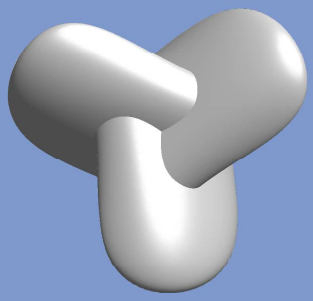

# Publicaciones Matemáticas del Uruguay



Volumen 14, Año 2013

# PUBLICACIONES MATEMÁTICAS DEL URUGUAY

---



Published by:
CMAT-Facultad de Ciencias
IMERL-Facultad de Ingeniería
**Universidad de la República**
and
**PEDECIBA** Matemática

web site *http://imerl.fing.edu.uy/pmu*

**ISSN: 0797-1443**

**Credits:**
**Cover design:** J. Rodriguez Hertz
LaTeX **editor:** J. L. Vieitez

## CONTENTS

## PREFACE

In the mid 80's, several mathematicians, who had been prevented from working in Uruguay by the civic-military putsch of June 1973, met with a group of young students who had already started their training there. This created much enthusiasm in both sides, the former having met a group of talented students, and the latter having studied without any contact with a proper researcher up to then. One of these mathematicians was Jorge Lewowicz. When he came back to Uruguay, he started giving a few basic courses on Dynamical Systems as well as developing the eagerness for research. Soon, a study group on that topic was created. Several students completed their initial training and then started their graduate studies, most of them abroad. It should be stressed thas such a development would not have been possible without the unshakeable and faithful support of the IMPA-Rio de Janeiro, and in particular of Jacob Palis and Ricardo Mañé. Already at the end of the 80's, much collaboration existed between researchers and students of the Universidad de la República. They would belong to the Instituto de Matemática de la Facultad de Ingeniería and the Departamento de Matemática de la Facultad de Humanidades y Ciencias. They developed their own and original style of doing maths, organizing courses and seminars under Lewowicz's general guidance. After 25 years of continuous activity, a School of Dynamical Systems has consolidated as the main maths research area in Uruguay. The recognition of its different subteams is due to the many international conferences they have been invited to, the numerous foreign professors they have welcomed, as well as the courses they have themselves given abroad. This way a large number of students was trained, and to this day, every week a seminar brings together researchers and students. At present the School has multiplied its international links and extended its ambitions. The sum of its members' interests now covers a large spectrum of Mathematics. A final aspect of this to be highlighted, more than for its subjective implication, but because it supports the progress of research quality, is the awareness of the common roots of all members of the School, as well as its members' subteam identities.

And here we are now, with this special issue of the Publicaciones Matemáticas del Uruguay, dedicated to to the International Congress on Dynamical Systems held in Montevideo, Uruguay, from August 13th to August 17th, 2012 and the Doctor Honoris Causa to Jorge Lewowicz. The Congress was a satellite conference of the 4th Latin American Congress of Mathematicians (CLAM, organized by the UMALCA) which took place in Córdoba, Argentina. It was Organized by *Grupo de Sistemas Dinámicos IMERL-CMAT - CSIC 618/2010- Universidad de la República - Uruguay.* In parallel the Universidad de la República titled Dr. Jorge Lewowicz, Doctor Honoris Causa on Wednesday 15th. August, 2012. Most articles of this volume are related to topics in Dynamical Systems to which Lewowicz contributed. We are indebted to CSIC, PEDECIBA-Matemática, and to IMERL-Facultad de Ingeniería and CMAT-Facultad de Ciencias from Universidad de la República, for partially supporting the edition of this volume. Last but not least, we counted on the generous collaboration of the authors and the referees, without whom this volume would not have been possible. We wish to express our gratitude to all of them.

**List of conferences presented at**
**Montevideo Dynamical Systems Congress 2012**

Alexander Arbieto- UFRJ, Río de Janeiro, Brazil
*On flows without points accumulated by periodic orbits of different indices*

Thierry Barbot- Université d'Avignon, France
*Structure and examples of pseudo-Anosov flows in graph-manifolds and Seifert fibered pieces*

Jairo Bochi- PUC, Río de Janeiro, Brazil
*Robust vanishing of all central Lyapunov exponents*

Keith Burns- Northwestern, Univ. Chicago, USA
*Ergodicity of the Weil-Petersson geodesic flow*

Sõnia Pinto de Carvalho- UFMG, Belo Horizonte, Brazil
*Dynamics near an invariant horizontal circle*

Thiago Catalan- UFU, Uberlãndia, Brazil
*A lower bound for topological entropy and some generic properties for symplectic diffeomorphisms*

Ruben Chaer- Universidad de la República, Montevideo, Uruguay
*Dimensionality reduction in optimal operation of hydrothermal power generation dynamical systems*

Alain Chenciner- IMCCE, France
*Angular momentum and Horn's problem*

Lorenzo Díaz- PUC, Río de Janeiro, Brazil
*Fragile and stable cycles*

Pedro Duarte- ULisboa, Portugal
*Dissipative polygonal billiards*

Davide Ferrario- UMilano, Bicocca, Italy
*Some symmetric orbits for the n-body problems*

Todd Fisher- Brigham - Youngh University, Utah, USA
*Equilibrium states for robustly transitive systems*

Jason Gallas- UFRGS, Porto Alegre, Brazil
*Accumulation cascades in families of stable self-induced oscillations in lasers: regularities in peak-adding sequences generated by DDES, ODES, and MAPS*

Katrin Gelfert- UFU, Uberlãndia, Brazil
*Spectral decomposition in non-hyperbolic dynamics*

Vadim Kaloshin- UMD, Maryland, USA
*Arnold Diffusion via invariant cylinders and Mather variational method*

Sergey Kryzhevich- Chebyshev Lab, Saint Petersburg, Russia
*On the plaque expansivity conjecture*

François Ledrappier- UND, Indiana, USA
*Regularity of the entropy of random walks on hyperbolic groups*

Marco Lenci- Unibo, Bologna, Italy
*Global observables and the question of mixing in infinite ergodic theory*

André L.P Livorati -I. Física - USP, Brazil
*On dynamics of a family of stadium billiards: phase transitions, Fermi acceleration and decay of energy*

Cristina Lizana- PUC, Río de Janeiro, Brazil
*Robust Transitiviy for endomorphisms*

Artur Lopes- UFRGS, Porto Alegre, Brazil
*Entropy and pressure for one-dimensional spin lattices with a general a priori measure: positive and zero temperature*

Arturo Martí- Universidad de la República, Montevideo, Uruguay
*Synchronization of chaotic maps in delayed complex networks: advances, perspectives and applications*

Carlos Gustavo Moreira- IMPA, Río de Janeiro, Brazil
*On geometric properties of horseshoes in arbitrary dimensions*

Maria José Pacifico- UFRJ, Río de Janeiro, Brazil
*A toy model for flows presenting equilibria accumulated by regular orbits*

Alberto Pinto- UPorto, Porto, Portugal
*Anosov tilings*

Silvina Ponce Dawson- UBA, Buenos Aires, Argentina
*Diffusion of messages and messengers and the formation of morphogen gradients*

Rafael Potrie- Universidad de la República, Montevideo, Uruguay
*On the work of Jorge Lewowicz on expansive systems*

Rafael Oswaldo Ruggiero- PUC, Río de Janeiro, Brazil
*Expansive and topologically stable geodesic flows: From dynamics to global geometry*

*and rigidity*

Radu Saghin- PUC, Valparaíso, Chile
*Volume growth and entropy for partially hyperbolic diffeomorphisms*

H. Sánchez Morgado- UNAM, México
*Exponential convergence of the solutions of the time-periodic Hamilton-Jacobi equation on the torus*

Samuel Senti- UFRJ, Río de Janeiro, Brazil
*The unicity of the T-Conformal measure for Hénon maps at the first bifurcation*

Nándor Simányi- UAB, Alabama, USA
*Singularities and non-hyperbolic manifolds do not coincide*

Vctor Sirvent- USB, Caracas, Venezuela
*Space filling curves, expanding maps and geodesic laminations*

Alfonso Sorrentino- Cambridge, UK
*Rigidity of Birkhoff billiards*

Jorge Sotomayor- USP, São Paulo, Brazil
*Singularities in geometrically defined foliations*

Domokos Szász- Budapest, Hungary
*On Dettmann's 'Horizon' conjectures*

Sergei Tabachnikov- PSU, Pennsylvania, USA
*Tire tracks geometry, Hatchet planimeter, Menzin's conjecture, and complete integrability*

Ali Tahzibi USP, São Carlos, São Paulo, Brazil
*Central Lyapunov exponents of partially hyperbolic diffeomorphisms of 3-torus*

Sergei Tikhomirov- Free University of Berlin, Germany
*Shadowing lemma for partially hyperbolic systems*

Sandro Vaienti- Univ.Toulon and CTPM, Marseille, France
*A survey on new results about statistical properties of deterministic and random dynamical systems*

Carlos Vásquez- UCV, Valparaíso, Chile
*Removing zero central Lyapunov exponents for partially hyperbolic diffeomorphisms on nilmanifold*

Maciej Wojtkowski- UWM, Olsztyn, Poland
*1-dim tilings and bi-partitions*

Photo taken at the end of the ceremony of the appointment of Jorge Lewowicz as *Doctor Honoris Causa* of the Universidad de la República with part of the Dynamical System group at Uruguay.

# LAUDATIO HONORIS CAUSA JORGE LEWOWICZ

MARTÍN SAMBARINO

Hoy es la ceremonia de entrega del título de Dr. Honoris Causa al Prof. Jorge Lewowicz y he sido designado para hacer la Laudatio; es un honor y un placer decir estas palabras en este homenaje a mi Profesor, colega y amigo Jorge.

Esta tarde, en el Congreso de Sistemas Dinámicos hemos escuchado dos conferencias, una por Rafael Potrie y otra por Rafael Ruggiero sobre los trabajos científicos de Jorge Lewowicz, su influencia y posterior desarrollo de su obra científica. Me voy a referir a su legado científico, pero de forma mucho menos técnica, y también a otro aspecto de la obra del Prof. Lewowicz, que no es independiente de lo anterior: la continuación de la escuela matemática uruguaya originada por Massera y Laguardia y la conformación de un grupo numeroso y destacado de investigadores en Sistemas Dinámicos.

Jorge Lewowicz comenzó sus estudios de Ingeniería a mediados de la década del 50 e ingresa como ayudante del Instituto de Matemática y Estadística en el año 58. Bajo la orientación del Prof. Jose Luis Massera inició sus estudios sobre Ecuaciones Diferenciales y en 1961 publica su primer trabajo científico: *Sobre un teorema de Szmydtowna*. Posteriormente, mediante una beca Fullbright, viaja a Estados Unidos en 1964 y en 1966 obtiene su título de PhD en Matemática en Brown University.

Desde sus inicios Lewowicz se interesa por cuestiones de estabilidad y de dinámica topológica, temas que ha desarrollado y plasmado en más de una treintena de artículos científicos. Pero por sobre todas las cosas, es mayormente reconocido por su desarrollo de la teoría de sistemas expansivos, temas en los cuales Lewowicz hizo escuela. En el área de Sistemas Dinámicos, el Uruguay fue ampliamente reconocido por tener un grupo muy fuerte en dinámica de expansivos. Entre los reconocimientos académicos como científico, Lewowicz fue designado miembro de la Academia de Ciencias del Tercer Mundo, miembro de la Academia de Ciencias de Uruguay y miembro correspondiente de la Academia de Ciencias de Argentina.

No voy a hacer aquí una lista ni enumeración de sus trabajos. Sí voy a decir que todos sus trabajos comparten las siguientes características: originalidad (y con originalidad me refiero además a una linea propia de investigación, a su búsqueda íntima de la armonía de la Matemática), la interrelación entre diversos métodos e ideas profundas en la Matemática, y que además son de una gran elegancia y belleza.

De todas formas, voy a referirme a tres trabajos fundamentales de su obra en particular. Primero, a su artículo *Lyapunov functions and topological stability* publicado en Journal of Differential Equations en 1980. En este trabajo, Lewowicz introduce la noción de funciones de Lyapunov para sistemas dinámicos, y demuestra un teorema de estabilidad topológica bajo las condiciones de existencia de una función de Lyapunov no degenerada. Esto dió lugar a nuevos ejemplos, fuera del mundo hiperbólico, de sistemas

CMAT, Facultad de Ciencias, Universidad de la República, Uruguay. email: *samba@cmat.edu.uy*.

topológicamente estables. Además, este resultado representa una analogía notable, e insospechada a priori, con el resultado de estabilidad asintótica de puntos de equilibrio para ecuaciones diferenciales. En el mismo trabajo también se caracterizan los difeomorfismos de Anosov o conjuntos hiperbólicos mediante una forma cuadrática no degenerada que crece a lo largo de las trayectorias. Esta caracterización y forma de pensar, ha sido usada y desarrollada después por Roberto Markarian para el estudio de billares.

El segundo trabajo al que me voy a referir es *Persistence in Expansive Systems*, publicado en Ergodic Theory and Dynamical Systems en 1983. La expansividad es una propiedad que aparece naturalmente en difeomorfismos de Anosov y en conjuntos hiperbólicos y es el concepto básico de lo que hoy se conoce como caos o impredictibilidad. Su definición es extremadamente sencilla y general en su contexto. En este trabajo, Lewowicz comienza con el estudio sistemático de sistemas expansivos y prueba una propiedad esencial de estos: no hay puntos Lyapunov estables. También introduce la noción de persistencia, y prueba, bajo ciertas condiciones que no voy a explicar aquí, que los sistemas expansivos son persistentes.

Y en tercer lugar, quiero referirme a su artículo *Expansive Homeormorphisms of Surfaces*, publicado en 1989. Este es su principal contribución y es uno de los trabajos más importantes y célebres del área, en particular en dinámica topológica. En este trabajo no solo resuelve un viejo problema abierto (sobre la no existencia de homeomorfismos expansivos en la esfera), sino que hace una clasificación completa de los homeomorfismos expansivos en superficies. En este trabajo Lewowicz exhibe de manera notable la dialéctica entre la topología y la dinámica. Pero más allá de la importancia del resultado y su profundidad, tanto el enunciado, el resultado en sí, es de una gran belleza como lo es su demostración: a partir de una simple definición y a través de argumentos simples pero profundos se va tejiendo la relación entre la dinámica y la topología hasta llegar a una descripción de las propiedades de los sistemas expansivos que permiten su clasificación.... es como si fuera una sinfonía que comienza con un solo intrumento y este, a través de su melodía y armonía, va despertando y contagiando a toda la orquesta para el *Grande Finale*.

Como dije al principio, quiero referirme ahora a otro aspecto del legado académico de Lewowicz: su carácter como formador. Fiel a la tradiciones de la escuela matemática uruguaya, la formación y el estímulo a los jóvenes y a la iniciación de la investigación es y ha sido una de sus preocupaciones centrales. Fiel al rigor académico, estimuló con generosidad la salida de muchos estudiantes a realizar el doctorado en lugares de excelencia. Ha dirigido 6 tesis de Doctorado (2 en Brasil, 4 en Uruguay), 15 tesis de Maestría (8 en Venezuela, 7 en Uruguay) y una decena de monografías de Licenciatura. Pero más allá de sus alumnos directos, Lewowicz ha influenciado el desarrollo como matemáticos e investigadores de otros tantos que no fueron sus doctorandos, ya sea en el Uruguay como en el exilio. Hoy nuestra Universidad cuenta con 17 investigadores en sistemas dinámicos y la conformación de este grupo se debe en gran o total medida a Jorge Lewowicz.

Lewowicz trasmitió a lo largo de estos años, tanto en sus clases como en las célebres caminatas por el pasillo del IMERL y en diversas reuniones, valores fundamentales de la escuela matemática uruguaya: la investigación temprana, la calidad y rigurosidad científica, el desarrollo de un ámbito propicio para la discusión, la trasmisión y creación de conocimiento, el compromiso social e institucional. Pero no solo dentro de

la matemática, sino en la Facultad de Ingeniería y en la Universidad, estimuló, inspiró e influenció el quehacer científico de muchos jóvenes y no tan jóvenes, generación tras generación. Defendió y abogó por la calidad académica con estándares internacionales, tanto en la Facultad de Ingeniería como en nuestra Universidad, en particular desde la CSIC. La Facultad de Ingeniería es lo que es hoy en parte gracias a la semilla del Instituto de Matematica y Estadística Prof. Rafael Laguardia, y de alguna forma el papel que jugaron Massera y Laguardia antes de la dictadura en la Facultad, lo desempeñó Jorge después de ésta.

Quiero contar una anécdota, que muy pocos conocen pero que ilustra las cosas que genera Jorge en su entorno. José Vieitez fue alumno de Jorge y el primer Doctor en Matemática por la Universidad de la República-PEDECIBA. Una tarde, llegando Jorge a su casa, abre el buzón de correspondencia y se encuentra con una copia del título de Dr de José Vieitez que había sido expedido ese día y con una dedicatoria en el reverso: Al Maestro con Cariño.

Cuando uno entraba en una clase de Lewowicz inmediatamente se sorprendía. Se sorprendía por la pasión, amor y placer que tenía con la Matemática. Se sorprendía por la profundidad y la trascendencia filosófica que le daba al objeto de estudio del curso. Y se sorprendía también por la importancia que le daba a los alumnos: parecía que no había cosa más importante en el mundo que las dudas e inquietudes que podrían tener estos y dejaba bien en claro que estaba a entera disposición en cualquier momento o lugar; hasta nos daba el número de teléfono de su casa (que claro, muchas veces después no atendía!). Y se sorprendía también, si aún no lo conocía, por su singular personalidad, agudeza y sentido del humor.

Aquellos que tuvimos el privilegio de estar en una clase con Lewowicz, la disfrutamos minuto a minuto, incluso cuando luego de enunciar un resultado nos decía: *Señoras y señores, tienen 3 minutos para pensar la demostración...* se hacía un silencio total y Jorge caminaba de lado a lado del salón. Y si alguien esbozaba alguna idea para la demostración, entonces Jorge la seguía, sin importar si había una camino más corto o más fácil: lo más importante era respetar la libertad y los caminos de pensamiento de cada uno.

Y como decía anteriormente, defendía la investigación temprana: uno no tenía que ser erudito para ser creativo, más aún, lo que había que estimular era la creatividad y en todo caso, la erudición venía de la mano de la necesidad de resolver y plasmar las ideas. Y así, les daba a los alumnos problemas abiertos o simplificaciones de estos, o incluso bajaba a tierra diversos problemas técnicos de sus investigaciones para que los alumnos los pudieran atacar. Y cuando alguien, fulano de tal, le hacía una pregunta de Matemática que Lewowicz entendía que el mismo la podía responder, decía, no exento de picardía: esa es una pregunta que debe responder fulano de tal.

Hablar de Matemática con Jorge ha sido, y es, iluminador y un placer, aún para los más jóvenes, bien en el Instituto, bien en una reunión o bien en las visitas que recibe ahora en su casa. Y no solo de Matemática, sino de lo que han sido sus preocupaciones durante toda su vida: la Ciencia, la Educación, la Universidad, el País, el Ser Humano. Y también, como hicieron otros matemáticos que volvieron del exilio, a través de anécdotas e historias, nos fue pincelando la figuras de Laguardia y Massera y de la vida del Instituto previa a la dictadura del 73, de forma que las generaciones más jóvenes aprendimos a

respetar y querer entrañablemente, a sentirnos parte de una Historia y a generar el compromiso de continuarla...

La Universidad debe reconocer, para sí misma y para la sociedad, quienes son sus hombres de valía, y tú Jorge, vaya que sí lo sos. ¡Muchas gracias!

Muchas gracias.

Jorge Lewowicz

# ON THE WORK OF JORGE LEWOWICZ ON EXPANSIVE SYSTEMS

RAFAEL POTRIE

Abstract. We will try to give an overview of one of the landmark results of Jorge Lewowicz: his classification of expansive homeomorphisms of surfaces. The goal will be to present the main ideas with the hope of giving evidence of the deep and beautiful contributions he made to dynamical systems. We will avoid being technical and try to concentrate on the tools introduced by Lewowicz to obtain these classification results such as Lyapunov functions and the concept of persistence for dynamical systems. The main contribution that we will try to focus on is his conceptual framework and approach to mathematics reflected by the previously mentioned tools and fundamentally by the delicate interaction between topology and dynamics of expansive homeomorphisms of surfaces he discovered in order to establish his result.

*The value of a person resides in his major contribution*
Arab proverb freely translated[1].

## 1. Introduction

Among the contributions of Lewowicz to mathematics, it is hard to ignore what I believe to be his major one: The creation of a school of dynamical systems in Montevideo. This school is also highly influenced by his way of looking at mathematics which I hope will be illustrated in this brief note. The main point is that it is not only the people who work in expansive systems that has been influenced by him. I recommend reading [Sam] for a global panorama of Lewowicz's contributions. The goal of this note is not to describe this aspect of Lewowicz contributions, it is devoted to describe some of his mathematical contributions.

As a disclaimer, I mention that I am by far not the best qualified to write about Lewowicz's work and that this note does not pretend to be a summary of all of his contributions to mathematics. However, as a member of the above mentioned school, and having been strongly influenced by him, I happily accepted this task and will try to give a panorama of the results of Lewowicz concerning expansive homeomorphisms.

Let us start with a simple and elementary definition:

*Definition* (Expansive homeomorphism). Let $f : M \to M$ be a homeomorphism of a compact metric space $M$. We say that $f$ is *expansive* if there exists $\alpha > 0$ such that given $x \neq y \in M$ there exists $n \in \mathbb{Z}$ such that $d(f^n(x), f^n(y)) \geq \alpha$. The largest possible constant $\alpha$ is called the *expansivity constant* of $f$ for the metric $d$.

$\diamondsuit$

Expanded version of a talk given by the author in the conference Dynamical Systems in Montevideo held at Montevideo from 13 to 17 of August 2012. The author was partially supported by CSIC group 618/2010.

[1]Translated from a phrase in the entrance of the Institut du Monde Arab in Paris , France.

It is important to remark that although the definition depends on the metric, the notion of expansivity is purely topological and can be stated for general topological spaces by demanding that points outside the diagonal $\Delta \subset M \times M$ escape by iteration of $f \times f$ from a fixed neighborhood[2] of $\Delta$.

There are many well known examples of expansive homeomorphisms: subshifts of finite type (as well as hyperbolic sets of Smale's theory) are expansive; and the dynamics restricted to the minimal set of a Denjoy counter-example is also expansive. We will focus mainly on other kind of examples, those whose phase space is a manifold.

Quoting Lewowicz himself in [$L_3$]

> (...)*expansivity means, from the topological point of view, that any point of the space $M$ has a distinctive dynamical behavior. Therefore, a stronger interaction between the topology of $M$ and the dynamics could be expected.*

Examples of expansive homeomorphisms on manifolds are given by Anosov and quasi-Anosov diffeomorphisms (see [Fr, FR]) as well as the well known pseudo-Anosov maps introduced by Thurston ([Th]). Of course, products of expansive homeomorphisms are expansive. In [OR] it is proved that every surface of positive genus admits an expansive homeomorphism.

We are now ready to state a landmark result of the work of Lewowicz ([$L_3$]):

**Theorem.** *There are no expansive homeomorphisms on the two-dimensional sphere $S^2$.*

This theorem is highly non-trivial, yet, its statement is completely simple. Let us remark that there is an independent proof of this result and the rest of the results in [$L_3$] by Hiraide ([H]).

The concrete purpose of this note is to explain the main ideas behind this result as well as the classification theorem of expansive homeomorphisms on surfaces obtained by Lewowicz in [$L_3$]. Other contributions will be covered by Ruggiero, specially those concerning geodesic flows and quotient dynamics (see also [Ru]), of course, both presentations will have substantial overlap.

It would not be fare to write about Lewowicz's work without giving motivations for the study of expansive homeomorphisms. We will start by introducing some motivations in the first sections by reviewing some of Lewowicz's previous work. Other motivations can be found along the literature, in particular [$L_4$] has a chapter devoted to that.

## 2. Lyapunov functions and topological stability

We start with a quotation from the introduction of [$L_1$] "*This paper contains some results on topological stability (see [2,3]) that generalize those obtained in [2] much in the same way as Lyapunov's direct theorem generalizes the asymptotic stability results of the hyperbolic case: if at a critical point, the linear part of a vector field has proper values with negative real parts, the point is asymptotically stable and the vector field has a quadratic Lyapunov function; however, asymptotic stability may also be proved for vector fields with non-hyperbolic linear approximations, provided they have a Lyapunov function. In a way this is what we do here, letting Anosov diffeomorphisms play the role of the hyperbolic critical point and replacing stability by topological stability; we get this time a class of topological stable diffeomorphisms wider than the class of Anosov diffeomorphisms.*"

[2]In more technical wording, that $\Delta$ is a locally maximal set for $f \times f$.

Lyapunov functions, introduced by Lewowicz in [$L_1$] play the role of a metric, which in the case of expansive homeomorphisms is a type of adapted metric which allows to distinguish the stable and unstable parts of the points which are nearby a given orbit. Other kinds of adapted metrics have been then proposed (see [Re, Fa]) but we will focus on Lyapunov functions that are present transversally in much of Lewowicz's work (and also in some of his students, see [$V_2$, $Gr_1$, $Gr_2$]).

One important tool introduced in order to construct Lyapunov functions is that of quadratic forms (or infinitesimal Lyapunov functions) which have had a strong impact in different directions well beyond expansive systems as we will explain below.

*Definition* (Lyapunov function). A continuous function $V : U \to \mathbb{R}$ from a neighborhood $U$ of the diagonal in $M \times M$ is said to be a Lyapunov function for $f : M \to M$ iff:

- $V(x, x) = 0$ for every $x \in M$.
- $V(f(x), f(y)) - V(x, y) > 0$ for every $x \neq y$.

♢

It can be seen as a function which "sees" the expansivity in one step. It is proved in [$L_3$] (Theorem 1.3) that these functions characterize expansive homeomorphisms (see [Fa] for a different approach):

**Theorem 2.1.** *A homeomorphism of a compact metric space is expansive if and only if it admits a Lyapunov function.*

Lyapunov functions also provide a way of establishing topological stability of diffeomorphisms (see [Wa] for the Anosov case) which may not be Anosov.

We recall the definition of topological stability. We say that a homeomorphism $f : M \to M$ of a compact manifold $M$ is *topologically stable* if there exists $\varepsilon > 0$ such that for every homeomorphism $g : M \to M$ which is at $C^0$-distance smaller than $\varepsilon$ of $g$ there exists a continuous surjective map $h : M \to M$ which semiconjugates $f$ and $g$, that is:

$$f \circ h = h \circ g$$

Thurston's pseudo-Anosov maps (see [OR, Th]) do admit Lyapunov functions (see [$L_2$] Lemma 3.4 or apply the previous theorem[3]), however, they are not topologically stable: One can make perturbations of a pseudo-Anosov map making that some points have their orbit going "across" the singularities and which will not be shadowed by an orbit of the pseudo-Anosov map. See [$L_2$] or look at the figures in [$L_4$] (Figure 2 in page 11) or [$LC_1$] (Figure 3).

Therefore, the existence of Lyapunov functions alone is not enough to get topological stability. One must add a new hypothesis which can be thought of as a weak topological version of hyperbolicity (see [$L_1$] Section 5 for a more general and precise definition):

*Definition.* We say that a Lyapunov function $V : U \to \mathbb{R}$ is *non-degenerate* if for every $x \in M$ there exists a splitting $T_xM = S_x \oplus U_x$ such that if $C_S(x)$ (resp. $C_U(x)$) is a cone around $S_x$ (resp. $U_x$) then $V(\cdot, x)$ is positive (resp. negative) in $\hat{C}_S(x)$ (resp. $\hat{C}_U(x)$), the projection of $C_S(x)$ (resp. $C_U(x)$) by the exponential map in a small neighborhood.

♢

[3] In fact Lewowicz uses his construction of a Lyapunov function to obtain an alternative proof of expansiveness of pseudo-Anosov maps.

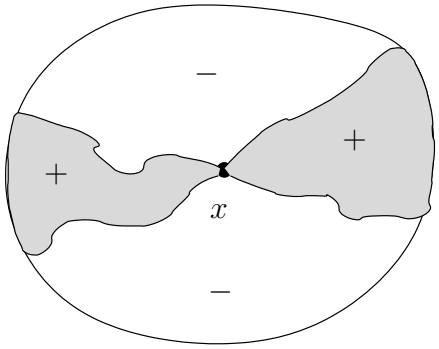


FIGURE 1. Positive and negative regions of $V(\cdot, x)$ when $V$ is non-degenerate in dimension 2.

In a nutshell, the requirement is that the positive and negative regions of the Lyapunov function in a neighborhood of a point resemble topologically to the positive and negative part of a quadratic function (see [4] Figure 1). This is exactly what forbids pseudo-Anosov maps to have non-degenerate Lyapunov functions in neighborhood of their singularities.

The following theorem is the main theorem of [$L_1$]:

**Theorem 2.2.** *Let $f$ be a $C^1$-diffeomorphism with a non-degenerate Lyapunov function, then $f$ is topologically stable.*

In [$L_1$] a characterization of Anosov diffeomorphisms in terms of quadratic forms is also given. With this approach Lewowicz is able to recover classical results on structural stability of Anosov and characterization of Anosov systems in terms of cone-families. Quadratic forms turn out to be, in some applications, better suited for the study of the tangent map dynamics than cone-fields as we will try to explain in the next subsection.

2.1. **Quadratic forms, Lyapunov functions and Pesin's theory.** In [$L_1$] the following example of diffeomorphism of $\mathbb{T}^2$ which is not Anosov and yet admits a non-degenerate Lyapunov function is proposed:

$$F_c(x,y) = \left(2x - \frac{c}{2\pi}\sin(2\pi x) + y, x - \frac{c}{2\pi}\sin(2\pi x) + y\right)$$

For $c < 1$ the diffeomorphism is Anosov (being linear for $c = 0$). On the other hand, for $c = 1$ there is no invariant splitting by the differential in the tangent space of the fixed point $(0, 0)$. However, it can be proved that $F_1$ admits a non-degenerate Lyapunov function, it is volume preserving and also ergodic.

In [CE] it is proven that $F_1$ as well as many other examples in the boundary of Anosov diffeomorphisms of $\mathbb{T}^2$ are ergodic and non-uniformly hyperbolic. The proof of non-uniform hyperbolicity relies on the existence of certain quadratic forms which instead of verifying that their first difference is everywhere positive, they verify this almost-everywhere extending the results of [$L_1$]. The following result was stated without a complete proof in [LL] and the proof was completed in the appendix of [$Mar_1$]:

[4] The lines in my drawings are all crooked on purpose in order to show the topological nature of the objects, :).

**Theorem 2.3.** *Let $f$ be a volume preserving diffeomorphism admitting a continuous quadratic form $B : TM \to \mathbb{R}$ such that the quadratic form $f^\sharp(B) - B$ is definitely positive almost everywhere. Then, $f$ is non-uniformly hyperbolic, i.e. Lyapunov exponents are almost-everywhere non-vanishing.*

Here, we denote $f^\sharp(B)_x(v) = B_{f(x)}(D_x f v)$.

This Theorem was later extended in [$\text{Mar}_2$, K] and is quite related to a cone-criterium ([Wo]) but works better in some situations (see [$\text{Mar}_1$, K] and references therein). We will not enter into details about these important results, but we refer the reader to [CM] for further developments and applications to billiard systems. Let us just mention that in the spirit of Lewowicz phrase in the introduction of [$\text{L}_1$] and quoted above, the work of [$\text{Mar}_2$] proves a reciprocal statement to the quadratic form criterium, completing the parallelism with Lyapunov method and Massera's theorem (an important mentor for Lewowicz), see [Mas].

Let us close this section by mentioning a problem which Lewowicz has always insisted on:

**Question** (Problem 10.3 of [$\text{LC}_2$])**.** *For $c > 1$ does the Pesin region of $F_c$ has positive measure?*

The latter is a typical coexistence question which has always interested many mathematicians. The maps $F_c$ proposed by Lewowicz are similar to those of [Pry] (see also [Li]). See also the work of Pesin ([Pes]) on the coexistence problem which is one of the central problems in dynamics.

## 3. Persistence

3.1. **Persistence vs Topological stability.** The concept of persistence was introduced by Lewowicz in [$\text{L}_2$] in order to study some robust properties of certain expansive homeomorphisms under perturbations. In a certain way, it is a property which can be thought of as a dual property to shadowing.

If an expansive homeomorphism has the shadowing property then it is topologically stable (see [$\text{L}_4$]); nevertheless, not every expansive homeomorphism is topologically stable as we have already seen. All known expansive homeomorphisms do verify though this weaker notion of stability which is called *persistence* (or *semi-persistence*) a term coined by Lewowicz in [$\text{L}_2$] (see also [$\text{L}_5$]).

*Definition* (Persistence). We say that $f : M \to M$ is *persistent* if for every $\varepsilon > 0$ there exists a $C^0$-neighborhood $\mathcal{U}$ of $f$ such that for every $g \in \mathcal{U}$ and $x \in M$ there exists $y \in M$ such that

$$d(f^n(x), g^n(y)) \leq \varepsilon \qquad \forall n \in \mathbb{Z}$$

$\diamondsuit$

In Lewowicz words ([$\text{L}_2$]):

> "(...)*roughly, the dynamics of $f$ may be found in each $g$ close to $f$ in the $C^0$-topology; however, these $g$ may present dynamical features with no counterpart in $f$.*"

In his paper [$\text{L}_2$] Lewowicz proves some results concerning persistence such as persistence for pseudo-Anosov maps and more generally, for those expansive diffeomorphisms having a dense set of hyperbolic periodic points with codimension one. Those results

can be thought of as the germ of further developments in higher dimensions such as [$V_1$, $V_2$, $V_3$, ABP].

In particular, he shows that a small $C^1$-perturbation of a pseudo-Anosov map preserving the singularities must be conjugated to the original map; a kind of structural stability result for pseudo-Anosov maps. See also [Ha] for further developments.

Before we continue with the results of [$L_2$] and some of the consequences found by Lewowicz and coauthors, we are tempted to add another quote of [$L_2$]:

> "*We believe that, apart from such applications, there is another reason for studying these persistence properties: it seems plausible to think that if a theory of asymptotic behavior is possible, then semi-persistence (i.e. persistence of positive or negative semi-trajectories) should hold on big subsets of $M$ for large classes of dynamical systems*".

See [$L_5$] for advances in that hope. He posed a precise question about this problem:

**Question** (Problem 10.2 of [$LC_2$])**.** *For an expansive homeomorphism is every semitrajectory persistent in the future (in the past)?*

Another question which is motivated by this persistency property can be stated as follows:

**Question.** *Does an expansive homeomorphism minimize the entropy in its isotopy class?*

This is true for every known example, and it is true after Lewowicz classification result for surface homeomorphisms (see also [Ha]). If every expansive homeomorphism is persistent, then nearby homeomorphisms should have at least the same topological entropy. However, it seems that the answer of this fundamental question is at present far out of reach.

3.2. **Expansive systems and stable points.** Probably the first interaction found by Lewowicz between the topology of the phase space and expansive dynamics is the fact that a non-trivial compact connected and locally connected set admitting an expansive homeomorphism cannot have Lyapunov stable points. If connectedness is not required, this is clearly false as can be seen by considering an heteroclinic orbit between two fixed points. A more delicate example, where the phase space is connected but not locally connected can be found in [RR].

As a way to pave the way of some results in low dimensions which required hyperbolic periodic points to have codimension one in dimensions 2 and 3, Lewowicz proved in [$L_2$] the following result:

**Theorem 3.1** (No Stable Points)**.** *Let $f : M \to M$ be an expansive homeomorphism of a non-trivial compact connected and locally connected metric space, then $f$ has no Lyapunov stable points.*

Recall that a point $x$ is Lyapunov stable if for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $d(x, y) < \delta$ then $d(f^n(x), f^n(y)) < \varepsilon$ for every $n \geq 0$.

We give here a sketch of the proof of this important result:

Sketch Let $\alpha > 0$ be the expansivity constant of $f$. The proof is divided into 3 steps:

**Step 1:** For $\varepsilon < \alpha$, if

$$S_\varepsilon(x) = \{y \ : \ d(f^n(x), f^n(y)) \leq \varepsilon \ \ n \geq 0\}$$

then we have that the diameter of $f^n(S_\varepsilon(x))$ converges to zero uniformly on $x$ and $n$.

**Step 2:** If $x$ is a Lyapunov stable point, and $\varepsilon > 0$, there exists $\sigma > 0$ such that for $n \geq 0$ we have that $f^{-n}(S_\varepsilon(x))$ contains the ball of radius $\sigma$ of $f^{-n}(x)$.

**Step 3:** The previous step implies that every point in the $\alpha$-limit of $x$ is Lyapunov stable. One can prove using this fact and the first step that the $\alpha$-limit set must consist of periodic attractors which are only $\alpha$-limit points of their own orbit. This gives a contradiction, since it implies that the whole space is a periodic orbit, and being connected a unique point (contradicting that the space was non-trivial).

The hardest step is Step 2 and it is where local connectedness is used in an essential way. Roughly, using local connectedness, if the uniform ball cannot be obtained, one finds a sequence of points $x_n$, $y_n$ such that they are at distance larger than $\delta$ and remain at distance less than $\varepsilon$ for all future iterates and for arbitrarily large number of iterates in the past. Taking limits, one contradicts expansivity.

Let us explain briefly how to find such pair of points: If when iterating backwards the $\delta$-ball of $x$ there is no uniform ball, given $n > 0$ one can choose an arc $\gamma$ (or a connected set) with length smaller than $1/n$ and containing $f^{-k_n}(x)$ (where $k_n$ must necessarily tend to $+\infty$ as $n \to +\infty$) such that $f^{k_n}(\gamma)$ is not contained in $B_\delta(x)$. By connectedness there exist a point $y_n$ and a backward iterate $x_n = f^{-m_n}(x)$ at distance larger than $\delta$ and such that $d(f^j(x_n), f^j(y_n)) \leq \varepsilon$ for every $j \geq -k_n + m_n$.

Since the points $x_n$ and $y_n$ are at distance larger than $\delta$ and its $k_n - m_n$ backward iterate sends them at distance less than $1/n$ we get that $k_n - m_n$ also goes to $+\infty$ as $n \to \infty$. Taking convergent subsequences of $x_n$ and $y_n$ one obtains different points whose orbits remain at less than $\varepsilon$ for all iterates contradicting expansivity.

□

3.3. **Analytic models of pseudo-Anosov maps.** In his paper [LL] with E. Lima de Sa, they provide a new construction of analytic models of pseudo-Anosov maps that had been obtained by Gerber ([Ge]) based on previous work by Gerber with Katok ([GeK]).

The idea is to replace their conditional stability results by the structural stability theorem of Lewowicz ([$L_2$]) for pseudo-Anosov maps involving the concept of persistence.

It is important to remark that constructing analytic (even smooth) models of pseudo-Anosov maps is not easy since by a change of coordinates which is $C^1$ out of a neighborhood of the singularities one cannot obtain a smooth model (this was shown in [GeK]), so a more global modification must be made.

The idea involves "slowing down" in a neighborhood of the singularities (much as one does if one wants to smooth the parametrization of a curve having a corner in its image without altering the image) and then approximating by analytic maps which preserve the singularities as well as some $r$-jets of the derivative of the map in the singularity. This allows to use the mentioned Lewowicz's results on persistence ([$L_2$]).

To show how this creation of models is far from being trivial, let me state an open problem which we are far from understanding. This question was strongly motivated by discussions with Jorge Lewowicz and his constant insistence on the lack of understanding we have of the role of the dynamics of the tangent map (see also the next subsection for related problems):

**Question.** *Let $f : M \to M$ be a* topological Anosov *(i.e. A homeomorphism of $M$ which preserves two topologically transverse foliations one of which contracts distances uniformly and the other one contracts them for backward iterations). Does there exist*

*a smooth model for $f$? And analytic?. Assuming the previous questions have positive answers, can these models be made Anosov?.*

♢

The question admits a positive answer both in the codimension one case and in the case where $M$ is a nilmanifold due to the fact that the classification results of Newhouse-Franks-Manning only use the fact that the map is a topological Anosov. However, the question is completely independent a priori of the classification of Anosov systems, and a proof in dimension 2 without using the global classification theory would already be interesting.

One of the main contributions of [LL], though lateral to the paper has already been explained in this note, and has to do to the way they prove that the resulting approximation maps is still Bernoulli with respect to Lebesgue measure (which can be thought of as the counterpart of the second part of the question above). To do this, they use quadratic forms and that is the germ of further results on non-uniform hyperbolicity as we have already mentioned.

3.4. **The $C^0$-boundary of Anosov diffeomorphisms.** In this section we state a result obtain by Lewowicz in colaboration with J. Tolosa about the $C^0$-boundary of codimension one Anosov diffeomorphisms (see [LT]).

They prove:

**Theorem 3.2.** *Let $f$ be an expansive homeomorphism in the $C^0$-boundary of Anosov diffeomorphisms of codimension one in $\mathbb{T}^d$. Then, $f$ is conjugated to an Anosov.*

With his classification result for expansive homeomorphisms of the torus, this can be further improved to get:

**Theorem 3.3.** *Let $f : \mathbb{T}^2 \to \mathbb{T}^2$ be an expansive homeomorphism. Then $f$ is contained in the $C^0$-closure of the set of Anosov diffeomorphisms of $\mathbb{T}^2$.*

Proof. Consider $h : \mathbb{T}^2 \to \mathbb{T}^2$ a homeomorphism isotopic to the identity such that $f = h \circ A \circ h^{-1}$ where $A$ is a linear Anosov automorphism. The existence of such an $h$ is given by Theorem 4.1 below.

Then there exists a sequence of diffeomorphisms $h_n$ converging to $h$ in the $C^0$-topology and such that $h_n^{-1}$ also converges to $h^{-1}$. Since conjugating an Anosov diffeomorphism by a diffeomorphism gives an Anosov diffeomorphism we get that $f$ is approximated in the $C^0$-topology by Anosov diffeomorphisms.

□

An important open question that is motivated by this result is the following:

**Question** (Problem 10.1 of [LC$_2$])**.** *Does the $C^1$-closure of Anosov diffeomorphisms contains all expansive diffeomorphisms of $\mathbb{T}^2$?*

Notice also that Mañé has proved that the $C^1$-interior of expansive diffeomorphisms consists of Quasi-Anosov ones[5], in particular in $\mathbb{T}^2$ of Anosov ones ([Ma$_1$]).

Notice that the set of Anosov diffeomorphisms in an given isotopy class of $\mathbb{T}^2$ forms a connected set (see [FG]).

[5]Is in this paper that Mañé introduces the concept of *dominated splitting*.

## 4. Classification theorem in surfaces

It can be shown easily that the only closed one dimensional manifold, namely the circle, admits no expansive homeomorphisms. This can be proved using the Poincare's classification of homeomorphisms of the circle by discussing depending on the rotation number. Other than that, some examples and some results on the non-existence of expansive homeomorphisms of other one dimensional continua, nothing was known about the existence or structure of expansive homeomorphisms. It is to be remarked that Mañé proved ([$Ma_2$]) that if a compact metric space admits an expansive homeomorphism, then it must have finite topological dimension.

Examples in every orientable surface different from the sphere were already known ([OR]), but there was no clue for example on which isotopy classes admitted them. The classification of expansive homeomorphisms of surfaces was thus meant to be started from scratch and that was what Lewowicz did ([$L_3$]): He gained an impressive understanding of their dynamics and their relation with the topology of the phase space and one of the most striking aspects of his study is that he relied only on some well known and almost elementary properties of plane topology. Of course, once he got a classification of expansive homeomorphisms in terms of their dynamics and local behavior, the final form of the result, giving conjugation to already known models, used some less elementary techniques ([Fr, Th]).

The starting point was the non existence of stable points proved by him in [$L_2$] and reviewed in the previous section. In this section we will give an overview of the classification results for expansive homeomorphisms of surfaces and the main ideas involved in the proof. We recall that as we said in the introduction, these results were obtained independently by Hiraide [H].

What we will provide is far from a complete proof of this classification result, but we hope that the outline here can be used as a guide to read the original paper [$L_3$] and to obtain some insight on the proof.

### 4.1. **Statement of the result.**

Along this section, $S$ will denote an orientable closed (compact, connected, without boundary) surface. It is well known that these surfaces are well characterized by their Euler characteristic, and consist of the sphere $S^2$, the torus $\mathbb{T}^2$ and the higher genus surfaces $S_g$ with $g \geq 2$.

The main result of [$L_3$] is the following:

**Theorem 4.1** (Classification of expansive homeomorphisms of surfaces)**.** *Let $f : S \to S$ an expansive homeomorphism. Then, $S \neq S^2$ and:*

- *If $S = \mathbb{T}^2$ then $f$ is conjugate to a linear Anosov automorphism.*
- *If $S = S_g$ then $f$ is conjugate to a pseudo-Anosov map (*[Th]*).*

As we mentioned, Lewowicz result has two parts, first, he gives a complete dynamical classification of expansive homeomorphisms by a detailed study of the stable and unstable sets of all the points in $S$, obtaining for them a local product structure outside some finite set of "singularities" which have a local behavior much like those of pseudo-Anosov maps. Then, by using global arguments and shadowing results he obtains the desired conjugacy.

Lewowicz result can be though of in a now very fashionable way called *rigidity*: Rigidity results (or non-existence results) are those which give strong restrictions from a priori very weak ones. In the words of Frederic Le Roux in [Ler]: "(...)*a simple dynamical property can imply a strong rigidity. The most striking result here is probably*

*Hiraide-Lewowicz theorem that an expansive homeomorphism on a compact surface is conjugate to a pseudoAnosov homeomorphism*".

4.2. **Stable and unstable sets.** This and the next will be the more technical sections of this note. However, we will try to first give a statement which will be proved in these two sections which will allow the reader to continue. Then, we will enter in some details.

Let $f : S \to S$ be an expansive homeomorphism with expansivity constant equal to $\alpha$. Consider the following sets:

$$S_\varepsilon(x) = \{y \in S \ : \ d(f^n(x), f^n(y)) \leq \varepsilon \ \ n \geq 0\}$$

$$U_\varepsilon(x) = \{y \in S \ : \ d(f^{-n}(x), f^{-n}(y)) \leq \varepsilon \ \ n \geq 0\}$$

Expansivity can be reformulated as $S_\alpha(x) \cap U_\alpha(x) = \{x\}$ for every $x \in S$. We call $S_\varepsilon(x)$ (resp. $U_\varepsilon(x)$) the $\varepsilon$-stable set of $x$ (resp. $\varepsilon$-unstable set of $x$).

As we mentioned in the previous section, it can be easily proved that the diameter of $f^n(S_\varepsilon(x))$ converges to zero uniformly independently of $x$ if $\varepsilon < \alpha$. The key technical result in the classification of expansive homeomorphisms of surfaces can be stated in terms of these sets:

**Theorem 4.2** (Classification Theorem Local Version)**.** *Let $f : S \to S$ be an expansive homeomorphism. Then, there exists a finite set $F$ (possibly empty) such that for every $x \in S \backslash F$ we have that there exists $\varepsilon > 0$ such that $S_\varepsilon(x)$ is a continuous arc having $x$ in its interior. Moreover, there exists a neighborhood $U$ of $x$ having local product structure. For $x \in F$ we have that the sets $S_\varepsilon(x) \backslash \{x\}$ and $U_\varepsilon(x) \backslash \{x\}$ are both a finite number ($\geq 3$) of arcs which are alternated and in each angle they form, there is also local product structure.*

We must explain some of the terminology appearing in the statement (see also Figure 2 for a visual explanation).

Local product structure means the following: We say that in an open set $U$ centered in $x$ there is *local product structure* if there is a homeomorphism

$$h : [-1, 1] \times [-1, 1] \to \overline{U}$$

such that $h(0, 0) = x$ and $h(\{t_0\} \times [-1, 1])$ is contained in a stable set $S_\varepsilon(h(t_0, 0))$ and $h([-1, 1] \times \{s_0\})$ is contained in an unstable set $U_\varepsilon(h(0, s_0))$.

In a similar way, given a point $x$, we can consider a connected component $L^s$ of $S_\varepsilon(x) \backslash \{x\}$ and a connected component $L^u$ of $U_\varepsilon(x) \backslash \{x\}$. If $U$ is a neighborhood of $x$ and $A$ is a connected component of $U \backslash (L^s \cup L^u \cup \{x\})$ which does not intersect $S_\varepsilon(x) \cup U_\varepsilon(x)$ we say that $A$ is an *angle*. We say that the angle has local product structure if a similar property as above holds except that $h : [0, 1] \times [0, 1] \to \overline{A}$ and it sends $h(0, 0) = x$ with the rest of the properties being equal (see Figure 2).

Before we continue with a sketch of the proof of this result, let us make some comments on some existing extensions. First, similar properties have been obtained for expansive flows in dimension 3 ([$\text{Pat}_1$]). Also, by assuming the existence of a dense set of topologically hyperbolic periodic points these results can be extended to any dimension ([$\text{V}_1$, ABP]), except that the behavior in the singularities is not well understood[6] except in dimension 3 or in the codimension one case ([$\text{V}_2$, ABP]) where one can show that they do not exist. With some differentiability assumptions, the hypothesis of the

[6]It seems that we lack examples of "genuine" pseudo-Anosov maps in higher dimensions.

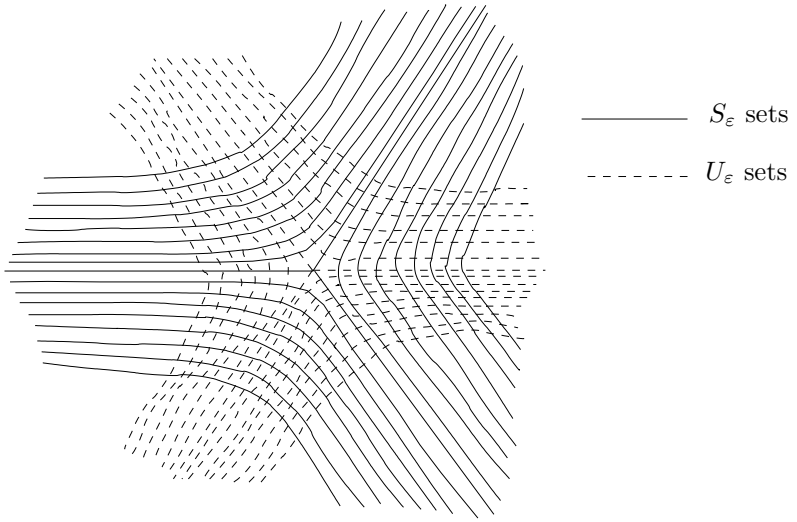


FIGURE 2. Local picture at a singular point with $p = 3$ "legs".

existence of periodic points can be removed, at least in dimension 3 ([V$_3$]). This has also been extended to the plane under certain conditions of the behavior at infinity ([Gr$_1$, Gr$_2$]).

Just to show how far we are from obtaining a similar result in higher dimensions let me state the following open question (it is known in the smooth case in $\mathbb{T}^3$, see [V$_3$]):

**Question.** *Has every expansive homeomorphism of a manifold a periodic point?*

Now, let us discuss the main points of the proof of Theorem 4.2. Let us remark that each of this steps are interesting by themselves, and some of them hold in higher dimensions.

The first step of the proof consists on showing that every point in $S$ has a local stable and unstable set of uniform size.

**Proposition 4.1.** *For $f : S \to S$ expansive homeomorphism and $\varepsilon < \alpha$ the expansivity constant, there exists $\delta > 0$ such that for every $x \in S$ we have that the connected component of $S_\varepsilon(x) \cap B_\delta(x)$ containing $x$ intersects $\partial B_\delta(x)$.*

SKETCH The proof of this proposition holds in any dimension. The key point is the non existence of Lyapunov stable (in fact, Lyapunov unstable) points proven in Theorem 3.1.

Once this is obtained, one can construct large connected sets by considering the sets $D_n$ build as the connected component containing $x$ of $n$-th preimage of the ball of radius $\varepsilon$ centered at $f^n(x)$ by $f^{-n}$. The fact that there are no Lyapunov unstable points allows one to prove that these sets have all diameter bounded from below and allow to construct the desired set as

$$C^s_\varepsilon(x) = \bigcap_N \overline{\bigcup_{n \geq N} D_n}$$

One has to check that this has the desired properties (see [L$_3$] Lemma 2.1), in particular that the sets $D_n$ have diameter bounded from bellow. This can be done using Lyapunov

functions and the metric they define, or using the metric introuduced in [Fa]. Also, it can be done by barehanded arguments (see [L$_4$]).

□

We remark that the previous result gives a conceptual proof that $S^1$ does not admit expansive homeomorphisms: On the one hand they cannot have Lyapunov stable points, but on the other hand the stable set of a point must contain a connected set of large diameter, thus, non-empty interior, a contradiction.

We make a remark on stable and unstable sets which is of importance in many steps of the proof. It can be thought of as a "big angles" result. The proof is not difficult (see Lemma 3.3 of [ABP]).

**Proposition 4.2** (Big Angles)**.** *Let $f : S \to S$ be an expansive homeomorphism with expansivity constant $\alpha$. Given $V \subset U$ neighborhoods of $x$ and $\rho > 0$ small enough, there exists a neighborhood $W \subset V$ of $x$ such that if $y, z \in W$ we have that $d(S_\varepsilon(y) \cap U \backslash V, U_\varepsilon(z) \cap U \backslash V) > \rho$.*

The next step of the proof is probably the deepest and it is really dependent on the two-dimensionality of the problem. Here one sees a clear manifestation of the already quoted phrase of "*a stronger interaction of the topology of $M$ and the dynamics of $f$ could be expected*".

**Theorem 4.3.** *For an expansive homeomorphism $f : S \to S$ with expansivity constant $\alpha$ and $\varepsilon < \alpha/10$, the connected component of $S_\varepsilon(x)$ containing $x$ is locally connected at each of its points and therefore arc-connected.*

Sketch We will only give a brief outline with an heuristic idea of this subtle proof. We refer the reader to [L$_3$] pages 119-121 for details (see also [L$_4$] pages 21-25).

Consider $C^s_\varepsilon(x)$ the connected component of $S_\varepsilon(x)$ containing $x$. We first show that it is locally connected at $x$ and then a clever argument allows to show local connectedness at every point. Once this is proved, arc-connectedness follows since a compact connected and locally connected set is arc-connected.

The proof is by contradiction. Roughly, the idea is that if it is not locally connected at $x$ we can think that in an arbitrarily small ball of $x$ the set $S_\varepsilon(x)$ is a sequence of connected sets approaching $x$ but connecting to $C^s_\varepsilon(x)$ outside the ball. Using separation properties of the plane (which are extensions of Jordan's curve theorem) we obtain some point $z$ which is trapped in both sides by connected components of $S_\varepsilon(x)$. Since the unstable set of $z$ has a large connected component containing $z$, we know it must leave the neighborhood, however, it can intersect $S_\varepsilon(x)$ only once, so we obtain that it leaves forming "small angles" with $S_\varepsilon(x)$ a contradiction with Proposition 4.2.

In fact, there are some subtleties in what we have just said, since the fact that the unstable set of $z$ has a large connected component does not imply that it must have two sides, and there is no problem to have one side going out by intersecting $S_\varepsilon(x)$. To solve this, Lewowicz makes a clever argument that he then repeats several times in his proof and so we partially reproduce it here: He considers an arc joining two different connected components of $C^s_\varepsilon(x)$ locally and he divides the arc depending on which side the unstable set of the points leave the neighborhood: a connectedness argument allows him to conclude that either there is a point whose unstable intersects twice $S_\varepsilon(x)$ (contradicting expansivity) or a point whose unstable leaves forming small angles (also

a contradiction). This connectedness argument uses the fact that stable and unstable sets vary semicontinuously[7].

Now, to get local connectedness at every point, we use local connectedness at the centers at many scales. Consider $y \in C_\varepsilon(x) \subset C_{2\varepsilon}(y)$. Then $C_{2\varepsilon}(y)$ is locally connected at $y$, so for every $\sigma > 0$ and $z \in C_\varepsilon(x)$ close to $y$ there exists a connected set $C \subset C_{2\varepsilon}(y) \cap B_\sigma(y)$ containing $y$ and $z$. Since there are no stable points, we know that $C \cup C_\varepsilon(x) \subset C_{2\varepsilon}(y)$ cannot separate, so, by an extension of Jordan's separation theorem we get that $C_\varepsilon(x) \cap C$ is connected and we deduce that $C_\varepsilon(x)$ is locally connected at $y$.

This finishes the sketch of the proof.

□

We will give an outline of the rest of the proof of Theorem 4.2 in the next subsection. We will omit even more details.

4.3. **Singularities.** The purpose of this section is to outline the rest of the proof of Theorem 4.2. We will not enter in details here, we will only explain the main steps of the proof.

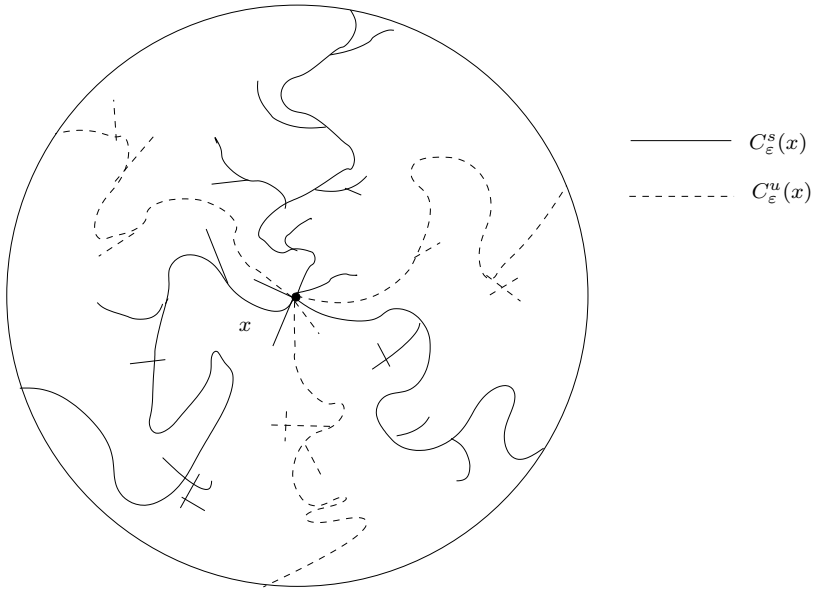


FIGURE 3. Structure of local stable and unstable sets. They are arc connected but may a priori be still "ugly".

Pick a point $x \in S$. As we have already seen, $S_\varepsilon(x)$ has at least one connected component intersecting $\partial B_\delta(x)$. If we consider the connected component $C^s_\varepsilon(x)$ of $S_\varepsilon(x) \cap B_\delta(x)$ and the connected component $C^u_\varepsilon(x)$ of $U_\varepsilon(x) \cap B_\delta(x)$ we know that there are arcs joining $x$ to $\partial B_\delta(x)$ contained in those sets. It is possible to make an equivalence relation between these arcs that identify arcs which start at $x$ and then bifurcate near the boundary of $B_\delta(x)$. By using this, the big angles property and connected arguments similar to the ones used in the previous section, Lewowicz shows:

[7]This is a general property that holds for any homeomorphism and it is not hard to check. See for example Lemma 3.2 of [ABP].

**Lemma 4.1.** *The number of (equivalence classes of) arcs in $C^s_\varepsilon(x)$ and $C^u_\varepsilon(x)$ joining $x$ to the boundary of $B_\delta(x)$ is the same, finite, and moreover, they are alternated in the order of $\partial B_\delta(x)$.*

This result together with the invariance of domain theorem and further application of the previous arguments give the following property around points which is almost the end of the proof of Theorem 4.2.

**Proposition 4.3.** *For every $x \in S$, there exists a neighborhood $U$ such that every point $y$ in $U \backslash \{x\}$ has a neighborhood with local product structure.*

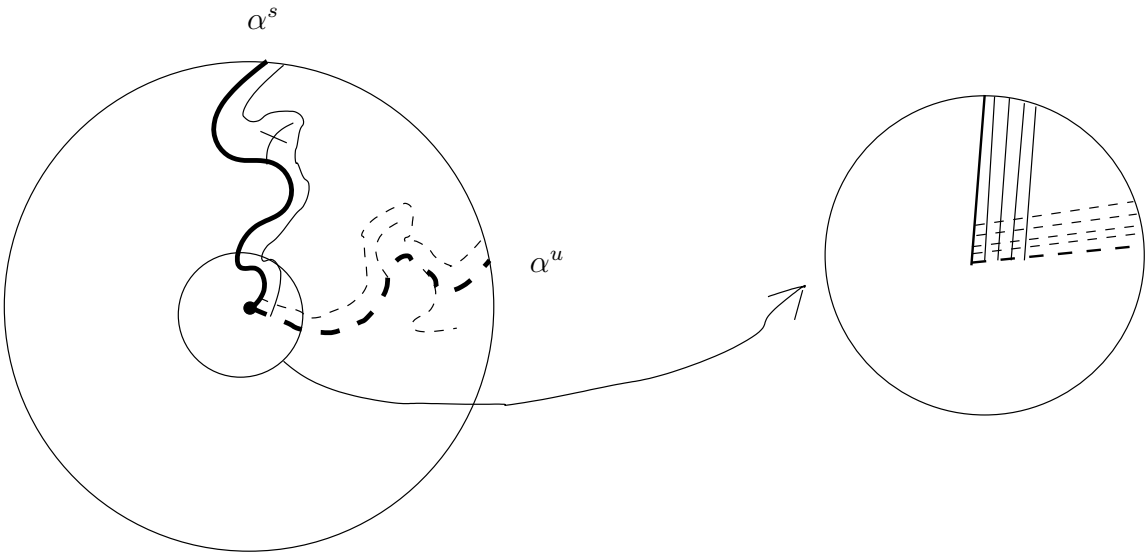


FIGURE 4. How to obtain local product structure.

This is proved as follows: Consider an arc $\alpha^s$ of $C^s_\varepsilon(x)$ and a consecutive one $\alpha^u$ of $C^u_\varepsilon(x)$. Now, for points of $\alpha^u$ close to $x$ we have that using semicontinuous variation[8] of stable sets and the "big angles" property (Proposition 4.2) that the stable set of the points near $x$ goes out of $B_\delta(x)$ near $\alpha^s$. The same happens for points in $\alpha^s$ near $x$ and their unstable sets. This allows to find a continuous and injective (due to expansivity) map from a neighborhood of $x$ in $\alpha^s$ times a neighborhood of $x$ in $\alpha^u$ into $S$. By the invariance of domain theorem this map is open and thus every point in this "angle" has local product structure. This can be done in all the angles formed by the stable and unstable arcs of $x$ (see Figure 4).

It is immediate to conclude that:

**Corollary 4.1.** *There exists a finite set $F \subset S$ such that every point outside of which every point has local product structure. Moreover, for $x$ in $F$ have a neighborhood such that their local stable and unstable sets are $p \geq 1$ (and different from 2 which would imply local product structure around $x$) arcs starting at $x$ and arriving at the boundary.*

[8]A disclaimer is that to be precise, this argument needs that there are at least two arcs of stable and two arcs of unstable for $x$. We will ignore this problem and "solve it" afterwards because we believe it gives a better heuristic of the global argument. See [L$_3$] for a correct proof.

It remains only to discard the possibility of having a unique arc in the stable set of $x$. This is an important issue since for example $S^2$ admits diffeomorphisms (even analytic, see [Ge, LL]) that have the local form we have obtained but with points having a singularity with a unique "leg". Needless to say, those examples are not expansive, since for points very near to $x$ in the stable set, very small horseshoes are created, contradicting expansivity. Building in this example, and using the arguments developed by Lewowicz for the other parts of the proof, one can give a general proof of the following (see also Figure 5):

**Proposition 4.4.** *The number $p$ in the above corollary is $\geq 3$ for every point in $F$.*

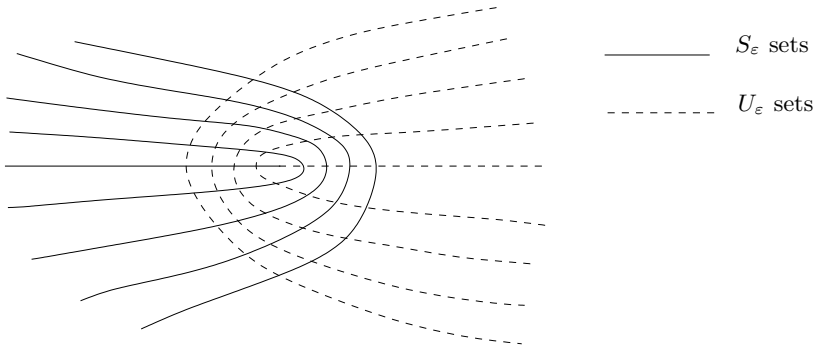


FIGURE 5. One leg implies that local stable and local unstable sets intersect in more than one point contradicting expansivity.

This concludes the outline of the proof of Theorem 4.2.

4.4. **Non-existence of expansive homeomorphisms on $S^2$.** We show here how Theorem 4.2 is enough to show that the two-dimensional sphere $S^2$ cannot admit expansive homeomorphisms.

The easiest way to see this is using index theory for foliations. The local product structure obtained allows one to see that stable and unstable sets foliate the surface admitting and expansive homeomorphisms giving rise to a continuous foliation with finitely many singularities of prong type. Since for every singularity the number of legs is $\geq 3$ we deduce that even if the foliation may be non-orientable then the index of the singularities is always negative (notice that if there were only one leg, then the index is positive and equal to $1/2$ so that one can make one example in $S^2$ with four such singularities). This implies that $S^2$ cannot support such a homeomorphism.

If the reader is not comfortable with the use of continuous (and not differentiable) foliations, one can go to [$L_3$] where a more elementary proof is given using Poincare-Bendixon's like arguments.

4.5. **Other surfaces.** In the torus case, essentially, due to the work of Franks, it is enough to show that there are no singularities (which is clear by the index argument shown above) and that the map is isotopic to a linear Anosov automorphism. Consider then $f : \mathbb{T}^2 \to \mathbb{T}^2$ an expansive homeomorphism.

Although he might have used the already known argument on the growth of periodic points and Lefshetz index, Lewowicz gives a different argument which is very beautiful[9].

I outline it here: Lift $f$ to the universal cover to obtain

$$\tilde{f} : \mathbb{R}^2 \to \mathbb{R}^2$$

Let $A \in GL(2, \mathbb{Z})$ be its linear part (i.e. $A$ is the matrix given by $\tilde{f}(\cdot) - \tilde{f}(0) : \mathbb{Z}^2 \to \mathbb{Z}^2$). If $A$ is not hyperbolic, then it has both eigenvalues of modulus 1 since it has determinant of modulus 1. Then, one obtains that the diameter of a set iterated by $A$ grows at most polynomially. Since $\tilde{f}$ is at bounded distance from $A$, the same holds for the iterates of a set by $\tilde{f}$. If $J$ is an unstable arc contained in a local product structure box, one gets that $\operatorname{diam}(\tilde{f}^n(J)) \leq p(n)$ where $p$ is a polynomial.

On the other hand, we know that the length[10] of an unstable arc by $\tilde{f}$ must grow exponentially due to expansivity, so, for the same $J$ we get that the length of $\tilde{f}^n(J)$ is comparable to $\lambda^n$ with $\lambda > 1$. Moreover, since there are no singularities, a Poincare-Bendixon's like type of argument implies that an arc of unstable cannot intersect the same box of local product structure twice. This implies, via the quadratic growth of volume of $\mathbb{R}^2$ that the diameter of an arc of unstable of length $L$ is comparable to $\sqrt{L}$ which will still be exponential. This gives a contradiction and completes the proof. See [$L_3$] Theorem 5.3 for more details.

In the higher genus case the proof is even more delicate. He again stands on previous conjugacy results by Handel [Ha] (improving the results of [$L_2$]) that state that in the isotopy class of a pseudo-Anosov map there exist certain semiconjugacies. Then, as in the torus case he must prove that the local classification theorem (Theorem 4.2) provides enough tools to show that $f$ is isotopic to pseudo-Anosov. He uses Thurston's classification and shows that no homotopy class of simple curves can be periodic (see Lemma 6.4 of [$L_3$]) which allows him to conclude.

□

**Acknowledgments:** My understanding of the theorem of classification on surfaces owes tremendously to conversations with A. Artigue and J. Brum. The writing of this note also benefited from many conversations, exchange and suggestions of: D.Armentano, E. Catsigeras, M. Cerminara, M.Delbracio, N. de Leon, H. Enrich, J. Groisman, P. Lessa, R. Markarian, A. Passeggi, M. Paternain, R. Ruggiero, A. Sambarino, M. Sambarino and J. Vieitez. Of course, I want to thank particularly Jorge Lewowicz, for his work and all the enlightening conversations we had, mathematically and otherwise.

## References

[ABP] A. Artigue, J. Brum, R. Potrie, Local product structure for expansive homeomorphisms, *Topology and its Applications* **156** (2009), no. 4, 674–685.

[CE] E. Catsigeras and H. Enrich, SRB measures of certain almost hyperbolic diffeomorphisms with a tangency. *Discrete Contin. Dynam. Systems* **7** (2001), no. 1, 177-202.

[CM] N. Chernov and R. Markarian, *Introduction to the ergodic theory of chaotic billiards*, 2nd. Edition, IMPA, Rio de Janeiro, 208 pp (2003).

[FG] F.T. Farrel, A. Gogolev, The space of Anosov diffeomorphisms, *preprint* arXiv:1201.3595.

[9] I could not trace a similar argument to before Lewowicz's paper. However, this kind of argument has been rediscovered many times so I do not claim that it is the first time it appeared. I show it in order to stress the continous search of Lewowicz for understanding and for conceptual and clean arguments.

[10] Since it is a continuous arc this is not really well defined. One can measure thus length by "counting" the number of local product structure boxes it intersects.

[Fa] A. Fathi, Expansiveness, hyperbolicity and Hausdorff dimension. *Comm. Math. Phys.* 126 (1989), no. 2, 249-262.

[Fr] J. Franks, Anosov Diffeomorphisms., *Proc. Sympos. Pure Math.*, vol. 14 (1970), 61–93.

[FR] J. Franks and C. Robinson, A Quasi-Anosov Diffeomorphism that is not Anosov. *Trans. Am. Math. Soc.*, **233** (1976), 267–278,

[Ge] M. Gerber, Conditional stability and real analytic pseudo-Anosov maps. *Mem. Amer. Math. Soc.* 54 (1985), no. 321, iv+116 pp.

[GeK] M. Gerber, A. Katok, Smooth models of Thurston's pseudo-Anosov maps. *Ann. Sci. cole Norm. Sup.* (4) 15 (1982), no. 1, 173-204.

[$Gr_1$] J. Groisman, Expansive homeomorphisms of the plane. *Discrete Contin. Dyn. Syst.* 29 (2011), no. 1, 213-239.

[$Gr_2$] J. Groisman, Expansive and fixed point free homeomorphisms of the plane. *Discrete Contin. Dyn. Syst.* 32 (2012), no. 5, 1709-1721.

[Ha] M. Handel, Global shadowing of pseudo-Anosov homeomorphisms. *Ergodic Theory Dynam. Systems* 5 (1985), no. 3, 373-377.

[H] K. Hiraide, Expansive homeomorphisms of compact surfaces are pseudo-Anosov. *Osaka J. Math* **27**, (1990), 117–162.

[K] A. Katok, Infinitesimal Lyapunov functions, invariant cone families and stochastic properties of smooth dynamical systems. With the collaboration of Keith Burns. *Ergodic Theory Dynam. Systems* **14** (1994), no. 4, 757-785.

[Ler] F. Le Roux, A topological characterization of holomorphic parabolic germs in the plane. *Fund. Math.* **198** (2008), no. 1, 77-94.

[$L_1$] J. Lewowicz, Lyapunov functions and topological stability, *J. of Diff. Equations* **38** (1980) 192209.

[$L_2$] J. Lewowicz, Persistence in expansive systems., *Ergodic Theory and Dynamical Systems*, **3** (1983), 567–578.

[$L_3$] J. Lewowicz, Expansive Homemorphisms of Surfaces, *Bol. Soc. Bras. de Mat.*, **20** (1989), 113–133,

[$L_4$] J. Lewowicz, *Dinámica de los homeomorphismos expansivos*, Monografias del IMCA **36** (2003).

[$L_5$] J. Lewowicz, Persistence of semi trajectories , *Journal of Dynamics and Differential Equations* **18** (2008), 1095-1102.

[$LC_1$] J.Lewowicz and M. Cerminara, Expansive systems, *Scholarpedia* **3**(12):2927 revision 91247.

[$LC_2$] J. Lewowicz and M. Cerminara, Some open problems concerning expansive systems. *Rend. Istit. Mat. Univ. Trieste* **42** (2010), 129-141.

[LL] J. Lewowicz, E. Lima de Sa, Analytic models of pseudo-Anosov maps. *Ergodic Theory Dynam. Systems* 6 (1986), no. 3, 385392.

[LT] J. Lewowicz, J. Tolosa, On expansive diffeomorphisms in the $C^0$-border of the set of Anosov diffeomorphisms. *Dynamical systems and partial differential equations* (Caracas, 1984), 57-64, Univ. Simon Bolivar, Caracas, 1986.

[Li] C. Liverani, Birth of an elliptic island in a chaotic sea. *Math. Phys. Electron. J.* 10 (2004), Paper 1, 13 pp (Electronic).

[$Ma_1$] R. Mañé, Expansive diffeomorphisms, *Lecture Notes in Math.* **468** (1975), 162-174.

[$Ma_2$] R. Mañé, Expansive homeomorphisms and topological dimension. *Trans. Amer. Math. Soc.* 252 (1979), 313-319.

[$Mar_1$] R. Markarian, Billiards with Pesin region of measure one. *Comm. Math. Phys.* **118** (1988), no. 1, 8797.

[$Mar_2$] R. Markarian, Non-uniformly hyperbolic billiards. *Ann. Fac. Sci. Toulouse Math.* (6) 3 (1994), no. 2, 223257.

[Mas] J.L. Massera,On Liapounoff's conditions of stability, *Ann. of Math.* 2 (50). pp. 705-721 (1949).

[OR] T. O'Brien,W. Reddy, Each compact orientable surface of positive genus admits an expansive homeomorphism. *Pacific J. Math.* **35** (1970) 737-741.

[$Pat_1$] M. Paternain, Expansive flows and the fundamental group. *Bol. Soc. Brasil. Mat.* (N.S.) **24** (1993), no. 2, 179-199.

[Pes] Y. Pesin, V. Climenhaga, Open problems in the theory of non-uniform hyperbolicity. *Discrete Contin. Dyn. Syst.* **27** (2010), no. 2, 589-607.

[Pry] F.Przytycki, Examples of conservative diffeomorphisms of the two-dimensional torus with coexistence of elliptic and stochastic behaviour, *Ergodic Theory Dynam. Systems*, 2, no. 34, 439-463 (1982).

[Re] W. Reddy, Expansive canonical coordinates are hyperbolic, *Topology and its Appl.*,**13** (1982), 327–334.

[RR] Reddy W, Robertson L, Sources, sinks and saddles for expansive homeomorphism with canonical coordinates, *Rocky Mt. J. Math.* **17** (1987) 673-681

[Ru] R. O. Ruggiero, *Dynamics and global geometry of manifolds without conjugate points*. Ensaios Matemticos [Mathematical Surveys], 12. Sociedade Brasileira de Matemtica, Rio de Janeiro, 2007. iv+181 pp.

[Sam] M. Sambarino, Laudatio of the Honoris Causa Diploma given by UdelaR to Lewowicz, 15 August 2012. Reprinted in this volume of the PMU.

[Th] W. Thurston, On the geometry and dynamics of diffeomorphisms of surfaces. *Bull. Amer. Math. Soc.* (N.S.) 19 (1988), no. 2, 417-431.

[$V_1$] J.L. Vieitez, Three dimensional expansive homeomorphisms. *Pitman Research Notes in Math.* **285**, (1993), 299–323.

[$V_2$] J.L. Vieitez, Expansive homeomorphisms and hyperbolic diffeomorphisms on three manifolds. *Ergodic Theory and Dynamical Systems*, **16** (1996), 591– 622.

[$V_3$] J.L. Vieitez, Lyapunov functions and expansive diffeomorphisms on 3D manifolds. *Ergodic Theory and Dynamical Systems*, **22** (2002), 601–632.

[Wa] P. Walters, Anosov diffeomorphisms are topologically stable, *Topology* **9** (1970) 71–78.

[Wo] M.P. Wojtkowski, Invariant families of cones and Lyapunov exponents, *Ergodic Theory Dynamical Systems* 5 (1985) 145-161.

CMAT, Facultad de Ciencias, Universidad de la República, Uruguay
*E-mail address*: rpotrie@cmat.edu.uy

# EXPANSIVE GEODESIC FLOWS: FROM THE WORK OF J. LEWOWICZ IN LOW DIMENSIONS TO GLOBAL GEOMETRY OF MANIFOLDS WITHOUT CONJUGATE POINTS

RAFAEL O. RUGGIERO

Abstract. The works of Jorge Lewowicz about expansive homeomorphisms of compact surfaces had remarkable impact in the theory of geodesic flows without conjugate points. We present a survey of results about expansive and weakly stable geodesic flows in compact manifolds without conjugate points, starting from Lewowicz's results and views about expansive dynamics in low dimensions, continuing with their generalizations in higher dimensions, and finishing with recent developments of the theory of weakly stable geodesic flows and their connections with Gromov hyperbolic spaces, control theory and Finsler rigidity.

## Introduction

Expansiveness is one of the most important features of hyperbolic dynamics. Its role in the study of the topological dynamics of hyperbolic systems is crucial, specially concerning stability theorems in weak or strong form (structural or topological stability, pseudo-orbit tracing properties, persistence, etc). Although hyperbolic dynamics implies in general expansiveness, the converse of this assertion is not true. It is not difficult to exhibit examples of expansive, non-hyperbolic systems some of which will be mentioned along the present exposition. The seminal works of Bowen [20] and Walters [129] in the 1970's showed many interesting and intriguing properties of expansive systems which are common to hyperbolic ones. Expansiveness played an important role in the whole body of work developed to prove the stability conjecture, a problem set in the 1960's by S. Smale and solved by Mañé [86] and Liao [74] in dimension 2 for diffeomorphisms and by Mañé [84] in any dimension in the 1980's. In those times and context the work of J. Lewowicz about expansive systems started by the end of the 1970's, after a long and fruitful experience with hyperbolic dynamics in differential equations going back to the 1960's. J. Lewowicz's academic legacy, not only as a researcher but also as a professor, advisor and colleague, left in many latin-american mathematicians (including me) a deep and definitive trace. The purpose of this survey is to pay a modest tribute to his work and academic life with a panorama of the theory of expansive and weakly stable geodesic flows, presented from the perspective of his pioneer work. Another survey by R. Potrie [101] contains a quite complete description of J. Lewowicz work with expansive and topologically stable non-conservative dynamics, so we shall devote ourselves to the conservative side of expansive systems theory (nevertheless, both surveys will have an unavoidable overlap).

*Date*: December 5th, 2012.

*Key words and phrases.* Expansive flow, geodesic flow, conjugate points, topological stability, shadowing property, Gromov hyperbolicity, accessibility, Finsler manifolds.

The author was partially supported by CNPq, Pronex de Geometria (Brazil), FAPERJ (Cientistas do nosso estado).

We can assert without doubts that J. Lewowicz contributions concern all relevant subjects in topological dynamics: topological stability, specification properties, the theory of invariant sets, classification of expansive and topologically stable homeomorphisms in surfaces up to semi-conjugacy and conjugacy, persistence sets and robustness. The originality of his contribution resides in the innovative tools and ideas to tackle certain problems in dynamics, providing elegant and simple proofs of deep results and revealing surprises in expansive, non-hyperbolic theory which led to many applications in other fields of dynamics. The use of Lyapunov forms and functions of two variables to classify hyperbolicity and to study stability of expansive systems is one of his main original ideas. The existence of invariant sets for expansive homeomorphisms in compact surfaces is certainly the most important result of his work. This result not only led to the classification of expansive homeomorphisms in surfaces - they are conjugate to pseudo-Anosov maps - but also to many different applications even in higher dimensions involving Riemannian global geometry, variational calculus and stability theory. The notion of persistence introduced in J. Lewowicz paper [70] admits a generalization for geodesic flows in any dimension which led to striking connections with geometric group theory and hyperbolic geometry in the large. In a paper about geodesic flows of surfaces with non-positive curvature Lewowicz drew his attention to a family of surfaces that proved to have many remarkable geometric and ergodic properties connected with subtle problems in non-positive curvature geometry.

The survey is divided in many sections, almost all of them devoted to explore in the context of geodesic flows one of the subjects of J. Lewowicz's research. A quite complete view of the theory of expansive and weakly stable geodesic flows is given with many references and open problems. Theorem 2.4 in Section 2 is new, it provides an extension of one of the results in [69] obtained for surfaces. Many of the results are due to the author, who as one of J. Lewowicz undergraduate students in the Universidad Simón Bolivar in Caracas, Venezuela, has been since then influenced by Lewowicz's ideas and points of view about dynamics. To finish the Introduction, I would like to give special thanks to the Instituto de Matemática y Estadística "Prof. Ingeniero Rafael Laguardia", Universidad de la República del Uruguay, for the hospitality received during the congress in the honor of J. Lewowicz, and for the invitation to write this humble tribute to his work.

## 1. Preliminaries

Let us start with some notations. $(M, g)$ will denote a $C^\infty$, compact n-dimensional Riemannian manifold; $\tilde{M}$ the universal covering of $M$, $p : \tilde{M} \longrightarrow M$ the covering map; $(\tilde{M}, \tilde{g})$ the pullback of $g$ by the map $p$, $TM$ the tangent bundle of $M$, $T_1M$ the unit tangent bundle of $(M, g)$; the canonical coordinates of a point $\theta \in TM$ are $\theta = (p, v)$, where $p \in M$, $v \in T_pM$; and $\pi : TM \longrightarrow M$ is the canonical projection $\pi(p, v) = p$.

**Definition 1.1.** The geodesic flow of $(M, g)$ is the one parameter family of diffeomorphisms $\phi_t : T_1M \longrightarrow T_1M$ defined by $\phi_t(p, v) = (\gamma_{(p,v)}(t), \gamma'_{(p,v)}(t))$, where $\gamma_\theta(t)$ is the geodesic having initial conditions $\gamma'_\theta(0) = v$, $\gamma_\theta(0) = p$. The parameter $t$ is the arc length parameter.

The unit tangent bundle $T_1M$ inherits a Riemannian metric $\bar{g}$ that is naturally associated to the Riemannian metric $g$, the so-called **Sasaki metric**. We refer to [41] for its definition, we shall state some of its main properties for the sake of completeness.

The first fundamental property of the Sasaki metric is that the canonical projection $\pi : (T_1M, \bar{g}) \longrightarrow (M, g)$ is a Riemannian submersion. For each $\theta \in T_1M$ there is a $n$-dimensional subspace $\mathcal{H}_\theta \subset T_\theta T_1M$ called the **horizontal subspace** where $d_\theta\pi$ is an isometry. The horizontal subspace can be identified with the subspace of parallel vector fields with respect to the metric $g$.

The kernel of $d_\theta\pi$ is called the **vertical subspace** $V_\theta \subset T_\theta T_1M$, it is an $n-1$ dimensional subspace and $\mathcal{H}_\theta$, $V_\theta$ are orthogonal with respect to the Sasaki metric. The unit vector tangent at $\theta$ to the geodesic flow will be denoted by $X(\theta)$, this is a horizontal vector field that is orthogonal to the vertical subspace with respect to the Sasaki metric. Let $N_\theta$ the orthogonal complement of $X(\theta)$, and let $H_\theta = \mathcal{H}_\theta \cap N_\theta$.

