectation when $N \to \infty$, while the Central Limit theorem gives the rate of convergence, which is $\frac{1}{\sqrt{N}}$. An alternative to Monte Carlo methods is given by quantization techniques.

3.1 Brief overview on quantization

We first provide some more technical details on vector quantization of a random variable. Consider an $\mathbb{R}^d$-valued random variable X defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with finite r-th moment and probability distribution $\mathbb{P}_X$. Quantization can be considered as a discretization of the probability space by at most N values, providing in some sense the best approximation to the original distribution. In other words, N-quantizing the random variable X, taking infinitely many values, boils down to approximating it by a discrete random variable $\widehat{X}$ valued in a set of cardinality N, $\Gamma = \{x_1, \cdots, x_N\}$. As a consequence, in view of our application to quantitative finance, integrals of the form $\mathbb{E}\left[h(X)\right]$ (for a given Borel function $h : \mathbb{R}^d \to \mathbb{R}$) can be approximated by the finite sum below

$$\mathbb{E}\left[h(X)\right] \widetilde{=} \mathbb{E}\left[h(\widehat{X})\right] = \sum_{i=1}^{N} h(x_i)\ \mathbb{P}(\widehat{X} = x_i)$$

Clearly it still remains to clarify how to get the optimal or at least a "good" grid Γ and the associated weights $\mathbb{P}(\widehat{X} = x_i), i = 1, \ldots, N$ and to estimate the error. More rigorously, quantizing X on a given grid $\Gamma = \{x_1, \cdots, x_N\}$ consists in projecting X on the grid Γ following the closest neighbor rule. An *N-quantizer* is a Borel function $f_N : \mathbb{R}^d \to \Gamma \subset \mathbb{R}^d$ projecting X on Γ. The induced mean L^r-error (for $r > 0$) is called *L^r-mean quantization error* and is given by

$$\|X - f_N(X)\|_r = \| \min_{1 \le i \le N} |X - x_i| \|_r$$

where $||X||_r := \left[\mathbb{E}(|X|^r)\right]^{1/r}$ is the usual norm in L^r. The projection of X on Γ, $f_N(X)$, is called *the quantization of X* (in the sequel, we will alternatively use $f_N(X)$ or $\mathrm{Proj}_\Gamma(X)$ to indicate the quantization of X). As a function of the grid Γ, the L^r-mean quantization error is continuous and reaches a minimum over all the grids with size at most N. A grid $\Gamma^\star$ minimizing the L^r-mean quantization error over all the grids with size at most N is called an *L^r-optimal quantizer*.
An optimal quantizer is then associated to an optimal grid of points $\Gamma^\star$ and to an optimal

Borel partition of the space $\mathbb{R}^d$, $(C_i(\Gamma^\star))_{1\le i\le N}$, and viceversa, so that the quantizer is defined as follows

$$f_N(X) = \sum_{i=1}^{N} x_i \mathbb{1}_{C_i(\Gamma^\star)}(X)$$

where the above partition $\{C_i(\Gamma^\star)\}_{i=1,\dots,N}$, with $C_i(\Gamma^\star) \subset \{\xi \in \mathbb{R}^d : ||\xi - x_i|| = \min_{1\le j\le N} ||\xi - x_j||\}$, is called the *Voronoi partition*, or *tessellation* induced by $\Gamma^\star$. Moreover, the L^r-mean quantization error vanishes as the grid size $N \to +\infty$ and the convergence rate has been obtained in the celebrated Zador theorem (see [5]):

$$\min_{\Gamma,\ |\Gamma|=N} \|X - \mathrm{Proj}_\Gamma(X)\|_r = Q_r(\mathbb{P}_X) N^{-1/d} + o\big(N^{-1/d}\big)$$

where $Q_r(\mathbb{P}_X)$ is a nonnegative constant ($r = 2$ of course will be of particular interest, with the corresponding quadratic optimal quantizer). From a numerical point of view, finding an optimal quantizer may be a very challenging task. This motivates the introduction of sub-optimal criteria, mostly because one is typically interested in quantizations which are close to X in distribution.
An N-quantizer $\Gamma^N = \{x_1, \cdots, x_N\}$ inducing the quantization f_N of X is said to be stationary if

$$\mathbb{E}[X | f_N(X)] = f_N(X)$$

In particular, if we introduce the distortion function associated with Γ^N

$$D(\Gamma^N) := \sum_{i=1}^{N} \int_{C_i(\Gamma^N)} |z - x_i|^2 d\mathbb{P}_X(z) \tag{7}$$

then it turns out that stationary quantizers are critical points of the distortion function (that is, a stationary quantizer Γ^N satisfies $\nabla D(\Gamma^N) = 0$). Computing quadratic optimal quantizers, or L^r-optimal (or stationary) quantizers in general, together with finding the associated weights and L^r-mean quantization errors, are important issues. Several algorithms are used in practice. In the one dimensional framework, the L^r-optimal quantizers are unique up to the grid size as soon as the density of X is strictly log-concave. In this case the Newton algorithm is commonly used to carry out the L^r-optimal quantizers when closed or semi-closed formulas are available for the gradient (and the Hessian matrix). From a numerical point of view, stationary quantizers are interesting insofar they can be found through zero search recursive procedures like Newton's algorithm that can be efficiently performed.

The reason why quantization is getting more interest in both academia ad the financial industry is because of the following reason:
If f is Lipschitz continuous with Lipschitz constant $[f]_{\mathrm{Lip}}$, then

$$\big|\mathbb{E}[f(X)] - \mathbb{E}[f(\mathrm{Proj}_\Gamma(X))]\big| \le [f]_{\mathrm{Lip}} ||X - \mathrm{Proj}_\Gamma(X)||_2.$$

$||X - \mathrm{Proj}_\Gamma(X)||_2 \to 0$ with linear convergence, due to the Zador theorem.
So if we whink of X as an underlying at maturity, and f as the payoff of a derivative, we are approximating an option price with rate of convergence much better than the one obtained using Monte Carlo methods.

4 Latest developments

The case of Geometric Brownian Motion (the Black - Scholes model) is too simple to be consistent with the market, so lots of different models have been studied in the literature:

- Local volatility: consider σ as a function of S_t
- Addition of jumps: in case of shocks, the prices can not be continuous
- Stochastic volatility: σ is not constant, but a stochastic process

When we add complexity to the models, we lose their analytical tractability, and as a result the distribution of the stock price can become impossible to determine in closed form. This is an issue in Numerical Probability because the knowledge of the distribution of a process is fundamental in the computation of expected values. Also, the possibility to compute optimal quantizers for this types of models has been, and still is, an open research problem.
One possible way to tackle the problem is using a a Euler discretization of the SDE of the price process. The most simple case, the local volatility models, have been studied for the CEV model $dS_t = rS_t dt + \sigma S_t^{\alpha} dW_t$ in [6] and for the QNV model $dS_t = rS_t dt + (\sigma_1 S_t^2 + \sigma_2 S_t + \sigma_3) dW_t$ in [1]. In the case where the volatility is itself a stochastic process, there has been studies on the SABR model, which is a CEV model where σ is a Geometric Brownian Motion, see [2] and on the Heston model, which is a lognormal model where σ has mean reversion, see [4]

4.1 A new approach to quantization

We introduce here a new development that can deal with more complicated models and also with the presence of jumps. This technique has been developed in [3].
The characteristic function of a process X_t is defined as

$$\phi_{t,T}^{x}(u) := \mathbb{E}\left[e^{iuX_T} \middle| X_t = x\right].$$

In the case of a price processes, it is useful to write the characteristic function of the log process, including the dependence on the initial value of the process, which is usually known:

$$\phi_{0,T}^{x}(u) = \mathbb{E}\left[e^{iu\log(S_T)} \middle| \log(S_0) = x\right]. \tag{8}$$

The characteristic function can also be seen as a Fourier transform (omitting the dependence on the initial value):

$$\phi_{0,T}^{x}(u) = \int_{-\infty}^{+\infty} e^{iu\log(z)} \mathbb{P}(S_T \in dz).$$

Indeed, using the Fourier inversion theorem, it is possible to get the density of the price process directly from the characteristic function:

$$\mathbb{P}(S_T \in dz) = \frac{1}{\pi}\frac{1}{z}\int_0^{+\infty} Re\left(e^{-iu\log(z)}\phi_{0,T}^{x}(u)\right) du$$

In practice, it is much easier to determine in closed form the characteristic function of a price process rather than the density. In addition, the characteristic function is known in closed form for different kind of models, like

- Local volatility
- Stochastic volatility
- Jump processes

The knowledge of the density of the price process is fundamental, because we can write the distortion function (7), and we can directly compute the minima, which are the stationary quantizers. In Figure 2 and 3 we present the results of this new type of quantization for respectively, the asset price and the volatility dynamics in a stochastic volatility model. It can be seen how the more we go through time, the more quantization is able to catch the randomness of the processes.

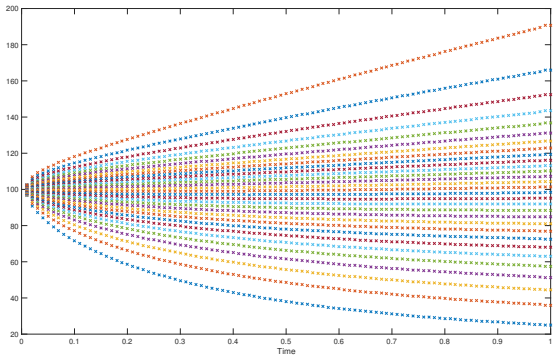


Figure 2. Quantization of the price process in the Heston model. $N = 30$.

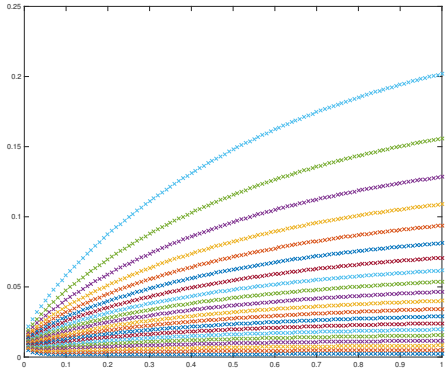


Figure 3. Quantization of the volatility process in the CIR model. $N = 20$.

References

[1] Callegaro, G., Fiorin L. and Grasselli M., *Quantized calibration in local volatility.*. Risk Magazine (2015), 62–67.

[2] Callegaro G., Fiorin L. and Grasselli M., *Pricing via Recursive Quantization in Stochastic Volatility Models*. Forthcoming in Quantitative Finance (2016).

[3] Callegaro G., Fiorin L. and Grasselli M., *Quantization meets Fourier. A New Technology for Pricing Options*. Preprint (2017).

[4] Fiorin L., Sagna A. and Pagès G., *Componentwise and Markovian product quantization of an $\mathbb{R}^d$-valued Euler diffusion process with applications*. Preprint (2017).

[5] Graf, S. and Luschgy, H., "Foundations of quantization for probability distributions". Springer, New York, 2000.

[6] Pagès, G. and Sagna, A., *Recursive marginal quantization of the Euler scheme of a diffusion process*. Preprint (2014).

An introduction to domain perturbation theory for elliptic eigenvalue problems

Francesco Ferraresso (*)

Abstract. How does the sound of a drum depend on its shape? This weak variant of the classical question "Can one hear the shape of a drum?" can be considered in the framework of domain perturbation theory for elliptic differential operators. The answer to this apparently harmless question is rather different in the case of regular perturbations and in the case of singular perturbations. We consider mainly the singular case, where the geometry of the problem is deeply interlaced with the differential structure, in particular with the boundary conditions. Finally, we give a survey of the main results in the study of the dumbbell domain perturbation for the Laplace operator and for the biharmonic operator.

1 Introduction

The relation between the eigenvalues and eigenfunctions of elliptic operators and the dynamics of vibrating objects is a classical topic (see e.g. [14], [20]), and it is nowadays understood in terms of spectral theory, (we refer to [15] for a general introduction to spectral theory with applications to PDEs). Let us introduce an example to fix ideas.
Let $u(x,t)$ be the vertical displacement of a point on the surface Ω of a drum. For small oscillations, its motion is described by the wave equation

$$\begin{cases} u_{tt}(x,t) = \Delta_x u(x,t) = \sum_{i=1}^N \partial^2_{x_i x_i} u(x,t), & \text{in } \Omega \times (0,+\infty) \\ u(x,t) = 0, & \text{on } \partial\Omega \times (0,+\infty). \end{cases}$$

which can be solved by separation of variables. Namely we claim the existence of a solution $u(x,t)$ such that $u(x,t) = \varphi(t)v(x)$ for almost all $(x,t) \in \Omega \times (0,+\infty)$. Then it must be

$$\varphi''(t)v(x) = \varphi(t)\Delta v(x)$$

which implies in turn

$$\frac{\varphi''(t)}{\varphi(t)} = \frac{\Delta v(x)}{v(x)} = -\lambda \in \mathbb{R}.$$

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: fferrare@math.unipd.it . Seminar held on March 29th, 2017.

It is easy then to deduce that

$$\varphi_n(t) = A\cos(\sqrt{\lambda_n}t) + B\sin(\sqrt{\lambda_n}t),$$

where λ_n is the n-th eigenvalue of the Helmholtz equation

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

If Ω has finite measure, then there exists a sequence of isolated eigenvalues of finite multiplicity

$$0 < \lambda_1[\Omega] \leq \lambda_2[\Omega] \leq \cdots \leq \lambda_n[\Omega] \leq \dots$$

If Ω represents a drum, then λ_n is what in music is usually called *n-th harmonic* of the instrument. Hence the eigenfunctions of the Laplace operator are exactly the principal modes of vibration of the drum corresponding to the harmonic frequencies, see Figures 1-2.

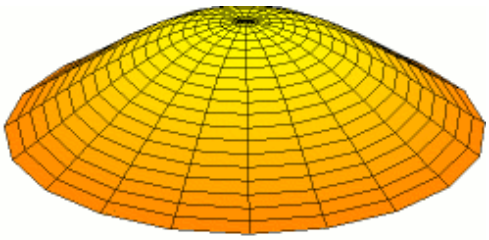


Figure 1. First mode of vibration of a circular drum.

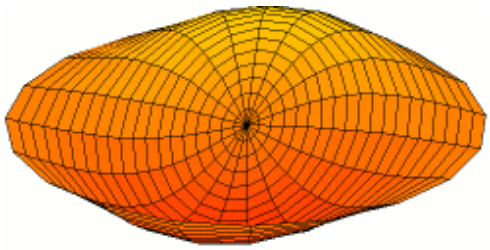


Figure 2. Second mode of vibration of a circular drum

2 Laplacian and bilaplacian.

We introduce here some notation and the main mathematical objects we will need in the sequel. Given a bounded open set $\Omega \subset \mathbb{R}^N$ we introduce the Dirichlet eigenvalue problem for the Laplace operator, which is defined by

$$\begin{cases} -\Delta u = \lambda u, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega. \end{cases} \tag{2.1}$$

and if $\partial\Omega$ is sufficiently regular (for example Lipschitz continuous) then we can define the Neumann eigenvalue problem for the Laplace operator, which is defined by

$$\begin{cases} -\Delta u = \lambda u, & \text{in } \Omega, \\ \frac{\partial u}{\partial n} = 0, & \text{on } \partial\Omega. \end{cases} \tag{2.2}$$

where we have denoted by n the unit outer normal to Ω.
From an operator-theoretical point of view, the set of eigenvalues $(\lambda_n)_n$ is the spectrum of

the Friedrichs self-adjoint extension of the operator $-\Delta$ in the Hilbert space $L^2(\Omega)$. The domain of this extension is a suitable Sobolev space on Ω. Let us recall some terminology.

Definition 1 A function $f \in L^2(\Omega)$ has a weak derivative $g \in L^2(\Omega)$ with respect to x_i if

$$\int_\Omega f \frac{\partial \varphi}{\partial x_i}\, dx = -\int_\Omega g\varphi\, dx$$

for all $\varphi \in C_c^\infty(\Omega)$.

Definition 2 The Sobolev space $H^k(\Omega)$ is the set of functions $f \in L^2(\Omega)$ such that the weak derivative $\frac{\partial^\alpha f}{\partial x^\alpha}$ exists in $L^2(\Omega)$ for all multiindexes $\alpha = (\alpha_1, \dots, \alpha_N) \in \mathbb{N}^N$ of length k. It is a Hilbert space with norm

$$\|f\|_{H^k(\Omega)} = \Big(\|f\|^2_{L^2(\Omega)} + \sum_{|\alpha|=k} \|\partial_\alpha f\|^2_{L^2(\Omega)}\Big)^{1/2}.$$

Moreover we define by $H_0^k(\Omega)$ the closure of $C_c^\infty(\Omega)$ with respect to the norm $\|\cdot\|_{H^k(\Omega)}$.

We recall that $C^\infty(\Omega) \cap H^k(\Omega)$ is dense in $H^k(\Omega)$ with respect to $\|\cdot\|_{H^k}$, whereas $C_c^\infty(\Omega)$ is dense in $H_0^k(\Omega)$ with respect to $\|\cdot\|_{H^k}$. Recall moreover that $C^\infty(\overline{\Omega})$ is dense in $H^k(\Omega)$ whenever the boundary of Ω is sufficiently regular.

In order to comprehend the relation between the eigenvalue problems (2.1), (2.2) and the Sobolev spaces we need to switch to the weak formulation. For simplicity let us assume Ω to be a bounded open set of $\mathbb{R}^N$ of class C^2. Recall that by standard regularity arguments the eigenfunctions solving problems (2.1), (2.2) are at least $H^2(\Omega)$, see e.g. [18]. We consider first the case of Neumann boundary conditions. We multiply equation (2.2) by a test function $\varphi \in C^\infty(\overline{\Omega})$ and we integrate over Ω in order to obtain

$$-\int_\Omega \Delta u\varphi = \int_\Omega \nabla u \nabla \varphi - \int_{\partial\Omega} \varphi \frac{\partial u}{\partial n} = \lambda \int_\Omega u\varphi$$

for all $\varphi \in C^\infty(\Omega)$. Then the boundary integral vanishes because of the boundary conditions. Hence we deduce that

$$\int_\Omega \nabla u \nabla \varphi = \lambda \int_\Omega u\varphi$$

for all $\varphi \in C^\infty(\Omega)$, hence for all $\varphi \in H^1(\Omega)$.

In the case of Dirichlet boundary conditions instead we note that it is sufficient to take $\varphi \in C_c^\infty(\Omega)$ to obtain that

$$\int_\Omega \nabla u \nabla \varphi = \lambda \int_\Omega u\varphi \tag{2.3}$$

for all $\varphi \in C_c^\infty(\Omega)$, hence for all $\varphi \in H_0^1(\Omega)$.

Assume now that u is a classical function. Then (2.3) and (2.1) turn out to be formulations equivalent to (2.1) and (2.2) respectively, since if

$$\int_\Omega (-\Delta u - \lambda u)\varphi dx = 0$$

for all $\varphi \in C^\infty(\Omega)$, then by the Fundamental Lemma of Calculus of Variations $-\Delta u - \lambda u = 0$.

We now consider the biharmonic operator Δ^2 which is the composition of $-\Delta$ with itself, or more explicitly

$$\Delta^2 u = \sum_{i,j=1}^{N} \frac{\partial^4 u}{\partial x_i^2 \partial x_j^2} = \sum_{i=1}^{N} \frac{\partial^4 u}{\partial x_i^4} + 2\sum_{i<j} \frac{\partial^4 u}{\partial x_i^2 \partial x_j^2}$$

In applications, the biharmonic operator usually models the vibration of a three-dimensional elastic plate whose thickness is very small if compared with the other dimensions. We remark here that in the last years polyharmonic operators have received attention from diverse authors and from different points of view. We refer to [17] as a general reference for polyharmonic theory. Let us mention [9], [11], [12], [13], [21], [6] where the authors have considered differential problems for higher order elliptic operators. Here we consider the eigenvalue problem for Δ^2 with either Dirichlet boundary conditions defined by

$$\begin{cases} \Delta^2 u = \lambda u, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \\ \frac{\partial u}{\partial n} = 0, & \text{on } \partial\Omega, \end{cases} \tag{2.4}$$

or with Neumann boundary conditions, given by

$$\begin{cases} \Delta^2 u = \lambda u, & \text{in } \Omega, \\ \sigma\Delta u + (1-\sigma)\frac{\partial^2 u}{\partial n^2} = 0, & \text{on } \partial\Omega, \\ (1-\sigma)\mathrm{div}_{\partial\Omega}(D^2 u \cdot n)_{\partial\Omega} + \frac{\partial \Delta u}{\partial n} = 0, & \text{on } \partial\Omega, \end{cases} \tag{2.5}$$

where σ is the Poisson coefficient of the material ($-\frac{1}{N-1} < \sigma < 1$), $D^2 u : D^2\varphi$ is the Frobenius product of the Hessian matrixes, $\mathrm{div}_{\partial\Omega}$ is the tangential divergence operator, and $(v)_{\partial\Omega}$ denotes the projection of the function v on the tangent space. We proceed to find the weak formulation of these problems. We first consider (2.4). Let $\varphi \in C_c^\infty(\Omega)$. Multiply equation (2.4) by φ, integrate over Ω and then integrate twice by parts in order to obtain

$$\int_\Omega \Delta^2 u\, \varphi = \int_\Omega \Delta u \Delta\varphi + \int_{\partial\Omega} \frac{\partial \Delta u}{\partial n}\varphi - \int_{\partial\Omega} \Delta u \frac{\partial \varphi}{\partial n} = \lambda \int_\Omega u^2$$

and the boundary integrals vanish since φ has compact support. We deduce the weak formulation:

$$\int_\Omega \Delta u \Delta\varphi = \lambda \int_\Omega u\varphi$$

for all $\varphi \in C_c^\infty(\Omega)$, hence for all $\varphi \in H_0^2(\Omega)$.

We now consider (2.5). In order to find the weak formulation we need to use the following biharmonic Green formula (see [9, Lemma 8.56]):

$$\int_\Omega \Delta^2 u\,\varphi = \int_\Omega D^2u : D^2\varphi - \int_{\partial\Omega} \frac{\partial^2 u}{\partial n^2}\frac{\partial \varphi}{\partial n} + \int_{\partial\Omega} \left(\mathrm{div}_{\partial\Omega}((D^2u\cdot n))_{\partial\Omega} + \frac{\partial \Delta u}{\partial n}\right)\varphi.$$

Let $\varphi \in C^\infty(\Omega)$. Write

$$\Delta^2 u\varphi = (1-\sigma)\Delta^2 u\varphi + \sigma\Delta^2 u\varphi = \lambda u\varphi$$

and integrate over Ω. Then

$$\begin{aligned}(1-\sigma)\int_\Omega \Delta^2 u\varphi &= (1-\sigma)\int_\Omega D^2u : D^2\varphi - (1-\sigma)\int_{\partial\Omega}\frac{\partial^2 u}{\partial n^2}\frac{\partial\varphi}{\partial n}\\ &+ (1-\sigma)\int_{\partial\Omega}\left(\mathrm{div}_{\partial\Omega}((D^2u\cdot n))_{\partial\Omega} + \frac{\partial\Delta u}{\partial n}\right)\varphi = (1-\sigma)\lambda\int_\Omega u\varphi\end{aligned} \tag{2.6}$$

and

$$\sigma\int_\Omega \Delta^2 u\varphi = \sigma\int_\Omega \Delta u\Delta\varphi - \sigma\int_{\partial\Omega}\frac{\partial\Delta u}{\partial n}\varphi + \sigma\int_{\partial\Omega}\Delta u\frac{\partial\varphi}{\partial n} = \sigma\lambda\int_\Omega u\varphi \tag{2.7}$$

and by summing up (2.6) and (2.7) we find that

$$\begin{aligned}\int_\Omega (1-\sigma)D^2u : D^2\varphi + \sigma\Delta u\Delta\varphi - \int_{\partial\Omega}\left((1-\sigma)\frac{\partial^2 u}{\partial n^2} + \sigma\Delta u\right)\frac{\partial\varphi}{\partial n}\\ + (1-\sigma)\int_{\partial\Omega}\left(\mathrm{div}_{\partial\Omega}((D^2u\cdot n))_{\partial\Omega} + \frac{\partial\Delta u}{\partial n}\right)\varphi + \sigma\int_{\partial\Omega}\frac{\partial\Delta u}{\partial n}\varphi = \lambda\int_\Omega u\varphi\end{aligned} \tag{2.8}$$

Keeping into account the boundary conditions in (2.5) we see that the boundary integrals vanish. Hence,

$$\int_\Omega (1-\sigma)D^2u : D^2\varphi + \sigma\Delta u\Delta\varphi = \lambda\int_\Omega u\varphi$$

for all $\varphi \in C^\infty(\overline{\Omega})$, hence for all $\varphi \in H^2(\Omega)$.

3 Domain perturbation theory

With domain perturbation theory we refer to the analysis of the dependence of the solutions and of the spectrum of a given elliptic differential problem on the properties of the underlying domain Ω. More precisely, let Ω be a bounded open set in $\mathbb{R}^N$. We consider elliptic operators of the type

$$Hu = (-1)^m \sum_{|\alpha|=|\beta|=m} D^\alpha\left(A_{\alpha\beta}(x)D^\beta u\right), \quad x\in\Omega, \tag{3.1}$$

subject to homogeneous boundary conditions. Under suitable assumptions, the spectrum is discrete

$$\lambda_1[\Omega] \le \lambda_2[\Omega] \le \cdots \le \lambda_n[\Omega] \le \dots$$

The main question in domain perturbation theory is to understand under which hypothesis the maps

$$\Omega \mapsto \lambda_n[\Omega], \qquad \Omega \to u_n[\Omega]$$

are continuous (or differentiable). Of course the answer to this question strongly relies on the choice of the topology on the class of open sets of $\mathbb{R}^N$. Let us remark that this is closely related to the stability (with respect to domain perturbations) of the Poisson problem

$$Hu = f, \quad \text{in } \Omega,$$

for given $f \in L^2(\Omega)$. In applications these stability problems are related to the following question:

Do drums/plates with "similar" shapes produce "similar" sounds?

In general it is known that the answer to this question is negative, since there exists examples of small perturbations (for example, in the sense of Lebesgue measure) of a given smooth domain producing strong instability (see Figure 3).

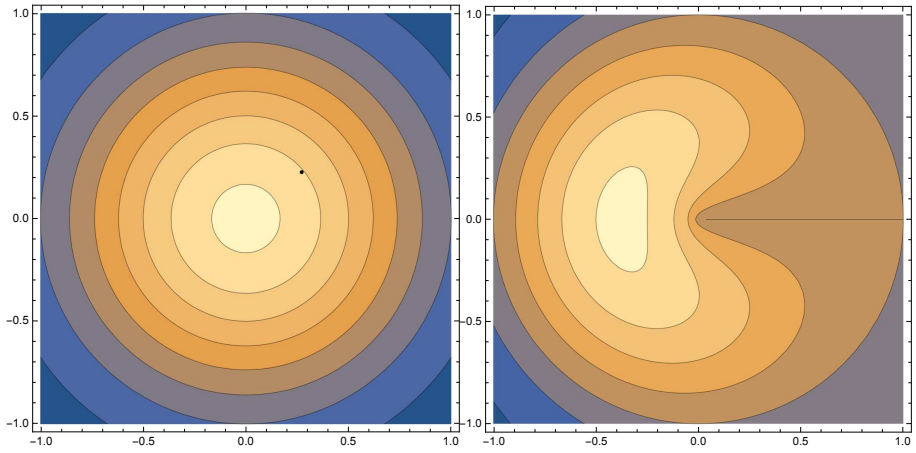


Figure 3. Dramatic change in the shape of the first eigenfunction caused by a slit in the drum.

Definition 3 We say that a domain perturbation $\Omega \mapsto \Omega_\varepsilon$ is *spectrally stable* if

(i) $\lambda_n[\Omega_\varepsilon] \to \lambda_n[\Omega]$ for all $n \in \mathbb{N}$

(ii) The spectral projections $P_a^{\Omega_\varepsilon}$ converge to P_a^{Ω} in L^2, i.e., for fixed $a \in \mathbb{R}^+ \setminus \{\lambda_j[\Omega]\}_{j=0}^{\infty}$, $\lambda_n[\Omega] < a < \lambda_{n+1}[\Omega]$ we define the projections $P_a^{\Omega_\varepsilon}$ from $L^2(\mathbb{R}^N)$ into $L^2(\Omega_\varepsilon)$ by

$$P_a^{\Omega_\varepsilon}(\psi) = \sum_{i=1}^{n} (u_i[\Omega_\varepsilon], \psi)_{L^2(\Omega_\varepsilon)} u_i[\Omega_\varepsilon]$$

and we ask that

$$\sup\{\|P_a^{\Omega_\varepsilon}(\psi) - P_a^{\Omega}(\psi)\|_{L^2(\Omega)} + \|P_a^{\Omega_\varepsilon}(\psi)\|_{L^2(\Omega_\varepsilon \setminus \overline{\Omega})} : \psi \in L^2(\mathbb{R}^N), \|\psi\|_{L^2(\mathbb{R}^N)} = 1\} \to 0,$$

as $\varepsilon \to 0$.

We address now the problem of characterizing the domain perturbations that are spectrally stable. For example, we may think about perturbations of the following types:

(i) *Exterior perturbations:* $\Omega \subset \Omega_\varepsilon$ for all $\varepsilon > 0$ and $|\Omega_\varepsilon \setminus \Omega| \to 0$;

(ii) *Interior perturbations:* $\Omega_\varepsilon \subset \Omega$ for all $\varepsilon > 0$ and $|\Omega \setminus \Omega_\varepsilon| \to 0$;

(iii) $\Omega_\varepsilon \to \Omega$ in the *complementary Hausdorff topology*, i.e.,

$$\max\{ \sup_{x\in\Omega_\varepsilon^C} \operatorname{dist}(x,\Omega^C);\ \sup_{y\in\Omega^C} \operatorname{dist}(y,\Omega_\varepsilon^C)\} \to 0)$$

(note that this condition is equivalent to the vanishing as $\varepsilon \to 0$ of the L^∞-distance between $d_{\partial\Omega}$ and $d_{\partial\Omega_\varepsilon}$, where d is the distance function);

(iv) Ω_ε is diffeomorphic to Ω via Φ_ε and $\|\Phi_\varepsilon - \mathbb{I}\|_{C^1(\mathbb{R}^N)} \to 0$

We remark here that in general perturbations satisfying condition (i) (or (ii) or even (iii)) may be quite irregular. For example, if we consider domains Ω_ε, Ω which are locally the hypograph of smooth functions g_ε, g, then it may happen that $\|g_\varepsilon - g\|_{L^\infty}$ tends to zero as $\varepsilon \to 0$, but $\|g_\varepsilon - g\|_{C^1}$ is not bounded as $\varepsilon \to 0$ (see Figure 4).

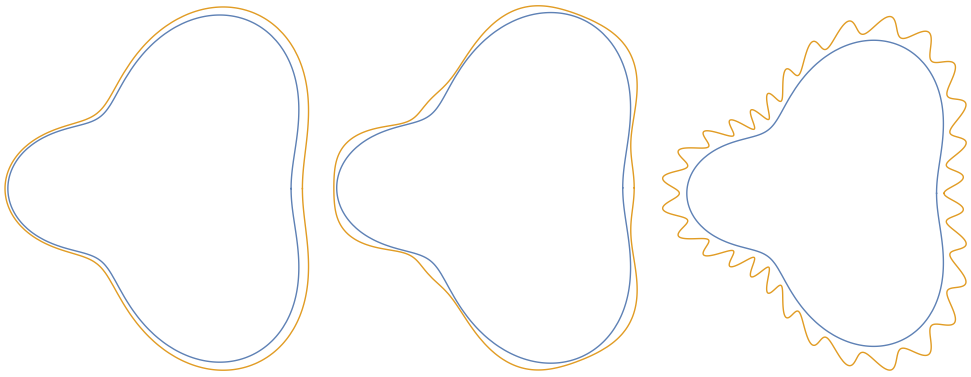


Figure 4. Examples of exterior perturbations.

3.1 Regular perturbations

We show here that the perturbations of type (iv) above are spectrally stable. Let us suppose $\Omega \subset \mathbb{R}^N$ is a fixed smooth bounded domain. Let $(\Phi_\varepsilon)_{\varepsilon>0}$ be diffeomorphisms of class C^m mapping Ω to $\Omega_\varepsilon := \Phi_\varepsilon(\Omega)$ such that $\|\mathbb{I} - \Phi_\varepsilon\|_{C^m(\mathbb{R}^N,\mathbb{R}^N)} \to 0$ as $\varepsilon \to 0$. Let H_Ω^{BC} be the differential operator H_Ω defined in (3.1) subject to some homogeneous boundary conditions. Then it is classical to prove the following

Theorem 1 *The perturbation $\Omega \mapsto \Omega_\varepsilon$ is spectrally stable. In particular $\lambda_n[\Omega_\varepsilon]$ converges to $\lambda_n[\Omega]$ as $\varepsilon \to 0$, for any $n \in \mathbb{N}$.*

Note that the spectral stability of the perturbation does not depend on the choice of boundary conditions.

3.2 Less regular perturbations

We consider now less restrictive domain perturbations, which may give singular spectral behavior in the limit. Indeed, in general Ω_ε and Ω may be non-diffeomorphic or the diffeomorphism Φ_ε mapping Ω to Ω_ε may fail to satisfy the condition

$$\|\mathbb{I} - \Phi_\varepsilon\|_{C^1(\mathbb{R}^N,\mathbb{R}^N)} \to 0, \qquad \text{as } \varepsilon \to 0.$$

Another possible problem in low regularity domain perturbations is given by the following example. Consider a smooth domain Ω and a monotonic decreasing exterior perturbation $(\Omega_\varepsilon)_{\varepsilon>0}$ (i.e., $\Omega \subset \Omega_\varepsilon \subset \Omega_{\varepsilon'}$ for all $\varepsilon < \varepsilon'$). Then the candidate "limiting domain" $\Omega_0 := \bigcap_{\varepsilon>0} \Omega_\varepsilon$ in general is not an open set, in particular it does not coincide with Ω, and it may be different from $\overline{\Omega}$ as well.
Finally, in all these cases *oscillation/concentrations issues* may appear: roughly speaking, the eigenfunctions $u_n[\Omega_\varepsilon]$ may converge weakly to $u_n[\Omega]$ but $\|u_n[\Omega_\varepsilon] - u_n[\Omega]\| \nrightarrow 0$ since either $\|u_n[\Omega_\varepsilon] - u_n[\Omega]\|$ is oscillating wildly around zero or $\|u_n[\Omega_\varepsilon]\|_{L^2(R_\varepsilon)} \to 1$, where $R_\varepsilon \subset \Omega_\varepsilon$ such that $|R_\varepsilon| \to 0$ (we say in this case that the L^2-mass of the eigenfunction u_n is *concentrating* on a lower dimensional manifold).
Importantly, for less regular perturbations spectral stability strongly depends on boundary conditions. It is widely known that Dirichlet boundary conditions allow a wider choice of stable perturbations.

Example: the Dirichlet Laplacian.
Consider here the problem of identifying the most general perturbations Ω_ε of Ω such that the spectrum associated with (2.1) is continuous with respect to the perturbation $\Omega \mapsto \Omega_\varepsilon$. This problem has been widely investigated in the literature (see e.g. [10]), giving the following results:

- Let $\Omega_{\varepsilon'} \subset \Omega_\varepsilon \subset \Omega$ for all $0 < \varepsilon < \varepsilon'$ and $\bigcup_{\varepsilon>0} \Omega_\varepsilon = \Omega$. Then $\Omega \mapsto \Omega_\varepsilon$ is a spectrally stable perturbation.

- Assume that the perturbation Ω_ε is made of monotonic decreasing sequences of sets "compact converging" to Ω, i.e., for every compact $K \subset \Omega \cup \overline{\Omega}^C$ there exists $\varepsilon_K > 0$ such that for all $\varepsilon < \varepsilon_K$, $K \subset \Omega_\varepsilon \cup \overline{\Omega_\varepsilon}^C$ (this is called *Keyldish's convergence*). Then if Ω is sufficiently regular (roughly speaking, it does not have cracks), then $\Omega \mapsto \Omega_\varepsilon$ is a spectrally stable perturbation.

- Assume that Ω_ε is converging in the Hausdorff complementary topology to Ω. In general, this is not enough to assure the spectral stability. However, if we impose a uniform geometric constraint on Ω_ε and Ω (for example, all these domains have Lipschitz boundaries, or satisfy a uniform exterior cone condition) then the perturbation is spectrally stable. In $\mathbb{R}^2$ an important result by Šverák (see [22]) states that the uniform geometric constraint can be replaced by a topological assumption on Ω_ε and Ω: namely Ω_ε^C must have an uniform bound on the number of connected components.

Note that convergence in measure of Ω_ε to Ω, $|\Omega_\varepsilon \,\Delta\, \Omega| \to 0$ is not enough to assure the spectral stability.

Non-Dirichlet boundary conditions. Generally speaking, non-Dirichlet boundary conditions may have a much more unstable behaviour with respect to boundary perturbations. Indeed, since we have non-zero boundary condition, the eigenfunctions are allowed to have quite wild behaviour near the boundary, resulting in oscillation/concentration issues combined with regularity issues. Moreover, there are examples in the literature of quite regular perturbations giving singular spectral behavior. The classical example - which can be found in [14] - is the following:

Example. Consider the eigenvalue problem for the Neumann Laplacian on Ω_ε (see Figure 5 below)

$$\begin{cases} -\Delta u_\varepsilon = \lambda(\Omega_\varepsilon) u_\varepsilon, & \text{in } \Omega_\varepsilon \\ \frac{\partial u_\varepsilon}{\partial n} = 0, & \text{on } \partial\Omega_\varepsilon. \end{cases}$$

where $\Omega \subset \mathbb{R}^2$ is the unit square $[-1/2, 1/2]^2$, and Ω_ε is as in the picture. Here η is much smaller than ϵ, take for example $\eta = \epsilon^4$. Since we are considering a Neumann problem, it is known that $\lambda_1(\Omega_\varepsilon) = \lambda_1(\Omega) = 0$ and the corresponding eigenfunctions are constant. Moreover by explicit calculations one can prove that $\lambda_2(\Omega) = \pi^2 > 0$. However, $\lambda_2(\Omega_\varepsilon) \to 0$ as $\varepsilon \to 0$.

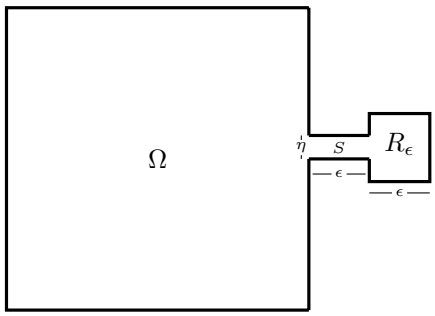


Figure 5. The "mushroom" perturbation Ω_ε.

To show that $\lambda_2(\Omega_\varepsilon) \to 0$ one can argue as follows. Recall that

$$\lambda_2(\Omega_\varepsilon) \leq \frac{\int_{\Omega_\varepsilon} |\nabla u|^2}{\int_{\Omega_\varepsilon} |u|^2}$$

for all $u \in H^1(\Omega_\varepsilon)$, $u \neq 0$, such that $\int_{\Omega_\varepsilon} u dx = 0$. Let $\varphi \in H^1(\Omega_\varepsilon)$ be defined by

$$\varphi = \begin{cases} -1/\epsilon, & \text{in } R_\varepsilon \\ c(\varepsilon), & \text{in } \Omega \\ \text{linear}, & \text{in } S \end{cases}$$

and we choose $c(\varepsilon)$ in such a way that $\int_{\Omega_\varepsilon} \varphi = 0$. Then

$$\int_S |\nabla\varphi|^2 dx \leq C\eta/\varepsilon^3, \quad \text{but} \quad \int_{\Omega_\varepsilon} |\varphi|^2 dx \to 1.$$

3.3 Dumbbell-type perturbations.

The *dumbbell domains* $\Omega_\varepsilon \subset \mathbb{R}^2$ are perturbations of a fixed bounded open set Ω with two (or more) connected components. More precisely, let $\Omega := \Omega_L \cup \Omega_R$, where Ω_L and Ω_R are the two connected components of Ω, and let the dumbbell Ω_ε be defined by $\Omega_\varepsilon := \Omega \cup R_\varepsilon$, where R_ε is a thin channel connecting Ω_L to Ω_R defined by

$$R_\varepsilon = \{(x, y) \in \mathbb{R}^2 : 0 < x < 1, 0 < y < \varepsilon g(x)\},$$

where g is a positive C^2 function. A general dumbbell domain is as in the following picture:

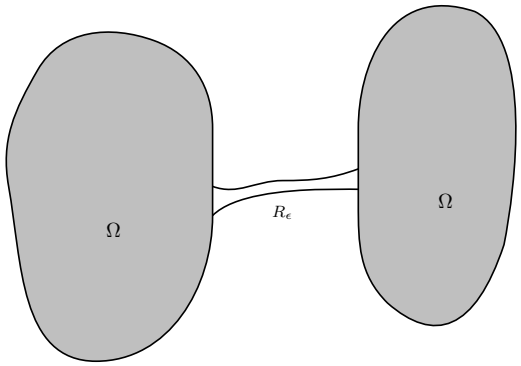


Figure 6. General dumbbell domain Ω_ε.

On the dumbbell $\Omega_\varepsilon \subset \mathbb{R}^2$ we consider the problem

$$\begin{cases} -\Delta u + u = \lambda_n(\Omega_\varepsilon)\, u, & \text{in } \Omega_\varepsilon, \\ \frac{\partial u}{\partial n} = 0, & \text{on } \partial\Omega_\varepsilon, \end{cases}$$

where $u \in H^1(\Omega_\varepsilon)$. The weak formulation for this problem is: *find* $u \in H^1(\Omega_\varepsilon)$ *such that for all* $\psi \in H^1(\Omega_\varepsilon)$ *the following equality holds true:*

$$\int_{\Omega_\varepsilon} \nabla u \nabla\psi + u\psi dx = \lambda(\Omega_\varepsilon) \int_{\Omega_\varepsilon} u\psi\, dx$$

Note that R_ε is collapsing to a lower dimensional manifold as $\varepsilon \to 0$ (R_ε is a *thin domain*).

This problem was first considered by Jimbo (see e.g. [19]) and then by Arrieta and collaborators (see [1], [2], [3], [5], [8]). An important result contained in [4] relates the spectral instability of the Neumann Laplacian to the concentration phenomenon, that we are now going to define.

Definition 4 We say that a sequence of functions $u_\varepsilon \in H^1(\Omega_\varepsilon)$ such that $\|u_\varepsilon\|_{L^2(\Omega_\varepsilon)} = c > 0$ for all $\varepsilon > 0$ is *concentrating* in $R_\varepsilon \subset \Omega_\varepsilon$ if there exists a constant C not depending on ε such that

$$\|u_\varepsilon\|_{H^1(\Omega_\varepsilon)} \leq C, \qquad \|u_\varepsilon\|_{L^2(R_\varepsilon)} \to c$$

as $\varepsilon \to 0$.

The existence of a sequence of functions $(u_\varepsilon)_\varepsilon$ as in Definition 4 turns out to be equivalent to the uniform boundedness of the sequence $(\tau_\varepsilon)_\varepsilon$ defined by

$$\tau_\varepsilon := \inf_{\substack{\phi_\varepsilon \in H^1(\Omega_\varepsilon) \\ \phi_\varepsilon = 0 \text{ in } \Omega}} \frac{\int_{\Omega_\varepsilon} |\nabla \phi_\varepsilon|^2 dx}{\|\phi_\varepsilon\|^2_{L^2(R_\varepsilon)}}. \tag{3.2}$$

By using the characterization of the spectral instability via τ_ε as defined in (3.2) it is easy to show that the continuity of the eigenvalues and of the eigenfunctions with respect to ε cannot hold in a wide class of dumbbell perturbations. Indeed, under suitable assumptions on the shape of R_ε it is possible to prove that there will be a contribution to the limit spectrum from the thin channel R_ε.
The suitable assumptions on R_ε are rather technical and we do not introduce them here. We only mention that they involve the eigenvalues of some auxiliary differential problems in small rectangles around the junctions of R_ε to Ω, see the Definition of (H)-Condition in [8], [3] and in [7] for more general operators. An easier condition on R_ε which implies the (H)-Condition is the following *monotonicity property*

(MP): there exists $1 >> \delta > 0$ *(not depending on* ε*) such that for any* $x \in (0, \delta)$, $g(x)$ *is decreasing, and for any* $x \in (1-\delta, 1)$, $g(x)$ *is increasing.*

Note that we are not imposing geometrical condition of the shape of the channel R_ε far from the junction, since we are only interested in the behaviour of the channel near the junctions (the red parts of the channel in Figure 7).