Let $T_\theta T_1M = H_\theta \oplus V_\theta \oplus X(\theta)$ be the horizontal-vertical splitting of $T_\theta T_1M$ (orthogonal in the Sasaki metric). The subspace $N_\theta = H_\theta \oplus V_\theta$ is invariant by the differential of the geodesic flow, whose action in $N_\theta$ is given by

$$D_\theta\phi_t(W) = D_\theta\phi_t(W_H, W_V) = (J_W(t), J'_W(t))$$

where $J_W(t)$ is a perpendicular Jacobi field of the geodesic $\gamma_\theta$ defined by the initial conditions

$$J_W(0) = W_H,\ J'_W(0) = W_V,\ g(J_W(0), \gamma'_\theta(0)) = 0.$$

Perpendicular Jacobi fields give matrix solutions of the differential equation $J''(t) + K_\theta(t)J(t) = 0$, where $K_\theta(t)$ is the matrix of sectional curvatures of planes containing $\gamma'_\theta(t)$. Indeed, if the choose an orthonormal, parallel frame $e_i(t)$ $i = 0, 1, .., n-1$ along a unit speed geodesic $\gamma_\theta(t)$, with $e_0(t) = \gamma'_\theta(t)$, then any collection $J_k(t)$ of $n-1$ linearly independent, perpendicular Jacobi fields of $\gamma_\theta(t)$ defines a curve of matrices $\mathcal{J}(t)$ with entries

$$\mathcal{J}_{ki}(t) = g(J_k(t), e_i(t))$$

that is a matrix solution of the above matrix Jacobi equation where

$$K_\theta(t)_{ki} = g(\mathcal{R}(\gamma'_\theta(t), e_k(t))\gamma'_\theta(t), e_i(t)).$$

In the above formula, the curvature tensor is $\mathcal{R}$. The curve of matrices $U(t) = \mathcal{J}'(t)\mathcal{J}(t)^{-1}$ defines a solution of the so-called Riccati equation $U'(t) + U^2(t) + K(t) = 0$, whenever $\mathcal{J}(t)$ is invertible. The Liouville 1-form, denoted by $\alpha$, is given at each point $\theta \in T_1M$ by $\alpha_\theta(Z) = \bar{g}(Z, X(\theta))$ for every $Z \in T_\theta T_1M$. The form $\alpha$ and its exterior differential $d\alpha$ are invariant by the geodesic flow. The two form $d\alpha_\theta$ is symplectic in $N_\theta$ and the exterior product $\alpha \wedge (d\alpha)^{n-1}$ provides a volume form for $T_1M$ that is invariant by the geodesic flow as well. A subspace $S$ of $N_\theta$ is called **Lagrangian** if $d\alpha(Z, Y) = 0$ for every pair of vectors $Z, Y$ in $S$ ($S$ is isotropic in notation of classical mechanics) and $S$ has maximal dimension with this property, $n-1$. It is not hard to show that

$$d\alpha(D\phi_t(Z), D\phi_t(Y)) = g(J_Z(t), J'_Y(t)) - g(J'_Z(t), J_Y(t))$$

for every $Z, Y \in N_\theta$ and every $\theta \in T_1M$. This expression is the well known Wronskian of the pair $J_Z, J_Y$ of solutions of the Jacobi equation.

The conservative nature of the geodesic flow can be viewed (and perhaps better understood) from the point of view of Lagrangian or Hamiltonian systems. Indeed, the geodesic flow is the Euler-Lagrange flow of the Lagrangian $L : TM \longrightarrow \mathbb{R}$, $L(p, v) = \frac{1}{2}g_p(v, v)$.

1.1. **Expansiveness and topological stability for flows.** The notion of expansiveness for flows poses some technical difficulties, concerning reparametrizations of the flow, which do not arise in discrete dynamics. Here we state the definition of expansive flow without singularities, for a general definition we refer to [20].

**Definition 1.2.** A non-singular smooth flow $\phi_t : \Sigma \longrightarrow \Sigma$ acting on a complete Riemannian manifold $\Sigma$ is $\epsilon$-**expansive** if given $x \in \Sigma$ we have that for each $y \in \Sigma$ such that there exists a continuous surjective function $\rho : \mathbb{R} \longrightarrow \mathbb{R}$ with $\rho(0) = 0$ satisfying

$$d(\phi_t(x), \phi_{\rho(t)}(y)) \leq \epsilon,$$

for every $t \in \mathbb{R}$ then there exists $t(y)$, $\mid t(y) \mid < \epsilon$ such that $\phi_{t(y)}(x) = y$. A smooth non-singular flow is called expansive if it is expansive for some $\epsilon > 0$.

The considerations about the reparametrization $\rho(t)$ are in many senses natural. If we just let $\rho(t) = t$ like in the discrete case where $t \in \mathbb{Z}$, many simple examples of "non-expansive" dynamical systems might turn into expansive ones. Take for instance a linear flow in the flat two torus where orbits are all periodic. It is possible to change the parametrization of the flow while keeping the same orbits as curves. So we can produce a "drift " in the dynamics, in a way that two close points move in different speeds along two periodic orbits which remain close as curves but not as orbits.

The same sort of considerations about reparametrizations appear in the definition of topological stability for flows.

**Definition 1.3.** A non-singular smooth flow $\phi_t : \Sigma \longrightarrow \Sigma$ acting on a complete Riemannian manifold $\Sigma$ is $C^k$ **topologically stable** if there exists an open neighborhood $U$ of $\phi_t$ in the $C^k$ topology such that for each flow $\psi_t$ in the neighborhood there exists a continuous surjective map $h : \Sigma \longrightarrow \Sigma$ such that for every $x \in \Sigma$ there exists $r : \mathbb{R} \longrightarrow \mathbb{R}$ continuous and surjective, $r(0) = 0$, with

$$h(\psi_t(x)) = \phi_{r(t)}(h(x)).$$

When $h$ is a homeomorphism (or conjugacy) the flow is $C^k$ structurally stable.

The reparametrization $\rho(t)$ cannot be the identity in many important families of continuous systems. In the set of Anosov geodesic flows in compact surfaces, a time preserving conjugacy $h$ isotopic to the identity between the geodesic flows of $(M, g)$ and $(M, \sigma)$ implies rigidity: $(M, g)$ is isometric to $(M, \sigma)$. This statement is known in the literature as marked length spectrum rigidity, it was first proved by Otal [94] in negative curvature, and by Croke, [35], Croke-Fathi [36] for surfaces without conjugate points.

1.2. **Lyapunov forms and functions for flows.**

**Definition 1.4.** Given a $C^\infty$ manifold $\Sigma$ and a smooth, non-singular vector field $Y$ in the tangent space $T\Sigma$ of $\Sigma$, a $C^k$ Lyapunov quadratic form $Q : T\Sigma \times T\Sigma \longrightarrow \mathbb{R}$ for the flow of $Y$ is given by the following properties:

(1) $Q$ is $C^k$.
(2) The Lie derivative $\mathcal{L}_Y Q$ of the form $Q$ is positive.

Lyapunov quadratic forms are powerful tools to find invariant cones of the dynamics of the differential of the flow of $Y$. The relevance of invariant cones is their close relationship with nonzero Lyapunov exponents, and hence with hyperbolicity and positive entropy (see for instance the works of Wojtkowsky [130], Markarian [87], [88] for billiards, Katok [61], Chernov-Markarian [29] for a complete exposition about the subject).

**Definition 1.5.** Given a smooth manifold $\Sigma$ and a neighborhood $U$ of the diagonal of $\Sigma \times \Sigma$, A $C^k$ Lyapunov function of two variables $f : U \longrightarrow \mathbb{R}$ for the flow of $Y$ is a $C^k$ non-negative function such that

(1) $f(x, x) = 0$ for every $x \in \Sigma$,
(2) The derivative of $f(x, y)$ with respect to the flow is positive for every $(x, y) \in \Sigma$.

Lewowicz introduces a special family of Lyapunov functions called **non-degenerate**: they can be obtained by applying an integration procedure to Lyapunov forms. Such Lyapunov functions imply topological stability. However, Lyapunov functions of two variables might exist independently of Lyapunov forms. In [71], Lewowicz shows that every expansive homeomorphism has a Lyapunov function of two variables that is not necessarily non-degenerate. The construction of such Lyapunov function can be viewed as a clever generalization of the ideas of Conley and Auslander in the 1960's to construct Lyapunov functions of one variable to localize recurrence in dynamical systems. The existence of a Lyapunov quadratic form is much stronger than the existence of a Lyapunov function of two variables.

The idea of invariant cones is behind one of the most remarkable results proved by J. Lewowicz: the characterization of Anosov dynamics in terms of Lyapunov quadratic forms. We shall state a version of this result for flows, that is in fact a continuous version of a discrete result by himself: Lyapunov functions and topological stability, J. of Diff. Equations (38), 1980.

**Theorem 1.6.** *: Let $\Sigma$ be $C^\infty$ compact manifold, then a smooth flow $Y_t$ acting on $\Sigma$ is Anosov if and only if there exists a non-degenerate Lyapunov quadratic form for the flow $Y_t$.*

Invariant cones were one of many different (equivalent) mechanisms developed in the 19070's and early 1980's to find hyperbolic behavior for the differential of the dynamics. Indeed, the notions of dominated splitting and quasi-Anosov system already present in Eberlein's work for Anosov geodesic flows [41], and formally introduced by R. Mañé [83], [84], are counterparts of the notion of invariant cones.

The application of Lyapunov quadratic forms to study local and global stability of hyperbolic systems proved to be very rich and enlightening. In the paper by J. Lewowicz : Invariant manifolds for regular points. Pacific J. of Math. 1981, a simple and very elegant proof of the stable manifold theorem is made using Lyapunov quadratic forms. In another article by J. Lewowicz, E. Lima de Sá, and J. Tolosa : Lyapunov functions of two variables and a conjugacy theorem for dynamical systems. Acta Científica Venezolana, 1981, a surprisingly simple proof of the Hartmann-Grobman Theorem arises as an application of the theory of Lyapunov quadratic forms. Another nice application of this theory is found in the paper by J. Lewowicz, J. Tolosa.: Local conjugacy of quasihyperbolic systems. Diff. equations, Qualitative theory I, II, 1984.

All the above results hold for general flows, but specific results for geodesic flows, the main object of our survey, were also obtained by J. Lewowicz in the case of surfaces. In those we can see the Riemannian structure of the flow in the construction of special Lyapunov quadratic forms. This will be the subject of the next section.

## 2. Lyapunov quadratic forms and functions for geodesic flows

The goal of this section is to discuss the proof and generalizations of the following result.

**Theorem 2.1.** *(J. Lewowicz): Let $(M, g)$ be a compact surface with non-positive curvature.*

1. *The geodesic flow is Anosov if and only if there exists an invariant sub-bundle $W$ of $T(T_1M)$, and a non-degenerate quadratic form $Q : W \times W \longrightarrow \mathbb{R}$ such that the Lie derivative is positive definite.*
2. *If the geodesic flow is not Anosov but the interior of the set of points with zero curvature is empty, then the flow has a Lyapunov function of two variables and is topologically stable in the set of non-positive curvature geodesic flows.*
3. *If the curvature fails to be negative just along a simple closed geodesic $\gamma$ where $\mid K(x) \mid$ decays to zero as $K(x) \approx d(x, \gamma)^2$ then the geodesic flow is $C^4$ topologically stable.*

This result is published in the article "Lyapunov functions and stability of geodesic flows" (1983) [69].

2.1. **A closer look at the paper: Lyapunov quadratic forms for Anosov geodesic flows.** The obtention of the Lyapunov quadratic form for geodesic flows of surfaces of negative curvature starts with Cartan's structural equations. Let us recall briefly Cartan's formulation.

Let $\alpha$ be the Liouville 1-form dual to the geodesic flow, $\omega$ the connection 1-form (dual to the vertical bundle in $T_1M$), $\omega^\perp$ the 1-form such that $\alpha \wedge \omega^\perp \wedge \omega$ is the volume form preserved by the geodesic flow. Such forms are called the Cartan forms, and they satisfy the following system of equations:

$$d\alpha = \omega \wedge \omega^\perp$$
$$d\omega^\perp = -\omega \wedge \alpha$$
$$d\omega = -(K \circ \pi)\alpha \wedge \omega^\perp,$$

where $K$ is the Gaussian curvature of the surface.

Let us define the quadratic form $Q$ given at each $\theta \in T_1M$ by the form $Q_\theta : N_\theta \times N_\theta \longrightarrow \mathbb{R}$,

$$Q_\theta(v, v) = \omega^\perp(v) \times \omega(v).$$

From Cartan's structural equations it is not difficult to deduce the first part of Theorem 2.1:

1. If the curvature is negative the geodesic flow is Anosov and the Lie derivative of $Q$ with respect to the geodesic flow, $\mathcal{L}_X Q = -(K \circ \pi)(\omega^\perp)^2 + (\omega)^2$, is positive.
2. If the curvature is nonpositive, the geodesic flow is Anosov if and only if there exists $T > 0$ such that the Lie derivative with respect to the geodesic flow of the form
$$\bar{Q}(v, v) = \int_0^T \phi_t^* Q(v, v) dt$$
is positive restricted to the unstable sub-bundle.

In item (1), the proof of the fact that negative curvature implies Anosov geodesic flow follows from standard Sturm-Liouville theory applied to the Jacobi equation and the representation of the differential of the geodesic flow in terms of Jacobi fields as stated in the Preliminaries (page 3). Indeed, negative curvature implies that the norms of perpendicular Jacobi fields are comparable to exponential functions. At the same

time, the Lie derivative of $Q$ with respect to the geodesic flow has positive sign whenever the curvature is non-positive.

To show item (2) we use the formula

$$(*)\ \mathcal{L}_X \phi_t^* Q(Z,Z) = -K(\gamma(t)) \parallel J_Z(t)) \parallel^2 + \parallel J_Z'(t) \parallel^2$$

where $\phi_t(p,v) = (\gamma(t), \gamma'(t))$, $Z \in N_{(p,v)}$, which follows from the definition of the forms $\omega$, $\omega^\perp$ and the representation of the differential of the geodesic flow in terms of Jacobi fields along the orbit $\phi_t(p,v)$. Since the curvature may vanish along the orbit of $(p,v)$ the proof of item (2) can be reduced to the following statement: $\parallel J_Z'(t) \parallel^2$ cannot be zero in arbitrarily large intervals for $(p,v)$ varying in $T_1M$. For otherwise we would get, after a limiting process, a geodesic and a perpendicular Jacobi field whose derivative is zero everywhere, a parallel vector field. But then, it is not difficult to deduce that the geodesic flow is not Anosov.

The goal of this subsection is to show a generalization to any dimension of the above statement. To do that, we shall explain in detail how formula (*) is deduced in the $n$-dimensional case. Let $(M,g)$ be a compact Riemannian manifold, let $\Omega$ and $\Omega^\perp$ be the 1-forms given respectively at each point by the orthogonal projection $\Omega_\theta : N_\theta \longrightarrow V_\theta$ in the vertical subspace with respect to the Sasaki metric, and the orthogonal projection $\Omega_\theta^\perp : N_\theta \longrightarrow \mathcal{H}_\theta$ in the horizontal subspace with respect to the Sasaki metric. Notice that in the case of surfaces, we have that $\omega = \Omega$, $\omega^\perp = \Omega^\perp$.

The tangent space $T_\theta TM$ for $\theta = (p,v)$ is isomorphic to $T_pM \times T_pM$ and the vertical subspace $V_\theta$ can be naturally identified with the second component $T_pM$ of the cartesian product. Indeed, any vector $v \in V_\theta$, being a vector tangent to a curve of vectors in $T_pM$, can be viewed itself as a tangent vector in $T_pM$. Let $i : V_\theta \longrightarrow T_pM$ be this isomorphism. We have the following statement that can be found in [41] for instance.

**Lemma 2.2.** *There exists an operator $\nabla : TM \longrightarrow TM$, called the connection operator, given for each $T_pM$ by a linear operator $\nabla_p : T_pM \longrightarrow T_pM$ such that*

1. *$\nabla_p(i(Z)) = \Omega(Z)$ for every $Z \in V_\theta$.*
2. *Let $\gamma_\theta(t)$ be a unit speed geodesic such that $\gamma_\theta(0) = p$, $\gamma_\theta'(0) = v$. Let $J(t)$ be a perpendicular Jacobi field defined along $\gamma_\theta$. Then*

$$\nabla_p(J(0)) = J'(0)$$

*while $\Omega_\theta(J(0), J'(0)) = (0, J'(0))$. Here, $(J(0), J'(0))$ are horizontal-vertical coordinates in $N_\theta$.*

The above formal discussion about the representation of $\Omega$ as an operator in $T_pM$ leads to the following definition: let $Q_\theta : N_\theta \times N_\theta \longrightarrow \mathbb{R}$ be the two form given by

$$Q_\theta(Z,Z) = g_{\pi(\theta)}(d_\theta\pi(Z), \nabla_p(i(Z))).$$

The two form $Q$ is a natural generalization of the two form given in Theorem 2.1 for surfaces. Notice that

**Lemma 2.3.** *Let $Z \in N_\theta$, let $J_Z$ be the Jacobi field defined along $\gamma_\theta$ whose initial condidions are $J_Z(0) = d\pi(Z) = \Omega_\theta^\perp(Z)$, $J_Z'(0) = \Omega_\theta(Z)$. Let $\phi_t$ be the geodesic flow of $(M,g)$. Then*

1. *$\phi_t^* Q(Z,Z) = \frac{1}{2}\frac{d}{dt}(\parallel J_Z(t) \parallel^2)$*
2. *The Lie derivative of $\phi_t^* Q(Z,Z)$ with respect to the geodesic flow at $\gamma_\theta(t)$ is*

$$\mathcal{L}_X \phi_t^* Q(Z,Z) = -g(\mathcal{R}(\gamma_\theta'(t), J_Z(t))\gamma_\theta'(t), J_Z(t)) + \parallel J_Z'(t) \parallel^2$$

*where $\mathcal{R}$ is the curvature tensor, $\| v \|$ is the norm of the metric $g$, and derivatives are taken with respect to the covariant derivative of $g$. In particular, if the sectional curvatures of $g$ are all negative, the Lie derivative of $Q$ is positive.*

*Proof.* The proof is almost tautological: by the definition of $Q$ we have that

$$\phi_t^* Q(Z,Z) = g(J_Z(t), J_Z'(t)) = \frac{1}{2}\frac{d}{dt}(\| J_Z(t) \|^2)$$

which is item (1). Item (2) follows from item (1) taking derivatives with respect to the covariant derivative of $g$ and replacing the second derivative of $J_Z(t)$ by its expression in the Jacobi equation. □

Lemma 2.3 extends the statement of Theorem 2.1 concerning surfaces of negative curvature to compact Riemannian manifolds of negative curvature. The main result of this subsection is the full generalization of Anosov geodesic flows in terms of the form $Q$ regardless of the sectional curvature signs of $(M, g)$.

**Theorem 2.4.** *Let $(M, g)$ be a compact Riemannian manifold. Then the geodesic flow is Anosov if and only if there exists $T_0 > 0$ and a Lagrangian, invariant subbundle $E^u$, where $E^u_\theta \subset N_\theta$ for each $\theta \in T_1M$, such that*

1. *The quadratic form $\bar{Q}(Z,Z) = \int_0^{T_0}(\int_0^t \phi_s^* Q(Z,Z)ds)dt$ is positive for every $Z \in E^u_\theta$ and every $\theta \in T_1M$.*
2. *The Lie derivative of the restriction of $\bar{Q}$ to $E^u_\theta$ with respect to the geodesic flow is positive for every $\theta \in T_1M$.*

Theorem 2.4 can be regarded as a (natural) version of Theorem 1.6 for geodesic flows, we think that symplectic diffeomorphisms and Hamiltonian flows should admit versions with analogous assumptions. Notice that the quadratic form $\bar{Q}$ has the same formula of the well known index form used in Morse theory to study minimizing properties of geodesics. While in Morse theory the index form is evaluated in Jacobi fields vanishing in at least two points, the form $\bar{Q}$ is relevant in Lyapunov form theory when evaluated at unstable Jacobi fields, which never vanish. The Anosov property implies that the manifold has no conjugate points. So these Jacobi fields generate the subbundle $E^u_\theta$, which is nothing but the unstable Green subspace (see Lemma 2.8 below). We start with some elementary calculations. Let $Q_t(Z,Z) = \int_0^t \phi_s^* Q(Z,Z)ds$.

**Lemma 2.5.** *We have the following identity:*

$$Q_t(Z,Z) = \frac{1}{2}(\| J_Z(t) \|^2 - \| J_Z(0) \|^2).$$

*Proof.* Just apply Lemma 2.3 item (1) to the definition of $Q_t$. □

**Definition 2.6.** Let $(M, g)$ be a complete Riemannian manifold. A geodesic $\gamma$ is said to have no conjugate points if every Jacobi field of $\gamma$ which vanishes at two different points must vanish everywhere. The manifold $(M, g)$ has no conjugate points if no geodesic has conjugate points.

The manifold $(M, g)$ has no conjugate points if and only if the exponential map is non-singular at every point. The following characterization of geodesics without conjugate points due to Green [50] will be useful in the section.

**Proposition 2.7.** *Let $(M, g)$ be a compact Riemannian manifold whose sectional curvatures are bounded below by a constant $-K_0$, where $K_0 > 0$. Then, a geodesic $\gamma(t)$ has*

*no conjugate points if and only if there exists a solution of the matrix Riccati equation $U(t)$ defined along $\gamma$ for every $t \in \mathbb{R}$. Moreover, any solution $U(t)$ of the Riccati equation defined for every $t \in \mathbb{R}$ satisfies*

$$\| U(t) \| \leq \sqrt{K_0}$$

*for every $t \in \mathbb{R}$. Here, $\| U \|$ is the usual norm of symmetric matrices.*

We continue with some results concerning the so-called Green bundles. We gather in the following statement some results proved by Green [50] and Eberlein [41].

**Lemma 2.8.** *Let $(M, g)$ be a compact manifold without conjugate points whose sectional curvatures are bounded below by a constant $-K_0$, where $K_0 > 0$. Then for every geodesic $\gamma_\theta$ we have:*

(1) *For every $w \in T_{\gamma_\theta(0)}M$ that is in the plane $\gamma'_\theta(0)^\perp$ of vectors perpendicular to $\gamma'_\theta(0)$, the limit*

$$\lim_{T \to +\infty} \mathbb{J}_T(t)(w) = J^s_w(t)$$

*exists for every $t \in \mathbb{R}$, and it is a perpendicular Jacobi field with $J^s_w(0) = V$. The collection of asymptotic Jacobi fields obtained in this way generates an invariant Lagrangian subspace $E^s_\theta \subset N_\theta$ given by*

$$E^s_\theta = \{(J^s_w(0), J^{s\prime}_w(0)),\ w \in \gamma'_\theta(0)^\perp\}.$$

(2) *Analogously, the limit $\lim_{T \to -\infty} J_T(t)(w) = J^u_w(t)$ exists, and it is a perpendicular Jacobi field with $J^u_w(0) = V$. The collection of asymptotic Jacobi fields obtained in this way generate an invariant Lagrangian subspace $E^u_\theta \subset N_\theta$ given by*

$$E^u_\theta = \{(J^u_w(0), J^{u\prime}_w(0)),\ w \in \gamma'_\theta(0)^\perp\}.$$

(3) *The Jacobi fields $J^s_w(t)$, $J^u_w(t)$ never vanish if $w \neq 0$, and we have*

$$\| J^{s\prime}_w(t) \| \leq \sqrt{K_0} \| J^s_w(t) \|$$
$$\| J^{u\prime}_w(t) \| \leq \sqrt{K_0} \| J^u_w(t) \|$$

*for every $t \in \mathbb{R}$.*

(4) *If a perpendicular Jacobi field $J(t)$ defined in the geodesic $\gamma_\theta$ satisfies*

$$\| J(t) \| \leq C$$

*for every $t \geq 0$, then $(J(0), J'(0)) \in E^s_\theta$. Analogously, if*

$$\| J(t) \| \leq C$$

*for every $t \leq 0$, then $(J(0), J'(0)) \in E^u_\theta$*

The Jacobi fields $J^s_w$ are called stable Jacobi fields, and the subspace $E^s_\theta$ is called the **stable Green subspace**. The Jacobi fields $J^u_w$ are called unstable Jacobi fields and $E^u_\theta$ is called the **unstable Green subspace**. Item (3) is a straightforward application of Proposition 2.7. Applying Rauch comparison theorem it is easy to check that in the case of negative curvature there exists $a > 0$ such that the norms of such Jacobi fields satisfy

$$\| J^s_w(t) \| = \| w \| e^{-at},$$
$$\| J^u_w(t) \| = \| w \| e^{at},$$

for every $t \in \mathbb{R}$. In the case of Anosov geodesic flows we have,

**Theorem 2.9.** *(Klingenberg* [63]*) Let $(M, g)$ be a compact Riemannian manifold whose geodesic flow is Anosov. Then $(M, g)$ has no conjugate points.*

Klingenberg's Theorem was improved by the following beautiful result due to R. Mañé [85].

**Theorem 2.10.** *Let $(M, g)$ be a compact Riemannian manifold such that the geodesic flow preserves a continuous subbundle $E : T_1M \longrightarrow TT_1M$ where each subspace $E(\theta)$ is a Lagrangian subspace of $N_\theta$ with respect to the symplectic form $d\alpha$. Then $(M, g)$ has no conjugate points.*

Combining Klingenberg's theorem with Lemma 2.8 and the divergence of Jacobi fields which vanish at one point (Green [49]) we get

**Proposition 2.11.** *Let $(M, g)$ be a compact Riemannian manifold whose geodesic flow is Anosov. Then $E^s_\theta$ is the dynamical stable subspace of the dynamics, and $E^u_\theta$ is the dynamical unstable subspace of the dynamics. So*

1. *$T_\theta T_1M = E^s_\theta \oplus E^u_\theta \oplus X(\theta)$ for every $\theta \in T_1M$.*
2. *There exist $C > 0$, $0 < a$ such that*

$$\| D_\theta\phi_t(Z) \|_S \leq Ce^{-at} \| Z \|_S$$

*for every $t > 0$, $Z \in E^s_\theta$, and*

$$\| D_\theta\phi_t(Z) \|_S \leq Ce^{at} \| Z \|_S$$

*for every $t < 0$, $Z \in E^u_\theta$, where $\| Z \|_S$ is the Sasaki norm.*

Now, we are ready to show Theorem 2.4.

**Proposition 2.12.** *Let $(M, g)$ be a compact Riemannian manifold. If the geodesic flow is Anosov then there exists $T_0 > 0$ such that for each $\theta \in T_1M$ we have,*

1. *The manifold has no conjugate points and the quadratic form $\bar{Q}(Z, Z) = \int_0^{T_0}(\int_0^t \phi_s^* Q(Z, Z)ds)dt$ is positive for every $Z \in E^u_\theta$ and every $\theta \in T_1M$, where $E^u_\theta$ is the unstable Green subspace.*
2. *The Lie derivative of the restriction of $\bar{Q}$ to $E^u_\theta$ with respect to the geodesic flow is positive for every $\theta \in T_1M$.*

*Proof.* This statement is the easy part of Theorem 2.4. Let us denote by $\| Y \|_S$ the Sasaki norm and let $\| J \|$ be the $g$-norm. By Proposition 2.11 and Lemma 2.8, we have that for every $Z \in E^u_\theta$,

$$\begin{aligned} \frac{1}{C}e^{bt} \| Z \|_S &\leq \| D_\theta\phi_t(Z) \|_S \\ &= (\| J_Z(t) \|^2 + \| J'_Z(t) \|^2)^{\frac{1}{2}} \\ &\leq \sqrt{1 + K_1^2} \| J_Z(t) \| . \end{aligned}$$

Applying again Lemma 2.8 we get,

$$\frac{1}{C}e^{bt} \| J_Z(0) \| \leq \sqrt{1 + K_0^2} \| J_Z(t) \| .$$

So the norm of unstable Jacobi fields growths exponentially with time. Since by Lemma 2.5 we have

$$Q_t(Z, Z) = \frac{1}{2}(\| J_Z(t) \|^2 - \| J_Z(0) \|^2),$$

clearly there exists $T_0 > 0$ such that for every $t > T_0$ the above both quadratic form is positive and growths exponentially fast with $t$ for every $Z \in E^u_\theta$, $\theta \in T_1M$. Since

$$\mathcal{L}_X\bar{Q}(Z,Z) = \int_0^T \phi_t^* Q(Z,Z)dt = Q_T(Z,Z)$$

we immediately conclude that the Lie derivative of $\bar{Q}$ restricted to $E^u_\theta$ is positive if $T > T_0 0$ and every $\theta \in T_1M$. □

Now, let us show the converse of Proposition 2.12. The next result due to Eberlein [41] will plays an important role in the proof.

**Theorem 2.13.** *Let $(M,g)$ be a compact Riemannian manifold without conjugate points. The following statements are equivalent:*

1. *The geodesic flow is Anosov.*
2. *There is no nontrivial perpendicular Jacobi field $J(t)$ such that $\| J(t) \| \leq C$ for every $t \in \mathbb{R}$.*
3. *Green subspaces are linearly independent.*

**Proposition 2.14.** *Let $(M,g)$ be a compact Riemannian manifold. Suppose that for every $\theta \in T_1M$ there exist an invariant, Lagrangian subspace $E_\theta \subset N_\theta$, $T > 0$ such that the Lie derivative of the restriction of $\bar{Q}(Z,Z) = \int_0^T (\int_0^t \phi_s^* Q(Z,Z)ds)dt$ to $E_\theta$ with respect to the geodesic flow is positive for every $\theta \in T_1M$. Then,*

1. *The manifold has no conjugate points.*
2. *There exists $\bar{T} > 0$ such that quadratic form $Q_T(Z,Z) = \int_0^T (\phi_s^* Q(Z,Z)ds$ is positive for every $T > \bar{T}$, $Z \in E_\theta$ and $\theta \in T_1M$.*
3. *The subbundle of subspaces $E_\theta$ is continuous.*
4. *$E_\theta = E^u_\theta$ for every $\theta \in T_1M$.*
5. *The geodesic flow is Anosov.*

*Proof.* Since the Lie derivative of $\bar{Q}$ is

$$\mathcal{L}_X\bar{Q}(Z,Z) = \int_0^T \phi_t^* Q(Z,Z)dt = Q_T(Z,Z) = \frac{1}{2}(\| J_Z(T) \|^2 - \| J_Z(0) \|^2),$$

by the assumption in the Proposition we have that there exists $\lambda(Z,\theta) > 1$ such that

$$\| J_Z(T) \| > \lambda(Z,\theta) \| J_Z(0) \|$$

for every $Z \in E_\theta$. Notice that this implies that $J_Z(t) \neq 0$ for every $Z \in E_\theta$, $Z \neq 0$, and $t \in \mathbb{R}$. Since the subspace $E_\theta$ is Lagrangian the dimension of the subspace of the Jacobi fields $J_Z$, $Z \in E_\theta$ is $n-1$. So a basis of these Jacobi fields provides a matrix solution $U(t)$ of the Riccati equation along $\gamma_\theta(t)$ that is defined for every $t \in \mathbb{R}$. By Proposition 2.7 the geodesic $\gamma_\theta$ has no conjugate points, and since this happens for every $\theta \in T_1M$ we conclude that the manifold has no conjugate points. This shows item (1).

Let us consider the unit ball $S(E_\theta)$ of vectors in $E_\theta$. The invariance of $E_\theta$ provides

$$\| J_Z(nT) \| > \lambda(\theta)^n,$$

$$\| J_Z(-nT) \| < \lambda(\theta)^{-n},$$

for every $Z \in S(E_\theta)$, $n \in \mathbb{N}$ and $\theta \in T_1M$.

**Claim:** By compactness of $(M,g)$, there exists $C > 0$ such that

$$(*) \quad \| J_Z(-t) \| \leq C,$$

for every $Z \in S(E_\theta)$, $t > 0$ and every $\theta \in T_1M$.

Indeed, by the choice of $Z$ we have that $\| J_Z(-nT) \| \leq 1$ for every $n > 0$. So it is enough to show that there exists $C > 1$ such that every non-vanishing Jacobi field $J_Z$ with $\max\{\| J_Z(0) \|, \| J_Z(T) \|\} \leq 1$ satisfies $\| J_Z(t) \| \leq C$ for every $t \in [0, T]$. To see this, observe first that applying Proposition 2.7 to the Riccati solutions associated to the Jacobi fields $J_Z$, $Z \in E_\theta$, we get

$$\| J_Z'(0) \|, \| J_Z'(T) \| \leq \sqrt{K_0}$$

where $K_0$ is the maximum absolute value of the sectional curvatures of $(M, g)$. So given $L > 0$, let $\mathcal{F}$ be the family of Jacobi fields of $(M, g)$ such that $\| J(0) \| \leq 1$, $\| J'(0) \| \leq L$. This is a co-compact family of Jacobi fields and by Rauch's comparison theorem, the norms of Jacobi fields with the same initial conditions in a Riemannian manifold with negative curvature $-K_0$ bound from above the norms of the Jacobi fields in $\mathcal{F}$. This yields that for every $J \in \mathcal{F}$ we have

$$\| J(t) \| \leq (1 + L)e^{T\sqrt{K_0}}$$

for every $t \in [0, T]$, from which we easily conclude the Claim.

**Claim:** The subspaces $E_\theta$ depend continuously on $\theta \in T_1M$.

The symplectic structure of the geodesic flow plays a key role in the proof. Observe that $(*)$ is a closed property in $T_1M$, as well as the Lagrangian character of the subspaces $E_\theta$. So if we consider a sequence $\theta_n$ of points converging to some $\theta_0$, we can choose a convergent sequence of subspaces $E_{\theta_{n_k}}$ converging to some Lagrangian subspace $E_0 \subset N_{\theta_0}$ where $\| J_Z(t) \| \leq C$ for every $t \leq 0$ and for every $Z \in E_0$. If $E_0$ is not equal to $E_{\theta_0}$, there exist two non-vanishing vectors $Z_0 \in E_0$, $Z_1 \in E_{\theta_0}$ such that $d\alpha(Z_0, Z_1) \neq 0$. Simply because $E_0$ is Lagrangian and hence it is a maximal isotropic subspace. On the other hand, the form $d\alpha$ is invariant by the action of the flow, so we have

$$\begin{aligned} 0 \neq d\alpha(Z_0, Z_1) &= d\alpha(D\phi_{-t}(Z_0), D\phi_{-t}(Z_1)) \\ &= g(J_{Z_0}(-t), J'_{Z_1}(-t)) - g(J'_{Z_0}(-t), J_{Z_1}(-t)) \end{aligned}$$

for every $t > 0$. But this implies that the expression at the right tends to zero as $t \to \infty$ since $\lim_{t\to-\infty} \| J_{Z_1}(t) \| = 0$ and therefore by Lemma 2.8 this yields $\lim_{t\to-\infty} \| J'_{Z_1}(t) \| = 0$ as well. This contradiction implies that $E_0$ coincides with $E_{\theta_0}$, and clearly implies the continuity of the subspaces $E_\theta$ and item (3).

By Lemma 2.8 we have that $Z \in E^u_\theta$ for every $Z \in E_\theta$ and every $\theta$. Since $E_\theta$ and $E^u_\theta$ are both Lagrangian, they must coincide thus showing item (4). Moreover, there is no nontrivial Jacobi field $J(t)$ such that $\| J(t) \| \leq C$ for every $t \in \mathbb{R}$. Because by Lemma 2.8 $J(t)$ should be unstable whilst the above inequalities imply that its norm is not bounded. So Theorem 2.13 implies that the geodesic flow of $(M, g)$ is Anosov showing item (5). Proposition 2.12 shows item (2). □

One interesting consequence of Theorem 2.4 combined with Lemma 2.3 is that the geodesic flow is Anosov if and only if the index form in an interval $[0, T]$ restricted to unstable Jacobi fields is positive.

2.2. **Lyapunov functions for nonpositive curvature.** The second part of J. Lewowicz's paper [69] contains an interesting construction of a Lyapunov function of two variables for compact surfaces with non-positive curvature such that int $(K^{-1}(0)) = \emptyset$. This Lyapunov function is obtained by a sort of integration process applied to the Lyapunov quadratic form in the statement of Theorem 2.1. The construction is quite technical and relies strongly in two dimensional arguments. With this tool in hand it is showed that such geodesic flows are $C^1$ topologically stable. Observe that the geodesic flow of a compact nonpositively curved surface where $\text{int}(K^{-1}(0)) = \emptyset$ is expansive.

The third part of [69] improves the Lyapunov function obtained in the second part for a certain family of interesting surfaces of nonpositive curvature: those where the curvature fails to be negative just along a simple closed geodesic and the way curvature decays to zero in a neighborhood of this geodesic has a non-degenerate analytic behavior. Namely, the second derivatives of the curvature with respect to Fermi coordinates orthogonal to the geodesic do not vanish along the geodesic. The non-degeneracy of the second jet of the curvature affects the second jet of the Lyapunov function obtained before, giving the function genericity enough to show topological stability of the geodesic flow in the $C^4$ topology.

In brief, the second and third parts of the paper [69] attempt to study topological stability of expansive, non-Anosov geodesic flows in surfaces with nonpositive curvature using Lyapunov functions of two variables. The construction of such a function showed to be so technical and particular of nonpositive curvature, bi-dimensional geometry, that a more topological approach seemed to be more promising than an analytical one. At the time, J. Lewowicz was working simultaneously on a topological approach to tackle stability problems of expansive, non-Anosov systems. This approach did not use Lyapunov functions and proved to be far more efficient and enlightening in stability theory of expansive systems, not only in surfaces but in higher dimensions. We shall discuss this branch of J. Lewowicz work and its applications to geodesic flows in the forthcoming sections.

## 3. Stability without Lyapunov functions: topological dynamics in low dimension and beyond

The goal of the section is to present the main results of J. Lewowicz about the topological dynamics of expansive homeomorphisms and its applications in the theory of geodesic flows of compact manifolds without restrictions in the dimension. This part of Lewowicz's work differs fundamentally from his first studies of topological stability because there is no use of Lyapunov forms and functions. The idea now is to get a version of the stable manifold theorem for expansive systems, then try to show the existence of a local product structure and from this a series of persistence properties of orbits would follow just as in the case of hyperbolic dynamics (pseudo-orbit tracing property, topological stability, persistence of some recurrent sets). The task does not seem easy: expansive systems might not be hyperbolic so the differential of an expansive diffeomorphism does not give any hint about the behavior of the dynamics, like in the smooth stable manifold theorem of hyperbolic dynamics and the tools involved in its proof (the well known lambda lemma for instance). A similar program to study stability of expansive systems was suggested by Bowen and Walters in the 1970's [20], [129], who made indeed some preliminary steps.

3.1. **Expansive homeomorphisms and hyperbolic topological dynamics.** Let us start with a result proved by J. Lewowicz: Expansive homeomorphisms of surfaces, Bol. Soc. Bras. Math. 1989, [71]. Obtained independently by Hiraide: Expansive homeomorphisms of compact surfaces are pseudo-Anosov, Osaka Math. Journal. 1990, it is in my opinion, the most important work of J. Lewowicz.

**Theorem 3.1.** *Expansive homeomorphisms of compact surfaces have local invariant sets with "product structure" in all but a finite number of periodic orbits, and are conjugate to pseudo-Anosov maps. In particular, there are no expansive homeomorphisms in the two sphere.*

Local invariant sets are connected arcs which provide two invariant families of curves of the dynamics. In one of them the dynamics approaches orbits with time, this would be a counterpart of the stable foliation of hyperbolic dynamics. In the other family the dynamics expands orbits with time, a counterpart of the unstable foliation. Of course, nothing about smoothness of these curves can be deduced from the expansiveness assumption. The local product structure refers to the following fact: in a neighborhood of every point but a finite number of periodic orbits, each stable set meets an unstable set at just one point. In the set of "exceptional" periodic orbits stable and unstable sets are ramified, like prone singularities of pseudo-Anosov maps. In the survey by R. Potrie [101], there is a very nice, complete and geometric exposition of the proof of Theorem 3.1, together with many references of subsequent applications of these ideas for homeomorphisms in two and three dimensional manifolds. Let us discuss some applications of Theorem 3.1 in the theory of geodesic and Hamiltonian flows.

3.2. **Persistently expansive geodesic flows.**

**Definition 3.2.** Given a $C^\infty$ manifold $N$ and a family of $C^\infty$ flows $\mathcal{F}$ defined in $N$, we say that a expansive flow $\psi_t : N \longrightarrow N$ is $C^k$ persistently expansive in $\mathcal{F}$ if there exists a $C^k$ neighborhood $W$ of $\psi_t$ in $\mathcal{F}$ such that every flow in $W$ is expansive.

Persistence of topological properties of orbits of systems was one of the main objects of study in the 1970's and 1908's in the context of the structural stability theory of Axiom A systems. Here we shall focus on the persistence of expansiveness, later on we shall comment about other types of persistent properties. The main result of the subsection is the following:

**Theorem 3.3.** *: (R. Ruggiero) Let $(M, g)$ be a compact Riemannian manifold. If the geodesic flow is $C^1$ persistently expansive in the family of $C^\infty$ Hamiltonian flows then the closure $\Omega$ of the set of periodic orbits is a hyperbolic set. If $M$ is a surface, the geodesic flow is Anosov.*

This result is published [108] in a paper entitled: Persistently expansive geodesic flows, Comm. Math. Physics, 1991. The problem was motivated by a result due to Mañé [81] where it is shown that persistently expansive diffeomorphisms are **quasi-Anosov**: the orbit of every nontrivial vector by the action of the differential of the diffeomorphism is not bounded. The work of Eberlein [41] in 1973 showed that the geodesic flow of a compact manifold without conjugate points is Anosov if and only if it is quasi-Anosov. So if we knew that persistently expansive geodesic flows have no conjugate points we could combine this property with a generalization of Mañé's result to flows and deduce that such flows are Anosov. This line of reasoning leads to the following interesting question: Do expansive geodesic flows have no conjugate points?

This problem was solved by M. Paternain [97] in 1994 for surfaces, and remains open in higher dimensions. We shall come back to the subject in the next section.

We shall give a sketch of proof of Theorem 3.3 to show where J. Lewowicz work about expansive homeomorphisms of surfaces plays a crucial role. The proof has several steps.

**Proposition 3.4.** *The $C^1$ persistent expansiveness in the family of geodesic flows implies that periodic orbits are $C^1$ persistently hyperbolic.*

*Proof.* The proof is by contradiction: if there is a nearby geodesic flow $\bar{\phi}_t$ with a non-hyperbolic periodic orbit $O$ then there exists eigenvalues of the Poincaré map of $O$ with modulus one. Then, by a theorem due to Klingenberg and Takens [65], we can perturb the flow in the family of geodesic flows to get another Hamiltonian flow $\hat{\phi}_t$ with a periodic $\hat{O}$ orbit having complex, generic eigenvalues with modulus one. Here generic refers to the set of generic properties of symplectic maps, which imply the assumptions on the Poincaré map of the orbit required by the Birkhoff-Lewis fixed point Theorem [63]. This theorem implies the existence of infinitely many periodic orbits in a tubular neighborhood of $\hat{O}$, contradicting the expansiveness of $\hat{\phi}_t$. A similar argument was used by Newhouse [93] to show that structurally stable symplectic diffeomorphisms are Anosov. □

We would like to point out that the genericity result by Klingenberg and Takens [65] has a recent, much simpler proof using ideas of control theory by L. Rifford and the author [105]. Actually, what is proved in [105] is that generic properties of symplectic maps are **Mañé generic** for Poincaré maps of closed orbits of Hamiltonians. A property $P$ of the Hamiltonian flow of $H : T^*M \longrightarrow \mathbb{R}$ is called Mañé $C^k$-generic if there exists a $C^k$-generic family of functions $U : M \longrightarrow \mathbb{R}$ such that the Hamiltonian flow of the Hamiltonians $H_U(p,q) = H(p,q)+U(p)$ have the property $P$. If the Hamiltonian is given by a Riemannian metric $g$ as in our case, Mañé's genericity is equivalent to genericity in the conformal class of $g$ by the Maupertuis' principle [5]. So Proposition 3.4 admits the following improvement:

**Proposition 3.5.** *The $C^1$ persistent expansiveness of the geodesic flow of $(M,g)$ in the family of geodesic flows of metrics in the conformal class of $(M,g)$ implies that periodic orbits are $C^1$ persistently hyperbolic.*

The following step is a version for Hamiltonian flows of a well known result due to R. Mañé [83]. Let $\mathcal{P}^1(M)$ be the collection of Hamiltonian flows in $M$ all of whose periodic orbits are hyperbolic endowed with the $C^1$ topology.

**Lemma 3.6.** *If $\phi_t$ is in the interior of $\mathcal{P}^1(M)$ then there exists a continuous, Lagrangian, invariant* **dominated splitting** *in the closure of the periodic orbits $\Omega$. Namely, there exist $C > 0$, $T > 0$, $\lambda \in (0,1)$, invariant, Lagrangian subspaces $E_\theta, U_\theta$ for every $\theta \in \Omega$, such that $E_\theta \oplus U_\theta \oplus X(\theta) = T_\theta T_1 M$, and*

$$\| D_\theta \phi_{nT} \mid_{E_\theta} \| \| D_{\phi_{nT}(\theta)} \phi_{-nT} \mid_{U_\theta} \| \leq C\lambda^n.$$

Let us remark that Lemma 3.6 has been proved by Contreras-Paternain [33] for surfaces and by Contreras [30] replacing $\mathcal{P}^1(M)$ by the set of geodesic flows of Riemannian metrics whose periodic orbits are all hyperbolic. In a work in progress with L. Rifford, we show Lemma 3.6 replacing $\mathcal{P}^1(M)$ by the set of geodesic flows of Riemannian metrics which are conformal to $(M,g)$ whose periodic orbits are hyperbolic.

A hyperbolic invariant set for a nonsingular flow of a compact manifold has a dominated splitting: the stable and unstable subspaces play the role of $E_\theta$ and $U_\theta$. The point is that this assertion has a converse in symplectic dynamics according to [108] Theorem 2.1:

**Theorem 3.7.** *Let $(M, g)$ be a compact Riemannian manifold. A Lagrangian, invariant, dominated splitting defined in a compact invariant set is hyperbolic.*

The third step of the proof of Theorem 3.3 is where J. Lewowicz's work about expansive dynamics comes into play. Indeed, if dim $(M) = 2$, dim $(T_1M) = 3$, and the product structure of an expansive homeomorphism extends to expansive flows without singularities. This fact has been used also by M. Paternain in his PhD thesis at IMPA, [97] 1990, and by Inaba-Matsumoto, in a paper published in the Japan J. Math. 1990 [57]. Moreover, the exceptional set of the flow where the local product structure might not exist is empty under our assumptions, since every orbit in this set is periodic and we know that periodic orbits are hyperbolic. So the stable manifold theorem holds for every periodic orbit. Thus, persistently expansive geodesic flows in compact surfaces have local product structure everywhere and by Poincaré 's recurrence lemma we get the density of periodic orbits. Therefore, the hyperbolic set $\Omega$ is the whole $T_1M$ and this proves Theorem 3.3.

Some final remarks. Theorem 3.3 actually holds for $C^1$-persistently expansive geodesic flows in the family of Riemannian geodesic flows, and it might be improved by assuming $C^1$ persistent expansiveness in the set of conformal perturbations of the metric (or equivalently, Mañé's perturbations).

## 4. Does expansiveness imply no conjugate points?

In this section we present a survey of results about the relationship between expansivity and absence of conjugate points. The starting point of our discussion is Klingenberg's Theorem 2.9 which states that Anosov geodesic flows have no conjugate points, establishing a link between expansive dynamics and the absence of conjugate points. Mañé's Theorem 2.10 implies Klingenberg's result since the stable and unstable subspaces of an Anosov geodesic flow form continuous, Lagrangian invariant bundles.

Of course, the existence of an invariant Lagrangian splitting does not imply expansiveness of the geodesic flow, as in the Anosov case. Nevertheless, Mañé's new, simpler approach to the proof of Klingenberg's Theorem combined with Lewowicz's work about expansive dynamics led to the first result linking expansiveness and conjugate points in the context of topological dynamics. The result was proved by M. Paternain, in Ergodic Theory and Dyn. sys. (1994), in a paper entitled "Expansive geodesic flows on surfaces" [97].

**Theorem 4.1.** *If the geodesic flow of a compact Riemannian surface is expansive then the surface has no conjugate points. In particular, there are no expansive geodesic flows in the two sphere.*

The proof starts with the three-dimensional extension of Lewowicz results about local product structure for expansive homeomorphisms. Geodesic flows of surfaces act without singularities on the unit tangent bundle that is three dimensional, local stable and unstable sets exist for every point in $T_1M$ and there is a local product structure everywhere but at a finite number of periodic orbits. The notion of index introduced by M. Paternain in [97] which mimics the Maslov index, allows to show that the exceptional

set is empty and that the index of every orbit is zero. This finally implies, as in Maslov index theory, that geodesics have no conjugate points.

There is a Hamiltonian version of the theorem by G. Paternain and M. Paternain published at Comptes Rendu de l' Academie de Sciences, Paris (1993) [96], always for surfaces. The n-dimensional version of Theorem 4.1 is still an open problem. Nevertheless, there are some partial positive results that will be presented next.

4.1. **Expansive geodesic flows in manifolds without conjugate points.** Based on the previous results linking expansiveness and absence of conjugate points, we focused on the study of geodesic flows of compact manifolds without conjugate points. The following results by the author, published in Ergodic Theory and Dynamical systems, 1996-1997, [111], [113], show that Lewowicz's theory for low dimensional expansive systems extends to geodesic flows in n-dimensional manifolds without conjugate points.

**Theorem 4.2.** *Let $(M, g)$ be a compact Riemannian manifold without conjugate points. If the geodesic flow is expansive. Then,*

1. *The flow has a local product structure, the pseudo-orbit tracing property, it is topologically stable, transitive and periodic orbits are dense and unique in each nontrivial homotopy class.*
2. *There exists a $C^0$ neighborhood of the flow in the family of geodesic flows such that any expansive geodesic flow in the neighborhood with the same expansivity constant has no conjugate points.*
3. *If the geodesic flow of $(M, g)$ is $C^1$-persistently expansive in the family of geodesic flows then it is Anosov.*

Item (1) shows that the topological dynamics of expansive geodesic flows in the absence of conjugate points is just like hyperbolic topological dynamics.

Item (2) can be viewed as a partial answer to the problem of absence of conjugate points in the presence of expansiveness. The idea of the proof is not difficult: expansive flows of metrics $(M, g)$, $(M, h)$ which are sufficiently close to each other with the same expansiveness constant are conjugate. So if $(M, g)$ has no conjugate points item one implies that closed orbits are dense. By Cartan's Theorem [28], in each homotopy class there is one closed orbit minimizing the $h$-length of closed curves in the class. Now, expansiveness allows to show that a periodic minimizer $\gamma$ of $(M, h)$ in its homotopy class must be a globally minimizing geodesic. Namely, each lift $\tilde{\gamma}$ of the geodesic in $(\tilde{M}, \tilde{h})$ has the property that the distance $d_{\tilde{h}}(\tilde{\gamma}(t), \tilde{\gamma}(s))$ is the $h$-length of $\tilde{\gamma}(s, t)$ for every $s \leq t$. In particular, the set of periodic minimizers of the length has no conjugate points. Since the flows of $(M, g)$ and $(M, h)$ are conjugate and periodic $g$-geodesics in each homotopy class are unique, the same happens with periodic $h$-geodesics. Therefore, $h$-geodesics without conjugate points are dense in the unit tangent bundle, and since geodesics without conjugate points form a closed set, there are no geodesics with conjugate points in $(M, h)$.

Item (2) implies as well that expansive geodesic flows in the boundary of geodesic flows without conjugate points must be accumulated by non-expansive geodesic flows. So such flows are not only boundary flows in the family of flows without conjugate points but they are also boundary flows for the family of expansive flows.

The proof of item (3) combines item (1) and the results in the previous section. Indeed, by Theorem 3.3, the closure of periodic orbits of a $C^1$-persistently expansive geodesic flow is a hyperbolic set. By item (1), the absence of conjugate points implies

the density of periodic orbits. So the whole unit tangent bundle is the closure of the periodic orbits and hence the flow is Anosov.

Theorem 4.2 gives a quite complete description of the topological dynamics of expansive geodesic flows. The ergodic theory of expansive geodesic flows is more complicated and less developed than topological dynamics theory. The most famous problem related with this field is the ergodicity of rank one geodesic flows in manifolds with nonpositive curvature, whose answer is not known even in surfaces. Expansiveness proved to be a rich source of links between dynamics and global geometry in the theory of manifolds without conjugate points. Theorem 4.2 is just one example, in the forthcoming sections we shall show many others.

## 5. Expansive geodesic flows and Gromov Hyperbolic geometry

Expansive dynamics of geodesic flows without conjugate points has a strong impact in the global geometry of the universal covering and in particular, in the algebraic structure of the fundamental group. We shall give the highlights of a theory developed by the author in many papers from 1994 to 2004 involving expansive and topologically stable dynamics. We start in this section with the link between expansivity and Gromov hyperbolic spaces. We say that a complete metric space $(X, d)$ is **geodesic** if for every pair of points $p, q \in X$ there exists a continuous curve $\gamma : [0, a] \longrightarrow X$ such that $\gamma(0) = p$, $\gamma(a) = q$, and $\gamma$ is an isometry of the interval $[0, a]$ endowed with the Euclidean length.

**Definition 5.1.** Given a complete geodesic space $(X, d)$, a geodesic triangle $\nabla$ with vertices $x_0, x_1, x_2$, is the union of three geodesics $I_0, I_1, I_2$ such that:

(1) The endpoints of $I_0$ are $x_0$ and $x_1$,
(2) The endpoints of $I_1$ are $x_1$ and $x_2$,
(3) The endpoints of $I_2$ are $x_2$ and $x_0$.

**Definition 5.2.** Let $(X, d)$ be a complete, geodesic metric space. $(X, d)$ is a *Gromov hyperbolic space* if there exists $\delta > 0$ such that every geodesic triangle $\nabla$ with sides $I_0, I_1, I_2$ satisfies the following property: the distance from any $p \in I_j$ to $I_k \cup I_s$, where $j \neq k$, $j \neq s$, $k \neq s$ is bounded above by $\delta$ (the indices are taken mod. 3).

A geodesic triangle with the property given in Definition 5.2 is called $\delta$-*t*hin. Geodesic triangles in manifolds of negative curvature bounded above by a negative constant are $\delta$-thin for some $\delta$ depending on the curvature bound. A tree is a Gromov hyperbolic space where triangles are 0-thin. Gromov hyperbolicity captures hyperbolic geometry in the large. The famous work of Gromov [51] in the 1980's started a fruitful field of research involving global analysis and geometry, group representations, combinatorics, graph theory, ergodic theory, complexity of algorithms and many other research areas. The rich structure of Gromov hyperbolic spaces is comparable with the structure of manifolds of negative curvature. Expansive dynamics is also related to Gromov hyperbolicity:

**Theorem 5.3.** *The fundamental group of a compact Riemannian manifold without conjugate points and expansive geodesic flow is Gromov hyperbolic.*

The above result was proved by the author and published in the Bulletin of the Brazilian Math. Soc. (1994) [110]. It shows that expansiveness not only provides for the geodesic flow all the features of the topological dynamics of Anosov flows, but also provides coarse hyperbolic geometry for the universal covering and the fundamental group. For instance, we have that the volume of balls increases exponentially with the

radius, the topological entropy of the flow is positive (already observed by M. Paternain in the case of compact surfaces in [97]), the universal covering has a compactification with a cone topology analogous to negative curvature manifolds, this compactification is homeomorphic to a $n$-ball if the manifold has dimension $n$, the action of the fundamental group extends to the boundary of this compactification (that is homeomorphic to a $(n-1)$-sphere), and many deep results of the theory of Kleinian and Fuchsian groups extend to the boundary action. A good survey of results of the theory of Gromov hyperbolic groups is [14]. In 3-dimensional manifolds, we get a topological classification of manifolds admitting expansive geodesic flows [120].

**Theorem 5.4.** *Let $(M, g)$ be a compact 3-manifold admitting a Riemannian metric without conjugate points and expansive geodesic flow. Then $(M, g)$ admits a hyperbolic geometric structure.*

The idea of the proof is based in the solution of the Poincaré conjecture by Perelmann [98]. Indeed, a compact Riemannian 3-manifold without conjugate points is a "prime" manifold (see the book of Hempel [52]). The word "prime" refers to certain decomposition of 3-manifolds in "minimal" pieces in a very precise sense. Milnor [91] shows that every smooth compact manifold can be decomposed "uniquely" in a connected sum of manifolds with the simplest possible topology. A manifold is called prime if each piece of a connected sum decomposition of the manifold is either diffeomorphic to the manifold itself or to a 3-sphere. The Poincaré conjecture according to the work of W. Thurston [127] is equivalent to the geometrization conjecture for prime manifolds: a prime manifold can be cut along tori such that each piece admits a geometric structure modeled in one of the eight geometries $\mathbb{R}^3$, $S^3$, $S^2 \times \mathbb{R}$, $\mathbb{H}^3$, $\mathbb{H}^2 \times \mathbb{R}$, the Heisenberg group, the so-called Solv-group, and $SL(2, \mathbb{R})$. To admit a geometric structure means to have a Riemannian covering that is an homogeneous, simply connected space. Since the universal covering of manifolds without conjugate points are diffeomorphic to $\mathbb{R}^n$, the manifold is prime. Moreover, Theorem 5.3 implies that there are no incompressible tori in the manifold. So the geometrization "conjecture" can be applied to the manifold, leaving us with $\mathbb{R}^3$, $\mathbb{H}^3$, $\mathbb{H}^2 \times \mathbb{R}$, the Heisenberg group, the Solv-group, and $SL(2, \mathbb{R})$. But the only one of them that is Gromov hyperbolic is $\mathbb{H}^3$. Since Gromov hyperbolicity does not depend on the Riemannian metric for compact manifolds, we get that the only possibility of geometric structure for our manifold is the hyperbolic 3-space.

A final remark of dynamical interest. The argument in E. Ghys work [47] to show that Anosov geodesic flows of compact surfaces are conjugated (not parameter preserving) to geodesic flows in constant negative curvature extends to expansive geodesic flows in compact manifolds without conjugate points which admit constant negative curvature structures. So Theorem 5.4 implies that expansive geodesic flows in compact 3-manifolds without conjugate points are conjugate to geodesic flows of constant negative curvature.

## 6. Surfaces with non-positive curvature and finite area ideal triangles

We come back to J. Lewowicz's paper [69] and Theorem 2.1. In item (3) of this theorem a family of nonpositive curvature surfaces is given with some properties prescribing the decay to zero of the absolute value of the curvature in a tubular neighborhood of a closed geodesic. This article is the first one to consider such surfaces, whose geodesic flows proved to enjoy very interesting properties revealed by many authors [34], [112], [46], [76]. The geodesic flows of these surfaces where about the first non-Anosov, ergodic examples of geodesic flows of rank one manifolds. The purpose of this section is to discuss some more subtle results in the literature.

6.1. **Prescribed decay of negative curvature and "fake" Anosov flows.** Let us start with a result inspired by the work of J. Barges and E. Ghys [23] published in 1988: if ideal geodesic triangles in the universal covering of a compact surface of negative curvature have constant area, the curvature is constant as well.

**Theorem 6.1.** *There exist expansive, non-Anosov, $C^2$ geodesic flows in compact surfaces with non-positive curvature with the following property: there exists a constant $C > 0$ such that every geodesic ideal triangle in the universal covering has area bounded above by $C$. The curvature of such surfaces is negative but along a simple closed geodesic $\gamma$ where $\mid K(x) \mid \approx d(x,\gamma)^\alpha$, for some $\alpha \in (1,2)$.*

The above theorem was proved by G. Contreras and the author and published in 1997 [34]. It says that if we replace the constant area assumption on ideal geodesic triangles proposed by Barges-Ghys by bounded area the geodesic flow might no longer be Anosov. And the examples are very close to the ones considered by J. Lewowicz in [69], the difference being the decay to zero of the absolute value of the curvature: it is of the type $d(x,\gamma)^\alpha$ where $\alpha$ is not integer as in item (3) of Theorem 2.1. We might have expected that the bounded area condition for ideal triangles would be sufficient to characterize Anosov flows, Theorem 6.1 gave us a surprising negative result in this sense. However, the following result due to the author and published in 1997 [112] gives a complete answer to the problem in the family of non-positive curvature surfaces.

**Theorem 6.2.** *If the geodesic flow of a compact surface $(M,g)$ is $C^3$ (namely, $(M,g)$ is a $C^4$ Riemannian manifold) and every geodesic ideal triangle has finite area (in the above sense) then the geodesic flow is Anosov.*

So the natural guess of the characterization of Anosov flows by the existence of a uniform bound for the area of ideal triangles holds if the flow is smooth enough. Moreover, in [34] it is proved that if the decay to zero of the absolute value of the curvature in the surfaces considered in Theorem 6.1 is of the order of $\mid K(x) \mid \approx d(x,\gamma)^\alpha$ where $\alpha \geq 2$ then the flow is of class $C^3$. Thus, the ideal triangles in the universal covering of the surfaces considered by Lewowicz in [69] do not have a uniform upper bound for the area.

6.2. **Hölder continuity of invariant foliations of expansive, non-Anosov geodesic flows.** The surfaces considered by Lewowicz and described in Theorem 2.1 item (3) provide examples of non-Anosov geodesic flows with invariant foliations (namely, center stable and center unstable foliations) which are transversal everywhere in $T_1M$ but along a vanishing curvature closed orbit. Tangencies of invariant foliations of partially hyperbolic systems pose serious technical problems in ergodic theory, the complexity of Pesin's theory [99] shows how difficult is to deal with smooth ergodic theory in the presence of zero Lyapunov exponents which appear naturally when such tangencies occur. The center stable foliation is obtained by saturation of the stable horocycle flow. The center unstable foliation is obtained by saturation of the unstable horocycle flow. However, the following result proved by Gerber and Nitica published in 1999 [46] shows that the regularity of invariant foliations of our expansive, non-Anosov surfaces is not that bad:

**Theorem 6.3.** *Let $(M,g)$ be a $C^2$ compact surface with non-positive curvature such that the curvature is negative but along a closed geodesic $\gamma$ where the decay to zero of its absolute value is of the order of $\mid K(x) \mid \approx d(x,\gamma)^\alpha$ where $\alpha > 1$. Then there exists*

*$\beta \in (0,1)$ such that the invariant foliations of the geodesic flow are $\beta$-Hölder continuous at the points of $\gamma$.*

The above result tells us that the regularity of invariant foliations of the geodesic flows of the considered surfaces may not be as good as Anosov regularity for geodesic flows of surfaces, that is $C^1$ by the work of Hopf [54], but is as good as Anosov regularity for higher dimensional manifolds according to the work of Anosov [4]. The Hölder regularity of the invariant foliations combined with the well known Hopf's argument to show the ergodicity of Anosov geodesic flows in compact surfaces, yield the ergodicity of the geodesic flow for the surfaces described in Theorem 6.3 (with a little help of Pesin's theory). In the forthcoming subsection we shall comment about other surprising resemblances of these expansive, non-Anosov geodesic flows with true Anosov flows.

6.3. **More about surfaces with "fake" Anosov flows: subactions and large deviations.** The subject now is ergodic theory for expansive, non-hyperbolic geodesic flows from a variational viewpoint. The variational study of the entropy of invariant measures goes back to the late 1960's and the 1970's, when the works of Bowen, Ruelle and Sinai started the application of the nowadays called Ruelle-Perron-Frobenius operator in the context of the thermodinamic formalism to find invariant measures which maximize the metric entropy [20]. This beautiful theory led naturally to the problem of finding invariant probabilities which maximize the action of Hölder continuous observables. Given a metric space $(X,d)$, a Hölder continuous observable with exponent $\alpha > 0$ is a continuous function $f : X \longrightarrow \mathbb{R}$ such that $d(f(x), f(y)) \leq d(x,y)^\alpha$. In smooth hyperbolic dynamics, the logarithm of the Jacobian of the differential restricted to the unstable bundle is the observable most commonly considered. In expansive dynamics, the above function is not Hölder continuous in general, it is typically quite singular due to the presence of zero Lyapunov exponents. However, the study of other types of observables has many applications in physics [100], [78], [79], [80], [19] and [32]. Based on the work of A. Lopes and P. Thieullien [79] for Anosov flows, A. Lopes, V. Rosas Meneses and the author get the following result for our expansive, non-Anosov geodesic flows that is published in Discrete and Continuous Dynamical systems (2004) [76]):

**Theorem 6.4.** *Let $(M,g)$ be a compact, $C^3$ surface with non-positive curvature such that the area of ideal triangles is finite. Then the Livsic's Theorem holds in its classical (continuous, Hölder) version and there exist continuous subaction functions associated to Hölder continuous observables.*

Livsic's Theorem is one of the most classical cohomological features of measure preserving Anosov dynamics. Briefly speaking, the theorem asserts that a continuous function that is cohomologous to zero along periodic orbits of a measure preserving Anosov flow (or diffeomorphism) acting on a compact manifold is cohomologous to zero in the whole manifold. The theorem has many interesting applications in spectral theory and rigidity (see for instance [66], [38]), [61], [100]).

A subaction function $F$ is defined as a solution of an inequality involving the same terms occurring in a cohomology equation: given a smooth flow $\psi_t : N \longrightarrow N$ and a Hölder continuous function $f : N \longrightarrow \mathbb{R}$, a continuous function $F : N \longrightarrow \mathbb{R}$ is a subaction function associated to $f$ if

$$F(\psi_t(p)) \geq F(p) + \int_0^t (f(\psi_s(p)) - m(f))ds,$$

where $m(f)$ is the supremum over all $\psi_t$-invariant probability measures of the action $\int f d\mu$.

Subaction functions "localize" the support of invariant measures which maximize the action of $f$ in the set of invariant probabilities: the set of zeroes contains the support of the measure.

This observation has a strong flavor of Aubry-Mather theory, as observed in many papers in the literature (see for instance [100], [78], [79], [19] and [32]). Subaction functions are in many respects counterparts of the so-called subsolutions of the Hamilton-Jacobi equation of Tonelli Lagrangians. Both functions attain equality in an invariant set that contains the support of an invariant measure that is critical, in the case of the subsolutions of the Hamilton-Jacobi equation we are talking about Mather measures. Subsolutions attain minimum values at the so-called Aubry set, which might contain strictly the Mather set. The set of vanishing points of a subaction can be compared with the Aubry set, and the support of a maximizing measure to the Mather set. It was conjectured by Mañé [86] that generically in the set of perturbations of Tonelli Lagrangians by potentials the Aubry set coincides with the Mather set. This conjecture has recent, positive partial answers [89], [15] which encourage to consider the same problem in the context of subactions (see [32] for a partial answer for shifts).

So the variational theory of measures from the Bowen-Ruelle point of view has many interesting links with Aubry-Mather theory, they might be considered dual of each other in many senses. Another remarkable link between both theories arises in the context of stochastic differential equations. N. Anantharamam in [2] proved that the family of stationary probabilities of twisted Brownian motions associated to the twisted Hamiltonians

$$H_\lambda(p,v) = \frac{1}{2} g_p(v,v) - \lambda \omega_p(v)$$

where $\omega$ is a closed one form, converges as $\lambda$ goes to $\infty$ to the projected Mather measure with cohomology class coinciding with the class of $\omega$, provided that this measure is unique in its class. The convergence is described by a very precise formula of large deviations depending on the so-called Peierl's barrier. The proof is a very interesting combination of what is called weak KAM theory, introduced by A. Fathi and A. Siconolfi [45] - an analytic approach to Aubry-Mather theory through fixed point theory in functional spaces - and stochastic partial differential equations. The parameter $\lambda$ can be compared to the inverse of the temperature in the one parameter, thermodynamic formalism: when we multiply the topological pressure by the temperature in Bowen-Ruelle formula, and let the temperature to go to zero, we get a family of equilibrium measures depending on the temperature that in many important systems converges to a ground state or equilibrium state. So the Mather measure can be viewed as a counterpart of a ground state in thermodynamics from this point of view.

When the support of the Mather measure is hyperbolic, an estimate of the Peierl's barrier can be obtained in terms of the Lyapunov exponents, providing an exponential large deviation principle for the above family of measures [3]. However, if the support is not hyperbolic, to obtain an estimate is much more subtle. In a paper published in 2011 [77], A. Lopes and the author gave an estimate for the large deviation in the case of nonpositive curvature surfaces having zero curvature just along a simple closed geodesic .. Once more, the examples considered by J. Lewowicz in his pioneering paper [69] give us a hint of how the lost of hyperbolicity might affect ergodic properties of a non-Anosov, expansive geodesic flow.

**Theorem 6.5.** *Let $(M, g)$ be a compact surface of non-positive curvature where the curvature fails to be negative just along a simple closed geodesic $\gamma$ where $\mid K(x) \mid \simeq d(x,\gamma)^n$, $n \geq 2$, that is the support of the Aubry-Mather measure associated to its homology class. If the geodesic coincides with the Aubry set of the homology class then the stationary measures of the Brownian motions of the twisted Hamiltonians $H_\lambda(p,v) = \frac{1}{2}g_p(v,v) - \lambda\omega_p(v)$ converge, as $\lambda \to +\infty$ to the projected Aubry-Mather measure with a complete large deviation law of polynomial type.*

The main idea of the proof is to obtain a formula for the Peierl's barrier in terms of the Busemann functions of a lift of the geodesic $\gamma$ in the universal covering. The Busemann functions can be estimated using the precise analytic expression of the curvature in a tubular neighborhood of $\gamma$ and comparison theory taking surfaces of revolution as models. We would like to point out that an analytic description of the curvature near the vanishing curvature set is absolutely necessary to get a large deviation formula, without any specification of this sort it is impossible to get any interesting estimate. Theorem 6.5 reminds us the ergodic theory of Manneville-Pommeau maps of the interval, a family of expansive, non-hyperbolic maps with one indifferent fixed point given by $f(x) = x + x^\alpha$, $\alpha > 1$. These maps were studied in great detail by L. S. Young [131] who gets a large deviation principle of polynomial type with exponent depending on $\alpha$.

As a conclusion for the section, the surfaces considered by J. Lewowicz in [69] are good examples to study how the lost of hyperbolicity might affect a variety of subtle properties of the dynamics: from the rigidity results of the first subsection, the Hölder regularity for the invariant foliations presented in the second subsection, to ergodic optimization and Aubry-Mather theory in the third subsection. There are still many open problems concerning the surfaces considered by Lewowicz, like the extension of Theorem 6.5 to surfaces with a finite number of vanishing curvature geodesics, and the decay of correlations and large deviations of the Liouville measure for such surfaces. We hope to have motivated the interest of the reader in this rich field of research.

## 7. Control Theory and Accessibility

Control theory usually applies to a category of problems with three elements: a source set (or set of initial conditions), a target set and a family of processes depending on time which start at a point in the source set and ends in the target set. The family of processes is usually given by a family of differential equations, ordinary or partial, and the problem is said to be controllable if for each $p$ in the source set and $q$ in the target set there exists a process in the family which starts at $p$ and ends at $q$. M. Brin in [21] introduced a notion related to controllability called **Accessibility** in recent works about persistent ergodicity (see for instance [25]). The initial setting is a smooth dynamical system with one or two invariant foliations, and the system has the local accessibility property if for every point $p$ there exists an open neighborhood $V(p)$ such that every $x \in V(p)$ can be connected to $p$ with a finite number of arcs each of which is contained in one of the foliations. A well known theorem due to Hörmander [56] allows to link accessibility with totally non-integrable distributions: if the Lie brackets of a smooth $k$-dimensional distribution defined in a $C^\infty$ manifold generate the tangent space at every point of the manifold, then for each point $x$ there exists an open neighborhood of $x$ where each point can be joined to $x$ by a smooth arc tangent to the distribution. So contact structures are related to accessible systems, and in particular Anosov geodesic flows are accessible with respect to the stable and unstable foliations: these foliations are not jointly integrable because their tangent planes generate the contact plane field of the

geodesic flow. Brin's notion of accessibility by arcs can be deduced for contact Anosov flows with $C^1$ invariant foliations by means of the relation between the Lie brackets of stable vector fields $X^s$ (i.e., tangent to the stable bundle) with unstable vector fields $X^u$, and the commutators of their flows. The main result of the section is a version due to the author [121], published in 2008, of this statement for expansive geodesic flows in manifolds without conjugate points.

**Theorem 7.1.** *Expansive geodesic flows in compact manifolds without conjugate points have the accessibility property with respect to stable and unstable sets. Namely, given a point $\theta \in T_1M$ there exists an open neighborhood $V(\theta)$ such that each $z \in V(\theta)$ can be joined to $\theta$ by a continuous arc formed by a finite number of arcs, each of which is contained either in a stable set (an s-arc) or in an unstable set (u-arc). Moreover, expansive geodesic flows are accessible: every two points can be joined by a continuous arc formed by a finite number of arcs each of which is either a s-arc or a u-arc.*

Theorem 7.1 can be viewed as a sort of continuous version of the non-joint-integrability of stable and unstable foliations of Anosov geodesic flows. The invariant foliations of an expansive, non-Anosov geodesic flow might not be smooth in general, so a definition of accessibility using arcs tangent to distributions might not make sense in this setting. The main idea of the proof is a surprising application of the Gromov hyperbolic structure of the universal covering of the manifold (Theorem 5.3). A continuous path formed by s-arcs and u-arcs is associated to a system of horospheres which are tangent to each other in the universal covering. The Gromov hyperbolic structure of the universal covering allows to describe in a very precise way the structure of the set of horospheres which are simultaneously tangent to two given horospheres: the ideal centers of such horospheres in the ideal boundary of the universal covering form a continuous, codimension 1 submanifold that separates the ideal boundary into two disjoint connected components. This statement is used to show that given a point $\theta \in T_1M$, the endpoints of continuous arcs starting at $\theta$ and formed by 4 continuous s-arcs or u-arcs fill an open neighborhood of $\theta$.

The accessibility of expansive geodesic flows in manifolds without conjugate points shows somehow that the contact structure of the geodesic flow can be seen even at a topological, non-smooth level. Of course, the lack of regularity of invariant foliations of expansive systems makes unfeasible a theory of persistent ergodicity. But perhaps accessibility might be used to explore the persistence of ergodic properties of the geodesic flow restricted to the Pesin set ...

Many questions remain open. For instance, does the geodesic flow of a compact manifold without conjugate points and Gromov hyperbolic fundamental have the accessibility property? Does accessibility imply the Gromov hyperbolicity of the fundamental group?

## 8. Rigidity in Finsler surfaces

Let us change completely the subject: Finsler metrics. A $C^k$ **Finsler metric** in a smooth manifold $M$ is a function $F : TM \longrightarrow [0, +\infty)$ such that

- $F(p, tv) = tF(p, v)$ for every $t > 0$, and $(p, v) \in TM$.
- $F$ is $C^k$ in $TM - (M, 0)$.
- The Hessian of $F^2$ in the vertical variables is positive definite.

The term metric is just a convention for Finsler, in fact it is possible to define a "distance" $d_F(x, y)$ by taking the infimum of the Finsler lengths $\int_0^1 F(c(t), c'(t))dt$ of continuous rectifiable paths $c : [0, 1] \longrightarrow M$ such that $c(0) = x, c(1) = y$. Although

this function satisfies the triangle inequality, it is not symmetric with respect to $x, y$. The convexity of the Hessian of the Finsler metric allows to solve the Euler-Lagrange problem for Finsler geometry and we get, as in Riemannian geometry, Finsler geodesics and a Finsler geodesic flow $\phi_t : T_1M \longrightarrow T_1M$ acting with constant speed in the set of unit vectors of the Finsler metric.

Finsler metrics are important in classical mechanics because the Hamiltonian flow of a Tonelli Hamiltonian in a sufficiently high energy level can be parametrized in a way that it becomes the geodesic flow of some Finsler metric. The minimum level above which this holds is the so-called Mañé critical level, after [31].

The local geometry of Finsler metrics is more complicated than the local Riemannian geometry. In the case of surfaces, there is a generalization of Cartan's structural equations (see for instance [11]) where we can see three shape operators instead of just one like in Riemannian surfaces. Let $\omega^1$, $\omega^2$, $\omega^3$ the Cartan forms, $\omega^3$ is the connection form (dual to the vertical bundle, the kernel of the canonical projection), $\omega^2$ is the canonical one form (dual to the geodesic flow). Then the exterior derivatives of the forms give,

$$\begin{aligned} d\omega^1 &= -I\omega^1 \wedge \omega^3 + \omega^2 \wedge \omega^3 \\ d\omega^2 &= -\omega^1 \wedge \omega^3 \\ d\omega^3 &= K\omega^1 \wedge \omega^2 - J\omega^1 \wedge \omega^3. \end{aligned}$$

The functions $I$, $J$, $K$ are respectively, the Cartan scalar, the Landsberg scalar and the flag curvature, the generalization of the Gaussian curvature in Finsler geometry. All these shape operators depend on the horizontal and vertical variables of the tangent space since the Finsler metric is defined in the tangent space of the manifold. If the flag curvature does not depend on the vertical variables the Finsler metric is called **k-basic**. Notice that if $I$ and $J$ are identically zero, we get the Riemannian Cartan's equations. It is known that $I$ vanishes everywhere if and only if the metric is Riemannian. It is also known that $J$ is the derivative of $I$ with respect to the geodesic flow. When $J$ vanishes everywhere the Finsler metric is called Landsberg metric. A huge body of work in Finsler geometry is devoted to find under what weaker assumptions on $I, J, K$ the metric is actually Riemannian. This field of the theory is usually called **rigidity theory**.

Before introducing expansive dynamics in the exposition, we would like to introduce some rigidity results whose proofs have a certain dynamical flavor. For instance:

**Theorem 8.1.** *(Akbar-Zadeh* [1] *) Let $(M, F)$ be a $C^\infty$ compact Finsler manifold with constant sectional flag curvatures. Then the manifold is Riemannian.*

**Theorem 8.2.** *(Paternain* [95] *Let $(M, F)$ be a compact, analytic Finsler surface with genus greater than one that is either Landsberg or k-basic. Then the metric is Riemannian.*

**Theorem 8.3.** *(Barbosa-Ruggiero* [13]*) Let $(M, F)$ be a $C^4$ compact Finsler, Landsberg surface with genus greater than one and no conjugate points. Then the metric is Riemannian.*

The proof of Akbar-Zadeh's Theorem relies on the following fact: if the flag sectional curvatures are constant, the Cartan scalar satisfies the Jacobi equation. Since the negative curvature implies that non-trivial solutions of the Jacobi equation are not bounded, the Cartan scalar must vanish since it is a continuous function defined in compact manifold. Therefore, the Finsler metric is in fact Riemannian. In the case of surfaces, it

is relatively easy to show that the Cartan scalar vanishes in the presence of hyperbolic dynamics and some additional assumptions on the flag curvature. The so-called Bianchi identity is given by,

$$I''(t) + K_v(t)I'(t) + K(t)I(t) = 0,$$

where $I(t)$ is the Cartan scalar evaluated at a point $\phi_t(\theta)$ of an orbit of the geodesic flow, $K_v$ is the derivative of the flag curvature with respect to a unit vertical field, and derivatives are taken with respect to the geodesic flow. If the flag curvature is constant, $K_v = 0$ everywhere, and we get the Jacobi equation. In the case of k-basic Finsler metrics $K_v$ vanishes as well. There are well known examples of Finsler surfaces which are k-basic and non-Riemannian: Randers metrics for instance [11]. The proof of Theorem 8.2 for k-basic metrics combines the existence of a hyperbolic set for the geodesic flow and the fact that the Cartan tensor must vanish in this set by the above argument. Since the metric is analytic, the Cartan scalar is analytic as well and hence it must be zero everywhere. The proof of Theorem 8.2 for Landsberg metrics applies the same dynamical feature of the geodesic flow and the fact that the Cartan scalar is a first integral of the flow since $0 = J = I'$. Thus, $I$ must be constant in the hyperbolic set and analyticity implies that it must be constant everywhere. From this and some local Finsler geometry we deduce that the Cartan scalar vanishes everywhere. The proof of Theorem 8.3 is a sort of extension of Theorem 8.2 without assuming analyticity but imposing the absence of conjugate points. The proof once more relies on the fact that a first integral of the geodesic flow must be constant in this case, but the lack of analyticity makes the argument much harder. The main result proved in Theorem 8.3 is that Finsler compact surfaces without conjugate points have a continuous, center stable foliation that is minimal, a fact that holds already for Riemannian surfaces without conjugate points. Combining this property with the fact that the geodesic flow restricted to the center stable leaf of a hyperbolic closed geodesic has expansive behavior, we conclude that every first integral in such a leaf must be constant in the leaf. Since each leaf is dense, every continuous first integral must be constant in the unit tangent bundle.

Following this dynamical line of arguments, J. Barbosa Gomes and the author [12] proved the next result involving expansiveness.

**Theorem 8.4.** *Every k-basic Finsler metric in a compact surface with expansive geodesic flow is Riemannian.*

The combination of Theorems 8.3 and 8.4 without the assumption of expansiveness would provide a complete extension of Theorem 8.2 for compact Finsler surfaces without conjugate points.

## 9. What about weakly stable geodesic flows?

The last sections of the survey will be devoted to study geodesic flows which are in some sense close to expansive. We shall look at geodesic flows which have some persistent properties similar to those introduced by J. Lewowicz in a very nice paper published in Ergodic Theory and Dynamical systems in 1983 [70]. Let us recall the notion of persistence stated in [70] to motivate the results of the section.

**Definition 9.1.** Let $(X, g)$ be a compact Riemannian manifold and let $f : X \longrightarrow X$ be a homeomorphism. An orbit $O(x)$ of $f$ is said to be **persistent** if given $\epsilon > 0$ there exists an open $C^0$ neighborhood $U$ of $f$ such that for every $h \in U$ there exists $y \in M$ whose orbit $\epsilon$-shadows the orbit of $x$. Namely, $d(f^n(x), f^n(y)) \leq \epsilon$ for every $n \in \mathbb{N}$.

In [70], expansive homeomorphisms with persistent sets of orbits are considered. We shall consider an extension of this sort of persistence in the set of geodesic flows and show how this will reflect on the topology and the global geometry of the manifold.

### 9.1. $C^k$-C-shadowing property.

**Definition 9.2.** The geodesic flow $\phi_t : T_1M \longrightarrow T_1M$ of a complete Riemannian manifold $(M, g)$ satisfies the $C^k$-$C$-**shadowing property** if there exists a $C^k$ neighborhood of $(M, g)$ flow such that the orbits of the geodesic flow $\phi_t^h$ of any metric $(M, h)$ in the neighborhood can be $C$-shadowed by orbits of $\phi_t$. Namely, given $x \in M$, there exist $y \in M$, and a continuous surjective function $\rho_x : \mathbb{R} \longrightarrow \mathbb{R}$ with $\rho_x(0) = 0$ such that

$$d(\phi_t(y), \phi^h_{\rho_x(t)}(x)) \leq C,$$

for every $t \in \mathbb{R}$.

The analogies with Definition 9.1 are clear, although we are considering flows instead of homeomorphisms and the constant $C$ might not be small, as the term "persistent" suggests.

**Definition 9.3.** The geodesic flow $\phi_t : T_1M \longrightarrow T_1M$ satisfies the **lifted** $C^k$-$C$-shadowing property if there exists an open neighborhood of $(M, g)$ in the $C^k$ topology such that for each metric $(M, h)$ in the neighborhood, the lift in $T_1\tilde{M}$ of each orbit of the lifted $C^k$-close geodesic flow $\tilde{\phi}_t^h$ of $(\tilde{M}, \tilde{h})$ can be $C$-shadowed by an orbit of the lifted flow $\tilde{\phi}_t$ in the above sense.