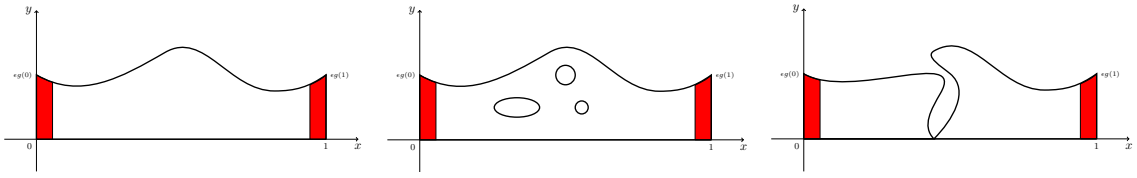


Figure 7. Examples of channels R_ε satisfying the (H)-Condition.

Under this assumption on the junctions one can prove that τ_ε defined in (3.2) is converging to τ as $\varepsilon \to 0$, where τ is the first eigenvalue of

$$\begin{cases} -\frac{1}{g}(g v_x)_x + v = \tau v, & \text{in } (0,1), \\ v(0) = v(1) = 0. \end{cases}$$

Remark 1 The boundedness of the sequence τ_ε is rather unusual in thin domain problems since whenever a Poincar inequality holds in R_ε one can prove that $\tau_\varepsilon \to \infty$ (the first Dirichlet eigenvalue on a thin domain always diverges, see e.g. [16]).

Denote now by $(\lambda_n(\Omega_\varepsilon))_n$ the eigenvalues of

$$\begin{cases} -\Delta u + u = \lambda_n(\Omega_\varepsilon)\, u, & \text{in } \Omega_\varepsilon, \\ \frac{\partial u}{\partial n} = 0, & \text{on } \partial\Omega_\varepsilon. \end{cases} \tag{3.3}$$

and by $(\omega_i)_i$ and (τ_i) the eigenvalues of

$$\begin{cases} -\Delta w + w = \omega\, w, & \text{in } \Omega, \\ \frac{\partial w}{\partial n} = 0, & \text{on } \partial\Omega. \end{cases} \qquad \begin{cases} -\frac{1}{g}(g v_x)_x + v = \tau v, & \text{in } (0,1), \\ v(0) = v(1) = 0. \end{cases}$$

We order them in a unique sequence $(\lambda_i)_i = (\omega_k)_k \cup (\tau_j)_j$, where it is understood that each eigenvalue is repeated according to its multiplicity, keeping in account possible 'resonance cases' when $\omega_k = \tau_j$ for some k, j. Then it is possible to prove that

$$\lambda_n(\Omega_\varepsilon) \to \lambda_n \qquad \text{for all } n \in \mathbb{N}$$

and moreover the spectral projections $P_a^{\Omega_\varepsilon}$ converge to P_a^{Ω} in H^1, i.e., if the projections $P_a^{\Omega_\varepsilon}$ from $L^2(\mathbb{R}^N)$ into $H^1(\Omega_\varepsilon)$ are defined by

$$P_a^{\Omega_\varepsilon}(\psi) = \sum_{i=1}^{n} (u_i[\Omega_\varepsilon], \psi)_{L^2(\Omega_\varepsilon)} u_i[\Omega_\varepsilon],$$

then

$$\sup\{\|P_a^{\Omega_\varepsilon}(\psi) - P_a^{\Omega}(\psi)\|_{H^1(\Omega)} + \|P_a^{\Omega_\varepsilon}(\psi)\|_{H^1(\Omega_\varepsilon \setminus \overline{\Omega})} : \psi \in L^2(\mathbb{R}^N), \|\psi\|_{L^2(\mathbb{R}^N)} = 1\} \to 0,$$

as $\varepsilon \to 0$. Hence the spectrum of problem (3.3) has a dichotomic behavior in the limit, as shown in Figure 8.

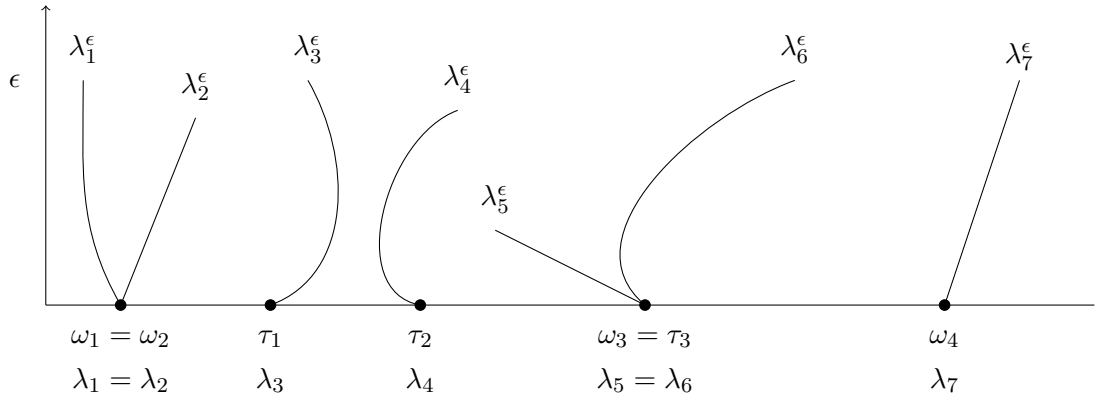


Figure 8. Convergence of the eigenvalues in the dumbbell to two different families of limit eigenvalues.

3.4 Free plate

On the dumbbell $\Omega_\varepsilon \subset \mathbb{R}^2$ we consider the problem

$$\begin{cases} \Delta^2 u - \tau\Delta u + u = \lambda_n(\Omega_\varepsilon)\, u, & \text{in } \Omega_\varepsilon, \\ (1-\sigma)\frac{\partial^2 u}{\partial n^2} + \sigma\Delta u = 0, & \text{on } \partial\Omega_\varepsilon, \\ \tau\frac{\partial u}{\partial n} - (1-\sigma)\,\mathrm{div}_{\partial\Omega_\varepsilon}(D^2u\cdot n)_{\partial\Omega_\varepsilon} - \frac{\partial(\Delta u)}{\partial n} = 0, & \text{on } \partial\Omega_\varepsilon. \end{cases}$$

where $\sigma \in (-1,1)$, $\tau \geq 0$, and $u \in H^2(\Omega_\varepsilon)$. The weak formulation is: *find* $u \in H^2(\Omega_\varepsilon)$ *such that for any* $\psi \in H^2(\Omega_\varepsilon)$

$$\int_{\Omega_\varepsilon} (1-\sigma)D^2u : D^2\psi + \sigma\Delta u\Delta\psi + \tau\nabla u\cdot\nabla\psi + u\psi\, dx = \lambda(\Omega_\varepsilon)\int_{\Omega_\varepsilon} u\psi\, dx$$

We denote by $(\varphi_n^\varepsilon, \lambda_n(\Omega_\varepsilon))$ the eigenpairs $\forall n \in \mathbb{N}$. We introduce the eigenpairs $(\varphi_k^\Omega, \omega_k)$ of

$$\begin{cases} \Delta^2 w - \tau\Delta w + w = \omega_k\, w, & \text{in } \Omega, \\ (1-\sigma)\frac{\partial^2 w}{\partial n^2} + \sigma\Delta w = 0, & \text{on } \partial\Omega, \\ \tau\frac{\partial w}{\partial n} - (1-\sigma)\mathrm{div}_{\partial\Omega}(D^2w\cdot n)_{\partial\Omega} - \frac{\partial(\Delta w)}{\partial n} = 0, & \text{on } \partial\Omega, \end{cases}$$

and the eigenpairs (h_i, θ_i) of

$$\begin{cases} \frac{1-\sigma^2}{g}(gv'')'' - \frac{\tau}{g}(gv')' + v = \theta v, & \text{in } (0,1), \\ v(0) = v(1) = 0, \quad v'(0) = v'(1) = 0 \end{cases}$$

Let $(\lambda_n)_n = (\omega_k)_k \cup (\theta_i)_i$, where the eigenvalues are arranged in increasing order and repeated according to the multiplicity. Under suitable assumptions on R_ε (stricter than in the Laplacian case), one can prove that

$$\lambda_n(\Omega_\varepsilon) \to \lambda_n, \qquad \text{as } \varepsilon \to 0,$$

with convergence of the spectral projections. We refer to [7] for the proofs of these facts. We remark that the main difference with the Laplacian case is the presence of σ in the limit problem

$$\begin{cases} \frac{1-\sigma^2}{g}(gv'')'' - \frac{\tau}{g}(gv')' + v = \theta v, & \text{in } (0,1), \\ v(0) = v(1) = 0, \quad v'(0) = v'(1) = 0 \end{cases}$$

which causes several technical difficulties in the passage to the limit. In order to pass successfully to the limit a mixed approach is needed. The main ingredients of this approach are:

- *rescaling/homogenization techniques*,
- *abstract compact convergence results* for the operators $(\Delta^2_{\Omega_\varepsilon})^{-1}$

Define the extension operator $\mathcal{E}_\varepsilon : L^2((0,1); g(x)dx) \to L^2(R_\varepsilon)$ by $\mathcal{E}_\varepsilon\varphi(x,y) = \varphi(x)$ for all $(x,y) \in R_\varepsilon$. Moreover let $N(x)$ be the counting function defined by

$$N(x) = \#\{\lambda_i : i \in \mathbb{N}, \lambda_i \leq x\}.$$

Finally one can prove the following

Theorem 2 *Let $\Omega_\varepsilon \subset \mathbb{R}^2$ be a dumbbell domain satisfying the (H)-condition. The eigenvalues $\lambda_n(\Omega_\varepsilon)$ converge either to ω_k or to θ_l. Moreover, if $\lambda_n(\Omega_\varepsilon) \to \omega_k$ for some $k \in \mathbb{N}$, then $\|\varphi_n^\varepsilon|_\Omega\|_{L^2(\Omega)} \to 1$, and*

$$\left\| \varphi_n^\varepsilon|_\Omega - \sum_{i=1}^{N(\omega_k)} (\varphi_n^\varepsilon, \varphi_i^\Omega)_{L^2(\Omega)} \varphi_i^\Omega \right\|_{H^2(\Omega)} \to 0$$

as $\varepsilon \to 0$. Otherwise, if $\lambda_n(\Omega_\varepsilon) \to \theta_l$ for some $l \in \mathbb{N}$, then $\varphi_n^\varepsilon|_\Omega \to 0$ in $L^2(\Omega)$ and

$$\left\| \varphi_n^\varepsilon - \sum_{i=1}^{N(\theta_l)} (\varphi_n^\varepsilon, \varepsilon^{-1/2}\mathcal{E}_\varepsilon h_i)_{L^2(R_\varepsilon)} \varepsilon^{-1/2}\mathcal{E}_\varepsilon h_i \right\|_{L^2(R_\varepsilon)} \to 0$$

as $\varepsilon \to 0$.

References

[1] J.M. Arrieta, "Spectral properties of Schrödinger operators under perturbations of the domain". Doctoral Dissertation, Georgia Institute of Technology, 1991.

[2] J.M. Arrieta, *Neumann eigenvalue problems on exterior perturbations of the domain.* J. Differential Equations, 117 (1995).

[3] J.M. Arrieta, *Rates of eigenvalues on a dumbbell domain. Simple eigenvalue case.* Trans. Amer. Math. Soc., 347/9 (1995), 3503–3531.

[4] J.M. Arrieta and A.N. Carvalho, *Spectral convergence and nonlinear dynamics of reaction-diffusion equations under perturbations of the domain.* J.Differential Equations 199 (2004), 143–178.

[5] J.M. Arrieta and A.N. Carvalho, *Dynamics in dumbbell domains I. Continuity of the set of equilibria.* J. Differential Equations 231/2 (2006), 551–597.

[6] J.M. Arrieta, F. Ferraresso and P.D. Lamberti, *Boundary homogenization for a triharmonic intermediate problem.* To appear in Mathematical Methods in the Applied Sciences.

[7] J.M. Arrieta, F. Ferraresso and P.D. Lamberti, *Spectral analysis of the biharmonic operator subject to Neumann boundary conditions on dumbbell domains.* Preprint (2017).

[8] J.M. Arrieta, J.K. Hale, and Q. Han, *Eigenvalue problems for nonsmoothly perturbed domains.* J. Differential Equations 91 (1991), 24–52.

[9] J.M. Arrieta and P.D. Lamberti, *Higher order elliptic operators on variable domains. Stability results and boundary oscillations for intermediate problems.* Preprint, online at arXiv:1502.04373v2 [math.AP].

[10] D. Bucur and G. Buttazzo, "Variational methods in some shape optimization problems". Appunti dei Corsi Tenuti da Docenti della Scuola. [Notes of Courses Given by Teachers at the School].

[11] D. Buoso and P.D. Lamberti, *Eigenvalues of polyharmonic operators on variable domains.* ESAIM. Control, Optimisation and Calculus of Variations 19/4 (2013), 1225–1235.

[12] D. Buoso and L. Provenzano, *A few shape optimization results for a biharmonic Steklov problem.* J. Differential Equations 259/5 (2015), 1778–1818.

[13] V. Burenkov and P.D. Lamberti, *Sharp spectral stability estimates via the Lebesgue measure of domains for higher order elliptic operators.* Rev. Mat. Complut. 25 (2012), no. 2, 435–457.

[14] R. Courant and D. Hilbert, "Methods of Mathematical Physics Vol.I". Wiley-Interscience, New York, 1953.

[15] E.B. Davies, "Spectral theory and differential operators". Cambridge Studies in Advanced Mathematics, 42. Cambridge University Press, Cambridge, 1995.

[16] J. Hale and G. Raugel, *Partial differential equations on thin domains.* Differential Equations and Mathematical Physics, Math. Sci. Engrg. 186 (1992), 63–97.

[17] F. Gazzola, H.-C. Grunau and G. Sweers, "Polyharmonic boundary value problems - Positivity preserving and nonlinear higher order elliptic equations in bounded domains". Lecture Notes in Mathematics, Springer-Verlag, Berlin, 2010.

[18] D. Gilbarg and N.S. Trudinger, "Elliptic partial differential equations of second order". Springer-Verlag, Berlin, 2001.

[19] S. Jimbo, *The singularly perturbed domain and the characterization for the eigenfunctions with Neumann boundary conditions.* J. Differential Equations 77 (1989), 322–350.

[20] J. Nečas, "Les méthodes directes en théorie des équations elliptiques". Masson et Cie, Éditeurs, Paris; Academia, Éditeurs, Prague 1967.

[21] L. Provenzano, *A note on the Neumann eigenvalues of the biharmonic operator.* To appear in Mathematical Methods in the Applied Sciences (2016).

[22] V. Šverák, *On optimal shape design.* J. Math. Pures Appl. (9), 72/6 (1993), 537–551.

Variational Approaches in Shape Partitioning

Martin Huska (*)

Abstract. The rapid development of 3D scanning technology has incredibly increased the availability of digital models exploited for a wide range of applications varying from computer graphics and medical imaging up to industrial production. One fundamental procedure that processes the raw acquired data for further manipulation, e.g. in product design, animation, deformation and reverse engineering, is the shape partitioning. This process consists in the decomposition of an object into non-overlapping salient sub-parts determined by a shape attribute. In this seminar, we will introduce the concept of Shape Partitioning together with the wide range of partitioning methods. Next, we will observe a few partitioning/segmentation models in the field providing some results. At last, if the time allows, we will introduce the concept of Convex-Nonconvex segmentation over surfaces.

1 Introduction

The digital models of 3D physical objects simply obtained by 3D scanning are represented by a set of 3D points on the surface of the object, that we call cloud of points. These raw 3D data can be by default connected by the scanner into spatial triangulations, calling 3D meshes. However, triangulated meshes provide only local information on the structure of the surface. A high level insight of the raw 3D data is required to make the digital model useful for further processing required in a variety of applications including surface processing, CAD, CAM and CAE.

Mesh segmentation is fundamental for many computer graphics and animation techniques such as modeling, rigging, shape-retrieval, and deformation. Given an object with arbitrary topology and a discrete manifold representing the object's boundary, this process consists in the decomposition of an object in K-disjoint salient sub-parts and it relies mostly on surface geometric attributes of the object's boundary. This process provides the first step to high level representation of the raw data.

Specific criteria dictate which elements belong to the same partition and these criteria are built upon the segmentation objective which in turn depends on the application.

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: martin@math.unipd.it . Seminar held on May 3rd, 2017.

Convexity/Concavity and *thickness* are popular shape criteria used in mesh decomposition. The convexity-driven segmentation of a shape finds a very intuitive match with the decomposition of an object made by the human vision system. This is due to the fact that an approximate convex decomposition can more accurately represent the important structural features of the model by ignoring insignificant features, such as wrinkles and other surface texture. Unlike, the thickness of parts of a shape is a less intuitive detection strategy for a human eye. Nevertheless, this geometry feature represents a strategic quantity in shape analysis in the context of industrial design and production.

In addition to the criteria that dictate the rules of the division into parts, the segmentation methods can be grouped into a few categories according to their computational methodology: (i) Region growing; (ii) Watershed-based; (iii) Reeb graphs; (iv) Model-based; (v) Skeleton-based; (vi) Clustering; (vii) Spectral analysis; (viii) Explicit Boundary Extraction; (ix) Critical points-based; (x) Multiscale Shape Descriptors; (xi) Markov Random Fields and (xii) Variational segmentation. A detailed analysis of the aforementioned categories is given in [1] and exhaustive surveys are provided in [2, 7]. The presented works falls into the latter category, which commonly shares the concept of iteratively seeking a partition that minimizes a given error metric.

2 Segmentation features

Let us consider a triangle mesh $\Omega := (V, T)$ which discretizes manifold $\mathcal{M}$ embedded in $\mathbb{R}^3$, where $V = \{X_i\}_{i=1}^n \in \mathbb{R}^{n\times 3}$ is the set of vertices, and $T \in \mathbb{N}^{n_T\times 3}$ is the connectivity matrix. Each vertex X_i has thus an immediate neighbours $X_j\,, j \in N(X_i)$ connected by an edge e_{ij} that create set of edges E. Naturally, $N(X_i) = N_i$ defines the first ring neighbours to vertex X_i, and $N_\triangle(X_i)$ denotes the neighbouring triangles τ_j containing X_i. By $A(\tau_j)$ we denote the area of τ_j.

Regarding the convexity, mesh associated affinity matrix should encode its structural information which reflects how vertices are grouped in accordance with human perception. Taking into account the curvature as shape information, we want to determine a perceptual partition of Ω such that the edges between different parts have very low weights (vertices in different clusters are dissimilar from each other), and the edges belonging to the same part have high weights (vertices within the same cluster are similar to each other). At this aim we define the affinity matrix $\mathcal{L} \in \mathbb{R}^{n\times n}$:

$$\mathcal{L}_{i,j} = \begin{cases} -w_{ij}, & i \neq j \ \text{ and } \ e_{ij} \in \mathcal{E} \\ \sum_{j\in N(X_i)} w_{ij}, & i \ = \ j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

with the following similarity non-negative weights:

$$w_{ij} \ := \ \frac{|N_\triangle(X_j)|}{\#N(X_i)}\, e^{-\|H(X_i)-H(X_j)\|_2^2/(2\sigma^2)}, \tag{2}$$

where the parameter $\sigma \in (0, 1]$ in (2) controls the width of the local neighborhoods. The

mean curvature field H in (2) is obtained by exploiting the well-known relation

$$\Delta_{LB} X = -2H\,\mathbf{N}, \tag{3}$$

between the vector field $H\,\mathbf{N}$ and the Laplace-Beltrami differential operator Δ_{LB}, applied to the coordinate functions X of a surface. Discretization of the Laplace-Beltrami operator (3) reads as

$$\begin{aligned} L(X_i) &= \frac{1}{2|N_{\triangle}(X_i)|} \sum_{j \in N(X_i)} \omega_{ij}\,(X_j - X_i)\,, \\ \omega_{ij} &= \tfrac{1}{2}(\cot\gamma_j + \cot\delta_j)\,, \end{aligned} \tag{4}$$

where γ_j, δ_j are the opposite angles to the edge $e_{i,j}$ in the triangles tuple connected by the edge.

The spectral decomposition of $\mathcal{L}$, defined in the following, provides a set of $(n-1)$ non trivial, smooth, shape intrinsic isometric-invariant maps.

The matrix $\mathcal{L} \in \mathbb{R}^{n\times n}$ defined in (1) associated to a connected mesh Ω of n vertices, satisfies the following properties:

1) $\mathcal{L}$ is symmetric and positive semi-definite;

2) $\mathcal{L} = U\,\Lambda\,U^T$, $\Lambda = diag(\lambda_i)$, $0 = \lambda_0 < \lambda_1 < \ldots < \lambda_n$;

3) $\lambda_i, \forall i$ are real eigenvalues, $U^T U = I_n$ with I_n the identity matrix of order n, $U = \{v_0, v_1, \ldots, v_n\}$ form an orthogonal basis of $\mathbb{R}^n$;

4) If $f = \sum_{i=1}^{n} \langle f, v_i\rangle\, v_i$, the k-term approximation of f is given by

$$f_k = \sum_{i=1}^{k} \langle f, v_i\rangle\, v_i.$$

The first k eigenvectors associating to the smallest nonzero eigenvalues, correspond to smooth and slowly varying functions, while the last one show more rapid oscillations. Property 4) defines the truncated spectral approximation of the $\mathcal{L}$ matrix, that considers the contribution of the first k eigenpairs related to the smallest eigenvalues which hold for identifying the main shape features at different scale forming a signature for shape characterization.

In case of eigen-decomposition-based segmentation, a vector function f is simply the truncated spectral coordinates of a vertex X_i, denoted by

$$f(X_i) = (v_1(X_i),\, v_2(X_i),\, \ldots, v_d(X_i)\,), \quad d \le k, \tag{5}$$

where each v_j is normalized in the range $[-1, 1]$.

In Fig. 1 the first $k = 6$ eigenvectors of the affinity matrix (1) corresponding to the first six nonzero eigenvalues are illustrated for the `horse` mesh, visualized in false colors in the range [blue,red].

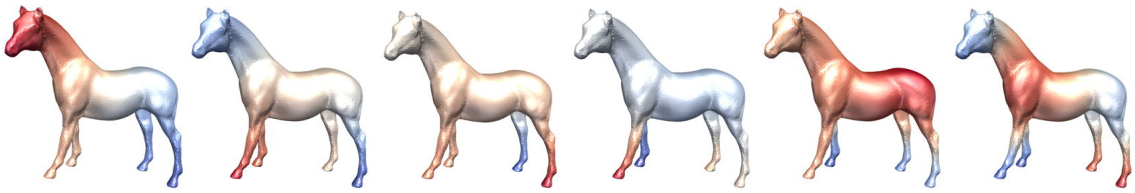


Figure 1. The smallest $k = 6$ eigenfunctions of the `horse` mesh.

On the opposite, as a measure of thickness that recovers volumetric information from the surface boundaries, thus providing a natural link between the object's volume and its boundary, the Shape Diameter Function (SDF) was introduced in [8]. The SDF is a scalar function which maps for every point on the surface its distance to the opposite inner part of the object. We can understand it as a shape diameter in selected points. Moreover, there is also a noticeable connection between SDF and Medial Axis, since distance from a point to Medial Axis is approximately half of the SDF value in the point.

In the original proposal in [8] the basic operation in the SDF computation is a ray-mesh intersection. Given a point on a mesh, several rays inside a cone centered around the inward-normal direction are sent towards the opposite side of the mesh. The intersected points are filtered to remove false intersections and finally their ray-lengths are summed as weighted contributions to the final SDF value of that point.

Alternatively, a simple particle flow can replace the ray-tracing procedure, as was successfully shown in [3]. The proposed method, which can be used also in case of clouds of points, evolve the object's vertices X_i, $i = 1, \ldots, n$ in the inward normal direction given by $-\mathbf{N}_i$ as particles $x_i(t)$. At each time step t_k, the particles are placed in the voxelized envelope and a simple collision test is performed:

Two evolving particles x_i and x_j collide if their trajectories intersect and they have the same, but opposite, velocity direction under a certain tolerance α, that is:

$$\measuredangle\,(\mathbf{N}_i, -\mathbf{N}_j) \leq \alpha. \tag{6}$$

The angle value α allows for a cone of admissible directions that, relaxing the ideal case where $\mathbf{N}_i = -\mathbf{N}_j$, better fits the spatial discretization of the surface.

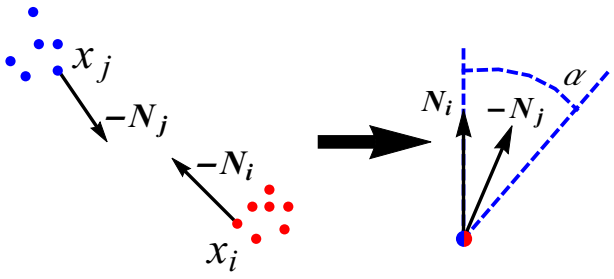


Figure 2. Collision test (6) for two particles x_i and x_j.

The evolution of a particle is stopped by the successful collision test, illustrated in Fig. 2. When the particle x_i (and x_j) stops its flow, a scalar function value $f_{SDF}(\cdot)$,

$f_{SDF} : V \to \mathbb{R}^+$, is computed as the distance between X_i and X_j, with $X_i, X_j \in V$, and associated both to the particle X_i and X_j.

In Figure 3 we plot the f_{SDF} function over mesh vertices with false colours in [blue,red] assigned to the SDF values.

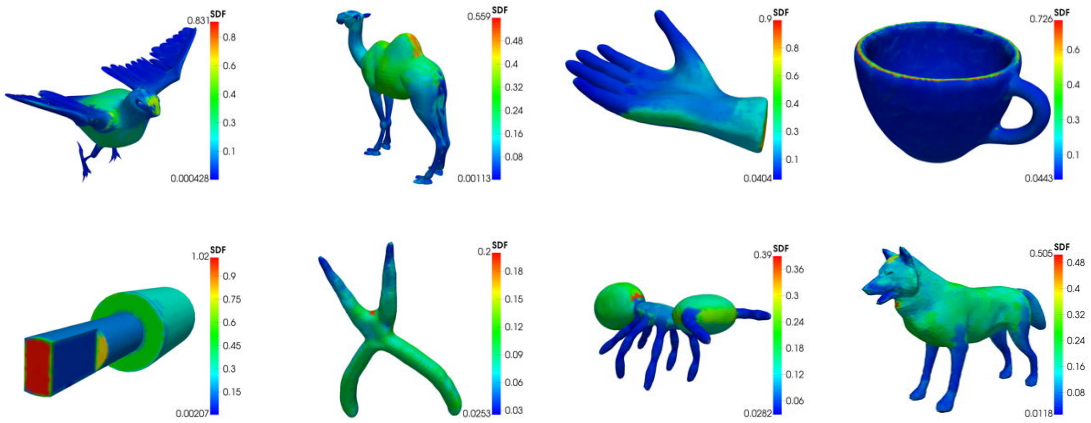


Figure 3. Examples of the scalar SDF function computed by [3].

3 Variational Shape Partitioning – SMCMR

Sparsity-inducing Multi-Channel Multiple Region algorithm [5] presents an object partitioning framework which consists of the two following steps:

Step 1 is based on a variational formulation with a sparsity-forcing penalty, providing single- or a multi-channel partitioning function;

Step 2 uses the partitioning function to split the object into K parts. To that aim, simple K-means-based algorithms are sufficient.

This formulation makes the algorithm efficient since the first, computationally costly step does not depend on the number of partitions required.

The strategy for the partitioning of meshes in Step 1 is based on a new variant of the regularized Mumford-Shah models where we adopt an L_p-norm approximation of the total length of the boundaries.

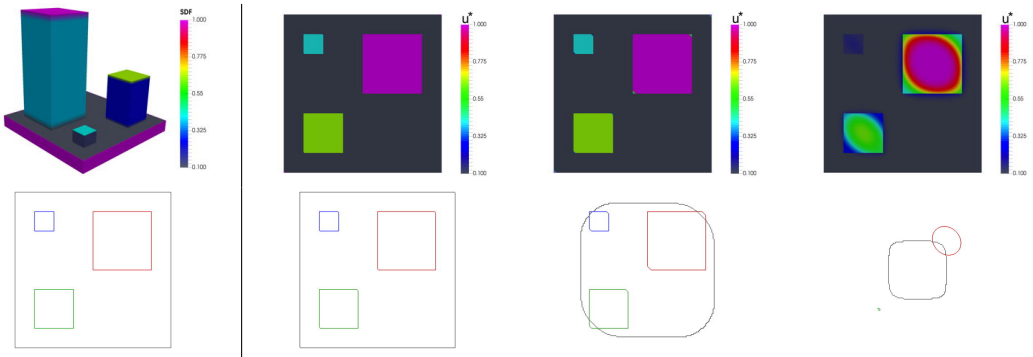
Let $f = (f_1, \ldots, f_d)$ be a given vector-valued function with channels $f_i : \Omega \to \mathbb{R}$, $i = 1, .., d$, and $u = (u_1, \ldots, u_d)$ be a vector function on Ω, eventually nonsmooth, named the *partition function*. Unlike for the color image segmentation process where all image channels participate jointly in driving the segmentation process, here we apply the variants of the Mumford-Shah models to each channel u_i of u, for $i = 1, \ldots, d$. In particular, in the first step, each channel u_i is separately computed by solving the following optimization problem:

$$u^* \leftarrow \arg\min_{u_i} \{\mathcal{J}_s(u_i)\} \tag{7}$$

$$\mathcal{J}_s(u_i) := \frac{1}{2}\int_\Omega |f_i - u_i|^2 d\Omega + \frac{\alpha}{p}\int_\Omega \phi\left(\|\nabla u_i\|\right)\, d\Omega + \frac{\beta}{2}\int_\Omega |\nabla u_i|^2 d\Omega, \tag{8}$$

where $\phi(t) := |t|^p$, is the penalty function with $p \in (0, 2]$, sparsity-inducing for $p < 1$, α is a trade-off parameter and $\beta := \beta(x)$, $\beta : \Omega \to [0, 1]$ is an adaptive function which approaches to zero at the high curvature points of Ω.

In (8), the first term is called fidelity and penalizes 'fitness' of the partition function u_i to input f_i. The second term penalizes the partition boundary lengths and the last term penalizes non-smoothness of the inner regions, thus we can call $\mathcal{J}_s$ piecewise-smooth functional. In case $\beta = 0$, we obtain as a special case the piecewise-constant functional which we use in case of the single-channel f_{SDF} input function.



(a) SDF partitioning ground truth (b) $p = 0.2$ (c) $p = 0.8$ (d) $p = 1.0$

Figure 4. Effect of the ℓ_p regularizer wrt the ℓ_1 regularizer for the SDF partitioning of the `blocks` mesh.

The benefit of using $p < 1$ is illustrated in Fig. 4 for the segmentation of a mesh composed of variable size boxes (Fig. 4(a) top). Since the thickness property is used as criteria for partitioning, from the top view, the expected results are four boxes which are shown in Fig. 4(a) bottom. The true thicknesses (heights) were used as thresholds. The top row of Fig. 4(b-c-d), shows the results obtained from the proposed variational model for different p, which are used in Step 2 to produce the simple partitions, according to the given thresholds, which represent the true heights. In the bottom row, we plot the partition boundaries, obtained as iso-contours of u^* in the top row, according to the thresholds. For the choice $p < 1$ in (8) our model preserves the sharp boundary shape, as illustrated in Fig. 4(b-c), while for $p = 1$, recovering L_1 norm, the boundaries shrink and the small features disappear as illustrated in Fig. 4(d). In particular for p approaching to zero the boundary shape improves and the original intensities are preserved.

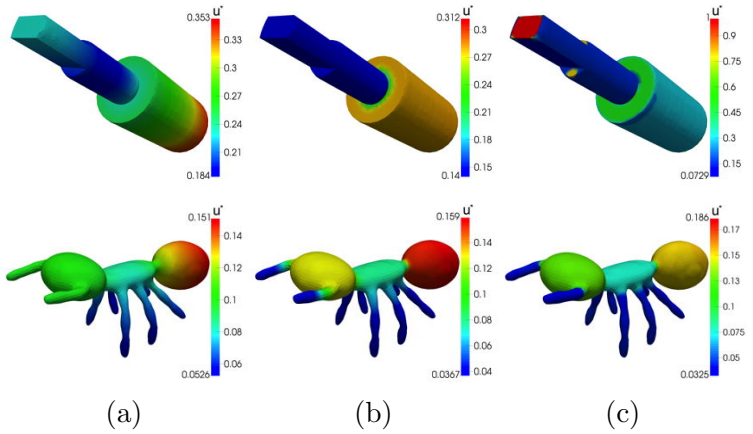


Figure 5. Effect of the parameter p on the Step 1 in **Algorithm SMCMR**. (a) $p = 2$, (b) $p = 1$, (c) $p = 0.8$.

Another benefit of L_p norm regularization, the sparser solutions, is shown in Fig. 5. In particular, we compared the results recovering L_2 norm, for $p = 2$, the L_1 norm, for $p = 1$, and the L_p norm for $p = 0.8$. While the L_2 norm regularization behaves as diffusion, Fig. 5(a), the partitioning function u^* obtained with $p = 1$, Fig. 5(b), is accompanied with varying gradient seen especially on the `ant` mesh legs. However, decreasing p, Fig. 5(c), the partitioning function resembles already a result of the partitioning thanks to the very few jumps in the gradient magnitude of u^*.

The discretized variational problem was solved by an iterative Proximal Forward-Backward-based scheme where we take use of lagged diffusivity fixed point algorithm for gradient linearization.

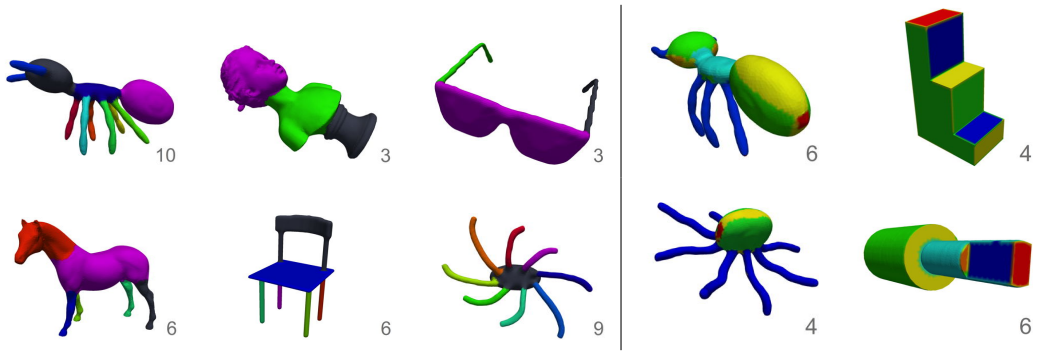


Figure 6. Examples of SMCMR Partitioning in K parts. Left, using eigenvectors of $\mathcal{L}$ in (1). Right, using Shape Diameter Function f_{SDF}.

In Fig. 6 we show some examples obtained by SMCMR framework. On the left side, we used the eigen-decomposition-based segmentation mimicking the human-vision perception. The right side represents thickness-based segmentation via f_{SDF} function. In this

example, one can appreciate the difference between human- and thickness-based segmentation, especially in case of the `octopus` meshes in the bottom row. While a person sees head and eight legs, from the thickness point of view, all the legs are of similar thickness, thus, one part.

4 L_p Compressed Modes

In the shape partitioning context, rather than a multiresolution representation of the shape, which is the peculiarity of the MHB on manifolds, the focus is on identifying the observable features of the manifold which represent for example protrusions, ridges, details in general localized in small regions.

Hence, in the partitioning context, a more suitable alternative to the MHB is represented by the Compressed Manifold Basis (CMB), introduced in [6], which is characterized by compact support quasi-eigenfunctions of the Laplace-Beltrami Operator (LBO) obtained by imposing sparsity constraints.

Motivated by the advantages in terms of control on the compact support obtained by using the L_1 norm to force the sparsity of the solution, we devised to replace the L_1 norm by a more effective sparsity-inducing L_p norm term, with $0 < p \leq 1$, which stronger enforces the locality of the resulting basis functions. The set of functions $\Psi = \{\psi_k\}_{k=1}^N$, that we will call L_p Compressed Modes (L_pCMs) [4], is computed by solving the following variational model

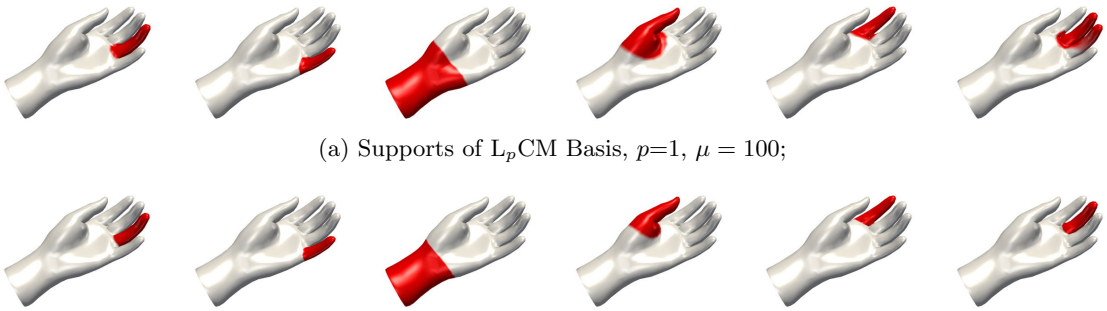
$$\min_{\Psi} \sum_{k=1}^{N} \left(\frac{1}{\mu} \int_\Omega |\psi_k|^p \, dx - \frac{1}{2} \int_\Omega \psi_k \Delta \psi_k dx \right) \quad s.t. \quad \int_\Omega \psi_j \, \psi_k \, dx \, = \, \delta_{jk}, \tag{9}$$

where we denoted the L_p norm of a function by $\|f\|_p = \left(\int_\Omega |f|^p dx\right)^{1/p}$; δ_{jk} is the Kronecker delta, and $\mu > 0$ is a penalty parameter. They form an orthonormal basis for the $L^2(\Omega)$ space, where Ω is the domain in consideration, and they represent a set of quasi-eigenfunctions of the Laplace-Beltrami operator.

The second term in the objective function of (9) is the fidelity term which represents the accuracy of the shape approximation provided by the set of functions Ψ, while the first term, so-called penalty term, forces the sparsity in the functions Ψ thus imposing spatially sparse solutions.

The penalty parameter μ controls the compromise between the two aspects. It is well known that the sparsity is better induced by the L_p norm for $0 < p < 1$, rather than the L_1 norm. For $p = 1$ model (9) reduces to the proposal in [6], where the sparsity is forced only by acting on the μ value to increase the contribution of the penalty term, thus decreasing the shape approximation guaranteed by the fidelity term.

The parameter p plays a crucial role since it allows for forcing the sparsity while maintaining the approximation accuracy without excessively stressing the penalty via the μ value. The accuracy is fundamental to localize the support of the functions in specific local features of the shape such as protrusions and ridges.



(a) Supports of L_pCM Basis, p=1, $\mu = 100$;

(b) Supports of L_pCM Basis, $p = 0.8$, $\mu = 1000$;

Figure 7. Partitioning of the 2-manifold `hand` using L_1CM basis (a) and L_pCM basis (b).

Some evidence of the benefit obtained by the sparsity-inducing proposal, is shown in Fig. 7 where we try to answer the following question: Can we identify the most salient parts of the manifold `hand` using only six compressed modes? Fig. 7 compares the supports of the compressed modes determined as solution of the variational problem (9) with $p = 1$ (Fig. 7(a)) and $p = 0.8$ (Fig. 7(b)) where the μ parameter has been selected to provide the best results. The supports of the six quasi-eigenfunctions are colored in red and we can observe that using $p < 1$ strengthens the sparsity, while if $p = 1$ no μ value has allowed to correctly identify all the fingers.

In Fig. 8 we present the role of parameter μ described above. It is easily noticeable that increasing μ, we decrease the sparsity-inducing effect of the regularization parameter in (9), thus, increasing the support of the basis function.

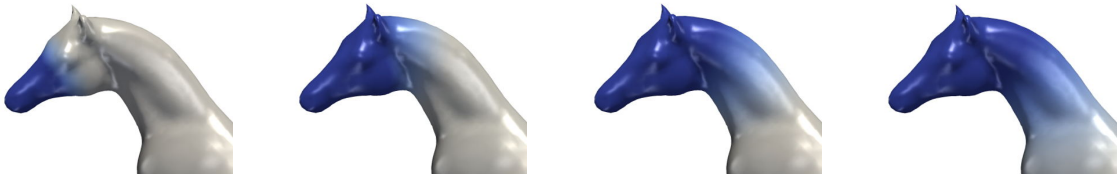


Figure 8. L_pCM localizing protrusion – head in a `horse` mesh. From left to right: enlarging of the support obtained by increasing the parameter μ.

An efficient solution of the orthogonality constrained problem (9) represents a challenging task due to the nonlinear, non-convex orthogonality constraints combined with the non-smooth and non-convex objective function, which may lead to many different local minimizers as solutions. Non-trivial iterative approaches are commonly used to solve this kind of optimization problems. We propose a variant of the basic Alternating Direction Method of Multipliers (ADMM) approach where the non-convex orthogonality constraints defined for the L_p Compressed Modes Ψ in (9) are preserved by means of a SVD matrix factorization, and a suitable proximal operator is devised to deal with the non-convex penalty.

Having the L_pCM basis, the partitioning is rather straightforward and can be realized by a simple region-growing algorithm driven by the compressed modes.

4.1 Discretization of the model

We first recall the popular discretization of the Laplace-Beltrami operator for a triangle mesh, which may be realized by $D^{-1}L$, where $L \in \mathbb{R}^{n\times n}$ is a symmetric, positive semi-definite, sparse matrix (weight matrix) defined as

$$L(i,j) := \begin{cases} \omega_{ij} = \frac{1}{2}(\cot\gamma_j + \cot\delta_j) & j \in N(X_i) \\ -\sum_{k\in N(X_i)} \omega_{ik} & i = j \\ 0 & otherwise \end{cases} \tag{10}$$

where γ_j, δ_j are the opposite angles to the edge $e_{i,j}$ in the triangles tuple connected by the edge; D is a lumped mass matrix defined as

$$D = diag\{|N_\triangle(X_1)|, \cdots, |N_\triangle(X_n)|\}$$

which relates to the area/volume around the vertices of the discretized manifold.

By applying the discretization $D^{-1}L$ for the LBO on given mesh, and arranging the discretized L_pCMs in columns of a matrix $\Psi = [\psi_1, \ldots, \psi_N]$, with $\Psi \in \mathbb{R}^{n\times N}$, the constrained minimization problem (9) reads as follows

$$\Psi^* \;=\; \arg\min_{\Psi} \frac{1}{\mu}\|\Psi\|_p^p + \mathrm{Tr}\left(\Psi^T L\Psi\right) \quad s.t. \quad \Psi^T D\Psi = I, \tag{11}$$

where $Tr(\cdot)$ denotes the trace operator, and $\|\Psi\|_p^p = \sum_{i,j} |\Psi_{i,j}|^p$. The orthogonality constraints in problem (11) are bounded above by quadratic functions.

The Lagrangian function of (11) is defined as

$$\mathcal{J}(\Psi, \Lambda) = \frac{1}{\mu}\|\Psi\|_p^p + \mathrm{Tr}\left(\Psi^T L\Psi\right) - \mathrm{Tr}\left(\Lambda(\Psi^T D\Psi - I)\right), \tag{12}$$

where Λ is the matrix of Lagrangian multipliers. The function (12) is proper, lower semi-continuous, bounded from below and coercive.

4.2 Applying ADMM to the model

We illustrate the ADMM-based iterative algorithm used to numerically solve the proposed model (11).