The $C^k$-$C$-shadowing property implies the lifted $C^k$-$C$-shadowing property if $C$ is small enough. The lifted shadowing property is based on Morse's work about globally minimizing geodesics in the universal covering of compact surfaces with genus greater than one [92]. The work of Morse implies that the lifts of geodesics of a compact surface without conjugate points and genus greater than one are $C$-shadowed by the geodesics of metric of constant negative curvature $-1$, where $C$ depends on the metric in the surface. So the geodesic flow of such a compact surface without conjugate points satisfies the $L$-shadowing property for every $L \geq L_0$ in the family of geodesic flows without conjugate points, where $L_0$ is a constant depending on the surface.

Morse's shadowing generalizes to a compact manifold $(M, g)$ without conjugate points whose universal covering is a visibility manifold. Indeed, the fundamental group of $M$ is Gromov hyperbolic according to [110], so the universal covering $(\tilde{M}, \tilde{M})$ is a Gromov hyperbolic space itself. This means that quasi-geodesics of $(\tilde{M}, \tilde{g})$ are shadowed by geodesics [51], and since each two metrics in $M$ are equivalent the lifts of geodesics of any metric in $M$ are shadowed by the geodesics of $(\tilde{M}, \tilde{g})$.

All the above considerations lead to many conjectures linking the $C^k$-$C$-shadowing property with "hyperbolic" global geometry and topology. M. Bonk [17] showed that a complete geodesic metric space is Gromov hyperbolic if and only if for every $A, B$, there exist $C = C(A, B)$ such that every $A, B$-quasi-geodesic of the space is contained in a tubular neighborhood of radius $C$ of a geodesic. So we might guess that the $C^k$-$C$-shadowing property would imply Gromov hyperbolicity of the universal covering. However, the $C^k$-$C$-shadowing property is weaker than the shadowing of all quasi-geodesics, since it requires the $C^k$-$C$-shadowing of geodesics which arise from perturbations of the given metric.

The next subsection contains some results illustrating what happens.

9.2. **Shadowing, Gromov hyperbolicity and Preissmann's property.** The discussion in the previuos subsection was the basis of the following results by the author, which were published in a series of journals (Ergodic Theory and Dyn. sys. [115] (1999), [116] (2000), Bull. Braz. Math. Soc. [117] (1999), Discrete and Continuous Dynamical Systems [119] (2006)).

**Theorem 9.4.** *Let $(M, g)$ be a compact manifold without conjugate points, let $r(M)$ be the injectivity radius. Then we have:*

1. *If $(M, g)$ has non-positive curvature, the lifted $C^\infty$-$C$-shadowing property for the geodesic flow implies that every abelian subgroup of the fundamental group is infinite cyclic.*
2. *If $(M, g)$ has non-positive curvature and is analytic, then the lifted $C^\infty$-$C$-shadowing property implies that the fundamental group is Gromov hyperbolic.*
3. *If $(\tilde{M}, \tilde{g})$ is a quasi-convex space, and $C \leq \frac{1}{5} r(M)$, the $C^\infty$-$C$-shadowing property implies that every abelian subgroup of the fundamental group is infinite cyclic.*

A discrete group $G$ is said to have the **Preissmann property** if every abelian subgroup is infinite cyclic. The definition is clearly based on the celebrated Preissmann's theorem for the fundamental group of manifolds with negative curvature. Item (2) is so far the closest we get to Gromov hyperbolicity assuming the $C\infty$-$C$-shadowing property.

The proof of item (2) in brief is as follows: By Eberlein's work [40] the universal covering of a compact manifold with non-positive curvature is a visibility manifold if and only if there is no flat, totally geodesic plane in the universal covering. If the metric is analytic, a beautiful result due to Bangert and Schröeder [9] implies that the existence of a flat plane in $\tilde{M}$ implies that there exists a immersed flat torus in the manifold $M$. Using some ideas of Mather theory we then show that a flat metric in the torus does not have the lifted $C^\infty$-$C$-shadowing property and using the global geometry of non-positive curvature manifolds we extend this to the universal covering. We conclude that the lifted $C^\infty$-$C$-shadowing property implies that there is no flat plane in $\tilde{M}$ and therefore, $(M, g)$ is a visibility manifold. Finally, we know that visibility manifolds without focal points are Gromov hyperbolic [110].

Bangert-Schöeder theorem is only known for analytic non-positively curved manifolds. So we cannot apply this theorem to show items (1) and (3). Nevertheless, we succeed in showing that the $C^\infty$-$C$-shadowing property implies the Preissmann's property. The proof of item (1) is by contradiction: suppose that the fundamental group has an abelian subgroup that is not infinite cyclic. Then there exists a flat immersed torus in the manifold by the geometry of manifolds of non-positive curvature. So the second part of the proof of item (2) leads to a contradiction: the geodesic flow does not have the lifted $C$-shadowing property. The proof of item (3) is more subtle and we refer to the reader to [119], [120] for details.

Many interesting questions remain open: Does the $C^k$-$C$-shadowing property (lifted or not) for geodesic flows of compact manifolds without conjugate points imply Gromov hyperbolicity? Does the Preissmann property in the fundamental group of compact manifolds without conjugate points implies Gromov hyperbolicity? If we replace the shadowing property by topological stability would it be possible to extend the above results?

## 10. Are Expansive geodesic flows in the closure of Anosov dynamics?

Lewowicz's results about expansive dynamics in low dimensional manifolds show how close these systems are to hyperbolic systems. Moreover, there are many natural ways to modify a hyperbolic system in order to get an expansive, non-hyperbolic one. A sort of converse of this statement is the subject of the section: are all expansive systems in the closure of hyperbolicity? Namely, given an expansive system there exists a certain topology in the set of flows and a sequence of hyperbolic systems (structurally stable) such that they approach the expansive system in this topology? In the category of geodesic flows some partial answers arise from the theory of evolution equations. We shall restrict ourselves to the theory of surfaces.

The Ricci flow for surfaces is a curve $(M, g_t)$ of Riemannian metrics of a surface $M$ defined by the following partial differential equation:

$$\frac{\partial g_t}{\partial t} = -2K_t \cdot g_t,$$

where $K_t$ is the Ricci curvature of the surface, that is the Gaussian curvature actually. The Ricci flow theory comes from relativity, it would give the path to deform a metric in order to get the "best" metric in a manifold. The word "best" refers to spaces with symmetries or constant curvature. The works of Hamilton [**?**], [**?**] for surfaces and three-spheres with positive Ricci curvature showed that the Ricci flow was a powerful tool to find metrics of constant curvature. The Ricci flow on compact surfaces with genus greater than one tends to a metric of constant negative curvature, in the sphere case it tends to a metric of positive constant curvature. Perelman [98], [16] used the Ricci flow to show the famous Poincaré conjecture: a compact, simply connected three manifold is diffeomorphic to the sphere. The solution of this problem led to the classification of three manifolds after the work of Thurston [127]. All these results concern the asymptotic evolution of the Ricci flow, while we are rather looking at boundary points in the set of metrics. Indeed, since Anosov geodesic flows are persistent and expansive, we would like to consider expansive, non-Anosov geodesic flows in $(M, g)$ and show that there exists a curve $g_t$ of metrics such that $g_0 = g$ and such that $(M, g_t)$ is Anosov for short time $t$.

What we know is the following:

**Lemma 10.1.** *Compact surfaces of non-positive curvature and genus greater than one are in the closure of surfaces of negative curvature.*

The proof of this result relies on the application of the Maximum principle for parabolic equations (see for instance [103]) applied to the Ricci flow. A complete argument can be found in [59]. Lemma 10.1 seems to encourage the use of the Ricci flow to study surfaces without conjugate points, however a control of the curvature sign of $g_t$ becomes extremely difficult if the surface $(M, g)$ has regions of positive curvature. Without an accurate control over the regions of positive curvature to show that there are no conjugate points seems to be out of reach.

Other evolution flows have been considered in the literature in the last 15 years, this time arising from magnetic field theory. The so called **Ricci-Yang-Mills** flow considers the coupled evolution of a Riemannian metric $(M, g)$ and a smooth function $m : M \longrightarrow \mathbb{R}$ that plays in physics the role of a Lorentz force, according to the following equations:

$$\frac{\partial g_t}{\partial t} = (m_t^2 - 2K_t) \cdot g_t, \tag{1a}$$

$$\frac{\partial m_t}{\partial t} = \triangle_t m_t + 2K_t m_t - m_t^3. \tag{1b}$$

The scalar function $m$ is also called the magnetic potential. The occurrence of terms which depend on the derivatives of $m$ has a short term effect on the sign of the derivative of the curvature. Indeed, in a paper by D. Jane and the author [59], it is proved that given a compact surface $(M, g)$ with genus greater than one, and a magnetic potential $m$ that is subharmonic in the complement of a small ball of negative curvature, then the Ricci-Yang-Mills flow shrinks the regions of positive curvature and decreases the curvature in these regions. Combining this fact and subtle properties of Jacobi fields in surfaces without focal points we show the following result:

**Theorem 10.2.** *Compact surfaces without focal points such that the region of positive curvature consists of a finite number of "isolated" bubbles are in the closure of Anosov metrics.*

A bubble is a region of positive curvature that is simply connected, whose boundary has zero curvature, and whose closure is surrounded by a thin annulus of negative curvature. Many examples of surfaces without conjugate points are obtained in this way through surgery.

Notice that neither in Lemma 10.1 nor in Theorem 10.2 the assumption of expansiveness of the geodesic flow was considered. This shows that in the case of surfaces, we might expect not only that expansive geodesic flows are in the closure of Anosov metrics, but that every metric without conjugate points is in the closure as well.

## 11. Quotient spaces

The last section of the survey is somehow related to the previous one. In a paper by J. Lewowicz and R. Ures, On Smale diffeomorphisms close to pseudo-Anosov maps, [73] (2001), they consider quotient spaces of maps which are close to pseudo-Anosov as expansive models of such systems up to semi-conjugacy. Applying this idea they show that expansive homeomorphisms which are in the closure of pseudo-Anosov maps in the $C^0$ topology are in fact conjugate to such maps. These quotient spaces have finite topological dimension (Mañé [82]) and the quotient dynamics inherits the topology of the initial dynamics (namely, isotopy to a Smale diffeomorphism implies transitivity, density of periodic orbits, etc).

The idea of an expansive quotient as a model dynamics up to semi-conjugacy suits well in the theory of compact surfaces without conjugate points and higher genus. Stable and unstable sets always exist, they are the stable and unstable horospheres respectively. And although the dynamics might not be expansive the saturation by the flow of these sets gives rise to two continuous foliations of the unit tangent bundle by Lipschitz codimension one submanifolds: the center stable and the center unstable foliations. This fact follows essentially from the work of Morse [92] and the divergence of geodesic rays in the universal covering of surfaces without conjugate points proved by Green [49]. We can lift the foliations to the unit tangent bundle of the universal covering and get a pair of foliations by embeddings of the plane. The intersection of a lifted stable leaf with a lifted unstable leaf might not be a single orbit as in the expansive case, but the structure

of such intersections is quite nice. In the case of surfaces without focal points, the intersection of a center stable leaf with a center unstable leaf is a union of flat strips, which have many strong topological properties. In a work in progress with A. De Carvalho we show,

**Theorem 11.1.** *If $(M, g)$ is a compact surface without focal points, the quotient $\Sigma$ of $T_1M$ by flats is a 3-dimensional manifold and the quotient $\bar{\phi}_t : \Sigma \longrightarrow \Sigma$ of the geodesic flow $\phi_t$ is semi-conjugate to $\phi_t$ by a time-preserving map.*

The existence of non-time-preserving semi-conjugacies between geodesic flows of compact surfaces without conjugate points and Anosov geodesic flows is well known since the works of M. Gromov and E. Ghys [47]. Theorem 11.1 tells us that there is an expansive, time preserving model for the dynamics of the geodesic flow of the surface. Because the quotient of $T_1M$ by flats eliminates precisely the non-expansive subset of the initial dynamics. A rough idea of the proof is the following: according to the work of Morse [92], expansiveness is lost precisely when there are strips in the universal covering foliated by geodesics. The structure of strips in surfaces with no focal points is, like in non-positive curvature geometry, quite simple: they are flat and its union is a set of empty interior. Moreover, each strip is contractible to any one of the geodesics in it. General topology of low dimensional manifolds, the semi-continuity of the flat strips and the trivial topology of the strips yield that the quotient of $T_1M$ obtained by identifying the orbits in a strip with just one of them is homeomorphic to $T_1M$ and carries a continuous expansive flow.

We think that Theorem 11.1 may be extended to compact surfaces without conjugate points and certain manifolds without conjugate points in higher dimensions. However, the proof in those cases might be more technical than the proof of Theorem 11.1 since the structure of the non-expansiveness set of the flow might exhibit more complicated geometric features.

## References


[1] Akbar-Zadeh, H.: Sur les espaces de Finsler à courbures sectionnelles constantes. Acad. Roy. Belg. Bull. C1. Sci. (5) **LXXIV** (1988), 281–322.

[2] Anantharamam, N.: Counting geodesics which are optimal in homology. Erg. Theo. and Dyn. Syst., 23 (2): 353-388, 2003.

[3] Anantharamam, N., R. Iturriaga, P. Padilla, H. Snchez-Morgado: Physical solutions of the Hamilton-Jacobi equation. Disc. Contin. Dyn. Syst. Ser. B 5 (3) 513-528, 2005.

[4] Anosov, D.: Geodesic flow on closed Riemannian manifolds of negative curvature. Tr. Mat. Inst. Steklova 90 (1967).

[5] Arnold, V. I.: Mathematical Methods of Classical Mechanics. Second Edition. Graduate Texts in Mathematics, 60. Springer-Verlag, New York, Berlin, Heidelberg.

[6] Ballmann, W., Brin, M., Burns, K.: On surfaces with no conjugate points. J. Diff. Geom. 25 (1987), 249-273.

[7] Ballman, W., Gromov, M., Schroeder, V.: Manifolds of Non-positive curvature. Boston, Birkhausser 1985.

[8] Ballmann, W., Brin, M., Burns, K.: On the differentiability of horocycles and horocycle foliations. Journal of Dif. Geom. 26 (1987) 337-347.

[9] Bangert, V., Schroeder, V.: Existence of flat tori in analytic manifolds of nonpositive curvature. Ann. Sci. Ec. Norm. Sup. 24 (1991) 605-634.

[10] Bao, D.,Chern, S.S., Shen, Z.: Rigidity issues on Finsler surfaces, Rev. Roumaine Math. Pures Appl. **42** (1997), 707–735.

[11] Bao, D., Chern, S.-S., Shen, Z.: An Introduction to Riemann-Finsler Geometry, Springer, New York, 2000.

[12] Barbosa Gomes, J. B., Ruggiero, R. O.: Smooth k-basic Finsler compact surfaces with expansive geodesic flows are Riemannian. Houston Journal of Mathematics (2011) 37(3) 793–806.

[13] Barbosa Gomes, J. B., Ruggiero, R. O.: On Finsler surfaces without conjugate points. Ergodic Theory and Dynamical Systems. Available on CJO 2012, doi: 10.1017/SO143385711001027.
[14] Benakli, N., Kapovich, I.: Boundaries of hyperbolic groups. Contemporary Mathematics (2002) American Math. Society.
[15] Bernard, P., Contreras, G.: A generic property of families of Lagrangian systems. Ann. of Math. (2) 167 (2008) n.3. 1099-1108.
[16] Besson, G.: Preuve de la conjecture de Poincaré en déformant la métrique par la courbure de Ricci, d'après G. Perelman. Astérisque 307 (2006). Société Mathématique de France.
[17] Bonk, M.: Quasi-geodesic segments and Gromov hyperbolic spaces. Geom. Ded. (1996) 62, 281-298.
[18] Bowen, R.: Equilibrium states and ergodic theory of Anosov diffeomorphisms. Lecture Notes in Math. Springer, Berlin, 1975.
[19] Bousch, T.: Le Poisson n'a pas d'arête. Ann. Inst. Henry Poincaré, 36 (2000), 459-508.
[20] Bowen, R.: Symbolic dynamics for hyperbolic flows. American Journal of Mathematics 95 (1972) 429-459.
[21] Brin, M.: Topological transitivity of one class of dynamical systems and flows of frames on manifolds of negative curvature. Funk. Anal. Appl. 9 (1975) 9-19.
[22] Brin, M.: The topology of group extensions of $C$ systems. Mat. Zametki 18 (1975) 453-465.
[23] Brin, M., Gromov, M.: Brin, M., Gromov, M.: On the ergodicity of frame flows. Invent. Math. 60 (1980) 1-7.
[24] Brin, M., Pesin, Ya.: Flows of frames on manifolds of negative curvature. Uspehi Math. Nauk. 28 (1973) 169-170.
[25] Burns, K., Pugh, C., Wilkinson, A.: Stable ergodicity and Anosov flows. Topology 39 (2000) 149-159.
[26] Busemann, H.: The Geometry of Finsler spaces, Bulletin of the AMS **56** (1950), 5–16.
[27] Busemann, H.: The geometry of geodesics. New York, Academic Press. 1955.
[28] Cartan, E.: Leçons sur la Géométrie des Espaces de Riemann, Gauthiers-Villars, Paris, 1951.
[29] Chernov, N., Markarian, R.: Introduction to the ergodic theory of chaotic billiards. Publicações Matemáticas do IMPA. 24o. Colóquio Brasileiro de Matemática, Instituto de Matemática Pura e Aplicada, Rio de Janeiro (2003) 207 pp.
[30] Contreras, G.: Geodesic flows with positive topological entropy, twist maps and hyperbolicity. Preprint 2008.
[31] Contreras, G., Iturriaga, R., Paternain, G. P., Paternain, M.: Lagrangian Graphs, Minimizing Measures and Mañé's Critical Values, Geometric And Functional Analysis **8** (1998), 788–809.
[32] Contreras, G., Lopes A. O. and Thieullen, P: Lyapunov Minimizing measures for expanding maps of the circle. Ergod. Th. and Dynam. Sys. Vol 21 (2001) Issue 5, 1379–1409.
[33] Contreras, G., Paternain, G.: Genericity of geodesic flows with positive topological entropy on $S^2$. Journal of Differential Geometry 61 (2002) 1-49.
[34] Contreras, G., Ruggiero, R.: Non-hyperbolic surfaces having all ideal triangles of finite area. Bul. Braz. math. Soc. 28, 1 (1997) 43-71.
[35] Croke, C.: Rigidity for surfaces of nonpositive curvature. Comment. Math. Helv. 65 (1990) n.1, 150-169.
[36] Croke, F., Fathi, A.: An inequality between energy and intersection. Bull. London Math. Soc. 22 (1990) n. 5, 489-494.
[37] De La Harpe, P., Ghys, E.: Sur les groupes hyperboliques d'après M. Gromov. Progress in Mathematics, 83. Birkhauser. Zurich.
[38] De la Llave, R., Marco, J. M., Morillón, R.: Canonical perturbation theory of Anosov systems and regularity results for the Livsic cohomology equation. Annals of Mathematics, 123 (1986), 537-611.
[39] Hasselblat, B., Katok, A.: Introduction to the Modern Theory of Dynamical Systems. Encyclopedia of Mathematics and its applications, vol. 54 (1995), G.-C. Rota Editor. Cambridge University Press.
[40] Eberlein, P.: Geodesic flows in certain manifolds without conjugate points. Trans. Amer. Math. Soc. 167 (1972) 151-170.
[41] Eberlein, P.: When is a geodesic flow of Anosov type I. J. Diff. Geom. 8 (1973) 437-463.
[42] Eberlein, P.: Geodesic flows in manifolds of nonpositive curvature. Proceedings of Symposia in Pure Mathematics. Volume 69 (2001) 525-571, AMS Providence, Rhode Island.
[43] Eberlein, P., O'Neil, B.: Visibility manifolds. Pacific J. Math. 46 (1973) 45-109.
[44] Fanai, H. R.: Spectre marqu des longueurs et mtriques conformment quivalentes. [Marked length spectrum and conformally equivalent metrics] Bull. Belg. Math. Soc. Simon Stevin 5 (1998), no. 4, 525–528.

[45] Fathi, A., Siconolfi, A.: Existence of $C^1$ critical subsolutions of the Hamilton-Jacobi equation. Invent. Math. 155 (2004) 2, 363-388.
[46] Gerber, M., Nitica, V. : Hölder exponents of horocycles foliations on surfaces. Ergod. Th. Dyn. Sys. 9 (1999) 5, 1247-1254.
[47] Ghys, E.: Flots d'Anosov sur les 3-variétés fibrées en cercles. Ergod. Th. and Dynam. Sys., 4 (1984) 67–80.
[48] Ghys, E.: Flots d'Anosov dont les feuilletages stables sont différentiables. Ann. Scient. c. Norm. Sup. 20 (1987), 251-270.
[49] Green, L.: Geodesic instability. Proc. Amer. Math. Soc. 7 (1956) 438-448.
[50] Green, L.: A theorem of E. Hopf. Michigan Math. Journal 5 (1958) 31-34.
[51] Gromov, M.: Hyperbolic groups. Essays in group theory 75-263. Gersten Ed. Springer-Verlag, New York.
[52] Hempel, J.: 3-manifolds. Providence, RI, American Mathematical Society, ISBN 0-8218-3695-1.
[53] Hirsch, M., Pugh, C., Shub, M.: Invariant manifolds. Vol. 583 of Lecture Notes in Mathematics, Springer-Verlag 1977.
[54] Hopf, E.: Statistik der geodätischen Linien in Mannigfaltigkeiten negativer Krümmung. Ber. Verh. Sächs. Akad. Wiss. Leipzig 91, (1939). 261–304.
[55] Hopf, E.: Closed surfaces without conjugate points. Proceedings of the National Academy of Sciences of the United States of America (1948), 47-51.
[56] Hormander, L. : Hypoelliptic second order differential equations. Acta Math. 119 (1967) 147-171.
[57] Inaba, T., Matsumoto, S.: Nonsingular expansive flows on 3-manifolds and foliations with circle prone singularities. Japan J. of Math. (N.S) 16 (1990)n. 2, 329-340.
[58] Jaco, W., Shalen, P.: Seifert fibred spaces in 3-manifolds. Mem. Amer. Math. Soc. (1979) 220.
[59] Jane, D., Ruggiero, R: Boundary of Anosov dynamics and evolution equations for surfaces. Preprint PUC-Rio (2012).
[60] Johannson, K.: Homotopy equivalences of 3-manifolds with boundary. Lecture Notes in Mathematics, 761, Springer, Berlin 1979.
[61] Katok, A.: Infinitessimal Lyapunov functions, invariant cones families and stochastic properties of smooth dynamical systems. With the collaboration fo Keith Burns. Ergod. Th. Dynam. Sys. 14 (1994) 757-785.
[62] Kleiner, B., Lott, J.: Notes on Perelman's papers (2005). Available in electronic version at http://arxiv.org/abs/math.DG/0605667.
[63] Klingenberg, W.: Riemannian manifolds with geodesic flows of Anosov type. Ann. of Math. 99 (1974) 1-13.
[64] Klingenberg, W.: Lectures on closed geodesics. Springer, Berlin-Heidelberg-New York, 1974.
[65] Klingenberg, W., Takens, F.: Generic properties of geodesic flows. Math. Ann. 197 (1972) 323-334.
[66] Guillemin, V., Kazdan, D.: On the cohomology of certain dynamical systems. Topology, 19 (1980), 291-299.
[67] Lewowicz, J.: Lyapunov functions and topological stability. Journal of Differential equations, 38 (1980) 2, 192-209.
[68] Lewowicz, J.: Invariant manifolds for regular points. Pacific Journal of Mathematics 96 (1981) 1, 163-173.
[69] Lewowicz, J.: Lyapunov functions and stability of geodesic flows, Geometric Dynamics (Rio de Janeiro) Lecture Notes in Math. 1007, 1983.
[70] Lewowicz, J.: Persistence in expansive systems. Ergod. Th. Dynam. Sys. 3 (1983) 4, 567-578.
[71] Lewowicz, J.: Expansive homeomorphisms of surfaces. Bol. Soc. Bras. Mat. (1989) 20, 113-133.
[72] Lewowicz, J., Lima de Sá, E.: Analytic models of pseudo-Anosov maps. Ergod. Th. Dynam. Sys. 6 (1986) 3, 385-392.
[73] Lewowicz, J., Ures, R. : On Smale diffeomorphisms close to pseudo-Anosov maps. Comput. Appl. Math. 20 (2001) n. 1-2, 187-194.
[74] Liao, S. T.: On the stability conjecture. Chinese Annals of Mathematics 1 (1980) 9-30.
[75] Livsic, A.: Some homology properties of Y-systems. Mathematical notes of the USSR Academy of Sciences, 10 (1971) 758-763.
[76] Lopes, A., Meneses, V., Ruggiero, R.: Cohomology and subcohomology problems for expansive, non-Anosov geodesic flows. Discrete and Continuous Dynamical systems A, 17 (2007) 403-422.
[77] Lopes, A., Ruggiero, R.: Large deviations and Aubry-Mather measures supported in non-hyperbolic closed geodesics. Discrete Contin. Dyn. Syst. 29 (2011) n. 3, 1155-1174.

[78] Lopes A. O. and Thieullen, P: Subactions for Anosov Diffeomorphisms. Astérisque volume 287 (2003) *Geometric Methods in Dynamics (II)*, 135–146 .
[79] Lopes A. O. and Thieullen, P: Subactions for Anosov Flows. Ergod. Th. and Dynam. Sys. Vol 25 (2005) Issue 2, 605–628 .
[80] Lopes A. O. and Thieullen, P: Mather Theory and the Bowen-Series transformation. To appear in Annal. Inst. Henry Poincaré, Anal Non-lin. preprint (2002).
[81] Mañé, R.: Quasi-Anosov diffeomorphisms. Lecture Notes in Math. Vol. 468, pp. 27-29. Berlin, Heildelberg, New York, Springer 1974.
[82] Mañé, R.: Expansive homeomorphisms and topological dimension. Trans. Amer. Math. Soc. 252 (1979) 313-319.
[83] Mañé, R.: An ergodic closing lemma. Ann. of Math. 116 (1982) 503-540.
[84] Mañé, R.: A proof of the $C^1$ stability conjecture. Publications Mathématiques de l'IHES (1987) 66, 721-724.
[85] Mañé, R.: On a theorem of Klingenberg, In: M. I. Camacho, M. J. Pacífico, F. Takens (Ed.), DYNAMICAL SYSTEMS AND BIFURCATION THEORY, Longman Scientific & Technical, New York, 1987, pp. 319–345.
[86] Mañé, R.: On the minimizing measures of Lagrangian dynamical systems. Nonlinearity 5 (1992) n. 3, 623-638.
[87] Markarian, R.: Billiards with Pesin region of measure one. Comm. Math. Phys. 118 (1988) n.1, 87-97.
[88] Markarian, R.: Ergodic properties of plane billiard with symmetric potentials. Comm. Math. Phys. 145 (1992) n. 3, 435-446.
[89] Massart, D.: On Aubry sets and Mather's action functional. Israel J. Math. 134 (2003) 157-171.
[90] Mather, J.: Variational construction of connecting orbits. Ann. Inst. Fourier 43 (1993) 1349-1386.
[91] Milnor, J.: A unique factorization theorem for 3-manifolds. Amer. J. Math. 79 (1962) 1-7.
[92] Morse, M.: A fundamental class of geodesics on any closed surface of genus greater than one. Trans. Amer. Math. Soc. 26 (1924) 25-60.
[93] Newhouse, S.: Quasi-elliptic periodic points in conservative dynamical systems. Amer. J. Math. 99 (1977), no. 5, 1061–1087.
[94] Otal, J-P.: Le spectre marqué des longueurs des surfaces à courbure négative. Annals of Math. 2, 131 (1990) n.1, 151-162.
[95] Paternain, G.: Finsler structures on surfaces with negative Euler characteristic. Houston Journal of Mathematics **23** (1997), 421–426.
[96] Paternain, G. P., Paternain, M.: Expansivity for optical Hamiltonian systems with two degrees of freedom. C. R. Acad. Sci. Paris, Série I **316** (1993), 837–841
[97] Paternain, M.: Expansive geodesic flows on surfaces. Ergod. Th. Dynam. Sys. 13 (1993) 153-165.
[98] Perelman,G. Ricci Flow with Surgery on Three-Manifolds. Preprint, March 2003.
[99] Pesin, Ya. B.: Geodesic flows on closed Riemannian manifolds without focal points. Math. USSR Izvestija. 11 (1977) 1195-1228.
[100] Pollicott, M and Sharp, R.: Livsic theorems, Maximizing measures and the stable norm. Dynamical Systems, Volume 19 (2004) Number 1, 75–88.
[101] Potrie, R.: On the work of J. Lewowicz on expansive systems. Prepublicaciones Matemáticas del Uruguay. 2012/143.
[102] Preissmann, A.: Quelques propriétés globales des espaces de Riemann. Comm. Math. Helv. 15 (1943) 175-216.
[103] Protter, Murray H, Weinberger H.: Maximum principles in differential equations. Prentice-Hall Inc. Englewood Clifs, N.J. 1967.
[104] Pugh, C.: The closing lemma. Amer. J. Math. 89 (1967) 956-1009.
[105] Rifford, L., Ruggiero, R.: Generic properties of closed orbits of Hamiltonian flows from Mañé's viewpoint. International Mathematical Research Notices (2011) doi: 10/1093/imnr/rnr231
[106] Robbin, J. W.: A structural stability theorem. Annals of Mathematics (1971) 94, 447-493.
[107] Robinson, C.: Structural stability of $C^1$ diffeomorphisms. Journal of differential equations (1976) 22, 28-73.
[108] Ruggiero, R.: Persistently expansive geodesic flows. Commun. Math. Phys. 140 (1991) 203-215.
[109] Ruggiero, R.: On the creation of conjugate points. Math. Z. 208 (1991) 41-55.
[110] Ruggiero, R, Expansive dynamics and hyperbolic geometry. Bul. Braz. Math. Soc. vol. 25, n. 2 (1994) 139-172.

[111] Ruggiero, R.: On a conjecture about expansive geodesic flows. Ergod. Th. Dynam. Sys. 16 (1996) 545-553.
[112] Ruggiero, R.: Flatness of Gaussian curvature and area of ideal triangles. Bul. Braz. Math. Soc. 28, 1 (1997) 73-87.
[113] Ruggiero, R.: Expansive geodesic flows in manifolds with no conjugate points. Ergod. Th. Dynam. Sys. 17 (1997) 211-225.
[114] Ruggiero, R.: On nonhyperbolic quasiconvex spaces. Trans. Amer. Math. Soc. 350, 2 (1998) 665-687.
[115] Ruggiero, R.: Topological stability and Gromov hyperbolicity. Ergod. Th. Dynam. Sys. (1999), 19, 143-154.
[116] Ruggiero, R.: Weak stability of the geodesic flow and Preissmann's theorem. Ergod. Th. Dynam. Sys. (2000),20, 1231-1251.
[117] Ruggiero, R.: On the nonexistence of rational geodesic foliations in the torus, Mather sets and Gromov hyperbolic spaces. Bol. Soc. Bras. Mat. (2000) 31, 1, 93-111.
[118] Ruggiero, R.: On the divergence of geodesic rays in manifolds without conjugate points, dynamics of the geodesic flow and global geometry. Astérisque 287 (2003) 231-250. Geometric methods in dynamics, II, volume in honor of Jacob Palis. De Melo, Viana, Yoccoz Ed.
[119] Ruggiero, R.: Shadowing of geodesics, weak stability of the geodesic flow and global hyperbolic geometry. Discrete and Continuous Dyn. Sys. Vol. 14, n. 2 (2006).
[120] Ruggiero, R.: Dynamics and global geometry of manifolds without conjugate points. Ensaios Matemáticos, vol. 12. Sociedade Brasileira de Matemática, (2007).
[121] Ruggiero, R.: The accessibility property of expansive geodesic flows without conjugate points. Ergod. Th. and Dynam. Sys. 28 (2008) 229-244.
[122] Sakai, K.: Diffeomorphisms with weak shadowing. Fundamenta Math. 168 (2001) 1, 57-75.
[123] Sakai, K.: Diffeomorphisms with $C^2$ stable shadowing. Dyn. Sys. 17 (2002) 3, 235-241.
[124] Schröeder, V.: Codimension one tori in manifolds of non-positive curvature. Geom. Dedicata (1990) 33, 251-263.
[125] Scott, P.: The geometries of 3-manifolds. Bull. London Math. Soc. 15 (1983) 401-487.
[126] Smale, S. Generalized Poincaré's Conjecture in Dimensions Greater than Four. Ann. Math. 74 (1961) 391-406.
[127] Thurston, W.: Hyperbolic structures on three manifolds I. Annals of Math. 124 (1986) 203-246.
[128] Thurston, W.: Hyperbolic structures on three manifolds II. Preprint (1987).
[129] Walters, P.: On the pseudo-orbit tracing property and its relationship to stability. Lecture Notes in Mathematics, 608 (1978) 231-244.
[130] Wojtkowski, M. : Monotonicity, J-algebra of Potapov and Lyapunov exponents. Smooth ergodic theory and its applications (Seattle, WA, 1999) Proc. Sympos. Pure Math 69. Amer. Math. Soc. Providence RI, 2001.
[131] Young, L-S. : Recurrence times and rates of mixing. Israel J. Math. 110, 1999, 153–188.

Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ, Brazil, 22453-900
*E-mail address*: rorr@mat.puc-rio.br

# EXPANSIVE MEASURES

A. ARBIETO, C. A. MORALES

Abstract. This survey is about *expansive measures*, namely, Borel probability measures for which the dynamical balls up to some prefixed radio have measure zero. Some properties dealing with ergodicity, variational principle, pressure, heteroclinic points etc will be considered.

The object of study in this survey, namely, the expansive measures, is motivated by the old definition of expansive system due to Utz [46] (bijective case) and Eisenberg [14] (general case). More precisely, we say that a map (resp. bijective map) $f : X \to X$ of a metric space $X$ is *positively expansive* (resp. *expansive*) if it has an *expansivity constant*, i.e., a positive number $\epsilon$ with the property that if $x, y \in X$ and $d(f^n(x), f^n(y)) \leq \epsilon$ for all $n \in \mathbb{N}$ (resp. $n \in \mathbb{Z}$), then $x = y$.

The concept of expansivity althouth simple has been very successful in the theory of dynamical systems. For instance, [50] proved that the set of points doubly asymptotic to a given point for expansive homeomorphisms is at most countable. Moreover, a homeomorphism of a compact metric space is expansive if it does in the complement of finitely many orbits [51]. In 1972 Sears proved the denseness of expansive homeomorphisms with respect to the uniform topology in the space of homeomorphisms of a Cantor set [42]. An study of expansive homeomorphisms using generators is given in [7]. Goodman [18] proved that every expansive homeomorphism of a compact metric space has a (nonnecessarily unique) measure of maximal entropy and Bowen [5] added specification to obtain unique equilibrium states. In another direction, [40] studied expansive homeomorphisms with canonical coordinates and showed in the locally connected case that sinks or sources cannot exist. Two years later, Fathi characterized expansive homeomorphisms on compact metric spaces as those exhibiting adapted hyperbolic metrics [16] (see also [41] or [12] for more about adapted metrics). Using this he was able to obtain an upper bound of the Hausdorff dimension and upper capacity of the underlying space using the topological entropy. In [26] it is computed the large deviations of irregular periodic orbits for expansive homeomorphisms with the specification property. The $C^0$ perturbations of expansive homeomorphisms on compact metric spaces were considered in [9]. Besides, the multifractal analysis of expansive homeomorphisms with the specification property was carried out in [45]. We can also mention [8] in which it is studied a new measure-theoretic pressure for expansive homeomorphisms.

From the topological viewpoint we can mention [35] and [38] proving the existence of expansive homeomorphisms in the genus two closed surface, the $n$-torus and the open disk. Analogously for compact surfaces obtained by making holes on closed surfaces different from the sphere, projective plane and Klein bottle [24]. In [22] it was proved

2010 *Mathematics Subject Classification.* Primary: 37A25; Secondary: 37A35.
*Key words and phrases.* Expansive Measure, Homeomorphism, Entropy.
Partially supported by CNPq, FAPERJ and PRONEX/DS from Brazil.

that there are no expansive homeomorphisms of the compact interval, the circle and the compact 2-disk.

On the other hand, on of the most important result in this direction, namely, the classification of expansive homeomorphisms on closed surfaces, was obtained independently by Lewowicz and Hiraide [28], [20]. In particular, they proved that there are no expansive homeomorphism in the 2-dimensional sphere. Mañé proved in [32] that a compact metric space exhibiting expansive homeomorphisms must be finite dimensional and, further, every minimal set of such homeomorphisms is zero dimensional. Previously he proved that the $C^1$ interior of the set of expansive diffeomorphisms of a closed manifold is composed by pseudo-Anosov (and hence Axiom A) diffeomorphisms. In 1993 Vieitez [47] obtained results about expansive homeomorphisms on closed 3-manifolds including the denseness of the topologically hyperbolic periodic points does imply constant dimension of the stable and unstable sets. As a consequence a local product property was obtained for such homeomorphisms. He also obtained that expansive homeomorphisms on closed 3-manifolds with dense topologically hyperbolic periodic points are both supported on the 3-torus and topologically conjugated to linear Anosov isomorphisms [48]. A nice account of the theory of expansive systems can be found in [10].

In light of these results it was natural to consider another notions of expansiveness. For example, $G$-expansiveness, continuouswise and pointwise expansiveness were defined in [11], [23] and [39] respectivelly. We also have the notion of *entropy-expansiveness* introduced by Bowen [4] in order to compute the metric and topological entropies for certain homeomorphisms.

The present authors were motivated by these results and considered the analogous concept but now involving Borel measures $\mu$ in $X$. As a motivation observe that a map (resp. bijective map) $f : X \to X$ is positively expansive (resp. expansive) if and only if there is $\epsilon > 0$ such that,

$$\{y \in X : d(f^n(x), f^n(y)) \leq \epsilon, \forall n \in \mathbb{N} \text{ (resp. } \forall n \in \mathbb{Z})\} = \{x\}, \qquad \forall x \in X.$$

This suggests the folowing definition:

**Definition 1.** *Let $f : X \to X$ be a measurable map (resp. measurable bijective map) of $X$. A Borel measure $\mu$ of $X$ is a* positively expansive measure *(resp.* expansive measure*) of $f$ if there is $\epsilon > 0$ such that,*

$$\mu(\{y \in X : d(f^n(x), f^n(y)) \leq \epsilon, \forall n \in \mathbb{N} \text{ (resp. } \forall n \in \mathbb{Z})\}) = 0, \qquad \forall x \in X.$$

Hereafter all maps or bijections will be measurable (with respect to the Borel $\sigma$-algebra). Let us present some remarks related to the above concept.

**Remark 2.** *Every positively expansive measure of a given bijective map is also an expansive measure of it (but the converse is false in general). Every positively expansive or expansive measure is nonatomic, i.e., has no point of positive mass. Conversely, if $f$ is a positively expansive map (resp. an expansive bijection), then every nonatomic Borel measure is positively expansive (resp. expansive) for $f$.*

For Borel probability measures it is possible to put the following equivalence.

**Lemma 3.** *A Borel probability measure $\mu$ of a compact metric space $f$ is a positively expansive (resp. expansive) measure of a map (resp. bijective map) $f : X \to X$ if and only if there is $\epsilon > 0$ such that*

$$\mu(\{y \in X : d(f^n(x), f^n(y)) \leq \epsilon, \forall n \in \mathbb{N} \text{ (resp. } \forall n \in \mathbb{Z})\}) = 0, \text{ for } \mu\text{-a.e. } x \in X.$$

We then call *expansivity constant* of a Borel probability measure any constant $\epsilon$ satisfying the conclusion of either Definition 1 or Lemma 3.

Let us present a list of examples.

**Example 4.** *As is well known* [37]*, every complete separable metric space which either is uncountable or has no isolated points exhibits nonatomic Borel probability measures. It follows that every positively expansive map (or expansive bijection) in such a space has an expansive measure.*

**Example 5.** *There are expansive homeomorphisms on certain compact metric spaces with no expansive measures.*

*Proof.* Indeed, consider the map $p(x) = x^3$ in $\mathbb{R}$ and define $X = \{0, 1, -1\} \cup \{p^n(c) : n \in \mathbb{N}, c \in \{-\frac{1}{2}, \frac{1}{2}\}\}$. We have that $X$ is an infinite (but countable) compact metric space with the induced metric $d(x, y) = |x - y|$. Observe that there are no nonatomic Borel probability measures in $X$ since every non-isolated set of $X$ must be contained in $\{-1, 0, 1\}$. Defining $f(x) = p(x)$ for $x \in X$ we obtain an expansive homeomorphism $f$ which is not $\mu$-expansive for every Borel probability measure $\mu$. □

Further examples of homeomorphisms without expansive measures can be obtained as follows. Recall that an *isometry* of a metric space $X$ is a homeomorphism $f$ such that $d(f(x), f(y)) = d(x, y)$ for all $x, y \in X$. We use the following notation (for bijective maps),

$$\Gamma_\delta(x) = \Gamma_\delta^f(x) = \{y \in X : d(f^n(x), f^n(y)) \leq \epsilon, \forall n \in \mathbb{Z}\}.$$

**Example 6.** *Every isometry of a separable metric space has no expansive measures. In particular, the identity map in these spaces (or the rotations in $\mathbb{R}^2$ or translations in $\mathbb{R}^n$) has no such measures.*

*Proof.* Suppose by contradiction that there is an isometry $f$ of a separable metric space $X$ with some expansive measure $\mu$. Since $f$ is an isometry we have $\Gamma_\delta(x) = B[x, \delta]$, where $B[x, \delta]$ denotes the closed $\delta$-ball around $x$. If $\delta$ is an expansivity constant of $f$, then $\mu(B[x, \delta]) = \mu(\Gamma_\delta(x)) = 0$ for all $x \in X$. Nevertheless, since $X$ is separable (and so Lindelof), we can select a countable covering $\{C_1, C_2, \cdots, C_n, \cdots\}$ of $X$ by closed subsets such that for all $n$ there is $x_n \in X$ such that $C_n \subset B[x_n, \delta]$. Thus, $\mu(X) \leq \sum_{n=1}^{\infty} \mu(C_n) \leq \sum_{n=1}^{\infty} \mu(B[x_n, \delta]) = 0$ which is a contradiction. This proves the result. □

**Example 7.** *Endow $\mathbb{R}^n$ with a metric space with the Euclidean metric. Then, the Lebesgue measure Leb in $\mathbb{R}^n$ is an expansive measure of a linear isomorphism $f : \mathbb{R}^n \to \mathbb{R}^n$ if and only if $f$ has eigenvalues of modulus less than or bigger than* 1*.*

*Proof.* Since $f$ is linear we have $\Gamma_\delta(x) = \Gamma_\delta(0) + x$ thus $Leb(\Gamma_\delta(x)) = Leb(\Gamma_\delta(0))$ for all $x \in \mathbb{R}^n$ and $\delta > 0$. If $f$ has eigenvalues of modulus less than or bigger than 1, then $\Gamma_\delta(0)$ is contained in a proper subspace of $\mathbb{R}^n$ which implies $Leb(\Gamma_\delta(0)) = 0$ thus $Leb$is an expansive measure. □

**Example 8.** *There are no expansive measures for any homeomorphism of a compact interval $I$. In the circle $S^1$ the only homeomorphism with expansive measures are the Denjoy ones.*

Recall that a subset $Y \subset X$ is *invariant* if $f(Y) = Y$.

**Example 9.** *A homeomorphism $f$ has an expansive measure $\mu$ if and only if there is an invariant borelian set $Y$ of $f$ for which the restriction $f/Y$ has some expansive measure.*

*Proof.* We only have to prove the only if part. Assume that $f/Y$ has an expansive measure $\nu$ in $Y$. Fix $\delta > 0$. Since $Y$ is invariant we have either $\Gamma^f_{\delta/2}(x) \cap Y = \emptyset$ or $\Gamma^f_{\delta/2}(x) \cap Y \subset \Gamma^{f/Y}_{\delta}(y)$ for some $y \in Y$. Therefore, either $\Gamma^f_{\delta/2}(x) \cap Y = \emptyset$ or $\mu(\Gamma^f_{\delta/2}(x)) \leq \mu(\Gamma^{f/Y}_{\delta}(y))$ for some $y \in Y$ where $\mu$ is the Borel probability of $X$ defined by $\mu(A) = \nu(A \cap Y)$. From this we obtain that for all $x \in X$ there is $y \in Y$ such that $\mu(\Gamma^f_{\delta/2}(x)) \leq \nu(\Gamma^{f/Y}_{\delta}(y))$. Taking $\delta$ as an expansivity constant of $f/Y$ we obtain $\mu(\Gamma^f_{\delta/2}(x)) = 0$ for all $x \in X$ thus $f$ is $\mu$-expansive with expansivity constant $\delta/2$. □

The next example implies that the property of having expansive measures is a conjugacy invariant. Given a Borel measure $\mu$ in $X$ and a homeomorphism $\phi : X \to Y$ we denote by $\phi_*(\mu)$ the pullback of $\mu$ defined by $\phi_*(\mu)(A) = \mu(\phi^{-1}(A))$ for all borelian $A$.

**Example 10.** *Let $\mu$ be an expansive measure of a homeomorphism $f : X \to X$ of a compact metric space $X$. If $\phi : X \to Y$ is a homeomorphism of compact metric spaces, then$\phi_*(\mu)$ is an expansive measure of $\phi \circ f \circ \phi^{-1}$.*

*Proof.* Clearly $\phi$ is uniformly continuous so for all $\delta > 0$ there is $\epsilon > 0$ such that $\Gamma^{\phi\circ f\circ\phi}_{\epsilon}(y) \subset \phi(\Gamma^f_{\delta}(\phi^{-1}(y)))$ for all $y \in Y$. This implies

$$\phi_*(\mu)(\Gamma^{\phi\circ f\circ\phi}_{\epsilon}(y)) \leq \mu(\Gamma^f_{\delta}(\phi^{-1}(y))).$$

Taking $\delta$ as the expansivity constant of $\mu$ we obtain that $\epsilon$ is also an expansivity constant $\phi_*(\mu)$. □

For the next example recall that a *periodic point* of a map $f : X \to X$ is a point $x \in X$ such that $f^n(x) = x$ for some $n \in \mathbb{N}^+$ (the minimum of which is the so-called period denoted by $n_p$). The nonwandering set of $f$ is the set $\Omega(f)$ formed by those points $x \in X$ such that for every neighborhood $U$ of $x$ there is $n \in \mathbb{N}^+$ satisfying $f^n(U) \cap U \neq \emptyset$. Clearly a periodic point belongs to $\Omega(f)$ but not every point in $\Omega(f)$ is periodic. If $X = M$ is a *closed* (i.e. compact connected boundaryless) manifold and $f$ is a diffeomorphism we say that an invariant set $H$ is *hyperbolic* if there are a continuous invariant tangent bundle decomposition $T_H M = E^s_H \oplus E^u_H$ and positive constants $K$, $\lambda > 1$ such that

$$\|Df^n(x)/E^s_x\| \leq K\lambda^{-n} \qquad \text{and} \qquad m(Df^n(x)/E^u_x) \geq K^{-1}\lambda^n,$$

for all $x \in H$ and $n \in \mathbb{N}$ ($m$ denotes the co-norm operation in $M$). We say that $f$ is *Axiom A* if $\Omega(f)$ is hyperbolic and the closure of the set of periodic points.

**Example 11.** *Every Axiom A diffeomorphism with infinite nonwandering set of a closed manifold has expansive measures.*

*Proof.* Consider an Axiom A diffeomorphism $f$ of a closed manifold. It is well known that there is a spectral decomposition $\Omega(f) = H_1 \cup \cdots \cup H_k$ consisting of finitely many disjoint homoclinic classes $H_1, \cdots, H_k$ of $f$ (see [19] for the corresponding definitions). Since $\Omega(f)$ is infinite we have that $H = H_i$ is infinite for some $1 \leq i \leq k$. As is well known $f/H$ is expansive. On the other hand, $H$ is compact without isolated points since it is a homoclinic class. It follows that $f/H$ has an expansive measure, so, $f$ also has an expansive measure by Example 9. □

In the sequel we present some results about expansive measures. To state them we need some basic definitions.

For any map $f : X \to X$ and $p \in X$ we define the *stable set* of $p$ by

$$W^s(p) = \left\{ x \in X : \lim_{n\to\infty} d(f^n(x), f^n(p)) = 0 \right\}.$$

A *stable class* is a subset equals to $W^s(p)$ for some $p \in X$.

Let $\mu$ be a Borel probability measure of $X$. If $f$ is Borel measurable, we say that $\mu$ is *invariant* if $\mu \circ f^{-1} = \mu$ and *ergodic* if every invariant measurable set of $f$ has measure zero or one. The *entropy* of $\mu$ with respect to $f$ is defined by

$$h_\mu(f) = \sup \left\{ -\lim_{n\to\infty} \frac{1}{n} \sum_{\xi\in P_{n-1}} \mu(\xi)\log(\mu(\xi)) : P \text{ is a finite partition of } X \right\},$$

where $P_n$ is the pullback partition of $P$ under $f^n$ (see [49] for details).

Our first result is the following.

**Theorem 12.** *The stable classes of a continuous map of a compact metric space have measure zero with respect to any positively expansive measure.*

On the other hand, if $f$ is continuous we define its *topological entropy* by

$$h(f) = \sup \left\{ \lim_{n\to\infty} \frac{1}{n} \log(N(\alpha_n)) : \alpha \text{ is an open cover of } X \right\}$$

where $\alpha_n$ is the pullback of $\alpha$ under $f^n$ and $N(\alpha_n)$ is the cardinality of a finite subcover with minimal cardinality of $\alpha_n$ for all $n \in \mathbb{N}$ (see [49] for details).

Recall also that the *recurrent set* of $f$ is defined by $R(f) = \{x \in X : x \in \omega(x)\}$ where $\omega(x)$ is the *omega-limit set*

$$\omega(x) = \left\{ y \in X : y = \lim_{k\to\infty} f^{n_k}(x) \text{ for some sequence } n_k \to \infty \right\}.$$

(Notation $\omega_f(x)$ will indicate dependece on $f$). Following [17] we say that $f$ is *Lyapunov stable* on a subset $A \subset X$ if for any $x \in A$ and any $\epsilon > 0$ there is a neighborhood $U(x)$ of $x$ such that $d(f^n(x), f^n(y)) < \epsilon$ whenever $n \geq 0$ and $y \in U(x) \cap A$. Notice that this definition is implied by the corresponding one in [43]. We have our second result.

**Theorem 13.** *A continuous map with positive topological entropy of a compact metric space cannot be Lyapunov stable on its recurrent set.*

In the invertible case we also define the *alpha-limit set* $\alpha(x) = \alpha_f(x) = \omega_{f^{-1}}(x)$. We then say that $x \in X$ is a *heteroclinic point* if both $\alpha(x)$ and $\omega(x)$ reduce to periodic orbits. Our third result deals with the measure of the set of heteroclinic points with respect to ergodic measures with positive entropy.

**Theorem 14.** *The set of heteroclinic points of a homeomorphism of a compact metric space has measure zero with respect to any expansive measure.*

Following Definition 2.1 in [40] for every map $f : X \to X$ and every point $x \in X$ we define the *local stable set* for $\delta \geq 0$ by

$$W^s(x, \delta) = \{y \in X : d(f^n(x), f^n(y)) \leq \delta \text{ for all } n \in \mathbb{N}\}.$$

In the invertible case we also define the *local unstable set* for $\delta \geq 0$ by

$$W^u(x, \delta) = \{y \in X : d(f^{-n}(x), f^{-n}(y)) \leq \delta \text{ for all } n \in \mathbb{N}\}.$$

In such a case we say that $x \in X$ is a *sink* of $f$ if $W^u(x, \delta) = \{x\}$ for some $\delta > 0$.

Following [4] we say that an invertible map $f$ *has canonical coordinates* if for every $\epsilon > 0$ there is $\delta > 0$ such that $W^s(x,\delta) \cap W^u(y,\delta) \neq \emptyset$ whenever $d(x,y) \leq \delta$. Our last result is about the measure of the set of sinks, for invertible maps with canonical coordinates, with respect to ergodic invariant measures with positive entropy:

**Theorem 15.** *The set of sinks of any homeomorphism with canonical coordinates on a compact metric space has zero measure with respect to any expansive measure.*

Let $\mu$ be a Borel probability measure of $X$. If $f$ is Borel measurable, we say that $\mu$ is *invariant* if $\mu \circ f^{-1} = \mu$ and *ergodic* if every invariant measurable set of $f$ has measure zero or one. The *entropy* of $\mu$ with respect to $f$ is defined by

$$h_\mu(f) = \sup\left\{ -\lim_{n\to\infty} \frac{1}{n} \sum_{\xi \in P_{n-1}} \mu(\xi)\log(\mu(\xi)) : P \text{ is a finite partition of } X \right\},$$

where $P_n$ is the pullback partition of $P$ under $f^n$ (see [49] for details).

Based on the fundamental result by Brin and Katok [6] we can prove the following result.

**Theorem 16.** *Every ergodic invariant measure with positive entropy of a continuous map on a compact metric space is positively expansive.*

Let us recall the notion of topological pressure [49]. Given a continuous map $f : X \to X$ of a compact metric space $X$ and $T : X \to \mathbb{R}$ we define

$$S_n(f,T)(x) = \sum_{i=0}^{n-1} T(f^i(x)).$$

We call a subset $E \subset X$ $(n,\epsilon)$-separated for a given $(n,\epsilon) \in \mathbb{N} \times \mathbb{R}^+$ if for every pair of points $x \neq y$ in $E$ there is and integer $0 \leq i \leq n-1$ such that $d(f^i(x), f^i(y)) \geq \epsilon$. Let $s(n,\epsilon)$ the maximal cardinality of any $(n,\epsilon)$-separated subset $E$. Define

$$Z(f,T,\epsilon,n) = \sup\left\{ \sum_{x\in E} e^{S_n(f,T)(x)} : E \text{ is } (n,\epsilon)\text{-separated} \right\}.$$

The *topological pressure* of $T$ with respect to $f$ is defined by

$$P(f,T) = \lim_{\epsilon\to 0} \limsup_{n\to\infty} \frac{1}{n} \log Z(T,f,\epsilon,n).$$

The *variational principle* [49] says that

$$P(f,T) = \sup\left\{ h_\mu(f) + \int_X T d\mu : \mu \in \mathcal{M}_e(f) \right\},$$

where $\mathcal{M}_e(f)$ denotes the set of ergodic invariant measures of $f$. Combining this with Theorem 16 we obtain the following:

**Theorem 17.** *If $f : X \to X$ and $T : X \to \mathbb{R}$ are continuous maps, then*

$$P(f,T) = \sup\left\{ h_\mu(f) + \int_X T d\mu : \mu \in \mathcal{M}_{pex}(f) \right\},$$

*where $\mathcal{M}_{pex}(f)$ denotes the set of positively expansive measures of $f$.*

An analogous statement was announced by T. Fisher in his recent talk [15]. More consequences of Theorem 16 and the above results are given below.

**Corollary 18.** *The stable classes of a continuous map of a compact metric space have measure zero with respect to any ergodic invariant measure with positive entropy.*

**Corollary 19.** *A continuous map with positive topological entropy of a compact metric space has uncountably many stable classes.*

**Corollary 20.** *A continuous map with positive topological entropy of a compact metric space cannot be Lyapunov stable on its recurrent set.*

**Corollary 21.** *The set of heteroclinic points of a homeomorphism of a compact metric space has measure zero with respect to any ergodic invariant measure with positive entropy.*

**Corollary 22.** *The set of sinks of any homeomorphism with canonical coordinates on a compact metric space has zero measure with respect to any ergodic invariant measure of positive entropy.*

Let us now use the notion of positively expansive measure to study the chaoticity in the sense of Li and Yorke [30]. Recall that if $\delta \geq 0$ a *$\delta$-scrambled set* of $f : X \to X$ is a subset $S \subset X$ satisfying

$$\liminf_{n\to\infty} d(f^n(x), f^n(y)) = 0 \quad \text{and} \quad \limsup_{n\to\infty} d(f^n(x), f^n(y)) > \delta \tag{1}$$

for all different points $x, y \in S$.

**Theorem 23.** *A continuous map of a Polish space carrying an uncountable $\delta$-scrambled set for some $\delta > 0$ also carries positively expansive measures.*

*Proof.* Let $X$ a Polish space and $f : X \to X$ be a continuous map carrying an uncountable $\delta$-scrambled set for some $\delta > 0$. Then, by Theorem 16 in [3], there is a *closed* uncountable $\delta$-scrambled set $S$. As $S$ is closed and $X$ is Polish we have that $S$ is also a Polish space with respect to the induced metric. As $S$ is uncountable we have from [37] that there is a nonatomic Borel probability measure $\nu$ in $S$. Let $\mu$ be the Borel probability induced by $\nu$ in $X$, i.e., $\mu(A) = \nu(A \cap S)$ for all Borelian $A \subset X$. We shall prove that this measure is expansive. If $x \in S$ and $y \in \Phi_{\frac{\delta}{2}}(x) \cap S$ we have that $x, y \in S$ and $d(f^n(x), f^n(y)) \leq \frac{\delta}{2}$ for all $n \in \mathbb{N}$ therefore $x = y$ by the second inequality in (1). We conclude that $\Phi_{\frac{\delta}{2}}(x) \cap S = \{x\}$ for all $x \in S$. As $\nu$ is nonatomic we obtain $\mu(\Phi_{\frac{\delta}{2}}(x)) = \nu(\Phi_{\frac{\delta}{2}}(x) \cap S) = \nu(\{x\}) = 0$ for all $x \in S$. On other hand, it is clear that every open set which does not intersect $S$ has $\mu$-measure 0 so $\mu$ is supported in the closure of $S$. As $S$ is closed we obtain that $\mu$ is supported on $S$. We conclude that $\mu(\Phi_{\frac{\delta}{2}}(x)) = 0$ for $\mu$-a.e. $x \in X$, so, $\mu$ is expansive by Lemma 3. □

A continuous map is *Li-Yorke chaotic* if it has an uncountable 0-scrambled set. Until the end $M$ will denote either the interval $I = [0, 1]$ or the unit circle $S^1$.

**Corollary 24.** *Every Li-Yorke chaotic map in $M$ carries positively expansive measures.*

*Proof.* Theorem in p. 260 of [13] together with theorems A and B in [27] imply that every Li-Yorke chaotic map in $I$ or $S^1$ has an uncountable $\delta$-scrambled set for some $\delta > 0$. Then, we obtain the result from Theorem 23. □

It follows that there are continuous maps with zero topological entropy in the circle exhibiting expansive *invariant* measures. This leads to the question whether the same result is true on compact intervals. The following consequence of the above corollary gives a partial positive answer for this question.

**Example 25.** *There are continuous maps with zero topological entropy in the interval carrying positively expansive measures.*

Indeed, by [21] there is a continuous map of the interval, with zero topological entropy, exhibiting a $\delta$-scrambled set of positive Lebesgue measures for some $\delta > 0$. Since sets with positive Lebesgue measure are uncountable we obtain an expansive measure from Theorem 23.

Another interesting example is this.

**Example 26.** *The Lebesgue measure is an ergodic invariant measure with positive entropy of the tent map $f(x) = 1 - |2x - 1|$ in $I$. Therefore, this measure is positively expansive by Theorem 16.*

It follows from this example that there are continuous maps in $I$ carrying expansive measures $\mu$ *with full support* (i.e. $supp(\mu) = I$). These maps also exist in $S^1$ (e.g. an expanding map). Now, we prove that Li-Yorke and positive topological entropy are equivalent properties among these maps in $I$. But previously we need a result based on the following well-known definition.

A *wandering interval* of a map $f : M \to M$ is an interval $J \subset M$ such that $f^n(J) \cap f^m(J) = \emptyset$ for all different integers $n, m \in \mathbb{N}$ and no point in $J$ belongs to the stable set of some periodic point.

**Lemma 27.** *If $f : M \to M$ is continuous, then every wandering interval has measure zero with respect to every expansive measure.*

*Proof.* Let $J$ a wandering interval and $\mu$ be an expansive measure with expansivity constant $\epsilon$. To prove $\mu(J) = 0$ it suffices to prove $Int(J) \cap supp(\mu) = 0$ since $\mu$ is nonatomic. As $J$ is a wandering interval one has $\lim_{n\to\infty} |f^n(J)| = 0$, where $|\cdot|$ denotes the length operation. From this there is a positive integer $n_0$ satisfying

$$|f^n(J)| < \epsilon, \qquad \forall n \geq n_0. \tag{2}$$

Now, take $x \in Int(J)$. Since $f$ is clearly uniformly continuous and $n_0$ is fixed we can select $\delta > 0$ such that $B[x, \delta] \subset Int(J)$ and $|f^n(B[x, \delta])| < \epsilon$ for $0 \leq n \leq n_0$. This together with (2) implies $|f^n(x) - f^n(y)| < \epsilon$ for all $n \in \mathbb{N}$ therefore $B[x, \delta] \subset \Phi_\epsilon(x)$ so $\mu(B[x, \delta]) = 0$ since $\epsilon$ is an expansivity constant. Thus $x \notin supp(\mu)$ and we are done. □

From this we obtain the following corollary.

**Corollary 28.** *A continuous map with expansive measures of the circle or the interval has no wandering intervals. Consequently, a continuous map of the interval carrying expansive measures with full support is Li-Yorke chaotic if and only if it has positive topological entropy.*

*Proof.* The first part is a direct consequence Lemma 2 while, the second, follows from the first since a continuous interval map without wandering intervals is Li-Yorke chaotic if and only if it has positive topological entropy [44]. □

Now we turn our attention to smooth ergodic theory. The motivation is the well-known fact that a diffeomorphism restricted to a hyperbolic basic set is expansive. In fact, it is tempting to say that every hyperbolic ergodic measures of a diffeomorphism is positively expansive (or at least expansive) but the Dirac measure supported on a hyperbolic periodic point is a counterexample. This shows that some extra hypotheses are necessary for a hyperbolic ergodic measure to be positively expansive. Indeed, by

the results above, we only need to recognize which conditions imply positive entropy. Let us state some basic definitions in order to present our result.

Assume that $X$ is a compact manifold and that $f$ is a $C^1$ diffeomorphism. We say that point $x \in X$ is a *regular point* whenever there are positive integers $s(x)$ and $\{\lambda_1(x), \cdots, \lambda_{s(x)}(x)\} \subset \mathbb{R}$ such that for every $v \in T_xM \setminus \{0\}$ there is $1 \leq i \leq s(x)$ such that

$$\lim_{n\to\infty} \frac{1}{n} \log \|Df^n(x)v\| = \lambda_i(x).$$

An invariant measure $\mu$ is called *hyperbolic* if there is a measurable subset $A$ with $\mu(A) = 1$ such that $\lambda_i(x) \neq 0$ for all $x \in A$ and all $1 \leq i \leq s(x)$.

On the other hand, the Eckmann-Ruelle conjecture [2] asserts that every hyperbolic ergodic measure $\mu$ is *exact-dimensional*, i.e., the limit below

$$d(x) = \lim_{r\to 0^+} \frac{\mu(B(x,r))}{r}$$

exists and is constant $\mu$-a.e. $x \in X$. This constant is the so-called *dimension* of $\mu$.

With these definitions we can state the following corollary.

**Theorem 29.** *Let $f$ be a $C^2$ diffeomorphism of a compact manifold.*

1. *Every hyperbolic ergodic measure of $f$ which either has positive dimension or is absolutely continuous with respect to Lebesgue is positively expansive.*
2. *If $f$ has a nonatomic hyperbolic ergodic measure, then $f$ also has a positively expansive ergodic invariant measure.*

*Proof.* Let us prove (1). First assume that the measure has positive dimension. As noticed in [2] p. 761 Theorem C′ p. 544 in [29] implies that if the entropy vanishes, then the stable and unstable dimension of the measure also do. In such a case we have from Theorem F p. 548 in [29] that the measure has zero dimension, a contradiction. Therefore, the measure has positive entropy and then we are done by Theorem 16.

Now assume that the measure is absolutely continuous with respect to the Lebesgue measure. Then, it is nonatomic so the argument in the proof of Theorem 4.2 p. 167 in [25] implies that it has at least one positive Lyapunov exponent. Therefore, the Pesin formula (c.f. p. 139 in [25]) implies positive entropy so we are done by Theorem 16.

To prove (2) we only have to see that Corollary 4.2 in [25] implies that every diffeomorphism as in the statement of (2) has positive topological entropy. Then, we are done by the variational principle (e.g. [49] or Theorem 17). □

Detailed proofs of some of the above results will appear in [1]. The concept of expansive measure has been generalized to measurable spaces in [34].

## References


[1] Arbieto, A., Morales, C., A., Some properties of positive entropy maps, Preprint 2012.

[2] Barreira, L., Pesin, Y., Schmeling, J., *Dimension and product structure of hyperbolic measures*, Ann. of Math. (2) 149 (1999), no. 3, 755–783.

[3] Blanchard, F., Huang, W., Snoha, L., Topological size of scrambled sets *Colloq. Math.* 110 (2008), no. 2, 293–361.

[4] Bowen, R., *Entropy-expansive maps*, Trans. Amer. Math. Soc. 164 (1972), 323–331.

[5] Bowen, R., *Some systems with unique equilibrium states*, Math. Systems Theory 8 (1974/75), no. 3, 193–202.

[6] Brin, M., Katok, A., *On local entropy*, Geometric dynamics (Rio de Janeiro, 1981), 30–38, Lecture Notes in Math., 1007, Springer, Berlin, 1983.

[7] Bryant, B. F., Walters, P., *Asymptotic properties of expansive homeomorphisms*, Math. Systems Theory 3 (1969), 60–66.

[8] Cao, Y.,, Zhao, Y., *Measure-theoretic pressure for subadditive potentials*, Nonlinear Anal. 70 (2009), no. 6, 2237–2247.
[9] Cerminara, M., Sambarino, M., *Stable and unstable sets of $C^0$ perturbations of expansive homeomorphisms of surfaces*, Nonlinearity 12 (1999), no. 2, 321–332.
[10] Cerminara, M., Lewowicz, J., *Expansive systems*, (2008) Scholarpedia, 3(12):2927.
[11] Das, R., *Expansive self-homeomorphisms on G-spaces*, Period. Math. Hungar. 31 (1995), no. 2, 123–130.
[12] Dai, X., Geng, X., Zhou, Z., *Some relations between Hausdorff-dimensions and entropies*, Sci. China Ser. A 41 (1998), no. 10, 1068–1075.]
[13] Dinaburg, E., I., *A correlation between topological entropy and metric entropy*, Dokl. Akad. Nauk SSSR 190 (1970), 19–22.
[14] Eisenberg, M., *Expansive transformation semigroups of endomorphisms*, Fund. Math. 59 (1966), 313–321.
[15] Fisher, T., Equilibrium states for robustly transitive systems, talk at *Dynamical Systems in Montevideo*, 2012.
[16] Fathi, A., *Expansiveness, hyperbolicity and Hausdorff dimension*, Comm. Math. Phys. 126 (1989), no. 2, 249–262.
[17] Fedorenko, V., V., Smital, J., Maps of the interval Ljapunov stable on the set of nonwandering points, *Acta Math. Univ. Comenian.* (N.S.) 60 (1991), no. 1, 11–14.
[18] Goodman, T., N., T., *Maximal measures for expansive homeomorphisms*, J. London Math. Soc. (2) 5 (1972), 439–444.
[19] Hasselblatt, B., Katok, A., *Introduction to the modern theory of dynamical systems. With a supplementary chapter by Katok and Leonardo Mendoza*, Encyclopedia of Mathematics and its Applications, 54. Cambridge University Press, Cambridge, 1995.
[20] Hiraide, K., *Expansive homeomorphisms of compact surfaces are pseudo-Anosov*, Proc. Japan Acad. Ser. A Math. Sci. 63 (1987), no. 9, 337–338.
[21] Jimenez Lopez, V., Large chaos in smooth functions of zero topological entropy, *Bull. Austral. Math. Soc.* 46 (1992), no. 2, 271–285.
[22] Jakobsen, J., F., Utz, W., R., *The non-existence of expansive homeomorphisms on a closed 2-cell*, Pacific J. Math. 10 (1960), 1319–1321.
[23] Kato, H., *Chaotic continua of (continuum-wise) expansive homeomorphisms and chaos in the sense of Li and Yorke*, Fund. Math. 145 (1994), no. 3, 261–279.
[24] Kato, H., *Expansive homeomorphisms on surfaces with holes*, Special volume in memory of Kiiti Morita. Topology Appl. 82 (1998), no. 1-3, 267–277.
[25] Katok, A., *Lyapunov exponents, entropy and periodic orbits for diffeomorphisms*, Inst. Hautes Études Sci. Publ. Math. No. 51 (1980), 137–173.
[26] Kifer, Y., *Large deviations, averaging and periodic orbits of dynamical systems*, Comm. Math. Phys. 162 (1994), no. 1, 33–46.
[27] Kuchta, M., Characterization of chaos for continuous maps of the circle, *Comment. Math. Univ. Carolin.* 31 (1990), no. 2, 383–390.
[28] Lewowicz, J., *Expansive homeomorphisms of surfaces*, Bol. Soc. Brasil. Mat. (N.S.) 20 (1989), no. 1, 113–133.
[29] Ledrappier, F., Young, L.-S., *The metric entropy of diffeomorphisms. II. Relations between entropy, exponents and dimension*, Ann. of Math. (2) 122 (1985), no. 3, 540–574.
[30] Li, T.,Y., Yorke, J., A., Period three implies chaos, *Amer. Math. Monthly* 82 (1975), no. 10, 985–992.
[31] Mañé, R., *Ergodic theory and differentiable dynamics*. Translated from the Portuguese by Silvio Levy. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], 8. Springer-Verlag, Berlin, 1987.
[32] Mañé, R., *Expansive homeomorphisms and topological dimension*, Trans. Amer. Math. Soc. 252 (1979), 313–319.
[33] Mañé, R., *Expansive diffeomorphisms*, Dynamical systems—Warwick 1974 (Proc. Sympos. Appl. Topology and Dynamical Systems, Univ. Warwick, Coventry, 1973/1974; presented to E. C. Zeeman on his fiftieth birthday), pp. 162–174. Lecture Notes in Math., Vol. 468, Springer, Berlin, 1975.
[34] Morales, C., A., Partition's sensitivity for measurable maps, Preprint 2011.
[35] O'Brien, T., *Expansive homeomorphisms on compact manifolds*, Proc. Amer. Math. Soc. 24 (1970), 767–771.
[36] Parry, W., *Aperiodic transformations and generators*, J. London Math. Soc. 43 (1968), 191–194.

[37] Parthasarathy, K., R., Ranga R., R., Varadhan, S., R., S., *On the category of indecomposable distributions on topological groups*, Trans. Amer. Math. Soc. 102 (1962), 200–217.
[38] Reddy, W., *The existence of expansive homeomorphisms on manifolds*, Duke Math. J. 32 (1965), 627–632.
[39] Reddy, W., *Pointwise expansion homeomorphisms*, J. London Math. Soc. (2) 2 (1970), 232–236.
[40] Reddy, W., Robertson, L., *Sources, sinks and saddles for expansive homeomorphisms with canonical coordinates*, Rocky Mountain J. Math. 17 (1987), no. 4, 673–681.
[41] Sakai, K., *Hyperbolic metrics of expansive homeomorphisms*, Topology Appl. 63 (1995), no. 3, 263–266.
[42] Sears, M., *Expansive self-homeomorphisms of the Cantor set*, Math. Systems Theory 6 (1972), 129–132.
[43] Sindelarova, P., A counterexample to a statement concerning Lyapunov stability *Acta Math. Univ. Comenianae* 70 (2001), 265–268.
[44] Smital, J., Chaotic functions with zero topological entropy, *Trans. Amer. Math. Soc.* 297 (1986), no. 1, 269–282.
[45] Takens, F.,Verbitski, E., *Multifractal analysis of local entropies for expansive homeomorphisms with specification*, Comm. Math. Phys. 203 (1999), no. 3, 593–612.
[46] Utz, W., R., *Unstable homeomorphisms*, Proc. Amer. Math. Soc. 1 (1950), 769–774.
[47] Vietez, J., L., *Three-dimensional expansive homeomorphisms*, Dynamical systems (Santiago, 1990), 299–323, Pitman Res. Notes Math. Ser., 285, Longman Sci. Tech., Harlow, 1993.
[48] Vieitez, J., L., *Expansive homeomorphisms and hyperbolic diffeomorphisms on 3-manifolds*, Ergodic Theory Dynam. Systems 16 (1996), no. 3, 591–622.
[49] Walters, P., *Ergodic theory—-introductory lectures*, Lecture Notes in Mathematics, Vol. 458. Springer-Verlag, Berlin-New York, 1975.
[50] Williams, R., *Some theorems on expansive homeomorphisms*, Amer. Math. Monthly 73 (1966), 854–856.
[51] Williams, R., *On expansive homeomorphisms*, Amer. Math. Monthly 76 (1969), 176–178.

Instituto de Matemática, Universidade Federal do Rio de Janeiro, P. O. Box 68530, 21945-970 Rio de Janeiro, Brazil.

*E-mail address*: arbieto@im.ufrj.br, morales@impa.br

# HYPER-EXPANSIVE HOMEOMORPHISMS

ALFONSO ARTIGUE

Abstract. A homeomorphism on a compact metric space is said hyper-expansive if every pair of different compact sets are separated by the homeomorphism in the Hausdorff metric. We characterize such dynamics as those with a finite number of orbits and whose non-wandering set is the union of the repelling and the attracting periodic orbits. We also give a characterization of compact metric spaces admiting hyper-expansive homeomorphisms.

## 1. Introduction

It is an important goal in topological dynamics to understand the global behavior of expansive homeomorphisms. In light of the work of J. Lewowicz and K. Hiraide of expansive homeomorphisms of surfaces (see [2, 5]) it seems that the key point is to determine the topological properties of stable sets. On manifolds of arbitrary dimension it is proved in the above mentioned papers that the topological dimension of stable sets is positive (i.e., contains non-trivial connected sets) and smaller than the dimension of the manifold (i.e., there are no stable points). But it seems that more technology is needed in order to understand the topological structure of such sets.

A proof of the cited results is based in the following tool. Take an arc or a continuum on the manifold, iterate it with the homeomorphism and consider an accumulation point in the Hausdorff metric on compact subsets. So, it seems that is of interest to consider the dynamics of sets instead of single points. This fact was noticed by H. Kato, who introduced the notion of continuum-wise expansiveness. Consider $f\colon X \to X$ a homeomorphism on a compact metric space. We say that $f$ is *continuum-wise expansive* if there is $\delta > 0$ such that if $C \subset X$ is a compact connected set (i.e., continuum) such that $\operatorname{diam}(f^n C) < \delta$ for all $n \in \mathbb{Z}$ then $C$ is a singleton.

One can try the following definition: a homeomorphism is *compact-wise expansive* if there is $\delta > 0$ such that if $C \subset X$ is compact and $\operatorname{diam} f^n(C) < \delta$ for all $n \in \mathbb{Z}$ then $C$ is a singleton. But it is easy to see that this is equivalent with expansiveness. One just has to notice that $\operatorname{diam}(\{x, y\}) = \operatorname{dist}(x, y)$ and that every non-trivial compact set has at least two different points. Expansiveness is also equivalent with what could be called *set-wise expansiveness* (with analogous definition and proof). It is interesting to remark that *open-wise expansiveness* is some kind of sensitive dependence on initial conditions.

Given a compact metric space we consider the space of all compact subsets $A \subset X$. That space is called the *hyperspace* of $X$ and is denoted as $2^X$. The topology of $2^X$ is defined by the *Hausdorff metric* $\operatorname{dist}_H$ defined as

$$\operatorname{dist}_H(A, B) = \inf\{\varepsilon > 0 : A \subset B_\varepsilon(B) \text{ and } B \subset B_\varepsilon(A)\}$$

for all $A, B \in 2^X$. As usual $B_\varepsilon(A)$ denotes the set $\cup_{x \in A} B_\varepsilon(x)$ where $B_\varepsilon(x) = \{y \in X : \operatorname{dist}(x, y) < \varepsilon\}$. The hyperspace has very nice properties. For example, it is known that

*Date*: July 9, 2013.

$2^X$ inherits the compactness of $X$. Also, if $X$ is connected then $2^X$ is arc-wise connected (see [8]). So, it is natural to extend the action of $f$ to $2^X$, simply as $\hat{f}\colon 2^X \to 2^X$ defined by $\hat{f}(A) = f(A)$. It gives a homeomorphism as can be easily verified. Some relationships are known between the dynamics of $f$ and $\hat{f}$. For example, if $f$ has positive topological entropy then $\hat{f}$ has infinite topological entropy (see Proposition 6 in [1]).

The purpose of the present note is to study the expansiveness of the induced map $\hat{f}$, that is what we call *hyper-expansiveness* of $f$. Notice that hyper-expansiveness is a stronger condition than expansiveness. The following facts are known:

- if a compact metric space admits an expansive homeomorphism then its topological dimension is finite (see [6]) and
- if $\dim_{top} X > 0$ then $\dim_{top} 2^X = \infty$ (this fact was first proved in [7], see also [8] Theorem 1.95).

Hence, if $2^X$ admits an expansive homeomorphism then $\dim_{top} X = 0$.

It is known that expansiveness does not imply hyper-expansiveness. Indeed, in [1] it is noticed that the shift map is not hyper-expansive (this can be deduced from the fact that the shift map has infinite periodic points) while the shift map itself is expansive. Those remarks on hyper-expansiveness were rediscovered in [9] (Proposition 2.23 and Example 2.24). We give a simple characterization of hyper-expansiveness, statements and proofs are in the following Section.

Another important problem in topological dynamics is to determine what spaces admit expansive homeomorphisms. In [4] this problem is solved for countable compact spaces. As we will see, spaces admiting a hyper-expansive homeomorphism are countable. In this note we also give a characterization of compact spaces admiting hyper-expansive homeomorphisms.

In terms of the hyperspace, expansiveness can be characterized as follows. Let $F_1 = \{\{x\} : x \in X\} \subset 2^X$ be the space of singletons. Notice that $F_1$ is $\hat{f}$-invariant, in fact $\hat{f}\colon F_1 \to F_1$ is conjugated with $f\colon X \to X$. By definition we have that $f$ is expansive if and only if $F_1$ is an isolated set for $\hat{f}$, i.e., there is an open set $U$ of $2^X$ such that $F_1 = \cap_{n\in\mathbb{Z}} \hat{f}^n U$.

I would like to thank Damián Ferraro, Mario González and Ignacio Monteverde for useful conversations on these topics, José Vieitez for his corrections in the preliminary version of the note and the referee for his or her remarks.

## 2. Hyper-expansiveness

Let $(X, \mathrm{dist})$ be a compact metric space.

**Definition 2.1.** A homeomorphism $f\colon X \to X$ on a compact metric space is *hyper-expansive* if $\hat{f}\colon 2^X \to 2^X$ is expansive, that is, there is $\delta > 0$ such that if $\mathrm{dist}_H(f^n A, f^n B) < \delta$ for all $n \in \mathbb{Z}$, with $A$ and $B$ compact subsets of $X$, then $A = B$.

We need some definitions. Given a point $p \in X$ we say that it is *(Lyapunov) stable* if for all $\varepsilon > 0$ there is $\delta > 0$ such that if $\mathrm{dist}(x, p) < \delta$ then $\mathrm{dist}(f^n x, f^n p) < \varepsilon$ for all $n \geq 0$. A point $p$ is said to be *unstable* if it is stable for $f^{-1}$. We say that $p$ is *asymptotically stable* if it is stable and there is $\gamma > 0$ such that if $\mathrm{dist}(x, p) < \gamma$ then $\mathrm{dist}(f^n x, f^n p) \to 0$ as $n \to \infty$. If $p$ is an asymptotically stable periodic orbit then the orbit of $p$ is said to be an *attractor*. A *repeller* is an attractor for $f^{-1}$. Notice that isolated periodic points are stable and unstable by definition. Let us denote

- $\Omega f$ the set of non-wandering points, i.e., $x \in \Omega f$ if for all $\varepsilon > 0$ there is $n > 0$ such that $f^n(B_\varepsilon x) \cap B_\varepsilon x \neq \emptyset$,

- $\mathrm{Per}_r$ the set of repeller periodic points and $\mathrm{Per}_a$ the set of attracting periodic points.

Now we can state the main result of this note.

**Theorem 2.2.** *A homeomorphism $f\colon X \to X$ is hyper-expansive if and only if $f$ has a finite number of orbits and $\Omega f = Per_r \cup Per_a$.*

**Remark 2.3.** *It is easy to see that every expansive homeomorphism has a finite number of fixed points. Also, every compact $f$-invariant set $K \subset X$ (i.e., $f(K) = K$) is a fixed point of $\hat{f}$. So, if $f$ is hyper-expansive then $f$ has a finite number of compact invariant sets (in particular, it has finitely many periodic points).*

A compact $f$-invariant set $K \subset X$ is said to be *minimal* if for all $x \in K$ the orbit $\{x, fx, \dots, f^n x, \dots\}$ is dense in $K$.

**Lemma 2.4.** *If $f\colon X \to X$ is hyper-expansive and $K \subset X$ is minimal then $K$ is finite (i.e., a periodic orbit).*

*Proof.* Minimality implies that for all $\varepsilon > 0$ there is $n \geq 0$ such that for all $x \in X$ the set $O_n x = \{x, fx, \dots, f^n x\}$ is $\varepsilon$-dense in $K$ (i.e., for all $y \in K$ there is $j \in \{0, 1, \dots, n\}$ such that $\mathrm{dist}(y, f^j x) < \varepsilon$). Therefore, $f^j(O_n x)$ is $\varepsilon$-dense for all $j \in \mathbb{Z}$ because $f^j(O_n x) = O_n(f^j x)$. If $\varepsilon$ is an expansive constant for $\hat{f}$ then $\mathrm{dist}_H(f^j(O_n x), f^j(K)) < \varepsilon$ for all $j \in \mathbb{Z}$. Then $K = O_n x$ and $K$ is finite. □

**Remark 2.5.** *In the previous proof the expansiveness was contradicted with two sets $K_1 \subset K_2$. Notice that $\mathrm{dist}_H(A, B) \geq \mathrm{dist}_H(A, B \cup A)$, so $f$ is hyper-expansive if and only if there is $\delta > 0$ such that if $A \subset B$, $A, B \in 2^X$ and $\mathrm{dist}_H(\hat{f}^n A, \hat{f}^n B) < \delta$ for all $n \in \mathbb{Z}$ then $A = B$.*

We have that if $f$ is hyper-expansive then $f$ has a finite number of periodic points. Eventually taking a power of $f$ we can suppose that every periodic point is a fixed point. Recall that if a homeomorphism is expansive then its non-trivial powers are expansive too. In the following Lemma we will need the next well known result.