First, we replace the orthogonality constraint in (11) using an indicator function

$$\iota(\Psi) = \begin{cases} 0 & \text{if} \quad \Psi^T D\Psi = I \\ \infty & \text{otherwise.} \end{cases}$$

Then problem (11) can be rewritten as:

$$\Psi^* \;=\; \arg\min_{\Psi} \frac{1}{\mu}\|\Psi\|_p^p + \mathrm{Tr}(\Psi^T L\Psi) + \iota(\Psi). \tag{13}$$

We can resort to the variable splitting technique for the orthogonality constraint and introduce two new auxiliary matrices, $E, S \in R^{n\times N}$, the problem (13) is then rewritten as:

$$\min_{\Psi,S,E} \frac{1}{\mu}\|S\|_p^p + \mathrm{Tr}(E^T LE) + \iota(\Psi) \quad s.t. \quad \Psi = S,\ \Psi = E. \tag{14}$$

To solve problem (14), we define the augmented Lagrangian functional

$$\begin{aligned}\mathcal{L}(\Psi, S, E; U_E, U_S; \mu) \;\; &= \;\; \frac{1}{\mu}\|S\|_p^p + \mathrm{Tr}(E^T L E) + \iota(\Psi) \\ &\quad - \langle\, U_S, \Psi - S \,\rangle \;\; + \frac{\rho}{2}\,\|\Psi - S\|_F^2 \\ &\quad - \langle\, U_E, \Psi - E \,\rangle + \frac{\rho}{2}\,\|\Psi - E\|_F^2 \;\;, \end{aligned} \tag{15}$$

where $\rho > 0$ is scalar penalty parameter and $U_S \in \mathbb{R}^{n\times N}$, $U_E \in \mathbb{R}^{n\times N}$ are the matrices of Lagrange multipliers associated with the linear constraints $\Psi = S$ and $\Psi = E$ in (14), respectively.

We then consider the following saddle-point problem:

$$\begin{aligned} &\text{Find} && (\Psi^*, S^*, E^*; U_S^*, U_E^*) \;\in\; \mathbb{R}^{n\times N}\times\; \mathbb{R}^{n\times N}\times\; \mathbb{R}^{n\times N}\times\; \mathbb{R}^{n\times N}\times\; \mathbb{R}^{n\times N} \\ &\text{s.t.} && \mathcal{L}\,(\Psi^*, S^*, E^*; U_E, U_S; \mu) \;\le\; \mathcal{L}\,(\Psi^*, S^*, E^*; U_E^*, U_S^*; \mu) \;\le\; \mathcal{L}\,(\Psi, S, E; U_E^*, U_S^*; \mu) \\ &&& \forall\,(\Psi, S, E; U_E, U_S) \;\in\; \mathbb{R}^{n\times N}\times\; \mathbb{R}^{n\times N}\times\; \mathbb{R}^{n\times N}\times\; \mathbb{R}^{n\times N}\times\; \mathbb{R}^{n\times N}\,, \end{aligned} \tag{16}$$

with the augmented Lagrangian functional $\mathcal{L}$ defined in (15).

Given the previously computed (or initialized for $k = 0$) matrices $S^{(k)}$, $E^{(k)}$, $U_S^{(k)}$ and $U_E^{(k)}$, the k-th iteration of the proposed ADMM-based iterative scheme applied to the solution of the saddle-point problem (15)–(16) reads as follows:

$$\Psi^{(k+1)} \;\leftarrow\; \arg\min_{\Psi\in\mathbb{R}^{n\times N}} \mathcal{L}(\Psi, S^{(k)}, E^{(k)}; U_S^{(k)}, U_E^{(k)}) \tag{17}$$

$$S^{(k+1)} \;\leftarrow\; \arg\min_{S\in\mathbb{R}^{n\times N}} \mathcal{L}(\Psi^{(k+1)}, S, E^{(k)}; U_S^{(k)}, U_E^{(k)}) \tag{18}$$

$$E^{(k+1)} \;\leftarrow\; \arg\min_{E\in\mathbb{R}^{n\times N}} \mathcal{L}(\Psi^{(k+1)}, S^{(k+1)}, E; U_S^{(k)}, U_E^{(k)}) \tag{19}$$

$$U_S^{(k+1)} \;\leftarrow\; U_S^{(k)} - \rho\left(\,\Psi^{(k+1)} - S^{(k+1)}\,\right) \tag{20}$$

$$U_E^{(k+1)} \;\leftarrow\; U_E^{(k)} - \rho\left(\,\Psi^{(k+1)} - E^{(k+1)}\,\right) \tag{21}$$

Let us have a look at the subproblems (17)–(19) since the updates in (20) and (21) have a closed-form solutions.

The subproblem for Ψ also have a closed-form solution

$$\Psi^{(k+1)} = YV\Sigma^{-1/2}V^T, \tag{22}$$

where $Y = \frac{1}{2}(S + \frac{1}{\rho}U_s + E + \frac{1}{\rho}U_E)$, $V \in \mathbb{R}^{N\times N}$ is a orthogonal matrix and Σ is a diagonal matrix satisfying the SVD factorization $Y^T D Y = V\Sigma V^T$.

The minimization sub-problem for S in (18) can be rewritten as follows:

$$S^{(k+1)} \leftarrow \arg\min_S \frac{1}{\mu}\|S\|_p^p + \frac{\rho}{2}\|\Psi - (S + \frac{1}{\rho}U_S)\|_F^2 \tag{23}$$

We can use the Generalized Iterated Shrinkage (GISA) strategy for Non-convex Sparse Coding proposed in [9], where the authors extended the popular soft-thresholding operator to l_p-norm. Rewriting component-wise Eq. (23), the minimization problem is equivalent to the following $n \times N$ independent scalar problems:

$$s_{i,j}^{(k+1)} \leftarrow \arg\min_{s_{i,j}\in\mathbb{R}} \left\{ f(s_{i,j}) = \frac{1}{\rho\mu}|s_{i,j}|^p + \frac{1}{2}(s_{i,j} - q_{i,j})^2 \right\}, \qquad \begin{array}{l} i = 1,\dots,n\,, \\ j = 1,\dots,N \end{array} \tag{24}$$

where $q_{i,j} = \psi_{i,j} - \frac{1}{\rho}(U_S)_{i,j}$. Following Theorem 1 in [9] each of the optimization problems (24) has a unique minimum given by

$$\mathrm{prox}_f^{\rho\mu}(q_{i,j}) \;=\; \begin{cases} 0 & \text{if } |q_{i,j}| \le \hat{s} \\ \mathrm{sign}(q_{i,j})\, s^* & \text{if } |q_{i,j}| > \hat{s}\ , \end{cases} \tag{25}$$

where the thresholding value is $\hat{s} = (\frac{2}{\rho\mu}(1-p))^{1/(2-p)} + \frac{p}{\rho\mu}(\frac{2}{\rho\mu}(1-p))^{(p-1)/(2-p)}$ and s^* is the unique solution of the following nonlinear equation:

$$s_{i,j} - q_{i,j} + \frac{p}{\rho\mu}(s_{i,j})^{p-1} \;=\; 0 \tag{26}$$

that can be easily solved by a few iterations of an iterative zero-finding algorithm.

The minimization problem of the augmented Lagrangian functional in (15) with respect to E in (19) reduces to the solution of N linear systems for E in the following form

$$(\rho I - 2L)E = \rho\left(\Psi - \frac{1}{\rho}U_E\right). \tag{27}$$

In Fig. 9 we show some examples obtained by the region-growing algorithm driven by computed L_p Compressed Modes basis Ψ. At the bottom right corner the number of partitions is identified which correspond to the number of basis vectors N in Ψ.

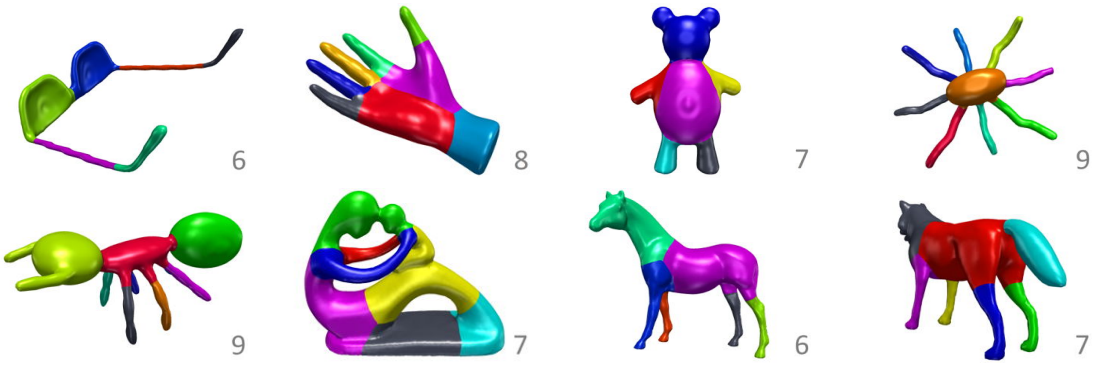


Figure 9. Mesh partitioning into salient parts obtained by L_pCMs [4].

References

[1] A. Agathos, I. Pratikakis, S. Perantonis, N. Sapidis, and P. Azariadis, *3D mesh segmentation methodologies for CAD applications.* Computer-Aided Design and Applications, 4 (2007), 827–841.

[2] M. Attene, S. Katz, M. Mortara, G. Patane, M. Spagnuolo, and A. Tal, *Mesh segmentation - a comparative study.* In Proceedings of the IEEE International Conference on Shape Modeling and Applications 2006, SMI '06, Washington, DC, USA, 2006, IEEE Computer Society, 7–20 (2006).

[3] M. Huska, S. Morigi, *A meshless strategy for shape diameter analysis.* Visual Computer, 33/3 (2017), 303–315.

[4] M. Huska, D. Lazzaro, S. Morigi, *Shape Partitioning via L_p Compressed Modes.* Submitted (2017).

[5] M. Huska, S. Morigi, *Sparsity-inducing variational shape partitioning.* Electronic Transactions on Numerical Analysis 46 (2017), 36–54.

[6] T. Neumann, K. Varanasi, C. Theobalt, M. Magnor, and M. Wacker, *Compressed Manifold Modes for Mesh Processing.* Computer Graphics Forum (Proc. of Symposium on Geometry Processing SGP), Eurographics Association, Vol. 33/5 (2014), 35–44.

[7] A. Shamir, *A survey on mesh segmentation techniques.* Computer Graphics Forum, 27 (2008), 1539–1556.

[8] L. Shapira, A. Shamir, and D. Cohen-Or, *Consistent mesh partitioning and skeletonisation using the shape diameter function.* The Visual Computer, 24 (2008), 249–259.

[9] W. Zuo, D. Meng, L. Zhang, X. Feng, and D. Zhang, *A Generalized Iterated Shrinkage Algorithm for Non-convex Sparse Coding.* International Conference of Computer Vision (ICCV), pp. 217–224, 2013.

The influence of network structure in neuronal information transmission

Giacomo Baggio (*)

Abstract. Understanding how neurons communicate is one of the most challenging open problems in neuroscience. In this note, we present some recent results aiming at formulating this problem from a mathematical and information-theoretic viewpoint. After an overview on neuronal network dynamical models, we introduce a digital communication framework for studying how "information" propagates in a neuronal network driven by linear dynamics. Within this framework, a novel metric for measuring the information capacity of a neuronal network based on Shannon's information capacity and the notion of inter-symbol interference is discussed. Finally, we investigate how the structure of the network matrix and, in particular, its departure from normality, affects the information capacity of a network.

1 Introduction

Human brain contains approximately 100 billion neurons, each of which has up to 10^5 synapses, i.e. anatomical connections between neurons through which "information" flows from one neuron to another. Individually, neurons are noisy and synaptic transmission is, in general, unreliable. However, groups of neurons that are arranged in specialized modules can collectively perform complex information processing tasks in a robust and efficient way. A classical example is given by the information processed within the motor cortex region for the execution and control of voluntary movements [4, 7]. Understanding what are the mechanisms that support complex yet reliable information routing and processing in the brain has been a topic of intense interest in the computational neuroscience community. As a matter of fact, many works have addressed this problem under different perspectives, assumptions, and constraints, see e.g. [2, 11, 13, 17] to cite just a few contributions.

In this note, we investigate the problem of information transmission in a neuronal network via a system and information-theoretic approach. More specifically, we consider a neuronal network driven by linear firing rate dynamics and we introduce a digital communication framework to examine information transmission in such a network. In this framework, Shannon's information capacity provides a measure of how much information

(*)Università di Padova, D.E.I., via Giovanni Gradenigo 6, I-35131 Padova, Italy; E-mail: baggio@dei.unipd.it . Seminar held on May 31st, 2017.

can be reliably propagated through a communication system. By taking into account the phenomenon of inter-symbol interference and by considering the Gaussian noise case, we derive an analytic expression for the Shannon's information capacity of a neuronal network. In the last part of the note, we illustrate how the structure of the synaptic connectivity matrix affects the information capacity of the network. In particular, we compare the classes of normal and non-normal connectivity matrices in terms of information capacity and optimal information transmission sampling time.

In order to make the note introductory in nature and as self-contained as possible, a substantial part of the present note will be devoted to review notions and results from neuronal modelling and information theory.

Notation. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, we denote by $\boldsymbol{A}^\top$ its transpose. $\boldsymbol{A}$ is said to be *normal* if $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{A}^\top\boldsymbol{A}$. Otherwise, $\boldsymbol{A}$ is said to be *non-normal*. Examples of normal matrices include symmetric, orthogonal, and skew-symmetric matrices. We denote by $\sigma(\boldsymbol{A})$, $\mathrm{tr}(A)$, $\det(A)$, and $\|\boldsymbol{A}\|_F$ the spectrum, trace, determinant, and Frobenius norm of $\boldsymbol{A}$, respectively. It is well-known that a normal matrix can be unitarily diagonalized, i.e., there exists a unitary matrix $\boldsymbol{U} \in \mathbb{C}^{N\times N}$, s.t. $\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U} = \boldsymbol{D}$, where $\boldsymbol{D}$ is a diagonal matrix with the eigenvalues of $\boldsymbol{A}$ in its diagonal and $\boldsymbol{U}^*$ denotes the conjugate transpose of $\boldsymbol{D}$.

If $\boldsymbol{A}$ is non-normal then it is possible to reduce it to a lower triangular form via unitary transformations, i.e., there exists a unitary matrix $\boldsymbol{U} \in \mathbb{C}^{N\times N}$ s.t. $\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U} = \boldsymbol{T}$ with $\boldsymbol{T} \in \mathbb{C}^{N\times N}$ being a lower triangular matrix with the eigenvalues of $\boldsymbol{A}$ on its diagonal. This form is known as the *Schur form* of $\boldsymbol{A}$. Finally, $\boldsymbol{A}$ is said to be *(Hurwitz) stable* if $\mathrm{Re}\,\lambda < 0$ for every $\lambda \in \sigma(\boldsymbol{A})$.

2 Neuronal network dynamics

In this section, we describe the dynamical network model that we will consider in the communication framework presented in the next section. We start by reviewing a simplified single neuron dynamical model and then we move to network models. The material of this section is mainly taken from [9, Ch. 14], [5, Ch. 5,7], [6].

2.1 Single neuron leaky integrate-and-fire model

The leaky integrate-and-fire (LIF) model is a simplified model describing the dynamics of the membrane potential of a neuron, that is, the difference in electrical potential between the interior of a neuron and the surrounding extracellular medium. In its basic form this model was proposed by Lapique at the beginning of the 1900s [10].

We denote by $V(t)$ the momentary value of the membrane potential of the neuron, by V_{rest} the resting potential of the neuron (which is approximately equal to -70 mV), by $I_{\mathrm{in}}(t)$ the synaptic input current provided by other neurons, and by $I_{\mathrm{out}}(t)$ the output current generated by the neuron. The LIF model behaves like an electrical circuit consisting of a resistor R and a capacitor C in parallel (see Fig. 1). In this model, whenever the integrated membrane potential crosses a prefixed threshold, denoted by V_{th}, a spike is generated in the output current and $V(t)$ is reset to a fixed value V_{reset}. The dynamics of

the membrane potential is given by

$$(1) \qquad \begin{cases} \tau_m \dot{V}(t) = -(V(t) - V_{\text{rest}}) + RI_{\text{in}}(t), & \text{if } V(t) < V_{\text{th}}, \\ V(t) = V_{\text{reset}}, & \text{if } V(t) \geq V_{\text{th}}, \end{cases}$$

where $\tau_m := RC$ is called the *membrane time constant* of the neuron.

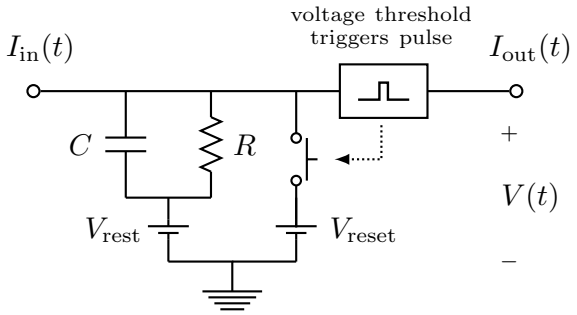


Figure 1. Electrical circuit describing the LIF model.

The output current is modelled as a spike train signal

$$I_{\text{out}}(t) = \sum_i \delta(t - t_i),$$

where $\delta(\cdot)$ denotes the Dirac's delta function and the t_i's are the time instants corresponding to membrane potential resets. Another quantity of interest arising from the LIF model is the so-called *firing rate* signal, which is defined as the trial and temporal averaged number of spikes within the time window $[t - \Delta T, t]$, $\Delta T > 0$, namely

$$(2) \qquad r(t) := \frac{1}{n} \sum_{j=1}^{n} \frac{1}{\Delta T} \int_{t-\Delta T}^{t} I_{j,\text{out}}(\tau)\, \mathrm{d}\tau,$$

where the sum in j is executed over n identical output current trials $\{I_{j,\text{out}}\}_{j=1}^{n}$.

In case of a constant input current $I_{\text{in}}(t) = I_0 = \text{const.}$, the interspike interval T_{th}, which is defined as the time it takes for the membrane potential to reach the fixed threshold V_{th}, can be analytically computed as a function of I_0 via (1),

$$\begin{cases} T_{\text{th}} = \tau_m \ln\left(\frac{RI_0 + V_{\text{rest}} - V_{\text{reset}}}{RI_0 + V_{\text{rest}} - V_{\text{th}}}\right), & \text{if } RI_0 > V_{\text{th}} - V_{\text{rest}}, \\ T_{\text{th}} = 0, & \text{otherwise.} \end{cases}$$

Using the latter equation and considering the approximation $r \approx 1/T_{\text{th}}$ obtained by choosing in (2) a time window ΔT large enough, the firing rate behavior is described by the expression

$$(3) \qquad r = \left(\tau_m \ln\left(\frac{RI_0 + V_{\text{rest}} - V_{\text{reset}}}{RI_0 + V_{\text{rest}} - V_{\text{th}}}\right)\right)^{-1} =: f(I_0).$$

This expression is valid if $RI_0 > V_{\text{th}} - V_{\text{rest}}$, otherwise $r(t) = 0$. The function $f(\cdot)$ introduced above is called *activation function* and relates the synaptic input current to the firing rate of the neuron in the steady-state regime. It is worth noticing that for sufficiently large values of I_0, $f(I_0)$ can be approximated by the linear function[(1)]

$$f(I_0) \approx \alpha + \beta I_0, \tag{4}$$

with $\alpha := \frac{V_{\text{rest}} - V_{\text{th}}}{\tau_m (V_{\text{th}} - V_{\text{reset}})}$ and $\beta := \frac{R}{\tau_m (V_{\text{th}} - V_{\text{reset}})}$.

In case of a time-varying input current $I_{\text{in}}(t)$, the firing rate evolution is often modelled as a "low-pass" filtered version of the steady-state firing rate via the dynamical equation

$$\tau_r \dot{r}(t) = -r(t) + f(I_{\text{in}}(t)), \tag{5}$$

where τ_r is a suitable parameter related to properties of the neuronal spike train (see [1] for further details).

2.2 Network firing rate model

Consider a network of N neurons, whose firing rates obey the dynamics in (5). Let us define the vector of firing rates by $\boldsymbol{r}(t) := [r_1(t), r_2(t), \dots, r_N(t)]^\top$, the vector of input currents by $\boldsymbol{I}_{\text{in}}(t) := [I_{\text{in},1}(t), I_{\text{in},2}(t), \dots, I_{\text{in},N}(t)]^\top$ and $\boldsymbol{f}(\cdot)$ as the function $f(\cdot)$ in (3) acting element-wise on a given vector. The input current vector $\boldsymbol{I}_{\text{in}}(t)$ is made of the sum of two contributions, namely

$$\boldsymbol{I}_{\text{in}}(t) = \boldsymbol{W}\boldsymbol{I}_{\text{out}}(t) + \boldsymbol{I}_{\text{ext}}(t),$$

where $\boldsymbol{W} \in \mathbb{R}^{N \times N}$ denotes the *synaptic connectivity matrix* of the network, $\boldsymbol{I}_{\text{out}}(t) := [I_{\text{out},1}(t), I_{\text{out},2}(t), \dots, I_{\text{out},N}(t)]^\top$ is the vector of output currents generated by the neurons of the network, while $\boldsymbol{I}_{\text{ext}}(t)$ is vector containing the currents provided by neurons external to the network.

The collective firing rate dynamics can be described in matrix form as

$$\begin{aligned} \tau_r \dot{\boldsymbol{r}}(t) &= -\boldsymbol{r}(t) + \boldsymbol{f}(\boldsymbol{I}_{\text{in}}(t)) \\ &= -\boldsymbol{r}(t) + \boldsymbol{f}(\boldsymbol{W}\boldsymbol{I}_{\text{out}}(t) + \boldsymbol{I}_{\text{ext}}(t)). \end{aligned}$$

By considering a network composed by a large number of neurons featuring slow synaptic or membrane dynamics, one can replace, under certain assumptions on the trial-to-trial variability in the spike sequences, the signal $\boldsymbol{I}_{\text{out}}(t)$ by its trial and temporal average, that is, $\boldsymbol{r}(t) \approx \boldsymbol{I}_{\text{out}}(t)$ (see [5, Chap. 7.2] for further details). Using this approximation, we end up with the following network firing rate model

$$\tau_r \dot{\boldsymbol{r}}(t) = -\boldsymbol{r}(t) + \boldsymbol{f}(\boldsymbol{W}\boldsymbol{r}(t) + \boldsymbol{I}_{\text{ext}}(t)). \tag{6}$$

Now, consider an equilibrium point $(\bar{\boldsymbol{r}}, \bar{\boldsymbol{I}}_{\text{ext}})$ of equation (6). Assuming that the entries of $\boldsymbol{W}\bar{\boldsymbol{r}} + \bar{\boldsymbol{I}}_{\text{ext}}$ are sufficiently large, we can exploit the linear approximation (4) of $\boldsymbol{f}(\cdot)$.