**Remark 2.6.** *If $f$ is expansive and $p$ is a stable (unstable) periodic point then $p$ is an attractor (repeller). It can be proved as follows. Without loss of generality we can suppose that $p$ is a fixed point. By contradiction suppose that $p$ is stable but it is not asymptotically stable. Let $\delta > 0$ be an expansive constant of $f$. Therefore, there is a point $q \in B_\delta(p)$ such that $f^n q \in B_\delta(p)$ for all $n \geq 0$ but the $\omega$-limit set of $q$ is not $\{p\}$. So $p$ and a point $p' \in \omega(q)$, $p' \neq p$, contradict the expansiveness of $f$.*

**Lemma 2.7.** *If $f$ is hyper-expansive then every fixed point of $f$ is an attractor or a repeller.*

*Proof.* By contradiction suppose that $p$ is a fixed point of $f$ that is neither attractor nor repeller. Since $p$ is not an attractor, $p$ is not stable (Remark 2.6). So, there is $\varepsilon > 0$ and a sequence $x_n$ such that $x_n \to p$ as $n \to \infty$ and for some $k_n > 0$, $f^{k_n}(x_n) \notin B_\varepsilon(p)$. Suppose that for all $k < k_n$, $f^k(x_n) \in B_\varepsilon(p)$. Assume that $a_n = f^{k_n - 1}(x_n)$ converges to $a \in \operatorname{clos} B_\varepsilon(p)$. It is easy to see that $f^j(a) \to p$ as $j \to -\infty$ and $a \neq p$.

Similarly, using that $p$ is not unstable, one can prove that there is $b \neq p$ such that $f^j(b) \to p$ as $j \to \infty$. Let $\delta > 0$ be an expansive constant for $\hat{f}$. Take $n \geq 0$ such that $f^m(b), f^{-m}(a) \in B_\delta(p)$ for all $m \geq n$. Let $A = \{f^n(b), f^{-n}(a)\}$ and $B = A \cup \{p\}$. So, $A \neq B$ and $\mathrm{dist}_H(f^n A, f^n B) < \delta$ for all $n \in \mathbb{Z}$. That contradicts the expansiveness of $\hat{f}$. □

Now we prove our main result.

*Proof.* (of Theorem 2.2) *Direct.* Suppose that $f$ is hyper-expansive. We have proved that there is a finite number of periodic points. So, eventually taking a power of $f$ we can suppose that every periodic point is in fact a fixed point. If there are only fixed points, there is nothing to prove ($X$ is finite). So, suppose that $x \in X$ is not a fixed point. Consider the $\omega$-limit set $\omega(x)$. It is a compact invariant set, therefore it contains a minimal set, say $K$. We have proved that every minimal set is a periodic orbit, so, it is a fixed point $K = \{p\}$. It is easy to see that $\omega(x) = \{p\}$, since $p$ must be an attractor. In particular $x$ is a wandering point. Then, we have proved that $\Omega(f) = \mathrm{Per}_a \cup \mathrm{Per}_r$.

Now we will prove that there is a finite number of orbits. It is easy to see that for all $\varepsilon > 0$ there is $N \geq 0$ such that if $x \notin B_\varepsilon(\Omega(f))$ then $f^j x, f^k x \in B_\varepsilon(\Omega(f))$ if $j \leq -N$ and $k \geq N$.

If $f$ has an infinite number of orbits and $\varepsilon > 0$ is smaller than an expansive constant for $\hat{f}$, then $X \setminus B_\varepsilon(\Omega(f))$ is a compact infinite set. So, there are $x, y \notin B_\varepsilon(\Omega(f))$ and $p, q \in \Omega(f)$ such that $\omega(x) = \omega(y) = \{p\}$ and $\alpha(x) = \alpha(y) = \{q\}$. Then, if $\mathrm{dist}(x, y)$ is small, this two points contradicts the expansiveness of $f$ (and hyper-expansiveness too). This contradiction proves that there is a finite number of orbits.

*Converse.* Again, eventually taking a power of $f$, we can assume that every periodic point of $f$ is a fixed point. Let $\delta_1 > 0$ be such that $\cap_{n \geq 0} f^n(B_{\delta_1}(\mathrm{Per}_a)) = \mathrm{Per}_a$ and $\cap_{n \leq 0} f^n(B_{\delta_1}(\mathrm{Per}_r)) = \mathrm{Per}_r$. Take $x_1, \ldots, x_n$ one point of each wandering orbit of $f$. Let $\delta_2 > 0$ be such that $B_{\delta_2}(x_i) = \{x_i\}$ for all $i = 1, \ldots, n$. We will show that $\delta = \min\{\delta_1, \delta_2\}$ is an expansive constant for $\hat{f}$. Let $A, B$ be two compact sets such that $\mathrm{dist}_H(f^n A, f^n B) < \delta$ for all $n \in \mathbb{Z}$. If there is a wandering point $x$ such that $x \in A \setminus B$ then there is $k \in \mathbb{Z}$ and $i \in \{1, \ldots, n\}$ such that $f^k x = x_i$. So, $\mathrm{dist}_H(f^k A, f^k B) > \delta_2$. This contradiction proves that the wandering points of $A$ and $B$ coincide. If $A \neq B$ then there is a fixed point $p \in A \setminus B$ (similarly for $p \in B \setminus A$). Without loss of generality suppose that $p$ is a repeller. Since $p \notin B$ then there is $\varepsilon > 0$ such that $B_\varepsilon(p) \cap B = \emptyset$. Take $n$ such that $B_{\delta_1}(p) \cap f^n B = \emptyset$. Since $p \in f^n A$ for all $n \in \mathbb{Z}$, we have that $\mathrm{dist}_H(f^n A, f^n B) > \delta_1$, wich is a contradiction. So $f$ is hyper-expansive. □

A simple consequence of the previous result is that if $X$ admits a hyper-expansive homeomorphism then $X$ is countable. As we will see, the converse is not true. Let

$$\mathrm{Iso}(X) = \{x \in X : \text{there is } \varepsilon > 0 \text{ such that } B_\varepsilon(x) \cap X = \{x\}\}$$

and

$$\mathrm{Lim}(X) = X \setminus \mathrm{Iso}(X).$$

The cardinality of a set $A$ is denoted as $|A|$.

**Theorem 2.8.** *A compact metric space $X$ admits a hyper-expansive homeomorphism if and only if $2 \leq |\mathrm{Lim}(X)| < \infty$ or $\mathrm{Lim}(X) = \emptyset$ (i.e., $X$ is finite).*

*Proof.* By Theorem 2.2 we have that $\mathrm{Lim}(X) \subset \Omega(f)$ that is because wandering points must be isolated. So, $\mathrm{Lim}(X)$ is finite. If $X$ is infinite, there must be at least one attractor and one repeller, so $\mathrm{Lim}(X) \geq 2$.

In order to prove the converse notice that if the set of limit points is finite then $X$ is countable. Consider an infinite countable space $X$ (the finite case is trivial). Since every infinite continuum is uncontable, we have that $\dim_{top}(X) = 0$. It is known that if $\dim_{top}(X) \leq n$ then $X$ is homeomorphic to a compact subset of $\mathbb{R}^{2n+1}$, see Theorem V2 in [3]. So, without loss of generality, we can assume that $X \subset \mathbb{R}$. Let $p_1 < \cdots < p_n \in X$,

$n \geq 2$, be the limit points of $X$. We can also suppose that $X \subset [p_1, p_n]$ and for all $\varepsilon > 0$ we have that

- $X \cap (p_j, p_j + \varepsilon) \neq \emptyset$ for all $j = 1, \ldots, n-1$ and
- $X \cap (p_j - \varepsilon, p_j) \neq \emptyset$ for all $j = 2, \ldots, n$

Define $I_j = X \cap (p_j, p_{j+1})$ for $j = 1 \ldots, n-1$. Now we define $f\colon X \to X$ as follows:

- $f(p_j) = p_j$ for all $j = 1, \ldots, n$,
- if $x \in I_j$ and $j$ is odd then $f(x)$ is the first point of $X$ at the right of $x$ and
- if $x \in I_j$ and $j$ is even then $f(x)$ is the first point of $X$ at the left of $x$.

In this way $p_j$ is a repeller fixed point if $j$ is odd and it is an attractor if $j$ is even. So, by Theorem 2.2 we have that $f$ is hyper-expansive. □

Since hyper-expansiveness is a very strong condition, we have that *most* homeomorphisms satisfy the following result.

**Corollary 2.9.** *If $f\colon X \to X$ is a homeomorphism of a compact metric space $X$ and $|\operatorname{Lim}(X)| = \infty$ then for all $\varepsilon > 0$ there are two different compact sets $A, B \subset X$ such that*

$$\operatorname{dist}_H(f^n A, f^n B) < \varepsilon, \text{ for all } n \in \mathbb{Z}.$$

It is a simple consequence of our previous result. It holds for example if $X$ is a manifold of positive dimension, a non-trivial connected space or a Cantor set.

Let us now give some examples and final remarks.

**Example 2.10.** *Let $X = \{0\} \cup \{1/n : n \in \mathbb{N}\}$. Since $X$ has just one limit point we have that $X$ does not admit hyper-expansive homeomorphisms, but it is easy to see that it admits an expansive one.*

Countable compact spaces admiting expansive homeomorphisms can be characterized as follows. Recall that $\operatorname{Lim}^{\lambda+1}(X) = \operatorname{Lim}(\operatorname{Lim}^{\lambda}(X))$, $\operatorname{Lim}^1(X) = \operatorname{Lim}(X)$ and

$$\operatorname{Lim}^{\lambda}(X) = \bigcap_{\alpha<\lambda} \operatorname{Lim}^{\alpha}(X)$$

for every limit ordinal number $\lambda$. The *limit degree* of $X$ is the ordinal number $d(X) = \lambda$ if $\operatorname{Lim}^{\lambda}(X) \neq \emptyset$ and $\operatorname{Lim}^{\lambda+1}(X) = \emptyset$. In [4] (Theorem 2.2) it is shown that a countable compact space $X$ admits an expansive homeomorphism if and only if $d(X)$ is not a limit ordinal number.

**Remark 2.11.** *Applying Theorem 2.8 we have that $X$ admits a hyper-expansive homeomorphism if and only if $d(X) \leq 1$ and $|\operatorname{Lim}(X)| \neq 1$.*

It seems to be of interest to provide an example of a countable compact space do not admiting expansive homeomorphisms.

**Example 2.12.** *Given $A \subset \mathbb{R}$ we say that $(a, b) \in A \times A$ is an* adjacent pair *if there are no points of $A$ in the open interval $(a, b)$. The set of adjacent pairs is denoted as*

$$\operatorname{Adj}(A) = \{(a, b) \in A \times A : a < b, (a, b) \cap A = \emptyset\}.$$

*Let $A_0 = \{0\} \cup \{1/n : n \in \mathbb{N}\}$ and*

$$A_{n+1} = A_n \cup \bigcup_{(a,b) \in \operatorname{Adj}(A_n \cap [0,1/n])} \{a + (b-a)/m : m \in \mathbb{N}\}.$$

*Define $X = \cup_{n=0}^{\infty} A_n$. It is easy to see that it is a compact set and it is countable by construction. Notice that $d(X)$ is the first infinite ordinal number and therefore it is*

*a limit ordinal number. To see that it does not admit an expansive homeomorphism notice that if* $f\colon X \to X$ *is a homeomorphism then* $\mathrm{Lim}^{\lambda}(X)$ *is an* $f$*-invariant set for all ordinal number* $\lambda$*. Now notice that for all* $\varepsilon > 0$ *there is a finite ordinal number* $\lambda$ *such that* $\mathrm{Lim}^{\lambda}(X) \subset [0, \varepsilon]$ *and* $\mathrm{Lim}^{\lambda}(X)$ *is an infinite set. Therefore, every pair of different points* $x, y \in \mathrm{Lim}^{\lambda}(X)$ *contradict the* $\varepsilon$*-expansiveness of* $f$*. Since* $\varepsilon$ *is arbitrary we have that* $X$ *does not admit expansive homeomorphisms.*

## References


[1] W. Bauer and K. Sigmund, *Topological dynamics of transformations induced on the space of probability measures*, Monatsh Math, **79**, 81–92, (1975).

[2] K. Hiraide, *Expansive homeomorphisms of compact surfaces are pseudo-Anosov*, Osaka J. Math., **27** (1990), 117–162.

[3] W. Hurewicz and H. Wallman, *Dimension Theory*, Princeton Univ. Press, 1948.

[4] H. Kato and J. Park, *Expansive homeomorphisms of countable compacta*, Top. and its App., **95**, (1999), 207–216.

[5] J. Lewowicz, *Expansive homeomorphisms of surfaces*, Bol. Soc. Bras. Mat., **20**, 113–133, (1989).

[6] R. Mañé, *Expansive homeomorphisms and topological dimension*, Trans. of the AMS, **252**, 313–319, (1979).

[7] S. Mazurkiewicz, *Sur le type de dimension de l'hyperespace d'un continu*, C. R. Soc. Sc. Varsovie, **24**, 191–192, (1931).

[8] S. Nadler Jr., Hyperspaces of Sets, Marcel Dekker Inc. New York and Basel (1978).

[9] P. Sharma and A. Nagar, *Topological dynamics on hyperspaces*, Applied general topology, **11**, 1–19, (2010).



*E-mail address*: aartigue@fing.edu.uy

Departamento de Matemática y Estadística del Litoral,
Regional Norte, Universidad de la República,
Salto-Uruguay

# INVARIANT MEASURES FOR RANDOM TRANSFORMATIONS EXPANDING ON AVERAGE

JOCHEN BRÖCKER AND GIANLUIGI DEL MAGNO

Abstract. The thermodynamic formalism for random transformations expanding on average is revisited. We consider the associated random transfer operators with Hölder continuous random potentials, and prove a random version of the Ruelle–Perron–Frobenius Theorem. This result allows us to construct random invariant measures for the transformations considered. These measures are ergodic, and enjoy fiberwise exponential decay of correlations. As a method of proof, we construct a family of cones of positive functions for which the transfer operator is a strict contraction. Application of a random fixed point theorem then yields a maximal random eigenvalue and eigenvector of the transfer operator.

September 28, 2013

## 1. Introduction

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with the $\sigma$-algebra $\mathcal{F}$ endowed with an automorphism $T : \Omega \to \Omega$ preserving the probability $\mathbb{P}$. Furthermore, let $M$ be a separable and complete metric space with its Borel algebra $\mathcal{B}$. A discrete time random dynamical system acting on $(M, \mathcal{B})$ and driven by $(\Omega, \mathcal{F}, \mathbb{P}, T)$ is a family $\{\phi_n\}_{n\in\mathbb{N}}$ of measurable transformations $\phi_n : \Omega \times M \to M$ enjoying the cocycle property

$$\phi_{n+k}(\omega, x) = \phi_k(T^n\omega, \phi_n(\omega, x))$$

for $\omega \in \Omega$, $x \in M$ and $k, n \in \mathbb{N}$. In this paper, we consider random dynamical systems generated by a single measurable transformation $\phi : \Omega \times M \to M$. This means that

$$\phi_n(\omega, \cdot) = \phi(T^{n-1}\omega, \cdot) \circ \cdots \circ \phi(\omega, \cdot)$$

for $\omega \in \Omega$ and $n \in \mathbb{N}$. Such a system will be denoted by $\phi$ rather than by $\{\phi_n\}_{n\in\mathbb{N}}$.

A measure $\mu$ on the product space $\Omega \times M$ is called $\phi$-invariant if $\mu$ is invariant for the transformation $\Omega \times M \ni (\omega, x) \mapsto (T\omega, \phi(\omega, x))$, and the marginal of $\mu$ on $\Omega$ is equal to $\mathbb{P}$. If the family $\{\mu_\omega\}_{\omega\in\Omega}$ of measures on $M$ denotes the disintegration of $\mu$ with respect to $\mathbb{P}$, then the $\phi$-invariance of $\mu$ is equivalent to the property that $\phi(\omega, \cdot)_*\mu_\omega = \mu_{T\omega}$ for $\mathbb{P}$-a.e. $\omega \in \Omega$.

When $\phi(\omega, \cdot) : M \to M$ is continuos for each $\omega \in \Omega$, a standard approach for constructing $\phi$-invariant measures with interesting properties consists in studying the family $\{\mathcal{L}_{\phi,\gamma}(\omega)\}_{\omega\in\Omega}$ of operators on the space of continuous functions $C(M)$ defined

2000 *Mathematics Subject Classification.* Primary 37H99, 37D35; Secondary 47B80.

*Key words and phrases.* Random Dynamical Systems, Thermodynamic Formalism, Transfer Operators.

The second author was supported by Fundação para a Ciência e a Tecnologia through the Program POCI 2010 and the Project 'Randomness in Deterministic Dynamical Systems and Applications' (PTDC-MAT-105448-2008).

by

$$(\mathcal{L}_{\phi,\gamma}(\omega)\varphi)(x) = \sum_{y\in\phi(\omega,\cdot)^{-1}(x)} e^{\gamma(\omega,y)}\varphi(y)$$

for $\omega \in \Omega$, $x \in M$, $\varphi \in C(M)$ and a measurable function $\gamma : \Omega \times M \to (0, +\infty)$ such that $\gamma(\omega, \cdot) \in C(M)$ for each $\omega \in \Omega$.

In this paper, we further assume that $M$ is a connected compact Riemannian manifold, and that $\phi(\omega, \cdot) : M \to M$ is surjective and a local diffeomorphism expanding on average. Precise definitions are postponed until Section 2. Under these hypotheses, we prove a Ruelle–Perron–Frobenius Theorem (RPFT) for the random operator $\mathcal{L}_{\phi,\gamma}$. For a detailed account on RPFTs, we refer the reader to the excellent monograph [2]. Roughly speaking, our RPFT states that there exists a triple $(\Lambda, h, \nu)$ consisting of a measurable function $\Lambda : \Omega \to (0, +\infty)$, a measurable function $h : \Omega \times M \to (0, +\infty)$, and a family $\nu = \{\nu_\omega\}_{\omega\in\Omega}$ of measures on $M$ such that for $\mathbb{P}$-a.e. $\omega \in \Omega$,

$$\mathcal{L}_{\phi,\gamma}(\omega)h(\omega, \cdot) = \Lambda(\omega)h(T\omega, \cdot),$$
$$\mathcal{L}^*_{\phi,\gamma}(\omega)\nu_{T\omega} = \Lambda(\omega)\nu_\omega,$$

where $\mathcal{L}^*_{\phi,\gamma}(\omega)$ is the dual operator of $\mathcal{L}_{\phi,\gamma}(\omega)$ acting on Borel measures on $M$. This results can be considered as the random counterpart of the classical Perron–Frobenious Theorem, with $\Lambda$ and $h$ playing the roles of the maximal eigenvalue and eigenvector of the random operator $\mathcal{L}_{\phi,\gamma}$, respectively. As the Perron–Frobenius Theorem allows to prove the existence of stationary measures for finite state Markov chains, so the RPFT will allow us to prove that the measure $\mu$ whose disintegration on $\mathbb{P}$ is given by the measures $\mu_\omega = h(\omega, \cdot)\nu_\omega$ is $\phi$-invariant. We then show that $\mu$ enjoys fiberwise exponential decay of correlations, and is ergodic for the transformation $(\omega, x) \mapsto (T\omega, \phi(\omega, x))$.

If the potential is given by $\gamma(\omega, x) = -\log|\det D_x\phi(\omega, \cdot)|$, then the transfer operator $\mathcal{L}_{\phi,\gamma}(\omega)$ is the usual Frobenius–Perron operator associated to the random map $\phi(\omega, \cdot)$. In this case, we further show that the measures $\mu_\omega$ are all absolutely continuous with respect to the Riemannian volume $m$.

The proof of our RPFT consists of three major steps. Firstly, we show that under our assumptions on the system $\phi$, the operators $\mathcal{L}_{\phi,\gamma}(\omega)$ preserve certain cones consisting of positive Hölder continuous functions. By equipping these cones with Hilbert's projective metric, we show that the operators $\mathcal{L}_{\phi,\gamma}(\omega)$ are contractions. It was shown first by Birkhoff that a linear map preserving a cone equipped with Hilbert's projective metric is a contraction [5, 7]. The existence of the function $h$ is then proved employing a Fixed Point Theorem for random Lipschiz maps contracting on average, which is a generalization of a theorem of Bougerol [4]. Finally, closely following an argument of Birkhoff (see also [10]), we prove the existence of the measures $\nu_\omega$.

Results similar to those presented in this paper were already obtained by several authors. Ferrero and Schmitt considered the case of subshifts of finite type with random potential [10]. Their results were then extended by Bogenschütz and Gundlach to the case of random subshifts [3]. Finally, Kifer examined the symbolic transformations of Bogenschütz and Gundlach, and the case of random maps expanding on average [13] (see also [12, 11]), the latter being essentially the subject of the present paper. Thus, our results are to some extent not original, and a comparison with the approach and results of Kifer is in order.

Kifer obtains the RPFT under essentially the same assumptions as in the present paper, except that he further requires the function $\omega \mapsto \sup_{x\in M}|\gamma(\omega, x)|$ to be $\mathbb{P}$-integrable. This condition is required to prove that the resulting random invariant

measures satisfy the celebrated variational principle, something we do not consider here. On the other hand, we derive the RPFT by proving that, asymptotically as $n \to +\infty$, the operator

$$\frac{1}{\Lambda(T^n\omega)\cdots\Lambda(\omega)}\mathcal{L}_{\phi,\gamma}(T^n\omega)\circ\cdots\circ\mathcal{L}_{\phi,\gamma}(\omega)$$

behaves like a projector onto the 1-dimensional subspace of $C(M)$ generated by $h(T^n\omega,\cdot)$ for $\mathbb{P}$-a.e. $\omega\in\Omega$. This implies, in particular, fiberwise exponential decay of correlations. This is stronger than the corresponding result by Kifer, which provides exponential decay of correlations only for a random subsequence with positive density [13].

As far as the difference between the proofs is concerned, we both use families of cones consisting of positive Hölder continuous functions, and Hilbert's projective metrics to prove the existence of the function $h$. The difference between the two proofs stands in the difference between the cones and their use. We construct a family of cones $\{\mathcal{C}(\omega)\}_{\omega\in\Omega}$ which is invariant under $\{\mathcal{L}_{\phi,\gamma}(\omega)\}_{\omega\in\Omega}$, that is,

$$\mathcal{L}_{\phi,\gamma}(\omega)\mathcal{C}(\omega)\subset\mathcal{C}(T\omega)\qquad\text{for }\mathbb{P}\text{-a.e. }\omega\in\Omega,$$

and furthermore, this inclusion is strict in the sense that the projective diameter of $\mathcal{L}_{\phi,\gamma}(\omega)\mathcal{C}(\omega)$ in $\mathcal{C}(T\omega)$ is bounded from above by a proper random variable on $\Omega$. Our approach also differs from Kifer's in the proof of the existence of the family of measures $\nu$. Whereas Kifer's uses a Krylov–Bogoliubov type argument, we exploit properties of the Hilbert projective metric. We follow [10], where an argument originally devised by Birkhoff for deterministic operators [5] is extended to the random case.

The paper is organized as follows. In Section 2, we return to the definition of the main mathematical objects studied in this paper: random dynamical systems, invariant measures and random transfer operators. In Section 3, we formulate our main results in Theorems 3.1-3.3. As already explained, these theorems are proved by using Hilbert's projective metric on cones and a Fixed Point Theorem for random maps. The latter is proved in Section 4. That section further contains other results on the uniqueness and measurability of the random fixed point, which might be of independent interest. In Section 5, we introduce Hilbert's projective metric on cones and recall Birkhoff's results on transformations preserving cones. We also give a detailed proof of the completeness of Hilbert's projective metric, which we were not able to find in the literature. Finally, Section 6 contains the proofs of Theorems 3.1-3.3. Because of their length, these proofs are split into several propositions.

## 2. Main definitions and statement of results

In this section, we define the objects investigated in this paper: random dynamical systems, invariant measures and random transfer operators.

2.1. **Random dynamical systems.** The notion of random dynamical system introduced below coincides essentially with the one of a discrete time random dynamical system generated by a random map $\phi$ as in Arnold's book [1]. The difference between the two definitions is that ours does not assume the map $\phi$ to be measurable. Later on, we will impose more restrictive conditions on our random dynamical systems.

Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space together with an automorphism $T:\Omega\to\Omega$ preserving the probability $\mathbb{P}$. We will always assume that the abstract dynamical system $(\Omega,\mathcal{F},\mathbb{P},T)$ is ergodic. Let $X$ be a set, and suppose that $\phi$ is a transformation from $\Omega\times X$ to $X$. At this stage, we do not impose any condition on $X$ and $\phi$.

The iterations $\phi_n : \Omega \times X \to X$ of the transformation $\phi$ are defined as follows

$$\phi_n(\omega, x) = \begin{cases} x & \text{if} \quad n = 0, \\ \phi(T^{n-1}\omega, \cdot) \circ \cdots \phi(T\omega, \cdot) \circ \phi(\omega, x) & \text{if} \quad n > 0. \end{cases}$$

If $\phi(\omega, \cdot) : X \to X$ is a bijection for every $\omega \in \Omega$, then we can also define $\phi_n$ for $n < 0$ as follows

$$\phi_n(\omega, \cdot) = (\phi_{-n}(T^n\omega, \cdot))^{-1} = \phi(T^n\omega, \cdot)^{-1} \circ \cdots \circ \phi(T^{-1}\omega, \cdot)^{-1} \qquad \text{for } \omega \in \Omega.$$

We refer to the case when $\phi(\omega, \cdot)$ is a bijection as the *invertible case*.

It follows immediately from their definition, that the maps $\{\phi_n\}$ satisfies the *cocycle property*:

$$\phi_{n+m}(\omega, x) = \phi_n(T^m\omega, \cdot) \circ \phi_m(\omega, x) \qquad \text{for } \omega \in \Omega, x \in X \text{ and } m, n \geq 0.$$

In the invertible case, the cocycle property holds for every $m, n \in \mathbb{Z}$.

The sequence of transformations $\{\phi_n\}_{n\geq 0}$ in the invertible case and the sequence of transformations $\{\phi_n\}_{n\in\mathbb{Z}}$ in the non-invertible case are both called a *random dynamical system on $X$ over $(\Omega, \mathcal{F}, \mathbb{P}, T)$ generated by $\phi$*. To avoid a cumbersome terminology, we will often avoid mentioning the set $X$ and the dynamical system $(\Omega, \mathcal{F}, \mathbb{P}, T)$, and simply refer to $\{\phi_n\}$ as the *random dynamical system $\phi$*.

A random dynamical system $\phi$ defines a skew-product $F : \Omega \times X \to \Omega \times X$ by setting

$$(\omega, x) \mapsto (T\omega, \phi(\omega, x)) \qquad \text{for } \omega \in \Omega \text{ and } x \in X.$$

We say that $F$ is the *skew-product associated to $\phi$*. Note that, conversely, every skew-product transformation defines a random dynamical system. Therefore, a skew-product provides an alternative way to describe a random dynamical system.

2.2. **Invariant measures.** Suppose now that $X$ is a compact metric space with its Borel $\sigma$-algebra $\mathcal{X}$, and that the transformation $\phi : (\Omega \times X, \mathcal{F} \otimes \mathcal{X}) \to (X, \mathcal{X})$ is measurable. Then, of course, the skew-product $F$ associated to $\phi$ is measurable as well.

Let $\pi_\Omega$ be the natural projection of $\Omega \times X$ onto $\Omega$. We say that a probability measure $\mu$ on $(\Omega \times X, \mathcal{F} \otimes \mathcal{X})$ is *$\phi$-invariant* if the following conditions are satisfied

(1) $\mu$ is invariant under $F$, i.e., $F_*\mu = \mu$,
(2) $\pi_{\Omega *}\mu = \mathbb{P}$.

We also say that a $\phi$-invariant probability $\mu$ is *ergodic* if $\mu$ is ergodic for the skew-product $F$.

Because $X$ is compact metric, it is separable. It follows that a probability $\mu$ on $(\Omega \times X, \mathcal{F} \otimes \mathcal{X})$ admits a disintegration $\{\mu_\omega\}$ with respect to $\mathbb{P}$ that is $\mathbb{P}$-almost surely unique [1, Proposition 1.4.3]. This means that up to a set of zero $\mathbb{P}$-measure, there exists a unique family $\{\mu_\omega\}$ of nonnegative set functions on $\mathcal{X}$ such that

(1) $\omega \mapsto \mu_\omega(A)$ is measurable for every $A \in \mathcal{X}$,
(2) $\mu_\omega$ is a probability on $(X, \mathcal{X})$ for $\mathbb{P}$-a.e. $\omega \in \Omega$,
(3) for every $B \in \mathcal{F} \otimes \mathcal{X}$, we have

$$\mu(B) = \int_\Omega \int_X I_B(\omega, x) \mathrm{d}\mu_\omega(x) \mathrm{d}\mathbb{P}(\omega),$$

where $I_B$ is the characteristic function of $B$.

Under our assumptions on $(\Omega, \mathcal{F}, \mathbb{P}, T)$ and $X$, it turns out that a probability $\mu$ on $(\Omega \times X, \mathcal{F} \otimes \mathcal{X})$ is $\phi$-invariant if and only if $\phi(\omega, \cdot)_*\mu_\omega = \mu_{T\omega}$ for $\mathbb{P}$-a.e. $\omega \in \Omega$ (see [1, Theorem 1.4.5]).

2.3. **Differentiable random transformations.** Let $M$ be a compact connected Riemannian manifold. We denote by $\|\cdot\|$ and $d$ the norm and the distance, respectively, generated by the Riemannian metric. We also denote by $\mathcal{B}$ the Borel $\sigma$-algebra of $M$ and by $m$ the volume measure on $M$, which is normalized so that $m(M) = 1$. Given a measurable function $\varphi : M \to \mathbb{R}$ that is integrable with respect to $m$, the expression $\|\varphi\|_1$ denotes the integral $\int_M |\varphi(x)|\mathrm{d}m(x)$, i.e., the $L^1$-norm of $\varphi$.

As usual, $C(M)$ denotes the collection of all real continuous functions on $M$. Endowed with the sup norm $\|\cdot\|_\infty$, the space $C(M)$ is a Banach algebra with respect to the pointwise multiplication. The set of strictly positive continuous functions on $M$ is denoted by $C_+(M)$. A central role in our analysis is played by the subset of $C(M)$ consisting of Hölder functions. Let $C^\alpha(M)$ be the set of all real Hölder continuous functions with exponent $0 < \alpha \leq 1$. If $\varphi \in C^\alpha(M)$, we say that $\varphi$ is $\alpha$-Hölder, and its Hölder constant $|\varphi|_\alpha$ is the smallest constant $a \geq 0$ such that $|\varphi(x) - \varphi(y)| \leq a \cdot d(x, y)^\alpha$ for every $x, y \in M$. The set of strictly positive $\alpha$-Hölder functions on $M$ is denoted by $C^\alpha_+(M)$.

The results obtained in this paper concern random dynamical systems generated by measurable maps $\phi : (\Omega \times X, \mathcal{F} \otimes \mathcal{X}) \to (X, \mathcal{X})$ with the following properties:

(1) $X = M$ and $\mathcal{X} = \mathcal{B}$,
(2) $\phi(\omega, \cdot)$ is a surjective local $C^1$ diffeomorphism of $M$ for every $\omega \in \Omega$.

We call such a map $\phi$ a *random local diffeomorphism of* $M$.

Observe that every local diffeomorphism $\psi$ from a connected smooth manifold $M$ to itself is an $N$-covering map of $M$ with $N$ being the number of preimages $\psi^{-1}(x)$ of any point $x \in M$, which is independent of $x$ (see [15, Proposition 2.19]). This means that for every $x \in M$, there exists a connected open set $U$ containing $x$ such that $\psi$ maps diffeomorphically every connected component of $\psi^{-1}(U)$ onto $U$. In particular, an $N$-covering map is surjective. The constant $N$ is called the *degree of* $\psi$.

2.4. **Random transfer operators.** A *random potential on* $M$ is a measurable function $\gamma : \Omega \times M \to \mathbb{R}$. We say that a random potential $\gamma$ on $M$ is *of class* $C^\alpha$ with $0 \leq \alpha \leq 1$ if $\gamma(\omega, \cdot) \in C^\alpha(M)$ for every $\omega \in \Omega$. The term 'potential' has its origin in statistical mechanics where the transfer operator approach to the study of translation invariant states originated (for more details, see [2] and further references therein).

The random transfer operator $\mathcal{L}_{\phi,\gamma}(\omega) : C(M) \to C(M)$ associated to the random local diffeomorphism $\phi$ and the continuous (i.e., of class $C^0$) random potential $\gamma$ is given by

$$(\mathcal{L}_{\phi,\gamma}(\omega)\varphi)(x) = \sum_{y:\phi(\omega,y)=x} e^{\gamma(\omega,y)}\varphi(y) \qquad \text{for } \omega \in \Omega, x \in M \text{ and } \varphi \in C(M).$$

It is not hard to see that, under our assumptions, $\mathcal{L}_{\phi,\gamma}(\omega)$ is a positive bounded linear operator on $C(M)$. Positivity means that $\mathcal{L}_{\phi,\gamma}(\omega)\varphi \geq 0$ whenever $\varphi \in C(M)$ with $\varphi \geq 0$. Note that if $\gamma(\omega, x) = -\log|\det D_x\phi(\omega, \cdot)|$, then $\mathcal{L}_{\phi,\gamma}(\omega)$ is the Perron-Frobenius operator of the transformation $\phi(\omega, \cdot) : M \to M$.

The Riesz Representation Theorem [19] states that the dual space of $C(M)$ is isomorphic to the space $\mathcal{M}(M)$ of regular signed Borel measures on $M$. We adopt the notation $\nu(\varphi)$ to denote the integral $\int_M \varphi(x)\mathrm{d}\nu(x)$ with $\nu \in \mathcal{M}(M)$ and $\varphi$ being a bounded measurable function on $M$. Then the dual operator $\mathcal{L}^*_{\phi,\gamma}(\omega) : \mathcal{M}(M) \to \mathcal{M}(M)$ of $\mathcal{L}_{\phi,\gamma}(\omega)$ is defined through

$$(\mathcal{L}^*_{\phi,\gamma}(\omega)\nu)(\varphi) = \nu(\mathcal{L}_{\phi,\gamma}(\omega)\varphi) \qquad \text{for } \nu \in \mathcal{M}(M) \text{ and } \varphi \in C(M).$$

We also introduce another operator $\hat{\mathcal{L}}_{\phi,\gamma}(\omega)$ that represents a normalization of $\mathcal{L}_{\phi,\gamma}(\omega)$ on $C_+(M)$. Namely, let $\hat{\mathcal{L}}_{\phi,\gamma}(\omega) : C_+(M) \to C_+(M)$ be the operator given by

$$\hat{\mathcal{L}}_{\phi,\gamma}(\omega)\varphi = \frac{1}{\|\mathcal{L}_{\phi,\gamma}(\omega)\varphi\|_\infty} \cdot \mathcal{L}_{\phi,\gamma}(\omega)\varphi \qquad \text{for } \omega \in \Omega \text{ and } \varphi \in C_+(M).$$

The iterates of the operator $\mathcal{L}_{\phi,\gamma}(\omega)$ are defined as follows

$$\mathcal{L}_{n,\phi,\gamma}(\omega) = \begin{cases} \mathrm{Id}_{C(M)} & \text{if } n = 0, \\ \mathcal{L}_{\phi,\gamma}(T^{n-1}\omega) \circ \cdots \circ \mathcal{L}_{\phi,\gamma}(\omega) & \text{if } n > 0. \end{cases}$$

The iterates $\hat{\mathcal{L}}_{n,\phi,\gamma}(\omega)$ of the operator $\hat{\mathcal{L}}_{\phi,\gamma}(\omega)$ are defined in a similar fashion. Whenever there is no risk of confusion, we will often drop the subscripts $\phi$ and $\gamma$ in $\mathcal{L}_{\phi,\gamma}$ and $\hat{\mathcal{L}}_{\phi,\gamma}$ for simplicity's sake.

## 3. Main results

In this section, we line out the assumptions made throughout the whole paper, and formulate our main results in Theorems 3.1 and 3.2.

Suppose that $\phi$ is a random local diffeomorphism of a manifold $M$, and that $\gamma$ is a random potential on $M$ of class $C^\alpha$ with $0 < \alpha \leq 1$. For every $\omega \in \Omega$, the quantity

$$\sigma(\omega) = \min_{x \in M} \|(D_x\phi(\omega, \cdot))^{-1}\|^{-1}$$

is the least expansion coefficient of $\phi(\omega, \cdot)$. Since $\phi(\omega, \cdot)$ is a local diffeomorphism, $\sigma(\omega)$ is strictly positive for every $\omega \in \Omega$. Define

$$b(\omega) = \max\{|\gamma(\omega, \cdot)|_\alpha, 1\} \qquad \text{for } \omega \in \Omega.$$

The separability of $M$ implies that the functions $\sigma$ and $b$ are measurable, as will be shown in Lemma 6.6. We further impose on $\phi$ and $\gamma$ the following conditions:

**C1 (Expansion on average):** The random map $\omega \mapsto \phi(\omega, \cdot)$ expands on average, i.e., $\log \sigma(\omega) \in L^1(\mathbb{P})$ and

$$\int_\Omega \log \sigma(\omega) \mathrm{d}\mathbb{P}(\omega) > 0.$$

**C2 (Integrability):** $\log b \in L^1(\mathbb{P})$.

In fact, condition C2 can be weakened somewhat. It is enough to assume that $b$ is a measurable and tempered function, see the comments immediately after Lemma 6.2. Note further that $b$ and $\sigma$ depend on the choice of the metric on $M$.

We are now ready to formulate our main results. Their proofs follows from the series of results proved in Section 6. The symbol $\mathbb{1}$ denotes the characteristic function of $M$.

**Theorem 3.1.** *Let $\phi$ be a random local diffeomorphism on a manifold $M$, and let $\gamma$ be a random potential on $M$ of class $C^\alpha$ with $0 < \alpha \leq 1$. Suppose that $\phi$ and $\gamma$ satisfy conditions C1 and C2. Then there exist a constant $\xi \in [-\infty, 0)$, a full $\mathbb{P}$-measure $T$-invariant set $\Omega_0 \subset \Omega$, a measurable function $h : \Omega \times M \to \mathbb{R}$, a family $\nu = \{\nu_\omega\}_{\omega \in \Omega_0}$ of finite regular Borel measures on $\mathcal{B}$ and a measurable function $\Lambda : \Omega_0 \to (0, +\infty)$ such that for every $\omega \in \Omega_0$ and $\varphi \in C^\alpha_+(M)$, we have*

*(1) $h(\omega, \cdot) \in C^\alpha_+(M)$ and $\|h(\omega, \cdot)\|_\infty = 1$,*
*(2) $\mathcal{L}(\omega)h(\omega, \cdot) = \Lambda(\omega)h(T\omega, \cdot)$,*
*(3) $\mathcal{L}^*(\omega)\nu_{T\omega} = \Lambda(\omega)\nu_\omega$,*
*(4) $\nu_\omega(h(\omega, \cdot)) = 1$,*
*(5) $\Omega_0 \ni \omega \mapsto \nu_\omega(A)$ is measurable for every $A \in \mathcal{B}$,*

*(6)* $\limsup_{n\to+\infty} n^{-1}\log\|\hat{\mathcal{L}}_n(T^{-n}\omega)\mathbb{1} - h(\omega,\cdot)\|_\infty \le \xi$,
*(7)* $\limsup_{n\to+\infty} n^{-1}\log\|\hat{\mathcal{L}}_n(\omega)\varphi - h(T^n\omega,\cdot)\|_\infty \le \xi$.

*Furthermore, the quadruple* $(\Omega_0,\Lambda,h,\nu)$ *is unique in the following sense. If there exist another quadruple* $(\tilde\Omega_0,\tilde\Lambda,\tilde h,\tilde\nu)$ *formed by a full* $\mathbb{P}$*-measure* $T$*-invariant set* $\tilde\Omega_0\subset\Omega$*, a measurable function* $\tilde h:\Omega\times M\to\mathbb{R}$*, a family* $\tilde\nu=\{\tilde\nu_\omega\}_{\omega\in\tilde\Omega_0}$ *of finite regular Borel measures on* $\mathcal{B}$ *and a measurable function* $\tilde\Lambda:\tilde\Omega_0\to(0,+\infty)$ *satisfying conclusions (1)-(5), then* $\tilde h(\omega)=h$*,* $\tilde\Lambda(\omega)=\Lambda(\omega)$ *and* $\tilde\nu_\omega=\nu_\omega$ *for every* $\omega\in\tilde\Omega_0\cap\Omega_0$.

Let $(\Omega_0,\Lambda,h,\nu)$ be the quadruple derived in the previous theorem. We define $\Lambda_n(\omega)=\Lambda(T^{n-1}\omega)\cdots\Lambda(\omega)$ for $\omega\in\Omega_0$ and $n>0$, We also define a probability measure $\mu$ on $\mathcal{F}\otimes\mathcal{M}$ by specifying its disintegration $\{\mu_\omega\}_{\omega\in\Omega_0}$ on $\mathbb{P}$,

$$\mu_\omega = h(\omega,\cdot)\nu_\omega \qquad \text{for } \omega\in\Omega_0.$$

Finally, given $\varphi\in C^\alpha(M)$ and $\psi:M\to\mathbb{R}$ measurable and bounded, we define the *random correlation function of* $\varphi$ *and* $\psi$ as follows

$$C_n(\omega,\varphi,\psi) = \int_M \varphi(\omega,x)\psi(\omega,\phi_n(\omega,x))\mathrm{d}\mu_\omega(x) - \int_M \varphi(\omega,x)\mathrm{d}\mu_\omega(x)\int_M \psi(\omega,x)\mathrm{d}\mu_{T^n\omega}(x) \qquad \text{for } \omega\in\Omega_0 \text{ and } n>0.$$

**Theorem 3.2.** *Suppose that the hypotheses of Theorem 3.1 are satisfied, and let* $(\xi,\Omega_0,\Lambda,h,\nu)$ *be the quintuple obtained in Theorem 3.1. Then for every* $\omega\in\Omega_0$*, every* $\varphi\in C^\alpha(M)$ *and every function* $\psi:M\to\mathbb{R}$ *measurable and bounded, we have*

*(1)* $\limsup_{n\to+\infty} n^{-1}\log\|\Lambda_n(\omega)^{-1}\mathcal{L}_n(\omega)\varphi - \nu_\omega(\varphi)h(T^n\omega,\cdot)\|_\infty \le \xi$,
*(2)* $\limsup_{n\to+\infty} n^{-1}\log|C_n(\omega,\varphi,\psi)| \le \xi$,
*(3) the probability* $\mu$ *is* $\phi$*-invariant and ergodic.*

As already detailed in the introduction, Theorems 3.1 and 3.2 constitute a random generalization of the classical Ruelle–Perron–Frobenius Theorem. Note that conclusion (6) of Theorem 3.1 provides information as to how the eigenfunction $h$ is constructed, namely as the limit of the sequence $\{\hat{\mathcal{L}}_n(T^{-n}\omega)\mathbb{1}\}_{n\in\mathbb{N}}$. The second part of conclusion (6) of Theorem 3.1 gives a related but slightly different result, namely that all forward orbits of the form $\{\hat{\mathcal{L}}_n(\omega)\varphi\}_{n\in\mathbb{N}}$ with $\varphi\in C^\alpha_+(M)$ are asymptotically attracted by the orbit $\{h(T^n\omega,.)\}_{n\in\mathbb{N}}$ with exponential speed of convergence. Finally, conclusion (1) of Theorem 3.2 states that the iterates of the operator $\mathcal{L}(\omega)$ asymptotically behave like a projector onto the eigenfunction $h$.

If $\gamma(\omega,x)=-\log|\det D_x\phi(\omega,\cdot)|$ in Theorems 3.1 and 3.2, then the random transfer operator $\mathcal{L}(\omega)$ becomes the usual Frobenius–Perron operator associated to the random map $\phi(\omega,\cdot)$. In this case, the measures $\{\mu_\omega\}_{\omega\in\Omega_0}$ are all absolutely continuous with respect to the Riemannian volume $m$. Of course, to guarantee that such a potential $\gamma$ is of class $C^\alpha$, we may assume that $\phi(\omega,\cdot)\in C^{1+\alpha}(M)$ for all $\omega\in\Omega$.

**Theorem 3.3.** *Suppose the transformation* $\phi(\omega,\cdot)$ *is of class* $C^{1+\alpha}(M)$ *for every* $\omega\in\Omega$*, and the random potential is given by* $\gamma(\omega,x)=-\log|\det D_x\phi(\omega,\cdot)|$*. Then the probabilities* $\mu_\omega$ *defined in* (10) *are absolutely continuous with respect to* $m$*. It follows that the probability* $\mu$ *is absolutely continuous with respect to* $\mathbb{P}\times m$*. Moreover,* $\mu$ *is the unique* $F$*-invariant probability with this property.*

## 4. A random fixed point theorem for Lipschitz maps

In this section, we present a random fixed point theorem building on a result of Bougerol [4, Theorem 3.1]. This theorem can be regarded as a random version of Banach's contracting principle. We consider here random dynamical systems generated by Lipschitz continuous transformations fulfilling certain growth hypotheses (assumptions A1 and A2). We then investigate several properties of the random fixed point, namely measurability, uniqueness, and temperedness. The last property plays an important role in the proof of Theorem 3.1. For the sake of brevity, this section will not contain any proofs, as our fixed point theorem is but a technical device used in the proof of Theorem 3.1. Further to this, the proofs provide little additional insight with regards to the main theme of this paper.

4.1. **Lipschitz random maps.** Let $\phi$ be random dynamical system on a set $X$. In this section, we dot not assume that $\phi$ is smooth. Instead, we assume that there exist a collection $\{X_\omega\}_{\omega\in\Omega}$ of subsets of $X$ and a complete metric $d_\omega$ on each $X_\omega$ such that $\phi(\omega, X_\omega) \subseteq X_{T\omega}$ and the transformation $\phi(\omega,\cdot)|_{X_\omega} : (X_\omega, d_\omega) \to (X_{T\omega}, d_{T\omega})$ is Lipschitz for every $\omega \in \Omega$. We call $\phi$ a *Lipschitz random dynamical system*. In the invertible case, we assume that $\phi(\omega, X_\omega) = X_{T\omega}$ and that the inverse transformation $\phi(\omega,\cdot)^{-1}|_{X_{T\omega}} : (X_{T\omega}, d_{T\omega}) \to (X_\omega, d_\omega)$ is Lipschitz as well.

For every $\omega \in \Omega$ and $n \geq 0$, denote by $\rho_n(\omega)$ the Lipschitz constant of $\phi(\omega,\cdot)|_{X_\omega}$, i.e.,

$$\rho_n(\omega) = \sup\left\{\frac{d_{T^n\omega}\left(\phi_n(\omega,x),\phi_n(\omega,y)\right)}{d_\omega\left(x,y\right)} : x, y \in X_\omega \text{ and } x \neq y\right\}.$$

Note that in the invertible case, this definition makes sense also for $n < 0$. In that case, we further define

$$r_n(\omega) = \inf\left\{\frac{d_{T^n\omega}\left(\phi_n(\omega,x),\phi_n(\omega,y)\right)}{d_\omega\left(x,y\right)} : x, y \in X_\omega \text{ and } x \neq y\right\}.$$

From the cocycle property of $\{\phi_n\}$ and the definition of $\rho_n$ and $r_n$, we obtain several relations linking $\rho_n$ and $r_n$. These are displayed in the next lemma, whose proof is left to the reader.

**Lemma 4.1.** *Let $\phi$ be an invertible Lispchitz random dynamical system. Then for every $\omega \in \Omega$ and every $m, n \in \mathbb{Z}$, we have*

*(1)* $\rho_{-n}(T^n\omega) \cdot r_n(\omega) = 1$,
*(2)* $r_m(\omega) \cdot \rho_n(T^m\omega) \leq \rho_{n+m}(\omega)$,
*(3)* $r_{m+n}(\omega) \leq r_n(T^m\omega) \cdot \rho_m(\omega)$,
*(4)* $\rho_{n+m}(\omega) \leq \rho_n(T^m\omega) \cdot \rho_m(\omega)$.

4.2. **Tempered functions.** We now introduce the concept of a tempered function. Contrary to the commonly used definition (for example, see [1, Definition 4.1.1]), we do not require the function to be measurable.

**Definition 4.2.** We say that a function $f : \Omega \to (0, +\infty)$ is *tempered with respect to* $(\Omega, \mathcal{F}, \mathbb{P}, T)$ if

$$\lim_{n\to\pm\infty} \frac{1}{|n|} \log^+ f(T^n\omega) = 0 \qquad \text{for a.e. } \omega \in \Omega. \tag{1}$$

**Remark 1.** If $f$ is measurable and $\log^+ f \in L^1(\mathbb{P})$, then a straightforward application of the Borel-Cantelli lemma shows that $f$ is tempered (see [1, Proposition 4.1.3]).

The next lemma gives an equivalent characterisation of tempered functions; see for example [1, Proposition 4.3,.3] for a proof.

**Lemma 4.3.** *Suppose that $f : \Omega \to (0, +\infty)$ is a tempered function. Then for every $\epsilon > 0$, there exists a function $C_\epsilon : \Omega \to (0, +\infty)$ such that for $\mathbb{P}$-a.e. $\omega \in \Omega$ and every $m \in \mathbb{Z}$, we have*

*(1) $f(\omega) \leq C_\epsilon(\omega)$,*
*(2) $C_\epsilon(T^m\omega) \leq C_\epsilon(\omega)e^{\epsilon|m|}$.*

*Moreover, if $f$ is measurable, then $C_\epsilon$ is measurable as well.*

4.3. **The random fixed point theorem.** Let $\phi$ be a Lipschitz random dynamical system. Before stating our random fixed point point theorem, we formulate a series of conditions.

**A1 (Non–uniform contraction):** There exists a constant $\beta \in [-\infty, 0)$ such that for $\mathbb{P}$-a.e. $\omega \in \Omega$, we have
- $\limsup_{n\to+\infty} \frac{1}{n} \log \rho_n(\omega) \leq \beta$,
- $\limsup_{n\to+\infty} \frac{1}{n} \log \rho_n(T^{-n}\omega) \leq \beta$.

**A2 (Temperedness):** There exist a map $x_0 : \Omega \to X$ such that
- $x_0(\omega) \in X_\omega$ for every $\omega \in \Omega$,
- the function $\omega \mapsto d_\omega\left(x_0(\omega), \phi(T^{-1}\omega, x_0(T^{-1}\omega))\right)$ is tempered.

**A3 (Measurability):** The set $X$ is equipped with a metric $d$ such that
- $d|_{X_\omega} \leq d_\omega$ for every $\omega \in \Omega$. This implies that the topology generated by $d$ on $X_\omega$ is not stronger than the one generated by the metric $d_\omega$ for every $\omega \in \Omega$,
- the transformation $\phi : (\Omega \times X, \mathcal{F} \otimes \mathcal{X}) \to (X, \mathcal{X})$ is measurable with $\mathcal{X}$ being the Borel $\sigma$-algebra of $X$ generated by $d$.

**Remark 2.** Suppose that each function $\rho_n : \Omega \to [0, +\infty)$ is measurable and that $\int_\Omega \log^+ \rho_1(\omega)\mathrm{d}\mathbb{P}(\omega) < +\infty$. Since the process $\{\log \rho_n\}_{n\geq 0}$ is subadditive, an easy application of the Subadditive Ergodic Theorem shows that assumption A1 is satisfied.

We are now ready to formulate our random fixed point theorem.

**Theorem 4.4.** *Let $\phi$ be a Lipschitz random dynamical system satisfying assumptions A1 and A2. Let $\beta < 0$ be the constant in assumption A1. Then there exist a full $\mathbb{P}$-measure $T$-invariant set $\Omega_1 \subset \Omega$ and a map $Z : \Omega \to X$ (random fixed point) such that for every $\omega \in \Omega_1$, we have*

*(1) $Z(\omega) \in X_\omega$,*
*(2) $\limsup_{n\to+\infty} \frac{1}{n} \log d_\omega\left(Z(\omega), \phi_n(T^{-n}\omega, x_0(T^{-n}\omega))\right) \leq \beta$,*
*(3) $Z(T^{n+1}\omega) = \phi(T^n\omega, Z(T^n\omega))$ for every $n \in \mathbb{Z}$,*
*(4) $\limsup_{n\to+\infty} \frac{1}{n} \log d_{T^n\omega}(Z(T^n\omega), \phi_n(\omega, x)) \leq \beta$ for every $x \in X_\omega$.*

**Remark 3.** Suppose that $X = \mathbb{R}$ and $\phi(\omega, x) = a(\omega)x + b(\omega)$ for $x \in \mathbb{R}$ with $a : \Omega \to (0, +\infty)$ and $b : \Omega \to \mathbb{R}$. Such an affine random dynamical system $\phi$ is of course Lipschitz. If $\phi$ satisfies Conditions A1 and A2, then Theorem 4.4 applies to $\phi$ (see also [1, Chapter 5]).

We now address the issue of the measurability and the uniqueness of the random fixed point $Z$ obtained in Theorem 4.4.

**Corollary 1.** *Let $\phi$ be a Lipschitz random dynamical system satisfying Conditions A1-A3, and assume that the map $x_0 : \Omega \to X$ in assumption A2 is measurable. Then the*

*map $Z : \Omega \to X$ given by Theorem 4.4 is measurable, and is the only measurable map (up to a set of $\mathbb{P}$-measure zero) satisfying conclusions (1) and (3) of Theorem 4.4.*

Finally, we provide sufficient conditions for the function $\omega \mapsto d_\omega(x_0(\omega), Z(\omega))$ to be tempered. We start with the following remark.

**Remark 4.** Let $\phi$ be an invertible Lipschitz random dynamical system. Suppose that that there exist a constant $\beta \in \mathbb{R}$ such that for $\mathbb{P}$-a.e. $\omega \in \Omega$, we have

$$\lim_{n\to\pm\infty} \frac{1}{n} \log r_n(\omega) = \lim_{n\to\pm\infty} \frac{1}{n} \log \rho_n(\omega) = \beta. \tag{2}$$

Then Lemma 4.1 implies that $\rho_{-n}(T^{-n}\omega) = 1/r_{-n}(\omega)$ for $n \in \mathbb{Z}$. Hence, it is immediate to see that the hypotheses 2 together with the extra assumption that $\beta < 0$ implies assumption A1.

**Proposition 1.** *Let $\phi$ be an invertible Lipschitz random dynamical system satisfying the hypotheses of Remark 4 with $\beta < 0$ and assumption A2. Then the conclusions of Theorem 4.4 hold. Moreover, if $Z : \Omega \to X$ is the map as in Theorem 4.4, then the function $\omega \mapsto d_\omega(x_0(\omega), Z(\omega))$ is tempered, and $Z$ is the unique map (up to a set of $\mathbb{P}$-measure zero) with this property satisfying conclusions (1) and (3) of Theorem 4.4.*

## 5. Hilbert's projective metric

In this section, we recall the essential definitions and properties concerning Hilbert's projective metric on cones, which plays a very important role in our proof of Theorem 3.1. For a more complete account on the subject, the reader is referred to [2, 5, 7, 17, 18, 20].

5.1. **Hilbert's metric on cones.** Let $(\mathcal{B}, |\cdot|)$ be a real normed space with the topology induced by the norm $|\cdot|$. A subset $\mathcal{C}$ of $\mathcal{B}\setminus\{0\}$ is called a *cone* if $\lambda\varphi \in \mathcal{C}$ for every $\varphi \in \mathcal{C}$ and every $\lambda > 0$. We say that the cone $\mathcal{C}$ is *convex* if $\lambda\varphi + \mu\psi \in \mathcal{C}$ for every $\varphi, \psi \in \mathcal{C}$ and every $\lambda, \mu > 0$. We say that $\mathcal{C}$ is *closed* if $\mathcal{C} \cup \{0\}$ is closed.

Any convex cone $\mathcal{C}$ defines a partial ordering on $\mathcal{B}$ by the rule $\varphi \preceq \psi$ with $\varphi, \psi \in \mathcal{B}$ if and only if $\psi - \varphi \in \mathcal{C} \cup \{0\}$. If $\mathcal{C}$ is also closed, then $\preceq$ is continuous, i.e., if $\varphi_n, \psi \in \mathcal{B}$ such that $\psi \preceq \varphi_n$ for every $n > 0$ and $\lim_{n\to+\infty} \varphi_n = \varphi$, then $\psi \preceq \varphi$.

We say that two elements $\varphi, \psi$ of a convex cone $\mathcal{C}$ are *comparable* and write $\varphi \sim \psi$ if and only if $\lambda\varphi \preceq \psi \preceq \mu\varphi$ for some $\lambda, \mu > 0$. The relation $\sim$ is an equivalence relation, and the equivalence classes of $\mathcal{C}$ are called *components* of $\mathcal{C}$. We denote by $\mathcal{C}_\psi$ the component of $\mathcal{C}$ containing the element $\psi \in \mathcal{C}$. A component of $\mathcal{C}$ has all the property of the cone $\mathcal{C}$ except possibly the closedness.

The *Hilbert metric* $\theta$ on a convex cone $\mathcal{C}$ is defined as follows. Let

$$a(\varphi, \psi) = \sup\{\lambda > 0 : \lambda\varphi \preceq \psi\},$$
$$b(\varphi, \psi) = \inf\{\mu > 0 : \psi \preceq \mu\varphi\}.$$

Then for any pair $\varphi, \psi \in \mathcal{C}$, define

$$\theta(\varphi, \psi) = \begin{cases} \log \frac{b(\varphi,\psi)}{a(\varphi,\psi)} & \text{if } \varphi \sim \psi, \\ +\infty & \text{otherwise.} \end{cases} \tag{3}$$

It is easy to check that the restriction of $\theta$ to each component $\mathcal{C}_\psi$ of $\mathcal{C}$ with $\psi \in \mathcal{C}$ is a pseudo-metric, and the restriction of $\theta$ to the set $\{\varphi \in \mathcal{C}_\psi : |\varphi| = 1\}$ is a metric.

5.2. **Birkhoff's contraction coefficient.** Let $(\mathcal{C}_1, \theta_1)$ and $(\mathcal{C}_2, \theta_2)$ be convex cones of the real normed spaces $(\mathcal{B}_1, |\cdot|_1)$ and $(\mathcal{B}_2, |\cdot|_2)$, respectively, with their respective Hilbert metrics. Suppose that $L : \mathcal{B}_1 \to \mathcal{B}_2$ is a linear transformation such that $L\mathcal{C}_1 \subset \mathcal{C}_2$. Then it can be shown that the restriction of $L$ to $\mathcal{C}_1$ is a contraction with respect to $\theta_1$ and $\theta_2$ [20, Section 2.1], i.e.

$$\theta_2(L\varphi, L\psi) \leq \theta_1(\varphi, \psi) \qquad \text{for } \varphi, \psi \in \mathcal{C}_1.$$

The next result is due originally to Birkhoff [7, Lemma 1, Section 4], and shows that $L$ is a strict contraction if the diameter of $L\mathcal{C}_1$ is finite.

**Proposition 2.** *Suppose that $D = \sup\{\theta_2(L\varphi, L\psi) : \varphi, \psi \in \mathcal{C}_1\}$ is finite, then*

$$\theta_2(L\varphi, L\psi) \leq \tanh\left(\frac{D}{4}\right) \theta_1(\varphi, \psi) \qquad \textit{for } \varphi, \psi \in \mathcal{C}_1. \tag{4}$$

For the proof of this proposition, see [20, Proposition 2.3]. The factor $\tanh(D/4)$ is called the *Birkhoff coefficient of $L$*.

5.3. **Completeness of Hilbert's metric.** We now turn our attention to the critical issue of the completeness of the Hilbert metric.

Let $\mathcal{C}$ be a convex cone of the real normed space $(\mathcal{B}, |\cdot|)$. We say that $\mathcal{C}$ is *normal* if there exists $A > 0$ such that $|\psi| \leq A|\varphi|$ whenever $0 \preceq \psi \preceq \varphi$.

A mapping $K : \mathcal{C} \to (0, +\infty)$ is called a *functional on the cone $\mathcal{C}$*. We say that a functional $K$ on $\mathcal{C}$ is *continuous* if $K$ is continuous with respect to the topology induced on $\mathcal{C}$ by the norm $|\cdot|$ of $\mathcal{B}$. A functional $K$ on $\mathcal{C}$ is called *monotone* if $K(\lambda\varphi) = \lambda K(\varphi)$ for every $\lambda > 0$ and every $\varphi \in \mathcal{C}$, and $K(\psi) \leq K(\varphi)$ whenever $0 \preceq \psi \preceq \varphi$.

The proof of the next lemma is but a minor refinement of the proof of [17, Lemma 1.3] (see also [18, Relation (1.20a)]).

**Lemma 5.1.** *Let $\mathcal{C}$ be a convex closed cone of a real normed space $(\mathcal{B}, |\cdot|)$. Suppose that $\mathcal{C}$ is normal and that there exists a monotone functional $K$ on $\mathcal{C}$. Let $A > 0$ be the constant as in the definition of a normal cone. If $\varphi, \psi \in \mathcal{C}$ and $K(\varphi) = K(\psi) = 1$, then*

$$|\varphi - \psi| \leq (1 + 2A)\left(e^{\theta(\varphi,\psi)} - 1\right) \min\{|\varphi|, |\psi|\}.$$

We can state in what sense Hilbert's metric is complete. Note that in the following proposition, we strengthen our assumptions, as we now require that $(\mathcal{B}, |\cdot|)$ is a Banach space, and that the monotone functional $K$ is continuous. The proof is essentially [6, Theorem 5].

**Proposition 3.** *Let $\mathcal{C}$ be a convex closed cone of a real Banach space $(\mathcal{B}, |\cdot|)$. Suppose that $\mathcal{C}$ is normal and $K$ is a continuous monotone functional on $\mathcal{C}$. Let $\psi \in \mathcal{C}$, and let $\Sigma = \{\varphi \in \mathcal{C}_\psi : K(\varphi) = 1\}$. Then the pair $(\Sigma, \theta|_\Sigma)$ is a complete metric space.*

5.4. **Cones of Hölder continuous functions.** We now apply the results of the previous subsection to a special family of cones, which are used in the proof of Theorem 3.1. The notation here is as in Subsection 2.3.

Let $M$ be a connected compact smooth Riemannian manifold, and consider the Banach space $(C(M), \|\cdot\|_\infty)$. For every $t \geq 0$, let

$$\mathcal{C}(t) = \{\varphi \in C_+(M) : \varphi(x) \leq e^{td(x,y)^\alpha} \varphi(y) \text{ for } x, y \in M\}$$

be the set of all continuous functions on $M$ whose logarithm is an $\alpha$-Hölder function with Hölder constant less or equal than $t$. One can easily check that each $\mathcal{C}(t)$ is a

closed convex cone of $(C(M), \|\cdot\|_\infty)$ and a subset of $C^\alpha_+(M)$. Indeed, we have $C^\alpha_+(M) = \bigcup_{t\geq 0} \mathcal{C}(t)$.

The Hilbert projective metric $\theta_t$ on $\mathcal{C}(t)$ is given by

$$\theta_t(\varphi,\psi) = \log\left(\sup_{\substack{x\neq y\\ u\neq v}} \frac{e^{td(x,y)^\alpha}\varphi(x)-\varphi(y)}{e^{td(x,y)^\alpha}\psi(x)-\psi(y)} \cdot \frac{e^{td(u,v)^\alpha}\psi(u)-\psi(v)}{e^{td(u,v)^\alpha}\varphi(u)-\varphi(v)}\right)$$

for $\varphi,\psi\in\mathcal{C}(t)$. For a proof, see [2, Theorem 2.1].

Let $\preceq_t$ be the partial ordering on $C(M)$ generated by $\mathcal{C}(t)$. We will simply write $\preceq$ when there is no danger of ambiguity.

**Corollary 2.** *Let $\psi\in\mathcal{C}(t)$, and define $\Sigma^{(k)} = \{\varphi\in\mathcal{C}_\psi(t) : \|\varphi\|_k = 1\}$ for $k = 1,\infty$. Then*

*(1) $\|\varphi_1-\varphi_2\|_\infty \leq 3\left(e^{\theta_t(\varphi_1,\varphi_2)}-1\right)\min\{\|\varphi_1\|_\infty,\|\varphi_2\|_\infty\}$ for $\varphi_1,\varphi_2\in\Sigma^{(k)}$ for each $k=1,\infty$,*

*(2) $\|\varphi_1-\varphi_2\|_1 \leq 3\left(e^{\theta_t(\varphi_1,\varphi_2)}-1\right)$ for $\varphi_1,\varphi_2\in\Sigma^{(1)}$,*

*(3) $(\Sigma^{(k)},\theta_t|_{\Sigma^{(k)}})$ is a complete metric space for each $k=1,\infty$.*

*Proof.* This follows easily from Lemma 5.1 and Proposition 3. □

Denote by $\mathbb{1}$ the characteristic function on $M$. Clearly, $\mathbb{1}\in\mathcal{C}(t)$ for every $t\geq 0$. The next lemma shows that the cone $\mathcal{C}(s)$ is contained in $\mathcal{C}_{\mathbb{1}}(t)$ (the connected component of $\mathcal{C}$ containing $\mathbb{1}$) for every $0\leq s<t$. The proof can be found (with minor modifications) in [20].

**Lemma 5.2.** *Let $\varphi\in\mathcal{C}(\delta t)$ for some $t\geq 0$ and $0<\delta<1$. Then*

$$\theta_t(\mathbb{1},\varphi) \leq \log\frac{1+\delta}{1-\delta} + (\operatorname{diam} M)^\alpha t\delta.$$

## 6. Proof of Theorems 3.1-3.3

Let $\phi$ be a random local diffeomorphism on a manifold $M$, and let $\gamma$ be a random potential on $M$ of class $C^\alpha$ with $0<\alpha\leq 1$. We also assume that $\phi$ and $\gamma$ satisfy Conditions C1 and C2. In this section, we prove Theorem 3.1. The main idea of the proof is to look at the random dynamical system generated by the random transfer operator $\mathcal{L} = \mathcal{L}_{\phi,\gamma}$, and apply Theorem 4.4 to it. Since the proof is quite long, we split it into several parts. The notation used throughout this section is as in Sections 2-5.

### 6.1. Preliminaries.

In this subsection, we give a precise definition of the random dynamical system generated by $\mathcal{L}$, and obtain some preliminary results required for the proof of Theorem 3.1. Let $\{\mathcal{C}(t)\}_{t\geq 0}$ be the family of cones defined in Subsection 5.4.

**Lemma 6.1.** *Suppose that $\varphi\in\mathcal{C}(t)$ with $t\geq 0$. Then*

$$\mathcal{L}(\omega)\varphi \in \mathcal{C}\left(\frac{t+b(\omega)}{\sigma^\alpha(\omega)}\right) \qquad \text{for } \omega\in\Omega. \tag{5}$$

*Proof.* see [20]. □

By Condition C1, we can choose $0<\delta<1$ such that

$$\beta := -\log\delta - \alpha\int_\Omega \log\sigma(\omega)\mathrm{d}\mathbb{P}(\omega) < 0.$$

Now, consider the transformation $q : \Omega \times \mathbb{R} \to \mathbb{R}$ given by

$$q(\omega, t) = \frac{t + b(\omega)}{\delta \sigma^{\alpha}(\omega)} \qquad \text{for } \omega \in \Omega \text{ and } t \in \mathbb{R}.$$

This map $q$ generates a random dynamical system on $\mathbb{R}$ that is measurable, invertible and Lipschitz. Indeed, using the notation introduced in Section 2, for every $\omega \in \Omega$, we have

- $X_\omega = X = \mathbb{R}$,
- $d_\omega = d$ is the Euclidean metric,
- $r_n(\omega) = \rho_n(\omega) = \delta^{-n}(\sigma(\omega) \cdots \sigma(T^{n-1}\omega))^{-1}$ for $n \in \mathbb{Z}$.

Denote by $q_n$ the $n$-iterates of $q$ as in Subsection 2.1.

**Lemma 6.2.** *The random dynamical system $q$ satisfies the hypotheses of Remark 4 as well as assumptions A2 and A3.*

*Proof.* Assumption A3 is trivially satisfied. Next, it is easy to see that $\{\log \rho_n\}_{n\in\mathbb{Z}}$ is an additive process, and $\log \rho_n \in L^1(\mathbb{P})$ as a consequence of condition C1. The Birkhoff Ergodic Theorem then implies that

$$\begin{aligned} \lim_{n\to\pm\infty} \frac{1}{n} \log \rho_n(\omega) &= \int_\Omega \log \rho_1(\omega) \mathrm{d}\mathbb{P}(\omega) \\ &= -\log \delta - \alpha \int_\Omega \log \sigma(\omega) \mathrm{d}\mathbb{P}(\omega) \\ &= \beta \qquad \text{for } \mathbb{P}\text{-a.e. } \omega \in \Omega. \end{aligned}$$

The same conclusion is obviously true for $r_n$ so that $q$ satisfies the hypotheses of Remark 4.

We now show that $q$ satisfies assumption A2 with $x_0 \equiv 0$. Clearly $0 \in X_\omega = \mathbb{R}$ for every $\omega$. So what remains to do is to show that the function $\omega \mapsto q(\omega, 0) = b(\omega)\delta^{-1}\sigma(\omega)^{-\alpha}$ is tempered. To see this, note that

$$\log^+(q(\omega, 0)) \le \log \frac{1}{\delta} + \log^+ b + \log^- \sigma, \tag{6}$$

and $\log^- \sigma$ as well as $\log^+ b$ are in $L^1(\mathbb{P})$ by conditions C1 and C2, respectively. Temperedness of $q(\cdot, 0)$ therefore follows by Remark 1. This completes the proof. □

We gather from the proof that the condition C2 is only needed to ensure the temperedness of $b$, in order to show that $q(\cdot, 0)$ is tempered. But this would follow already if instead of condition C2 we merely demand that $b$ is tempered (the measurability of $b$ was needed as well to show that $q$ is is measurable). Indeed, if $b$ and $1/\sigma$ are tempered (the latter because $\log^+(1/\sigma) = \log^- \sigma \in L^1(\mathbb{P})$), we can apply Lemma 4.3 to these quantities to show that $q$ is tempered as well. Remark 4 applies to $q$ and so the hypotheses of Remark 4 imply assumption A1. By Lemma 6.2, we then see that Theorem 4.4, Remark 3, Corollary 1 and Proposition 1 all apply to $q$, and we obtain the next proposition.

**Proposition 4.** *There exist a full $\mathbb{P}$-measure $T$-invariant set $\Omega_1 \subset \Omega$ and a unique measurable tempered map $Z : \Omega \to [0, +\infty)$ such that for every $\omega \in \Omega_1$, we have*

*(1)* $Z(\omega) = \sum_{k=1}^{+\infty} \delta^{-k-1} \left(\sigma(T^{-k}\omega) \cdots \sigma(T^{-1}\omega)\right)^{-\alpha} b(T^{-k}\omega) > 0$,
*(2)* $Z(T^{n+1}\omega) = q(T^n\omega, Z(T^n\omega))$ *for every* $n \in \mathbb{Z}$,
*(3)* $\limsup_{n\to+\infty} n^{-1} \log |Z(T^n\omega) - q_n(\omega, t)| \le \beta$ *for every* $t \in \mathbb{R}$.

The next lemma is an easy consequence of Lemma 6.1, Proposition 4 and Lemma 5.2.

**Lemma 6.3.** *Let $\Omega_1$ and $Z$ be as in Proposition 4. Then for every $n > 0$, we have*

*(1) $\mathcal{L}_n(\omega)\mathcal{C}(t) \subset \mathcal{C}(\delta q_n(\omega, t)) \subset \mathcal{C}_{\mathbb{1}}(q_n(\omega, t))$ for $\omega \in \Omega$ and $t \geq 0$,*
*(2) $\mathcal{L}_n(\omega)\mathcal{C}(Z(\omega)) \subset \mathcal{C}(\delta Z(T^n\omega)) \subset \mathcal{C}_{\mathbb{1}}(Z(T^n\omega))$ for $\omega \in \Omega_1$.*

The following proposition will allow us to reduce the proof of Theorem 3.1 for the general case $\varphi \in C^\alpha_+(M)$ to the case $\varphi \in \mathcal{C}(Z(\omega))$.

**Proposition 5.** *Let $\Omega_1$ and $Z$ be as in Theorem 4. Then*

*(1) for every $t \in \mathbb{R}$ and $\mathbb{P}$-a.e. $\omega \in \Omega_1$, there exists $n > 0$ such that $\delta q_n(\omega, t) < Z(T^n\omega)$;*
*(2) for every $\varphi \in C^\alpha_+(M)$ and $\mathbb{P}$-a.e. $\omega \in \Omega_1$, there exists an integer $n > 0$ such that $\mathcal{L}_n(\omega)\varphi \in \mathcal{C}(Z(T^n\omega))$.*

*Proof.* By conclusion (1) of Theorem 4, we can find $\epsilon > 0$ such that the set

$$\tilde{\Omega} = \left\{\omega \in \Omega_1 : Z(\omega) > \epsilon \frac{\delta}{1-\delta}\right\}$$

has positive $\mathbb{P}$-measure. Since $(\Omega, \mathcal{F}, \mathbb{P}, T)$ is ergodic, there exists a full $\mathbb{P}$-measure set $\Omega_2 \subset \Omega_1$ such that for every $\omega \in \Omega_2$, we can find an increasing sequence of positive integers $\{n_k(\omega)\}_{k>0}$ such that $T^{n_k(\omega)}\omega \in \tilde{\Omega}$ for every $k > 0$.

Fix $\omega \in \Omega_2$ and $t \in \mathbb{R}$. Write $n_k$ for $n_k(\omega)$. By conclusion (3) of Theorem 4, we have

$$\lim_{k\to+\infty} (q_{n_k}(\omega, t) - Z(T^{n_k}\omega)) = 0.$$

Hence, for $k$ sufficiently large,

$$q_{n_k}(\omega, t) \leq Z(T^{n_k}\omega) + \epsilon.$$

But since $T^{n_k}\omega \in \tilde{\Omega}$, it follows that

$$q_{n_k}(\omega, t) \leq Z(T^{n_k}\omega) + (\frac{1}{\delta} - 1)Z(T^{n_k}\omega) \leq \frac{1}{\delta} Z(T^{n_k}\omega),$$

which proves conclusion (1) of the proposition.

Since $\varphi \in C^\alpha_+(M)$, there exists $t > 0$ such that $\varphi \in \mathcal{C}(t)$. By Lemma 6.3, we know that $\mathcal{L}_n(\omega)\varphi \subset \mathcal{C}(\delta q_n(\omega, t))$ for every $\omega \in \Omega$ and every $n > 0$. Conclusion (2) of the proposition now follows from conclusion (1). □

6.2. **The random dynamical system generated by $\hat{\mathcal{L}}_{\phi,\gamma}$.** We now show that the random dynamical system generated by $\hat{\mathcal{L}} = \hat{\mathcal{L}}_{\phi,\gamma}$ is Lipschitz. For this claim to make sense, the first thing to do is to specify the family of metric spaces $\{(X_\omega, d_\omega)\}_{\omega\in\Omega}$.

Let $X$ be space Banach space $(C(M), \|\cdot\|_\infty)$. Let $\Omega_1$ be the set as in Proposition 4. Without loss of generality, we can assume that $\Omega_1 = \Omega$ and so that the function $Z$ is strictly positive everywhere on $\Omega$. Recall that $\mathbb{1}$ is the characteristic function on $M$. Clearly, we have that $\mathbb{1} \in \mathcal{C}(t)$ for every $t \geq 0$. Let as before $\mathcal{C}_{\mathbb{1}}(t)$ be the component of $\mathcal{C}(t)$ containing $\mathbb{1}$. Define

$$X_\omega = \{\varphi \in \mathcal{C}_{\mathbb{1}}(Z(\omega)) : \|\varphi\|_\infty = 1\} \qquad \text{for } \omega \in \Omega.$$

Denote by $\theta_\omega$ the Hilbert metric of $\mathcal{C}(Z(\omega))$. By Part (3) of Corollary 2, the metric space $(X_\omega, \theta_\omega|_{X_\omega})$ is complete. We will simply write $\theta_\omega$ instead of $\theta_\omega|_{X_\omega}$. By Lemma 6.3, we have $\mathcal{L}(\omega)\mathcal{C}_{\mathbb{1}}(Z(\omega)) \subseteq \mathcal{C}_{\mathbb{1}}(Z(T\omega))$, and so $\hat{\mathcal{L}}(\omega)X_\omega \subseteq X_{T\omega}$.

**Lemma 6.4.** *The random dynamical system generated by $\hat{\mathcal{L}}$ is Lipschitz and satisfies assumptions A1 and A2.*

*Proof.* Fix $\omega \in \Omega$. By Lemma 5.2, the diameter of $\mathcal{C}(\delta Z(T\omega))$ computed with respect to the metric $\theta_{T\omega}$ is bounded above by the measurable function

$$D(\omega) := 2\log\frac{1+\delta}{1-\delta} + 2\delta\,(\operatorname{diam} M)^{\alpha}\, Z(T\omega).$$

Proposition 2 then implies that $\hat{\mathcal{L}}(\omega) : (X_\omega, \theta_\omega) \to (X_{T\omega}, \theta_{T\omega})$ is a strict contraction with Lipschitz constant $\rho(\omega) \leq \tanh(D(\omega)/4)$. This proves that $\hat{\mathcal{L}}$ generates a Lipschitz random dynamical system.

Now, for every $\omega \in \Omega$ and every $n > 0$, define

$$\tau_n(\omega) = \sum_{j=0}^{n-1} \log\left(1 - e^{-D(T^j\omega)}\right).$$

It is not hard to see that $\{\tau_n\}_{n>0}$ and $\{\tau_n \circ T^{-n})\}_{n>0}$ are additive processes with respect to $T$ and $T^{-1}$, respectively. Moreover, we have $\tau_1 < 0$, because $Z(\omega) < +\infty$ for $\omega \in \Omega$. It follows that we can apply the Subadditivite Ergodic Theorem [14] to the processes $\{\tau_n\}_{n>0}$ and $\{\tau_n \circ T^{-n})\}_{n>0}$. Let

$$\xi = \int_\Omega \tau_1(\omega)\mathrm{d}\mathbb{P}(\omega) = \int_\Omega \log\left(1 - e^{-D(\omega)}\right)\mathrm{d}\mathbb{P}(\omega) \in [-\infty, 0).$$

Since $T$ and $T^{-1}$ are ergodic, the Subadditive Ergodic Theorem implies that

$$\begin{aligned}
\lim_{n\to+\infty} \frac{\tau_n(T^{-n}\omega)}{n} &= \lim_{n\to+\infty} \frac{\tau_n(\omega)}{n} \\
&= \inf_{n>0} \frac{1}{n} \int_\Omega \tau_n(\omega)\mathrm{d}\mathbb{P}(\omega) \\
&\leq \int_\Omega \tau_1(\omega)\mathrm{d}\mathbb{P}(\omega) \\
&= \xi \qquad \text{for } \mathbb{P}\text{-a.e. } \omega \in \Omega.
\end{aligned} \tag{7}$$

From the inequality $\tanh(t/4) \leq (1 - e^{-t})$ for $t \geq 0$, it follows immediately that

$$\log \rho_n(\omega) \leq \sum_{j=0}^{n-1} \log\left(\tanh\left(\frac{D(T^j\omega)}{4}\right)\right) \leq \tau_n(\omega).$$

This together (7) implies that for $\mathbb{P}$-a.e. $\omega \in \Omega$, we have

$$\limsup_{n\to+\infty} \frac{1}{n}\log\rho_n(\omega) \leq \xi \qquad \text{and} \qquad \limsup_{n\to+\infty} \frac{1}{n}\log\rho_n(T^{-n}\omega) \leq \xi,$$

which is assumption A1.

To complete the proof, we prove that assumption A2 is satisfied with $x_0 \equiv \mathbb{1}$. Note that $\mathbb{1} \in X_\omega$ for every $\omega \in \Omega$. By Lemma 5.2, we have $\omega \to \theta_\omega(\mathbb{1}, \hat{\mathcal{L}}(T^{-1}\omega)\mathbb{1}) \leq D(\omega)$. Since $Z$ is tempered, also $D$ is tempered as can be shown using Lemma 4.3. Thus we see that $\omega \to \theta_\omega(\mathbb{1}, \hat{\mathcal{L}}(T^{-1}\omega)\mathbb{1})$ is tempered as well. □

We are now going to discuss assumption A3. Note that A3 is not needed in the random fixed point theorem 4.4 proper, but rather in Corollary 1 to prove the measurability of the fixed point. In the present situation, this required proving the measurability of $\hat{\mathcal{L}}$ as a mapping from $\Omega \times C(M)$ to $C(M)$, and we would obtain the measurability of the fixed point $Z$ as a function from $\Omega$ to $C(M)$. All of this though would make sense only if we had measurable structures on $C(M)$, which we are not going to implement. Rather, we will demonstrate the 'measurability of the fixed point' in the sense that $Z(\omega)(x)$ is

measurable as a function from $\Omega \times M$ to $\mathbb{R}$. The property proved in Lemma 6.5 together with the conclusion of Lemma 6.6 will be sufficient for this purpose.

**Lemma 6.5.** *Let $\gamma$ be a continuous random potential on $M$. Then the function $(\omega, x) \mapsto (\mathcal{L}_{\phi,\gamma}(\omega)\varphi(\omega,\cdot))(x)$ is measurable for every measurable function $\varphi : \Omega \times M \to \mathbb{R}$ such that $\varphi(\omega,\cdot) \in C(M)$ for every $\omega \in \Omega$.*

*Proof.* Let $\hat{\gamma}(\omega, x) = -\log|\det D_x\phi(\omega,\cdot)|$ for every $\omega \in \Omega$ and $x \in M$. Also, for every pair of functions $\varphi$ and $\gamma$ as in the hypotheses of the lemma, define

$$\tilde{\varphi}(\omega, x) = e^{\gamma(\omega,x)-\hat{\gamma}(\omega,x)}\varphi(x,\omega) \qquad \text{for } \omega \in \Omega \text{ and } x \in M.$$

It is easy to see that $\hat{\gamma}$ and $\tilde{\varphi}$ are both measurable and $\hat{\gamma}(\omega,\cdot), \tilde{\varphi}(\omega,\cdot) \in C(M)$ for every $\omega \in \Omega$. In other words, $\hat{\gamma}$ and $\tilde{\varphi}$ satisfy the hypotheses of the lemma. Next, we observe that

$$\mathcal{L}_{\phi,\gamma}(\omega)\varphi = \mathcal{L}_{\phi,\hat{\gamma}}(\omega)\tilde{\varphi}.$$

If we write $\tilde{\varphi} = \tilde{\varphi}_+ - \tilde{\varphi}_-$ with $\tilde{\varphi}_\pm = (|\tilde{\varphi}| \pm \varphi)/2 \geq 0$, then $\mathcal{L}_{\phi,\hat{\gamma}}(\omega)\tilde{\varphi} = \mathcal{L}_{\phi,\hat{\gamma}}(\omega)\tilde{\varphi}_+ - \mathcal{L}_{\phi,\hat{\gamma}}(\omega)\tilde{\varphi}_-$ by the linearity of $\mathcal{L}_{\phi,\hat{\gamma}}(\omega)$. To prove the lemma, we can therefore assume without loss of generality that $\gamma = \hat{\gamma}$ and $\varphi \geq 0$.

Let $\varphi$ be as in the hypotheses of the lemma, and suppose that $\varphi \geq 0$. Define

$$Q_\omega(B) = \int_M I_B(x)(\mathcal{L}_{\phi,\hat{\gamma}}(\omega)\varphi(\omega,\cdot))(x)\mathrm{d}m(x) \qquad \text{for } B \in \mathcal{B}.$$

Since $\mathcal{L}_{\phi,\hat{\gamma}}(\omega)$ is the standard Frobenius–Perron operator of the map $\phi(\omega,\cdot)$, it follows that $Q_\omega(B) = \int_M I_B(\phi(\omega,x))\varphi(\omega,x)\mathrm{d}m(x)$. Since the function $(\omega, x) \mapsto I_B(\phi(\omega,x))\varphi(\omega,x)$ is non-negative and measurable, $\omega \mapsto Q_\omega(B)$ is measurable for every $B \in \mathcal{B}$ by Fubini's Theorem. Let $B_r(x)$ be the ball of $M$ with center at $x \in M$ and of radius $r > 0$. We claim that

$$\lim_{r\to 0^+} \frac{Q_\omega(B_r(x))}{m(B_r(x))} = (\mathcal{L}_{\phi,\hat{\gamma}}(\omega)\varphi(\omega,\cdot))(x) \qquad \text{for } \omega \in \Omega \text{ and } x \in M.$$

Indeed, since $\mathcal{L}_{\phi,\hat{\gamma}}(\omega)\varphi(\omega,\cdot) \in C(M)$, for every $\epsilon > 0$, there exists $r > 0$ such that $|(\mathcal{L}_{\hat{\gamma}}(\omega)\varphi(\omega,\cdot))(x) - (\mathcal{L}_{\phi,\hat{\gamma}}(\omega)\varphi(\omega,\cdot))(y)| < \epsilon$ provided that $y \in B_r(x)$. Hence,

$$\begin{aligned}
&\left|(\mathcal{L}_{\phi,\hat{\gamma}}(\omega)\varphi(\omega,\cdot))(x) - \frac{1}{m(B_r(x))}\int_M I_B(x)(\mathcal{L}_{\phi,\hat{\gamma}}(\omega)\varphi(\omega,\cdot))(y)\mathrm{d}m(y)\right| \\
&= \frac{1}{m(B_r(x))}\left|\int_M I_B(x)\left((\mathcal{L}_{\phi,\hat{\gamma}}(\omega)\varphi(\omega,\cdot))(x) - (\mathcal{L}_{\phi,\hat{\gamma}}(\omega)\varphi(\omega,\cdot))(y)\right)\mathrm{d}m(y)\right| \\
&\leq \frac{1}{m(B_r(x))}\int_M I_B(x)\left|(\mathcal{L}_{\phi,\hat{\gamma}}(\omega)\varphi(\omega,\cdot))(x) - (\mathcal{L}_{\phi,\hat{\gamma}}(\omega)\varphi(\omega,\cdot))(y)\right|\mathrm{d}m(y) \\
&< \epsilon.
\end{aligned}$$

Now, since $\omega \mapsto Q_\omega(B_r(x))/m(B_r(x))$ is measurable for every $r > 0$ and $x \in M$, the function $\omega \mapsto (\mathcal{L}_{\phi,\hat{\gamma}}(\omega)\varphi(\omega,\cdot))(x)$ is measurable for every $x \in M$. Finally, the joint-measurability of $(\omega, x) \mapsto (\mathcal{L}_{\phi,\hat{\gamma}}(\omega)\varphi(\omega,\cdot))(x)$ follows from the continuity of the function $\mathcal{L}_{\phi,\hat{\gamma}}(\omega)\varphi(\omega,\cdot)$ for every $\omega \in \Omega$ (see e.g. [9, Theorem 4.2.2]). □

**Lemma 6.6.** *Suppose that $\varphi : \Omega \times M \to \mathbb{R}$ is a measurable function such that $\varphi(\omega,\cdot) \in C(M)$ for every $\omega \in \Omega$. Then the function $\omega \mapsto \|\varphi(\omega,\cdot)\|_k$ is measurable for $k \in \{1, \infty\}$. Further, if $\varphi(\omega,\cdot) \in C^\alpha(M)$ for every $\omega \in \Omega$, then $\omega \mapsto |\varphi|_\alpha$ is measurable.*

*Proof.* The case $k = 1$ follows directly from Fubini's Theorem. For the case $k = \infty$, note that the manifold $M$ admits a dense sequence $\{x_n\}_{n>0}$ of points. Since $\varphi$ is measurable, so is each function $\omega \mapsto |\varphi(\omega, x_n)|$. But since $\varphi(\omega, \cdot)$ is continuous for $\omega \in \Omega$, we have $\|\varphi(\omega, \cdot)\|_\infty = \sup_{n>0}\{L_n(\omega)\}$ for $\omega \in \Omega$. Thus $\omega \mapsto \|\varphi(\omega, \cdot)\|_\infty$ is the supremum of measurable functions and hence measurable. The case of the Hölder norm $|\cdot|_\alpha$ works analogously. □

### 6.3. **Random fixed point.**

**Proposition 6.** *Let $\phi$ be a random local diffeomorphism on a manifold $M$, and let $\gamma$ be a random potential on $M$ of class $0 < \alpha \leq 1$. Suppose that $\phi$ and $\gamma$ satisfy conditions C1 and C2. Let $\xi \in [-\infty, 0)$ be the constant as in the proof of Lemma 6.4. Then there exist a full $\mathbb{P}$-measure $T$-invariant set $\Omega_0 \subset \Omega$, a measurable function $h : \Omega \times M \to \mathbb{R}$ and a measurable function $\Lambda : \Omega \to (0, +\infty)$ such that for every $\omega \in \Omega_0$, we have*

*(1) $h(\omega, \cdot) \in C^\alpha_+(M)$ and $\|h(\omega, \cdot)\|_\infty = 1$,*
*(2) $\limsup_{n\to+\infty} n^{-1} \log \|h(\omega, \cdot) - \hat{\mathcal{L}}_n(T^{-n}\omega)\mathbb{1}\|_\infty \leq \xi$,*
*(3) $\mathcal{L}(\omega)h(\omega, \cdot) = \Lambda(\omega)h(T\omega, \cdot)$,*
*(4) $\limsup_{n\to+\infty} n^{-1} \log \|h(T^n\omega, \cdot) - \hat{\mathcal{L}}_n(\omega)\varphi\|_\infty \leq \xi$ for $\varphi \in C^\alpha_+(M)$.*

*Proof.* Lemma 6.4 allows to apply Theorem 4.4 to the random dynamical system $\hat{\mathcal{L}}$. Thus, there exist a full $\mathbb{P}$-measure $T$-invariant set $\Omega_0 \subset \Omega$ and a function $h : \Omega \times M \to \mathbb{R}$ such that for every $\omega \in \Omega_0$, we have

(i) $h(\omega, \cdot) \in X_\omega$,
(ii) $\limsup_{n\to+\infty} n^{-1} \log \theta_\omega(h(\omega, \cdot), \hat{\mathcal{L}}_n(T^{-n}\omega)\mathbb{1}) \leq \xi$,
(iii) $\hat{\mathcal{L}}_n(\omega)h(\omega, \cdot) = h(T^n\omega, \cdot)$ for every $n \in \mathbb{Z}$,
(iv) $\limsup_{n\to+\infty} n^{-1} \log \theta_{T^n\omega}(h(T^n\omega, \cdot), \hat{\mathcal{L}}_n(\omega)\varphi) \leq \xi$ for $\varphi \in X_\omega$.

Conclusions (1)-(4) of the proposition are straightforward consequences of statements (i)-(iv) above. Indeed, conclusion (1) is just statement (i). Conclusion (2) follows from (ii) and the second part of Corollary 2. To prove conclusion (3), we define $\Lambda(\omega) = \|\mathcal{L}(\omega)h(\omega, \cdot)\|_\infty$ for every $\omega \in \Omega$, and then use (iii). Conclusion (4) for $\varphi \in X_\omega$ follows from (iv) and the second part of Corollary 2. Conclusion (4) is then extended to every $\varphi \in C^\alpha_+(M)$ by using the second part of Proposition 5.

To complete the proof, we have to show that $h$ and $\Lambda$ are measurable. Let $g_n(\omega, x) = (\mathcal{L}(T^{-n}\omega)\mathbb{1})(x)$ for $\omega \in \Omega$, $x \in M$ and $n > 0$. By Lemma 6.5, each function $g_n$ is measurable. Using Lemma 6.6, we obtain that also $\omega \mapsto \|g_n(\omega, \cdot)\|_\infty$ is measurable. Hence, $h_n(\omega, x) := g_n(\omega, x)/\|g_n(\omega, \cdot)\|_\infty$ is measurable as well. Now, by conclusion (2), the sequence $\{h_n(\omega, \cdot)\}_{n>0}$ converges uniformly to $h(\omega, \cdot)$ for $\omega \in \Omega_0$. Therefore, $h_n$ converges pointwise to $h$ on $\Omega_0 \times M$, and so the restriction $h|_{\Omega_0\times M}$ is measurable. By construction of $h$ (see the proof of Theorem 4.4) and our choice $x_0(\omega, \cdot) = \mathbb{1}$, it follows that $h(\omega, \cdot) = \mathbb{1}$ for $\Omega \setminus \Omega_0$. We conclude that $h$ is measurable on the entire set $\Omega \times M$. Note that this implies that conclusion (1) holds actually for every $\omega \in \Omega$.

It remains to show that $\Lambda$ is measurable. Since $h$ is measurable, Lemma 6.5 implies that $(\omega, x) \mapsto (\mathcal{L}(\omega)h(\omega, \cdot))(x)$ is measurable. Moreover, since $h(\omega, \cdot) \in C(M)$ is continuous for every $\omega \in \Omega$, also $\mathcal{L}(\omega)h(\omega, \cdot)$ is continuous for every $\omega \in \Omega$. This means that $(\omega, x) \mapsto (\mathcal{L}(\omega)h(\omega, \cdot))(x)$ satisfies the hypotheses of Lemma 6.6, and so $\Lambda$ is measurable. □

In the next lemma, we prove the uniqueness of $h$ and $\Lambda$.

**Lemma 6.7.** *Suppose that $h, \tilde{h} : \Omega \times M \to \mathbb{R}$ and $\Lambda, \tilde{\Lambda} : \Omega \to (0, +\infty)$ are measurable functions that satisfy conclusions (1) and (3) of Proposition 6 on full $\mathbb{P}$-measure sets $\Omega_0, \tilde{\Omega}_0 \subset \Omega$. If $h$ also satisfies conclusion (4) of Proposition 6, then $\tilde{h}(\omega, \cdot) = h(\omega, \cdot)$ and $\tilde{\Lambda}(\omega) = \Lambda(\omega)$ for $\mathbb{P}$-a.e. $\omega \in \Omega$.*

*Proof.* Suppose that $\tilde{h}(\omega, \cdot)$ and $h(\omega, \cdot)$ differ on a positive $\mathbb{P}$-measure subset of $\Omega_0 \cap \tilde{\Omega}_0$. Since $h(\omega, \cdot)$ and $\tilde{h}(\omega, \cdot)$ are both continuous for every $\omega \in \Omega_0 \cap \tilde{\Omega}_0$, we must have

$$\mathbb{P}\left(\left\{\omega \in \Omega_0 \cap \tilde{\Omega}_0 : \left\|h(\omega, \cdot) - \tilde{h}(\omega, \cdot)\right\|_\infty > 0\right\}\right) > 0.$$

It is then not difficult to see that there exists $\epsilon > 0$ such that the set

$$\hat{\Omega} = \left\{\omega \in \Omega_0 \cap \tilde{\Omega}_0 : \left\|h(\omega, \cdot) - \tilde{h}(\omega, \cdot)\right\|_\infty > \epsilon\right\},$$

has positive $\mathbb{P}$-measure. Now, since $\Omega_0 \cap \tilde{\Omega}_0$ is a full $\mathbb{P}$-measure set, and $T$ is ergodic, we can find $\omega \in \Omega_0 \cap \tilde{\Omega}_0$ and an increasing sequence $\{n_j\}_{j>0}$ such that $T^{n_j}\omega \in \hat{\Omega}$ for every $j > 0$. On the other hand, since $\tilde{h}$ satisfies conclusions (1) and (3) of Proposition 6 on $\tilde{\Omega}_0$, conclusion (4) of Proposition 6 implies that the sequence $\|h(T^{n_j}\omega, \cdot) - \tilde{h}(T^{n_j}\omega, \cdot)\|_\infty$ vanishes as $j \to +\infty$. Putting all together, we obtain the contradiction

$$\epsilon < \left\|h(T^{n_j}\omega, \cdot) - \tilde{h}(T^{n_j}\omega, \cdot)\right\|_\infty < \epsilon,$$

for $j$ sufficiently large. Hence, $\tilde{h}(\omega, \cdot) = h(\omega, \cdot)$ for $\mathbb{P}$-a.e. $\omega \in \Omega$.

The uniqueness of $\Lambda$ is an obvious consequence of the uniqueness of $h$. Indeed, from conclusion (2) of Proposition 6, it follows that $\Lambda(\omega) = \|\mathcal{L}(\omega)h(\omega, \cdot)\|_\infty = \|\mathcal{L}(\omega)\tilde{h}(\omega, \cdot)\|_\infty = \tilde{\Lambda}(\omega)$ for every $\omega \in \Omega_0 \cap \tilde{\Omega}_0$. □

6.4. **Random measures.** Let $\Omega_0, h, \Lambda$ and $\xi$ be as in Proposition 6, and define $\Lambda_n(\omega) = \Lambda(T^{n-1}\omega) \cdots \Lambda(\omega)$ for $\omega \in \Omega$ and every $n > 0$.

The proof of the next proposition, which we will skip, closely follows the proof of [10, Theorem 2], which is based on a deterministic version of that theorem originally due to Birkhoff [5, Lemma 3].

**Proposition 7.** *Under the same hypotheses of Proposition 6, there exists a family $\{\nu_\omega\}_{\omega \in \Omega_0}$ of positive functions on $C^\alpha_+(M)$ such that for $\mathbb{P}$-a.e. $\omega \in \Omega_0$, we have*

$$\limsup_{n \to +\infty} \frac{1}{n} \log \left\| \frac{1}{\Lambda_n(\omega)} \mathcal{L}_n(\omega)\varphi - \nu_\omega(\varphi) h(T^n\omega, \cdot) \right\|_\infty \leq \xi \qquad \text{for } \varphi \in C^\alpha_+(M). \tag{8}$$

**Lemma 6.8.** *Let $\{\nu_\omega\}_{\omega \in \Omega_0}$ be the family of functions as in Proposition 7. Then*

*(1) $\nu_\omega(h(\omega, \cdot)) = 1$ for $\omega \in \Omega_0$,*
*(2) $\nu_{T\omega}(\mathcal{L}(\omega)\varphi) = \Lambda(\omega)\nu_\omega(\varphi)$ for $\omega \in \Omega_0$ and $\varphi \in C^\alpha_+(M)$,*
*(3) $\omega \mapsto \nu_\omega(\varphi)$ is measurable for every $\varphi \in C^\alpha_+(M)$.*

*Proof.* Fix $\omega \in \Omega_0$. To prove conclusion (1), use the fact that $\lambda_n(\omega, h(\omega, \cdot)) = 1$, and then pass to the limit as $n \to +\infty$. Next, observe that $\mathcal{L}_{n+1}(\omega)\varphi = \mathcal{L}_n(T\omega)(\mathcal{L}(\omega)\varphi)$. Then using the definition of $\lambda_n$ and $\Lambda_n$, we deduce that the largest $A > 0$ for which $A\lambda_{n+1}(\omega)\Lambda_{n+1}(\omega)h(T^{n+1}\omega, \cdot)$ is a lower bound of $\mathcal{L}_{n+1}(\omega)$ is simultaneously equal to $\lambda_{n+1}(\omega, \varphi)$ and to $\lambda_n(T\omega, \mathcal{L}(\omega)\varphi)/\Lambda(\omega)$. By passing to the limit as $n \to +\infty$, we obtain conclusion (2). This argument makes sense, because the invariance of $\Omega_0$ guarantees that $T\omega \in \Omega_0$.

Proposition 7 implies that $\nu_\omega(\varphi) = \lim_{n \to +\infty} \|\mathcal{L}_n(\omega)\varphi\|_\infty / \Lambda_n(\omega)$. Since we know that $\Lambda_n$ is measurable by Proposition 6, to prove conclusion (3), it suffices to prove that

$\omega \mapsto \|\mathcal{L}_n(\omega)\varphi\|_\infty$ is also measurable. But this is so because of Lemma 6.6, and the fact $\omega \mapsto \mathcal{L}_n(\omega)(\varphi)$ is measurable by Lemma 6.5. □

**Proposition 8.** *Each function $\nu_\omega$ with $\omega \in \Omega_0$ extends uniquely to a positive bounded linear functional on the Banach space $(C(M), \|\cdot\|_\infty)$. Furthermore, Proposition 7 extends to all $\varphi \in C^\alpha(M)$, and Lemma 6.8 extends to all $\varphi \in C(M)$.*

*Proof.* Fix $\omega \in \Omega_0$. First of all, we prove that $\nu_\omega$ is linear on $C^\alpha_+(M)$, which means that $\nu_\omega(c\varphi) = c\nu_\omega(\varphi)$ for $\varphi \in C^\alpha_+(M)$ and $c > 0$, and $\nu_\omega(\varphi_1 + \varphi_2) = \nu_\omega(\varphi_1) + \nu_\omega(\varphi_2)$ for $\varphi_1, \varphi_2 \in C^\alpha_+(M)$. It follows easily from the definitions of $\lambda_n$ and $\zeta_n$ that $\lambda_n$ is superlinear, i.e.,

- $\lambda_n(\omega, c\varphi) = c\lambda_n(\omega, \varphi)$ for $\varphi \in C^\alpha_+(M)$ and $c > 0$,
- $\lambda_n(\omega, \varphi_1 + \varphi_2) \geq \lambda_n(\omega, \varphi_1) + \lambda_n(\omega, \varphi_2)$ for $\varphi_1, \varphi_2 \in C^\alpha_+(M)$,

and $\zeta_n$ is sublinear, i.e.,

- $\zeta_n(\omega, c\varphi) = c\zeta_n(\omega, \varphi)$ for $\varphi \in C^\alpha_+(M)$ and $c > 0$,
- $\zeta_n(\omega, \varphi_1 + \varphi_2) \leq \zeta_n(\omega, \varphi_1) + \zeta_n(\omega, \varphi_2)$ for $\varphi_1, \varphi_2 \in C^\alpha_+(M)$.

Let $\varphi, \varphi_1, \varphi_2 \in C^\alpha_+(M)$, and let $c > 0$. By definition of $\nu_\omega$, we have

$$c\lambda_n(\omega, \varphi) \leq \nu_\omega(c\varphi) \leq c\zeta_n(\omega, \varphi),$$
$$\lambda_n(\omega, \varphi_1) + \lambda_n(\omega, \varphi_2) \leq \nu_\omega(\varphi_1 + \varphi_2) \leq \zeta_n(\omega, \varphi_1) + \zeta_n(\omega, \varphi_2).$$

Thus, passing to the limit as $n \to +\infty$, we obtain

$$\nu_\omega(c\varphi) = c\nu_\omega(\varphi),$$
$$\nu_\omega(\varphi_1 + \varphi_2) = \nu_\omega(\varphi_1) + \nu_\omega(\varphi_2).$$

Next, we extend $\nu_\omega$ from $C^\alpha_+(M)$ to $C(M)$, and we do it in two steps: first, we extend $\nu_\omega$ from $C^\alpha_+(M)$ to $C^\alpha(M)$ and then from $C^\alpha(M)$ to $C(M)$. The argument is standard, but we include it anyway for the sake of completeness.

Step 1. Observe that given $\varphi \in C^\alpha(M)$, we can write $\varphi = \varphi^+ - \varphi^-$, where $\varphi^\pm = (|\varphi| \pm \varphi)/2$. We extend the functional $\nu_\omega$ to the whole space $C^\alpha(M)$, by defining

$$\hat{\nu}_\omega(\varphi) = \nu_\omega(\varphi^+ + d) - \nu_\omega(\varphi^- + d) \qquad \text{for every } \varphi \in C^\alpha(M),$$

where $d$ is a positive constant which makes sure that the argument of $\nu_\omega$ is in $C^\alpha_+(M)$. We leave it to the reader to show with the help of the linearity of $\nu_\omega$ on $C^\alpha_+(M)$ that $\hat{\nu}$ does not depend on $d$. We now prove that $\hat{\nu}_\omega$ is linear on $C^\alpha(M)$. Let $\varphi, \varphi_1, \varphi_2 \in C^\alpha(M)$, and let $c \in \mathbb{R}$. To prove that $\hat{\nu}_\omega(c\varphi) = c\hat{\nu}_\omega(\varphi)$, first note that $(c\varphi)^\pm = c\varphi^\pm$ when $c > 0$. Therefore,

$$\begin{aligned}
\hat{\nu}_\omega(c\varphi) &= \nu_\omega((c\varphi)^+ + d) - \nu_\omega((c\varphi)^- + d) \\
&= \nu_\omega\left(c\left(\varphi^+ + \frac{d}{c}\right)\right) - \nu_\omega\left(c\left(\varphi^- + \frac{d}{c}\right)\right) \\
&= c\left(\nu_\omega\left(\varphi^+ + \frac{d}{c}\right) - \nu_\omega\left(\varphi^- + \frac{d}{c}\right)\right) \\
&= c\hat{\nu}_\omega(\varphi).
\end{aligned}$$

The case of $c < 0$ is similar, noting that $(c\varphi)^\pm = -c\varphi^\mp$ in this situation. Since $0^\pm = 0$, we immediately see that $\hat{\nu}_\omega(c\varphi) = 0$ when $c = 0$. Now, let $\psi = \varphi_1 + \varphi_2$. It is easy to check that $\psi^+ - \psi^- = \varphi_1^+ - \varphi_1^- + \varphi_2^+ - \varphi_2^-$, or equivalently that

$$\psi^+ + \varphi_1^- + \varphi_2^- = \psi^- + \varphi_1^+ + \varphi_2^+.$$

Adding a constant $d > 0$ to each term of the previous equality, all the functions become elements of $C^\alpha_+(M)$. We then apply $\nu_\omega$ and use its linearity on $C^\alpha_+(M)$ to obtain

$$\nu_\omega(\psi^+ + d) + \nu_\omega(\varphi_1^- + d) + \nu_\omega(\varphi_2^- + d) \\ = \nu_\omega(\psi^- + d) + \nu_\omega(\varphi_1^+ + d) + \nu_\omega(\varphi_2^+ + d)$$

Rearranging the terms and using the definition of $\hat{\nu}$, we finally get $\hat{\nu}_\omega(\psi) = \hat{\nu}_\omega(\varphi_1) + \hat{\nu}_\omega(\varphi_2)$, showing that $\hat{\nu}_\omega$ is linear on $C^\alpha(M)$.

Note that the extension $\omega \mapsto \hat{\nu}_\omega(\varphi)$ is measurable for every $\varphi \in C^\alpha(M)$, because so are $\omega \mapsto \nu_\omega(\varphi^\pm + d)$ by conclusion (3) of Lemma 6.8.

We now prove that $\hat{\nu}_\omega$ is bounded on $C^\alpha(M)$ endowed with the norm $\|\cdot\|_\infty$. First, suppose that $\varphi \in C^\alpha(M)$ with $0 \neq \varphi \geq 0$. Then

$$\begin{aligned}\hat{\nu}_\omega(\varphi) &= \nu_\omega(\varphi + d) - \nu_\omega(d) \\ &\leq \|\varphi\|_\infty \left(\nu_\omega\left(1 + \frac{d}{\|\varphi\|_\infty}\right) - \nu_\omega\left(\frac{d}{\|\varphi\|_\infty}\right)\right) \\ &= \|\varphi\|_\infty \hat{\nu}_\omega(1).\end{aligned}$$

This conclusion remains trivially true when $\varphi = 0$. Now, let $\varphi \in C^\alpha(M)$. By the previous conclusion and the obvious fact that $\|\varphi^\pm\|_\infty \leq \|\varphi\|_\infty$, we obtain

$$\begin{aligned}|\hat{\nu}_\omega(\varphi)| &\leq |\hat{\nu}_\omega(\varphi^+ - \varphi^-)| \\ &\leq |\hat{\nu}_\omega(\varphi^+)| + |\hat{\nu}_\omega(\varphi^-)| \\ &\leq 2\|\varphi\|_\infty \hat{\nu}_\omega(1).\end{aligned}$$

Therefore, the functional $\hat{\nu}_\omega$ is bounded. In the following, we drop the hat on $\hat{\nu}$.

Step 2. Since the Banach space $(C(M), \|\cdot\|_\infty)$ is the completion of the real normed space $(C^\alpha(M), \|\cdot\|_\infty)$ by the Stone-Weierstrass Theorem, the functional $\nu_\omega$ extends uniquely to a bounded linear functional on $(C(M), \|\cdot\|_\infty)$. The extended functional, which we keep denoting by $\nu_\omega$, is positive. Indeed, suppose that $\varphi \in C(M)$ with $\varphi \geq 0$. Let $\{\varphi_n\}_{n>0}$ be a sequence of elements of $C^\alpha(M)$ such that $\varphi_n \to \varphi$ uniformly on $M$ as $n \to +\infty$. Next, define

$$\psi_n = \varphi_n + \left(\|\varphi - \varphi_n\|_\infty + \frac{1}{n}\right)\mathbb{1} \qquad \text{for } n > 0.$$

It is easy to see that $\psi_n \in C^\alpha_+(M)$ and $\psi_n \to \varphi$ uniformly on $M$ as $n \to +\infty$. Since $\nu_\omega$ is positive on $C^\alpha_+(M)$ and continuous, it follows at once that $\nu_\omega(\varphi) = \lim_{n\to+\infty} \nu_\omega(\psi_n) \geq 0$.

We now prove the measurability of $\omega \mapsto \nu_\omega(\varphi)$ for every $\varphi \in C(M)$. So let $\varphi \in C(M)$. From Step 2, we know that there exists a sequence $\varphi_n \in C^\alpha(M)$ such that $\nu_\omega(\varphi) = \lim_{n\to+\infty} \nu_\omega(\varphi_n)$, and from Step 1, we know that $\omega \mapsto \nu_\omega(\varphi_n)$ is measurable for every $n > 0$. We can then conclude that $\omega \mapsto \nu_\omega(\varphi)$ is measurable as well.

At this point, it is not hard to see that by the linearity of the operator $\mathcal{L}(\omega)$ and the extended functional $\nu_\omega$, Proposition 7 extends to all $\varphi \in C^\alpha(M)$, and Lemma 6.8 extends to all $\varphi \in C(M)$. □

**Proposition 9.** *The family of functionals $\{\nu_\omega\}_{\omega\in\Omega_0}$ is the unique family of positive bounded linear functionals on $(C(M), \|\cdot\|_\infty)$ satisfying conclusions (1)-(3) of Lemma 6.8 for every $\varphi \in C(M)$.*

*Proof.* We argue by contradiction. Suppose that the exist another full $\mathbb{P}$-measure $T$-invariant set $\Omega_1 \subset \Omega$ and another family of positive bounded linear functionals $\{\tilde{\nu}_\omega\}_{\omega\in\tilde{\Omega}_0}$

on $C(M)$ satisfying conclusions (1)-(3) of of Lemma 6.8 for every $\varphi \in C(M)$, and furthermore suppose that if

$$\Omega_1 = \{\omega \in \Omega_0 \cap \tilde{\Omega}_0 : d(\tilde{\nu}_\omega, \nu_\omega) > 0\},$$

where $d(\tilde{\nu}_\omega, \nu_\omega) = \sup\{|\tilde{\nu}_\omega(\varphi) - \nu_\omega(\varphi)| : \varphi \in C(M) \text{ and } 0 \leq \varphi \leq 1\}$, then $\mathbb{P}(\Omega_1) > 0$. We observe that the measurability of the set $\Omega_1$ is a consequence of the measurability of the function $\omega \mapsto d(\tilde{\nu}_\omega, \nu_\omega)$, which can be proved by an argument exploiting the separability of $(C(M), \|\cdot\|_\infty)$ similarly to the argument used in the proof of Lemma 6.6.

Since $h(\omega, \cdot) \in C^\alpha_+(M)$, the norm $\|1/h(\omega, \cdot)\|_\infty$ is finite for $\omega \in \Omega$. This implies immediately that $1 \leq \|1/h(\omega, \cdot)\|_\infty h(\omega, \cdot)$, and therefore $\tilde{\nu}_\omega(1) \leq \|1/h(\omega, \cdot)\|_\infty$ for $\omega \in \Omega_1$. Moreover, $\omega \mapsto 1/h(\omega, \cdot)$ satisfies the hypotheses of Lemma 6.6, and so $\omega \mapsto \|1/h(\omega, \cdot)\|_\infty$ is measurable. As a consequence, we see that there exists a bound $H > 0$ and a positive $\mathbb{P}$-measure subset $\hat{\Omega} \subset \Omega_1$ such that $\|1/h(\omega, \cdot)\|_\infty < H$ for every $\omega \in \hat{\Omega}$. Hence, $\tilde{\nu}_\omega(1) \leq H$ for $\omega \in \hat{\Omega}$.

Now let $\omega \in \Omega_1$. Since $C^\alpha(M)$ is dense in $C(M)$, the fact that $\mathbb{P}(\Omega_1) > 0$ implies that we can find $\varphi \in C^\alpha(M)$ such that $\tilde{\nu}_\omega(\varphi) \neq \nu_\omega(\varphi)$. Write $\varphi = \nu_\omega(\varphi)h(\omega, \cdot) + \psi$ with $\psi = \varphi - \nu_\omega(\varphi)h(\omega, \cdot) \in C^\alpha(M)$. Since $\tilde{\nu}_\omega(h(\omega, \cdot)) = 1$, it follows that $\tilde{\nu}_\omega(\varphi) = \nu_\omega(\varphi) + \tilde{\nu}_\omega(\psi)$. But $\tilde{\nu}_\omega(\varphi) \neq \nu_\omega(\varphi)$, and hence $\tilde{\nu}_\omega(\psi) \neq 0$. On the other hand, $\nu_\omega(\psi) = 0$.

Using the part of Proposition 8 which extends conclusion (2) of Lemma 6.8 to all $C(M)$, we obtain

$$\begin{aligned} |\tilde{\nu}_\omega(\psi)| &= \left| \frac{1}{\Lambda_n(\omega)} \mathcal{L}^*_n(\omega) \tilde{\nu}_{T^n\omega}(\psi) \right| \\ &= \left| \tilde{\nu}_{T^n\omega} \left( \frac{1}{\Lambda_n(\omega)} \mathcal{L}_n(\omega)\psi \right) \right| \\ &\leq \tilde{\nu}_{T^n\omega}(1) \left\| \frac{1}{\Lambda_n(\omega)} \mathcal{L}_n(\omega)\psi \right\|_\infty \qquad \text{for } n > 0. \end{aligned} \tag{9}$$

Since $T$ is ergodic, the point $\omega \in \Omega_1$ can be chosen in such a way that there exists a divergent subsequence $\{n_k\}_{k>0}$ such that $T^{n_k}\omega \in \hat{\Omega}$ for every $k > 0$. This fact together with (9) implies that

$$|\tilde{\nu}_\omega(\psi)| \leq H \left\| \frac{1}{\Lambda_{n_k}(\omega)} \mathcal{L}_{n_k}(\omega)\psi \right\|_\infty \qquad \text{for } k > 0.$$

Passing to the limit as $k \to +\infty$, and using the part of Proposition 8 which extends conclusion (2) of Proposition 6 to the entire set $C^\alpha(M)$, we obtain $\tilde{\nu}_\omega(\psi) = 0$. But this is a contradiction, and so $\tilde{\nu}_\omega = \nu_\omega$ for $\mathbb{P}$-a.e. $\omega \in \Omega_0 \cap \Omega_1$. □

In the next corollary, we will use the Riesz Representation Theorem to prove that each functional $\nu_\omega$ is a regular Borel Measure. Furthermore, we recall the properties of $\nu_\omega$ obtained previously, and prove that family $\{\nu_\omega\}_{\omega\in\Omega_0}$ has the property of being measurable, i.e., the function $\omega \mapsto \nu_\omega(A)$ is measurable for every $A \in \mathcal{B}$.

**Corollary 3.** *The family functional $\{\nu_\omega\}_{\omega\in\Omega_0}$ has the following properties:*

*(1) $\nu_\omega$ is a regular Borel measure on $M$ for $\omega \in \Omega_0$,*
*(2) $\nu_\omega(h(\omega, \cdot)) = 1$ for $\omega \in \Omega_0$,*
*(3) $\mathcal{L}^*(\omega)\nu_{T\omega} = \Lambda(\omega)\nu_\omega$ for $\omega \in \Omega_0$,*
*(4) $\omega \mapsto \nu_\omega(A)$ is measurable for every $A \in \mathcal{B}$.*

*Proof.* Property (1) follows from Proposition 8 and the Riesz Representation Theorem [19]. Property (2) is precisely conclusion (1) of Lemma 6.8. Property (3) is just the

extension of conclusion (2) of Lemma 6.8 to all $\varphi \in C(M)$ (see Proposition 8), once we have noticed that $\mathcal{L}^*(\omega)\nu_{T\omega}(\varphi) = \nu_{T\omega}(\mathcal{L}(\omega)\varphi)$ for every $\varphi \in C(M)$.

It remains to prove Property (4). First of all, we observe that since the manifold $M$ is compact and connected, the topology of $M$ has a countable base, and the space $(C(M), \|\cdot\|_\infty)$ is separable. Let $V$ be an open subset of $M$. By the Riesz Representation Theorem (see [19, Theorem 2.14]), we have $\nu_\omega(V) = \sup\{\nu_\omega(\varphi) : 0 \leq \varphi \leq I_V\}$, where $I_V$ is the characteristic function of $V$. Since $C(M)$ is separable, there exists a sequence of continuous functions $\{\varphi_n\}_{n>0}$ with $0 \leq \varphi_n \leq I_V$ such that $\nu_\omega(V) = \sup_n \nu_\omega(\varphi_n)$. The extension of conclusion (3) of Lemma 6.8 to all $\varphi \in C(M)$ (see Proposition 8) then implies that $\omega \mapsto \nu_\omega(V)$ is measurable. Now, let $A \in \mathcal{B}$. Since $\nu_\omega$ is regular for $\omega \in \Omega_0$ , we have $\nu_\omega(A) = \inf\{\nu_\omega(V) : A \subset V \subset M \text{ and } V \text{ is open}\}$. Since the topology of $M$ has a countable base, we can find a sequence of open subsets $\{V_n\}_{n>0}$ of $M$ such that $\nu_\omega(A) = \inf_{n>0} \nu_\omega(V_n)$. But each $\omega \mapsto \nu_\omega(V_n)$ is measurable, and so $\omega \mapsto \nu_\omega(A)$ is measurable as well. This completes the proof. □

6.5. **Invariant measure.** Let us define a new family $\{\mu_\omega\}_{\omega\in\Omega_o}$ of non-negative set functions on $(M, \mathcal{B})$ by

$$\mu_\omega(A) = \int_A h(\omega, x)\mathrm{d}\nu_\omega(x) \qquad \text{for } \omega \in \Omega_0 \text{ and } A \in \mathcal{B}. \tag{10}$$

From Corollary 3, it follows immediately that $\mu_\omega$ is a regular Borel probability on $M$, and from conclusion (iv) of Corollary 3 and [8, conclusion (i) of Proposition 3.3], it follows that $\omega \mapsto \mu_\omega(A)$ is measurable for every $A \in \mathcal{B}$.

Now, define

$$\mu(B) = \int_\Omega \int_X I_B(\omega, x)\mathrm{d}\mu_\omega(x)\mathrm{d}\mathbb{P}(\omega) \qquad \text{for } B \in \mathcal{F} \otimes \mathcal{B},$$

By using a monotone class argument, one can show that $\mu_\omega$ is a probability on $(\Omega \times M, \mathcal{F}\otimes\mathcal{B})$ whose marginal on $\Omega$ is $\mathbb{P}$ (see [8, conclusion (ii) of Proposition 3.3]). Moreover, $\{\mu_\omega\}_{\omega\in\Omega_0}$ is the disintegration of $\mu$ with respect to $\mathbb{P}$ (see [8, Proposition 3.6]).

**Proposition 10.** *The probability $\mu$ is $\phi$-invariant.*

*Proof.* It is clear that $\pi_{\Omega *}\mu = \mathbb{P}$, and since $\Omega_0$ is a full $\mathbb{P}$-measure $T$-invariant set of $\Omega$, it suffices to prove that $\phi(\omega, \cdot)_*\mu_\omega = \mu_{T\omega}$ for $\omega \in \Omega_0$.

Fix $\omega \in \Omega_0$ and $\varphi \in C(M)$. By definition of $\mu_\omega$, we have

$$\phi(\omega, \cdot)_*\mu_\omega(\varphi) = \mu_\omega\left(\varphi(\phi(\omega, \cdot))\right) = \nu_\omega\left(h(\omega, \cdot) \cdot \varphi(\phi(\omega, \cdot))\right).$$

Using conclusion (3) of Corollary 3, we obtain

$$\begin{aligned}
\nu_\omega\left(h(\omega, \cdot) \cdot \varphi(\phi(\omega, \cdot))\right) &= \frac{1}{\Lambda(\omega)}\mathcal{L}^*(\omega)\nu_{T\omega}\left(h(\omega, \cdot) \cdot \varphi(\phi(\omega, \cdot))\right) \\
&= \frac{1}{\Lambda(\omega)}\nu_{T\omega}\left(\mathcal{L}(\omega)(h(\omega, \cdot) \cdot \varphi(\phi(\omega, \cdot))\right) \\
&= \frac{1}{\Lambda(\omega)}\nu_{T\omega}\left(\varphi \cdot \mathcal{L}(\omega)h(\omega, \cdot)\right).
\end{aligned}$$

Finally, conclusion (3) of Proposition 6 implies that

$$\frac{1}{\Lambda(\omega)}\nu_{T\omega}\left(\varphi \cdot \mathcal{L}(\omega)h(\omega, \cdot)\right) = \nu_{T\omega}(\varphi \cdot h(T\omega, \cdot)) = \mu_{T\omega}(\varphi).$$

Hence, $\phi(\omega, \cdot)_*\mu_\omega(\varphi) = \mu_{T\omega}(\varphi)$, and the proof is complete. □

6.6. **Exponential decay of correlations.** Let $\varphi_1, \varphi_2 : \Omega \times M \to \mathbb{R}$ be measurable functions such that $\varphi_1(\omega, \cdot) \in C^\alpha(M)$ and that $\varphi_2(\omega, \cdot)$ is bounded for every $\omega \in \Omega_0$. For every $\omega \in \Omega_0$ and $n > 0$, we define the *random correlation function of* $\varphi_1$ *and* $\varphi_2$ as follows

$$C_n(\omega, \varphi_1, \varphi_2) = \int_M \varphi_1(\omega, x)\varphi_2(\omega, \phi_n(\omega, x))\mathrm{d}\mu_\omega(x) \\ - \int_M \varphi_1(\omega, x)\mathrm{d}\mu_\omega(x) \int_M \varphi_2(\omega, x)\mathrm{d}\mu_{T^n\omega}(x).$$

In the next proposition, we prove that the random correlation function $C_n(\omega, \varphi_1, \varphi_2)$ decays exponentially fast for $\mathbb{P}$-a.e. $\omega \in \Omega$.

**Proposition 11.** *Let $\varphi_1$ and $\varphi_2$ as above, and let $\xi < 0$ be the constant as Proposition 6. Then*

$$\limsup_{n\to+\infty} \frac{1}{n} \log |C_n(\omega, \varphi_1, \varphi_2)| \le \xi \qquad \text{for } \omega \in \Omega_0.$$

*Proof.* Fix $\omega \in \Omega_0$. Then

$$\begin{aligned}
&|C_n(\omega, \varphi_1, \varphi_2)| \\
&= \left| \int_M h(\omega, x)\varphi_1(\omega, x)\varphi_2(\omega, \phi_n(\omega, x))\mathrm{d}\nu_\omega(x) \right. \\
&\qquad \left. - \int_M h(\omega, x)\varphi_1(x)\mathrm{d}\nu_\omega(x) \int_\Omega h(T^n\omega, x)\varphi_2(\omega, x)\mathrm{d}\nu_{T^n\omega}(x) \right| \\
&= \left| \int_M \frac{1}{\Lambda_n(\omega)} \mathcal{L}_n(\omega)(h(\omega, \cdot)\varphi_1(\omega, \cdot)\varphi_2(\omega, \phi_n(\omega, \cdot)))(x)\mathrm{d}\nu_{T^n\omega}(x) \right. \\
&\qquad \left. - \int_M h(\omega, x)\varphi_1(\omega, x)\mathrm{d}\nu_\omega(x) \int_M h(T^n\omega, x)\varphi_2(\omega, x)\mathrm{d}\nu_{T^n\omega}(x) \right| \\
&= \left| \int_M \varphi_2(\omega, x) \frac{1}{\Lambda_n(\omega)} \mathcal{L}_n(\omega)(h(\omega, \cdot)\varphi_1(\omega, \cdot))(x)\mathrm{d}\nu_{T^n\omega}(x) \right. \\
&\qquad \left. - \int_M h(\omega, x)\varphi_1(\omega, x)\mathrm{d}\nu_\omega(x) \int_M h(T^n\omega, x)\varphi_2(\omega, x)\mathrm{d}\nu_{T^n\omega}(x) \right| \\
&= \left| \int_M \varphi_2(\omega, x) \left( \frac{1}{\Lambda_n(\omega)} \mathcal{L}_n(\omega)(h(\omega, \cdot)\varphi_1(\omega, \cdot))(x) \right. \right. \\
&\qquad \left. \left. - h(T^n\omega, x) \int_M h(\omega, x)\varphi_1(\omega, x)\mathrm{d}\nu_\omega(x) \right) \mathrm{d}\nu_{T^n\omega}(x) \right| \\
&\le \|\varphi_2(\omega, \cdot)\|_\infty \\
&\qquad \times \left\| \frac{1}{\Lambda_n(\omega)} \mathcal{L}_n(\omega)(h(\omega, \cdot)\varphi_1(\omega, \cdot)) - \nu_\omega(h(\omega, \cdot)\varphi_1(\omega, \cdot))h(T^n\omega, \cdot) \right\|_\infty .
\end{aligned}$$

The second equality is obtained by using the part of Proposition 8 relative to the extension of conclusion (2) of Proposition 6.8. We now observe that $\|\varphi_2(\omega, \cdot)\|_\infty$ is finite by hypothesis, and that $h(\omega, \cdot)\varphi_1(\omega, \cdot) \in C^\alpha(M)$ because $h(\omega, \cdot) \in C^\alpha(M)$ by Proposition 6 and $\varphi_1(\omega, \cdot) \in C^\alpha(M)$ by hypothesis. To obtain the wanted conclusion, we just need to apply the part of Proposition 8 about the extension of Proposition 7 to the function $\varphi = h(\omega, \cdot)\varphi_1(\omega, \cdot)$. □

6.7. **Ergodicity.** We now prove that the measure $\mu$ is ergodic.

**Proposition 12.** *The measure $\mu$ is ergodic.*

*Proof.* To prove that $\mu$ is ergodic, we can show that either $\mu(A) = 0$ or $\mu(A) = 1$ for every $F$-invariant set $A \in \mathcal{F} \otimes \mathcal{B}$. It is quite easy to see that such a set $A$ is completely characterized by the relation

$$\phi(\omega, \cdot)^{-1} A_{T\omega} = A_\omega \qquad \text{for every } \omega \in \Omega_0. \tag{11}$$

Therefore, to prove that $\mu$ is ergodic, we can equivalently show that either $\mu_\omega(A_\omega) = 0$ for $\mathbb{P}$-a.e. $\omega \in \Omega_0$, or $\mu_\omega(A_\omega) = 1$ for $\mathbb{P}$-a.e. $\omega \in \Omega_0$.

Let $A \in \mathcal{F} \otimes \mathcal{B}$ be an $F$-invariant set. Using relation (11) and the fact that $\phi_*(\omega, \cdot)\mu_{T\omega} = \mu_\omega$ for $\omega \in \Omega_0$, we obtain

$$\begin{aligned} \mu_\omega(A_\omega) &= \mu_\omega(\phi(\omega, \cdot)^{-1} A_{T\omega}) \\ &= \phi_*(\omega, \cdot)\mu_{T\omega}(A_{T\omega}) \\ &= \mu_{T\omega}(A_{T\omega}) \qquad \text{for } \omega \in \Omega_0, \end{aligned} \tag{12}$$

which shows that the function $\omega \mapsto \mu_\omega(A_\omega)$ is $T$-invariant. The measurability of this function follows from [8, Proposition 3.3]. Since $T$ is ergodic, we can conclude that there exists a constant $c \in [0, 1]$ such that $\mu_\omega(A_\omega) = c$ for $\mathbb{P}$-a.e. $\omega \in \Omega_0$.

It remains to show that $c$ is equal to either 0 or 1. Fix $\omega \in \Omega_0$ and $\varphi \in C^\alpha(M)$. Using relations (11) and (12), we get

$$\begin{aligned} \mu_\omega\left(\varphi(I_{A_\omega} - \mu_\omega(A_\omega))\right) &= \mu_\omega(\varphi I_{A_\omega}) - \mu_\omega(\varphi)\mu_\omega(A_\omega) \\ &= \mu_\omega(\varphi I_{A_{T^n\omega}} \circ \phi_n(\omega, \cdot)) - \mu_\omega(\varphi)\mu_{T^n\omega}(A_{T^n\omega}) \end{aligned}$$

for every $n > 0$. The same calculations as in the proof of Proposition 11 give

$$\begin{aligned} \mu_\omega(\varphi I_{A_{T^n\omega}} \circ \phi_n(\omega, \cdot)) - \mu_\omega(\varphi)\mu_{T^n\omega}(A_{T^n\omega}) &= \|I_{T^n A_\omega}\|_\infty \\ &\cdot \left\| \frac{1}{\Lambda_n(\omega)} \mathcal{L}_n(\omega)(h(\omega, \cdot)\varphi(\omega, \cdot)) - \nu_\omega(h(\omega, \cdot)\varphi(\omega, \cdot))h(T^n\omega, \cdot) \right\|_\infty, \end{aligned}$$

which together Proposition 8 (the part regarding the extension of Proposition 7) shows that

$$\lim_{n \to +\infty} \left(\mu_\omega(\varphi I_{A_{T^n\omega}} \circ \phi_n(\omega, \cdot)) - \mu_\omega(\varphi)\mu_{T^n\omega}(A_{T^n\omega})\right) = 0.$$

Hence,

$$\mu_\omega(\varphi(I_{A_\omega} - \mu_\omega(A_\omega))) = 0 \qquad \text{for all } \varphi \in C^\alpha(M).$$

Since $C^\alpha(M)$ is dense in $C(M)$, we can finally conclude that

$$\mu_\omega(A_\omega) = I_{A_\omega}(x) \qquad \mu_\omega\text{-a.e. } x \in M,$$

and so $c \in \{0, 1\}$. This completes the proof. □

6.8. **Absolutely continuous invariant measure.** We now consider the special situation when the random potential is given by $\gamma(\omega, x) = -\log|\det D_x\phi(\omega, \cdot)|$ for $\omega \in \Omega$ and $x \in M$. For such a potential, the operator $\mathcal{L}(\omega)$ is the usual Perron–Frobenius operator associated the transformation $\phi(\omega, \cdot)$, and it is easy to check that $\mathcal{L}^*(\omega)m = m$, or equivalently that

$$\|\mathcal{L}(\omega)\varphi\|_1 = \|\varphi\|_1 \qquad \text{for } \varphi \in L^1(m) \text{ with } \varphi \geq 0. \tag{13}$$

**Proposition 13.** *Suppose the transformation $\phi(\omega,\cdot)$ is of class $C^{1+\alpha}(M)$ for every $\omega\in\Omega$, and the random potential is given by $\gamma(\omega,x)=-\log|\det D_x\phi(\omega,\cdot)|$. Then the probabilities $\mu_\omega$ defined in (10) are absolutely continuous with respect to $m$. It follows that the probability $\mu$ is absolutely continuous with respect to $\mathbb{P}\times m$.*

*Proof.* Let $(\Omega_0,\Lambda,h,\nu)$ be the quadruple obtained in Proposition 6 for the random potential $\gamma$ introduced above. Write as before $\Lambda_n(\omega)=\Lambda(T^{n-1}\omega)\cdots\Lambda(\omega)$ for $\omega\in\Omega_0$ and $n>0$. From (13) and conclusion (3) of Proposition 6, we get

$$\Lambda_n(\omega)=\frac{\|h(\omega,\cdot)\|_1}{\|h(T^n\omega,\cdot)\|_1}.$$

It follows that

$$\begin{aligned}\frac{\|h(T^n\omega,\cdot)\|_1}{\|h(\omega,\cdot)\|_1}\cdot\|\mathcal{L}_n(\omega)\varphi\|_1-\nu_\omega(\varphi)\|h(\omega,\cdot)\|_1&\\ \le\left\|\frac{1}{\Lambda_n(\omega)}\mathcal{L}_n(\omega)\varphi-\nu_\omega(\varphi)h(T^n\omega,\cdot)\right\|_1&\\ \le\left\|\frac{1}{\Lambda_n(\omega)}\mathcal{L}_n(\omega)\varphi-\nu_\omega(\varphi)h(T^n\omega,\cdot)\right\|_\infty.&\end{aligned}$$

Since $\|\mathcal{L}_\omega\varphi\|_1=\|\varphi\|_1$ by (13), Proposition 7 implies that

$$\lim_{n\to+\infty}\frac{\|h(T^n\omega,\cdot)\|_1}{\|h(\omega,\cdot)\|_1}\,|\|\varphi\|_1-\nu_\omega(\varphi)\|h(\omega,\cdot)\|_1|=0. \tag{14}$$

Since $T$ is ergodic, using an argument similar to that one in the proof of Lemma 6.7, one can show that for $\mathbb{P}$-a.e. $\omega\in\Omega_0$, there exist $\epsilon>0$ and a divergent sequence of positive integers $\{n_k\}_{k>0}$ such that $\|h(T^{n_k}\omega,\cdot)\|_1/\|h(\omega,\cdot)\|_1>\epsilon$ for every $k>0$. This fact combined with (14) implies that

$$\nu_\omega(\varphi)=\frac{1}{\|h(\omega,\cdot)\|_1}\|\varphi\|_1\qquad\text{for }\mathbb{P}\text{-a.e. }\omega\in\Omega_0\text{ and }\varphi\in C^\alpha_+(M).$$

By Proposition 8, this measure has unique extension to $C(M)$, which is clearly given by

$$\nu_\omega(\varphi)=\frac{1}{\|h(\omega,\cdot)\|_1}\int_M\varphi(x)\mathrm{d}m(x)\qquad\text{for }\varphi\in C(M).$$

We can then conclude that

$$\mu_\omega=\frac{h(\omega,\cdot)}{\|h(\omega,\cdot)\|_1}m\qquad\text{and}\qquad\mu=\frac{h(\omega,\cdot)}{\|h(\omega,\cdot)\|_1}\mathbb{P}\times m,$$

which completes the proof. □

In the next proposition, we show that $\mu$ is the only $F$-invariant probability measure that is absolutely continuous with respect to $\mathbb{P}\times m$.

**Proposition 14.** *Under the hypotheses of Proposition 13, the probability $\mu$ is the unique $F$-invariant probability that is absolutely continuous with respect to $\mathbb{P}\times m$.*

*Proof.* Suppose that $\tilde\mu$ is an $F$-invariant probability measure that is absolutely continuous with respect to $\mathbb{P}\times m$. Let $g$ be the Radon-Nikodym derivative of $\tilde\mu$ with respect

to $\mathbb{P} \times m$. Given a bounded measurable function $\psi : \Omega \times M \to \mathbb{R}$, the $F$-invariance of $\tilde{\mu}$ implies that

$$\begin{aligned} \tilde{\mu}(\psi) &= \frac{1}{n} \sum_{k=0}^{n-1} \tilde{\mu}(\psi) = \tilde{\mu} \left( \frac{1}{n} \sum_{k=0}^{n-1} \psi \circ F^k \right) \\ &= (\mathbb{P} \times m) \left( g \cdot \frac{1}{n} \sum_{k=0}^{n-1} \psi \circ F^k \right) \qquad \text{for } n > 0. \end{aligned} \tag{15}$$

The probability $\mu$ is ergodic by Proposition 12, and so

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{k=0}^{n-1} \psi \circ F^k = \mu(\psi) \qquad \mu\text{-a.e. on } \Omega \times M. \tag{16}$$

From the proof of Proposition 13, we know that the Radon-Nikodym derivative of $\mu$ with respect to $\mathbb{P} \times m$ is equal to $h(\omega, x)/\|h(\omega, \cdot)\|_1$ for $(\mathbb{P} \times m)$-a.e. $(\omega, x) \in \Omega \times M$. Since $h(\omega, \cdot)$ is strictly positive for $\mathbb{P}$-a.e. $\omega \in \Omega$, it is easy to see that $\mathbb{P} \times m \ll \mu$. In other words, the probabilities $\mathbb{P} \times m$ and $\mu$ are equivalent. We can then conclude that limit (16) holds almost everywhere with respect to $\mathbb{P} \times m$ as well. Next, note that

$$\left| g \cdot \frac{1}{n} \sum_{k=0}^{n-1} \psi \circ F^k \right| \leq \|\psi\|_\infty \cdot g \in L^1(\mathbb{P} \times m).$$

This fact together with (16) allows us to apply the Dominated Convergence Theorem to the integral in (15) and obtain

$$\tilde{\mu}(\psi) = (\mathbb{P} \times m)(\mu(\psi)) = \mu(\psi).$$

Since the function $\psi$ is arbitrary, we infer that $\tilde{\mu} = \mu$. □

## References


[1] L. Arnold, *Random Dynamical Systems*, Springer, 1998.

[2] V. Baladi, *Positive Transfer Operators and Decay of Correlations*, volume 16 of *Advanced Series in Nonlinear Dynamics*, World Scientific, 2000.

[3] T. Bogenschütz and V. M. Gundlach, *Ruelle's transfer operator for random subshifts of finite type*, Ergodic Theory Dynam. Systems **15** (1995), no. 3, 413–447.

[4] P. Bougerol, *Kalman filtering with random coefficients and contractions*, SIAM J. Control Optim. **31**(4) (1993), 942–959.

[5] G. Birkhoff, *Extension of Jentzsch's Theorem*, Trans. Amer. Math. Soc. **85** no. 1 (1957), 219–227.

[6] G. Birkhoff, *Uniformly Semi-Primitive Multiplicative Processes*, Trans. Amer. Math. Soc. **104** no. 1 (1962), 37–51.

[7] G. Birkhoff, *Lattice Theory*, volume 25 of *AMS Colloquium Publications* American Mathematical Society, Providence, Rhode Island, 3rd edition, 1967.

[8] H. Crauel, *Random probability measures on Polish spaces*, CRC Press, 2002.

[9] R. M. Dudley, *Real analysis and probability*, Cambridge University Press, 2002.

[10] P. Ferrero and B. Schmitt, *Produits aléatoires d'opérateurs matrices de transfert*, Probab. Th. Rel. Fields **79** (1988), 227–248.

[11] K. Khanin and Yu. Kifer, *Thermodynamic formalism for random transformations and statistical mechanics*, Amer. Math. Soc. Transl. Ser. 2 **171** (1996), 107–140.

[12] Yu. Kifer, *Equilibrium states for random expanding transformations*, Rand. Comput. Dyn. **1** (1992), 1–31.

[13] Yu. Kifer, *Thermodynamic formalism for random transformations revisited*, Stoch. Dyn. **8** no. 1 (2008), 77–102.

[14] U. Krengel, *Ergodic Theorems*, de Gruyter, 1985.

[15] J. M. Lee, *Introduction to smooth manifolds*, Springer, 2002.

[16] J. M. Lee, *Introduction to topological manifolds*, Springer, 2010.

[17] C. Liverani, *Decay of Correlations*, Ann. of Math. **142** (1995), 239–301.
[18] R. Nussbaum, *Iterated nonlinear maps and Hilbert's projective metric*, Mem. Amer. Math. Soc. **79** no. 401, 1989.
[19] W. Rudin, *Real and complex analysis*, McGraw-Hill, 3rd edition, 1986.
[20] M. Viana, *Stochastic dynamics of deterministic systems*, Volume 21 of *Colóquio Brasileiro de Matemática*, IMPA, Rio de Janeiro, 1997.

School of Mathematical and Physical Sciences, University of Reading, Whiteknights, PO Box 220, Reading RG6 6AX, UK
*E-mail address*: j.broecker@reading.ac.uk

CEMAPRE, ISEG, Universidade Técnica de Lisboa, Rua do Quelhas 6, 1200-781 Lisbon, Portugal
*E-mail address*: delmagno@iseg.utl.pt

# THE BOUNDARY OF A DIVISIBLE CONVEX SET

MICKAËL CRAMPON

Abstract. We try to describe the boundary of a divisible convex set at an infinitesimal level. The geodesic flow of the Hilbert metric is the main tool in this study, because its asymptotic exponential behaviour (Lyapunov exponents) is related to the shape of the boundary of the convex set.

## 1. Introduction

I have studied in [Craar] the local asymptotic behaviour of the geodesic flow of Hilbert metrics. It naturally led me to introduce what seems to be a new regularity property of convex functions. I gave it the not-so-good name of *approximate regularity*.

**Definition 1.1.** *Let $U$ be an open convex subset of $\mathbb{R}^{n-1}$ and $f : U \longrightarrow \mathbb{R}$ a $\mathcal{C}^1$ strictly convex function. For $x_0 \in U$ and small $v \in \mathbb{R}^{n-1}$, define $f_{x_0}(v) = f(x_0 + v) - f(x_0) - d_{x_0}f(v)$. We say that the function $f$ is* approximately regular *at the point $x_0 \in U$ if, for all $v \in \mathbb{R}^{n-1}$, the limit*

$$\lim_{t\to 0} \frac{\log(f_{x_0}(tv) + f_{x_0}(-tv))}{\log t}$$

*exists.*

The property is here defined for strictly convex $\mathcal{C}^1$ functions but it has a trivial extension to general convex functions. The main result of [Craar] about this property is the following decomposition theorem, that I proved using the geodesic flow of Hilbert metrics:

**Theorem 1.2** ([Craar], Theorem 6.1)**.** *Let $f : U \longrightarrow \mathbb{R}$ be a $\mathcal{C}^1$ strictly convex function. The following propositions are equivalent:*

(i) *$f$ is approximately regular at the point $x_0 \in U$;*

(ii) *there exist $1 \leqslant p \leqslant n-1$, a splitting $\mathbb{R}^{n-1} = \oplus_{i=1}^{p} G_i$ and numbers $+\infty \geqslant \alpha_1 > \cdots > \alpha_p \geqslant 1$ such that for all $v \in G_i$,*

$$\lim_{t\to 0} \frac{\log(f_{x_0}(tv_i) + f_{x_0}(-tv_i))}{\log t} = \alpha_i;$$

(iii) *there exist $1 \leqslant p \leqslant n-1$,a filtration $\{0\} = H_0 \subsetneq H_1 \subsetneq \cdots \subsetneq H_p = \mathbb{R}^{n-1}$ and numbers $+\infty \geqslant \alpha_1 > \cdots > \alpha_p \geqslant 1$ such that, for any $v_i \in H_i \setminus H_{i-1}$,*

$$\lim_{t\to 0} \frac{\log(f_{x_0}(tv_i) + f_{x_0}(-tv_i))}{\log t} = \alpha_i.$$

When $f$ is approximately regular at $x_0$, we call the numbers $\alpha_i$ the *Lyapunov exponents* of $f$ at $x_0$. It will be more convenient in this work to count the Lyapunov exponents with multiplicities, taking into account the dimension of the subsets $G_i$. We thus define

The author was partially supported by the Fondecyt project N° 3120071 of CONICYT (Chile).

the vector of Lyapunov exponents $\alpha = (\alpha_i)_{i=1\cdots n-1}$, with $\alpha_1 \geqslant \cdots \geqslant \alpha_{n-1}$ and we say that $f$ is *approximately $\alpha$-regular* at $x_0$.

Apart from the previous theorem, I do not know what else can be said about approximate regularity. For example, I asked the question to know whether a convex function is Lebesgue-almost everywhere approximately regular, and to describe the range of possible Lyapunov exponents. Actually, I do not even know if any convex function is approximately regular at at least one point.

In this note, I study these questions for the boundary of a divisible convex set for which lots of properties can be deduced from its numerous symmetries (see section 2.3). As an example, let us give the following result.

**Theorem 1.3.** *Let $\Omega \subset \mathbb{RP}^n$ be a divisible strictly convex set. There exists $\alpha = (\alpha_i)_{i=1\cdots n-1} \in \mathbb{R}^{n-1}$ such that its boundary $\partial\Omega$ is approximately $\alpha$-regular at Lebesgue-almost every point. Furthermore, $\alpha_1 > 2$ unless $\Omega$ is an ellipsoid.*

**Acknowledgements:** I would like to thank François Ledrappier for shadowing orbits together.

## 2. Hilbert geometry and divisible convex sets

### 2.1. Hilbert geometry.

A *Hilbert geometry* is a metric space $(\Omega, d_\Omega)$ where

- $\Omega$ is a *properly convex open set* of the real projective space $\mathbb{RP}^n$, $n \geqslant 2$; *properly* means that there exists a projective hyperplane which does not intersect the closure of $\Omega$, or, equivalently, that there is an affine chart in which $\Omega$ appears as a relatively compact set;
- $d_\Omega$ is the distance on $\Omega$ defined, for two distinct points $x, y$, by

$$d_\Omega(x,y) = \frac{1}{2}|\log[a,b,x,y]|,$$

where $a$ and $b$ are the intersection points of the line $(xy)$ with the boundary $\partial\Omega$ (see Figure 1) and $[a,b,x,y]$ denotes the cross ratio of the four points : if we identify the line $(xy)$ with $\mathbb{R}\cup\{\infty\}$, it is defined by $[a,b,x,y] = \frac{|ax|/|bx|}{|ay|/|by|}$ .

These geometries had been introduced by Hilbert at the end of the nineteenth century as examples of spaces where lines are geodesics, which one can see as a motivation for the fourth of his famous problems: roughly speaking, this problem consisted in finding all geometries for which lines are geodesics.
When $\Omega$ is an ellipsoid, one recovers in this way the Beltrami model of the hyperbolic space. This is the only case where a Hilbert geometry is Riemannian. Otherwise, it is only a Finsler space: The Hilbert metric $d_\Omega$ is generated by a field of norms $F$ on $\Omega$, the norm $F(x,u)$ of a tangent vector $u \in T_x\Omega$ being given by the formula

$$F(x,u) = \frac{|u|}{2}\left(\frac{1}{|xu^+|} + \frac{1}{|xu^-|}\right),$$

where $|\ .\ |$ is an arbitrary Euclidean metric, and $u^+$ and $u^-$ are the intersection points of the line $x + \mathbb{R}.u$ with the boundary $\partial\Omega$ (see Figure 1).

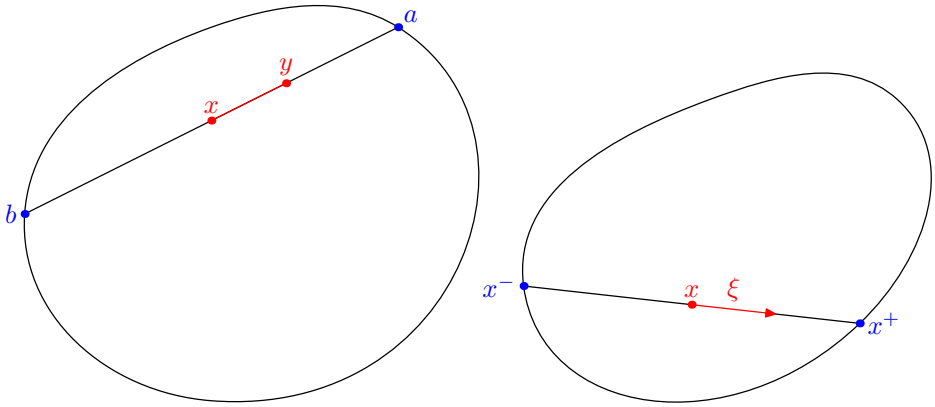


FIGURE 1. The Hilbert distance and the Finsler metric

2.2. **Horospheres.** Assume $\Omega$ is strictly convex with $\mathcal{C}^1$ boundary. In this case, Busemann functions and horospheres can be defined as in hyperbolic geometry.
The *Busemann function* based at $x \in \partial\Omega$ is defined by

$$b_x(z,y) = \lim_{p \to x} d_\Omega(z,p) - d_\Omega(y,p), \ \ z,y \in \Omega$$

which, in some sense, measures the (signed) distance from $z$ to $y$ in $\Omega$ as seen from the point $x \in \partial\Omega$.

The *horosphere* passing through $z \in \Omega$ and based at $x \in \partial\Omega$ is the set

$$\mathcal{H}_x(z) = \{y \in \Omega, \ b_x(z,y) = 0\}.$$

$\mathcal{H}_x(z)$ is also the limit when $p$ tends to $x$ of the metric spheres $S(p, d_\Omega(p,z))$ about $p$ passing through $z$. In some sense, the points on $\mathcal{H}_x(z)$ are those which are as far from $x$ as $z$ is.

2.3. **Divisible convex sets.** Since projective transformations preserve cross-ratios, the group of projective symmetries of $\Omega$,

$$\mathrm{Aut}(\Omega) = \{g \in \mathrm{PSL}_{n+1}(\mathbb{R}), \ g(\Omega) = \Omega\},$$

is a subgroup of isometries of the Hilbert geometry $(\Omega, d_\Omega)$[1]. A discrete subgroup $\Gamma$ of $\mathrm{Aut}(\Omega)$ acts then properly discontinuously on $\Omega$; by Selberg's lemma, it contains a finite index subgroup which has no torsion. The quotient $\Omega/\Gamma$ is thus an orbifold in general, a manifold if $\Gamma$ has no torsion.

**Definition 2.1.** *We say that a properly convex open set $\Omega$ or the corresponding Hilbert geometry $(\Omega, d_\Omega)$ is* divisible *if there exists a discrete subgroup $\Gamma$ of $\mathrm{Aut}(\Omega)$ with compact quotient $\Omega/\Gamma$.*

The first example of divisible convex set is the ellipsoid, that is, the hyperbolic space. Y. Benoist proved the following alternative in [Ben04].

[1]It is conjectured that, for most Hilbert geometries, all isometries are projective.

**Theorem 2.2.** *Let $\Omega$ be a divisible convex set, divided by a discrete subgroup $\Gamma$ of* $\mathrm{Aut}(\Omega)$. *The following properties are equivalent:*

- *the convex set $\Omega$ is strictly convex;*
- *the boundary $\partial\Omega$ is of class $\mathcal{C}^1$;*
- *the Hilbert geometry $(\Omega, d_\Omega)$ is Gromov-hyperbolic;*
- *the group $\Gamma$ is Gromov-hyperbolic.*

An important argument of duality is used to prove this theorem, that we recall now. Consider one of the two convex cones $C \subset \mathbb{R}^{n+1}$ whose trace is $\Omega$. The dual convex set $\Omega^*$ is the trace of the dual cone

$$C^* = \{f \in (\mathbb{R}^{n+1})^*,\ \forall x \in C,\ f(x) > 0\}.$$

The set $\Omega^*$ can be identified with the set of projective hyperplanes which do not intersect $\overline{\Omega}$: to such a hyperplane corresponds the line of linear maps whose kernel is the given hyperplane. For example, we can see the boundary of $\Omega^*$ as the set of tangent spaces to $\partial\Omega$. In particular, when $\Omega$ is strictly convex with $\mathcal{C}^1$ boundary, there is a homeomorphism between the boundaries of $\Omega$ and $\Omega^*$: to the point $x \in \partial\Omega$ we associate the (projective class of the) linear map $x^*$ such that $\ker x^* = T_x\partial\Omega$. The group $\mathrm{Aut}(\Omega)$ acts on the dual convex set $\Omega^*$ via $g.y = ({}^t g)^{-1}(y)$, $g \in \mathrm{Aut}(\Omega)$.

**Lemma 2.3** ([Ben04], Lemme 2.8)**.** *Let $\Gamma$ be a discrete subgroup of* $\mathrm{Aut}(\Omega)$. *The action of $\Gamma$ on $\Omega$ is cocompact if and only if the action of $\Gamma$ on $\Omega^*$ is also cocompact.*

Apart from the ellipsoid, various examples of strictly convex divisible sets have been given. Some can be constructed using Coxeter groups ([KV67], [Ben06b]), some by deformations of hyperbolic manifolds (based on [JM87] and [Kos68], see also [Gol90] for the 2-dimensional case); we should also quote the exotic examples of M. Kapovich [Kap07] of divisible convex sets in all dimensions which are not quasi-isometric to the hyperbolic space (Y. Benoist [Ben06b] had already given an example in dimension 4).
Non-strictly convex examples are more difficult to find. The trivial ones are given by the symmetric spaces of the groups $\mathrm{SL}_n(\mathbb{K})$ ($\mathbb{K}$ being the set of complex, quaternionic or octonionic numbers[2]) or by products (see the historical remarks in [Ben03]). The only other known examples have been constructed by Y. Benoist [Ben06a] and L. Marquis [Mar10] in dimension 3 using Coxeter groups.

2.4. **Properties of the dividing group.** Let $\Omega \subset \mathbb{RP}^n$ be a properly convex strictly convex set, divided by a torsion-free discrete group $\Gamma$. All elements $g \in \Gamma$ are *hyperbolic isometries* of the Hilbert geometry $(\Omega, d_\Omega)$. That means the following.
The element $g$ fixes exactly two points $x_g^+$ and $x_g^-$ on $\partial\Omega$; the point $x_g^+$ is the attractive point of $g$, $x_g^-$ is the repulsive point of $g$ : for any point $x \in \overline{\Omega} \smallsetminus \{x_g^-, x_g^+\}$, $\lim_{n\to\pm\infty} g^n(x) = x_g^\pm$.
Denote by $(\ell_i(g))_{i=0\cdots n}$ the complex eigenvalues of $g$, counted with multiplicities and ordered such that $|\ell_0(g)| \geqslant |\ell_1(g)| \geqslant \cdots \geqslant |\ell_n(g)|$. The largest and smallest eigenvalues $\ell_0$ and $\ell_n(g)$ are simple, real and positive, and the points $x_g^+$ and $x_g^-$ are the corresponding eigenvectors.
Let $\lambda_i(g) = \log|\ell_i(g)|$, $i = 0\cdots n$. The isometry $g$ acts as a translation of length

[2]In the case of octonions, the only possibility is $n = 3$.

$\frac{1}{2}(\lambda_n(g) - \lambda_0(g))$ on the open segment $]x_g^- x_g^+[$.

The following result will be crucial to deduce some rigidity results.

**Theorem 2.4** (Y. Benoist [Ben00]). *Let $\Omega \subset \mathbb{RP}^n$ be a properly convex strictly convex set, divided by a discrete group $\Gamma$. The group $\Gamma$ is Zariski-dense in $\mathrm{SL_{n+1}}(\mathbb{R})$, unless $\Omega$ is an ellipsoid.*

Recall that the *Zariski-closure* of a subgroup $\Gamma$ of $\mathrm{SL_{n+1}}(\mathbb{R})$ is the smallest algebraic subgroup $G$ of $\mathrm{SL_{n+1}}(\mathbb{R})$ which contains $\Gamma$. We then say that $\Gamma$ is *Zariski-dense* in $G$.
The hypothesis of strict convexity in the last theorem is actually unnecessary, but the proof in the general case is far more involved [Ben03].
This last theorem will be useful through the following characterization of Zariski-dense subgroups of semisimple Lie groups, which is also due to Y. Benoist, and that we explain in the case of the group $\mathrm{SL_{n+1}}(\mathbb{R})$. To each element $g$ in $\mathrm{SL_{n+1}}(\mathbb{R})$, we associate the vector $\log(g) = [\lambda_0(g) : \cdots : \lambda_n(g)] \in \mathbb{RP}^n$ and for a subgroup $\Gamma$ of $\mathrm{SL_{n+1}}(\mathbb{R})$, we set $\log\Gamma = \overline{\{\log g,\ g \in \Gamma\}}$.

**Theorem 2.5** (Y. Benoist, [Ben97]). *Let $\Gamma$ be a subgroup of $\mathrm{SL_{n+1}}(\mathbb{R})$. If $\Gamma$ is Zariski-dense in $\mathrm{SL_{n+1}}(\mathbb{R})$, then $\log\Gamma$ has nonempty interior.*

## 3. Curvature of the boundary

3.1. **What is curvature.** Let us begin with an old theorem of A. D. Alexandrov [Ale39] about convex functions:

**Theorem 3.1.** *Let $f : U \subset \mathbb{R}^{n-1} \longmapsto \mathbb{R}$ be a convex function defined on a convex open set $U$ of $\mathbb{R}^{n-1}$. The Hessian matrix $\mathrm{Hess}(f) = \left(\frac{\partial^2 f}{\partial_i \partial_j}\right)_{ij}$ exists Lebesgue almost everywhere in $U$.*

Let $\Omega$ be a bounded convex set of the Euclidean space $\mathbb{R}^n$. It is then possible to compute the Hessian of its boundary at Lebesgue almost every point $x \in \partial\Omega$. We will call a $\mathcal{C}^2$ *point* a point $x$ where this is possible.
The Hessian is a positive symmetric bilinear form on the tangent space $T_x\partial\Omega$. It represents the curvature of the boundary at $x$. When it is degenerate, that means the curvature of the boundary is zero in some tangent direction.
The Hessian is a Euclidean notion, but its degeneracy is not. Namely, if $\Omega$ is a properly convex open set of $\mathbb{RP}^n$ and $x$ a point of $\partial\Omega$, we can choose an affine chart centered at $x$ and a metric on it and compute the Hessian of $\partial\Omega$ at $x$; its degeneracy does not depend on the choice of the affine chart and the metric.

We can *measure* the vanishing of the curvature of $\partial\Omega$ in the following way. Fix a smooth measure $\lambda^*$ on the boundary of the dual convex set $\Omega^*$, and call $\lambda$ its pull-back to $\partial\Omega$. Then $\lambda$ can be seen as a measure of the curvature of $\partial\Omega$. It can be decomposed as

$$\lambda = \lambda^{ac} + \lambda^{sing},$$

where $\lambda^{ac}$ is an absolutely continuous measure and $\lambda^{sing}$ is singular with respect to any Lebesgue measure on $\partial\Omega$. For example, in dimension 2, if $\partial\Omega$ is not $\mathcal{C}^1$ at some point $x$ then $\lambda$ will have an atom at $x$. The support of $\lambda^{ac}$ is the closure of the set of $\mathcal{C}^2$ points with nondegenerate Hessian.
Though $\Omega$ is convex, it may happen that $\lambda^{ac} = 0$, that is, $\lambda$ is singular with respect to some (hence any) smooth measure on $\partial\Omega$. This is equivalent to the fact that the Hessian

is degenerate at Lebesgue-*almost all* $\mathcal{C}^2$ point of $\partial\Omega$. We then say that *the curvature of the boundary is supported on a set of zero Lebesgue measure.*

3.2. **Curvature of the boundary of a divisible convex set.** The curvature of the boundary of a divisible convex set has been investigated by J.-P. Benzécri [Ben60].

**Lemma 3.2** (J.-P. Benzécri [Ben60])**.** *Let $X_n$ denote the set of properly convex open sets of $\mathbb{RP}^n$, equipped with the Hausdorff topology. Let $\Omega \in X_n$.*

- *If there exists a $\mathcal{C}^2$ point $x \in \partial\Omega$ with nondegenerate Hessian, then the closure of the orbit $\mathrm{PSL}_{n+1}(\mathbb{R}) \cdot \Omega$ in $X_n$ contains an ellipsoid.*
- *If $\Omega$ is divisible then the orbit $\mathrm{PSL}_{n+1}(\mathbb{R}) \cdot \Omega$ is closed in $X_n$.*

*Proof.* These two results are respectively Propositions 5.3.10 and 5.3.3 of [Ben60]. Let us recall the proofs.
Choose an affine chart and a Euclidean metric on it such that $\Omega$ appears as a bounded convex open set of $\mathbb{R}^n$. Let $x$ be a point of $\partial\Omega$ with nondegenerate Hessian. Let $\mathcal{E}$ be the osculating ball of $\partial\Omega$ at $x$. It defines a hyperbolic geometry $(\mathcal{E}, d_{\mathcal{E}})$. Pick a point $y \in \partial\mathcal{E}$ distinct from $x$, and choose a hyperbolic isometry $g$ of $\mathcal{E}$ whose attracting fixed point $y$ and repulsive one $x$. Now, since $\partial\mathcal{E}$ and $\partial\Omega$ are tangent up to order 2, it is not difficult to see that $g^n \cdot \Omega$ converges to $\mathcal{E}$ when $n$ goes to $+\infty$. This proves the first point.
The second point is a consequence of another result of Benzécri, which says that the action of $\mathrm{PSL}_{n+1}(\mathbb{R})$ on the set $\dot{X}_n = \{(\Omega, x),\ \Omega \in X_n, x \in \Omega\}$ is proper (this is Théorème 3.2.1 of [Ben60]). Each orbit $\mathrm{PSL}_{n+1}(\mathbb{R}) \cdot (\Omega, x)$ is thus closed. Now, the orbit $\mathrm{PSL}_{n+1}(\mathbb{R}) \cdot \Omega$ is closed in $X_n$ if and only if the union $\cup_{x\in\Omega}\mathrm{PSL}_{n+1}(\mathbb{R}) \cdot (\Omega, x)$ is closed in $\dot{X}_n$. Since $\Omega$ is divisible, divided, say, by the group $\Gamma$, there is a compact subset $K$ of $\Omega$ such that $\Gamma \cdot K = \Omega$. So the union

$$\bigcup_{x\in\Omega} \mathrm{PSL}_{n+1}(\mathbb{R}) \cdot (\Omega, x) = \bigcup_{x\in K} \bigcup_{g\in\Gamma} \mathrm{PSL}_{n+1}(\mathbb{R}) \cdot (\Omega, g(x)) = \bigcup_{x\in K} \mathrm{PSL}_{n+1}(\mathbb{R}) \cdot (\Omega, x)$$

is closed in $\dot{X}_n$. □

More about Benzécri's contributions can be found in L. Marquis's survey [Mar13]; the proof of the second point above is actually taken from it. As a consequence of the last lemma, we get the following

**Proposition 3.3.** *Let $\Omega \subset \mathbb{RP}^n$ be a divisible convex set, and assume $\Omega$ is not an ellipsoid. Then any $\mathcal{C}^2$ point has degenerate Hessian. In particular, the curvature of $\partial\Omega$ is supported on a subset of zero Lebesgue measure.*

*Proof.* Assume that the Hessian of $\partial\Omega$ is not degenerate at some $\mathcal{C}^2$ point. Lemma 3.2 implies that the orbit $\mathrm{PSL}_{n+1}(\mathbb{R}) \cdot \Omega$ is closed and contains an ellipsoid. So $\Omega$ itself is an ellipsoid. □

When the convex set is strictly convex, the geodesic flow of the Hilbert metric allows to say more about the properties of the boundary. The rest of this paper is dedicated to this case.

## 4. The geodesic flow and the boundary

4.1. **The geodesic flow.** When $\Omega$ is strictly convex, the metric space $(\Omega, d_\Omega)$ is uniquely geodesic, and the geodesics are lines. The geodesic flow is then well defined on the homogeneous bundle $\pi : H\Omega \longrightarrow \Omega$ of tangent directions: To find the image by $\varphi^t$ of a

point $w = (x, [\xi]) \in H\Omega$, consisting of a point and a direction, one follows the geodesic line $c_w$ leaving $x$ in the direction $[\xi]$, and one has $\varphi^t(w) = (c_w(t), [c'_w(t)])$.
By projection, this also defines the geodesic flow on $HM$, the homogeneous bundle of $M = \Omega/\Gamma$.
The geodesic flow has the same regularity as the boundary of $\Omega$. So, if $\Omega$ is strictly convex and divisible, by Theorem 2.2, it is a $\mathcal{C}^1$ flow. We will denote by $X$ the generator of the geodesic flow (both on $H\Omega$ or $HM$).

**Theorem 4.1** (Y. Benoist, [Ben04]). *Let $M = \Omega/\Gamma$ a compact manifold quotient of a strictly convex set $\Omega \subset \mathbb{RP}^n$. The geodesic flow on $HM$ is an Anosov flow: There exist a $\varphi^t$-invariant splitting of the tangent bundle*

$$THM = \mathbb{R}.X \oplus E^u \oplus E^s$$

*and constants $C, \alpha > 0$ such that, for any $t \geqslant 0$,*

$$\|d\varphi^t(Z^s)\| \leqslant Ce^{-\alpha t}\|Z^s\|,\ Z^s \in E^s,$$
$$\|d\varphi^{-t}(Z^u)\| \leqslant Ce^{-\alpha t}\|Z^u\|,\ Z^u \in E^u.$$

Here the norm $\|\ \cdot\ \|$ denotes an arbitrary Finsler metric on $HM$; because $HM$ is compact, the Anosov property of the flow does not depend on the metric, even if the constants $C$ and/or $a$ do.
In our situation, the stable and unstable bundles $E^s$ and $E^u$ can be geometrically understood using horospheres. For $w \in H\Omega$, define the sets

$$W^s(w) = \{v \in H\Omega \mid v^+ = w^+,\ \pi(v) \in \mathcal{H}_{w^+}(\pi(w))\},$$

and

$$W^u(w) = \{v \in H\Omega \mid v^- = w^-,\ \pi(v) \in \mathcal{H}_{w^-}(\pi(w))\}.$$

The sets $W^s(w)$ and $W^u(w)$ are $\mathcal{C}^1$ submanifolds of $H\Omega$ and it is not difficult to see that they are the stable and unstable sets of the geodesic flow ($d$ denotes the distance generated by $\|\ \cdot\ \|$):

$$W^s(w) = \{v \in H\Omega,\ \lim_{t\to+\infty} d(\varphi^t w, \varphi^t v) = 0\}$$

and

$$W^u(w) = \{v \in H\Omega,\ \lim_{t\to-\infty} d(\varphi^t w, \varphi^t v) = 0\}.$$

Both families $W^s(w), w \in H\Omega$ and $W^u(w), w \in H\Omega$ form a $\varphi^t$-invariant foliation of $H\Omega$. Everything projects down on $HM$ where we will use the same notation. The stable and unstable bundles are then the tangent spaces to the stable and unstable foliations : $E^s(w) = T_w W^s(w)$, $E^u(w) = T_w W^u(w)$.