[(1)] Here we used the approximation of the logarithm $\ln(1 + z) \approx z$ for small z.

Hence, by defining $\tilde{\boldsymbol{r}}(t) := \boldsymbol{r}(t) - \bar{\boldsymbol{r}}$ and $\tilde{\boldsymbol{I}}_{\text{ext}}(t) := \boldsymbol{I}_{\text{ext}}(t) - \bar{\boldsymbol{I}}_{\text{ext}}$, the linearization of (6) around the equilibrium point $(\bar{\boldsymbol{r}}, \bar{\boldsymbol{I}}_{\text{ext}})$ yields

$$\tau_r \dot{\tilde{\boldsymbol{r}}}(t) = -\tilde{\boldsymbol{r}}(t) + \beta \boldsymbol{W} \tilde{\boldsymbol{r}}(t) + \beta \tilde{\boldsymbol{I}}_{\text{ext}}(t). \tag{7}$$

For later convenience, we rename $\tilde{\boldsymbol{r}}(t)$, $\frac{\beta}{\tau_r}\tilde{\boldsymbol{I}}_{\text{ext}}(t)$ to $\boldsymbol{x}(t)$, $\boldsymbol{u}(t)$, respectively, and we define $\boldsymbol{A} := \frac{1}{\tau_r}(\beta \boldsymbol{W} - \boldsymbol{I})$, being $\boldsymbol{I}$ the $N \times N$ identity matrix. In this way, Equation (7) becomes

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{u}(t). \tag{8}$$

3 Modelling and measuring information transmission in neuronal networks

In this section, we present a digital communication framework for information transmission in neuronal networks driven by the linearized firing rate dynamics (8). Before doing so, we review some notions of information and communication theory. The review part is mainly based on [3, Chap. 7–9].

3.1 Communication channel and information capacity

A digital communication system can be schematically represented as in Fig. 2.

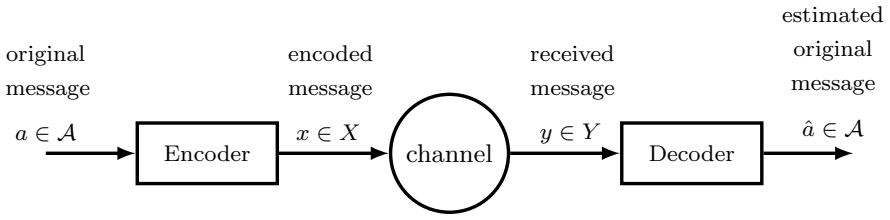


Figure 2. Schematic representation of a digital communication system.

In this scheme, a to-be-transmitted message a belonging to a finite cardinality alphabet $\mathcal{A}$ is first encoded by an encoder into a message $x \in X$ and then transmitted via a physical *communication channel*. At the receiving end, the received message $y \in Y$ is first decoded in the decoder, and then the transmitted message is recovered as an estimate $\hat{a} \in \mathcal{A}$.

Mathematically, the communication channel is defined by the *transition probability* $p(y|x)$, that expresses the probability of observing the output symbol y given that we send the symbol x. The channel is said to be *memoryless* if the probability distribution of the output depends only on the input at that time and is conditionally independent of previous channel inputs or outputs.

Definition 1 The information capacity of a memoryless communication channel is defined as

$$\mathcal{C} = \max_{p(x)} \mathcal{I}(X;Y)$$

where $\mathcal{I}(X;Y)$ denotes the mutual information[(2)] between random variables X, Y, and the maximum is taken over over all possible input distributions $p(x)$.

The notion of information capacity was introduced by Shannon in the late 1940s [14]. Remarkably, Shannon showed that the notion of channel capacity introduced in Definition 1 coincides with the highest rate in bits per channel use at which information can be sent with arbitrarily low probability of error.

One of the most important and used memoryless communication channels is the *Additive White Gaussian Noise (AWGN) channel*. In the scalar case, at a given transmission time $t \in \mathbb{Z}$, this channel is described by an output $\boldsymbol{y}(t)$ which is generated by adding a noise term $\boldsymbol{n}(t)$ drawn from a zero-mean Gaussian distribution with variance σ_n^2, i.e. $\boldsymbol{n}(t) \sim \mathcal{N}(0, \sigma_n^2)$, to the input $\boldsymbol{x}(t)$. Further, the sequence of noise terms is independent and identically distributed in time, i.e. $\mathbb{E}[\boldsymbol{n}(t)\boldsymbol{n}(s)] = \sigma_n^2 \delta(t-s)$, and $\boldsymbol{n}(t)$ is assumed to be uncorrelated with $\boldsymbol{x}(t)$ at each time $t \in \mathbb{Z}$.

The capacity of the AWGN channel is infinite unless we add a "power" constraint on the input variance, i.e., we require that $\mathbb{E}[\boldsymbol{x}(t)^2] \leq P$, $P > 0$. Under this constraint, the information capacity of the scalar AWGN channel can be analytically computed and has the form

$$\mathcal{C} = \frac{1}{2} \log_2 \left(1 + \frac{P}{\sigma_n^2} \right),$$

where the ratio P/σ_n^2 is usually known as the *signal-to-noise ratio (SNR)*. In the vector zero-mean Gaussian case, it can be shown that the capacity of the AWGN channel with power constraint $\operatorname{tr} \mathbb{E}[\boldsymbol{x}(t)\boldsymbol{x}^\top(t)] \leq P$, $P > 0$, is given by

$$\mathcal{C} = \frac{1}{2} \max_{\operatorname{tr} \boldsymbol{\Sigma}_x \leq P} \log_2 \frac{\det(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_n)}{\det \boldsymbol{\Sigma}_n}, \tag{9}$$

where we defined $\boldsymbol{\Sigma}_x := \mathbb{E}[\boldsymbol{x}(t)\boldsymbol{x}^\top(t)]$ and $\boldsymbol{\Sigma}_n := \mathbb{E}[\boldsymbol{n}(t)\boldsymbol{n}^\top(t)]$.

3.2 Transmitting information via linear dynamical networks

We describe here a digital communication framework for analyzing and measuring the amount of "information" that can be reliably propagated in a neuronal network driven by the linear firing-rate dynamics derived in (8). The framework we are going to present has

[(2)] For discrete random variables X, Y the mutual information is defined as

$$\mathcal{I}(X;Y) := \sum_{y \in Y} \sum_{x \in X} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right),$$

where $p(x,y)$ is the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y, respectively. For continuous random variables X, Y the mutual information is defined as

$$\mathcal{I}(X;Y) := \int_Y \int_X p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right) \mathrm{d}x \, \mathrm{d}y,$$

where $p(x,y)$ is now the joint probability density function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability density functions of X and Y, respectively.

been inspired by the recent work [7] and is schematically reported in the block diagram of Fig. 3.

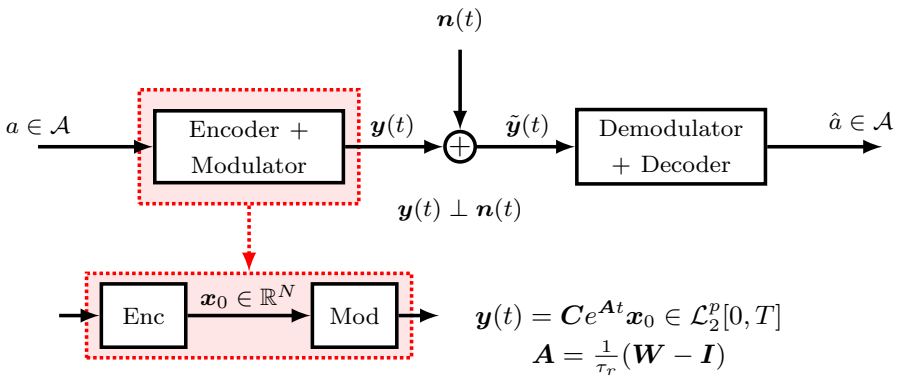


Figure 3. Block diagram of the digital communication protocol.

We assume that a symbol a belonging to a finite cardinality alphabet $\mathcal{A}$ is transmitted at the source. The information contained in this symbol is then encoded in a vector of firing rates $\boldsymbol{x}_0 \in \mathbb{R}^N$, which is assumed to be drawn from a zero-mean Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}$, i.e., $\boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$.

This vector is thought of as the initial condition for the evolution of the linear system (8) with no input term and output matrix $\boldsymbol{C} \in \mathbb{R}^{p \times N}$, namely

$$\begin{cases} \dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t), \\ \boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{x}(t), \end{cases} \quad \boldsymbol{x}_0 \in \mathbb{R}^N, \ t \geq 0.$$

This step represents the *modulation* stage of the communication protocol. A modulator is a function mapping a discrete symbol to a continuous-time trajectory. In our case, assuming that each transmission is performed within a time interval $[0, T]$, $T > 0$, this mapping is given by

$$\begin{aligned} \text{Mod}\colon \mathbb{R}^N &\to \mathcal{L}_2^p[0,T] \\ \boldsymbol{x}_0 &\mapsto \boldsymbol{C}e^{\boldsymbol{A}t}\boldsymbol{x}_0, \ t \in [0,T], \end{aligned}$$

where $\mathcal{L}_2^p[0,T]$ denotes the Hilbert space of p-dimensional square integrable signals in the time interval $[0,T]$ equipped with the inner product $\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{L}_2} := \int_0^T \boldsymbol{f}^\top(t)\boldsymbol{g}(t)\,\mathrm{d}t$, for $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{L}_2^p[0,T]$.

After this stage, the modulated signal $\boldsymbol{y}(t)$ goes through a vector AWGN channel, which gives the corrupted output $\tilde{\boldsymbol{y}}(t)$. Eventually, a demodulation and decoding stage produces an estimate of the original transmitted symbol $\hat{a} \in \mathcal{A}$.

Now, we assume that at time $t = kT$, with $k \in \mathbb{Z}$, a new transmission is done by generating the signal

$$\boldsymbol{y}_k(t) = \boldsymbol{C}e^{\boldsymbol{A}(t-kT)}\boldsymbol{x}(kT)\boldsymbol{1}(t - kT),$$

which is non-zero only for $t \geq kT$.[3] We consider, without loss of generality, the case $k = 0$. In this case, the total information signal will be the superposition of the $\boldsymbol{y}_k(t)$, namely $\sum_{k\in\mathbb{Z}} \boldsymbol{y}_k(t)$, while the signal that contains the "useful" information is given by $\boldsymbol{y}_0(t)$. The covariance of $\boldsymbol{y}_0(t)$ is

$$\boldsymbol{\Sigma}_{y_0} := \mathbb{E}[\boldsymbol{y}_0(t)\boldsymbol{y}_0^\top(t)] = \boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}^\top, \tag{10}$$

where $\boldsymbol{F} \in \mathbb{R}^{N\times N}$ is such that $\boldsymbol{F}^\top\boldsymbol{F} = \int_0^T e^{\boldsymbol{A}^\top t}\boldsymbol{C}^\top\boldsymbol{C}e^{\boldsymbol{A}t}\mathrm{d}t =: \boldsymbol{\mathcal{O}}$. The noise $\boldsymbol{n}(t)$ is modelled as the sum of two terms, namely

$$\boldsymbol{n}(t) = \boldsymbol{r}(t) + \boldsymbol{i}(t),$$

where

i) $\boldsymbol{r}(t)$ represents a zero-mean background white noise term that satisfies $\mathbb{E}[\boldsymbol{r}(t)\boldsymbol{r}^\top(s)] = \sigma_r^2\boldsymbol{I}\delta(t-s)$, for all $t, s \in \mathbb{R}$, and

ii) $\boldsymbol{i}(t) := \sum_{k=1}^{\infty} \boldsymbol{C}e^{\boldsymbol{A}(t+kT)}\boldsymbol{x}(-kT)$, this term is made of the sum of contributions due to previous transmissions at time $-kT$, $k \in \mathbb{Z}$, $k > 1$. In communication theory, this term is known under the name of *inter-symbol interference*.

Supposing that $\mathbb{E}[\boldsymbol{x}(k_1T)\boldsymbol{x}^\top(k_2T)] = \boldsymbol{\Sigma}\,\delta(k_1 - k_2)$, for all $k_1, k_2 \in \mathbb{Z}$, that is the input code statistics is identically distributed at every transmission and independent of past and future transmissions, it can be shown that

$$\boldsymbol{\Sigma}_i := \mathbb{E}[\boldsymbol{i}(t)\boldsymbol{i}^\top(t)] = \boldsymbol{F}(\boldsymbol{\mathcal{W}} - \boldsymbol{\Sigma})\boldsymbol{F}^\top,$$

where we have defined $\boldsymbol{\mathcal{W}} := \sum_{k=0}^{\infty} e^{\boldsymbol{A}kT}\boldsymbol{\Sigma}e^{\boldsymbol{A}^\top kT}$. Finally, by assuming that, for every $k \in \mathbb{Z}$ and $t \in \mathbb{R}$, $\boldsymbol{x}(kT)$ and $\boldsymbol{r}(t)$ are independent random variables, the overall noise covariance takes the form

$$\boldsymbol{\Sigma}_n := \mathbb{E}[\boldsymbol{n}(t)\boldsymbol{n}^\top(t)] = \boldsymbol{\Sigma}_i + \sigma_r^2\boldsymbol{I}. \tag{11}$$

By taking into account the expression of the channel capacity for vector AWGN channel (9) and the covariance of the useful signal (10) and noise (11), we arrive at the following result.

Proposition 1 *The information capacity of the above introduced channel under power constraint* $\operatorname{tr}\boldsymbol{\Sigma} \leq P$, $P > 0$, *is given by*

$$\mathcal{C}_T = \frac{1}{2}\max_{\operatorname{tr}\boldsymbol{\Sigma}\leq P} \log_2 \frac{\det(\boldsymbol{\mathcal{O}}\boldsymbol{\mathcal{W}} + \sigma_r^2\boldsymbol{I})}{\det(\boldsymbol{\mathcal{O}}(\boldsymbol{\mathcal{W}} - \boldsymbol{\Sigma}) + \sigma_r^2\boldsymbol{I})}. \tag{12}$$

Notice that we added a subscript T to the capacity in (12) in order to stress its dependence on the time T. The latter corresponds to the data transmission sampling

[3] Here $\mathbf{1}(t)$ denotes the step function, i.e., $\mathbf{1}(t) = 0$ for $t < 0$ and $\mathbf{1}(t) = 1$ for $t \geq 1$.

time of the communication system. Moreover, in the jargon of linear system theory, $\boldsymbol{\mathcal{O}}$ is called finite-horizon continuous-time *observability Gramian* of the pair $(\boldsymbol{A}, \boldsymbol{C})$ while $\boldsymbol{\mathcal{W}}$ infinite-horizon discrete-time *controllability Gramian* of the pair $(e^{\boldsymbol{A}T}, \boldsymbol{\Sigma}^{1/2})$. The first one is related to the amount of energy generated in the output response of an autonomous linear system in the time $[0, T]$, while the second one quantifies the energy needed by a linear system to steer the initial state to a final target one (see e.g. [8]).

To conclude this section, we observe that information capacity is measured in bits per channel use. In order to express the capacity w.r.t. a time unit of measure, we introduce the following information transmission metric

$$\mathcal{R}_T := \frac{1}{T}\mathcal{C}_T = \frac{1}{2T} \max_{\operatorname{tr}\boldsymbol{\Sigma} \leq P} \log_2 \frac{\det(\boldsymbol{\mathcal{O}}\boldsymbol{\mathcal{W}} + \sigma_r^2 \boldsymbol{I})}{\det(\boldsymbol{\mathcal{O}}(\boldsymbol{\mathcal{W}} - \boldsymbol{\Sigma}) + \sigma_r^2 \boldsymbol{I})}. \tag{13}$$

We call this metric *information transmission rate*. If T is measured in seconds, $\mathcal{R}_T$ measures the highest information rate in bits per second at which information can be reliably (i.e., with arbitrarily low error probability) sent through the channel within the time interval $[0, T]$.

4 The role of connectivity structure in information transmission

In this section, we investigate how the structure of the connectivity matrix $\boldsymbol{A}$ affects the information capacity and rate (12)–(13). We will start by analyzing the information capacity and rate of a single neuron, for which the network architecture does not matter. Next, we study the behavior of information capacity and rate for the classes of normal matrices. Finally, we numerically compute the optimal connectivity structure w.r.t. information rate, for different values of the sampling time T.

4.1 Single neuron information capacity and rate

The information capacity and rate derived in the previous section in Equations (12)–(13) can be computed in closed-form for the case of a single neuron, i.e., for $\boldsymbol{A} = -a$, $a \in \mathbb{R}$, $a > 0$. These have the form

$$\mathcal{C}_T = \frac{1}{2} \log_2 \left(\frac{\mathrm{SNR} + 2a}{\mathrm{SNR}\, e^{-2aT} + 2a} \right), \tag{14}$$

$$\mathcal{R}_T = \frac{1}{2T} \log_2 \left(\frac{\mathrm{SNR} + 2a}{\mathrm{SNR}\, e^{-2aT} + 2a} \right), \tag{15}$$

where we defined the signal-to-noise ratio $\mathrm{SNR} := P/\sigma_r^2$.

Fig. 4 shows some plots of these $\mathcal{C}_T$ and $\mathcal{R}_T$ as functions of $T \geq 0$, for different values of a and SNR. From expressions (14)–(15) and the plots in Fig. 4, it is interesting to note that:

i) $\mathcal{C}_T$ is an increasing function of $T \geq 0$, starting from zero and tending to a constant value as $T \to \infty$. Further, the larger SNR and/or the smaller a, the larger the

steady-state capacity

$$\mathcal{C}_\infty = \lim_{T\to\infty} \frac{1}{2}\log_2\left(\frac{\mathrm{SNR}+2a}{\mathrm{SNR}\,e^{-2aT}+2a}\right) = \frac{1}{2}\log_2\left(1+\frac{\mathrm{SNR}}{2a}\right).$$

ii) $\mathcal{R}_T$ is a decreasing function of $T \geq 0$, which tends to zero asymptotically. Hence, $\mathcal{R}_T$ attains its maximum value at $T = 0$; the latter is given by

$$\mathcal{R}_0 = \lim_{T\to 0} \frac{1}{2T}\log_2\left(\frac{\mathrm{SNR}+2a}{\mathrm{SNR}\,e^{-2aT}+2a}\right) = \frac{1}{\ln 2}\frac{a\mathrm{SNR}}{\mathrm{SNR}+2a}.$$

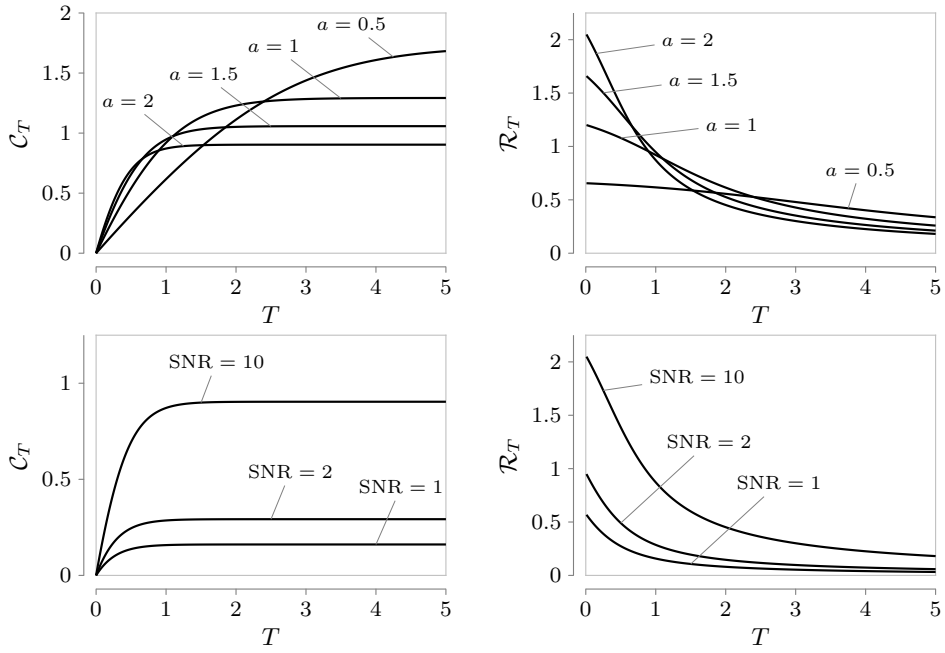


Figure 4. Information capacity (14) and rate (15) for single neuron as a function of the sampling time T. In the top plots: SNR = 10. In the bottom plots: $a = 2$.

4.2 Information capacity and rate of normal networks

If $\boldsymbol{A}$ is stable and normal and $\boldsymbol{C} = \boldsymbol{I}$ (i.e., information is transmitted to all neurons of the network), then it is possible to show that the optimal $\boldsymbol{\Sigma}$ which maximizes (12)–(13) can be always taken to be diagonal. This fact yields the following result.

Proposition 2 *If $\boldsymbol{C} = \boldsymbol{I}$ and $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is a normal and stable matrix with eigenvalues $\{\lambda_i\}_{i=1}^N$, then*

$$\mathcal{C}_T = \max_{\substack{\{P_i\}_{i=1}^N \text{ s.t.} \\ P_i\geq 0,\ \sum_{i=1}^N P_i \leq P}} \sum_{i=1}^N \mathcal{C}_{T,i}, \tag{16}$$

where

$$\mathcal{C}_{T,i} := \frac{1}{2}\log_2 \frac{\mathrm{SNR}_i - 2\mathrm{Re}\,\lambda_i}{\mathrm{SNR}_i\, e^{2T\mathrm{Re}\,\lambda_i} - 2\mathrm{Re}\,\lambda_i}, \quad i = 1, 2, \ldots, N, \tag{17}$$

with $\mathrm{SNR}_i := P_i/\sigma_r^2$. *Similarly, under the previous assumptions,*

$$\mathcal{R}_T = \max_{\substack{\{P_i\}_{i=1}^N \text{ s.t.} \\ P_i \geq 0,\ \sum_{i=1}^N P_i \leq P}} \sum_{i=1}^N \mathcal{R}_{T,i}, \tag{18}$$

where

$$\mathcal{R}_{T,i} := \frac{1}{2T}\log_2 \frac{\mathrm{SNR}_i - 2\mathrm{Re}\,\lambda_i}{\mathrm{SNR}_i\, e^{2T\mathrm{Re}\,\lambda_i} - 2\mathrm{Re}\,\lambda_i}, \quad i = 1, 2, \ldots, N. \tag{19}$$

Proposition 2 shows that, in terms of information capacity and rate, an $N \times N$ normal communication network is equivalent to N independent scalar communication channels. Indeed, the terms in Eqs. (17)–(19) correspond to the capacity and rate, respectively, of a single neuron, as defined in Eqs. (14)–(15), for $a = -2\mathrm{Re}\,\lambda_i$.