The asymptotic behaviour of the geodesic flow is encoded in the boundary of $\Omega$: When we look at the behaviour of the norm $\|d\varphi^t Z\|$ when $t$ goes to $+\infty$, for some $Z \in T_w H\Omega$, we see appearing naturally the graph of the boundary at the extremal point $w^+$. This observation is at the basis of this work. To illustrate this observation, notice the following "consequence" of Theorem 4.1:

**Proposition 4.2** ([Ben04], Proposition 4.6). *The boundary of a divisible strictly convex set is $\mathcal{C}^\alpha$ and $\beta$-convex for some $1 < \alpha \leqslant 2$, $\beta \geqslant 2$. In particular, the geodesic flow is $\mathcal{C}^\alpha$ for some $\alpha > 1$.*

To understand the last statement, we recall the following definitions:

**Definition 4.3.** *Let $1 < \alpha < 2, \beta > 1$ and $U$ an open subset of $\mathbb{R}^n$. A $\mathcal{C}^1$-function $f : U \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ is*

- *of class $\mathcal{C}^\alpha$ if, for some constant $C > 0$,*
$$|f(x) - f(y) - d_x f(y - x)| \leqslant C|x - y|^{1+\varepsilon},\ x, y \in U;$$
- *$\beta$-convex if, for some constant $C > 0$,*
$$|f(x) - f(y) - d_x f(y - x)| \geqslant C|x - y|^{\beta},\ x, y \in U.$$

4.2. **Approximate regularity and Lyapunov exponents.** Recall the following definition:

**Definition 4.4.** *Let $\Omega \subset \mathbb{RP}^n$ be a strictly convex set with $\mathcal{C}^1$ boundary. A point $w \in H\Omega$ is* weakly regular *if, for any $Z \in T_w H\Omega \smallsetminus \{0\}$, the limit*
$$\chi(Z) = \lim_{t \to \pm\infty} \frac{1}{t} \log \|d\varphi^t(Z)\|$$
*exists. It is said to be* forward *or* backward weakly regular *if the limits exist only when $t$ goes to $+\infty$ or $-\infty$ (or if both limits differ). The number $\chi(Z)$ is called the Lyapunov exponent of $Z$.*

Because stable and unstable manifolds $W^s(w)$ and $W^u(w)$ at $w$ have the same projection on $\Omega$, there is a symmetry between the action of the flow on stable and unstable vectors (see for example Lemma 2.3 in [Craar]). In particular, we can see that if $Z^s \in E^s(w), Z^u \in E^u(w)$ project on the same vector $z \in \mathcal{H}_w$, then
$$\chi(Z^u) = 2 + \chi(Z^s).$$
The complete behaviour is then encoded in the behaviour of unstable vectors, and we will be only interested in these vectors by looking at the restriction of the differential $d\varphi^t$ to the bundle $E^u$. Because the geodesic flow is an Anosov flow, all Lyapunov exponents of unstable vectors are positive.
Given a forward weakly regular point $w \in H\Omega$, the numbers $\chi(Z)$, for any $Z \in E^u(w)$, can take only a finite number $0 < \chi_1 < \cdots < \chi_p$ of values, which are called the *positive Lyapunov exponents* of $w$. There is then a $\varphi^t$-invariant splitting
$$TH\Omega = E_1 \oplus \cdots \oplus E_p$$
along the orbit $\varphi.w$, called *Lyapunov splitting*, such that, for any vector $Z_i \in E_i \smallsetminus \{0\}$,
$$\lim_{t \to +\infty} \frac{1}{t} \log \|d\varphi^t(Z_i)\| = \chi_i.$$
As for the exponents $(\alpha_i)$ appearing in the definition of approximate regularity, we will count the $(\chi_i)$ with multiplicities. We thus have $n - 1$ positive Lyapunov exponents $(\chi_i)_{i=1\cdots n-1}$ ordered as $\chi_1 \leqslant \cdots \leqslant \chi_n$. The main result of [Craar] is the following

**Theorem 4.5** ([Craar], Theorem 1)**.** *Let $\Omega \subset \mathbb{RP}^n$ be a strictly convex set with $\mathcal{C}^1$ boundary. A point $w \in H\Omega$ is forward weakly regular if and only if the boundary $\partial\Omega$ is approximately regular at the point $w^+ = \varphi^{+\infty}(w)$. If $0 \leqslant \chi_1 \leqslant \cdots \leqslant \chi_n$ are the positive Lyapunov exponents of $w$, then $\partial\Omega$ is approximately $\alpha$-regular with $\alpha = (\alpha_i)_{i=1\cdots n-1}$ given by $\alpha_i = 2/\chi_i$.*

## 5. The set of approximately regular points and the range of Lyapunov exponents

Let $\Omega \subset \mathbb{RP}^n$ be a divisible strictly convex set. Our interest now will lie on the set of approximately regular points $\Lambda \subset \partial\Omega$, as well as the set of all possible Lyapunov exponents

$$\mathcal{A} = \{\alpha(x) \in \mathbb{R}^{n-1},\ x \in \Lambda\}.$$

By projective invariance of the notion of approximate-regularity, the set of approximately $\alpha$-regular points of $\partial\Omega$ is $\Gamma$-invariant, for any vector $\alpha$. Since the action of $\Gamma$ on $\partial\Omega$ is minimal, it is either empty (if $\alpha \notin \mathcal{A}$) or a dense subset of $\partial\Omega$ (if $\alpha \in \mathcal{A}$).

5.1. **Oseledets' theorem.** The following result is a version of Osedelets' ergodic multiplicative theorem [Ose68]:

**Theorem 5.1.** *Let $M = \Omega/\Gamma$ a manifold quotient of a strictly convex set $\Omega \subset \mathbb{RP}^n$ with $\mathcal{C}^1$ boundary. Let $\mu$ a $\varphi^t$-invariant probability measure on $HM$. The set of weakly regular points has full $\mu$-measure.*

It allows us to deduce the following

**Corollary 5.2.** *Let $\Omega \subset \mathbb{RP}^n$ be a divisible strictly convex set. The set $\mathcal{A}$ is nonempty and the set $\Lambda$ is dense in $\partial\Omega$.*

*Proof.* The set of $\varphi^t$-invariant probability measures on $HM$ is nonempty. In particular, by Oseledets's theorem, there exists a weakly regular point $w \in H\Omega$. By Theorem 4.5, the boundary $\partial\Omega$ is approximately regular at the point $w^+$, so $\mathcal{A}$ is nonempty and $\Lambda$ is dense in $\partial\Omega$. □

5.2. **Hyperbolic isometries and closed orbits.** Recall that any element $g \in \Gamma$ is a hyperbolic isometry of the Hilbert geometry $(\Omega, d_\Omega)$. We use the notation introduced in section 2.4.
For $g \in \Gamma$, pick a point $w \in H\Omega$ such that $w^- = x_g^-,\ w^+ = x_g^+$. The projection on $HM$ of the orbit of $w$ under the flow is a closed orbit of the flow, of length $\frac{1}{2}(\lambda_n(g) - \lambda_0(g))$. Two elements $g$ and $g'$ yield the same closed orbit if and only if they are conjugated. Conversely, any closed orbit is obtained in this way. In other words: Closed orbits of the geodesic flow on $HM$ are in bijection with conjugacy classes of $\Gamma \smallsetminus \{1\}$.

**Proposition 5.3.** *Let $\Omega \subset \mathbb{RP}^n$ be a divisible strictly convex set, divided by a torsion-free discrete group $\Gamma < \mathrm{Aut}(\Gamma)$. Let $g \in \Gamma$. The boundary $\partial\Omega$ is approximately $\alpha(g)$-regular at the point $x_g^+$, with $\alpha(g) = (\alpha_i(g))_{i=1\cdots n-1}$ given by*

$$\alpha_i(g) = \frac{1 - \lambda_n(g)/\lambda_0(g)}{1 - \lambda_i(g)/\lambda_0(g)}. \tag{5.1}$$

*Proof.* In [Cra09], I showed that the positive Lyapunov exponents $(\chi_i(g))_{i=1\cdots n-1}$ of the closed orbit corresponding to the (conjugacy class of the) element $g \in \Gamma$ were given by

$$\chi_i(g) = 2\ \frac{\lambda_0(g) - \lambda_i(g)}{\lambda_0(g) - \lambda_n(g)}.$$

Theorem 4.5 gives the result. □

The element $g \in \Gamma$ acts on the dual convex set $\Omega^*$ by $g.y = ({}^t g)^{-1}(y)$. To $g \in \Gamma$, we thus associate the isometry $g^* = ({}^t g)^{-1} \in \mathrm{Aut}(\Omega^*)$. The dual point to $x_g^+$ is the point

$x^-_{g^*}$, at which $\partial\Omega^*$ is approximately $\alpha(g^*)$-regular, with $\alpha(g^*) = (\alpha_i(g^*))_{i=1\cdots n-1}$ given by

$$\alpha_i(g^*) = \frac{1-\lambda_n(g)/\lambda_0(g)}{1-\lambda_n(g)/\lambda_{n-i}(g)}.$$

Remark that

$$\frac{1}{\alpha_{n-i}(g^*)} + \frac{1}{\alpha_i(g)} = 1,\ i = 1\cdots n-1.$$

In general, if $\partial\Omega$ is approximately $\alpha$-regular at some point $x$ with $\alpha = (\alpha_i)_{i=1\cdots n-1}$, one can expect $\partial\Omega$ to be approximately $\alpha^*$-regular at the dual point $x^* \in \partial\Omega^*$ with $\alpha^* = (\alpha_i^*)_{i=1\cdots n-1}$ satisfying to the previous relation: $1/\alpha^*_{n-i}+1/\alpha_i = 1,\ i = 1\cdots n-1.$. I was able to prove this fact only for $\Omega \subset \mathbb{RP}^2$ in [Craar].

If $\Omega$ is an ellipsoid, then obviously $\Lambda = \partial\Omega$ and $\mathcal{A} = \{2\}$. This second property is characteristic of the ellipsoid. (The first one will be treated in section 5.5.)

**Corollary 5.4.** *Let $\Omega \subset \mathbb{RP}^n$ be a divisible strictly convex set. The closure $\overline{\mathcal{A}}$ of $\mathcal{A}$ has empty interior if and only if $\Omega$ is an ellipsoid.*

*Proof.* Assume $\Omega$ is not an ellipsoid. Then, by Theorems 2.4 and 2.5, the set

$$\log\Gamma = \overline{\{[\lambda_0(g) : \cdots : \lambda_n(g)],\ g \in \Gamma\}} \subset \mathbb{RP}^n$$

has nonempty interior.
Now, the set $\mathcal{A}$ contains the vectors $\alpha(g) = (\alpha_i(g))$, $g \in \Gamma$, defined by $\alpha_i(g) = \frac{1-\lambda_n(g)/\lambda_0(g)}{1-\lambda_i(g)/\lambda_0(g)}$. Hence, $\overline{\mathcal{A}}$ contains the image of the well-defined continuous function

$$\begin{array}{rcl} \log\Gamma & \longrightarrow & \mathbb{R}^{n-1} \\ [\lambda_0 : \cdots : \lambda_n] & \longmapsto & (\frac{1-\lambda_n/\lambda_0}{1-\lambda_1/\lambda_0}, \cdots, \frac{1-\lambda_n/\lambda_0}{1-\lambda_{n-1}/\lambda_0}). \end{array}$$

This gives the result. □

It is likely that one can replace $\overline{\mathcal{A}}$ by $\mathcal{A}$ in the last proposition. A way to prove that would be to see that the set $\mathcal{A}_{\mathcal{M}}$ defined in section 5.3 contains the interior of $\mathcal{A}$.

5.3. **Ergodic measures.** Let $\Lambda(H\Omega)$ be the set of forward weakly regular points of $H\Omega$, which is obviously $\Gamma$-invariant. By Theorem 4.5, the set $\Lambda$ is given by

$$\Lambda = \{w^+ \in H\Omega,\ w \in \Lambda(H\Omega)\}.$$

Let $\mathcal{M}$ be the set of invariant probability measures of the flow on $HM$. Each measure $m \in \mathcal{M}$ defines by lifting it a measure $\tilde{m}$ on $H\Omega$ which is invariant under the actions of $\Gamma$ and the flow.
Oseledets' theorem tells us that, for any $m \in \mathcal{M}$, $\Lambda(H\Omega)$ has full $\tilde{m}$-measure, hence Lyapunov exponents are defined $\tilde{m}$-almost everywhere. If $m$ is an ergodic measure, that is invariant sets have zero or full measure, then Lyapunov exponents are constant almost everywhere: to each ergodic measure $m$ we can thus associate its positive Lyapunov exponents $\chi_1(m) \leqslant \cdots \leqslant \chi_{n-1}(m)$.

We can associate, in a one-to-one way, to each invariant probability measure $m$ on $HM$ a $\Gamma$-invariant Radon measure $M = M(m)$ on the space of oriented geodesics of $\Omega$ given by $\partial^2\Omega = (\partial\Omega \times \partial\Omega) \smallsetminus \Delta$, where $\Delta = \{(x,x),\ x \in \partial\Omega\}$ (see [Kai90] for example). If $m$ is ergodic, Oseledets' theorem implies that for $M$-almost all $(x,y) \in \partial^2\Omega$, the geodesic from $x$ to $y$ is weakly regular with positive Lyapunov exponents $\chi_1(m) \leqslant \cdots \leqslant \chi_{n-1}(m)$;

thus, for $M$-almost all $(x,y) \in \partial^2\Omega$, the boundary $\partial\Omega$ is approximately $\alpha(m)$-regular at $x$, with $\alpha(m) = (\alpha_i(m)))_{i=1\cdots n-1}$ given by

$$\alpha_i(m) = \frac{2}{\chi_i(m)}.$$

The set $\mathcal{A}_{\mathcal{M}} = \{\alpha(m),\ m \in \mathcal{M}\}$ is an interesting subset of $\mathcal{A}$. As I said before, it might contain the interior of $\overline{\mathcal{A}}$.
The diversity of ergodic measures gives an idea of the complexity of the boundary of a divisible strictly convex set which is not an ellipsoid. Here are some examples.

5.3.1. *Closed orbits.* The easiest examples of ergodic measures are the Lebesgue measures $l_g$ supported by a closed orbit $g$, associated to a conjugacy class of a hyperbolic element $g \in \Gamma$. The corresponding subset of $\partial^2\Omega$ of full $M(l_g)$-measure is precisely the orbit of $(x_g^-, x_g^+)$ under $\Gamma$. This has been treated in the previous part.
Denote by $\mathcal{M}_{Per} = \{l_g,\ g \in \Gamma\}$ the set of ergodic measures supported on closed orbits and define $\mathcal{A}_{\mathcal{M}_{Per}} = \{\alpha(m),\ m \in \mathcal{M}_{Per}\}$. It is a consequence of the Anosov closing lemma that $\mathcal{M}_{Per}$ is dense in the set of ergodic measures, so we could expect the following

**Proposition 5.5.** *Let $\Omega \subset \mathbb{RP}^n$ be a divisible strictly convex set. The set $\mathcal{A}_{\mathcal{M}_{Per}}$ is dense in $\mathcal{A}_{\mathcal{M}}$.*

*Proof.* This is a consequence of a nontrivial result one can find in [Kal11] (Theorem 1.4): it states that the vector $\chi(m)$ associated to an ergodic measure can be approximated by a sequence of vectors $(\chi(g_n))$ with $g_n \in \mathcal{M}_{Per}$. The vector $\alpha(m)$ is thus approximated by the sequence $(\alpha(g_n))$. □

5.3.2. *Gibbs measures.* A Gibbs measure is the equilibrium state of a Hölder continuous potential $f : HM \longrightarrow \mathbb{R}$: it is the unique invariant probability measure $\mu_f$ such that

$$h_{\mu_f} + \int f\ d\mu_f = \sup\{h_m + \int f\ dm,\ m \in \mathcal{M}\}.$$

The corresponding measure $M_f$ on $\partial^2\Omega$ can always be written as $M_f = FM_f^s \times M_f^u$, where $F$ is a continuous function on $\partial^2\Omega$, and $M_f^s$ and $M_f^u$ are two finite measures on $\partial\Omega$. The three objects are determined by the potential; in particular, $M_f^u$ and $M_f^s$ are given by the Patterson-Sullivan construction, associated to the potentials $f$ and $\sigma * f$, where $\sigma$ is the flip map, defined on $H\Omega$ by $\sigma(x,[\xi]) = (x,[-\xi])$ (see [Cou03] or [Led95]).
Among Gibbs measures is for instance the Bowen-Margulis measure $\mu_{BM}$ which is the measure of maximal entropy of the flow, that is, the equilibrium state associated to the potential $f = 0$. The corresponding measure $M_{BM}$ is given by

$$dM_{BM}(\xi^+,\xi^-) = e^{2\delta(\xi^+|\xi^-)_o} d\mu_o^2(\xi^+,\xi^-),$$

where $\mu_o$ is the Patterson-Sullivan measure at an arbitrary point $o \in \Omega$, and $(\xi^+|\xi^-)_o$ is the Gromov product $\xi^+$ and $\xi^-$ based at the point $o$: we have $(\xi^+|\xi^-)_o = \frac{1}{2}(b_{\xi^-}(o,x) + b_{\xi^-}(o,x))$ for any point $x \in (\xi^-\xi^+)$ (see [Sul79]).
In [Cra09], I had proved that $\chi^+(\mu_{BM}) = \sum \chi_i(\mu_{BM}) = n-1$. Thus, we get that $\mu_o$-almost every point of $\partial\Omega$ is approximately $\alpha(\mu_{BM})$-regular with $\alpha(\mu_{BM}) = (\alpha_i(\mu_{BM}))_{i=1\cdots n-1}$, such that $(\sum_i 1/\alpha_i(\mu_{BM}))^{-1} = 2(n-1)$. For example, in dimension 2, $\mu_o$-almost every point of $\partial\Omega$ is approximately 2-regular. A question I am not able to answer is to know if, in dimension $n \geqslant 3$, the $\alpha_i$ are all equal to 1 if and only if $\Omega$ is an ellipsoid.

5.4. **Shape of the boundary at Lebesgue almost every point.** The Sinai-Ruelle-Bowen (SRB) measure $\mu^+$ is the equilibrium state associated to the potential

$$f^+ = \frac{d}{dt}|_{t=0} \log \det d\varphi^t_{|E^u}.$$

This potential is Hölder continuous because the geodesic flow is $\mathcal{C}^\alpha$ for some $\alpha > 1$. The measure $\mu^+$ is the only measure whose conditional measures $(\mu^+)^u$ along unstable manifolds are absolutely continuous.
Closely related to this measure is the "reverse" SRB measure $\mu^- = \sigma * \mu^+$, which is the equilibrium state of the potential

$$f^- = -\frac{d}{dt}|_{t=0} \log \det d\varphi^t_{|E^s}.$$

The measure $\mu^-$ is the only invariant measure whose conditional measures along stable manifolds are absolutely continuous.
In the case of the ellipsoid, $\mu^+$, $\mu^-$ and $\mu_{BM}$ all coincide, since $f^+ = f^- = 0$, and they are all absolutely continuous; indeed, they coincide with the Liouville measure of the flow. When $\Omega$ is not an ellipsoid, the Zariski-density of the cocompact group $\Gamma$ implies via Livschitz-Sinai theorem that there is no absolutely continuous measure (see [Ben04]). So the three measures are distinct.

The measure $\mu^+$ is also the only one which satisfies the equality in the Ruelle inequality (see [LY85]). Recall that the Ruelle inequality relates the entropy of an invariant measure $m$ to the sum of positive Lyapunov exponents $\chi^+$ of the flow:

$$h_m \leqslant \int \chi^+ \, dm.$$

For example, the topological entropy $h_{top}$ of the flow satisfies

$$h_{top} = h_{\mu_{BM}} \leqslant n-1,$$

with equality if and only if $\Omega$ is an ellipsoid (this is the main result of [Cra09]). The measures $\mu^+$ and $\mu^-$ have the same entropy $h_{SRB}$ given by

$$h_{SRB} = \int \chi^+ \, d\mu^+ = -\int \chi^- \, d\mu^-,$$

where $\chi^-$ is the sum of negative Lyapunov exponents. In particular, if $\Omega$ is not an ellipsoid, we have $\int \chi^+ \, d\mu^+ = h_{SRB} < h_{\mu_{BM}} < n-1$. Hence the $\mu^+$-almost sure value $\chi^+(\mu^+)$ of the sum of positive Lyapunov exponents satisfies $\chi^+(\mu^+) < n-1$.
The measure $\mu^+$ corresponds to the measure $M^+$ on $\partial^2\Omega$ which can be written $M^+ = F^+M^s \times M^u$, with $M^u$ absolutely continuous, while the measure $\mu^-$ corresponds to $M^- = F^-M^u \times M^s$. In particular, we have the following

**Proposition 5.6.** *Let $\Omega \subset \mathbb{RP}^n$ be a divisible strictly convex set. Then Lebesgue-almost every point of $\partial\Omega$ is approximately $\alpha$-regular with $\alpha = (\alpha_i)_{i=1\cdots n-1}$ given by*

$$\alpha_i = \frac{2}{\chi_i(\mu^+)}.$$

Since $\partial\Omega$ is also Lebesgue almost-everywhere 2-differentiable by Alexandrov's theorem, we have that $\alpha_i \leqslant 2$, $i = 0 \cdots n-1$. When $\Omega$ is an ellipsoid, we have $\alpha_i(SRB) = 2$, $i = 0 \cdots n-1$. Otherwise, the fact that $\chi^+(\mu^+) < 0$ implies that $\chi_1(\mu^+) < 1$ hence $\alpha_1 > 2$. In particular, we recover the fact that the curvature of $\partial\Omega$ is supported on a set of zero Lebesgue-measure.

5.5. **The 2-dimensional case.** In dimension 2, we can understand better the sets $\Lambda$ and $\mathcal{A}$.

5.5.1. *The set of approximately regular points.* We will see here that the property that $\Lambda = \partial\Omega$ characteristic of the ellipsoid. This is probably true in higher dimensions but we would need a more careful approach.

**Proposition 5.7.** *Let $\Omega \subset \mathbb{RP}^2$ be a divisible strictly convex set. If $\Omega$ is not an ellipse, then there is a point of $\partial\Omega$ at which $\partial\Omega$ is not approximately regular.*

To prove this proposition, we will use the specification property of an Anosov flow, that we recall now (see [KH95]). It roughly means that given a family of pieces of orbits ($S$ below), there exists an orbit that follows these pieces.
A *specification* is a family $S = (S_i)_{i=0\cdots N}$, for some $N \in \mathbb{N}\cup\{+\infty\}$, of pairs $S_i = (w_i, I_i)$ with $w_i \in HM$, $I_i = [t_i, T_i]$, $t_i < T_i$ which satisfy $t_i > T_{i-1}$. For $T > 0$, we say that the specification $S$ is *$T$-spaced* if $t_i - T_{i-1} \geqslant T$, $i = 1\cdots N$. Given $\varepsilon > 0$, we say that the orbit of $w \in HM$ *$\varepsilon$-shadows* $S$ if for any $i = 0\cdots N$, $t \in [t_i, T_i]$, we have $d(\varphi^t(w), \varphi^t(w_i)) \leqslant \varepsilon$.

**Theorem 5.8.** *The Anosov flow $\varphi^t : HM \longrightarrow HM$ has the specification property: given $\varepsilon > 0$, there exists $T(\varepsilon)$ such that, for any $T(\varepsilon)$-spaced specification $S$, there exists a point $w \in HM$ whose orbit $\varepsilon$-shadows $S$.*

We can now give a

*Proof of Proposition 5.7.* Fix $\varepsilon > 0$, and let $T = T(\varepsilon)$ given by the last theorem.
Choose two periodic points $w_1$ and $w_2$ in $HM$, with distinct positive Lyapunov exponent $\chi_1 < \chi_2$. This is possible if $\Omega$ is not an ellipsoid, by Corollary 5.4. For $k \geqslant 0$, let $S'_k$ be the specification

$$S'_k = ((w_1, [0, 2^{2^k}]), (w_2, [T + 2^{2^k}, T + 2^{2^k} + 2^{2^{k+1}}])).$$

If $S = (w_i, [t_i, T_i])_{i=1\cdots N}$ is a specification, we set $\max S = T_N$. For $t \geqslant 0$, we denote by $t + S$ the specification $S = (w_i, [t + t_i, t + T_i])_{i=1\cdots N}$. We set $S_0 = S'_0$, $S_k = T + \max(S_{k-1}) + S'_k$, $k \geqslant 1$.We finally define the infinite specification $S$ by

$$S = (S_0, S_1, \cdots).$$

Since $S$ is $T$-spaced by construction, there is a point $w$ whose orbit $\varepsilon$-shadows $S$, and, for $Z \in E^u$, we have

$$|\limsup_{t\to-\infty} \frac{1}{t} \log \|d\varphi^t Z\| - \chi_2| < \eta(\varepsilon), \ |\liminf_{t\to+\infty} \frac{1}{t} \log \|d\varphi^t Z\| - \chi_1| < \eta(\varepsilon),$$

with $\lim_{\varepsilon\to 0} \eta(\varepsilon) = 0$. So, if $\varepsilon$ is taken so that $\eta(\varepsilon) < (\chi_2 - \chi_1)/27$, then $w$ is not forward weakly regular. Theorem 4.5 implies that the boundary $\partial\Omega$ is not approximately regular at the point $w^+$. □

Proposition 5.7 yields the following

**Corollary 5.9.** *For any $n \geqslant 2$, there exists a $\mathcal{C}^1$ strictly convex function $f : U \subset \mathbb{R}^{n-1} \longrightarrow \mathbb{R}$ which is not approximately regular at some point.*

Notice that is possible to construct by hand a function which is not approximately regular at some point, but this is somehow funny to construct one in this way.

5.5.2. *The range of Lyapunov exponents.* We now turn to the study of $\mathcal{A}$ which benefits from the following observation, which has no equivalent in dimension higher than 2. If $\mu \in \mathcal{M}$ is ergodic, the positive Lyapunov exponent of $\mu$ is given by

$$\chi(\mu) = \int \frac{d}{dt}|_{t=0} \log \|d\varphi^t\| \ d\mu,$$

hence the application $\mu \longrightarrow \chi(\mu)$ is continuous. In this case for example, proposition 5.5 is immediate.

**Proposition 5.10.** *Let $\Omega \subset \mathbb{RP}^2$ be a divisible strictly convex set. Then $\mathcal{A}$ is a closed interval.*

*Proof.* First, remark that $\mathcal{A}_{\mathcal{M}}$ is the image of the set of ergodic measures by the continuous application

$$\mu \longmapsto \alpha(\mu) = \frac{2}{\int \frac{d}{dt}|_{t=0} \log \|d\varphi^t\| \ d\mu}.$$

As the set of ergodic measures is compact, $\mathcal{A}_{\mathcal{M}}$ is compact.
We now see that $\mathcal{A}_{\mathcal{M}}$ is convex. For that, recall that $\mathcal{A}_{\mathcal{M}_{Per}}$ is dense in $\mathcal{A}_{\mathcal{M}}$. So it suffices to prove that for any $g,\ g' \in \mathcal{M}_{Per}$, $\varepsilon > 0$ and $\lambda \in [0,1]$, we can find $g_\varepsilon \in \mathcal{M}_{Per}$ so that

$$|\chi(g_\varepsilon) - (\lambda\chi(g) + (1-\lambda)\chi(g'))| < \varepsilon.$$

This is a simple application of the shadowing lemma (a particular case of the specification property, see [KH95]).
It remains to see that $\mathcal{A} = \mathcal{A}_{\mathcal{M}}$. Pick a point $w \in \Lambda$ with Lyapunov exponent $\chi(w)$. Consider the measures $\mu_T$ defined for $T > 0$ by

$$\int f \ d\mu_T = \frac{1}{T}\int_0^T f(\varphi^t w) \ dt.$$

For $f = \frac{d}{dt}|_{t=0} \log \|d\varphi^t\|$, we have

$$\lim_{T\to+\infty} \frac{1}{T}\int \log \frac{d}{dt}|_{t=0}\|d\varphi^t\| \ d\mu_T = \lim_{T\to+\infty} \frac{1}{T} \log \|d\varphi^T\| = \chi(w).$$

Hence, any accumulation point $\mu$ of the family $(\mu_T)_{T>0}$ is an invariant measure such that $\chi(\mu) = \chi(w)$. □

Remark that, in fact, the same proof would prove that

$$\mathcal{A} = \{\limsup_{t\to+\infty} \frac{1}{t} \log \|d_w\varphi\|, \ w \in \Omega\}.$$

## References


[Ale39] A. D. Alexandrov. Almost everywhere existence of the second differential of a convex function and some properties of convex surfaces connected with it. *Leningrad State Univ. Annals [Uchenye Zapiski] Math. Ser. 6*, pages 3–35, 1939.

[Ben60] Jean-Paul Benzécri. Sur les variétés localement affines et localement projectives. *Bull. Soc. Math. France*, 88:229–332, 1960.

[Ben97] Y. Benoist. Propriétés asymptotiques des groupes linéaires. *Geom. Funct. Anal.*, 7(1):1–47, 1997.

[Ben00] Y. Benoist. Automorphismes des cônes convexes. *Invent. Math.*, 141(1):149?193, 2000.

[Ben03] Y. Benoist. Convexes divisibles 2. *Duke Math. Journ.*, 120:97–120, 2003.

[Ben04] Y. Benoist. Convexes divisibles 1. *Algebraic groups and arithmetic, Tata Inst. Fund. Res. Stud. Math.*, 17:339–374, 2004.

[Ben06a] Y. Benoist. Convexes divisibles 4. *Invent. Math.*, 164:249–278, 2006.

[Ben06b] Y. Benoist. Convexes hyperboliques et quasiisométries. *Geom. Dedicata*, 122:109–134, 2006.
[Cou03] Y. Coudene. Gibbs measures on negatively curved manifolds. *J. Dynam. Control Systems*, 9(1):89–101, 2003.
[Cra09] M. Crampon. Entropies of strictly convex projective manifolds. *Journal of Modern Dynamics*, 3(4):511–547, 2009.
[Craar] M. Crampon. Lyapunov exponents in Hilbert geometry. *Ergodic Theory and Dynamical Systems*, to appear.
[Gol90] W. M. Goldman. Convex real projective structures on compact surfaces. *J. Diff. Geom.*, 31:791–845, 1990.
[JM87] D. Johnson and J. J. Millson. Deformation spaces associated to compact hyperbolic manifolds. In *Discrete groups in geometry and analysis, 1984)*, volume 67 of *Progr. Math.*, pages 48–106. Birkhäuser Boston, 1987.
[Kai90] V. A. Kaimanovich. Invariant measures of the geodesic flow and measures at infinity on negatively curved manifolds. *Ann. Inst. Henri Poincaré*, 53, n.4:361–393, 1990.
[Kal11] B. Kalinin. Livšic theorem for matrix cocycles. *Ann. of Math. (2)*, 173(2):1025–1042, 2011.
[Kap07] M. Kapovich. Convex projective structures on Gromov-Thurston manifolds. *Geom. Topol.*, 11:1777–1830, 2007.
[KH95] A. Katok and B. Hasselblatt. *Introduction to the modern theory of dynamical systems*. Cambridge Univ. Press, 1995.
[Kos68] J.-L. Koszul. Déformations de connexions localement plates. *Ann. Inst. Fourier (Grenoble)*, 18(fasc. 1):103–114, 1968.
[KV67] V. G. Kac and È. B. Vinberg. Quasi-homogeneous cones. *Mat. Zametki*, 1:347–354, 1967.
[Led95] F. Ledrappier. Structure au bord des variétés à courbure négative. In *Séminaire de Théorie Spectrale et Géométrie, No. 13, Année 1994–1995*, volume 13 of *Sémin. Théor. Spectr. Géom.*, pages 97–122. Univ. Grenoble I, 1995.
[LY85] F. Ledrappier and L.-S. Young. The metric entropy of diffeomorphisms. *Ann. of Math.*, 122:509–574, 1985.
[Mar10] L. Marquis. Espace des modules de certains polyèdres projectifs miroirs. *Geom. Dedicata*, 147:47–86, 2010.
[Mar13] L. Marquis. About groups in Hilbert geometry. In A. Papadopoulos G. Besson and M. Troyanov, editors, *Handbook of Hilbert Geometry*. European Mathematical Society Publishing House, Zürich, 2013.
[Ose68] V. I. Osedelec. A multiplicative ergodic theorem. *Trans. Moscow Math. Soc.*, 19:197–231, 1968.
[Sul79] D. Sullivan. The density at infinity of a discrete group of hyperbolic isometries. *Publ. Math. IHES*, 50:171–209, 1979.

*E-mail address*: mickael.crampon@usach.cl

Departamento de Matemática y Ciencia de la Computación, Av. Las Sophoras 173, Universidad de Santiago de Chile, Santiago de Chile

# ON THE GEOMETRY OF QUADRATIC MAPS OF THE PLANE

J. DELGADO, J.L. GARRIDO, N. ROMERO, A. ROVELLA, AND F. VILAMAJÓ

Abstract. In this article we give a geometric classification of the set of quadratic maps of the plane. The fundamental step is the proof that the restriction of the map to the critical set is injective, from which it follows that there are finitely many classes of geometrically equivalent maps. In the last sections we apply this geometric knowledge to obtain some simple dynamical properties of a particular family of quadratic maps.

## 1. Introduction

Let $Q$ be the set of quadratic self-mappings of the real plane endowed with the topology of coefficients. In [1] it is proved that six parameters are enough to describe an open and dense subset $Q_g$ of $Q$; in addition, every map in $Q_g$ without fixed points has trivial dynamics. This constitutes a version, for non-invertible mappings, of the well known Brouwer's theorem [2], which states that an orientation preserving homeomorphism of the plane having no fixed points has empty limit sets; on this topic see the article of J. Franks [4].

This paper is devoted to show a geometric classification of that open and dense set. We took adventage of this classification to analyze some interesting properties of a real one-parameter family of endomorphisms on the complex plane. The meaningful concept in our approach is the geometric equivalence of maps. We recall that two smooth maps $f, g : M \to N$ are *(geometrically) equivalent* if there exist smooth diffeomorphisms $\varphi : M \to M$ and $\psi : N \to N$ such that $f \circ \varphi = \psi \circ g$. A map is *stable* if it has a neighborhood consisting of equivalent maps. Clearly $\varphi$ (resp. $\psi$) carries critical points (resp. critical values) of $g$ to critical points (resp. critical values) of $f$; further, critical sets of equivalent maps are diffeomorphic.

We briefly describe some other geometric invariants that we will consider throughout this paper. If $f : M \to M$ is a smooth and proper map, then the number of preimages of every regular point is finite and constant in each connected component of the set of regular values of $f$. If the set of regular values of such a map $f$ has $k$ components and $a_1 \leq \cdots \leq a_k$ are the number of preimages in each one of these components, then we say that $f$ has type $(a_1, \ldots, a_k)$. We also recall that generically real planar maps have only two kind ofcritical points: *folds* and *cusps*, both having simple local canonical forms. The number of cusp points, the type of the map and the absolute value of the degree of the map are invariants of geometric equivalence. Each of these invariants is sufficient to characterize the geometric equivalence classes among the endomorphisms in $Q_g$. This is a consequence of the following proposition whose proof is contained in lemmas 1 and 2 below.

**Proposition 1.** *The restriction of $G \in Q_g$ to its critical set is injective.*

*Date*: September 29, 2013.
2010 *Mathematics Subject Classification.* 58K05, 37C05.

In the above referred lemmas it is also proved the existence of just two classes of geometric equivalent maps in the generic set $Q_g$; in addition, it can be proved that there exist finitely many classes of geometrically equivalent quadratic maps (see at the beginning of paragraph 6.1 in section 6). The proof of these lemmas rest on geometrical objects that represent the set of critical values of quadratic maps in that equivalent classes: deltoids and hypdeltoids. Deltoids, also called 3-cusped hypocycloids, were first studied by Euler in 1745 while considering optical problems, they have a simple parametrization with sine and cosine functions. Dual parametrizations with hyperbolic sine and cosine functions give rise to a geometrical object that we called hypdeltoids. The striking property of these curves in our context is that they describe the sets of critical values and its preimages, which gives an accurate geometric description of maps in $Q_g$:

**Theorem 1.** *For the open and dense set $Q_g$ the following properties hold:*

*(i) For every $G \in Q_g$, the point at $\infty$ is an attractor.*
*(ii) Every map in $Q_g$ is geometrically stable.*
*(iii) There exist only two classes, $Q_+$ and $Q_-$, of geometric equivalence in $Q_g$.*
*(iv) Every $G \in Q_-$ is of type $(2, 4)$, has degree $\pm 2$ and the set of critical points is an ellipse containing exactly three cusp points.*
*(v) Every $G \in Q_+$ is of type $(0, 2, 4)$, have degree $0$ and the set of critical points is a hyperbola containing exactly one cusp point.*

## 2. The generic set $Q_g$

Consider the set of all real planar maps defined, for every $(x, y) \in \mathbb{R}^2$, by

$$G(x, y) = (pxy + ax + by + k_1, rx^2 + sy^2 + txy + cx + dy + k_2), \tag{1}$$

where $prs \neq 0$. It was proved in [1] that the set of maps affinely conjugated to a map of this form is open and dense in $Q$. Let $Q_g$ be the set of maps $G_0$ satisfying:

- $G_0$ is affinely conjugated to a map of the form (1);
- The critical set of $G_0$ is either an ellipse or a hyperbola.

It is easy to see that $Q_g$ is open and dense in $Q$. Additionally, note that $G_0 \in Q_g$ has an ellipse (resp. a hyperbola) as its critical set if, and only if, there is a $G$ as in (1) with $rs < 0$ (resp. $rs > 0$) and affinely conjugated to $G_0$. This property splits $Q_g$ into two disjoint subsets: $Q_-$, consisting of maps in $Q_g$ whose the critical set is an ellipse, and $Q_+$, consisting of those maps whose critical set is a hyperbola.

Take $G_0 \in Q_-$ and $G$ as in (1) which is affinely conjugated to $G_0$. After the change of variables $(X, Y) = \left(\sqrt{-rs}\, x, -sy\right)$ and an appropriate traslation, the map $G$ is written as:

$$G(x, y) = \left(pxy + ax + by + k_1, x^2 - y^2 + txy + \frac{at - 2b}{p}x + \frac{bt + 2a}{p}y + k_2\right). \tag{2}$$

Let $\Theta_-$ be the family of maps in $Q_-$ and defined as in (2). Notice that if $G \in \Theta_-$ is as above, then its critical set $\ell$ is given by the circle with Cartesian equation $x^2 + y^2 = (a^2 + b^2)/p^2$; obviously $a^2 + b^2 > 0$. We refer this kind of maps as the normal form for $Q_-$.

In analogous way, maps in $Q_+$ are affinely conjugated to a map of the family $\Theta_+$ given by the normal form:

$$G(x, y) = \left(pxy + ax + by + k_1, x^2 + y^2 + txy + \frac{at - 2b}{p}x + \frac{bt - 2a}{p}y + k_2\right), \tag{3}$$

whose critical set $\ell$ is the hyperbola given by $y^2 - x^2 = (a^2 - b^2)/p^2$; note that $a^2 > b^2$, otherwise $G \notin Q_+$.

## 3. Deltoids and Hypdeltoids

We begin this section by recalling generic properties related to critical sets of smooth planar maps and singularities of smooth parametrized curves in $\mathbb{R}^2$.

Take a smooth map $G : \mathbb{R}^2 \to \mathbb{R}^2$. The following notions and statements were introduced by H. Whitney in [9]. The map $G$ is said to be *good* if every point $p \in U$ is either regular or the gradient of the Jacobian matrix of $G$ at $p$ is non-null. If $G$ is a good map, then its critical set $\ell$ is a 1-manifold. In this case, if $p \in \ell$ and $\varphi$ is a parametrization of $\ell$ around $p$ ($\varphi(0) = p$), then this critical point is called a *fold point of* $G$ if $d(G \circ \varphi)/dt \neq 0$ at $t = 0$ and $p$ is a *cusp point of* $G$ whenever $d(G \circ \varphi)/dt = 0$ and $d^2(G \circ \varphi)/dt^2 \neq 0$ at $t = 0$; these definitions are independent of the choice of the parametrization. In that seminal article Whitney found a generic set of good maps: the set of *excellent* maps, which are characterized by the fact that the critical set is only composed by fold or cusp points. Furthermore, local normal forms for these critical points were constructed. If $p$ is a fold point, then the map $G$ is equivalent, in some neighborhood of $p$, to the map $(x, y) \mapsto (x^2, y)$ in a neighborhood of the origin; so $G$ is locally of type $(0, 2)$. For cusp points the normal form is given by $(x, y) \mapsto (xy - x^3, y)$, which implies that cusp points are isolated and the mapping is of type $(1, 3)$ around $p$. It is proved in [7] that the restriction of a generic map $G$ to any component of the complement of $G^{-1}(G(\ell))$ is a covering map whose image is a component of the complement of $G(\ell)$; see Lemma 3 in section 4. Therefore, determining the critical sets, the critical values and the preimages of the critical values is essential in the description of the geometry of a generic map.

Additionally to the notion of cusp point as critical point of smooth maps we deal with cusp points as singularities of plane smooth curves. In order to recall this notion we consider a parametrized smooth curve $\gamma(t) = (x(t), y(t))$, where $t$ is varying in an open interval. Take a singular point $p$ on this curve, that is, $p = (x(t_0), y(t_0))$ for some $t_0 \in I$, and $x'(t_0) = y'(t_0) = 0$. Hence it holds that

$$x(t) = x_0 + a(t - t_0)^2 + b(t - t_0)^3 + R_1(t), \text{ and}$$
$$y(t) = y_0 + c(t - t_0)^2 + d(t - t_0)^3 + R_2(t),$$

where $R_i(t)/(t - t_0)^3 \to 0$ when $t \to t_0$, $i = 1, 2$. Assuming $a^2 + c^2 > 0$, the curve $\gamma$ is tangent to the line through $p$ with slope $c/a$ if $a \neq 0$, and it is tangent to the vertical line $x = x_0$ at that point when $a = 0$. Observe that this assumption implies that near $p$ the curve is injective and this singularity is isolated. The singular point $p$ is said to be an *ordinary cusp* (or simply a cusp) on $\gamma$ when $ad - bc \neq 0$. It is easy to see that under this open condition on the derivatives of second and third order, the two branches of $\gamma$ near $p$, that is $\{\gamma(t) : t < t_0\}$ and $\{\gamma(t) : t > t_0\}$ with $|t - t_0|$ small, are located in different sides of the tangent line. The same notion of cusp point on simple and piecewise regular curves is introduced in [3].

### 3.1. Deltoids and maps in $\Theta_-$.

**Definition 1.** For $\alpha \in [0, 2\pi)$, the regular *$\alpha$-deltoid* (*deltoid*, for short) is the parametrized smooth closed curve $\Delta_\alpha$ given by:

$$\Delta_\alpha(\omega) = (\sin(2\omega) + 2\sin(\omega + \alpha), \cos(2\omega) - 2\cos(\omega + \alpha)), \quad \omega \in [0, 2\pi). \tag{4}$$

Notice that $\Delta_\alpha(\omega) = ie^{-2i\omega} - 2ie^{i(\omega+\alpha)}$. In this way it is easy to verify that:
• The function $\Delta_\alpha : [0, 2\pi) \to \mathbb{R}^2$ is injective; hence $\Delta_\alpha$ is a closed simple curve.
• Only at $\omega = (\pi - \alpha)/3, (3\pi - \alpha)/3, (5\pi - \alpha)/3$ it holds $d\Delta_\alpha(\omega)/dw = 0$. That is, $\Delta_\alpha$ has three singularities. Since $d^2\Delta_\alpha(\omega)/dw^2 \neq 0$ for all $\omega \in [0, 2\pi)$ and the imaginary part of the product $d^2\Delta_\alpha(\omega)/dw^2 . \overline{d^3\Delta_\alpha(\omega)/dw^3}$ is non-null at the values where $d\Delta_\alpha(\omega)/dw = 0$, then that three singularities are cusp on the deltoid $\Delta_\alpha$. Here $\overline{z}$ denotes the conjugate of the complex numbre $z$.

An implicit Cartesian equation of $\Delta_\alpha$ can be obtained by eliminating the variable $\omega$ in the equations $x = \sin(2\omega) + 2\sin(\omega + \alpha)$ and $y = \cos(2\omega) - 2\cos(\omega + \alpha)$. Indeed, with the proceduce described in [8, p. 206] one arrives to $D_\alpha(x, y) = 0$, where

$$D_\alpha(x, y) = (x^2 + y^2)(x^2 + y^2 + 18) + 8x(3y^2 - x^2)\sin(2\alpha) + 8y(3x^2 - y^2)\cos(2\alpha) - 27.$$

It is simple to check that for all $\alpha \in [0, 2\pi)$ the function $D_\alpha$ satisfies

$$D_\alpha = D_0 \circ J \circ R_{-2\alpha/3}, \tag{5}$$

where $J$ is the reflection with respect to the vertical axis and $R_{-2\alpha/3}$ is the rotation by angle $-2\alpha/3$.

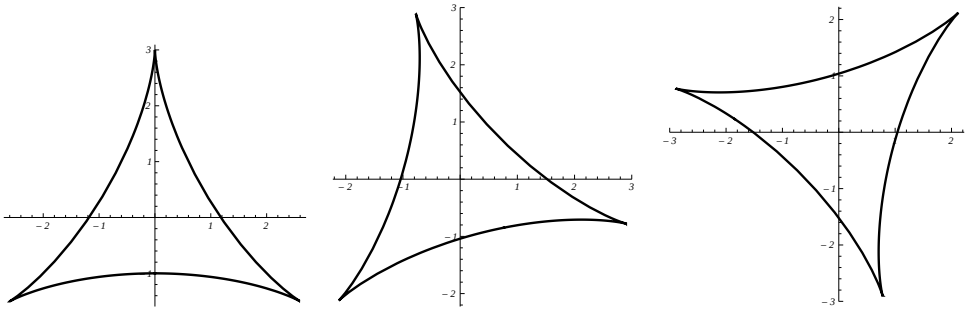

FIGURE 1. Regular deltoids with $\alpha = 0, \frac{\pi}{8}, \frac{13\pi}{8}$.

Take $G \in \Theta_-$ as in equation (2). Recall that its critical set $\ell$ is the ellipse given by $x^2 + y^2 = \rho^2/p^2$, where $\rho = \sqrt{a^2 + b^2}$. Fix $\alpha \in [0, 2\pi)$ such that

$$(a, b) = \rho(\cos(\alpha), \sin(\alpha)), \tag{6}$$

and parametrize $\ell$ as

$$\ell(\omega) = \frac{\rho}{p}(\sin(\omega), \cos(\omega)),\ \omega \in [0, 2\pi). \tag{7}$$

With these considerations $G(\ell(\omega))$ is written, for all $\omega \in [0, 2\pi)$, as:

$$G(\ell(\omega)) = A\left(\frac{\rho^2}{p^2}\Delta_\alpha(\omega)\right), \tag{8}$$

where $\alpha$ and $\rho$ are given by (6) and $A$ is the affine bijection:

$$A(x, y) = \frac{1}{2}((p, t)x + (0, -2)y) + (k_1, k_2). \tag{9}$$

Notice that (8) and the existence of the three cusp points on $\Delta_\alpha$ imply that $G$ has only three critical point of cusp type: $\ell((\pi - \alpha)/3), \ell((3\pi - \alpha)/3)$ and $\ell((5\pi - \alpha)/3)$. It is also concluded that the restriction of $G$ to $\ell$ is an injective function.

Now we will analize the preimage under $G$ of $G(\ell)$. First, it is clear that $(x,y)$ belongs to $G^{-1}(G(\ell))$ if and only if $\frac{p^2}{\rho^2}A^{-1}G(x,y) \in \Delta_\alpha$; that is,

$$\frac{p^2}{\rho^2}\left(2xy + \frac{2a}{p}x + \frac{2b}{p}y, -x^2 + y^2 + \frac{2b}{p}x - \frac{2a}{p}y\right) \in D_\alpha^{-1}(0).$$

Introducing the change of variable $(X,Y) = \frac{p}{\rho}(x,y)$, using (6) and defining

$$H(X,Y) = (2XY + 2X\cos(\alpha) + 2Y\sin(\alpha), -X^2 + Y^2 + 2X\sin(\alpha) - 2Y\cos(\alpha)),$$

it follows that $(x,y) \in G^{-1}(G(\ell))$ if and only if $(D_\alpha \circ H)(X,Y) = 0$. A straightforward calculation leads to the identity

$$(D_\alpha \circ H \circ R_{-\alpha/3})(X,Y) = (X^2+Y^2-1)^2 D_\alpha(X,Y).$$

From (5) one obtains $D_\alpha \circ R_{\alpha/3} = D_{\alpha/2}$, consequently

$$(D_\alpha \circ H)(X,Y) = (X^2+Y^2-1)^2 D_{\alpha/2}(X,Y).$$

This implies that $G^{-1}(G(\ell)) = \ell \cup \widetilde{\ell}$, where $\widetilde{\ell}$ is the deltoid obtained as the homothetic transformation with scale $\rho/p$ of the deltoid $\Delta_{\alpha/2}$. From this fact one can verify that $\ell$ is contained in the closure of the bounded component of the complement of $\widetilde{\ell}$; moreover, $\ell$ and $\widetilde{\ell}$ are tangent at the three cusp points in $\ell$.

The following lemma summarizes the preceding discussion.

**Lemma 1.** *If $G \in \Theta_-$, then its critical set $\ell$ is a circle having exactly three cusp points, the restriction of $G$ to $\ell$ is injective, the set $G(\ell)$ is a deltoid and $G^{-1}(G(\ell))$ is the union of $\ell$ and another deltoid $\widetilde{\ell}$ which is tangent to $\ell$ at the three critical points of cusp type.*

3.2. **Hypdeltoids and maps in $\Theta_+$.** Now we will proceed in very similar way as above to analize the geometry of the set of critical values of maps in $\Theta_+$.

**Definition 2.** Given $\alpha \in \mathbb{R}$, the regular *$\alpha$-hypdeltoid* (*hypdeltoid*, for short) is the pair of parametrized curves $\Lambda_\alpha^\pm$ given by:

$$\Lambda_\alpha^\pm(\omega) = (\sinh(2\omega) \pm 2\sinh(\omega+\alpha), -\cosh(2\omega) \pm 2\cosh(\omega+\alpha)),\ \omega \in \mathbb{R}. \tag{10}$$

For $i = 1,2$ and $\sigma = \pm$ we denote by $\varphi_i^\sigma(\omega)$ the $i$th-coordinate of $\Lambda_\alpha^\sigma(\omega)$. Suppose that for $\omega, \omega' \in \mathbb{R}$ are satisfied $\varphi_1^+(\omega) = \varphi_1^-(\omega')$ and $\varphi_2^+(\omega) = \varphi_2^-(\omega')$. This implies that $-\cosh(3\omega+\alpha) = \cosh(3\omega'+\alpha)$, which occurs when $\omega = \omega' = -\frac{\alpha}{3}$; but $\varphi_2^+(-\frac{\alpha}{3}) \neq \varphi_2^-(-\frac{\alpha}{3})$. Thus, the branches $\Lambda_\alpha^-$ and $\Lambda_\alpha^+$ are disjoint. On the other hand, since $\varphi_1^+$ is a function onto $\mathbb{R}$ and $d\varphi_1^+(\omega)/d\omega \neq 0$, it follows that $\Lambda_\alpha^+$ is an embbeding of $\mathbb{R}$; indeed, it is the graph of a smooth function. With respect to the branch $\Lambda_\alpha^-$, it is easy to see that $d\varphi_1^-(\omega)/d\omega = d\varphi_2^-(\omega)/d\omega$ if and only if $\omega = -\frac{\alpha}{3}$. Hence, on $\Lambda_\alpha^-$ there is only one singularity; moreover, by analyzing the values of the second and third derivatives of $\varphi_i^-$ at $\omega = -\frac{\alpha}{3}$ we conclude that this singularity is a cusp point. Furthermore, as $\varphi_1^-$ is a function onto $\mathbb{R}$, $d^2\varphi_2^-(\omega)/d\omega^2 < 0$ and $d\varphi_1^-(\omega)/d\omega = 0$ exactly at $\omega = -\frac{\alpha}{3}$ and $\omega = \alpha$, then the function $\omega \mapsto \Lambda_\alpha^-(\omega)$ is injective, and the branch $\Lambda_\alpha^-$ is topological immersion of $\mathbb{R}$.

Now we will obtain a Cartesian equation for $\Lambda_\alpha^\pm$. First we introduce

$$x = \sinh(2\omega) \pm 2\sinh(\omega) \text{ and } y = -\cosh(2\omega) \pm 2\cosh(\omega). \tag{11}$$

Since $y^2 - x^2 = 5 \mp 4\cosh(3\omega)$, the equation on the right side of (11) implies that $u = \frac{1}{2}(\sqrt{3-2y} \pm 1)$, by setting $u = \cosh(\omega)$. But $\cosh(3\omega) = 4u^3 - 3u$, then

$$(y^2-x^2)(y^2-x^2+18) + 8y(3x^2+y^2) - 27 = 0$$

is an implicit Cartesian equation of $\Lambda_0^{\pm}$. Thanks to this Cartesian representation and the identity

$$\Lambda_\alpha^{\pm}(\omega) = \varphi(\omega) + B_\alpha(\Lambda_0^{\pm}(\omega) - \varphi(\omega)),$$

where $\varphi(\omega) = (\sinh(2\omega), -\cosh(2\omega))$ and $B_\alpha$ is the linear map given by the matrix $\begin{pmatrix}\cosh(\alpha) & \sinh(\alpha)\\ \sinh(\alpha) & \cosh(\alpha)\end{pmatrix}$, we get (after a tedious computation) that the zero set of

$$H_\alpha(x,y) = (x^2-y^2)(x^2-y^2-18) - 8x(x^2+3y^2)\sinh(2\alpha) + 8y(3x^2+y^2)\cosh(2\alpha) - 27$$

is a the Cartesian description of $\Lambda_\alpha^{\pm}$. It is simple to check that $H_\alpha = H_0 \circ S_\alpha$, where $S_\alpha$ is the linear isomorphisms given by $\begin{pmatrix}\cosh(2\alpha/3) & -\sinh(2\alpha/3)\\ -\sinh(2\alpha/3) & \cosh(2\alpha/3)\end{pmatrix}$.

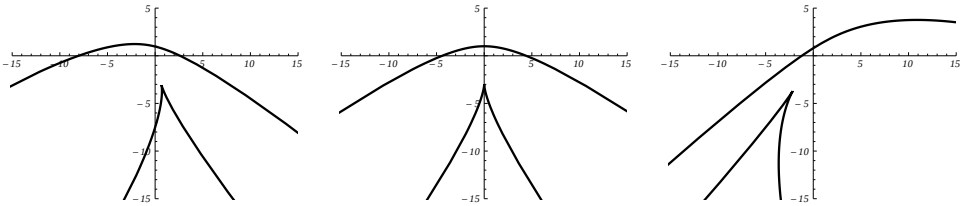

FIGURE 2. Hypdeltoids with $\alpha = -\frac{1}{3}, 0, 1$.

Take $G \in \Theta_+$ as in (3), recall that its critical set $\ell$ is given by the equation $y^2 - x^2 = \rho^2/p^2$, where $\rho = \sqrt{a^2-b^2}$ and $|a| > |b|$. Consider $\alpha \in \mathbb{R}$ such that the coefficients $a$ and $b$ in (3) satisfy

$$(a,b) = \rho(\cosh\alpha, \sinh\alpha). \tag{12}$$

We parametrize the branches $\ell_\pm$ of $\ell$ by

$$\ell_\pm(\omega) = \frac{\rho}{p}(\sinh(\omega), \pm\cosh(\omega)),\ \omega \in \mathbb{R}.$$

Then the image by $G$ of $\ell_\pm(\omega)$ is expressed as $G(\ell_\pm(\omega)) = A\left(\frac{\rho^2}{p^2}\Lambda_\alpha^{\pm}(\omega)\right)$, where $A$ is defined in (9). This expression allows to conclude that:

- The map $G$ restricted to each branch $\ell_\pm(\omega)$ is an injective function.
- There is only one cusp point in the critical set of $G$, which belongs to $\ell_-$. The remaining critical points are all of the fold type.
- A point $(x,y) \in G^{-1}(G(\ell))$ if and only if $\frac{p^2}{\rho^2}A^{-1}G(x,y) \in \Lambda_\alpha^{\pm}$, that is

$$\frac{p^2}{\rho^2}\left(2xy + \frac{2a}{p}x + \frac{2b}{p}y, -x^2 - y^2 + \frac{2b}{p}x + \frac{2a}{p}y\right) \in H_\alpha^{-1}(0).$$

Making $(X,Y) = \frac{p}{\rho}(x,y)$, equation (12) implies that $(x,y) \in G^{-1}(G(\ell))$ if and only if $(H_\alpha \circ h)(X,Y) = 0$, where

$$h(X,Y) = (2XY + 2\cosh(\alpha)X + 2\sinh(\alpha)Y, -X^2 - Y^2 + 2\sinh(\alpha)X + 2\cosh(\alpha)Y).$$

It can be checked that for all $X, Y \in \mathbb{R}$ it holds

$$(H_\alpha \circ h)(X,Y) = (1 + X^2 - Y^2)^2 H_{\alpha/2}(X, -Y).$$

Thus, the zero set of the polynomial $(H_\alpha \circ h)(X,Y)$ is the union of the hyperbola $1+X^2-Y^2=0$ and the hypdeltoid $H_{\alpha/2}(X,-Y)=0$. Therefore, $G^{-1}(G(\ell))$ is the union of the hyperbola $\ell$ and the hypdeltoid obtained as the homethetic transformation with scale $\rho/p$ of the reflection respect to the horizontal axis of the hypdeltoid $\Lambda^{\pm}_{\alpha/2}$.

We summarize the precedent exposition in the following lemma.

**Lemma 2.** *If $G \in \Theta_+$, then its critical set $\ell$ is a hyperbola containing only one cusp point, the mapping $G$ is injective when it is restricted to $\ell$, the set of critical values $G(\ell)$ is a hypdeltoid and its preimage is the union of $\ell$ and a hypdeltoid $\widetilde{\ell}$.*

## 4. Proof of Theorem 1

As all the statements in Theorem 1 are invariant under affine conjugation, we only consider generic quadratic maps.

*Proof of part (i) of Theorem 1.* Take a generic map $G$ as in equation (1), that is

$$G(x,y) = (pxy + ax + by + k_1, rx^2 + sy^2 + txy + cx + dy + k_2),$$

with $prs \neq 0$, by simplicity we assume $p > 0$. Let $|(x,y)| = \max\{|x|,|y|\}$. We show that there exists $K_0 > 0$ depending only on $G$ such that, for $K > K_0$ the condition $|(x,y)| > K$ implies $|G(x,y)| > 2K$. So it is clear that $\infty$ is an attracting fixed point for $G$. Indeed, assume that $|(x,y)| > K$ and $|x| \geq |y|$. If $|pxy + ax + by + k_1| < 2K$ and $K$ is large enough, then:

$$\begin{aligned} |y| &< \frac{2K + |ax + k_1|}{|px + b|} \leq \frac{2K + |a||x| + |k_1|}{p|x| - |b|} \\ &\leq \frac{2K + |k_1|}{pK - |b|} + \frac{|a|}{p - |b|/K} \leq \frac{3}{p} + \frac{2|a|}{p} \leq \frac{3 + 2|a|}{p}. \end{aligned}$$

This inequality implies that:

$$|G(x,y)| \geq |r|x^2 - \left( \left| \frac{t(3+2|a|)}{p} + c \right| \right) |x| - \frac{|s|(3+2|a|)^2}{p^2} - \frac{|d|(3+2|a|)}{p} - k_2;$$

since $r \neq 0$ it follows that $|G(x,y)| > 2K$ if $K$ is sufficiently large and $|x| > K$. The proof for the case $|y| \geq |x|$ is similar. □

For the proof of the other parts of theorem 1 we will use the following result, which can be found in [7].

**Lemma 3.** *Let $G$ be a smooth proper map on a manifold $M$. The restriction of $G$ to any component of the complement of $G^{-1}(G(\ell))$ is a covering map whose image is a component of the complement of $G(\ell)$.*

The proof of this lemma is based on the fact that every point $y$ in $G(C)$ has finitely many preimages, where $C$ is a component of the complement of $G^{-1}(G(\ell))$. Then the result holds even if the set of critical points is not bounded. Note that part (i) proved above implies that the restriction of $G$ to a component of $G^{-1}(G(\ell))$ is a proper map.

*Proof of part (iv) of Theorem 1.* Let $G$ be a map in $\Theta_-$ as in (2). Denote by $c_1, c_2$ and $c_3$ the cusp points of $G$. Besides the injectivity of $G$ when restricted to $\ell$, Lemma 1 describes the way as the sets $\ell$, $G(\ell)$ and $G^{-1}(G(\ell)) = \ell \cup \widetilde{\ell}$ are displayed, just as figure 3 shows.

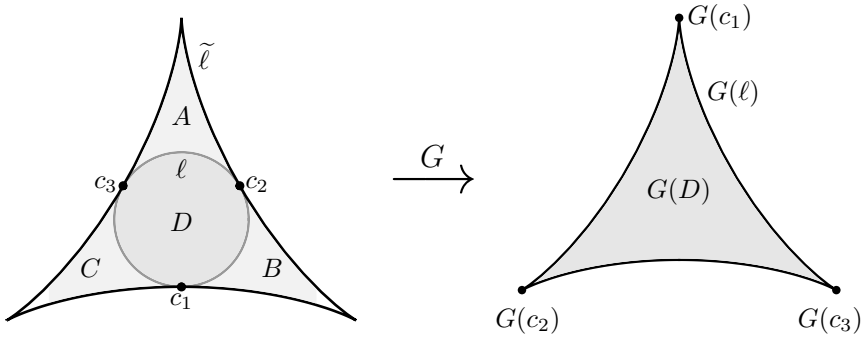


FIGURE 3. Critical set, critical values and its preimage for maps in $\Theta_-$.

Note that the regions $A$, $B$, $C$ and $D$ are topological discs and constitute the bounded components of the complement of $G^{-1}(G(\ell))$. It follows from Lemma 3 that the restriction of $G$ to each of these regions is a homeomorphism onto the bounded component of the complement of $G(\ell)$.

**Claim:** *Every point in the unbounded component of the complement of $G(\ell)$ has two preimages.*

Note that the first coordinate of $G(x, y)$ can be made $pxy + by + u$ by a translation in the second coordinate. The preimage of a point $(u, v)$ (with $v > 0$ large enough) satisfies $y = 0$ or $x = -b/p$. Substituting $x = -b/p$ in the second coordinate of $G$, and assuming $v$ large, there exist two solutions for $y$. On the other hand, substituting $y = 0$ in the second coordinate of $G(x, y)$, and taking $v$ large, it comes that two solutions for $x$ exist because $v > 0$. Hence, from Lemma 3, the restriction of $G$ to the unbounded component $E$ of the complement of $G^{-1}(G(\ell))$ is a two-to-one covering of the unbounded component of the complement of $G(\ell)$.

It remains to calculate the degree of $G$. As in the claim, take $(u, v)$ with $v > 0$ large enough and having preimages $(x_\pm, 0)$. The determinant of $DG$ at these points has the same sign of $-v$, so the degree is $-2$. □

*Proof of part (v) of Theorem 1.* Consider a map $G \in \Theta_+$ whose critical set is the hyperbola $\ell$. Let $\ell_1$ and $\ell_2$ be the branches of $\ell$; we assume that $\ell_1$ contains the unique cusp point $c_1$ of $G$, the remaining critical points of $G$ are fold points. As $G(\ell_1)$ and $G(\ell_2)$ are the branches of the hypdeltoid $G(\ell)$ (see Lemma 2), it follows that the set of regular values has three components: $Y_0, Y_2$ and $Y_4$. As in the proof of part (iv) above, one can take now a point of the form $(u, -v)$ with $v$ large enough to conclude that there are points with no preimages. It follows immediately that the degree of $G$ is zero in this case. Moreover, by the normal form at cusp points (see the remarks at the beginning of Section 3) there exists a basis of neighborhoods of a cusp whose images are open. Therefore, exactly one of the three regions contained in the complement of $G(\ell)$ has no preimage under $G$; this component will be denoted by $Y_0$. It follows also that the boundary of $Y_0$ is equal to $G(\ell_2)$, because there cannot be images of cusps points in the boundary of $Y_0$. Then denote by $Y_2$ the other region having $G(\ell_2)$ in its boundary. Points in $Y_2$ have two preimages because passing through $G(\ell_2)$ from $Y_0$ to $Y_2$ means an increasing in two units of the number of preimages; this follows by using the normal form at fold points. With similar arguments using the normal form at cusp points it

follows that every point in the region $Y_4$ whose boundary is $G(\ell_1)$, has four preimages ($\ell_1$ contains the cusp point).

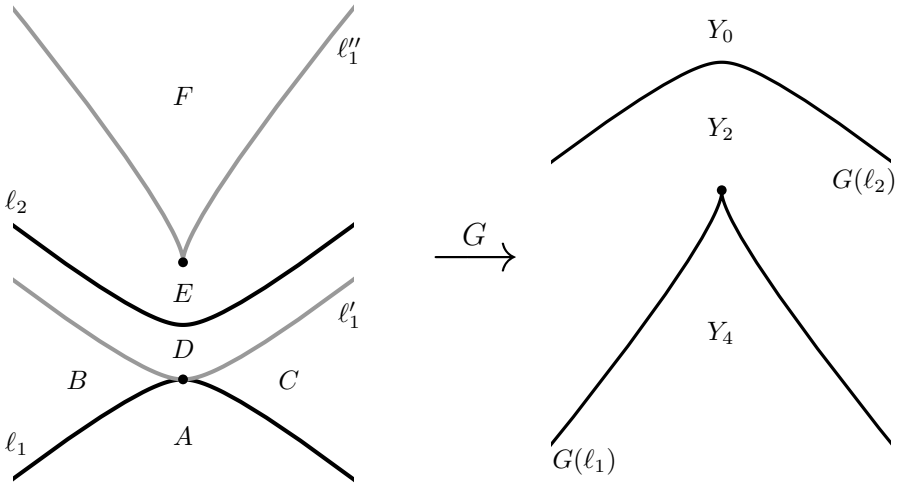


FIGURE 4. Critical set, critical values and its preimage for maps in $\Theta_+$.

It remains to describe the set $G^{-1}(G(\ell))$. Note that $G^{-1}(G(\ell_2))$ has only one component, the contrary assumption would imply that points in $Y_0$ have preimages. It follows that the hypdeltoid $\widetilde{\ell} = \ell_1' \cup \ell_1''$ in the preimage of $G(\ell)$ (see Lemma 2) is contained in $G^{-1}(G(\ell_1))$. Then $G^{-1}(G(\ell_1)) = \ell_1 \cup \ell_1' \cup \ell_1''$. Now we want to determine the location of the branches of $\widetilde{\ell}$. First note that as $\ell_1$ has a cusp, then using again normal forms, it comes that one of the branches of $\widetilde{\ell}$, say $\ell_1'$, is tangent to $\ell_1$ at $c_1$. Now take a simple curve $\gamma$ joining $G(\ell_1) \setminus \{G(c_1)\}$ to $G(\ell_2)$ and whose interior is contained in $Y_2$. It is claimed that $G^{-1}(\gamma)$ satisfies the following properties:

(1) Its interior is a simple arc, denoted from now on as $\gamma'$.
(2) One of the extreme points of $\gamma'$ belongs to $\ell_1'$, the other one belongs to $\ell_1''$.
(3) The preimage of $\gamma \cap Y_2$ does not intersect $\ell_1$.

Proof of these assertions: (1) Note that the two preimages of points in $Y_2$ are located at different sides of $\ell_2$. This is because $\ell_2$ only contains fold points. Recall from the normal form at a fold type critical point that the preimage of a simple curve intersecting $\ell_2$ at just one point is a simple curve.
(2) and (3). Note that $\ell_1'$ cannot intersect $\ell_2$ because their images are disjoint, recall that $G|_\ell$ is injective. The same thing occurs with $\ell_1''$ and $\ell_2$. Hence $\gamma'$ cannot have both extreme points in the same component of the preimage of $G(\ell)$. So, to complete the proof of both (2) and (3) it remains to show that its end points cannot belong to $\ell_1$. Assume that one of the extreme points of $\gamma'$ belongs to $\ell_1$. As $\ell_1$ is a set whose points (excepting $c_1$) are critical points of fold type, then points in $\gamma \cap Y_2$ would have preimages at both sides of $\ell_1$, but then these points would have more than two preimages contradicting the definition of $Y_2$.

Supported on these arguments we conclude that the complement of $G^{-1}(G(\ell))$ is the union of six regions: $A$, $B$, $C$, $D$, $E$ and $F$. The restrictions of $G$ to $A$, $B$, $C$ and $F$ are homeomorphisms onto the region $Y_4$, while the restrictions of $G$ to $D$ and $E$ are homeomorphisms onto the region $Y_2$; see figure 4. □

*Proof of part (ii) of Theorem 1.* For the proof of this parte one must find, for any perturbation $\widetilde{G}$ of $G$, diffeomorphisms $\varphi$ and $\psi$ such that $\psi \circ G = \widetilde{G} \circ \varphi$. Consider first a map $G \in \Theta_-$. It was shown above that this map has the geometrical structure described in figure 3. What must be shown now is that any perturbation $\widetilde{G}$ of $G$ has the same geometrical structure, that is, the set of critical points of $G$ and $\widetilde{G}$ must be diffeomorphic, as well as the sets of critical values; moreover, the number of components of the complement of $G^{-1}(G(\ell))$ must remain unchanged after perturbation. Indeed, assume that we have proved that for $\widetilde{G}$ the following picture is realized:

(1) The set of critical points $\ell$ of $\widetilde{G}$ is diffeomorphic to a circle and contains exactly three cusps.
(2) The image $\widetilde{G}(\ell)$ of $\ell$ is a simple closed curve of class $C^1$ except at the image of the cusp points of $\widetilde{G}$.
(3) The preimage of $\widetilde{G}(\ell)$ is equal to the union of $\ell$ and another simple closed curve $\delta$ which is of class $C^1$ except at three points. Moreover, $\ell$ is contained in the bounded component of the complement of $\delta$ except at the cusps of $\ell$, where a tangency between $\ell$ and $\delta$ occurs.
(4) The complement of the preimage of $\widetilde{G}(\ell)$ is equal to the union of five regions, the map $\widetilde{G}$ is injective in each one of the four bounded regions, and it is two-to-one in the unbounded one.

With these properties at hand, one can easily construct the diffeomorphisms making the equivalence. Begin with a diffeomorphism $\varphi$ carrying the closure of the bounded component $D$ of the complement of $\ell(G)$ to the closure of the bounded component $\widetilde{D}$ of the complement of $\ell(\widetilde{G})$. By property (1) above it is obvious that this can be done with the additional assumption that $\varphi$ carries cusps to cusps. Denote by $c_1, c_2, c_3$ the cusps of $G$ and by $c_i' = \varphi(c_i)$, $i = 1, 2, 3$. Properties (2) and (3) above imply that the bounded components of the complement of $\widetilde{G}^{-1}(\widetilde{G}(\ell))$ are four: $\widetilde{A}$, $\widetilde{B}$, $\widetilde{C}$ and $\widetilde{D}$, they are labeled as in figure 3; that is, $\widetilde{A}$ is the region containing in its boundary $c_2'$ and $c_3'$, $\widetilde{B}$ is the component containing in its boundary $c_2'$ and $c_1'$, $\widetilde{C}$ is the component containing $c_1'$ and $c_3'$, and $\widetilde{D}$ as described above. Use corresponding notations $(A, B, C, D)$ for the components of the complement of the preimage of $G(\ell(G))$. We proceed to extend $\varphi$ as follows: for a point $x \in A$, there exists a unique point in $y \in D$ such that $G(y) = G(x)$. Then define $\varphi(x)$ as the unique point in $\widetilde{x} \in \widetilde{A}$ such that $\widetilde{G}(\widetilde{x}) = \widetilde{G}(\varphi(y))$. Note that $\varphi$ was defined in $A$ as

$$\varphi = (\widetilde{G}|_{\widetilde{A}})^{-1} \circ \widetilde{G}|_{\widetilde{D}} \circ \varphi \circ (G|_D)^{-1} \circ G|_A.$$

Similar extension to $B$ and $C$. Thus, $\varphi$ is defined in the closure of the union $A \cup B \cup C \cup D$. Observe that the equation above implies that $\varphi$ is differentiable in the union of the interiors of these regions. It is also is smooth in the boundary of $D$. In common boundaries it is well defined because the common boundaries are critical points and $\varphi$ satisfies the symmetric property:

$$G(x) = G(y) \text{ implies } \widetilde{G}(\varphi(x)) = \widetilde{G}(\varphi(y)). \tag{13}$$

This formula also implies the smoothness of $\varphi$ at those boundaries. It remains to define $\varphi$ in $E$, the unbounded component of the complement of $G^{-1}(G(\ell(G)))$. That is, $\varphi$ must be any diffeomorphism from $E$ onto $\widetilde{E}$ with prescribed boundary values, and such that the symmetric property holds. To construct this, let $L$ be an unbounded simple line starting at $G(c_1)$, and note that the preimage of $L$ under $G$ has two components: $L_1$ and

$L_2$, recall that $L$ is simple and $G|_E$ is a covering of the annulus. Note that $G^{-1}(G(c_1))$ has two preimages, one of which is $c_1$, and assume that $L_1$ has $c_1$ as its unique extreme point. Let $\widetilde{L}_i$ $(i = 1, 2)$ be equally constructed for $\widetilde{G}$; as above $c_1'$ is the extreme point of $\widetilde{L}_1$. Let $\varphi$ be any diffeomorphism from $L_1$ to $\widetilde{L}_1$. Then extend $\varphi$ to $L_2$, making as before: for $x_2 \in L_2$ there is a unique $x_1$ in $L_1$ such that $G(x_2) = x_1$, hence one can define $\varphi(x_2)$ as the unique point $\widetilde{x}_2$ in $\widetilde{L}_2$ such that $\widetilde{G}(\widetilde{x}_2) = \widetilde{G}(\varphi(x_1))$. Observe that $\varphi$ was defined twice at the points $G^{-1}(G(c_1))$, but both definitions coincide. As $E$ is an annulus, it follows that $E \setminus (L_1 \cup L_2)$ equals the union of two connected components $V_1$ and $V_2$, and that $G$ is injective in each $V_i$. Define $\widetilde{V}_1$ and $\widetilde{V}_2$ in similar way. Note that $V_1$ and $\widetilde{V}_1$ are half-planes and that $\varphi$ was already defined as a diffeomorphism from the boundary of $V_1$ onto the boundary of $\widetilde{V}_1$. It is easy then to extend $\varphi$ to a diffeomorphism from $V_1$ onto $\widetilde{V}_1$. Making the same trick as above, extend $\varphi$ to the whole $L_2$.

At this point, we have constructed a diffeomorphism $\varphi$ from the plane onto itself satisfying the symmetry property (13). To define $\psi$ we proceed as follows: let $y$ be any point in the plane and choose any $x$ such that $G(x) = y$. Then define $\psi(y) = \widetilde{G}(\varphi(y))$. This definition does not depend on the choice of $x$ by the symmetry of $\varphi$. It is smooth since it is locally a composition of diffeomorphisms at any point $y \notin G(\ell_1)$, and every point in $G(\ell)$ has a preimage that is not critical. Clearly $\widetilde{G} \circ \varphi = \psi \circ G$. This finishes the construction of the equivalence between $G$ and $\widetilde{G}$.

Now it remains to prove that the properties (1) to (4) are satisfied for a perturbation $\widetilde{G}$ of $G$. A strong $C^\infty$ neighborhood of $G$ is given by a function $\epsilon : \mathbb{R}^2 \to \mathbb{R}^+$ and defined as the class of maps $\widetilde{G}$ of the plane such that every derivative of $\widetilde{G}$ at a point $z$ is at a distance less than $\epsilon(z)$ from the corresponding derivative of $G$ at $z$. If $\widetilde{G}$ is a $C^0$ strong perturbation of $G$, then $\widetilde{G}$ has an attractor at $\infty$, from which it follows that it is a proper map and has degree two. Moreover, given a neighborhood $U$ of the critical set of $G$, there exists a $C^1$ strong neighborhood of $G$ such that the set of critical points of any $\widetilde{G}$ in that neighborhood is contained in $U$. This is also easy to prove since critical points are determined by a $C^1$ condition: the Jacobian equal to zero. Furthermore, if the perturbation is of class $C^3$, and the initial map $G$ is generic, then the critical set of the perturbation is $C^1$ close to that of $G$, and the type of the critical points is preserved. That is, properties (1) and (2) are immediate application of the generic conditions imposed on the maps $G$ under consideration. Then property (3) follows from the fact that the map is two-to-one in the un bounded component of the complement of the preimage of the critical set, and finally this implies property (4).

The proof for $G \in \Theta_+$ is similar and will be omitted. □

*Proof of part (iii) of Theorem 1.* Until now it was proved that generic quadratic maps belong to one of two classes of geometric equivalence. It follows that no other class may contain an open set. □

## 5. A one-parameter family

In this section we analyze some global aspects of the dynamics of the one-parameter family $f_\mu(z) = z^2 - \mu\bar{z}$, where $\mu \in \mathbb{R}$, $z \in \mathbb{C}$ and $\bar{z}$ denotes the conjugated of the complex number $z$. Several properties about the dynamics of this family are exposed in [5] and [6]. If $I : \mathbb{C} \to \mathbb{R}^2$ is given by $I(x + iy) = (y, -x)$ and $G_\mu = I \circ f_\mu \circ I^{-1}$, then

$$G_\mu(x, y) = (-2xy + \mu x, x^2 - y^2 - \mu y). \tag{14}$$

Observe that for every $\mu$ the map $G_\mu$ belongs to $\Theta_-$ and verifies the symmetries: $G_\mu \circ J = J \circ G_\mu$ and $R_\alpha \circ G_\mu \circ R_\alpha = G_\mu$, where $J(x,y) = (-x,y)$ and $R_\alpha$ is the rotation of angle $\alpha = \frac{2\pi}{3}$.

The map $G_2$ has very interesting features: it is a two-dimensional analogue of the map $x \to -x^2 - 2x$ on the interval $[-3,1]$. The next theorem, which was proved in [5], emphasizes the importance of this map.

**Theorem 2.** *The following properties are satisfied by the mapping $G_2$:*

*(i) The basin of attraction of $\infty$ is the unbounded component of the complement of $G_2(\ell)$, where $\ell$ is the critical set of $G_2$.*

*(ii) The restriction of $G_2$ to the complement of this basin is conjugated to a Baker-like map.*

Let $T$ be an equilateral triangle; joining the middle points of the sides of $T$ one obtains an equilateral triangle $T'$. The Baker-like map mentioned in the statement above is obtained as follows (see [5]): first, carry $T$ into $T'$ by means of four affine maps with singularities at the sides of $T'$, then apply a symmetry with respect to one of the sides of $T'$ and finally multiply by two, to fit again on $T$. The map obtained has a fixed vertex while the other two are two-periodic. Observe that the map is expanding except for the singularities, each point in the interior of $T$ has four preimages, while the restriction to the boundary is two-to-one. Moreover, this map preserves the Lebesgue measure.

Denote by $B_\mu$ the basin of attraction of $\infty$ for the map $G_\mu$ and by $\partial B_\mu$ its boundary. The results stated in the theorem above imply that the deltoid $\widetilde{\ell}$, closure of $G_2^{-1}(G_2(\ell)) \setminus \ell$, is equal to $G_2(\ell)$ and also equal to $\partial B_2$, while the restriction of $G_2$ to the complement of $B_2$ preserves a smooth measure. Note that the restriction of $G_2$ to $\partial B_2$ is conjugated to the circle map $z \to z^2$. The map $G_2$ is highly unstable, the bifurcations of the dynamics around $G_2$ depends on the relative positions of the critical points and the basin of attraction of $\infty$. The study of the boundary of $B_2$ is determinant in the global behavior of the perturbations of $G_2$.

We will just consider perturbations of $G_2$ within the family $G_\mu$. The main goal in this section is to prove the next theorem:

**Theorem 3.** *The family $G_\mu$ has the following two properties:*

*(i) If $\mu < 2$, then $B_\mu$ is simply connected (if considered as $B_\mu \cup \{\infty\}$).*

*(ii) The complement of $B_\mu$ is a Cantor set for every $\mu$ large.*

Since $G_\mu$ belongs to $\Theta_-$, the following parameter values are obtained from (2): $p = -2$, $a = \mu$, $b = t = k_1 = k_2 = 0$. So, $\alpha = 0$, $\rho = \mu$ and $A = -Id$; see (6), (7) and (9). Keeping the meaning of $\ell$ and $\widetilde{\ell}$ for the mapping $G_\mu$, it follows that $G_\mu(\ell)$ and $\widetilde{\ell}$ are parametrized, respectively, by $\omega \mapsto -\frac{\mu^2}{4}\Delta_0(\omega)$ and $\omega \mapsto -\frac{\mu}{2}\Delta_0(\omega)$, with $\omega \in [0, 2\pi)$. From these facts it is easy to conclude that:

- If $\mu = 2$, then $\widetilde{\ell} = G_\mu(\ell)$.
- If $\mu < 2$, then $G_\mu(\ell) \subset \operatorname{int} \widetilde{\ell}$.
- If $\mu > 2$, then $G_\mu(\ell) \subset \operatorname{ext} \widetilde{\ell}$.

Here $\operatorname{int}\gamma$ and $\operatorname{ext}\gamma$ denote, respectively, the bounded and unbounded regions provided by the complement of the plane simple closed curve $\gamma$.

**Lemma 4.** *If $L$ is a simple closed curve such that $G_\mu(\ell) \subset \operatorname{int} L$, then $G_\mu^{-1}(L)$ is also a simple closed curve and $G_\mu : G_\mu^{-1}(L) \to L$ is two-to-one. Moreover, if $L_1 = G_\mu^{-1}(L) \subset$*

$\overline{\operatorname{ext} L}$, *then* $L_2 = G_\mu^{-1}(L_1) \subset \overline{\operatorname{ext} L_1}$ *and* $G_\mu(L) \subset \overline{\operatorname{int} L}$. *On the other hand, if* $L$ *is a simple closed curve and* $L \subset \operatorname{int} G_\mu(\ell)$, *then* $L_1$ *is the disjoint union of four simple closed curves.*

*Proof.* The first assertion is an immediate consequence of part (iv) in Theorem 1. Next assume by contradiction that $L_1 \subset \overline{\operatorname{ext} L}$ and suppose that there exists a point $x \in \operatorname{ext} L_2 \cap \operatorname{int} L_1$. Observe that if $x \in \operatorname{ext} L_2$, then $G_\mu(x) \in \operatorname{ext} L_1$. But if $x \in \operatorname{int} L_1$, then $G_\mu(x) \in \operatorname{int} G_\mu(L_1) = \operatorname{int} L \subset \operatorname{int} L_1$, which is absurd; thus $L_2 \subset \overline{\operatorname{ext} L_1}$. To prove that $G_\mu(L) \subset \overline{\operatorname{int} L}$, note that $G_\mu^{-1}(L) \subset \overline{\operatorname{ext} L}$ implies that $L \subset G_\mu(\overline{\operatorname{ext} L}) = \overline{\operatorname{ext} G_\mu(L)}$, therefore $G_\mu(L) \subset \overline{\operatorname{int} L}$. The last assertion is also direct consequence of Theorem 1. □

*Proof of part (i) of Theorem 3.* As exhibited above, the deltoids $G(\ell)$ and $\tilde{\ell}$ coincide for $\mu = 2$. The vertices of this deltoid are the fixed point $r_2 = (0, -3)$ and a two-periodic orbit $\{p_2, q_2\}$. Denote by $r_\mu$, $p_\mu$, $q_\mu$ the analytic continuation of these points. The circle $C$ through these three points is centered at the origin and has radius $1 + \mu$. Since $\mu < 2$ we have

$$G_\mu(\ell) \subset \operatorname{int} \widetilde{\ell} \subset \operatorname{int} C. \tag{15}$$

Consider the function $\chi : \mathbb{C} \to \mathbb{R}$ defined by $\chi(z) = |z|$ for all $z \in \mathbb{C}$. Observe that for every $z \in \operatorname{ext} C$, $\chi(f_\mu(z)) - \chi(z) \geq |z|^2 - (\mu + 1)|z| > 0$. This says that $\chi$ is a Lyapunov function for the restriction of $f_\mu$ to $\operatorname{ext} C$, therefore this set is contained in the basin of $\infty$ for this map.