Moreover, since in the scalar case $\mathcal{R}_T$, is a decreasing function of T, the same holds in the normal matrix case. Indeed, as shown in Eq. (18), in this case the rate is equal to the sum, optimized over the scalar powers P_i's, of N independent scalar channels. In light of this fact, if we define the *optimal sampling time* as

$$T^\star := \arg\max_{T \geq 0} \mathcal{R}_T, \tag{20}$$

it follows that for normal networks we have $T^\star = 0$. In practice, this means that in order to maximize the information rate, a neuronal network driven by a normal connectivity matrix must sample data very fast.

Fig. 5 shows the behavior, as a function of $T \geq 0$, of the information rate for a chain network of N nodes. In this case, $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ corresponds to a stable Toeplitz tridiagonal matrix. In the top plots, $\boldsymbol{A}$ is taken to be symmetric (and, hence, normal) of the form

$$\boldsymbol{A} = \begin{bmatrix} \gamma & \alpha & 0 & \ldots & 0 \\ \alpha & \gamma & \alpha & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \alpha & \gamma & \alpha \\ 0 & \cdots & 0 & \alpha & \gamma \end{bmatrix}, \quad \alpha \geq 0,\ \gamma \leq 0.$$

From these plots, it can be noticed that $\mathcal{R}_T$ is a decreasing function of T and the closer the spectrum of $\boldsymbol{A}$ to zero, the better $\mathcal{R}_T$ is (for sufficiently large T's). In the bottom

plots, we considered a non-normal Toeplitz chain network

$$\boldsymbol{A} = \begin{bmatrix} \gamma & \alpha/\beta & 0 & \dots & 0 \\ \beta\alpha & \gamma & \alpha/\beta & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \beta\alpha & \gamma & \alpha/\beta \\ 0 & \cdots & 0 & \beta\alpha & \gamma \end{bmatrix}, \quad \alpha \geq 0,\ \beta \geq 1,\ \gamma \leq 0.$$

where the parameter β quantifies, in a sense, the amount of "non-normality" of $\boldsymbol{A}$. From these plots, it can be seen that $\mathcal{R}_T$ exhibits a peak. Furthermore, fixing the spectrum of $\boldsymbol{A}$, the larger β, the larger this peak is. This simple example reveals that, in principle, for non-normal networks the optimal sampling time as defined in (20) can be different from zero.

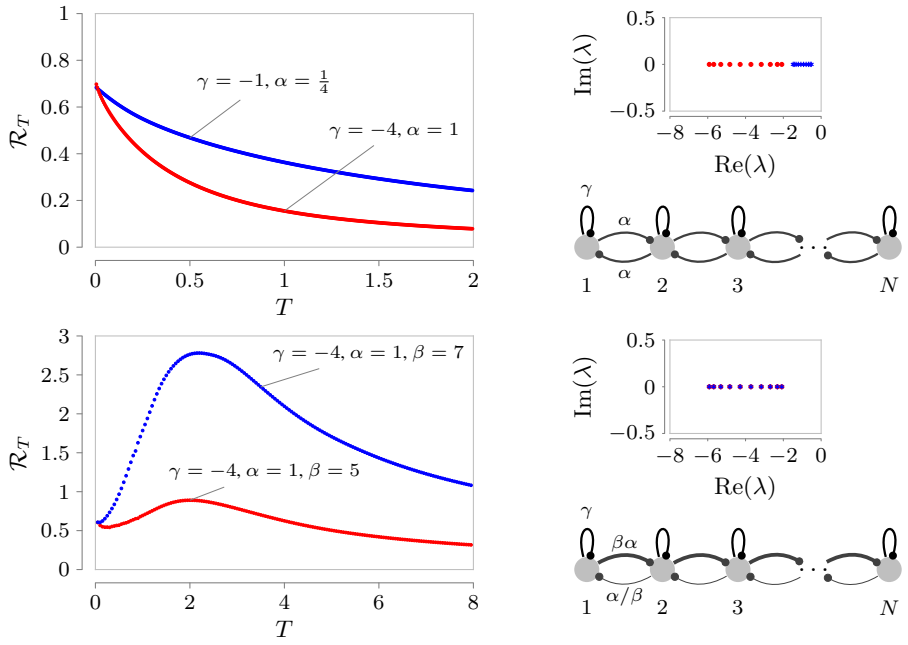


Figure 5. Information capacity (14) and rate (15) as a function of the sampling time T for a Toeplitz chain network of cardinality $N = 10$ and SNR $= P/\sigma_r^2 = 1$. In the small plots on the right, the spectrum of $\boldsymbol{A}$ is plotted for the values of parameters α, β, γ shown in the corresponding left plots.

4.3 Network architectures maximizing information rate

To conclude this section, we investigate the problem of finding the network architecture that maximizes the information rate (13) for a given sampling time $T > 0$ and $\boldsymbol{C} = \boldsymbol{I}$. The latter problem can be formally stated as follows

$$\boldsymbol{A}^\star := \arg\max_{\boldsymbol{A} \in \mathcal{A}} \mathcal{R}_T(\boldsymbol{A}), \tag{21}$$

where we denoted by $\mathcal{A} := \{\boldsymbol{A} \in \mathbb{R}^{N\times N} : \boldsymbol{A} \text{ stable}\}$ the set of stable connectivity matrices and we made explicit the fact that $\mathcal{R}_T$ depends on $\boldsymbol{A}$ using the notation $\mathcal{R}_T(\boldsymbol{A})$. In what follows, we will restrict the attention to the subset $\mathcal{A}_r \subset \mathcal{A}$ consisting of stable matrices $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ with real eigenvalues. The following lemma can be proved.

Lemma 1 *Let $\boldsymbol{C} = \boldsymbol{I}$ and $\boldsymbol{A} \in \mathcal{A}_r$. Let $\boldsymbol{T} \in \mathbb{R}^{N\times N}$ be the lower triangular Schur form of $\boldsymbol{A}$. Then*

$$\mathcal{R}_T(\boldsymbol{A}) = \mathcal{R}_T(\boldsymbol{T}).$$

In view of Lemma 1, without loss of generality, we can consider the Schur form of matrices in $\mathcal{A}_r$, and, therefore, reduce problem (21) to an optimization problem over the set of lower triangular connectivity matrices with negative diagonal entries. We denote the latter set by $\mathcal{A}_{S,r}$. In view of this fact, the new "simplified" optimization problem reads as

$$\boldsymbol{A}^\star := \arg \max_{\boldsymbol{A}\in\mathcal{A}_{S,r}} \mathcal{R}_T(\boldsymbol{A}) - \varepsilon\|\boldsymbol{A}\|_F, \tag{22}$$

where we added a regularization term $\varepsilon\|\boldsymbol{A}\|_F$, $\varepsilon > 0$, to the objective function $\mathcal{R}_T(\boldsymbol{A})$ in order to bound the entries of $\boldsymbol{A}$. Recall that the computation of the information rate $\mathcal{R}_T$ requires to solve a constrained optimization problem over input covariances $\boldsymbol{\Sigma}$'s. Consequently, in order to numerically solve the optimization problem (22), we exploited a *coordinate gradient ascent* strategy [12, Ch. 9] over trace-constrained covariance matrices $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$, $\boldsymbol{\Sigma} > 0$, $\operatorname{tr}\boldsymbol{\Sigma} = P$, and connectivity matrices belonging to $\mathcal{A}_{S,r}$. Numerical solutions of the optimization problem (22) are shown in Fig. 6, for a network of $N = 10$ neurons and different values of $T > 0$. From these plots, it is worth noting that

i) for small values of T, the optimal structure $\boldsymbol{A}^\star$ is diagonal and, therefore, normal.

ii) as long as T increases, the strictly lower triangular terms in $\boldsymbol{A}^\star$ becomes different from zero, yielding a non-normal optimal network structure. In particular, observe that, for T large enough, the entries in the main subdiagonal of the optimal matrix $\boldsymbol{A}^\star$ are much greater than the other strictly lower triangular entries of $\boldsymbol{A}^\star$.

To summarize, numerical simulations suggest that non-normality of the connectivity matrix $\boldsymbol{A}$ can be beneficial in terms of information rate $\mathcal{R}_T$, if the neuronal network cannot sample data sufficiently fast. Intuitively, this seems in agreement with the fact that, in contrast with the normal case, trajectories generated by a linear networked dynamical system described by a stable non-normal network matrix can feature a transient amplification and, at the same time, a quick decay to zero after a fixed time instant [16]. In the communication framework presented in Sec. 3, such a behavior can be beneficial since a large transient can help in increasing the signal-to-noise ratio of the channel, while a quick decay to zero can help in reducing the contribution of inter-symbol interference after a fixed positive time.

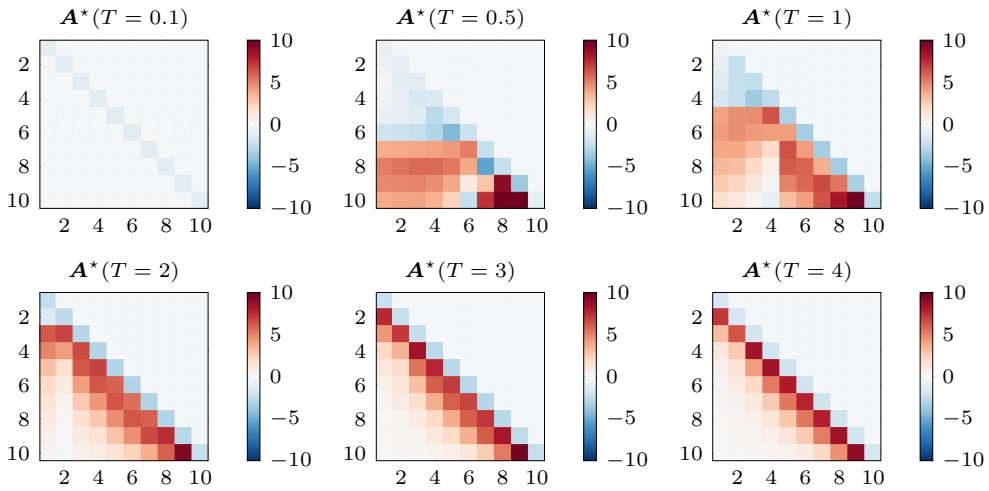


Figure 6. Optimal network architecture $\boldsymbol{A}^\star$ obtained by solving problem (22) for $N = 10$, $\mathrm{SNR} = P/\sigma_r^2 = 1$, $\varepsilon = 0.001$. The solution of problem (22) has been computed via unconstrained coordinate gradient ascent over unit-trace positive definite $\boldsymbol{\Sigma}$ and lower triangular $\boldsymbol{A} \in \mathcal{A}_{S,r}$. The simulations have been carried out in `Python` using `Theano` library [15].

5 Concluding remarks

In this note, we addressed the problem of information transmission in a neuronal network driven by linear dynamics. First, we briefly reviewed a linearized firing rate dynamical model of a neuronal network. Then, we introduced a digital communication framework for the propagation of information in such a linear dynamical neuronal network. Further, we discussed a novel metric for measuring the amount of information that a neuronal network can reliably transmit. Finally, we analyzed the relation between network structure and optimal information transmission rate. In particular, from our analysis, it turns out that a non-normal network architecture should be preferred if the sampling time of neurons in the network is not short enough.

Interesting directions for future work include the analysis of the role of the background noise on the optimal network structure and the extension to the case in which only some neurons of the network can read out transmitted information, i.e., the case $\boldsymbol{C} \neq \boldsymbol{I}$.

References

[1] Abbott, L.F., *Decoding neuronal firing and modelling neural networks.* Quarterly Reviews of Biophysics, 27/03 (1994), 291–331.

[2] Akam, T., Kullmann, D.M., *Oscillations and filtering networks support flexible routing of information.* Neuron, 67/2 (2010), 308–320.

[3] Cover, T.M., Thomas, J.A., "Elements of information theory". John Wiley & Sons, 2012.

[4] Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., Shenoy, K.V., *Neural population dynamics during reaching.* Nature, 487/7405 (2012), 51–56.

[5] Dayan, P., Abbott, L.F., "Theoretical neuroscience". Cambridge, MA: MIT Press, 2001.

[6] Gerstner, W., Kistler, W.M., Naud, R., Paninski, L., "Neuronal dynamics: From single neurons to networks and models of cognition". Cambridge University Press. 2014.

[7] Hennequin, G., Vogels, T.P., Gerstner, W., *Optimal control of transient dynamics in balanced networks supports generation of complex movements.* Neuron, 82/6 (2014), 1394–1406.

[8] Hespanha, J.P., "Linear systems theory". Princeton University Press, 2009.

[9] Koch, C., "Biophysics of computation: information processing in single neurons". Oxford University Press, 2004.

[10] Lapicque, L., *Recherches quatitatives sur l'excitation electrique des nerfs traitee comme polarisation.* J. Physiol. Pathol. Gen. 9 (1907), 620–635.

[11] Laughlin, S.B., Sejnowski, T.J., *Communication in neuronal networks.* Science, 301(/5641 (2003), 1870–1874.

[12] Nocedal, J., Wright, S., "Numerical optimization". Springer-Verlag New York, 2006.

[13] Salinas, E., Sejnowski, T.J., *Correlated neuronal activity and the flow of neural information.* Nature Reviews Neuroscience, 2/8 (2001), 539–550.

[14] Shannon, C.E., *A Mathematical Theory of Communication.* Bell System Technical Journal 27 (1948), 623–656.

[15] Theano Development Team, *Theano: A Python framework for fast computation of mathematical expressions.* arXiv preprint, `arXiv:1605.02688`.

[16] Trefethen, L.N., Embree, M., "Spectra and pseudospectra: the behavior of nonnormal matrices and operators". Princeton University Press, 2005.

[17] Vogels, T.P., Abbott, L.F., *Gating multiple signals through detailed balance of excitation and inhibition in spiking networks.* Nature Neuroscience, 12/4 (2009), 483–491.

[18] Vogels, T.P., Rajan, K., Abbott, L.F., *Neural network dynamics.* Annu. Rev. Neurosci. 28 (2005), 357–376.

Biodiversity: Mathematical Modelling and Statistics

Anna Tovo (*)

Life on Earth is diverse at many levels, beginning with genes and extending to the wealth and complexity of species, life forms, and functional roles.
Identifying and understanding the relationships between all the life on Earth are some of the greatest challenges in science.
Most people recognize biodiversity by species, which are group of living organisms that can interbreed. But actually, biodiversity is more than that.
It is straightforward that ecological systems are characterized by the recurrent emergency of patterns. This suggests that the mechanisms that structure them are insensitive to the details of the system. Such universality motivates the development of mathematical models able to grasp the basic ecological mechanisms at work and to faithfully reproduce the empirical data.
This fascinating intellectual challenge fits perfectly into the world of complex system theory, which teaches us that from interactions among small particles can emerge universal collective interesting behaviours.
When studying an ecological community, a first intuitive biodiversity indicator could be the total number of species S^* we observe in our sample. However, this is clearly a poor indicator.
Much more informative is the so-called *Species-Abundance Distribution* (or SAD), which tells us how rarity and commonness are distributed among species. More precisely, denoting with ϕ_n the number of species having exactly n individuals in our community, a typical representation of the SAD is the so-called Preston Plot, an histogram where the first column accounts for half of the singletons (species with only one individuals), $\phi_1/2$, the second column accounts for $\phi_1/2 + \phi_2/2$, i.e. the other half of the singletons and half of the doubletons, the third column for $\phi_2/2 + \phi_3 + \phi_4/2$ and so on. Generally, in the n^{th} column will fall half of the species with 2^{n-2} individuals, half of those having 2^{n-1} individuals and all the other species having abundance comprised between $2^{n-2}+1$ and $2^{n-1}-1$. In the first panel of Figure 1 we see the SAD of a 50ha sample of Barro Colorado Island rainforest, which is located in the middle of Panama Canal. Global patterns of em-

(*)Ph.D. course, Università di Padova, Dip. Matematica, via Trieste 63, I-35121 Padova, Italy; E-mail: anna.tovo90@gmail.com . Seminar held on June 14th, 2017.

pirical abundance distributions show that tropical forests vary in their absolute number of species but display surprising similarities in the distribution of populations across species. Indeed, different functional forms have been proposed in literature to describe the SAD of an ecosystem, but most of them have been introduced only to provide a good fit of empirical data.

A second pattern is the *Species-Area Relationships* (or SAR), which looks at how biodiversity changes with the sampled area. In particular, given a sub-area A of our ecosystem, $SAR(A)$ gives the expected number of species which will fall within A (see Figure 1, middle panel). There is empirical evidence that such curve shows a tri-phasic behaviour when plotted in a log-log scale. An intuitive explanation is the following. When looking at very small scales, it is likely that increasing even a little the surveyed area will lead to the observation of lots of new species, which are the most abundance ones occupying the vast majority of the ecosystem area. At contrast, when the sample size is big enough, all those species will be already have been accounted, while rare species are the only one missing from the count. The curve then saturates to the total number of species S^* present in our community. In the middle panel of Figure 1 we can see these two phases. If one could go to greater scales, the SAR would start growing up again because of species turnover.

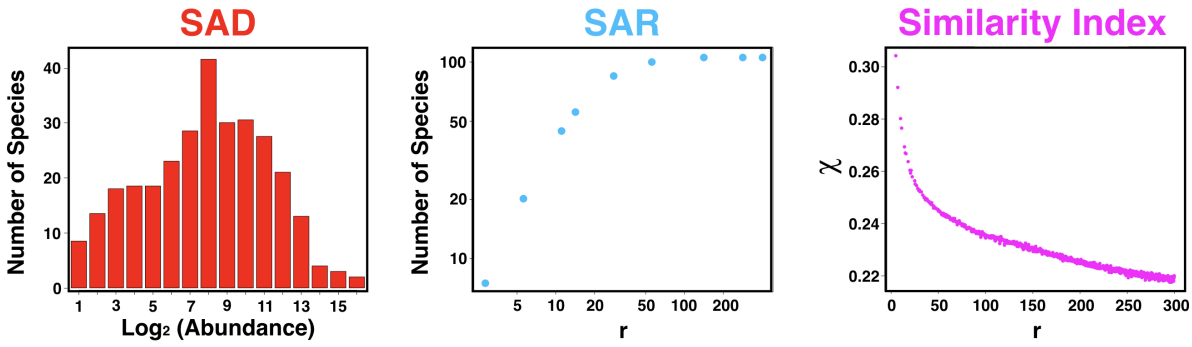


Figure 1. Universal patterns of ecological systems. From left to right: *Species-Abundance Distribution*, *Species-Area Relationship* and *Similarity Index* for a 50ha sample of Barro Colorado Island, in Panama.

Still connected with space, another important ecological pattern is the *Similarity Index*, which gives us information about how similar two sub-samples are in function of their distance. More precisely, given two disjoint sub-regions B_1 and B_2 of the same area A, let us denote with $S(B_1, B_2)$ the number of species which fall both in B_1 and B_2 and with $S(B_i),\ i \in \{1, 2\}$, the number of species residing in B_i, both shared and not. The classic Sørensen index (Chao et al., 2005) is defined as the ratio

$$\sigma(B_1, B_2) = \frac{S(B_1, B_2)}{\frac{1}{2}(S(B_1) + S(B_2))}.$$

We will later see how to reformulate this notion in the context of point processes theory, where it will be useful to define a function $\chi(r) = \sigma(B_1, B_2)/A$ of the distance r between

the two samples and which we will simply call *Similarity Index* (see Figure 1, third panel). The behaviour of all these patterns is surprisingly similar for different ecosystems, meaning that there are universal mechanisms at work. We wish now to link and explain them through mathematical models.

It is instructive to start with an extremely simple model, which was firstly developed by Coleman in 1981 (Coleman, 1981). Let us consider a collection of individuals belonging to S^* different species, each having abundance n_s, $s \in \{1, \dots, S^*\}$ and residing in a spatial region of area A^*. Coleman's model hypothesis is the following: whether or not the location of other individuals has been specified, the probability that a given individual resides in a given subregion of area A within A^* is equal to its relative area A/A^*.

This hypothesis is clearly very strong, since it is reasonable only when the interactions between individuals, whether intra-specific or inter-specific can be neglected and when the surveyed landscape does not present hard inhomogeneities.

In this framework, where no underlying ecological mechanism is driving our system, but only randomness, the probability that at least one individual of the s^{th} species falls within the sub-area A is given by

$$p_s = 1 - \left(1 - \frac{A}{A^*}\right)^{n_s}.$$

Then one can compute the expected number of species residing within the sub-area A by summing up the presence probability of each species:

$$SAR(A) = \sum_{s=1}^{S^*} p_s = S^* - \sum_{s=1}^{S^*} \left(1 - \frac{A}{A^*}\right)^{n_s}$$

We thus have an analytical formula for the Species-Area Relationship.

We tested Coleman's model in reproducing the SAR of the Barro Colorado Island (BCI) rainforest database. We found that, at any spatial scale, the species predicted by the model overestimates the empirical average.

The failure of the random placement model to capture the SAR is a clear indication that ecological patterns are driven by non-trivial mechanisms that still need to be appropriately identified. In particular, we have to model in a more realistic way the location of individuals in our surveyed sample.

Spatial point processes are particularly useful for this goal and they have been widely used in theoretical ecology to describe spatial patterns such as the location of trees or birds. An immediate description of such kind of databases can be given by defining, for every subregion B of the working space, which for our purposes will always be $\mathbb{R}^2$, the counting variable $\mathcal{N}(B)$ which gives the number of points falling within it.

Then the collection of such counting variables $\{\mathcal{N}(B)\}$ contains all information about a point process's realisation, since the locations of its points are uniquely determined by the set of $x \in \mathbb{R}^2$ such that $\mathcal{N}(\{x\}) > 0$. We will call a point process the collection of all random variables $\mathcal{N}(B)$ indexed by subsets B of the $2-$dimensional real space and we will denote it with $\mathbf{X}$. The starting point of exploratory analysis of point patterns is to characterise their first and second moments.

Definition 1 Given $B \subseteq \mathbb{R}^2$, the *intensity measure* $\nu_{\mathbf{X}}$ of $\mathbf{X}$ evaluated in B is

$$\nu_{\mathbf{X}}(B) = \mathbb{E}[\mathcal{N}(B)].$$

Moreover, if there exists a function $\lambda_{\mathbf{X}}$ such that

$$\nu_{\mathbf{X}}(B) = \int_B \lambda_{\mathbf{X}}(x)\mathrm{d}x,$$

then we call $\lambda_{\mathbf{X}}$ the *intensity function* of $\mathbf{X}$.

A local interpretation of the intensity function $\lambda_{\mathbf{X}}$ is the following. Let us consider a ball $B(x, r_x)$ in $\mathbb{R}^2$ centred in x, having radius r_x and infinitesimally small size $\mathrm{d}x$. Then $\lambda_{\mathbf{X}}\mathrm{d}x$ approximates the probability that one point of the process will fall within $B(x, r_x)$:

$$\lambda_{\mathbf{X}}(x)\mathrm{d}x \sim \mathbb{P}(\mathcal{N}_{\mathbf{X}}(B(x, r_x)) > 0).$$

As for the second moments, we firstly remark that if $\mathbf{X}$ is a point process defined on $\mathbb{R}^2$, then $\mathbf{X}\times\mathbf{X}$ is a point process defined on $\mathbb{R}^2\times\mathbb{R}^2$ whose points are all ordered pairs (x_1, x_2) with $x_i \in \mathbf{X}$ for $i = 1,\ 2$. Thus the product $\mathcal{N}_{\mathbf{X}}(B_1) \times \mathcal{N}_{\mathbf{X}}(B_2)$, with $B_1,\ B_2 \in \mathbb{R}^2$ is a random variable counting the number of ordered pairs (x_1, x_2) with $x_1 \in B_1$ and $x_2 \in B_2$. We can then introduce the following definitions.

Definition 2 Given $B_1,\ B_2 \subseteq \mathbb{R}^2$,the *second factorial moment measure* $\nu_{\mathcal{N},[2]}$ of $\mathbf{X}$ is:

$$\nu_{\mathbf{X},[2]}(B_1 \times B_2) = \mathbb{E}[\mathcal{N}(B_1)\mathcal{N}(B_2)] - \mathbb{E}[\mathcal{N}(B_1) \cap \mathcal{N}(B_2)].$$

Moreover, if there exists a function $\rho_{\mathbf{X}}$ such that

$$\nu_{\mathbf{X},[2]}(B_1 \times B_2) = \int_{B_1}\int_{B_2} \rho_{\mathbf{X}}(x, y)\mathrm{d}x\mathrm{d}y,$$

then we call $\rho_{\mathbf{X}}$ the *second moment density* of $\mathbf{X}$.

Again, a formal interpretation is possible. Denoting with $B(x, r_x)$ and $B(y, r_y)$ two infinitesimally small balls centred in x and y respectively, the quantity $\rho_{\mathbf{X}}(x, y)\mathrm{d}x\mathrm{d}y$ gives an approximation of the joint probability that one point of the point process will occur within $B(x, r_x)$ and a second point within $B(y, r_y)$:

$$\rho_{\mathbf{X}}(x, y)\mathrm{d}x\mathrm{d}y \sim \mathbb{P}(\mathcal{N}_{\mathbf{X}}(B(x, r_x)) > 0, \mathcal{N}_{\mathbf{X}}(B(y, r_y)) > 0).$$

We remark that in the special case where no correlations occur between points, then the second moment density of $\mathbf{X}$ splits into $\rho_{\mathbf{X}}(x, y) = \lambda_{\mathbf{X}}(x)\lambda_{\mathbf{X}}(y)$, since the joint probability $\mathbb{P}(\mathcal{N}_{\mathbf{X}}(B(x, r_x)) > 0, \mathcal{N}_{\mathbf{X}}(B(y, r_y)) > 0) = \mathbb{P}(\mathcal{N}_{\mathbf{X}}(B(x, r_x)) > 0)\mathbb{P}(\mathcal{N}_{\mathbf{X}}(B(y, r_y)) > 0)$. The notion of the second moment density leads to another object, fundamental when studying the spatial relation between points of a process $\mathbf{X}$, which is the pair correlation function:

Definition 3 Let $\mathbf{X}$ be a point process with intensity function $\lambda_{\mathbf{X}}$ and second moment density $\rho_{\mathbf{X}}$. The pair correlation function $g_{\mathbf{X}}$ of $\mathbf{X}$ is the ratio

$$g_{\mathbf{X}}(x, y) = \frac{\rho_{\mathbf{X}}(x, y)}{(\lambda_{\mathbf{X}}(x)\lambda_{\mathbf{X}}(y))}.$$

It follows that for randomly located points, the pair correlation function is constantly equal to 1. In what follows, we will focus on isotropic and homogeneous point processes. This implies that $\lambda_{\mathbf{X}}$ is constant and that second moments only depend on distance between points, not on their exact locations. We will therefore write $\rho_{\mathbf{X}}(x, y) = \rho_{\mathbf{X}}(r)$ and $g_{\mathbf{X}}(x, y) = g_{\mathbf{X}}(r)$ with r being the Euclidean distance between x and y.
A first basic example of a point process is the homogeneous Poisson process, which is sometimes called the zero or completely random process after its property that each point is stochastically independent from all the others.
Therefore, in ecological theory, it represents the special case where there are no intraspecific spatial interactions between individuals. Of course, in applications, this complete randomness hypothesis mostly comes to fail due both to environmental variables and to seed dispersal mechanisms. Anyway, in exploratory analysis, the first step to analyse a database is to compare it with the zero-process.
We say that spatial point process $\mathbf{X}$ defined in $\mathbb{R}^2$ is an homogeneous Poisson process of intensity $\lambda_{\mathbf{X}}$ if it satisfies the following properties:

- for every bounded closed set $B \in \mathbb{R}^2$ having Lebesgue measure $\mu(B)$, the number of points falling in B has a Poisson distribution with mean $\lambda_{\mathcal{N}} \cdot \mu(B)$:

$$\mathcal{N}_{\mathbf{X}}(B) = e^{-\lambda_{\mathbf{X}}\mu(B)} \frac{(\lambda_{\mathbf{X}}\mu(B))^n}{n}$$

- Given two disjoint closed bounded subsets B_1 and B_2 of $\mathbb{R}^2$, $\mathcal{N}_{\mathbf{X}}(B_1)$ and $\mathcal{N}_{\mathbf{X}}(B_2)$ are independent.

In particular, as already noticed, if $\mathbf{X}$ is a Poisson process, then its pair correlation function is constantly equal to one: $g_{\mathbf{X}}(r) \equiv 1,\ \forall r \in (0, \infty)$. Nevertheless, when looking at the empirical pair correlation functions of Barro Colorado Island species, they do not show such constant behaviour. At contrast, they present a monotonic decreasing behaviour reaching the asymptotic value of 1 only at distance big enough so that correlations between points become actually negligible. This is a common characteristic which shows the inadequacy of the Poisson model. Indeed, the complete spatial randomness hypothesis in real species mostly comes to fail due to several reasons: on one side, changing environmental conditions, such as physical heterogeneity of the landscape or chemical composition of the soil, may lead to inhomogeneous patterns. On the other side, different seed dispersal mechanisms may favour the formation of clumping structures as well as dispersed ones.
A much more interesting example of point process are the Neyman-Scott processes (NS), which have found large applications in ecological theory due to their ability to model the clumping mechanism of plants' species in which daughter seeds are spread around a parent tree's location (Tovo et al., 2016a). They are the result of a three steps procedure:

- We distribute parent points according to a homogeneous Poisson process with fixed intensity ρ.

- To each parent, a random number of offspring is assigned, drawn from a Poisson distribution of intensity μ.

- The offspring are identically and independently scattered around their parents according to a fixed radial probability density $d_\gamma(r)$, called dispersal kernel, which depends on some parameters γ.
- The resulting process is formed only by the offspring's locations.

We remark that, as for the homogeneous Poisson process, also NS processes are both homogeneous and isotropic. Moreover, independently of the offspring's distribution, the intensity function of the process is given by the product $\lambda_{\mathbf{X}} = \mu\rho$, while its pair correlation function is given by $g_{\mathbf{X}}(r) = 1+f_\gamma(r)/\rho$, where $f_\gamma(r)$ is the convolution of the 2-dimensional probability density $d_\gamma(r)$. The three Neyman-Scott processes I most apply in my research are the modified Thomas, the exponential and the Cauchy point processes, whose dispersal kernel and pair correlation functions are shown in Table 1.

Neyman-Scott process	**Dispersal kernel**	**Pair Correlation function**
Modified Thomas	$d_\sigma(r) = \dfrac{1}{2\pi\sigma^2} e^{-\frac{r^2}{2\sigma^2}}$	$1 + \dfrac{1}{\rho}\dfrac{1}{4\pi\sigma^2} e^{-\frac{r^2}{4\sigma^2}}$
Exponential	$d_\beta(r) = \dfrac{\beta^2}{2\pi} e^{-\beta r}$	$1 + \dfrac{1}{\rho}\dfrac{\beta^4 r^2}{16\pi} K_2(\beta r)$
Cauchy	$d_b(r) = \dfrac{1}{2\pi b^2\left(1+\frac{r^2}{b^2}\right)^{3/2}}$	$1 + \dfrac{1}{\rho}\dfrac{b^2}{\pi\left(4+\frac{r^2}{b^2}\right)^{\frac{3}{2}}}$

Table 1. Examples of Neyman-Scott processes.

In order to model BCI species through these models, the first step, for each species s, is to estimate the set of parameters $(\rho_{\mathbf{X}_s}, \mu_{\mathbf{X}_s}, \gamma_{\mathbf{X}_s})$ which best describe its pattern, where $\mathbf{X}_s$ denotes the point process with which we decided to model the species. To get the parameters a standard method is the one of minimum contrast, which relies on the minimisation of the following integral:

$$\int_0^{d_{max}} (\hat{g}_{\mathbf{X}_s}(r)^{1/4} - g_{\mathbf{X}_s}(r)^{1/4})^2 \mathrm{d}r,$$

where $\hat{g}_{\mathbf{X}_s}(\cdot)$ is the empirical pair correlation function of species s, $g_{\mathbf{X}_s}(\cdot)$ is the analytical formulation depending on the parameters and d_{max} is the maximum considered distance. Thanks to the cluster parameters one obtains from the fit, it is then possible to predict the macro-ecological patterns of interest. However, since each species may contribute differently to them, we firstly need to introduce the concept of the superposition process $\mathbf{X}$, which is the collection of all random variables $\mathcal{N}_{\mathbf{X}}(B)$ giving the total number of points,

regardless of the species, falling within each subset $B \in \mathbb{R}^2$. One can then compute the first and second moments of the superposition process as function of those of its component processes. In particular, if we denote with $\mathbf{X}_1, \ldots, \mathbf{X}_s$ these latter, we have that $\lambda_{\mathbf{X}} = \sum_{s=1}^{S^*} \lambda_{\mathbf{X}_s}$ and that

$$g_{\mathbf{X}}(r) = \frac{1}{\lambda_{\mathbf{X}}^2} \left(\sum_{\substack{s,t=1 \\ s \neq 1}}^{S^*} \rho_{\mathbf{X}_s}(r) + 2 \sum_{s=1}^{S^*} \lambda_{\mathbf{X}_s} \lambda_{\mathbf{X}_t} \right).$$

From all these ingredients, one can search for analytical formulas of macro-pattern in function of the superposition process's moments. Let us see, for example, how to obtain the Sørensen index of similarity.
Let $\mathcal{B}_x = \mathcal{B}(x, r_x)$ and $\mathcal{B}_y = \mathcal{B}(y, r_y)$ two infinitesimally small balls centred in x and y. Then, in the context of point process, one can give the following interpretation:

$$\chi(\mathcal{B}_x, \mathcal{B}_y) = \frac{1}{\mathrm{d}x} \frac{\mathbb{E}[S(\mathcal{B}_x, \mathcal{B}_y)]}{\frac{1}{2}\mathbb{E}[S(\mathcal{B}_x) + (\mathcal{B}_y)]},$$

where $\mathrm{d}x$ is the infinitesimal area of both $\mathcal{B}_x$ and $\mathcal{B}_y$, $S(\mathcal{B}_x, \mathcal{B}_y)$ are the species shared by the two balls and $S(\mathcal{B}_x)$ is the total number of species falling in $\mathcal{B}_x$ and analogously for $S(\mathcal{B}_y)$.
One can find (Tovo et al., Soon on arXiv) that the following approximation holds for isotropic and homogeneous processes as the NS ones:

$$\chi(r) = \lambda_{\mathbf{X}}(g_{\mathbf{X}}(r) - 1) + \chi_\infty,$$

with χ_∞ is a constant depending solely on the species' abundances and representing the similarity at a scale where the clustering of individuals has no effect, so that no correlations occur between points.

Up to now we have only worked in our relatively small surveyed sample of rainforest. But what can we say of the rest of it? Is there a way to extrapolate information on bigger scales? The problem of inferring total biodiversity when only scattered samples are observed is a long-story problem.
In the early 1940s, the English naturalist Corbet spent two years trapping butterflies in Malaya, noting down, for any species he observed, the number of times he trapped it. At the end of that time, he returned to England and asked R. A. Fisher, how many new species he would have seen if he had returned to Malaya for another two years. The father of statistics was only the first mathematician to tackle a problem which seems impossible to solve at sight.
Today, the quantification of tropical tree biodiversity worldwide remains an open and challenging problem. In fact, more than two-fifths of the number of worldwide trees can be found either in tropical or sub-tropical forests, but only 0.000067% of species identities are known.
For practical reasons, biodiversity is typically measured or monitored at fine spatial scales. However, important drivers of ecological change tend to act at large scales. An example

are conservation issues which apply to diversity at global, national or regional scales.
Extrapolating species richness from the local to the global scale is not straightforward. Indeed, a vast number of different biodiversity estimators have been developed under different statistical sampling frameworks, but most of them have been designed for local/regional-scale extrapolations, and they tend to be sensitive, for example, to the spatial distribution of trees or sampling methods.
As we said before, a common statistical tool used to describe the commonness and rarity of species in an ecological community is the SAD, which is measured at local scales, where the identities of the individuals living in the area are known. The sampled SAD can be fit to a given functional form at that scale. However, in general, that form changes at different spatial scales, thus hindering analytical treatment.
We developed an analytical framework that provides robust and accurate estimates of species richness and abundances in biodiversity-rich ecosystems, when only a sample of them have been observed (Tovo et al., 2016b).
Let us assume that our ecological community consists of S species and let $P_{n,s}$ be the probability that species s has exactly n individuals at time t. We also make the following assumptions:

- All species are independent one another, so that no inter-specific interactions occur;
- All species are compound of demographically equivalent individuals, in the sense that each of them has the same probability of giving birth, dying, speciating or migrating (neutrality hypothesis, see Hubbell (2001), Volkov et al. (2003)).

We then model each species according to a stochastic birth and death process ruled by the following Master Equation:

$$\frac{\partial}{\partial n}P_n(t) = \ P_{n-1}(t)b_{n-1} + P_{n+1}(t)d_{n+1} - P_n(t)\cdot(b_n + d_n)$$

Setting $b_{-1} = d_0 = 0$, one get the following steady-state solution:

$$P_n = P_0 \prod_{i=0}^{n-1} \frac{b_i}{d_{i+1}},$$

where P_0 is the normalization condition and where we have dropped the index of the species because of the neutrality hypothesis. We will refer to P_n as the *Relative-Species Abundance* (RSA) of the community. Let us now consider two cases.
If one assumes that the population dynamics in the community are governed by ecological drift and random speciation, then we have:

$$b_n = bn + \delta_{n,0}\nu, \qquad d_n = dn,$$

where the birth rate accounts for reproduction and speciation, this latter through the parameter ν ensuring that the system is always populated whenever species go extinct.
Then we find the following stationary solution

$$P_n = P_0 \frac{\nu}{b}\frac{x^n}{n},$$

which is known amongst ecologists as *Fisher log-series*, since it was first discovered experimentally in 1943 by the great statistician while studying Corbet's problem.
One can also find the SAD:

$$SAD(n) = \mathbb{E}[S_n] = \sum_{s=1}^{S} P_n = \theta \frac{x^n}{n},$$

where $x = b/d$ and where $\theta = SP_0\nu/b$ is called the *biodiversity parameter*. This SAD has no internal mode and therefore it predicts that singleton species, those with one individual only, are always the most frequent. This is not always the case, as many communities have species' abundances that are more frequent than singletons.
Let us thus consider a second case. If our community is not closed, but it is surrounded by a meta-community, it is more reliable that speciation phenomenon has a much smaller affect with respect to random migration. Let us then modify the birth and death rates of our process as follows:

$$b_{n,k} = b(n + r), \qquad d_{n,k} = dn,$$

where the r parameter now incorporates migration from the meta-community into our community.
Then the new stationary solution is a *Negative Binomial* of parameters r and $\xi = \frac{b}{d}$:

$$P_n = \binom{n + r - 1}{n} \xi^n (1 - \xi)^r.$$

In particular, we have the following key result for these two distributions.

Proposition 4 *Let P_n be the RSA at the global scale and $\mathbb{P}_p(k|n)$ be the sampling probability at a sub-scale $p \in (0, 1)$, i.e. the conditional probability that a species has k individuals in the sub-sample given that it has abundance n at the global scale.*
If $\mathbb{P}_p(k|n)$ is binomially distributed, then we have the following:

- *If $P_n \sim NB(r, \xi)$, with $r \geq 1$, $0 \leq \xi \leq 1$, then the RSA at the sub-scale p is given by:*

$$P_k^S(r, \xi) \sim NB(r, \xi_{(p)}), \text{ with } \xi_{(p)} = \frac{p\xi}{1 - \xi(1 - p)}$$

- *If $P_n \sim LS(x)$, with $0 \leq x \leq 1$, then the RSA at the sub-scale p is given by:*

$$P_k^S(x) \sim LS(x_{(p)}), \text{ with } x_{(p)} = \frac{px}{1 - x(1 - p)}.$$

We say that both the distributions satisfy the self-similarity property.
Let us see how to apply such result to extrapolate informations from samples to bigger areas.
We observe a fraction p^* of a forest and we observe S^* species within it. Our goal is to infer the number of species S which are present at the global scale. Given the abundances

of the S^* species at the local scale, we can construct the RSA of our sample and fit it with a negative binomial to get the parameters r and $\xi_{(p^*)}$ which best describe the empirical data. Clearly, one could do the same with a Log-Series distribution and get the $x_{(p^*)}$ parameters.
One can then analytically extract the values of the RSA parameters at the global scale and the corresponding total biodiversity S through the following equations (Tovo et al., 2016b):

$$\xi = \frac{\xi_{(p^*)}}{p^* + \xi_{(p^*)}(1-p^*)}$$
$$S = S^* \frac{1-(1-\xi)^r}{1-(1-\xi_{(p^*)})^r}$$

A great advantage of using a negative binomial instead of the Log-Series relies on its versatility. Indeed, it can display both monotonic log series-like behaviour and a unimodal shape, depending on its parameters. This is an important property, since we already underlined that empirical SADs, especially at large scales or with increasing sampling effort, often change shape and start displaying an interior mode that cannot be captured by the Log-Series distribution but that can now be explained and well described by the negative binomial.
Another important aspect of negative binomials comes from the fact that also linear combinations of negative binomials of the same scaling parameter ξ and different parameters r_i satisfy the self-similarity property, thus allowing to apply our method also when complex-behavioured SADs are encountered, case for which using one single negative binomial may provide a very poor fit of the data whereas using linear combinations of negative binomials can faithfully reproduce them.
We tested our framework both to synthetic and real data and it resulted to better perform with respect to previously proposed methods in the literature (Chao and Chiu, 2016, Slik et al., 2015), both in predicting biodiversity and macro-patterns as the SAR.
There are many directions one may take from here and on which we are now working on. First of all we would like to obtain a spatial-explicit model which may link and predict the profile of the various macro-ecological patterns cited above. Indeed, although various models have been proposed in literature to explain and capture some recurrent patterns, there still misses a unifying model able to explain all of them at the same time.
Lastly, we also wish to apply our framework to the very different scale of genetic diversity. In particular, we have a collection of all bacteria proteomes, which are the set of all proteins expressed by a genome. We would like to test our method's reliability in predicting the total number of proteomes present in a meta-genome, when only a fraction of them is observed. This would have clearly important application in medicine for example.

References

[1] A. Chao and C.-H. Chiu, "Species richness: estimation and comparison". Wiley StatsRef: Statistics Reference Online, 2016.

[2] A. Chao, R.L. Chazdon, R.K. Colwell, and T.-J. Shen, *A new statistical approach for assessing similarity of species composition with incidence and abundance data.* Ecology letters, 8/2 (2005), 148–159.

[3] B.D. Coleman, *On random placement and species-area relations.* Mathematical Biosciences, 54/3-4 (1981), 191–215.

[4] S. Hubbell, *The unified neutral theory of species abundance and diversity.* Princeton University Press, Princeton, NJ. Hubbell, SP (2004); Quarterly Review of Biology, 79 (2001), 96–97.

[5] J.F. Slik, V. Arroyo-Rodrguez, S.-I. Aiba, P. Alvarez-Loayza, L.F. Alves, P. Ashton, P. Balvanera, M.L. Bastian, P.J. Bellingham, E. Van Den Berg, et al, *An estimate of the number of tropical tree species.* Proceedings of the National Academy of Sciences, 112/24 (2015), 7472–7477.

[6] A. Tovo, M. Formentin, M. Favretti, and A. Maritan, *Application of optimal data-based binning method to spatial analysis of ecological datasets.* Spatial Statistics, 16: 137–151, 2016a.

[7] A. Tovo, S. Suweis, M. Formentin, M. Favretti, J.R. Banavar, S. Azaele, and A. Maritan, *Upscaling species richness and abundances in tropical forests.* bioRxiv, page 088534, 2016b.

[8] A. Tovo, M. Formentin, and M. Favretti, *Decay of similarity with point processes.* Soon on arXiv.

[9] I. Volkov, J. Banavar, S. Hubbell, and A. Maritan, *Neutral theory and relative species abundance in ecology.* Nature, 424(6952):1035?7, Aug. 2003. ISSN 1476-4687. doi: 10.1038/nature01883. URL `http://www.ncbi.nlm.nih.gov/pubmed/12944964`.