We claim that $\{G_\mu^{-n}(\operatorname{ext} C)\}_{n \geq 0}$ is an increasing sequence of simply connected sets. Note that the map $I$, defined at the beginnig of this section, leaves $C$ invariant; hence $\operatorname{ext} C$ is also invariant under $G_\mu$ and contained in $B_\mu$. It follows from Lemma 4 that $G_\mu^{-1}(C) \subset \operatorname{int} C$. On the other hand, equation (15) and the same lemma imply that $G_\mu^{-1}(C)$ is a simple closed curve. Joining these two facts we have that

$$G_\mu^{-1}(\operatorname{ext} C) = \operatorname{ext} G_\mu^{-1}(C) \supset \operatorname{ext} C.$$

As $G_\mu(\ell) \subset \operatorname{int} C$, Lemma 4 also implies that $\widetilde{\ell} \subset \operatorname{int} G_\mu^{-1}(C)$. It follows that $G_\mu(\ell)$ is also contained in $\operatorname{int} G_\mu^{-1}(C)$. Hence, the preceding argument implies that $G_\mu^{-2}(C)$ is a simple closed curve, which obviously is contained in $\operatorname{int} G_\mu^{-1}(C)$. Thus by a recursive discourse, the claim follows by induction. Finally, since the basin of $\infty$ satisfies $B_\mu = \bigcup_{n \geq 0} G_\mu^{-n}(C)$, the proof of this part of the theorem is complete. □

*Proof of part (ii) of Theorem 3.* A simple calculation shows that $G_\mu(\ell) \subset \operatorname{int} C$ for all $\mu > 2(1 + \sqrt{2})$; so, every critical value, and hence every critical point, belongs to $B_\mu$. In this case the complement $B_\mu^c$ of $B_\mu$ satisfies $B_\mu^c = \bigcap_{n \geq 0} G_\mu^{-n}(\operatorname{int} \widetilde{\ell})$. Then, by standard arguments one can prove that $B_\mu^c$ has uncountably many components, but to show that it is a Cantor set we need to make $\mu$ larger. Indeed, for $\mu$ sufficiently large, it will be showed that the distance between the critical set $\ell$ and the preimage of $\widetilde{\ell}$ is large and the differential at these points expands any vector at a constant rate. Note that $G_\mu^{-1}(\widetilde{\ell})$ has four connected components, denoted by $K_\mu^i$ $(i = 0, 1, 2, 3)$. One of these components, say $K_\mu^0$, is contained in $\operatorname{int} \ell$. By calculating the vertices of $K_\mu^0$ one can see that for every $R > 4$ there exists $\mu(R)$ such that $K_\mu^0$ is contained in the disc of center 0 and radius $R$, for every $\mu > \mu(R)$. Using the symmetries of the mapping, the same property holds for every $K_\mu^i$. Fix any $R > 4$. Since $DG_\mu(x, y) = \begin{pmatrix} -2y + \mu & -2x \\ 2x & -2y - \mu \end{pmatrix}$, it follows that for all $(x, y)$ in $K_\mu^0$, $DG_\mu(x, y)$ expands uniformly any nonzero vector for

every $\mu > \mu(R)$. On the other hand, let $K^1_\mu$ be the component contained in the exterior of $\ell$ and intersecting the vertical axis. For every $(x, y) \in K^1_\mu$, it holds that $|x| < R$ and $-\frac{\mu}{2} - y \sim \mu$. Then the expansion of $DG_\mu$ at points in $K^1_\mu$ can be equally obtained. Since the regions $K^2_\mu$ and $K^3_\mu$ can be obtained from $K^1_\mu$ by special rotations, the result follows using the symmetries of the map $G_\mu$. □

## 6. Conclusions and questions.

In this final section we discuss some problems related to the topics of this article.

6.1. **The geometry of critical sets.** In the wide world of planar quadratic maps there are finitely many classes of geometrically equivalent maps. In the generic set $Q_g$ such a classification was possible mainly by two reasons: first, the mechanism created to understand the preimages of the deltoids or hypdeltoids; second, because these maps are injective when restricted to its critical sets. However, for nongeneric quadratic maps these restrictions are not necessarily injective, but one can directly check the assertion in each one of the parts of the decomposition established in sections 7 to 9 in [1], where the nongeneric quadratic maps of plane were classified.

It is easy to see that even within the class of generic maps of the plane having just one component of critical points, there exist infinitely many nonequivalent maps. The next example illustrates this claim.

**Example 1.** Consider the one-parameter family of plane endomorphisms:

$$F_\mu(x, y) = (x^3 - 3xy, y + f_\mu(x)).$$

The critical set of $F_\mu$ is the curve $y = x^2 + xf'_\mu(x)$. These maps have nondegenerate critical points, a cusp point occurs at every $(x, y)$ in the critical set such that $2x+2f'_\mu(x)+ xf''_\mu(x) = 0$. For example, by choosing the function $f_\mu$ so that $xf''_\mu(x)+2f'_\mu(x) = \mu \sin x$, there exist values of $\mu$ for which the number of cusps is arbitrarily large. Then there exist infinitely many different classes of geometric equivalence within that family.

The problem of classifying under geometric equivalence degree three polynomial endomorphisms is not possible with similar techniques. At this point, we like to pose the following question: *Are there finitely many equivalence classes of generic polynomials of a given degree?*

6.2. **Dynamics of plane maps.** As was said in the introduction, it is known that generic quadratic maps of the plane having no fixed points, must have empty limit sets. The question arising is if the bifurcation giving rise to the appearance of a first fixed point occurs in the boundary of the basin of attraction of $\infty$. That is, if $f_\mu$ is a one-parameter family of generic maps, and $f_0$ is the first map having fixed points, then we ask wether there exists an interval $[0, \mu_0]$ such that $f_\mu$ has a fixed point in the boundary of the basin of $\infty$. This problem seems to be very difficult, and a positive answer would give a new element for understanding globally the dynamics of these maps. More generally, we state the following open question: *Does a generic quadratic map of the plane (having fixed points) necessarily have a fixed point in the boundary of the basin of $\infty$?*

6.3. **Dynamics of the family $G_\mu(z) = z^2 - \mu\bar{z}$ and its perturbations.** Some questions that would be interesting to answer concern also with the boundary of the basin of $\infty$. It is natural to ask if the fact that every critical point is contained in the basin of $\infty$ implies that the nonwandering set is a hyperbolic set, however this seems to be very difficult to prove. Being less ambitious, one can ask if at least for the one parameter family under consideration, it holds that for parameters $\mu$ little larger than two, the complement of the basin is an expanding Cantor set.

There is another interesting question concerning the basin of $\infty$. It is clear that the complement of the basin is a forward invariant set. It was sometimes conjectured (for this family and also for others families of endomorphisms of the plane appearing in diverse models) that, as some numerical experiments have shown, there is a unique attractor in the complement of the basin of $\infty$. In other words, it is asked if the plane is subdivided into three sets: the basin of $\infty$, the basin of another attractor and the boundary of both sets; see [6] and references therein.

We want to state another problem. Observe that the restriction of $G_2$ to the boundary of the basin of $\infty$ is a degree two map isotopic to the map $z \to z^2$ in the unit circle. Moreover, this map is a local homeomorphism, but has three degenerate critical points, whose images are periodic repellers. It is an interesting problem to solve if this invariant curve has some kind of persistence. When $\mu > 2$ the set of critical points is contained in the basin of $\infty$, and it is impossible for the curve to persist. But consider the case where $\mu < 2$. In this case the situation is different because the set of critical points does not intersect the closure of the basin of $\infty$. We finish this section with two questions: *Is it true when $\mu < 2$ that an invariant curve persists in the boundary of the basin of $\infty$?* If the answer to this question is yes, *what can be said about the dynamics of the restriction of $G_\mu$ to the invariant curve?*

*Acknowledgments.* First and third author are grateful for the hospitality and support received at the Universidad de La República, where part of this paper was written. The authors are grateful to the referee for careful reading and helpful comments.

## References

[1] F. Bofill, J.L. Garrido, F. Vilamajó, N. Romero and A. Rovella. *On the quadratic Endomorphisms of the Plane.* Advanced Nonlinear Studies. **4** (2004), 37–55.
[2] L.E.J. Brouwer. *Beweis des ebenen translationssatzes. Math. Ann.* **72** (1912), 37–54.
[3] M. do Carmo. *Differential Geometry of Curves and Surfaces* Prentice-Hall (1976).
[4] J. Franks. *A new proof of the Brouwer plane translation theorem.* Ergodic Theory Dynam. Systems **12** (1992), 217–226.
[5] J. King-Dávalos, H. Méndez-Lango and G. Sienra-Loera. *Some Dynamical Properties of $F(z) = z^2 - 2\overline{z}$.* Qual. Th. Dyn. Sys. 5, Art. **77**, (2004) 101-120.
[6] J. King-Dávalos. *Algunos aspectos dinámicos y bifurcaciones de la familia $F_a(z) = z^2 + 2a\overline{z}$.* Ph.D. Thesis. UNAM, México, (2006).
[7] I. Malta, N. Saldanha and C. Tomei. *The numerical inversion of functions from the plane to the plane.* Math. Comp. **65**, (1996), 1531-1552.
[8] A. Ostermann and G. Wanner. *Geometry by Its History.* Springer (2012).
[9] H. Whitney. *On singularities of mappings of Euclidean spaces, I. Mappings of the plane into the plane.* Ann. of Math. **62**, (1955), 374-410.

Universidade Federal Fluminense. Instituto de Matemática. Rua Mario Santos Braga S/N. Niterói, 24.020-140, Rio de Janeiro, Brasil.
*E-mail address*: jdelgado@mat.uff.br

Universitat Politècnica de Catalunya. Departament de Matemàtica Aplicada 2. Escola Tècnica Superior d'Enginyeria Industrial. Colom 11, 08222. Terrasa, Barcelona, Spain.
*E-mail address*: jose.luis.garrido@upc.edu
*E-mail address*: francesc.vilamajo@upc.edu

Universidad Centroccidental Lisandro Alvarado. Departamento de Matemática. Decanato de Ciencias y Tecnología. Apartado Postal 400. Barquisimeto, Venezuela.
*E-mail address*: nromero@ucla.edu.ve

Universidad de La República. Facultad de Ciencias. Centro de Matemática. Iguá 4225. C.P. 11400. Montevideo, Uruguay.
*E-mail address*: leva@cmat.edu.uy

# LINEAR COCYCLES OVER LORENZ-LIKE FLOWS

MOHAMMAD FANAEE

Abstract. We prove that the Lyapunov exponents of typical fiber bunched linear cocycles over Lorenz-like flows have multiplicity one: the set of exceptional cocycles has infinite codimention, i.e. it is locally contained in finite unions of closed submanifolds with arbitrarily high codimension.

Contents



## 1. Introduction

A linear cocycle over a flow $f^t : \Lambda \to \Lambda$ is a flow $F^t : \Lambda \times \mathbb{C}^d \to \Lambda \times \mathbb{C}^d$ of the form

$$F^t(x, v) = (f^t(x), A^t(x)v)$$

where each $A^t(x) : \mathbb{C}^d \to \mathbb{C}^d$ is a linear isomorphism. The Lyapunov exponents are the exponential rates

$$\lambda(x, v) = \lim_{|t|\to\infty} \frac{1}{t} \log ||A^t(x)v||, \ v \neq 0.$$

By Oseledets [14] this limit exists for every $v \in \mathbb{C}^d$ on a full measure set of $x \in \Lambda$, relative to any invariant measure $m$. There are at most $d$ Lyapunov exponents; they are constant on orbits and vary measurably with the base point. Thus Lyapunov exponents are constant if $m$ is ergodic.

Our main interest is to characterize when all exponents have multiplicity one i.e. the subspace of vectors $v \in \mathbb{C}^d$ that share the same value of $\lambda(x, v)$ has dimension one.

There has been much recent progress on this problem, specially when the base dynamics is hyperbolic, see [9,5,6,11]. Here, we extend the theory to the case when the base dynamics is a Lorenz-like attractor.

A Lorenz-like flow in 3-dimensions admits a cross section $S$ and a Poincaré transformation $P : S\backslash\Gamma \to S$ defined outside a curve $\Gamma$ which is contained in the intersection of $S$ with the stable manifold of some hyperbolic equilibrium. Trajectories through $\Gamma$ just converge to the equilibrium and the other trajectories through $S$ eventually return to

$S$. Their accumulation set is the so-called geometric Lorenz attractor. Moreover, there is an invariant splitting

$$T_\Lambda M = E^s \oplus E^{cu}$$

of the tangent bundle where the uniformly contracting bundle $E^s$ has dimension 1, and the volume-expanding bundle $E^{cu}$ which contains the flow direction has dimension 2. Another important feature is that the Poincaré transformation of this flow admits an invariant contracting foliation $\mathcal{F}$ through which the dynamics can be reduced to that of a map on the interval (leaf space of $\mathcal{F}$). A Lorenz-like flow admites an invariant physical probability measure which is ergodic.

1.1. **Cocycles over maps.** A linear cocycle over an invertible transformation $f : N \to N$ is a transformation $F : N \times \mathbb{C}^d \to N \times \mathbb{C}^d$ satisfying $f \circ \pi = \pi \circ F$ which acts by linear isomorphisms $A(x)$ on fibers. So, the cocycle has the form

$$F(x, v) = (f(x), A(x)v)$$

where

$$A : N \to \mathrm{GL}(d, \mathbb{C}).$$

Conversely, any $A : N \to \mathrm{GL}(d, \mathbb{C})$ defines a linear cocycle over $f$. Note that $F^n(x, v) = (f^n(x), A^n(x)v)$, where

$$A^n(x) = A(f^{n-1}(x)) \ ... \ A(f(x))A(x),$$

$$A^{-n}(x) = (A^n(f^{-n}(x)))^{-1},$$

for any $n \geq 1$, and $A^0(x) = \mathrm{id}$.

Let $\mu$ be a probability measure invariant by $f$. Oseledets Theorem [14] states that there exist a Lyapunov splitting

$$E_1(x) \oplus ... \oplus E_k(x), \ 1 \leq k = k(x) \leq d,$$

and Lyapunov exponents $\lambda_1(x) > ... > \lambda_k(x)$,

$$\lambda_i(x) = \lim_{|n| \to \infty} \frac{1}{n} \log \parallel A^n(x)v_i \parallel, \ v_i \in E_i(x), \ 1 \leq i \leq k,$$

at $\mu$-almost every point. Lyapunov exponents are invariant, uniquely defined at almost every $x$ and vary measurably with the base point $x$. Thus, Lyapunov exponents are constant when $\mu$ is ergodic. Then $\{\lambda_1, ..., \lambda_k\}$ is called the Lyapunov spectrum of $A$.

We recall that, for any $r \in \mathbb{N} \cup \{0\}$ and $0 \leq \rho \leq 1$, the $C^{r,\rho}$ topology is defined by

$$||A||_{r,\rho} = \max_{0 \leq i \leq r} \sup_x ||D^i A(x)|| + \sup_{x \neq y} \frac{||D^r A(x) - D^r A(y)||}{\mathrm{d}(x, y)^\rho}$$

(for $\rho = 0$ omit the last term) and then

$$\mathcal{C}^{r,\rho}(N, d, \mathbb{C}) = \{A : N \to \mathrm{GL}(d, \mathbb{C}) : \ ||A||_{r,\rho} < +\infty\}$$

is a Banach space. We assume that $r + \rho > 0$ which implies $\eta-$Hölder continuity:

$$\parallel A(x) - A(y) \parallel \leq \parallel A \parallel_{0,\eta} \mathrm{d}(x, y)^\eta,$$

with

$$\eta = \begin{cases} \rho & r = 0 \\ 1 & r \geq 1. \end{cases}$$

1.2. **Fiber bunching condition.** Suppose that $N = \mathbb{N}^{\mathbb{Z}}$, the full shift space with countably many symbols, and $f : N \to N$ the shift map

$$f((x_n)_{n\in\mathbb{Z}}) = (x_{n+1})_{n\in\mathbb{Z}}.$$

A cylinder of $N$ is any subset

$$[a_k, ...; a_0; ..., a_l] = \{x : \ x_j = \iota_j, \ j = k, ..., l\}$$

of $N$. We endowed $N$ with topology generated by cylinders. The local stable and local unstable sets of any $x \in N$ are defined as

$$W^s_{\rm loc}(x) = \{y : \ x_n = y_n, \ n \geq 0\}$$

and

$$W^u_{\rm loc}(x) = \{y : \ x_n = y_n, \ n < 0\}.$$

Assume that $N$ is endowed with a metric d for which
(i) $\mathrm{d}(f(y), f(z)) \leq \theta(x)\mathrm{d}(y, z)$, for all $y, z \in W^s_{\rm loc}(x)$,
(ii) $\mathrm{d}(f^{-1}(y), f^{-1}(z)) \leq \theta(x)\mathrm{d}(y, z)$, for all $y, z \in W^u_{\rm loc}(x)$,
where $0 < \theta(x) \leq \theta < 1$, for all $x \in N$.

Let $A$ be an $\eta$-Hölder continuous linear cocycle over $f$.

**Definition 1.1.** *$A$ is fiber bunched if there exists some constant $\tau \in (0, 1)$ such that*

$$||A(x)|| \ ||A(x)^{-1}||\theta(x)^{\eta} < \tau,$$

*for any $x \in N$.*

**Remark 1.1.** *Fiber bunching is an open condition in $C^{r,\rho}(N, d, \mathbb{C})$: if $A$ is a fiber bunched linear cocycle then any linear cocycle $B$ sufficiently $C^0$ close to $A$ is also fiber bunched, by definition.*

1.3. **Product structure.** Let $N_u = \mathbb{N}^{\{n\geq 0\}}$ and $N_s = \mathbb{N}^{\{n<0\}}$. The map

$$x \mapsto (x_s, x_u)$$

is a homeomorphism form $N$ onto $N_s \times N_u$ where $x_s = \pi_s(x)$ and $x_u = \pi_u(x)$, for natural projections $\pi_s : N \to N_s$ and $\pi_u : N \to N_u$. We also consider the maps $f_s : N_s \to N_s$ and $f_u : N_u \to N_u$ defined by

$$f_u \circ \pi_u = \pi_u \circ f,$$
$$f_s \circ \pi_s = \pi_s \circ f^{-1}.$$

Assume that $\mu_f$ is an ergodic probability measure for $f$. Let $\mu_s = (\pi_s)_*\mu_f$ and $\mu_u = (\pi_u)_*\mu_f$ be the images of $\mu_f$ under the natural projections. It is easy to see that $\mu_s$ and $\mu_u$ are ergodic probabilities for $f_s$ and $f_u$, respectively. Notice that $\mu_s$ and $\mu_u$ are positive on cylinders, by definition.

We say that $\mu_f$ has product structure if there exists a measurable density function $\omega : N \to (0, +\infty)$ such that

$$\mu_f = \omega(x)(\mu_s \times \mu_u).$$

Assuming a probability measure which has product strucrure, Bonatti and Viana [9] obtained a general criterion for simplicity of Lyapunov spectrum for cocycles over hyperbolic systems and used it to prove that simplicity holds for generic linear cocycles that satisfy the fiber bunching condition. This criterion has improved by Avila and Viana [5] who used it to prove the Zorich-Kontsevich conjecture [6]. In [11], by geometric tools, we prove

**Theorem 1.1.** [11] *Lyapunov exponents of typical fiber bunched linear cocycles over complete shift map have multiplicity 1.*

1.4. **Suspension flows.** Consider a suspension flow $f^t : \Lambda \to \Lambda$ of $f : N \to N$ and let $T : N \to \mathbb{R}$ be the corresponding return time to $N$. Assume that $A^t : \Lambda \to \mathrm{GL}(d, \mathbb{C})$ is a linear cocycle over $f^t$, and define

$$A_f(x) = A^{T(x)}(x),$$

for any $x \in N$. Note that $A_f : N \to \mathrm{GL}(d, \mathbb{C})$ is a linear cocycle over $f$.

Then we define a relative topology as

$$||A^t||_{r,\rho} = ||A_f||_{r,\rho}$$

for any $r \in \mathbb{N} \cup \{0\}$ and $0 \leq \rho \leq 1$ with $r + \rho > 0$, and let

$$\mathcal{C}^{r,\rho}(\Lambda, d, \mathbb{C}) = \{A^t : \Lambda \to \mathrm{GL}(d, \mathbb{C}) : \ ||A^t||_{r,\rho} < +\infty\}.$$

**Definition 1.2.** *$A^t$ is fiber bunched if the corresponding linear cocycle $A_f$ is a fiber bunched linear cocycle over $f$.*

**Remark 1.2.** *Note that fiber bunching is an open condition in $\mathcal{C}^{r,\rho}(\Lambda, d, \mathbb{C})$, by definition.*

Our main result is

**Main Theorem.** Typical fiber bunched linear cocycles over geometric Lorenz attractors have simple spectrum.

## 2. Lorenz-like flows

In this section, we recall the basic notions and strategies to construct a geometric Lorenz attractor and the unique physical probability measure and then, we study existence of a Markov structure on these flows.

The geometric Lorenz attractors were introduced in [18,12] as a precise model for the dynamical behavior of the equations

$$\begin{aligned} \dot{x} &= a(y - x), \\ \dot{y} &= bx - y - xz, \\ \dot{z} &= xy - cx, \end{aligned} \tag{1}$$

proposed by Lorenz [13], loosely related to fluid convection and weather prediction. Tucker [16] showed that the Lorenz equations exhibits a geometric Lorenz attractor, for classical parameters $a = 10,\ b = 28,\ c = 8/3$.

This system of equations is symmetric with respect to the $z-$axis. The singularity $\mathbf{0}$ has real eigenvalues $\alpha_{ss} < \alpha_s < 0 < -\alpha_{ss} < \alpha_u$ with $\alpha_s + \alpha_u > 0$. There are also two symmetric saddles $\sigma_1, \sigma_2$ with a real negative and two conjugate complex eigenvalues where the complex eigenvalues have positive real parts. The character of this flow is strongly dissipative, in particular, any maximally positively invariant subset has zero volume.

2.1. **The geometric model.** To construct a geometric Lorenz attractor, we should analyze the dynamics of Lorenz flow in a neighborhood of $\mathbf{0}$ imitating the effect of the pair of saddles.

2.1.1. *Poincaré transformation.* By construction, there is a cross section $S$ intersecting the stable manifold of 0 along a curve $\Gamma$ that separates $S$ into 2 connected components. We denote the corresponding Poincaré transformation

$$P : S\backslash\Gamma \to S.$$

Note that the future trajectories of points in $\Gamma$ do not came back to $S$.

We consider the smooth foliation $\mathcal{F}$ of $S$ into curves having $\Gamma$ as a leaf which are invariant and uniformly contracted by forward iterates of $P$. Indeed, every leaf $\mathcal{F}_{(x,y)}$ is mapped by $P$ completely inside the leaf $\mathcal{F}_{P(x,y)}$, and $P|_{\mathcal{F}_{(x,y)}}$ is a uniform contraction. Indeed, $P$ must have the form

$$P(x, y) = (g(x), h(x, y))$$

which by effect of saddles and singularity, we can assume that $h$ is a contraction along its second coordinate. The map $g$ is uniformly expanding with derivative tending to infinity as one approaches to $\Gamma$. We assume that $|g'| \geq \theta^{-1} > \sqrt{2}$ and since the rate of contraction of $h$ on the second coordinate should be much higher than the expansion of $g$, we can take $|\partial_y h| \leq \theta < 1$.

2.1.2. *Lorenz map.* Let $\pi$ be the canonical projection of section $S$ into $\mathcal{F}$, i.e. $\pi$ assigns to each point of $S$ the leaf that contained it. By invariance of $\mathcal{F}$, one dimensional Lorenz map

$$g : (\mathcal{F}\backslash\Gamma) \to \mathcal{F}$$

is uniquely defined so that

$$\begin{array}{ccc} S\backslash\Gamma & \xrightarrow{P} & S \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\pi} \\ \mathcal{F}\backslash\Gamma & \xrightarrow[g]{} & \mathcal{F} \end{array}$$

commutes, i.e. $g \circ \pi = \pi \circ P$ on $S\backslash\Gamma$.

One may identify quotient space $S/\mathcal{F}$ with a compact interval as $I = [-1, 1]$, and so

$$g : [-1, 1]\backslash\{0\} \to [-1, 1]$$

is smooth on $I\backslash\{0\}$ with a discontinuity and infinite left and right derivatives at 0. Note that the symmetry of the Lorenz equations implies $g(-x) = -g(x)$.

2.2. **The attractor.** The geometric Lorenz attractor $\Lambda$ is characterized as follows. Note that the restriction of $g$ to both $\{x < 0\}$ and $\{x > 0\}$ admits continuous extensions to the point 0. Hence, $g$ may be considered as an extension to a 2-valued map at 0 and continuous on both $\{x \leq 0\}$ and $\{x \geq 0\}$. Correspondingly, the restriction of the Poincaré transformation to each of the connected components of $S\backslash\Gamma$ admits a continuous extension to the closure, each one collapsing the curve $\Gamma$ to a single point. Thus, $P$ may also be considered as a 2-valued transformation defined on the whole cross section and continuous on the closure of each of the connected components. Let

$$\Lambda_P = \bigcap_{n\geq 0} P^n(S) \subset S.$$

We define $\Lambda$ to be the saturation of $\Lambda_P$ by the Lorenz flow, that is, the orbits of its points. Therefore, orbits in $\Lambda$ intersect the cross section infinitely often, both forward and backward.

Dynamical properties of $\Lambda$ may be deduced from corresponding properties for the quotient map $h$. More important, a quotient map with similar properties exists for all nearby vector fields, and so such properties are robust for these flows.

2.3. **Physical probability measure.** The existence of a unique absolutely continuous invariant probability $\mu_g$ which is ergodic and $0 < \frac{d\mu_g}{d(Leb)} < +\infty$ for Lorenz one-dimensional map $g$ is well-known (see [16] for more details).

One may construct an invariant probability measure $\mu_P$ on $\Lambda_P$, as the lifting of $\mu_g$. Indeed, we may think of $\mu_g$ as a probability measure on Borel subsets of $\mathcal{F}$. Since $P$ is uniformly contracting on leaves of $\mathcal{F}$, one concludes that the sequence

$$(P_*^n \mu_g)_{n\geq 1},$$

of push-forwards is weak$^*$-Cauchy: given any continuous $\varphi : S \to \mathbb{R}$,

$$\int \varphi \; d(P_*^n \mu_g) = \int (\varphi \circ P^n) \; d\mu_g, \; n \geq 1,$$

is a Cauchy sequence in $\mathbb{R}$. Define the probability measure $\mu_P$ as the weak$^*$-limit of this sequence that is

$$\int \varphi \; d\mu_P = \lim_{n\to+\infty} \int \varphi \; d(P_*^n \mu_g),$$

for each continuous function $\varphi$. Thus $\mu_P$ is invariant under $P$, and it is a physical probability measure on Borel subsets of $\Lambda_P$ which is ergodic.

Later, as the Poincaré transformation may be extended to the Lorenz flow through a suspension construction, the invariant probability $\mu_P$ corresponds to an ergodic physical probability measure $m$ on $\Lambda$: Denote by $R : S\backslash\Gamma \to (0, +\infty)$ the *first return time* to $S$ defined by

$$P(x) = f^{R(x)}(x).$$

The first return time $R$ is Lebesgue integrable, since $P(x) \approx |\log(\mathrm{d}(x, \Gamma))|$, for $x$ close to $\Gamma$. This follows that

$$\int R \; d\mu_P < +\infty.$$

Let $\sim$ be an equivalence relation on $S \times \mathbb{R}$ defined as $(x, R(x)) \sim (P(x), 0)$. Set $\tilde{S} = (S \times \mathbb{R})/\sim$ and define the finite measure

$$\tilde{\mu} = \pi_*(\mu_P \times dt)$$

where $\pi : S \times \mathbb{R} \to \tilde{S}$ is the quotient map and $dt$ is Lebesgue measure in $\mathbb{R}$. Define $\phi : \tilde{S} \to M$ as $\phi(x, t) = f^t(x)$, and let

$$m = \phi_* \tilde{\mu}.$$

One may check also that

$$\frac{1}{T} \int_0^T \varphi(f^t(x)) \; dt \; \to \; \int \varphi \; dm$$

as $T \to +\infty$, for any continuous function $\varphi : M \to \mathbb{R}$, and Lebesgue almost every $x \in \phi(\tilde{S})$.

## 3. A symbolic structure

Consider a Lorenz one dimensional map $g : I\backslash\{0\} \to I$.

**Theorem 3.1.** [10] *There exists a return map $\hat{g}$, an interval $\hat{I} = (-\delta, \delta)$, $0 < \delta < 1$, and a partition $\{\hat{I}(l) : \ l \in \mathbb{N}\}$ to subintervals of $\hat{I}$, Lebesgue mod 0, for which $\hat{g}$ maps any $\hat{I}(l)$ diffeomorphically onto $\hat{I}$, and the return time $\hat{r}$ is Lebesgue integrable. Moreover, there exists a constant $0 < c < 1$ such that, for all $x, y$ in any $\hat{I}(l)$,*

$$\log \frac{|\hat{g}'(x)|}{|\hat{g}'(y)|} \leq c^{n(x,y)}$$

*where $n(x, y) = \min\{n : \ \hat{g}^n(x) \in \hat{I}(l_i), \ \hat{g}^n(y) \in \hat{I}(l_j), \ i \neq j\}$.*

**Remark 3.1.** *Note that, as Lorenz map $g$ is uniformly expanding, the intersection of $(\hat{g}^{-n}(J(l_n)))$ over all $n \geq 0$ consists of exactly one point.*

Therefore, $\hat{g}$ may be seen as the shift map on $\hat{N} = \mathbb{N}^{\{n\geq 0\}}$: there exists a conjugation between the shift map $\hat{f} : \hat{N} \to \hat{N}$ and $\hat{g}$ presented by the next commuting diagram

$$\begin{array}{ccc} \hat{N} & \xrightarrow{\hat{f}} & \hat{N} \\ {\scriptstyle\hat{\phi}}\downarrow & & \downarrow{\scriptstyle\hat{\phi}} \\ \hat{I} & \xrightarrow[\hat{g}]{} & \hat{I} \end{array}$$

where the bijection $\hat{\phi}$ may be defined as

$$\hat{\phi} : (l_n)_{n\geq 0} \mapsto \bigcap_{n\geq 0} \hat{g}^{-n}(\hat{I}(l_n)).$$

3.1. **Bi-dimensional Markov structure.** Now, we consider the bi-dimensional domain $\hat{S} = \pi^{-1}(\hat{I}) \subset S$ and corresponding to the Markov partition of $\hat{I}$ define a Markov partition $\{\hat{S}(l) = \pi^{-1}(\hat{I}(l)) : \ l \in \mathbb{N}\}$ of $\hat{S}$. The return time is defined as

$$r(x) = \hat{r}(\pi(x)).$$

Hence, there exists a return map $\hat{P}$ to $\hat{S}$ as

$$\hat{P}(x) = P^{r(x)}(x),$$

for any $x \in \hat{S}$. Moreover

$$\hat{g} \circ \pi = \pi \circ \hat{P}.$$

Let

$$\Lambda_{\hat{P}} = \bigcap_{n\geq 0} \hat{P}^n(\hat{S}).$$

So $\Lambda_{\hat{P}}$ is homeomorphically equal to $N$. Indeed, since $\bigcap_{n\in\mathbb{Z}} \hat{P}^{-n}(\hat{S}(l_n))$ consists of exactly one point, one may define a bijection $\phi : N \to \Lambda_{\hat{P}}$ as

$$\phi : (l_n)_{n\in\mathbb{Z}} \mapsto \bigcap_{n\in\mathbb{Z}} \hat{P}^{-n}(\hat{S}(l_n))$$

which implies the commuting diagram

$$\begin{array}{ccc} N & \xrightarrow{f} & N \\ {\scriptstyle\phi}\downarrow & & \downarrow{\scriptstyle\phi} \\ \Lambda_{\hat{P}} & \xrightarrow[\hat{P}]{} & \Lambda_{\hat{P}}. \end{array}$$

3.2. **Lifting the probability measure.** The normalized restriction $\hat{\mu}$ of $\mu_g$ to the domain of $\hat{g}$ is an absolutely continuous ergodic probability for $\hat{g}$ and then for $\hat{f}$, by conjugacy.

As the natural extension of $\hat{f}$ realized as the complete shift map $f$ on $N$, the *lift* $\mu$ of $\hat{\mu}$ is the unique $f$-invariant ergodic probability measure on $N$ such that

$$\hat{\pi}_*\mu = \hat{\mu}.$$

**Proposition 3.1.** *The lift probability $\mu$ has product structure. Moreover, the density function $\omega$ is continuous and, bounded from zero and infinity*

*Proof.* Note that by Theorem 3.1, for all $\hat{x}, \hat{y}$ in the same cylinder

$$\log \frac{J\hat{f}(\hat{x})}{J\hat{f}(\hat{y})} \leq c^{n(x,y)}.$$

The rest of proof is based on 4 main steps *Step 1.* If $\hat{x}, \hat{y} \in \hat{N}$ then for any $x \in W^s_{\rm loc}(\hat{x})$ and $y \in W^u_{\rm loc}(x) \cap W^s_{\rm loc}(\hat{y})$, the limit

$$J_{\hat{x},\hat{y}}(x) = \lim_{n\to\infty} \frac{J\hat{f}^n(\hat{x}^n)}{J\hat{f}^n(\hat{y}^n)},$$

where $\hat{x}^n = \hat{\pi}(f^{-n}(x)),\ \hat{y}^n = \hat{\pi}(f^{-n}(y))$, exists uniformly on $\hat{x}, \hat{y}, x$. Moreover,

$$(\hat{x}, \hat{y}, x) \mapsto J_{\hat{x},\hat{y}}(x)$$

is continuous and uniformly bounded from zero and infinity.

Indeed, we observe that

$$\log \frac{J\hat{f}^n(\hat{x}^n)}{J\hat{f}(\hat{y}^n)} \leq \sum_{i=1}^{n} \log \frac{J\hat{f}(\hat{x}^i)}{J\hat{f}(\hat{y}^i)}.$$

Since $\hat{x}^i$ and $\hat{y}^i$ are in the same cylinder, the series is uniformly bounded by $\sum_i c^{n(\hat{x}^i,\hat{y}^i)}$. But $n(\hat{x}^i, \hat{y}^i)$ is strictly increasing that implies uniform convergence of the series.
*Step 2.* If $\{\mu_{\hat{x}} :\ \hat{x} \in \hat{N}\}$ be an integration of $\mu$ then, for $\mu$-almost every $\hat{x} \in \hat{N}$,

$$\mu_{\hat{x}}(\xi_n) = \frac{1}{J\hat{f}^n(\hat{x}^n)},$$

for every cylinder $\xi_n = [x_{-n}, ..., x_{-1}],\ n \geq 1$, and any $x \in \xi_n \times \{\hat{x}\}$.
*Step 3.* Given any disintegration, by the last step, one may find a disintegration $\{\mu_{\hat{x}} :\ \hat{x} \in \hat{N}\}$ of $\mu$ so that

$$\mu_{\hat{y}} = J_{\hat{x},\hat{y}}\mu_{\hat{x}}.$$

*Step 4.* Fixing any $\hat{x}_0 \in \hat{N}$, one may define

$$\hat{\omega}(x_s, x_u) = J_{\hat{x}_0,x_u}(x_s, x_u),$$

for every $x = (x_s, x_u) \in N$. By Step 2, $\mu_{x_u} = \hat{\omega}(x_s, x_u)$, for any $x_u \in \hat{N}$.

The lift measure $\mu$ projects to $\hat{\mu} = \mu_u$, but the projection $\mu_s$ to $N_s$ is given by

$$\mu_s = \mu_{\hat{x}_0} \int_{\hat{N}} \hat{\omega}(x_s, x_u)\ d\hat{\mu}.$$

Therefore

$$\mu = \omega(x_s, x_u)\mu_s \times \mu_u$$

where

$$\omega(x_s, x_u) = \frac{1}{\int_{\hat{N}} \hat{\omega}(x_s, x_u)\ d\hat{\mu}}\ \hat{\omega}(x_s, x_u).$$

As conditional probabilities vary continuously with the base point so the density function $\omega$ is continuous. Also, $\omega$ is bounded from zero and infinity.

The i of Proposition 3.1 is now completed.

3.3. **Suspending the bi-lateral shift.** The saturation of $N$ by the Lorenz flow $f^t$, by ergodicity of $m$ has full measure in $\Lambda$. Now on, by $\Lambda$ we mean this full measure subset. Henceforth, a return time to $N$ is defined as

$$T : N \to \mathbb{R}$$
$$T(x) = \sum_{j=0}^{r(x)-1} R(P^j(x)),$$

for any $x \in N$.

**Proposition 3.2.** *The return time $T$ is integrable with respect to the probability measure $\mu$.*

*Proof.* For almost every $x$,

$$\int T(x)\ d(Leb) = \int r(x)[\frac{1}{r(x)} \sum_{j=0}^{r(x)-1} R(P^j(x))]\ d(Leb)$$

converges to

$$\int r(x)(\int R\ d(Leb))\ d(Leb) < +\infty$$

which implies

$$\int T\ d(Leb) < +\infty.$$

The proof is now completed by absolute continuity.

## 4. The proof of Main Theorem

Now, we are in the setting to complete the proof of Main Theorem.

For any linear cocycle $A^t$ over $\Lambda$ consider the corresponding linear cocycle $A_f$ on $N$ by

$$A_f(x) = A^{T(x)}(x),$$

for any $x \in N$.

**Proposition 4.1.** *Lyapunov spectrum of $A^t$ is simple if and only if Lyapunov spectrum of $A_f$ is simple.*

*Proof.* The Lyapunov exponents of $A_f$ are obtained by multiplying those of $A^t$ by the average return time

$$s_n(x) = \sum_{j=0}^{n-1} T(\hat{P}^j(x)),\ x \in N.$$

Indeed, given any non zero vector $v$,

$$\lim_{n\to+\infty} \frac{1}{n} \log ||A_f^n(x)v|| = \lim_{n\to+\infty} \frac{1}{n} \log ||A^{s_n(x)}(x)v||$$

which, for $\mu$-almost every $x$, this is equal to

$$\lim_{n\to+\infty} \frac{1}{n} s_n(x) \lim_{m\to+\infty} \frac{1}{m} \log ||A^m(x)v||.$$

But $\frac{1}{n}s_n(x)$ converges to $\int T\ d\mu < +\infty$

The proof of Proposition 4.1 is now completed.

Let $A^t$ be a linear cocycle over $\Lambda$. We define a neighborhood $\mathcal{V}$ of $A^t$ as the subset of all cocycles $B^t$ over $\Lambda$ for which $B_f \in \mathcal{U}$.

**Proposition 4.2.** *The application*

$$\mathcal{V} \ni B^t \mapsto B_f \in \mathcal{U}$$

*is a submersion.*

*Proof.* By definition,

$$\partial_{B^t} B_f(\dot{B}_t) = \dot{B}_f.$$

Let $\dot{B} \in \mathcal{C}^{r,\rho}(N, d, \mathbb{C})$. Then the suspension $\dot{B}^t$ of $\dot{B}$ is defined by

$$\dot{B}^t(X^s(x)) = (\mathrm{id}, t+s),\ 0 < t+s \le T(x),$$

identifying $(\mathrm{id}, T(x))$ with $(\dot{B}(x), 0)$, for any $x \in N$, setting $\dot{B}^0 = \mathrm{id}$. $\dot{B}^t$ is an $\eta$-Hölder linear cocycle over $\Lambda$ for which $\dot{B}_f(x) = (\dot{B}(x), 0)$. This shows that the derivative is surjective.

The proof of Proposition 4.2 is now completed.

The proof of Main Theorem is then completed, by Theorem 1.1.

***Acknowledgements.*** I would like to thanks M. Viana for all supports during my PhD studies at IMPA. This work is supported by a doctoral grant from CNPq-TWAS.

## References

[1] V. Araujo and M. Pacifico, Three dimensional flows, Springer-Verlag (2010).
[3] V. Araujo, M. Pacifico, E. Pujals and M. Viana, Singular-hyperbolic attractors, Transactions of the American Mathematical Society 361 (2009) 2431-2485.
[5] A. Avila and M. Viana, Simplicity of Lyapunov spectra: a sufficient condition, Potugaliae Matematica 64 (2007) 311-376.
[6] A. Avila and M. Viana, Simplicity of Lyapunov spectra: Proof of the Zorich-Kontsevich conjecture Acta Mathematica 198 (2007) 1-56.
[7] C. Bonatti, L. Diaz and M. Viana, Dynamics Beyond Uniform Hyperbolicity: A Global Geometric and Probabilistic Perspective, Encyclopedia of Mathematical Sciences 102 Springer-Verlag (2004).
[8] C. Bonatti, X. Gomez-Mont and M. Viana, Généreicité d'exposants de Lyapunov non-nuls pour des produits déterministes de matrices, Annales de l'Institut Henri Poincaré 20 (2003) 579-624.
[9] C. Bonatti and M. Viana, Lyapunov exponents with multiplicity 1 for deterministic products of matrices, Ergodic Theory and Dynamical Systems 24 (2004) 1295-1330.

[10] K. Díaz-Ordaz, Decay of correlation for non-Hölder observables for one-dimensional expanding Lorenz-like maps, Discrete and Continuous Dynamical Systems 15 (2006) 159-176.
[11] M. Fanaee. Generic simple cocycles over Markov maps. arXiv.org
[12] J. Guckenheimer and R. Williams, Structural stability of Lorenz attractors, Publications Mathématiques de l'IHÉS 50 (1979) 59-72.
[13] E. Lorenz, Deterministic non periodic flow, Journal of the Atmospheric Sciences, 20 (1963) 130-141.
[14] V. Oseledets, A multiplicative ergodic theorem, Transactions of the Moscow Mathematical Society 19 (1968) 197-231.
[15] W. Tuker, The Lorenz attractor exists, Comptes Rendus de l'AcadÃ©mie des Sciences Paris. Série I. Mathématique 328 (1999) 1197-1202.
[16] M. Viana, Stochastic dynamics of deterministic systems, Brazilian Mathematics Colloquium IMPA (1997).
[17] M. Viana, What's new on Lorenz strange attractors?, The Mathematical Intelligencer 22 (2000), 6-19
[18] R. Williams, The structure of the Lorenz attractor, Publications Mathématiques de l'IHÉS 50 (1979) 73-99.

*Mohammad Fanaee*
*Instituto de Matemática e Estatística (IME)*
*Universidade Federal Fluminense (UFF)*
*Campus Valonguinho*
*24020-140 Niterói-RJ-Brazil*
*Email: mf@id.uff.br*

# REGULARITY OF THE DRIFT AND ENTROPY OF RANDOM WALKS ON GROUPS

LORENZ GILCH AND FRANÇOIS LEDRAPPIER

Random walks on a group $G$ model many natural phenomena. A random walk is defined by a probability measure $p$ on $G$. We are interested in global asymptotic properties of the random walks and in particular in the linear drift and the asymptotic entropy. If the geometry of the group is rich, then these numbers are both positive and the way of dependence on $p$ is some global property of $G$. In this note, we review recent results about the regularity of the drift and the entropy in some examples.

## 1. Entropy and linear drift

We recall in this section the main notations for the objects under consideration associated to a group $G$ and a probability measure $p$ on $G$. Background on random walks can be found in the survey papers [**KV**] and [**V**] and in the book by W. Woess ([**W**]).

Let $G$ be a finitely generated group and $S$ a symmetric finite generator. For $g \in G$, let $|g|$ denote the smallest $n \in \mathbb{N}$ such that $g$ can be written as $g = s_1 \cdots s_n$, where $s_1, \dots, s_n \in S$. We denote by $d(g,h) := |g^{-1}h|$ the left invariant associated metric. Let $p$ be a probability measure on $G$ with support $B$. Unless otherwise specified, we always assume that $B$ is finite and that $\bigcup_{n\in\mathbb{N}} B^n = G$. We denote by $\mathcal{P}(B)$ the set of probability measures with support $B$. The set $\mathcal{P}(B)$ is naturally identified with an open subset of the probabilities on $B$, which is a contractible open polygonal bounded convex domain in $\mathbb{R}^{|B|-1}$. We form, with $p^{(0)}$ being the Dirac measure at the identity $e$,

$$p^{(n)}(g) \;=\; [p^{(n-1)} \star p](g) \;=\; \sum_{h\in G} p^{(n-1)}(gh^{-1})p(h),$$

where $g \in G$. The spectral radius is given by $\varrho(p) = \limsup_{n\to\infty} p^{(n)}(e)^{1/n}$. Define the entropy $H_{n,p}$ and the drift $L_{n,p}$ of $p^{(n)}$ by:

$$H_{n,p} := -\sum_{g\in G} p^{(n)}(g) \ln p^{(n)}(g), \quad L_{n,p} := \sum_{g\in G} |g| p^{(n)}(g),$$

and the average entropy $h_p$ and the linear drift $\ell_p$ by

$$h_p := \lim_{n\to\infty} \frac{1}{n} H_{n,p}, \quad \ell_p := \lim_{n\to\infty} \frac{1}{n} L_{n,p}.$$

Both limits exist by subadditivity and Fekete's Lemma. The linear drift makes sense as soon as $\sum_{g\in G} |g|p(g) < +\infty$, the entropy under the slightly weaker condition $H_{1,p} < +\infty$. The entropy $h_p$ was introduced by Avez ([**Av**]) and is related to bounded solutions of the equation on $G$ of the form $f(g) = \sum_{h\in G} f(gh)p(h)$ (see e.g. [**KV**]). In particular, $h_p = 0$ if and only if the only bounded solutions are the constant functions ([**KV**], [**De2**]). The general relation is ([**Gu**])

$$h_p \;\le\; \ell_p v, \tag{1}$$

where $v := \lim_{n\to\infty} \frac{1}{n} \ln(\#\{g \in G; |g| \leq n\})$ is the *volume entropy* of $G$. In particular, if $\ell_p = 0$ then $h_p = 0$.

We say that $p$ is symmetric if $B = B^{-1}$ and $p(g) = p(g^{-1})$ for all $g \in B$. We call $p$ centered if $\sum_{g\in B} \chi(g)p(g) = 0$ for all group morphisms $\chi : G \to \mathbb{R}$. Clearly, symmetric probabilities are centered. If $p$ is centered and $h_p = 0$, then $\ell_p = 0$ ([**Va**], [**Ma1**]). If $p$ is not centered, we may have $h_p = 0$ and $\ell_p \neq 0$, for instance on $\mathbb{Z}$. If this is the case, there is a group morphism $\chi : G \to \mathbb{R}$ such that $\ell_p = \sum_{g\in B} \chi(g)p(g)$ ([**KL**], see also [**EKn**] for finite versions of this result).

Both $h_p$ and $\ell_p$ describe asymptotic properties of the *random walk* directed by $p$. Let $(\Omega, P) = (G^{\mathbb{N}}, p^{\otimes\mathbb{N}})$ be the infinite product space such that $\omega = (\omega_1, \omega_2, \ldots) \in G^{\mathbb{N}}$ is realized by a sequence of i.i.d. random variables with values in $G$ and distribution $p$. We form the right random walk by $X_n(\omega) := \omega_1\omega_2 \cdots \omega_n$. The probability $p^{(n)}$ is the distribution of $X_n$, and an application of Kingman's subadditive ergodic theorem ([**Ki**]) gives that, for $P$-a.e. $\omega$,

$$\lim_{n\to\infty} \frac{1}{n}|X_n| \;=\; \ell_p \quad \text{and} \quad \lim_{n\to\infty} -\frac{1}{n} \ln\big(p^{(n)}(X_n)\big) \;=\; h_p. \tag{2}$$

The random walk is said to be recurrent if, for $P$-a.e. $\omega$, there is a positive $n \in \mathbb{N}$ with $X_n(\omega) = e$. In this case there is an infinite number of integers $n$ with $X_n = e$ and, by (2), $\ell_p = 0$. Hence, $h_p = 0$. From here on, we assume that the random walk is transient, i.e. $|X_n| \to \infty$ for $P$-a.e. $\omega$. The *Green function* $G(g, h)$, $g, h \in G$, is defined by

$$G(g, h) \;:=\; \sum_{n\geq 0} p^{(n)}(g^{-1}h).$$

By decomposing of the first visit to $h$ and using transitivity of the random walk we get

$$G(g, h) \;=\; F(g, h)G(h, h) \;=\; F(g, h)G(e, e),$$

where $F(g, h)$ is the probability of reaching $h$ starting from $g$. If $p$ is symmetric, then the (left invariant) Green distance is defined by $d_G(g, h) := -\ln F(g, h)$. The drift $\ell_{p,G}$ for that distance coincide with the entropy $h_p$ ([**BP**], Proposition 6.2, [**BHM**]) and the volume entropy is 1, so that there is equality in (1) for that distance ([**BHM**]).

We now turn to another representation of the drift and entropy. Let $X$ be a compact space. $X$ is called a $G$-space if the group $G$ acts by continuous transformations on $X$. This action extends naturally to probability measures on $X$. We say that the measure $\nu$ on $X$ is stationary if $\sum_{g\in G}(g_*\nu)p(g) = \nu$. The *entropy* of a stationary measure $\nu$ is defined by

$$h_p(X, \nu) \;:=\; -\sum_{g\in G} \left( \int_X \ln \frac{dg_*^{-1}\nu}{d\nu}(\xi) d\nu(\xi) \right) p(g). \tag{3}$$

The entropy $h_p$ and the linear drift $\ell_p$ are given by variational formulas over stationary measures (see [**KV**] for the entropy, [**KL**] for the linear drift):

$$h_p \;=\; \max\{h_p(X, \nu); X \text{ } G\text{-space and } \nu \text{ stationary on } X\}, \tag{4}$$

$$\ell_p \;=\; \max\Big\{ \sum_{g\in G} \left( \int_{\overline{G}} \xi(g^{-1}) d\nu(\xi) \right) p(g); \nu \text{ stationary on } \overline{G} \Big\}, \tag{5}$$

where $\overline{G}$ is the Busemann compactification of $G$, the elements of which are horofunctions $\xi$ on $G$. A pair $(X, \nu)$, where $X$ is a $G$-space and $\nu$ a $p$-stationary measure is called a

*boundary* if, for $P$-a.e.$\omega$, $(X_n(\omega))_*\nu$ converge towards a Dirac measure. It is called a *Poisson boundary* if it is a boundary and it realizes the maximum in formula (4).

From the definition of $\ell_p$ and $h_p$, one sees that the mappings $p \mapsto \ell_p$ and $p \mapsto h_p$ are uppersemicontinuous on $\mathcal{P}(B)$. Erschler ([**Er**]) raised the question of continuity of these functions and gave examples where these mappings are not continuous on the closure of $\mathcal{P}(B)$. The question of continuity in general on the interior of $\mathcal{P}(B)$ is open. In the rest of the paper, we discuss several examples where one can prove stronger regularity results.

## 2. Nearest neighbour random walks on a free group

In the case when the group $G$ is a *free group* with $d$ generators, $d \geq 2$, and $p$ is supported by these generators, explicit computations can be made (see [**DM**]).

Let $G$ be the free group with set of generators $S = \{\pm i; i = 1, \dots, d\}$, where $-i = i^{-1}$ for $i \in S$. Let $\mathcal{P}(S)$ be the set of probability measures on $G$ with support $S$. Since $d \geq 2$, as $n$ goes to infinity, the reduced word representing $X_n(\omega)$ converges towards an infinite reduced word $X_\infty(\omega) = s_1(\omega)s_2(\omega)\cdots$ with $s_j(\omega) \neq -s_{j+1}(\omega)$. Denote by $G_\infty$ the space of infinite reduced words. The stationary measure is unique: it is the distribution $\nu$ of $X_\infty(\omega)$. Then $(G_\infty, \nu)$ is both the Poisson boundary and the Busemann boundary of $G$. Let $q_i = P\big(\{\omega; s_1(\omega) = i\}\big) = \nu([i])$, where $[i]$ consists of all infinite words in $G_\infty$ starting with letter $i \in S$. We have $\sum_{i\in S} q_i = 1$. Let $i_1 \dots i_k$ be a reduced word in $G$. Then $\nu$ is uniquely determined by the values $\nu([i_1 \dots i_k]) = F(e, i_1 \dots i_k)(1 - q_{-i_k})$, where $F(e, i_1 \dots i_k)$ is the probability of hitting $i_1 \dots i_k$ when starting at the identity $e$. Formula (5) writes:

$$\ell_p = 1 - 2\sum_{i\in S} p_i q_{-i}.$$

In order to write the formula for the entropy, we introduce $z_i := F(e, i)$ for $i \in S$. The density $\dfrac{dg_*^{-1}\nu}{d\nu}(\xi)$ gives the minimal positive harmonic function with pole at $\xi = i_1 i_2 \dots \in G_\infty$. The Green function satisfies the following multiplicative structure:

$$G(e, i_1 \dots i_k) = F(e, i_1)G(i_1, i_1 \dots i_k) = F(e, i_1)G(e, i_2 \dots i_k).$$

This yields together with [**L1**, Theorem 2.10]

$$\frac{di_*^{-1}\nu}{d\nu}(\xi) \;=\; \lim_{k\to\infty} \frac{G(-i, i_1 \dots i_k)}{G(e, i_1 \dots i_k)} = \begin{cases} z_i, & \text{if } i_1 \neq -i, \\ z_{-i}^{-1}, & \text{if } i_1 = -i. \end{cases}$$

Formula (4) writes:

$$h_p = \sum_{i\in S} p_i \big[q_{-i} \ln z_{-i} - (1 - q_{-i}) \ln z_i\big].$$

We can express the $q_i$ in terms of the $p_i$, and vice versa, thanks to the *traffic equations*: using the Markov property, we can write:

$$z_i \;=\; p_i + z_i \sum_{j\in S\setminus\{i\}} p_j z_{-j} \quad \text{and} \quad q_i \;=\; z_i(1 - q_{-i}).$$

Setting $Y := \sum_{j\in S} p_j z_{-j}$, we get

$$p_i = \frac{z_i(1 - Y)}{1 - z_i z_{-i}} \quad \text{and} \quad z_i \;=\; \frac{q_i}{1 - q_{-i}},$$

so that we find:

$$\ell_p \;=\; 1-\frac{2}{A}\sum_{i\in S}\frac{q_iq_{-i}(1-q_i)}{1-q_i-q_{-i}},\quad \text{where}\quad A\;=\;(1-Y)^{-1}\;=\;1+\sum_{i\in S}\frac{q_iq_{-i}}{1-q_i-q_{-i}},$$

which writes:

$$\ell_p=\frac{B}{A},\ \text{where } B:=1-\sum_{i\in S}\frac{q_iq_{-i}(1-2q_i)}{1-q_i-q_{-i}}.$$

Hence, in terms of the $q_i, i\in S$, $p_i$ and $\ell_p$ are rational, and the expression of $h_p$ involves rational functions and $\ln q_i, \ln(1-q_i)$.

**Proposition 2.1.** *The mappings $p\mapsto \ell_p$ and $p\mapsto h_p$ are real analytic on $\mathcal{P}(S)$.*

*Proof.* Since all formulas are explicit in terms of the $q_i$, we only have to check that the $q_i$ are real analytic functions on $\mathcal{P}(S)$. First, we can write $z_i$ as a power series in terms of the $p_i$'s and the additional variable $z\in\mathbb{C}$, namely as

$$z_i(z)=\sum_{(n_1,\dots,n_{2d})\in\mathbb{N}^{2d}} c(n_1,\dots,n_{2d})p_1^{n_1}p_{-1}^{n_2}p_2^{n_3}p_{-2}^{n_4}\cdot\ldots\cdot p_d^{n_{2d-1}}p_{-d}^{n_{2d}}z^{n_1+\dots+n_{2d}},$$

where $c(n_1,\dots,n_{2d})\geq 0$. Since the spectral radius is strictly smaller than 1 (see e.g. [**W**, Cor. 12.5]), the power series $G(e,i|z)=\sum_{n\geq 0}p^{(n)}(i)z^n$ has radius of convergence strictly bigger than 1 and satisfies $G(e,i|z)\geq z_i(z)$ for all real $z>0$. That is, for each $p\in\mathcal{P}(S)$, $z_i(z)$ has radius of convergence $R_i>1$. Choose now any $\delta>0$ with $1+\delta<R_i$. Then

$$\begin{aligned} z_i &= z_i(1)\leq z_i(1+\delta)\\ &= \sum_{(n_1,\dots,n_{2d})\in\mathbb{N}^{2d}} c(n_1,\dots,n_{2d})\big((1+\delta)p_1\big)^{n_1}\cdot\ldots\cdot\big((1+\delta)p_{-d}\big)^{n_{2d}}<\infty.\end{aligned}$$

In other words, $z_i=z_i(1)$ is real analytic in a neighbourhood of any $p\in\mathcal{P}(S)$. The equations $q_i=z_i(1-q_{-i}), q_{-i}=z_{-i}(1-q_i)$ give

$$q_i=\frac{z_i(1-z_{-i})}{1-z_iz_{-i}},$$

and this finishes the proof. □

Observe that for $d=1$, the group $G$ is $\mathbb{Z}$, $S=\{\pm 1\}$ and $p\mapsto\ell_p=|p_1-p_{-1}|$ is not a real analytic function on $\mathcal{P}(\pm 1)$.

The formulas are even simpler when the probability $p$ is symmetric. Let $\mathcal{P}_\sigma(S)$ be the set of symmetric probability measures on $S$; elements of $\mathcal{P}_\sigma(S)$ are described by $d$ positive numbers $\{p_1,\cdots,p_d\}$ such that $\sum_{i=1}^d p_i=1/2$. If $p\in\mathcal{P}_\sigma(S)$, $q_i=q_{-i}$ and we have:

$$\begin{aligned}\ell_p &= \frac{B}{A}\ \text{with } A=1+2\sum_{i=1}^d\frac{q_i^2}{1-2q_i}\ \text{and } B=1-2\sum_{i=1}^d q_i^2,\\ h_p &= -\frac{2}{A}\sum_{i=1}^d q_i(1-q_i)\ln\frac{q_i}{1-q_i},\text{whereas}\\ p_i &= \frac{q_i(1-q_i)}{A(1-2q_i)}\quad \text{for } i=1,\dots,d.\end{aligned}$$

**Proposition 2.2.** *The functions $p \mapsto \ell_p$ and $p \mapsto h_p$ reach their maxima on $\mathcal{P}_\sigma(S)$ at the constant vector $p_0 = (1/2d, \dots, 1/2d)$ and*

$$\ell_{p_0} = 1 - \frac{1}{d}, \quad h_{p_0} = \big(1 - \frac{1}{d}\big) \ln(2d-1).$$

*Proof.* By symmetry, the constant vector $p_0$ is a critical point for $\ell_p$. At $p_0$, $q_i = 1/2d$ by symmetry and $\ell_{p_0} = 1 - 1/d, h_{p_0} = (1-1/d)\ln(2d-1)$ by the formulas above (observe that these expressions are also valid for $d = 1$: the only point of $\mathcal{P}_\sigma(\pm 1)$ is $(1/2, 1/2)$, for which $\ell = 0 = 1 - 1/d$ and $h = 0 = (1-1/d)\ln(2d-1)$). Moreover, the volume entropy of $G$ is $\ln(2d-1)$. By (1), the result for $\ell_p$ implies that $(1-1/d)\ln(2d-1)$ is the maximal value that the entropy might take on $\mathcal{P}_\sigma(S)$. Since this is the entropy $h_{p_0}$, $p_0$ achieves the maximum of the entropy as well.

We are going to prove that the function $(q_1, \dots, q_d) \mapsto B/A$ has a unique critical point on the set $\{(q_1, \dots, q_d); q_j > 0, \sum_{j=1}^d q_j = 1/2\}$. Observe that the formula for $\ell_p$ is continuous on the domain $0 \le q_i \le 1/2$ and that the value of $\ell_p$ at the boundary of the domain $\{(q_1, \dots, q_d); q_j > 0, \sum_{j=1}^d q_j = 1/2\}$ is the one computed with only the non-zero $q_i$'s on a free group with a smaller set of generators. Since at the constant vector $p_0$, $\ell_{p_0} = 1 - 1/d$, it follows, by induction on the dimension, that the critical point $p_0$ is a maximum. The proof for $d = 2$ is the same as in the general case: there is only one critical point by the argument below and the limit of the expression for $\ell_p$ at $(0, 1/2), (1/2, 0)$ is 0.

Using a Lagrange multiplier, we are looking for the critical points of the function $F(q, \lambda) = \ell_p - \lambda(\sum_{j=1}^d q_j - 1/2)$ satisfying $0 \le q_j \le 1/2$ for $j = 1, \dots, d$. Setting as above

$$A = 1 + 2\sum_{i=1}^d \frac{q_i^2}{1-2q_i} \quad \text{and} \quad B = 1 - 2\sum_{i=1}^d q_i^2,$$

all equations $\frac{\partial F}{\partial q_i} = 0$ depend only on $A, B, \lambda$ and $q_i$.

Indeed, they write $G(A, B, \lambda, q_i) = 0$, where:

$$G(A, B, \lambda, q) = 16Aq^3 + 4q^2(\lambda A^2 - 4A - B) + 4q(-\lambda A^2 + A + B) + \lambda A^2.$$

If, for fixed $A, B, \lambda$, the equation $G(A, B, \lambda, q_i) = 0$ has only one solution $q \in [0, 1/2)$, then, for these values of $A, B, \lambda$, the only possible critical point of $F$ is $q_j = 1/2d$ for all $j$. Then, unless $A = \frac{2d-1}{2d-2}, B = \frac{2d-1}{2d}$, there is no critical point for $F$ with those values of $A, B$.

To summarize, we only have to verify that the equation $G(A, B, \lambda, q_i) = 0$ has at most one solution $q \in [0, 1/2)$ for all $A, B, \lambda$ with $0 < B < 1 < A$.

The function $q \mapsto G(A, B, \lambda, q)$ is a third degree polynomial with positive highest coefficient, $1/2$ is a critical point and $G(A, B, \lambda, 1/2) = B > 0$. Therefore, there is at most one solution $q \in [0, 1/2)$. □

It is likely that $p_0$ gives also the maximum of the entropy on the whole $\mathcal{P}(S)$, but we do not have a proof of that fact. We also conjecture that the mapping $p \mapsto \ell_p$ is a concave function; calculating the drift for small $d \in \mathbb{N}$ supports and confirms this conjecture, but we do not have a proof for general $d$.

## 3. Free products, Artin dihedral groups and braid groups

The computations in Section 2 have been known for fifty years (even if Proposition 2.2 seems to be formally new). There are very few other examples where it is possible

to describe geometrically the Poisson boundary and the Busemann boundary, and it is even rarer to be able to give useful formulas for the stationary measure. In this section, we review the examples we are aware of.

One important concept of constructing new groups from given ones is the free product of groups. The crucial point is that free products have a tree-like structure. More precisely, suppose we are given finitely generated groups $G_1, \dots, G_r$ equipped with finitely supported probability measures $p_1, \dots, p_r$. The identity of $G_i$ is denoted by $e_i$, and w.l.o.g. we assume that these groups are pairwise disjoint and we exclude the case $r = 2 = |G_1| = |G_2|$ (this case leads to recurrent random walks in our setting). The free product $G_1 * \dots * G_r$ is given by

$$G = *_{i=1}^r G_i = \big\{x_1 x_2 \dots x_n \big| x_j \in \bigcup_{i=1}^r G_i \setminus \{e_i\}, x_j \in G_k \Rightarrow x_{j+1} \notin G_k\big\} \cup \{e\},$$

the set of finite words over the alphabet $\bigcup_{i=1}^r G_i \setminus \{e_i\}$ such that two consecutive letters do *not* come from the same group $G_k$, where $e$ describes the empty word. A group operation on $G$ is given by concatenation of words with possible contractions and cancellations in the middle such that one gets a reduced word as above. For $x = x_1 \dots x_n \in G$ define the *block length* of $x$ as $\|x\| := n$.

A random walk on $G$ is constructed in a natural way as follows: we lift $p_i$ to a probability measure $\bar{p}_i$ on $G$: if $x = x_1 \dots x_n \in G$ with $x_n \notin G_i$ and $v, w \in G_i$, then $\bar{p}_i(xv, xw) := p_i(v, w)$. Otherwise we set $\bar{p}_i(x, y) := 0$. Choose $0 < \alpha_1, \dots, \alpha_r \in \mathbb{R}$ with $\sum_{i=1}^r \alpha_i = 1$. Then we obtain a new probability measure on $G$ defined by

$$p = \sum_{i=1}^r \alpha_i \bar{p}_i$$

with $B = \operatorname{supp}(p) = \bigcup_{i=1}^r \operatorname{supp}(p_i)$. We consider random walks $(X_n)_{n\in\mathbb{N}_0}$ on $G$ starting at $e$, which are governed by $p$. For $i \in \{1, \dots, r\}$, denote by $\xi_i$ the probability of hitting the set $G_i \setminus \{e_i\}$ when starting at $e$. The spectral radius $\varrho(p)$ is strictly less than 1 due to the non-amenability of $G$. Let $\partial G_i$ be the Martin boundary of $G_i$ with respect to $p_i$, and denote by $G_\infty$ the set of infinite words $x_1 x_2 \dots$ such that $x_i \in G_k$ implies $x_{i+1} \notin G_k$. Then the Martin boundary of $G$ is given by

$$\partial G = G_\infty \cup \bigcup_{i=1}^r \{x\xi; x = x_1 \dots x_n \in G, x_n \notin G_i, \xi \in \partial G_i\};$$

see e.g. [**W**, Proposition 26.21]. The random walk on $G$ converges almost surely to an infinite word in $G_\infty$, and the limit distribution $\nu$ is determined by

$$\nu\big(\{x_1 x_2 \dots \in G_\infty; x_1 = y_1, \dots, x_n = y_n\}\big) = F(e, y_1 \dots y_n)\big(1 - (1 - \xi_i)G_i(\xi_i)\big),$$

where $n \in \mathbb{N}$, $y_1 \dots y_n \in G$ with $y_n \notin G_i$, $F(e, y_1 \dots y_n)$ being the probability of hitting $y_1 \dots y_n$ and $G_i(e_i|z) = \sum_{n\geq 0} p_i^{(n)}(e_i) z^n$ with $z \in \mathbb{C}$.

The next propositions summarize results about regularity of drift and entropy. Explicit formulas can be found in the cited sources.

**Proposition 3.1** ([**Gi1**])**.** *The drift w.r.t. the block length $\ell_B = \lim_{n\to\infty} \frac{1}{n}\|X_n\|$ exists and varies real-analytically in $p \in \mathcal{P}(B)$.*

*Proof.* In [**Gi1**, Equ. (9)] a formula for $\ell_B$ is given:

$$\ell_B = \sum_{i=1}^{r} \alpha_i \, \frac{1-\xi_i}{\xi_i} \, \big(1 - (1-\xi_i)\, G_i(e_i|\xi_i)\big).$$

Let $d = |B| - 1$, and write $p = (q_1, \ldots, q_d) \in \mathcal{P}(B)$. Analogously to the proof of Proposition 2.1 one can write $\xi_i$ as a power series (evaluated at $z = 1$) in the form

$$\xi_i(z) = \sum_{(n_1,\ldots,n_d)\in\mathbb{N}^d} c(n_1, \ldots, n_d) q_1^{n_1} q_2^{n_2} \ldots q_d^{n_d} z^{n_1+\cdots+n_d}, \quad z \in \mathbb{C}.$$

Since $\varrho(p) < 1$ the Green functions $G(g|z) = \sum_{n\geq 0} p^{(n)}(e) z^n$, $g \in G$, have radii of convergence $R = 1/\varrho(p) > 1$ and dominate $\xi_i(z)$ for real $z > 0$. Hence, $\xi_i(z)$ has radius of convergence bigger than 1, which in turn – following the same argumentation as in Proposition 2.1 – yields real analyticity of $\xi_i = \xi_i(1)$ in a neighbourhood of any $p \in \mathcal{P}(B)$. Furthermore, $G_i(z)$ can be expanded in the same form as $\xi_i(z)$ and, for each real positive $z_0 < 1$, the mapping $p \mapsto G_i(z_0)$ is real analytic. Since $\xi_i < 1$ (see e.g. [**Gi1**, Lemma 2.3]) the mapping $p \mapsto G_i(\xi_i)$ is aslo real-analytic as a composition of real-analytic functions. This yields the proposed statement. □

**Proposition 3.2** ([**Gi1**])**.** *Let $p$ govern a nearest neighbour random walk on $G$, that is, the length $|g|$ is computed with respect to the generator $B$. Then the drift function $p \mapsto \ell_p = \lim_{n\to\infty} \frac{1}{n}|X_n|$ is real-analytic.*

*Proof.* By the formula for $\ell$ given in [**Gi1**, Section 7] we just have to check that the mapping

$$p \mapsto \widetilde{G}_j(y, z) := \sum_{m,n\geq 0} \sum_{x\in G_j : |x|=m} p_j^{(n)}(e_j, x) y^n z^m$$

is real-analytic for all $y \in (0, 1)$ and $z = 1$. For a moment fix $y < 1$ and choose $\delta > 0$ small enough such that $y(1 + 2\delta)^2 < 1$. Since $p_j^{(n)}(e_j, x) > 0$, $x \in G_j$ with $|x| = m$, implies $n \geq m$, we get

$$\widetilde{G}_j\big(y, (1+2\delta)^2\big) = \sum_{m,n\geq 0} \sum_{x\in G_j : |x|=m} p_j^{(n)}(e_j, x) \big(y(1+\delta)^2\big)^n \leq \frac{1}{1 - y(1+2\delta)^2} < \infty.$$

This yields $\frac{\partial}{\partial z}\widetilde{G}\big(y, (1+\delta)^2\big) < \infty$. Since each term $p_j^{(n)}(e_j, x)$ can be written as a polynomial

$$\sum_{\substack{(n_1,\ldots,n_d)\in\mathbb{N}^d: \\ n_1+\ldots+n_d=n}} c(n_1, \ldots, n_d) q_1^{n_1} \cdot \ldots \cdot q_d^{n_d},$$

where $p = (q_1, \ldots, q_d) \in \mathcal{P}(B)$ and $c(n_1, \ldots, n_d) \geq 0$, we can rewrite $\frac{\partial}{\partial z}\widetilde{G}\big(y, (1+\delta)^2\big)$ as

$$\frac{1}{1+\delta} \sum_{m,n\geq 0} \sum_{x\in G_j : |x|=m} m\big(p_j^{(n)}(e_j, x)(1+\delta)^n\big)\big(\xi_j(1+\delta)\big)^n.$$

That is, $\sum_{m,n\geq 0} \sum_{x\in G_j : |x|=m} m p_j^{(n)}(e_j, x) \xi_j^n$ is real-analytic in $\mathcal{P}(B)$ as a composition of real-analytic functions, and this yields the claim. □

Let us mention that – in contrast to Proposition 2.2 – simple random walk is not necessarily the fastest random walk. Namely, it can be verified with the help of MATHEMATICA that – with $p_i$ describing the simple random walk on $G_i$ – the simple random

walk on $(\mathbb{Z}/3\mathbb{Z}) * (\mathbb{Z}/2\mathbb{Z})$ (that is, $\alpha_1 = 2/3, \alpha_2 = 1/3$) is slower than the random walk on the free group with the parameters $\alpha_1 = \alpha_2 = 1/2$.

Furthermore, we have the following regularity result:

**Proposition 3.3** ([**Gi3**]). *Assume that $h_i := -\sum_{g\in G_i} p_i(g) \ln p_i(g) < \infty$ for all $i \in \{1, \dots, r\}$, that is, all random walks on the factors $G_i$ have finite single-step entropy. Then the mapping $p \mapsto h_p$ is real analytic.*

We now turn to another class of groups whose Cayley graphs have a tree-like structure. A group $G$ is called *virtually free* if it has a free subgroup of finite index. At this point we assume that $G$ has a free subgroup with at least $d \geq 2$ generators; otherwise, $G$ is a finite extension of $\mathbb{Z}$ where we either get recurrent random walks or non-regularity points on $\mathbb{P}(B)$. It is well-known that virtually free groups can be constructed from a finite number of finite groups by iterated amalgamation and HNN extensions. Each element of $G$ can be written as $x_1 \dots x_n h$, where $x_i \in \{\pm i; i = 1, \dots d\}$ and $h$ being one of finitely many representatives for the different cosets. Suppose we are given a weight or length function $l(\pm i) \in \mathbb{R}$ for $i \in \{1, \dots, d\}$. Then a natural length function on $G$ is defined by $l(x_1 \dots x_n h) = \sum_{j=1}^{n} l(x_j)$. We have the following result:

**Proposition 3.4** ([**Gi2**]). *Let $G$ be a virtually free group. Let $p$ govern a finite range random walk on $G$. Then the mapping $p \mapsto \lim_{n\to\infty} l(X_n)/n$ is real-analytic.*

*Proof.* Random walks on virtually free groups can be interpreted as a random walk on a regular language in the sense of [**Gi2**]. The claim follows from the formula for $\lim_{n\to\infty} l(X_n)/n$, the drift with respect to the length function $l$, given in [**Gi2**, Theorem 2.4]. Due to non-amenability of $G$ we have again $\varrho(p) < 1$. The rest follows analogously as in the proofs of Propositions 2.1 and 3.1. □

For the case $l$ being the natural word length the last proposition is also covered by Corollary 4.2.

At this point we want to mention the article [**MM2**], which uses similar techniques to establish statements about the drift of random walks on the braid group $B_3$ and on Artin groups of dihedral type. Traffic equations are established, whose unique solutions lead to formulas for the drift. For random walks on these groups there might occur transitions (when varying the probability measures of constant support), where one has no regularity. An explicit example for a non-differentiabilty point is given on the braid group $B_3$. However, [**MM2**] gives explicit formulas for the drift in terms of the solutions of the traffic equations splitted up into different branches. By methods similar to the above, one can show that the drift is real-analytic on each branch. Indeed, solutions of the traffic equations can be written as converging power series as in the proofs of Propositions 3 and 3.2.

## 4. Hyperbolic groups

A geodesic metric space is called *hyperbolic* if geodesic triangles are thin: there is $\delta \geq 0$ such that each side of a geodesic triangle is contained in a $\delta$-neighborhood of the union of the other two sides. A finitely generated group is called hyperbolic if the Cayley graph defined by some finite symmetric generator is hyperbolic. This property does not depend on the set of generators. Free groups are hyperbolic, as are fundamental groups of compact manifolds of negative curvature, and small cancellation groups. See e.g. [**GH**] for the main geometric properties of hyperbolic groups. The geometric boundary of a hyperbolic space is the space of equivalence classes of geodesic rays, where two geodesic

rays are equivalent if they are at a bounded Hausdorff distance. The geometric boundary $\partial G$ of the Cayley graph of a hyperbolic group $G$ is a compact $G$-space. It is endowed with the Gromov metric (see [**GH**]). The mapping $\Phi : G \to \mathbb{Z}^G, \Phi(g)(h) = |h^{-1}g| - |g|$ is an isometry such that $\Phi(G)$ is relatively compact for the product topology. The Busemann compactification $\overline{G}$ is the closure of $\Phi(G)$ in $\mathbb{Z}^G$. There is an equivariant homomorphism $\pi : \overline{G} \setminus G \to \partial G$ (see e.g. [**WW**]). The homomorphism $\pi$ is finite-to-one (see e.g. [**CP**]). Following [**Bj**], we say that $G$ with the generator $S$ satisfies (BA) if the homomorphism $\pi$ is one-to-one. In this case, we write, for $\xi \in \partial G, h \in G$, $\xi(h)$ for the value at $h$ of the sequence $\pi^{-1}\xi \in \mathbb{Z}^G$. Free groups and surface groups with their natural generators satisfy (BA). It is an open problem whether any hyperbolic group admits a symmetric generator with the property (BA).

Let $p$ be a probability on $G$ with finite support. Then, there is a unique $p$-stationary probability measure $\nu_p$ on $\partial G$ and $(\partial G, \nu_p)$ is a Poisson boundary for $(G, p)$([**An**], [**K**] Theorem 7.6). If (BA) is satisfied, the measure $\nu_p$ is the unique stationary probability measure on the Busemann compactification and formulas (4) and (5) write:

(6)

$$h_p \;=\; -\sum_{g\in B}\left(\int_{\partial G} \ln \frac{dg_*^{-1}\nu_p}{d\nu_p}(\xi)d\nu_p(\xi)\right)p(g),\;\; \ell_p \;=\; \sum_{g\in B}\left(\int_{\partial G} \xi(g^{-1})d\nu_p(\xi)\right)p(g).$$

**Proposition 4.1.** *Assume that $(G, S)$ is a non-elementary hyperbolic group and satisfies (BA). Let $p \in \mathcal{P}(B)$, $\alpha$ be small enough, and let $f$ be an $\alpha$-Hölder continuous function on $\partial G$. Then the mapping $p \mapsto \int_{\partial G} f(\xi)d\nu_p(\xi)$ is real analytic on a neighborhood of $p$ in $\mathcal{P}(B)$.*

*Proof.* Let $\mathcal{K}_\alpha$ be the space of $\alpha$-Hölder continuous functions on $\partial G$. The space $\mathcal{K}_\alpha$ is a Banach space with norm $\|f\|_\alpha$, where

$$\|f\|_\alpha \;=\; \max_{\xi\in\partial G}|f(\xi)| + \sup_{\xi,\eta\in\partial G:\xi\neq\eta}\frac{|f(\xi)-f(\eta)|}{(d(\xi,\eta))^\alpha}.$$

For $p \in \mathcal{P}(B)$, let $\mathcal{Q}_p$ be the operator on $\mathcal{K}_\alpha$ defined by

$$\mathcal{Q}_p f(\xi) \;=\; \sum_{g\in B} f(g^{-1}\xi)p(g).$$

Clearly, the mapping $p \mapsto \mathcal{Q}_p$ is real analytic from $\mathcal{P}(B)$ into $\mathcal{L}(\mathcal{K}_\alpha)$. If $G$ is not elementary and satisfies (BA), it can be shown (see [**Bj**], Lemma 4) that, for $\alpha$ small enough, $f \mapsto \int f d\nu_p$ is an isolated eigenvector for the transposed operator $\mathcal{Q}_p^*$ on the dual space $\mathcal{K}_\alpha^*$. The proposition follows by a perturbation lemma. □

**Corollary 4.2.** *Assume that $(G, S)$ is a non-elementary hyperbolic group and satisfies (BA). Then the mapping $p \mapsto \ell_p$ is real analytic.*

Indeed, the function $\xi(g^{-1})$ in formula (6) belongs to $\mathcal{K}_\alpha$ for all $\alpha$. Corollary 4.2 is due to [**L2**] in the case of the free group. P. Mathieu ([**Ma2**]) proved the $C^1$ regularity and gave a formula for $\nabla_p\nu_p$ and $\nabla_p\ell_p$ in the symmetric case, to be compared with formulas for linear response of dynamical systems (cf. [**R**]).

The formula (6) for the entropy is valid in general, even without the (BA) hypothesis, but observe that the integrand $\varphi_p(g, \xi) := -\ln \frac{dg_*^{-1}\nu_p}{d\nu_p}(\xi)$ is itself a function of $p$. To study this function, we use the description by A. Ancona ([**An**]) of the Martin boundary

of a random walk with finite support on a hyperbolic group. Recall that $F_p(g,h)$ is the probability of reaching $h$ starting from $g$ in dependence of $p$.

**Proposition 4.3** ([**An**]). *Assume that $G$ is hyperbolic and that $p$ has finite support. Then,*

$$\varphi_p(g,\xi) \;=\; \lim_{h\to\xi} \ln \frac{F_p(e,h)}{F_p(g^{-1},h)} \qquad \text{for all } g\in G, \xi\in\partial G.$$

A consequence of the proof of Proposition 4.3 is that, for all $g \in G$, for $\alpha$ small enough $\varphi_p(g,\xi) \in \mathcal{K}_\alpha$ (see [**INO**]). In the case of free groups, Proposition 4.3 goes back to Derriennic ([**De1**]) and using his arguments one can prove:

**Proposition 4.4** ([**L2**]). *If $G$ is a free group and $p$ has finite support $B$, there is $\alpha$ small enough that, for all $g \in B$, the mapping $p \mapsto \varphi_p$ is real analytic from a neighborhood of $p$ in $\mathcal{P}(B)$ into $\mathcal{K}_\alpha$.*

**Corollary 4.5** ([**L2**]). *If $G$ is a free group and $p$ has finite support $B$, the mapping $p \mapsto h_p$ is real analytic on $\mathcal{P}(B)$.*

For cocompact Fuchsian groups there is the following recent result:

**Proposition 4.6** ([**HMM**]). *Let $G$ be a cocompact Fuchsian group with planar presentation. Then the mapping $p \mapsto \ell_p$ is real analytic.*

For a general hyperbolic group, we have a weaker result:

**Proposition 4.7** ([**L3**]). *If $G$ is a hyperbolic group and $p$ has finite support $B$, there is $\alpha$ small enough that, for all $g \in B$, the mapping $p \mapsto \varphi_p$ is Lipschitz continuous from a neighborhood of $p$ in $\mathcal{P}(B)$ into $\mathcal{K}_\alpha$.*

**Corollary 4.8** ([**L3**]). *If $G$ is a hyperbolic group and $p$ has finite support $B$, the mappings $p \mapsto h_p$, $p \mapsto \ell_p$ are Lipschitz continuous on $\mathcal{P}(B)$.*

[**Gi4**] proves also continuity of the mapping $p \mapsto h_p$ for random walks on regular languages, which adapt, for instance, to the case of virtually free groups.

The best results of regularity to-date are due to P. Mathieu; in particular:

**Proposition 4.9** ([**Ma2**]). *If $G$ is a hyperbolic group satisfying (BA), $B$ is finite and symmetric and $\lambda \mapsto p_\lambda, \lambda \in [-\varepsilon,+\varepsilon]$ is a smooth curve in the set $\mathcal{P}_\sigma(B)$ of symmetric probability measures on $B$, then the mapping $\lambda \mapsto h_{p_\lambda}$ is differentiable .*

Let $\lambda \mapsto p_\lambda, \lambda \in [-\varepsilon,+\varepsilon]$ be a smooth curve in $\mathcal{P}(B)$. We write:

$$\begin{aligned}
&\lim_{\lambda\to 0} \frac{h(p_\lambda)-h(p_0)}{\lambda}\\
=\;& \lim_{\lambda\to 0} \frac{1}{\lambda} \sum_{g\in B} \left( \int_{\partial G} \varphi_{p_\lambda}(g,\xi) d\nu_{p_\lambda}(\xi) p_\lambda(g) - \int_{\partial G} \varphi_{p_0}(g,\xi) d\nu_{p_0}(\xi) p_0(g) \right)\\
=\;& \lim_{\lambda\to 0} \frac{1}{\lambda} \sum_{g\in B} \left( \int_{\partial G} (\varphi_{p_\lambda}(g,\xi) - \varphi_{p_0}(g,\xi)) d\nu_{p_\lambda}(\xi) \right) p_\lambda(g) +\\
&\quad + \sum_{g\in B} \left( \int_{\partial G} \varphi_{p_0}(g,\xi) [\lim_{\lambda\to 0} \frac{1}{\lambda} (d\nu_{p_\lambda} - d\nu_{p_0})](\xi) \right) p_\lambda(g) +\\
&\quad\quad + \sum_{g\in B} \left( \int_{\partial G} \varphi_{p_0}(g,\xi) d\nu_{p_0}(\xi) \right) [\lim_{\lambda\to 0} \frac{1}{\lambda} (p_\lambda(g) - p_0(g)].
\end{aligned}$$

The third line converges by definition. To prove that the second line converges, P. Mathieu observes that the Green metric $-\ln F_{p_0}(g,h)$ on $G$ satisfies (BA) and a form of hyperbolicity that allows him to extend Proposition 4.1. More precisely, he shows directly the differentiability of $\lambda \mapsto \int f(\xi) d\nu_{p_\lambda}(\xi)$, for $f \in \mathcal{K}_\alpha$, and gives a formula for the derivative. For the first line, P. Mathieu shows a general result for any non-amenable group $G$ and $p_\lambda$ with finite support. In our case, his result writes:

**Proposition 4.10** ([**Ma2**])**.** *Let $\lambda \mapsto p_\lambda, \lambda \in [-\varepsilon, +\varepsilon]$ be a smooth curve in $\mathcal{P}(B)$. Then,*

$$\lim_{\lambda \to 0} \frac{1}{\lambda} \sum_{g \in B} \left( \int (\varphi_{p_\lambda}(g,\xi) - \varphi_{p_0}(g,\xi)) d\nu_{p_\lambda}(\xi) \right) p_\lambda(g) \ = \ 0.$$

It is likely that the function $p \mapsto h_p$ has more regularity on $\mathcal{P}(B)$, but this is an open problem.

Another natural extension is towards more general families of probability measures on $G$. Proposition 2.1 is valid for $p$ varying in finite dimensional affine subsets of $\{p; \sum_{g \in G} e^{\gamma |g|} p(g) < +\infty\}$ for some $\gamma > 0$ (see [**L1**]). The other properties rest on Harnack inequality at infinity (see [**An**]), which has been proven only for probability measures with finite support on hyperbolic groups. Finally, let $\mathcal{P}^1(G)$ be the set of probabilities on $G$ satisfying $\sum_{g \in G} |g| p(g) < +\infty$ endowed with the topology of convergence on the functions which grow slower than $C|g|$ at infinity. The first observation on this topic of regularity of the entropy is the fact that, if $G$ is hyperbolic, $p \mapsto h_p$ is continuous on $\mathcal{P}^1(G)$ ([**EKc**]).

## References


[**An**] A. Ancona, Théorie du potentiel sur les graphes et les variétés, École d'été de Saint-Flour XVIII,1988, *Lecture Notes in maths,* **1427** (1990) 1–112.

[**Av**] A. Avez, Entropie des groupes de type fini, *C. R. Acad. Sc. Paris Sér A-B* **275** (1972), A1363–A1366.

[**Bj**] M. Bjorklund, Central Limit Theorem for Gromov Hyperbolic Groups, *J. Theo. Probability*, to appear.

[**BHM**] S. Blachère, P. Haïssinsky and P. Mathieu, Asymptotic entropy and Green speed for random walks on countable groups, *Ann. Prob.,* **36** (2008), 1134–1152.

[**BP**] I. Benjamini and Y. Peres, Tree-indexed random walks on groups and first passage percolation, *Probab. Theory Relat. Fields,* **98** (1994), 91–112.

[**CP**] M. Coornaert and A. Papadopoulos, Horofunctions and symbolic dynamics on Gromov hyperbolic groups, *Glasgow Math. Journal*, **43** (2001), 425–456.

[**De1**] Y. Derriennic, Marche aléatoire sur le groupe libre et frontière de Martin, *Z. Wahrscheinlichkeitstheorie verw. Geb.* **32** (1975), 261–276.

[**De2**] Y. Derriennic, Quelques applications du théorème ergodique sousadditif, *Astérisque* **74** (1980), 183–201.

[**DM**] E. B. Dynkin and M. B. Malyutov, Random walks on groups with a finite number of generators, *Dokl. Akad. Nauk SSSR*, **137** (1961), 1042–1045.

[**Er**] A. Erschler, On continuity of range, entropy and drift for random walks on groups, *in* Random Walks, Boundaries and Spectra, D. Lenz, F. Sobieszky and W. Woess ed., *Progress in Probability* bf 64, Birkhäuser Basel (2011), 55-64.

[**EKc**] A. Erschler and V. A. Kaimanovich, Continuity of asymptotic characteristics for random walks on hyperbolic groups, *Functional Anal. and Appl.* **47** (2013), 152–156.

[**EKn**] A. Erschler and A. Karlsson, Homomorphisms to $\mathbb{R}$ constructed from random walks, Ann. Inst. Fourier, **60** (2010), 2095–2113.

[**GH**] É. Ghys and P. de la Harpe, Sur les groupes hyperboliques d'après Mikhael Gromov, *Progress in Mathematics* Vol.83, Birkhäuser Boston Inc., Boston, MA, 1990.

[**Gi1**] L. A. Gilch, Rate of Escape of Random Walks on Free Products, *J. of Austral. Math. Soc.,* **83** (2007) 31 - 54.

[**Gi2**] L. A. Gilch, Rate of Escape of Random walks on Regular Languages and Free Products by Amalgamation of Finite Groups, *Fifth Colloquium on Mathematics and Computer Science,* Discrete Mathematics and Computer Science Proc. AI, Assoc. Discrete Math. Theor. Comput. Sci., Nancy (2008), 405–420.

[**Gi3**] L. A. Gilch, Asymptotic entropy of random walks on free products, *Elec. J. Prob.,* **16** (2011) 76–105.

[**Gi4**] L. A. Gilch, Asymptotic entropy of random walks on regular languages, preprint, 2012.

[**Gu**] Y. Guivarc'h, Sur la loi des grands nombres et le rayon spectral d'une marche aléatoire, *Astérisque*, **74** (1980), 47–98.

[**HMM**] P. Haïssinsky, P. Mathieu and S. Müller, Renewal theory for random walks on cocompact Fuchsian groups, preprint, (2012).

[**INO**] M.Izumi, S.Neshveyev and R. Okayasu, The ratio set of the hyperbolic measure of a random walk on a hyperbolic group, *Israel J. Math.*, **163** (2008), 285–316.

[**K**] V.A. Kaimanovich, The Poisson formula for groups with hyperbolic properties, *Ann. Math.,* **152** (2000), 659–692.

[**KV**] V.A. Kaimanovich and A. M. Verschik, Random walks on discrete groups: boundary and entropy, *Annals Prob.* **11** (1983), 457–490.

[**KL**] A. Karlsson and F. Ledrappier, Drift and entropy for random walks, *Pure Appl. Math. Quarterly,* **3** (2007), 1027–1036.

[**Ki**] J.F.C. Kingman, The Ergodic Theory of Subadditive Processes, *J. Royal Stat. Soc., Ser. B,* **30** (1968), 499–510.

[**L1**] F. Ledrappier, Some Asymptotic properties of random walks on free groups, *in* Topics in Probability and Lie Groups, J.C. Taylor, ed. *CRM Proceedings and Lecture Notes,* **28** (2001), 117–152.

[**L2**] F. Ledrappier, Analyticity of the entropy for some random walks, *Groups, Geometry and Dynamics*, **6** (2012), 317–333.

[**L3**] F. Ledrappier, Regularity of the entropy for random walks on hyperbolic groups, *Annals Prob.*, to appear.

[**MM1**] J. Mairesse and F. Mathéus, Random walks on free products of cyclic groups, *J. London Math. Soc.,* **75** (2007), 47–66.

[**MM2**] J. Mairesse and F. Mathéus, Randomly growing braid on three strands and the manta ray, *Ann. Applied Proba.,* **17** (2007), 502–536.

[**Ma1**] P. Mathieu, Carne-Varopoulos bounds for centered random walks, *Annals Prob.*, **34** (2006), 987–1011.

[**Ma2**] P. Mathieu, Differentiating the entropy of random walks on hyperbolic groups, *preprint.*

[**R**] D. Ruelle, A review of linear response fo rgeneral differentiable dynamical systems, *Nonlinearity,* **22** (2009), 855–870.

[**Va**] N. Th. Varopoulos, Long range estimates for Markov chains, *Bull. Sci. Math.*, **100** (1985), 225–252.

[**V**] A. M. Verschik, Dynamic Theory of growth in groups: entropy, boundaries examples. *Russ. Math. Surveys* **55** (2000), 667–733.

[**WW**] C. Webster and A. Winchester, Boundaries of hyperbolic metric spaces, *Pacific J. Math.,* **221**(2005), 147–158.

[**W**] W. Woess, Random walks on infinite graphs and groups, Cambridge Tracts in Mathematics, vol. 138, Cambridge University Press, Cambridge, 2000.

*E-mail address*: gilch@TUGraz.at

*E-mail address*: ledrappier.1@nd.edu

Institut für Mathematische Strukturtheorie, Graz University of Technology, Steyrergasse 30/III, 8010 Graz, Austria

Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556, USA

# EXACTNESS, K-PROPERTY AND INFINITE MIXING

MARCO LENCI

Abstract. We explore the consequences of exactness or K-mixing on the notions of mixing (a.k.a. *infinite-volume mixing*) recently devised by the author for infinite-measure-preserving dynamical systems.

Mathematics Subject Classification (2010): 37A40, 37A25.

## 1. Introduction

Currently, in infinite ergodic theory, there is a renewed interest in the issues related to mixing for infinite-measure-preserving (or just nonsingular) dynamical systems, in short *infinite mixing* (see [Z, DS, L1, I3, DR, MT, LP, A2, Ko, T1], and some applications in [I1, I2, AMPS, L2, T2]).

The present author recently introduced some new notions of infinite mixing, based on the concept of *global observable* and *infinite-volume average* [L1]. In essence, a global observable for an infinite, $\sigma$-finite, measure space $(X, \mathscr{A}, \mu)$ is function in $L^\infty(X, \mathscr{A}, \mu)$ that "looks qualitatively the same" all over $X$. This is in contrast with a *local observable*, whose support is essentially localized, so that the function is integrable.

Postponing the mathematical details to Section 2, the purpose of the global observables is basically twofold. First, the past attempts to a general definition of infinite mixing involved mainly local observables (equivalently, finite-measure sets), and the problems with such definitions seemed to depend on that. Second, seeking inspiration in statistical mechanics (which is the discipline of mathematical physics that has successfully dealt with the question of predicting measurements in very large, formally infinite, systems), one realizes that many quantities of interest are *extensive observables*, that is, objects that behave qualitatively in the same way in different regions of the phase space. (More detailed discussions about these points are found in in [L1, L2].)

Extensive observables are "measured" by taking averages over large portions of the phase space. We import that concept too, by defining the infinite-volume average of a global observable $F : X \longrightarrow \mathbb{R}$ as

$$\overline{\mu}(F) := \lim_{V \nearrow X} \frac{1}{\mu(V)} \int_V F \, d\mu. \tag{1.1}$$

Here $V$ is taken from a family of ever larger but finite-measure sets that somehow covers, or *exhausts* the whole of $X$. The precise meaning of the limit above will be given in Section 2.

*Date*: Apr 4, 2013.

Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato 5, 40126 Bologna, Italy. E-mail: marco.lenci@unibo.it .

Istituto Nazionale di Fisica Nucleare, Sezione di Bologna, Via Irnerio 46, 40126 Bologna, Italy.

Now, let us consider a measure-preserving dynamical system on $(X, \mathscr{A}, \mu)$. For the sake of simplicity, let us restrict to the discrete-time case: this means that we have a measurable map $T : X \longrightarrow X$ that preserves $\mu$. Choosing two suitable classes of global and local observables, respectively denoted $\mathcal{G}$ and $\mathcal{L}$, we give five definitions of infinite mixing. These fall in two categories, exemplified as follows.

Using the customary (abuse of) notation $\mu(g) := \int_X g \, d\mu$. we say that the system exhibits:

- *global-local mixing* if, $\forall F \in \mathcal{G}$, $\forall g \in \mathcal{L}$, $\lim_{n\to\infty} \mu((F \circ T^n)g) = \overline{\mu}(F)\mu(g)$;
- *global-global mixing* if, $\forall F, G \in \mathcal{G}$, $\lim_{n\to\infty} \overline{\mu}((F \circ T^n)G) = \overline{\mu}(F)\,\overline{\mu}(G)$.

Disregarding for the moment the mathematical issues connected to the above notions, we focus on the interpretation of global-local mixing. Restricting, without loss of generality, to local observables $g \geq 0$ with $\mu(g) = 1$, and defining $d\mu_g := g \, d\mu$, the above limit reads:

$$\lim_{n\to\infty} T^n_* \mu_g(F) = \overline{\mu}(F), \tag{1.2}$$

where the measure $T^n_* \mu_g$ is the push-forward of $\mu_g$ via the dynamics $T^n$ (in other words, $T^n_* \mu_g := \mu_g \circ T^{-n} = \mu_{P^n g}$, where $P$ is the Perron-Frobenius operator relative to $\mu$, cf. (3.2)-(3.3)). If (1.2) occurs for all $g \in \mathcal{L}$ and $F \in \mathcal{G}$, the above is a sort of "convergence to equilibrium" for all *initial states* given by $\mu$-absolutely continuous probability measures. In this sense the functional $\overline{\mu}$ (not a measure!) plays the role of the *equilibrium state*.

Exactness and K-mixing (a.k.a. the K-property) are notions that exist and have the same definition both in finite and infinite ergodic theory. In finite ergodic theory they are known to be very strong properties, as they imply mixing of all orders, cf. definition (3.1). The purpose of this note is to explore their implications in terms of the notions of infinite mixing introduced in [L1].

As we will see below (Theorem 3.5*(a)*), the most notable of such implications is a weak form of global-local mixing, whereby any pair of measures $\mu_g$, $\mu_h$, as introduced earlier, are *asymptotically coalescing*, in the sense that

$$\lim_{n\to\infty} (T^n_* \mu_g(F) - T^n_* \mu_h(F)) = 0, \tag{1.3}$$

for all $F \in \mathcal{G}$.

In the next section we review the five definitions of global-local and global-global mixing, together with the already known (though with a different name) definition of local-local mixing. In Section 3 we prepare, state and prove Theorem 3.5, which lists some consequences of exactness and the K-property. Finally, in Section 4, we introduce the space of the *equilibrium observables*, which is a purely ergodic-theoretical construct in which some information about global-local mixing can be recast.

## 2. Definitions of infinite mixing

Let $(X, \mathscr{A}, \mu, T)$ be a measure-preserving dynamical system, where $(X, \mathscr{A})$ is a measure space, $\mu$ an infinite, $\sigma$-finite, measure on it, and $T$ a $\mu$-endomorphism, that is, a measurable surjective map that preserves $\mu$ (i.e., $\mu(T^{-1}A) = \mu(A)$, $\forall A \in \mathscr{A}$).

Denoting by $\mathscr{A}_f := \{A \in \mathscr{A} \mid \mu(A) < \infty\}$ the class of finite-measure sets, we assume that the following additional structure is given for the dynamical system:

- A class of sets $\mathscr{V} \subset \mathscr{A}_f$, called the **exhaustive family**. The elements of $\mathscr{V}$ will be generally indicated with the letter $V$.

- A subspace $\mathcal{G} \subset L^\infty(X, \mathscr{A}, \mu; \mathbb{R})$, whose elements are called the **global observables**. These functions are indicated with uppercase Roman letters ($F, G$, etc.).
- A subspace $\mathcal{L} \subset L^1(X, \mathscr{A}, \mu; \mathbb{R})$ whose elements are called the **local observables**. These functions will be indicated with lowercase Roman letters ($f, g$, etc.).

A discussion on the role and the choice of $\mathscr{V}, \mathcal{G}, \mathcal{L}$ is given in [L1], together with the proofs of most assertions made in this section.

We assume that $\mathscr{V}$ contains at least one sequence $(V_j)_{j\in\mathbb{N}}$, ordered by inclusion, such that $\bigcup_j V_j = X$. (In actuality, this requirement is never used in the proofs, but, since the elements of $\mathscr{V}$ are regarded as large and "representative" regions of the phase space $X$, we keep it to give "physical" meaning to the concept of infinite-volume average, see below.) We also assume that $1 \in \mathcal{G}$ (with the obvious notation $1(x) := 1$, $\forall x \in X$).

**Definition 2.1.** *Let $\mathscr{V}$ be the aforementioned exhaustive family. For $\phi : \mathscr{V} \longrightarrow \mathbb{R}$, we write*

$$\lim_{V \nearrow X} \phi(V) = \ell$$

*when*

$$\lim_{M\to\infty} \sup_{\substack{V\in\mathscr{V} \\ \mu(V)\geq M}} |\phi(V) - \ell| = 0.$$

*We call this the '$\mu$-uniform infinite-volume limit w.r.t. the family $\mathscr{V}$', or, for short, the **infinite-volume limit**.*

We assume that, $\forall n \in \mathbb{N}$,

$$\mu(T^{-n}V \triangle V) = o(\mu(V)), \quad \text{as } V \nearrow X. \tag{2.1}$$

This is reasonable because, if a large $V \in \mathscr{V}$ is to be considered a finite-measure substitute for $X$, it makes sense to require that a finite-time application of the dynamics does not change it much. Finally, the most crucial assumption is that,

$$\forall F \in \mathcal{G}, \quad \exists\, \overline{\mu}(F) := \lim_{V \nearrow X} \frac{1}{\mu(V)} \int_V F \, d\mu. \tag{2.2}$$

$\overline{\mu}(F)$ is called the **infinite-volume average** of $F$ w.r.t. $\mu$. It easy to check that $\overline{\mu}$ is $T$-invariant, i.e., for all $F \in \mathcal{G}$ and $n \in \mathbb{N}$, $\overline{\mu}(F \circ T^n)$ exists and equals $\overline{\mu}(F)$ [L1].

With this machinery, we can give a number of definitions of infinite mixing for the dynamical system $(X, \mathscr{A}, \mu, T)$ endowed with the *structure of observables* $(\mathscr{V}, \mathcal{G}, \mathcal{L})$.

The following three definitions will be called **global-local mixing**, as they involve the coupling of a global and a local observable. We say that the system is mixing of type

**(GLM1):** if, $\forall F \in \mathcal{G}$, $\forall g \in \mathcal{L}$ with $\mu(g) = 0$, $\lim_{n\to\infty} \mu((F \circ T^n)g) = 0$;

**(GLM2):** if, $\forall F \in \mathcal{G}$, $\forall g \in \mathcal{L}$, $\lim_{n\to\infty} \mu((F \circ T^n)g) = \overline{\mu}(F)\mu(g)$;

**(GLM3):** if, $\forall F \in \mathcal{G}$, $\lim_{n\to\infty} \sup_{g\in\mathcal{L}\setminus 0} \|g\|_1^{-1} \, |\mu((F \circ T^n)g) - \overline{\mu}(F)\mu(g)| = 0$,

where $\|\cdot\|_1$ is the norm of $L^1(X, \mathscr{A}, \mu; \mathbb{R})$.

Clearly, **(GLM1–3)** are listed in increasing order of strength, with **(GLM2)** being possibly the most natural definition one can give for the time-decorrelation between a global and a local observable (recall that $\overline{\mu}(F \circ T^n) = \overline{\mu}(F)$). **(GLM3)** is a uniform

version of it, with important implications (cf. Proposition 2.4), while **(GLM1)** is a much weaker version, as will become apparent in the remainder.

Although this note is mostly concerned with global-local mixing, one can also consider the decorrelation of two global observables, namely **global-global mixing**. For this we need the following terminology:

**Definition 2.2.** *For $\mathscr{V}$ as defined above and $\phi : \mathscr{V} \times \mathbb{N} \longrightarrow \mathbb{R}$, we write*

$$\lim_{\substack{V \nearrow X \\ n \to \infty}} \phi(V, n) = \ell$$

*to mean*

$$\lim_{M \to \infty} \sup_{\substack{V \in \mathscr{V} \\ \mu(V) \geq M \\ n \geq M}} |\phi(V, n) - \ell| = 0.$$

*As $n$ will take the role of time, we refer to this limit as the 'joint infinite-volume and time limit'.*

For $F \in L^\infty$ and $V \in \mathscr{V}$, let us also denote $\mu_V(F) := \mu(V)^{-1} \int_V F d\mu$. We say that the system is mixing of type

**(GGM1):** if, $\forall F, G \in \mathcal{G}$, $\lim_{n \to \infty} \overline{\mu}((F \circ T^n)G) = \overline{\mu}(F)\,\overline{\mu}(G)$;

**(GGM2):** if, $\forall F, G \in \mathcal{G}$, $\lim_{\substack{V \nearrow X \\ n \to \infty}} \mu_V((F \circ T^n)G) = \overline{\mu}(F)\,\overline{\mu}(G)$.

Though **(GGM1)** seems the cleaner of the two versions, it has the serious drawback that, for $n \in \mathbb{N}$, $\overline{\mu}((F \circ T^n)G)$ might not even exist, for there is no provision in our hypotheses to guarantee the ring property for condition (2.2) (namely, $\exists \overline{\mu}(F), \overline{\mu}(G) \Rightarrow \exists \overline{\mu}(FG)$). Nor do we want one, if we are to keep our framework general enough. **(GGM2)** solves this question of wellposedness, and is in some sense stronger than **(GGM1)**:

**Proposition 2.3.** *If $F, G \in \mathcal{G}$ are such that $\overline{\mu}((F \circ T^n)G)$ exists for all $n$ large enough (depending on $F, G$), then*

$$\lim_{\substack{V \nearrow X \\ n \to \infty}} \mu_V((F \circ T^n)G) = \ell \quad \Longrightarrow \quad \lim_{n \to \infty} \overline{\mu}((F \circ T^n)G) = \ell. \tag{2.3}$$

*In particular, if the above hypothesis holds $\forall F, G \in \mathcal{G}$, then* **(GGM2)** *implies* **(GGM1)**.

*Proof.* From Definition 2.2, the left limit of (2.3) implies that, $\forall \varepsilon > 0$, $\exists M = M(\varepsilon)$ such that

$$\ell - \varepsilon \leq \mu_V((F \circ T^n)G) \leq \ell + \varepsilon \tag{2.4}$$

for all $V \in \mathscr{V}$ with $\mu(V) \geq M$ and all $n \geq M$. By hypothesis, if $M$ is large enough, the infinite-volume limit of the above middle term exists $\forall n \geq M$ and equals $\overline{\mu}((F \circ T^n)G)$. Upon taking such limit, what is left of (2.4) and its conditions of validity is the very definition of the right limit in (2.3). □

With reasonable hypotheses on the structure of $\mathcal{G}$ and $\mathcal{L}$, the strongest version of global-local mixing implies the "strongest" version of global-global mixing. The following proposition is a simplified version of a similar result of [L1] (for an intuitive understanding of the hypotheses, see Proposition 3.2 and Remark 3.3 there).

**Proposition 2.4.** *Suppose there exist a family $(\psi_j)_{j \in \mathbb{N}}$ of real-valued functions of $X$ (this will play the role of a partition of unity) and a family $(\mathbb{J}_V)_{V \in \mathscr{V}}$ of finite subsets of $\mathbb{N}$ such that:*

*(i)* $\forall j \in \mathbb{N}$, $\psi_j \geq 0$;
*(ii)* $\forall G \in \mathcal{G}$, $\forall j \in \mathbb{N}$, $G\psi_j \in \mathcal{L}$;
*(iii) in the limit* $V \nearrow X$, $\left\| \sum_{j \in \mathbb{J}_V} \psi_j - 1_V \right\|_1 = o(\mu(V))$,

*where* $1_V$ *is the indicator function of* $V$. *Then* **(GLM3)** *implies* **(GGM2)**.

Proof of Proposition 2.4. Since the limit in **(GGM2)** is trivial when $G$ is a constant, and since the global observables are bounded functions, it is no loss of generality to prove **(GGM2)** for the case $G \geq 0$ only.

The proof follows upon verification that the functions $g_j := G\psi_j$ verify all the hypotheses of Proposition 3.2 of [L1] (cf. also Remark 3.3). Notice that the identity $G = \sum_j g_j$ (which makes sense insofar as $(\psi_j)_j$ is a partition of unity) is illustrative and not really used in the proof there. □

Since the five definitions presented above deal with the decorrelation of, first, a global and a local observable, and then two global observables, symmetry considerations would induce one to give a definition of **local-local mixing** as well. A reasonable possibility would be to call a dynamical system mixing of type

**(LLM):** if, $\forall f \in \mathcal{L} \cap \mathcal{G}$, $g \in \mathcal{L}$, $\lim_{n\to\infty} \mu((f \circ T^n)g) = 0$.

In fact, this definition already exists, as it is easy to check that, in the most general case (that is, $\mathcal{G} = L^\infty$, $\mathcal{L} = L^1$), a dynamical system is **(LLM)** if and only if, $\forall A, B \in \mathscr{A}_f$, $\lim_{n\to\infty} \mu(T^{-n}A \cap B) = 0$, i.e., if and only if the system is of *zero type* [HK] (cf. also [DS, Ko]). Incidentally, this is the same definition that Krengel and Sucheston call 'mixing', for an infinite-measure-preserving dynamical system [KS].

## 3. Exactness and K-property

Two of the few definitions that are copied verbatim from finite to infinite ergodic theory are those of exactness and K-mixing. Though they are well known, we repeat them here for completeness. We state the versions for measure-preserving maps, but they can be given for nonsingular maps as well ($T$ is nonsingular if $\mu(A) = 0 \Rightarrow \mu(T^{-1}A) = 0$).

Let us denote by $\mathscr{N}$ the *null* $\sigma$*-algebra*, i.e., the $\sigma$-algebra that only contains the zero-measure sets and their complements. Also, given two $\sigma$-algebras $\mathscr{A}, \mathscr{B}$, we write $\mathscr{A} = \mathscr{B} \bmod \mu$ if $\forall A \in \mathscr{A}$, $\exists B \in \mathscr{B}$ with $\mu(A \triangle B) = 0$, and viceversa; equivalently, the $\mu$-completions of $\mathscr{A}$ and $\mathscr{B}$ are the same.

**Definition 3.1.** *The measure-preserving dynamical system* $(X, \mathscr{A}, \mu, T)$ *is called* ***exact*** *if*

$$\bigcap_{n=0}^{\infty} T^{-n}\mathscr{A} = \mathscr{N} \bmod \mu.$$

Since exactness implies that $T^{-1}\mathscr{A} \neq \mathscr{A} \bmod \mu$, a nontrivial exact $T$ cannot be an automorphism of the measure space $(X, \mathscr{A}, \mu)$—although in some sense an invertible map can still be exact, cf. Remark 3.3 below.

The counterpart of exactness for automorphisms is the following:

**Definition 3.2.** *The invertible measure-preserving dynamical system* $(X, \mathscr{A}, \mu, T)$ *possesses the* ***K-property*** *(from A. N. Kolmogorov) if* $\exists \mathscr{B} \subset \mathscr{A}$ *such that:*

(i) $\mathscr{B} \subset T\mathscr{B}$;

(ii) $\bigvee_{n=0}^{\infty} T^n\mathscr{B} = \mathscr{A} \text{ mod } \mu$;

(iii) $\bigcap_{n=0}^{\infty} T^{-n}\mathscr{B} = \mathscr{N} \text{ mod } \mu$.

*In this case, one also says that the dynamical system is K-mixing, or that $T$ is a K-automorphism of $(X, \mathscr{A}, \mu)$.*

**Remark 3.3.** Comparing Definition 3.1 with condition *(iii)* of Definition 3.2, one might be tempted to say that, if $(X, \mathscr{A}, \mu, T)$ has the K-property, then $(X, \mathscr{B}, \mu, T)$ is exact. This is not *technically* correct because, in all nontrivial cases, the inclusion in Definition 3.2*(i)* is strict, thus $T$ is not a self-map of the measure space $(X, \mathscr{B}, \mu)$. That said, if $(X, \mathscr{A}, \mu)$ is a Lebesgue space, $(X, \mathscr{B}, \mu, T)$ is still *morally* exact, in the following sense. Assume w.l.g. that $\mathscr{B}$ is complete, let $X_{\mathscr{B}}$ be the measurable partition of $X$ that generates $\mathscr{B}$. (In a Lebesgue space there is a one-to-one correspondence, modulo null sets, between complete sub-$\sigma$-algebras and measurable partitions [R].) $\mathscr{B}$ can be lifted to a $\sigma$-algebra for $X_{\mathscr{B}}$, which we keep calling $\mathscr{B}$. Also, defining $T_{\mathscr{B}}([x]) := [T(x)]$ (where $[x]$ denotes the element of $X_{\mathscr{B}}$ that contains $x$), we verify that $T_{\mathscr{B}}$ is well defined as a self-map of $(X_{\mathscr{B}}, \mathscr{B}, \mu)$ (in fact, from Definition 3.2*(i)*, $X_{T\mathscr{B}}$ is a sub-partition of $X_{\mathscr{B}}$) and $T_{\mathscr{B}}^{-1}A = T^{-1}A$, $\forall A \in \mathscr{B}$ (with the understandable abuse of notation whereby $A$ denotes both a subset of $X_{\mathscr{B}}$ and a subset of $X$). This and Definition 3.2*(iii)* show that $T_{\mathscr{B}}$ is an exact endomorphism of $(X_{\mathscr{B}}, \mathscr{B}, \mu)$. Of course, in all of the above, $\mathscr{B}$ can be replaced by $\mathscr{B}_m := T^m\mathscr{B}$, for all $m \in \mathbb{Z}$ (because $\mathscr{B}_m$ can be used in lieu of $\mathscr{B}$ in Definition 3.2).

In finite ergodic theory, both exactness and the K-property imply *mixing of all orders*, namely, $\forall k \in \mathbb{Z}^+$ and $A_1, A_2, \ldots, A_k \in \mathscr{A}$,

$$\mu(A_1 \cap T^{-n_2}A_2 \cap \cdots T^{-n_k}A_k) \longrightarrow \mu(A_1)\mu(A_2)\cdots\mu(A_k), \tag{3.1}$$

whenever $n_2 \to \infty$ and $n_{i+1} - n_1 \to \infty$, $\forall i = 2, \ldots k-1$ [Q]. (In (3.1) we have assumed $\mu(X) = 1$.)

One would expect such strong properties to have consequences also in infinite ergodic theory. This is the case, as we describe momentarily. But first we need some elementary formalism from the functional analysis of dynamical systems. For $F \in L^\infty$ and $g \in L^1$, let us denote

$$\langle F, g\rangle := \mu(Fg). \tag{3.2}$$

Define the *Koopman operator* $U : L^\infty \longrightarrow L^\infty$ as $UF := F \circ T$. Its adjoint for the above coupling is called the *Perron-Frobenius operator*, denoted $P : L^1 \longrightarrow L^1$. Its defining identity is

$$\langle UF, g\rangle = \langle F, Pg\rangle. \tag{3.3}$$

Let us explain in detail how $P$ is defined through (3.3). Take $g \in L^1$ and assume for the moment $g \geq 0$. Take also $F = 1_A$, with $A \in \mathscr{A}$. We see that $\langle UF, g\rangle = \int_{T^{-1}A} g\, d\mu$. Since $T$ preserves $\mu$ and is thus nonsingular w.r.t. it, and since the measure space is $\sigma$-finite, the Radon-Nykodim Theorem yields a locally-$L^1$, positive, function $Pg : X \longrightarrow \mathbb{R}$ such that $\int_{T^{-1}A} g\, d\mu = \int_A (Pg) d\mu = \langle F, Pg\rangle$. Using $F = 1_X = 1$, we see that $Pg \in L^1$ with $\|Pg\|_1 = \|g\|_1$. For a general $g \in L^1$, we write $g = g^+ - g^-$, where $g^+$ and $g^-$ are, respectively, the positive and negative parts of $g$. Then $Pg := Pg^+ - Pg^-$ is also in $L^1$

and

$$\|Pg\|_1 \leq \|g\|_1. \tag{3.4}$$

Therefore, through approximations of $F$ via simple functions (on finite-measure sets and in the $L^\infty$-norm), one can extend (3.3) to all $F \in L^\infty$.

In the process, we have learned that $P$ is a positive operator ($g \geq 0 \Rightarrow Pg \geq 0$) and $\|P\| = 1$, whereas, obviously, $U$ is a positive isometry. Moreover, it is easy to see that $Pg = g$, with $g \geq 0$, if and only if $g$ is an invariant density, i.e., if $\mu_g$ defined by $d\mu_g/d\mu = g$ is an invariant measure. (In fact, had we defined (3.2) for $F \in L^1$ and $g \in L^\infty$, (3.3) would have defined a positive operator $P : L^\infty \longrightarrow L^\infty$, with $\|P\| = 1$, and such that $P1 = 1$.)

Most of the remainder of this note will be based on an important theorem by Lin [Li] (see also [A1] for a nice short proof).

**Theorem 3.4.** *The nonsingular dynamical system $(X, \mathscr{A}, \mu, T)$ is exact if and only if, $\forall g \in L^1$ with $\mu(g) = 0$, $\lim_{n\to\infty} \|P^n g\|_1 = 0$.*

In the rest of the paper we assume to be in one of the following two cases:

(H1) $(X, \mathscr{A}, \mu, T)$ is exact. $\mathscr{V}$ is any exhaustive family that verifies (2.1). $\mathcal{G} = L^\infty$. $\mathcal{L} = L^1$. (Given the assumptions of Section 2, this corresponds to the most general choice of $\mathscr{V}, \mathcal{G}, \mathcal{L}$.)

(H2) $(X, \mathscr{A}, \mu, T)$ is K-mixing (thus $T$ is an automorphism). $\mathscr{V}$ is any exhaustive family that verifies (2.1). $\mathcal{G}$ is the closure, in $L^\infty$, of $\bigcup_{m>0} L^\infty(\mathscr{B}_m)$, where $\mathscr{B}_m = T^m\mathscr{B}$, as defined in Remark 3.3. Lastly, $\mathcal{L} = L^1$.

**Theorem 3.5.** *Under either* (H1) *or* (H2),

(a) **(GLM1)** *holds true;*

(b) **(LLM)** *holds true;*

(c) **(GGM2)** *implies* **(GLM2)**;

(d) *If, $\forall F \in \mathcal{G}$, $\exists g_F \in \mathcal{L}$, with $\mu(g_F) \neq 0$, such that*

$$\lim_{n\to\infty} \mu((F \circ T^n)g_F) = \overline{\mu}(F)\mu(g_F),$$

*then* **(GLM2)** *holds true.*

As anticipated in the introduction, **(GLM1)** (which is the most important assertion of the theorem) means that the evolutions of two absolutely continuous initial measures become indistinguishable, as time goes to infinity. We may call this phenomenon *asymptotic coalescence*. This implies that they will return the same measurements of global observables, but not that this measurements will converge (in which case we would have a sort of convergence to equilibrium). In fact, for many interesting systems, it is not hard to construct $F \in L^\infty$ such that $\langle F, P^n g\rangle$ does not converge for all $g \in L^1$.

This is not surprising, for, even in finite ergodic theory, certain proofs of mixing, or decay of correlation, are divided in two parts: asymptotic coalescence and the convergence of *one* initial measure. The difference there is that the latter is usually easy.

The remainder of this section is devoted to the following:

Proof of Theorem 3.5. Let us start by proving assertion *(a)*, namely **(GLM1)**. We use the formalism of functional analysis outlined earlier in the section.

If (H1) is the case, the proof is immediate: for $F \in L^\infty$ and $g \in L^1$, with $\mu(g) = 0$,

$$|\mu((F \circ T^n)g)| = |\langle F, P^n g\rangle| \leq \|F\|_\infty \|P^n g\|_1 \to 0, \tag{3.5}$$

as $n \to \infty$, by Theorem 3.4.

In the case (H2), let us observe that, by easy density arguments, all the definitions **(GLM1–3)** hold true if they are verified w.r.t. $\mathcal{G}'$ and $\mathcal{L}'$ which are subspaces of $\mathcal{G}$ and $\mathcal{L}$, respectively, in the $L^\infty$- and $L^1$-norms. We can take $\mathcal{G}' := \bigcup_{m>0} L^\infty(\mathscr{B}_m)$ (which is dense in $\mathcal{G}$ by definition) and $\mathcal{L}' := \bigcup_{m>0} L^1(\mathscr{B}_m)$, which is dense in $\mathcal{L} = L^1(\mathscr{A})$ by the K-property [A1]. Therefore, it suffices to show **(GLM1)** for a general $m > 0$ and $\forall F \in L^\infty(\mathscr{B}_m)$, $\forall g \in L^1(\mathscr{B}_m)$ with $\mu(g) = 0$.

Using the arguments and the notation of Remark 3.3, we denote by $\hat{F}$ the function induced by $F$ on $X_{\mathscr{B}_m}$ (i.e., $\hat{F}([x]) := F(x)$), and analogously for all the other $\mathscr{B}_m$-measurable functions. We observe that $F \circ T^n$ is $\mathscr{B}_m$-measurable and $\widehat{F \circ T^n} = \hat{F} \circ T^n_{\mathscr{B}_m}$. Thus

$$\mu((F \circ T^n)g) = \mu((\hat{F} \circ T^n_{\mathscr{B}_m})\hat{g}), \tag{3.6}$$

where the r.h.s. is regarded as an integral in $X_{\mathscr{B}_m}$. Since $(X_{\mathscr{B}_m}, \mathscr{B}_m, \mu, T_{\mathscr{B}_m})$ is exact, and $\mu(\hat{g}) = \mu(g) = 0$, we use (3.6) in (3.5) to prove that the l.h.s. of (3.6) vanishes, as $n \to \infty$.

The following is a corollary of **(GLM1)**.

**Lemma 3.6.** *Assume either* (H1) *or* (H2)*, and fix $F \in \mathcal{G}$. If, for some $\ell \in \mathbb{R}$ and $\varepsilon \geq 0$, the limit*

$$\limsup_{n\to\infty} \left| \frac{\mu((F \circ T^n)g)}{\mu(g)} - \ell \right| \leq \varepsilon$$

*holds for some $g \in \mathcal{L}$ (with $\mu(g) \neq 0$), then it holds for all $g \in \mathcal{L}$ (with $\mu(g) \neq 0$).*

Proof of Lemma 3.6. Suppose the above limit holds for $g_0 \in \mathcal{L}$. Take any other $g \in \mathcal{L}$, with $\mu(g) \neq 0$. We have:

$$\begin{aligned} &\left| \frac{\mu((F \circ T^n)g)}{\mu(g)} - \ell \right| \\ &\leq \left| \mu\left( (F \circ T^n) \left( \frac{g}{\mu(g)} - \frac{g_0}{\mu(g_0)} \right) \right) \right| + \left| \frac{\mu((F \circ T^n)g_0)}{\mu(g_0)} - \ell \right|. \end{aligned} \tag{3.7}$$

By **(GLM1)**, the first term of the above r.h.s. vanishes as $n \to \infty$, whence the assertion. □

Going back to the proof of Theorem 3.5, we see that Lemma 3.6 immediately implies assertion *(d)*.

As for *(b)*, again we prove it for both cases (H1) and (H2) at the same time. W.l.g., let us assume that $\mathscr{A} \neq \mathscr{N} \bmod \mu$ (otherwise $L^1$ would be trivial). We claim that

$$\sup_{A \in \mathscr{A}_f} \mu(A) = \infty. \tag{3.8}$$

In fact, since $\mathscr{A}$ is not trivial, the above sup is positive. If it equalled $M \in \mathbb{R}^+$, it would be easy to construct an invariant set $B$ with $0 < \mu(B) \leq M$. But $\mu(X) = \infty$, therefore $T$ would not be ergodic, contradicting both (H1) and (H2).

Now take $f \in L^1 \cap \mathcal{G}$ and $\varepsilon > 0$. By (3.8), $\exists A \in \mathscr{A}_f$ with $\mu(A) \geq \|f\|_1/\varepsilon$. Set $g_\varepsilon = 1_A/\mu(A)$. We have that

$$\left| \frac{\mu((f \circ T^n)g_\varepsilon)}{\mu(g_\varepsilon)} \right| = |\mu((f \circ T^n)g_\varepsilon)| \leq \|f\|_1 \, \|g_\varepsilon\|_\infty \leq \varepsilon. \tag{3.9}$$

By Lemma 3.6,

$$\limsup_{n\to\infty}\left|\frac{\mu((f\circ T^n)g)}{\mu(g)}\right|\leq\varepsilon \tag{3.10}$$

holds for *all* $g\in\mathcal{L}$ with $\mu(g)\neq 0$. Since $\varepsilon$ is arbitrary, we get that the above r.h.s. is zero. The case $\mu(g)=0$ is trivial because the same assertion comes directly from **(GLM1)**. This proves **(LLM)**, namely, assertion *(b)*.

Finally for *(c)*. Take a $G\in\mathcal{G}$ such that $\overline{\mu}(G)>0$. Since $\mu_V(G)\to\overline{\mu}(G)$, as $V\nearrow X$, **(GGM2)** implies that there exist a large enough $M$ and a $V\in\mathscr{V}$, with $\mu(V)\geq M$, such that

$$|\mu_V((F\circ T^n)G)-\overline{\mu}(F)\overline{\mu}(G)|\leq\varepsilon\mu_V(G) \tag{3.11}$$

for all $n\geq M$. Setting $g:=G1_V$, we can divide (3.11) by $\mu_V(G)=\mu(g)$ and take the lim sup in $n$:

$$\limsup_{n\to\infty}\left|\frac{\mu((F\circ T^n)g)}{\mu(g)}-\frac{\overline{\mu}(G)}{\mu_V(G)}\,\overline{\mu}(F)\right|\leq\varepsilon. \tag{3.12}$$

By Lemma 3.6, the above holds $\forall g\in\mathcal{L}$, with $\mu(g)\neq 0$. Since $\varepsilon$ can be taken arbitrarily close to 0 and $\overline{\mu}(G)/\mu_V(G)$ arbitrarily close to 1, we have that, for all $F\in\mathcal{G}$ and $g\in\mathcal{L}$, with $\mu(g)\neq 0$,

$$\lim_{n\to\infty}\mu((F\circ T^n)g)=\overline{\mu}(F)\mu(g). \tag{3.13}$$

The corresponding statement for $\mu(g)=0$ comes from **(GLM1)**. □

## 4. The equilibrium observables

The "pure" ergodic theorist might raise an eyebrow at the constructions of Section 2, especially at the ideas of the exhaustive family (which demands that one singles out some sets as more important than the others) and of the infinite-volume average (which is not a measure, or even guaranteed to always exist).

Though these issues (and more) have been addressed in [L1], one might still want to see if some of the concepts presented here can be viewed from the vantage point of traditional infinite ergodic theory. For what follows I am indebted to R. Zweimüller.

As we discussed in the introduction, the definition **(GLM2)** makes sense as a kind of convergence to equilibrium for a large class of initial distributions (see also the observation on **(GLM1)** after the statement of Theorem 3.5). Without worrying too much about predetermining good test functions for this convergence (namely, the global observables), and the value of any such limit (namely, the infinite-volume average), one might simply consider the space $\mathcal{E}=\mathcal{E}(X,\mathscr{A},\mu,T)$ of *all* the good test functions, in this sense:

$$\mathcal{E}:=\left\{F\in L^\infty\;\middle|\;\exists\rho(F)\in\mathbb{R}\text{ s.t. }\lim_{n\to\infty}\mu((F\circ T^n)g)=\rho(F)\mu(g),\,\forall g\in L^1\right\}. \tag{4.1}$$

(Occasionally, one might want to restrict the space of the initial distributions to some subspace of $L^1$.) Clearly, $\mathcal{E}$ is a vector space which contains at least the constant functions.

$\rho(F)$ represents a sort of *value at equilibrium* of $F$ and, in this context, it need not have anything to do with $\overline{\mu}(F)$ (which might or might not exist), $\mathscr{V}$, or the choice of $\mathcal{G}$ and $\mathcal{L}$. Thus, the elements of the vector space $\mathcal{E}$ may be called the **equilibrium observables** and $\rho:\mathcal{E}\longrightarrow\mathbb{R}$ the **equilibrium functional**.

If we are in either case (H1) or (H2), Theorem 3.5*(d)* shows that, for a given $F \in \mathcal{G}$, one only need find *one* local observable that verifies the limit in (4.1). Also, by Theorem 3.5*(b)*, any $f \in \mathcal{G} \cap L^1$ belongs to $\mathcal{E}$, with $\rho(f) = 0$. Therefore, in these cases, it makes sense to introduce $\hat{\mathcal{E}} := \mathcal{E}/(\mathcal{G} \cap L^1)$, and $\rho$ is well defined there. When talking about $\hat{\mathcal{E}}$, we write $F \in \hat{\mathcal{E}}$ to mean $F \in \mathcal{E}$, and $F = G$ to mean $[F] = [G]$ (where $[\cdot]$ denotes an equivalence class in $\mathcal{E}/(\mathcal{G} \cap L^1)$).

Determining $\hat{\mathcal{E}}$ for a given, say, exact dynamical system appears to be as complicated as proving **(GLM2)** for a truly large class of global observables, though occasionally some information can be obtained quickly. We conclude this note by giving some examples thereof.

*Boole transformation.* This is the transformation $T : \mathbb{R} \longrightarrow \mathbb{R}$ defined by $T(x) := x - 1/x$. This map preserves the Lebesgue measure on $\mathbb{R}$, as it is easy to verify, and is exact [A1]. We can use the fact that $T$ is odd to construct a nonconstant equilibrium observable. Set $F(x) := \text{sign}(x)$, and $g := 1_{[-1,1]}$. Clearly, for all $n \in \mathbb{N}$, $F \circ T^n$ is odd and $\mu((F \circ T^n)g) = 0$, so $F \in \hat{\mathcal{E}}$, with $F \neq$ constant, and $\rho(F) = 0$.

Evidently, the same reasoning can be applied to any exact map with an odd symmetry.

*Translation-invariant expanding maps of* $\mathbb{R}$. Take a $C^2$ bijection $\Phi : [0,1] \longrightarrow [k_1, k_2]$, with $k_1, k_2 \in \mathbb{Z}$, and $\Phi' > 1$, where $\Phi'$ denotes the derivative of $\Phi$. (Notice that these conditions imply $\Phi(0) = k_1$, $\Phi(1) = k_2$, and $k := k_2 - k_1 \geq 2$.) Define $T : \mathbb{R} \longrightarrow \mathbb{R}$ via

$$T|_{[j,j+1)}(x) := \Phi(x - j) + j, \tag{4.2}$$

for all $j \in \mathbb{Z}$. By construction $T(x + 1) = T(x) + 1$, $\forall x \in \mathbb{R}$, and so $T$ is a $k$-to-1 translation-invariant map, in the sense that it commutes with the natural action of $\mathbb{Z}$ in $\mathbb{R}$.

Suppose that $T$ preserves the Lebesgue measure, which we denote $m_{\mathbb{R}}$. (One can easily construct a large class of maps of this kind.) It can be proved that any such $T$ is exact [L3]. Now, define $I := [0,1)$ and $T_I : I \longrightarrow I$ as $T_I(x) := T(x) \mod 1$. Clearly, $(I, \mathscr{B}_I, T_I, m_I)$, where $\mathscr{B}_I$ and $m_I$ are, respectively, the Borel $\sigma$-algebra and the Lebesgue measure on $I$, is a probability-preserving dynamical system. It is easy to see that it is exact, and thus mixing.

Now consider a $\mathbb{Z}$-periodic, bounded, $F : \mathbb{R} \longrightarrow \mathbb{R}$. Evidently, $\forall x \in I$, $\forall n \in \mathbb{N}$, $F \circ T^n(x) = F \circ T_I^n(x)$. Hence, by the mixing of the quotient dynamical system, for any square-integrable $g$ supported in $I$,

$$\begin{aligned} \lim_{n\to\infty} m_{\mathbb{R}}((F \circ T^n)g) &= \lim_{n\to\infty} m_I((F \circ T_I^n)g) \\ &= m_I(F)\, m_I(g) \\ &= m_I(F)\, m_{\mathbb{R}}(g). \end{aligned} \tag{4.3}$$

By the exactness of $T$, the above holds for all $g \in L^1(\mathbb{R})$. Hence $F \in \hat{\mathcal{E}}$, with $\rho(F) = m_I(F) = \overline{m_{\mathbb{R}}}(F)$.

An analogous procedure (using $I_j := [0, j)$ instead of $I$) can be employed to prove that any $(j\mathbb{Z})$-periodic, bounded $F$ belongs in $\hat{\mathcal{E}}$, with $\rho(F) = \overline{m_{\mathbb{R}}}(F)$. In [L3] we extend this result to observables that are quasi-periodic w.r.t. any $j\mathbb{Z}$, and more.

*Random walks.* A special case of the above situation occurs when $\Phi$ is linear. The result is a piecewise linear Markov map that represents a random walk in $\mathbb{Z}$, in the following sense. Denote by $\lfloor x \rfloor$ the maximum integer not exceeding $x \in \mathbb{R}$. If an initial condition

$x \in I$ is randomly chosen with law $m_I$, then the stochastic process $(\lfloor T^n(x) \rfloor)_{n\in\mathbb{N}}$ is precisely the random walk starting in $0 \in \mathbb{Z}$, with uniform transition probabilities for jumps of $k_1, k_1+1, \dots, k_2-1$ units [L2].

A reelaboration of a result of [L1] shows that $\hat{\mathcal{E}}$ contains all $L^\infty$ functions such that the limit

$$\rho(F) := \lim_{M\to\infty} \int_{a-M}^{a+M} F(x)\,dx \tag{4.4}$$

exists independently of and uniformly in $a \in \mathbb{R}$. In fact, it is proved in [L1, Thm. 4.6(b)] (see also [L2, Thm. 9]) that, if $g \in L^1$, $F \in L^\infty(\mathscr{A}_0)$, where $\mathscr{A}_0$ is the $\sigma$-algebra generated by the partition $\{[j, j+1)\}_j$, and the limit

$$\lim_{j\to\infty} \int_{q-j}^{q+j} F(x)\,dx =: \overline{m_{\mathbb{R}}}(F) \tag{4.5}$$

($j \in \mathbb{Z}$) exists uniformly in $q \in \mathbb{Z}$, then $m_{\mathbb{R}}((F\circ T^n)g) \to \overline{m_{\mathbb{R}}}(F) m_{\mathbb{R}}(g)$, as $n \to \infty$. Obviously, comparing (4.4) with (4.5), $\rho(F) = \overline{m_{\mathbb{R}}}(F)$.

Now, for a general $F$, one can take $g = 1_{[0,1)} \in L^1(\mathscr{A}_0)$. It is easy to check that $P^n g$ is $\mathscr{A}_0$-measurable too, thus

$$\begin{aligned} \lim_{n\to\infty} m_{\mathbb{R}}((F\circ T^n)g) &= \lim_{n\to\infty} \langle \mathbb{E}(F|\mathscr{A}_0), P^n g\rangle \\ &= \overline{m_{\mathbb{R}}}(\mathbb{E}(F|\mathscr{A}_0))\, m_{\mathbb{R}}(g) \\ &= \rho(F)\, m_{\mathbb{R}}(g), \end{aligned} \tag{4.6}$$

which proves our claim.

If the random walk has a drift, say a positive drift, then a.e. orbit will converge to $+\infty$. Therefore, any bounded function $G$ that asymptotically shadows any of the above observables—meaning $\lim_{x\to+\infty}(G(x)-F(x)) = 0$, for some $F$ verifying (4.4)—will also belong to $\hat{\mathcal{E}}$, with $\rho(G) = \rho(F)$.

## References


[A1] J. Aaronson, *An introduction to infinite ergodic theory*, Mathematical Surveys and Monographs, 50. American Mathematical Society, Providence, RI, 1997.

[A2] J. Aaronson, *Rational weak mixing in infinite measure spaces*, to appear in Ergodic Theory Dynam. Systems (2013).

[AMPS] A. Arbieto, R. Markarian, M. J. Pacifico and R. Soares, *Scaling rate for semi-dispersing billiards with non-compact cusps*, Ergodic Theory Dynam. Systems **32** (2012), no. 6, 1836–1861.

[DR] A. I. Danilenko and V. V. Ryzhikov, *Mixing constructions with infinite invariant measure and spectral multiplicities*, Ergodic Theory Dynam. Systems **31** (2011), no. 3, 853–873.

[DS] A. I. Danilenko and C. E. Silva, *Ergodic Theory: nonsingular transformation*, in: R. E. Meyers (ed.), Encyclopedia of Complexity and Systems Science, pp. 3055–3083, Springer, 2009.

[HK] A. B. Hajian and S. Kakutani, *Weakly wandering sets and invariant measures*, Trans. Amer. Math. Soc. **110** (1964), 136–151.

[I1] S. Isola, *Renewal sequences and intermittency*, J. Statist. Phys. **97** (1999), no. 1-2, 263–280.

[I2] S. Isola, *On systems with finite ergodic degree*, Far East J. Dyn. Syst. **5** (2003), no. 1, 1–62.

[I3] S. Isola, *From infinite ergodic theory to number theory (and possibly back)*, Chaos Solitons Fractals **44** (2011), no. 7, 467–479.

[Ko] Z. Kosloff, *The zero-type property and mixing of Bernoulli shifts*, Ergodic Theory Dynam. Systems **33** (2013), no. 2, 549-559.

[KS] U. Krengel and L. Sucheston, *Mixing in infinite measure spaces*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **13** (1969), 150–164.

[LP] M. Lemańczyk and F. Parreau, *Lifting mixing properties by Rokhlin cocycles*, Ergodic Theory Dynam. Systems **32** (2012), no. 2, 763–784.

[L1] M. Lenci, *On infinite-volume mixing*, Comm. Math. Phys. **298** (2010), no. 2, 485–514.
[L2] M. Lenci, *Infinite-volume mixing for dynamical systems preserving an infinite measure*, Procedia IUTAM **5** (2012), 204–219.
[L3] M. Lenci, *Infinite mixing for piecewise expanding maps of the real line*, in preparation.
[Li] M. Lin, *Mixing for Markov operators*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **19** (1971), 231–242.
[MT] I. Melbourne and D. Terhesiu, *Operator renewal theory and mixing rates for dynamical systems with infinite measure*, Invent. Math. **189** (2012), no. 1, 61–110.
[Q] A. Quas, *Ergodicity and mixing properties*, in: R. E. Meyers (ed.), Encyclopedia of Complexity and Systems Science, pp. 2918–2933, Springer, 2009.
[R] V. A. Rokhlin, *Lectures on the entropy theory of measure-preserving transformations*, Russ. Math. Surv. **22** (1967), no. 5, 1–52.
[T1] D. Terhesiu, *Improved mixing rates for infinite measure preserving systems*, preprint (2012).
[T2] D. Terhesiu, *Mixing rates for intermittent maps of high exponent*, preprint (2012).
[Z] R. Zweimüller, *Mixing limit theorems for ergodic transformations*, J. Theoret. Probab. **20** (2007), no. 4, 1059–1071.

# A SURVEY ON THE MINIMAL SETS OF LEFSCHETZ PERIODS FOR MORSE–SMALE DIFFEOMORPHISMS ON SOME CLOSED MANIFOLDS

JAUME LLIBRE[1] AND VÍCTOR F. SIRVENT[2]

*Dedicated to Jorge Lewowicz in occasion of his* Doctorado Honoris Causa *conferred by* Universidad de La República.

Abstract. We present the actual state of the study of the minimal sets of Lefschetz periods $\mathrm{MPer}_L(f)$ for the Morse–Smale diffeomorphisms on some closed manifolds, as the connected compact surfaces (orientable or not) without boundary, the $n$–dimensional torus and some other manifolds. The results on $\mathrm{MPer}_L(f)$ are valid for $C^1$ self–maps on the mentioned closed manifolds with finitely many periodic points all of them hyperbolic such that all the eigenvalues of the induced maps on homology are roots of unity. This class of maps includes the Morse–Smale diffeomorphisms.

## 1. Introduction

In the study of the discrete dynamical systems and, in particular in the study of the orbits of self–maps defined on a given compact manifold, the periodic orbits play an important role. These last forty years there was many results showing that some simple assumptions force qualitative and quantitative properties (like the set of periods) of a map. One of the first results in this direction was the famous paper *Period three implies chaos* for the interval continuous self–maps, see [24].

One of the most used tool for studying the existence of fixed points and periodic points, for continuous self maps on compact manifolds, and more generally topological spaces which are retract of finite simplicial complexes, is the Lefschetz fixed point theorem and its improvements (*cf.* [1, 2, 7, 8, 9, 11, 18, 19, 25, 30]). The Lefschetz zeta function $\zeta_f(t)$ simplifies the study of the periodic points of $f$. It is a generating function for the Lefschetz numbers of all iterates of $f$. All these notions are defined in Section 3.

The Morse-Smale diffeormorphisms have simple dynamic behaviour, however they are an important class of discrete dynamical systems. Our objective is to describe the periodic structure of these systems, in particular their minimal sets of periods. The results that we present here are valid for a class of maps that includes the Morse-Smale diffeomorphisms, i.e. $C^1$ maps having finitely many periodic points all of them hyperbolic and with the same action on the homology as the Morse–Smale diffeomorphisms.

Many papers have been published analyzing the relationships between the dynamics of the Morse–Smale diffeomorphisms and the topology of the manifold where they are defined, see for instance [3, 4, 5, 11, 12, 13, 14, 15, 31, 33, 35, 36, 37]. The Morse–Smale

*Date*: September 29, 2013.

2010 *Mathematics Subject Classification.* 37D15, 37E15.

*Key words and phrases.* Morse-Smale diffeomorphisms, quasi-nilpotent maps, Lefschetz numbers, zeta functions, set of periods, minimal set of periods.

diffeomorphisms have a relatively simple orbit structure. In fact, their set of periodic orbits is finite, and their structure is preserved under small $C^1$ perturbations.

Let $X$ be a topological space. Let $f : X \to X$ be a continuous map. A point $x \in M$ is *nonwandering* of $f$ if for any neighborhood $\mathcal{U}$ of $x$ there exists some positive integer $m$ such that $f^m(\mathcal{U}) \cap \mathcal{U} \neq \emptyset$. The set of nonwandering points of $f$ is denoted by $\Omega(f)$.

We say that $x$ is a *periodic point* of $f$ of period $p$ if $f^p(x) = x$ and $f^j(x) \neq x$ for all $0 \leq j < p$. The set $\{x, f(x), \ldots, f^{p-1}(x)\}$ is called the *periodic orbit* of the periodic point $x$. If $X = M$ is a $C^1$ manifold and $f$ a $C^1$ map, we say that $x$ a periodic point of period $p$, is *hyperbolic* if the eigenvalues of $Df^p(x)$ have modulus different from 1.

If $x$ is a hyperbolic periodic point of $f$ of period $p$, the *stable manifold* of $x$ is

$$W^s(x) = \{y \in M \,:\, d(x, f^{pm}(y)) \to 0 \text{ as } m \to \infty\}$$

and the *unstable manifold* of $x$ is

$$W^u(x) = \{y \in M \,:\, d(x, f^{pm}(y)) \to 0 \text{ as } m \to -\infty\},$$

where $d$ is the distance on $M$ induced by the supremum norm.

We say that $M$ is a *closed manifold* if it is a connected compact manifold without boundary. A diffeomorphism $f : M \to M$ is *Morse–Smale* if

(i) $\Omega(f)$ is finite,
(ii) all periodic points are hyperbolic, and
(iii) for each $x, y \in \Omega(f)$, $W^s(x)$ and $W^u(y)$ have transversal intersections.

The first condition implies that $\Omega(f)$ is the set of all periodic points of $f$.

Two diffeomorphisms $f, g \in \mathrm{Diff}(M)$ are $C^1$ *equivalent* if and only if there exists a $C^1$ homeomorphism $h : M \to M$ such that $h \circ f = g \circ h$. A diffeomorphism $f$ is *structurally stable* provided that there exists a neighborhood $\mathcal{U}$ of $f$ in $\mathrm{Diff}(M)$ such that each $g \in \mathcal{U}$ is topologically equivalent to $f$. Since the class of Morse–Smale diffeomorphisms is structurally stable inside the class of all diffeomorphisms (see [33, 34, 32]), to understand the dynamics of this class is an interesting problem.

Let

$$\mathrm{Per}(f) = \{m \in \mathbb{N} \,:\, f \text{ has a periodic orbit of period } m\},$$

i.e. $\mathrm{Per}(f)$ is the *set of periods* of $f$.

The Lefschetz zeta function $\zeta_f(t)$ for a $C^1$ Morse–Smale diffeomorphism $f$ on a closed surface $M$ is introduced in Section 3. Using this function we define the minimal set of Lefschetz periods $\mathrm{MPer}_L(f)$ for a such diffeomorphism $f$ in Section 4. As we shall see the study of this set is important because any other $C^1$ Morse–Smale diffeomorphism $g$ on a manifold $M$ in the same homology class than $f$ satisfies

$$\mathrm{MPer}_L(f) \subseteq \mathrm{Per}(g).$$

The set $\mathrm{MPer}_L(f)$ is computable from the Lefschetz zeta function of $f$, and it consists of odd positive integers, see Proposition 7. In Section 5 we mention the results related to the $\mathrm{MPer}_L(f)$ for maps on orientable closed surfaces, in Section 6 on non-orientable closed surfaces, and in Section 7 on the $n$-dimensional torus.

The results are of two different types, some give an explicit description of the $\mathrm{MPer}_L(f)$ for all Morse-Smale diffeomorphisms on a given manifold, see Theorems 8, 13 and 15. Other results describe what type of subsets of odd positive integers can be $\mathrm{MPer}_L(f)$ for some Morse-Smale diffeomorphisms $f$, see Theorems 9, 10, 11, 12 14, 16 and 17.

## 2. Cyclotomic polynomials

In this section we describe some basic properties of the cyclotomic polynomials which we shall use in our study of the Lefschetz zeta function.

Let $n$ denote an integer. The $n$–th cyclotomic polynomial is given by

$$c_n(t) = \frac{1-t^n}{\prod\limits_{\substack{d|n\\ d<n}} c_d(t)}, \tag{1}$$

for $n > 1$ and $c_1(t) = 1 - t$. An alternative way to express $c_n(t)$ is

$$c_n(t) = \prod_k (w_k - t),$$

for $n \neq 2$, where $w_k = e^{2\pi ik/n}$ and $k$ runs over the relative primes to $n$ and smaller than $n$, for $c_2(t) = -(w_2 - t)$. For more details about these polynomials see [23].

Let $\varphi(n)$ be the degree of $c_n(t)$. Then $n = \sum\limits_{d|n} \varphi(d)$. So $\varphi(n)$ is the *Euler function*, which satisfies

$$\varphi(n) = n \prod_{\substack{p|n\\ p \text{ prime}}} \left(1 - \frac{1}{p}\right).$$

Therefore if the prime decomposition of $n$ is $p_1^{\alpha_1} \cdots p_k^{\alpha_k}$, then

$$\varphi(n) = \prod_{j=1}^{k} p_j^{\alpha_j - 1}(p_j - 1).$$

From the formula (1), we have

$$c_n(t) = \prod_{d|n} (1 - t^d)^{\mu(n/d)}$$

where $\mu$ is the *Möbius function*, i.e.

$$\mu(m) = \begin{cases} 1 & \text{if } m = 1, \\ 0 & \text{if } k^2 | m \text{ for some } k \in \mathbb{N}, \\ (-1)^r & \text{if } m = p_1 \cdots p_r \text{ has distinct primes factors.} \end{cases}$$

**Lemma 1** (Gauss)**.** *Irreducible polynomials whose roots are roots of unity are precisely the collection of cyclotomic polynomials.*

Here are some elementary properties of the cyclotomic polynomials (cf. [23]).

(p1) If $p > 1$ is prime then $c_p(t) = (1 - t^p)/(1 - t)$.
(p2) If $p = 2r$ with $r$ odd then $c_{2r}(t) = c_r(-t)$.
(p3) If $p = 2^\alpha$ with $\alpha$ a positive integer, then $c_p(t) = 1 + t^{2^{\alpha-1}}$.
(p4) If $p = r^\alpha$ with $r > 2$ prime and $\alpha$ a positive integer,, then $c_p(t) = c_r(t^{r^{\alpha-1}}) = (1 - t^{r^\alpha})/(1 - t^{r^{\alpha-1}})$.
(p5) If $p = 2^n r$ with $r$ odd and $n > 1$, then $c_n(t) = c_{2r}(t^{2^{n-1}})$.
(p6) For all $n$ we have that $c_n(0) = 1$, and the leading term of $c_n(t)$ is 1 if $n \geq 2$.
(p7) The degree of $c_n(t)$ is even for $n > 2$.

TABLE 1. The first thirty cyclotomic polynomials.

| | | |
|---|---|---|
| $c_1(t) = 1 - t$ | $c_2(t) = 1 + t$ | $c_3(t) = \dfrac{1-t^3}{1-t}$ |
| $c_4(t) = 1 + t^2$ | $c_5(t) = \dfrac{1-t^5}{1-t}$ | $c_6(t) = \dfrac{1+t^3}{1+t}$ |
| $c_7(t) = \dfrac{1-t^7}{1-t}$ | $c_8(t) = 1 + t^4$ | $c_9(t) = \dfrac{1-t^9}{1-t^3}$ |
| $c_{10}(t) = \dfrac{1+t^5}{1+t}$ | $c_{11}(t) = \dfrac{1-t^{11}}{1-t}$ | $c_{12}(t) = \dfrac{1+t^6}{1+t^2}$ |
| $c_{13}(t) = \dfrac{1-t^{13}}{1-t}$ | $c_{14}(t) = \dfrac{1+t^7}{1+t}$ | $c_{15}(t) = \dfrac{(1-t^{15})(1-t)}{(1-t^3)(1-t^5)}$ |
| $c_{16}(t) = 1 + t^8$ | $c_{17}(t) = \dfrac{1-t^{17}}{1-t}$ | $c_{18}(t) = \dfrac{1+t^9}{1+t^3}$ |
| $c_{19}(t) = \dfrac{1-t^{19}}{1-t}$ | $c_{20}(t) = \dfrac{1+t^{10}}{1+t^2}$ | $c_{21}(t) = \dfrac{(1-t^{21})(1-t)}{(1-t^3)(1-t^7)}$ |
| $c_{22}(t) = \dfrac{1+t^{11}}{1+t}$ | $c_{23}(t) = \dfrac{1-t^{23}}{1-t}$ | $c_{24}(t) = \dfrac{1+t^{12}}{1+t^4}$ |
| $c_{25}(t) = \dfrac{1-t^{25}}{1-t^5}$ | $c_{26}(t) = \dfrac{1+t^{13}}{1+t}$ | $c_{27}(t) = \dfrac{1-t^{27}}{1-t^9}$ |
| $c_{28}(t) = \dfrac{1+t^{14}}{1+t^2}$ | $c_{29}(t) = \dfrac{1-t^{29}}{1-t}$ | $c_{30}(t) = \dfrac{(1+t^{15})(1+t)}{(1+t^3)(1+t^5)}$ |

## 3. Lefschetz Zeta Function

Let $X$ be a topological space which is a retract of a finite simplicial complex [20]. The compact manifolds, the CW complexes are spaces of this type. Let $n$ be the topological dimension of $X$. If $f : X \to X$ is a continuous map on $X$, it induces a homomorphism on the $k$–th rational homology group of $X$ for $0 \leq k \leq n$, i.e. $f_{*k} : H_k(X, \mathbb{Q}) \to H_k(X, \mathbb{Q})$. The $H_k(X, \mathbb{Q})$ is a finite dimensional vector space over $\mathbb{Q}$ and it is torsion free, because it is a vector space over $\mathbb{Q}$. The map $f_{*k}$ is linear given by a matrix with integer entries, then the *Lefschetz number* of $f$ defined as

$$L(f) = \sum_{k=0}^{n} (-1)^k \text{trace}(f_{*k}),$$

is always an integer number.

The Lefschetz Fixed Point Theorem states that if $L(f) \neq 0$ then $f$ has a fixed point (cf. [7]). If we consider the Lefschetz number of $f^m$, then in general is not true that

$L(f^m) \neq 0$ implies that $f$ has a periodic point of period $m$; it only implies the existence of a periodic point with period a divisor of $m$.

The technique of using Lefschetz numbers to obtain information about the periods of a map is also used in many other papers, see for instance the book of Jezierski and Marzantowicz [22], the article of Gierzkiewicz and Wójcik [17] and the references quoted in both.

The *Lefschetz zeta function* of $f$ is defined as

$$\zeta_f(t) = \exp\left(\sum_{m\geq 1} \frac{L(f^m)}{m}\, t^m\right).$$

This function keeps the information of the Lefschetz number for all the iterates of $f$, so this function gives information about the set of periods of $f$. There is an alternative way to compute it:

$$\zeta_f(t) = \prod_{k=0}^{n} \det(Id_{*k} - tf_{*k})^{(-1)^{k+1}}, \tag{2}$$

where $n = \dim X$, $n_k = \dim H_k(X, \mathbb{Q})$, $Id := Id_{*k}$ is the identity map on $H_k(X, \mathbb{Q})$, and by convention $\det(Id_{*k} - tf_{*k}) = 1$ if $n_k = 0$ (cf. [11]).

A rational linear transformation is called *quasi–unipotent* if their eigenvalues are roots of unity. We say that a continuous map $f : X \to X$ is *quasi–unipotent* if the maps $f_{*k}$ are quasi–unipotent for $0 \leq k \leq n$.

**Proposition 2** (Shub [35])**.** *Let $M$ be a compact manifold. If $f : M \to M$ is a Morse–Smale diffeomorphism, then $f$ is quasi–unipotent.*

The following result shows that the class of $C^1$ quasi-unipotent maps are more general that the Morse-Smale diffeomorphisms.

**Theorem 3** ([29])**.** *Let $M$ be a $C^1$ closed manifold of dimension $n$ and $f : M \to M$ be a $C^1$ map with finitely many periodic points all of then hyperbolic. Then the eigenvalues of $f_{*k}$ are zero or roots of unity for $0 \leq k \leq n$.*

When $X$ is a surface, i.e. a 2-dimensional manifold, we can compute the Lefschetz zeta function of a quasi-unipotent self-map on $X$. If $X = M_g$ is an orientable surface of genus $g$ without boundary then $H_0(X, \mathbb{Q}) = \mathbb{Q}$, $H_2(X, \mathbb{Q}) = \mathbb{Q}$ and

$$H_1(X, \mathbb{Q}) = \underbrace{\mathbb{Q} \oplus \cdots \oplus \mathbb{Q}}_{2g}.$$

And if $X = N_g$ a non-orientable surface without boundary of genus $g$, i.e., $X$ is a connected sum of $g$ real projective planes, then $H_0(X, \mathbb{Q}) = \mathbb{Q}$, $H_2(X, \mathbb{Q}) \approx 0$ and

$$H_1(X, \mathbb{Q}) = \underbrace{\mathbb{Q} \oplus \cdots \oplus \mathbb{Q}}_{g-1}.$$

If the rational linear transformation $f_{*k}$ is quasi-unipotent. Then its characteristic polynomial is in $\mathbb{Z}[x]$ and its factors over $\mathbb{Z}$ are irreducible polynomials whose roots are roots of unity. So these factors are cyclotomic polynomials.

The following result is used to compute the Lefschetz zeta functions of $f$.

**Proposition 4** ([26])**.** *If $f_{*_1}$ is quasi–unipotent, then*

$$\det(Id_{*1} - tf_{*1}) = (-1)^{1+\det(f_{*1})} \det(f_{*1} - tId_{*1}).$$

**Proposition 5.** *Let $X$ be a closed surface and $f : X \to X$ be a continuous map such that $f_{*1}$ is quasi-unipotent and $p(t)$ its characteristic polynomial .*

*(1) If $X = N_g$ is a non–orientable closed surface of genus $g$, then*

$$\zeta_f(t) = \frac{p(t)}{1-t}, \tag{3}$$

*being $p(t)$ a product of cyclotomic polynomials of degree $g-1$.*

*(2) If $X = M_g$ is an orientable closed surface of genus $g$, then*

$$\zeta_f(t) = \begin{cases} \dfrac{p(t)}{(1-t)^2} & \text{if } f \text{ is orientation preserving,} \\ \dfrac{p(t)}{(1-t)(1+t)} & \text{if } f \text{ is orientation reversing} \end{cases} \tag{4}$$

*being $p(t)$ a product of cyclotomic polynomials of degree $2g$.*

We say that a rational function $\zeta_f(t)$ is a *possible zeta function*, if $\zeta_f(t)$ satisfies formula either (3), or (4) for a map $f : X \to X$ satisfying the hypothesis of Proposition 5.

Proposition 5 allows us to describe explicitly the possible Lefschetz zeta functions for quasi-unipotent maps on $M_g$ and $N_g$. By finding all possible products of cyclotomic polynomials of total degree $2g$ or $g-1$, in the orientable or non-orientable cases respectively. In the following paragraphs, we present the possible Lefschetz zeta functions for small $g$, see [26, 28] for details.

If $X = M_0 = \mathbb{S}^2$, then $\zeta_f(t) = (1-t)^{-2}$ when $f$ is orientation preserving, and $\zeta_f(t) = (1-t^2)^{-1}$ when $f$ is orientation reversing.

Let $X = M_1 = \mathbb{T}^2$. If $f$ preserves the orientation, then the possible characteristic polynomials of $f_{*1}$ are:

$$c_1^2(t), \quad c_2^2(t), \quad c_3(t), \quad c_4(t), \quad c_6(t).$$

So the possible zeta functions are:

$$1, \quad \frac{(1+t)^2}{(1-t)^2}, \quad \frac{1+t^2}{(1-t)^2}, \quad \frac{1-t^3}{(1-t)^3}, \quad \frac{1+t^3}{(1+t)(1-t)^2}.$$

If $f$ reverses the orientation, then the only possible characteristic polynomial of $f_{*1}$ is $c_1(t)c_2(t)$. Then $\zeta_f(t) = 1$ is the only possible zeta function.

Let $X = M_2$. If $f$ preserves the orientation, then the possible $\zeta_f(t)$ are:

$$\frac{1-t^5}{(1-t)^3}, \quad \frac{1+t^5}{(1+t)(1-t)^2}, \quad \frac{1+t^4}{(1-t)^2}, \quad \frac{1+t^6}{(1+t^2)(1-t)^2},$$

$$\frac{(1-t^3)^2}{(1-t)^4}, \quad \frac{(1-t^3)(1+t^3)}{(1-t)^3(1+t)}, \quad \frac{(1+t^2)(1-t^3)}{(1-t)^3}, \quad \frac{(1+t^2)^2}{(1-t)^2},$$

$$\frac{(1+t^2)(1+t^3)}{(1-t)^2(1+t)}, \quad \frac{(1+t^3)^2}{(1+t)^2(1-t)^2}, \quad \frac{(1-t^3)(1+t)^2}{(1-t)^3}, \quad \frac{(1+t^2)(1+t)^2}{(1-t)^2},$$

$$\frac{(1+t^3)(1+t)}{(1-t)^2}, \quad \frac{1-t^3}{1-t}, \quad 1+t^2, \quad \frac{1+t^3}{1+t},$$

$$\frac{(1+t)^4}{(1-t)^2}, \quad (1+t)^2, \quad (1-t)^2.$$

If $f$ reverses the orientation, then the possible $\zeta_f(t)$ are:

$$\frac{1-t^3}{1-t}, \quad 1+t^2, \quad \frac{1+t^3}{1+t}, \quad (1-t)^2, \quad (1+t)^2.$$

In Tables 2 and 3 are listed the possible Lefschetz zeta functions for quasi-unipotent maps on $M_3$.

Now we consider non–orientable surfaces. If $X = N_1$, i.e. the real projective plane, then the only possible zeta function is $\zeta_f(t) = (1-t)^{-1}$. When $X = N_2$ the Klein bottle, the possible zeta functions are $\zeta_f(t) = 1$ or $\zeta_f(t) = (1+t)(1-t)^{-1}$. When $X = N_3$ the possible Lefschetz zeta functions are

$$1-t, \quad 1+t, \quad \frac{(1+t)^2}{1-t}, \quad \frac{1-t^3}{(1-t)^2}, \quad \frac{1+t^3}{(1-t)(1+t)}, \quad \frac{1+t^2}{1-t}. \tag{5}$$

In Table 4 and 5 the possible Lezfchetz zeta functions on $N_4$ and $N_5$ are listed, respectively. For higher $g$, see [28].

The case of $X = \mathbb{D}_n$, the closed disc with $n$-holes, is similar to $N_{n+1}$; since they have the same homology groups, see [16].

The case of $X = \mathbb{T}^n$ is studied in [14] and [6]. The homology spaces of $\mathbb{T}^n$ with rational coefficients are

$$H_k(\mathbb{T}^n, \mathbb{Q}) = \underbrace{\mathbb{Q} \oplus \cdots \oplus \mathbb{Q}}_{n_k},$$

where $n_k = \binom{n}{k}$. Since the homology spaces of $\mathbb{T}^n$ form an exterior algebra, then the map $f_{*1}$ determines all the other $f_{*k}$, for $2 \leq k \leq n$, in the following way (*cf.* [38]): Let $p_1(t)$ be the characteristic polynomial of $f_{*1}$, then

$$p_1(t) = \prod_{j=1}^{n} (t - \lambda_j),$$

where the $\lambda_j$ are the eigenvalues of $f_{*1}$. Then the other $p_k(t)$ are expressed as:

$$\begin{aligned} p_2(t) &= \prod_{i<j} (t - \lambda_i \lambda_j), \\ p_3(t) &= \prod_{i<j<l} (t - \lambda_i \lambda_j \lambda_l), \\ &\vdots \\ p_n(t) &= t - (\lambda_1 \lambda_2 \cdots \lambda_n). \end{aligned}$$

TABLE 2. Possible Lefschetz zeta functions for orientation preserving quasi-unipotent maps on $M_3$

| | | | |
|---|---|---|---|
| $\frac{1-t^7}{(1-t)^3}$ | $\frac{1-t^9}{(1-t)^2(1-t^3)}$ | $\frac{1+t^7}{(1-t)^2(1+t)}$ | $\frac{1+t^9}{(1+t^3)(1-t)^2}$ |
| $\frac{(1-t^5)(1-t^3)}{(1-t)^4}$ | $\frac{(1-t^5)(1+t^2)}{(1-t)^3}$ | $\frac{(1-t^5)(1+t^3)}{(1-t)^3(1+t)}$ | $\frac{(1+t^5)(1-t^3)}{(1+t)(1-t)^3}$ |
| $\frac{(1+t^5)(1+t^2)}{(1+t)(1-t)^2}$ | $\frac{(1+t^5)(1+t^3)}{(1-t)^2(1+t)^2}$ | $\frac{(1+t^4)(1-t^3)}{(1-t)^3}$ | $\frac{(1+t^4)(1+t^2)}{(1-t)^2}$ |
| $\frac{(1+t^4)(1+t^3)}{(1-t)^2(1+t)}$ | $\frac{(1+t^6)(1-t^3)}{(1+t^2)(1-t)^3}$ | $\frac{1+t^6}{(1-t)^2}$ | $\frac{(1+t^6)(1+t^3)}{(1+t^2)(1-t)^2(1+t)}$ |
| $\frac{(1-t^3)^3}{(1-t)^5}$ | $\frac{(1-t^3)^2(1+t^2)}{(1-t)^4}$ | $\frac{(1-t^3)^2(1+t^3)}{(1-t)^4(1+t)}$ | $\frac{(1-t^3)(1+t^3)(1+t^2)}{(1-t)^3(1+t)}$ |
| $\frac{(1-t^3)(1+t^3)^2}{(1-t)^3(1+t)^2}$ | $\frac{(1+t^2)^2(1-t^3)}{(1-t)^3}$ | $\frac{(1+t^2)^3}{(1-t)^2}$ | $\frac{(1+t^2)^2(1+t^3)}{(1-t)^2(1+t)}$ |
| $\frac{(1+t^2)(1+t^3)^2}{(1-t)^2(1+t)^2}$ | $\frac{(1+t^3)^3}{(1+t)^3(1-t)^2}$ | $\frac{(1-t^3)^2(1+t)^2}{(1-t)^4}$ | $\frac{(1-t^3)(1+t)^2(1+t^2)}{(1-t)^3}$ |
| $\frac{(1-t^3)(1+t)(1+t^3)}{(1-t)^3}$ | $\frac{(1+t^2)^2(1+t)^2}{(1-t)^2}$ | $\frac{(1+t^2)(1+t)(1+t^3)}{(1-t)^2}$ | $\frac{(1+t^3)^2}{(1-t)^2}$ |
| $\frac{(1-t^3)^2}{(1-t)^2}$ | $\frac{(1-t^3)(1+t^2)}{1-t}$ | $\frac{(1-t^3)(1+t^3)}{(1-t)(1+t)}$ | $(1+t^2)^2$ |
| $\frac{(1+t^2)(1+t^3)}{1+t}$ | $\frac{(1+t^3)^2}{(1+t)^2}$ | $\frac{(1+t)^4(1-t^3)}{(1-t)^3}$ | $\frac{(1+t)^4(1+t^2)}{(1-t)^2}$ |
| $\frac{(1+t)^3(1+t^3)}{(1-t)^2}$ | $\frac{(1-t^3)(1+t)^2}{1-t}$ | $(1+t^2)(1+t)^2$ | $(1+t^3)(1+t)$ |
| $(1-t^3)(1-t)$ | $(1+t^2)(1-t)^2$ | $\frac{(1+t^3)(1-t)^2}{1+t}$ | $\frac{1-t^5}{1-t}$ |
| $\frac{(1-t^5)(1+t)^2}{(1-t)^3}$ | $\frac{1+t^5}{1+t}$ | $\frac{(1+t^5)(1+t)}{(1-t)^2}$ | $1+t^4$ |
| $\frac{(1+t^4)(1+t)^2}{(1-t)^2}$ | $\frac{1+t^6}{1+t^2}$ | $\frac{(1+t^6)(1+t)^2}{(1+t^2)(1-t)^2}$ | $\frac{(1-t^3)^2(1+t)^2}{(1-t)^4}$ |
| $\frac{(1+t^2)(1+t)^4}{(1-t)^2}$ | $(1+t)^4$ | $\frac{(1+t)^6}{(1-t)^2}$ | $(1-t)^2(1+t)^2$ |
| $(1-t)^4$ | | | |

Using this information, Proposition 4 and formula (2), we can compute explicitly some of the possible possible Lefschetz zeta functions for quasi-unipotent maps on $\mathbb{T}^n$ (*cf.* [6]). In [14] all possible $\zeta_f(t)$ for quasi-unipotent maps on $\mathbb{T}^3$ and $\mathbb{T}^4$ are listed.

TABLE 3. Possible Lefschetz zeta functions for orientation reversing quasi-unipotent maps on $M_3$.

| | | | |
|---|---|---|---|
| $\frac{(1-t^3)^2}{(1-t)^2}$ | $\frac{(1-t^3)(1+t^2)}{1-t}$ | $\frac{(1-t^3)(1+t^3)}{(1-t)(1+t)}$ | $(1+t^2)^2$ |
| $\frac{(1+t^2)(1+t^3)}{1+t}$ | $\frac{(1+t^3)^2}{(1+t)^2}$ | $\frac{(1-t^3)(1+t)^2}{(1-t)}$ | $(1-t^3)(1-t)$ |
| $(1+t^2)(1+t)^2$ | $(1+t^2)(1-t)^2$ | $(1+t^3)(1+t)$ | $\frac{(1+t^3)(1-t)^2}{1+t}$ |
| $(1+t)^4$ | $(1+t)^2(1-t)^2$ | $(1-t)^4$ | |

TABLE 4. Possible Lefschetz zeta functions for quasi-unipotent maps on $N_4$.

| | | | | |
|---|---|---|---|---|
| $(1-t)^2$, | $(1-t)(1+t)$, | $(1+t)^2$, | $\frac{(1+t)^3}{1-t}$, | $\frac{1-t^3}{1-t}$, |
| $\frac{1+t^3}{1+t}$, | $1+t^2$, | $\frac{(1+t)(1-t^3)}{(1-t)^2}$, | $\frac{(1+t)(1+t^2)}{1-t}$. | $\frac{1+t^3}{1-t}$ |

TABLE 5. Possible Lefschez zeta functions for quasi-unipotent maps on $N_5$.

| | | | |
|---|---|---|---|
| $\frac{1-t^5}{(1-t)^2}$, | $\frac{1+t^5}{(1-t)(1+t)}$, | $\frac{1+t^4}{1-t}$, | $\frac{(1-t^3)^2}{(1-t)^3}$, |
| $\frac{(1-t^3)(1+t^3)}{(1+t)(1-t)^2}$, | $\frac{(1-t^3)(1+t^2)}{(1-t)^2}$, | $\frac{(1+t^3)^2}{(1+t)^2(1-t)}$, | $\frac{(1+t^3)(1+t^2)}{(1+t)(1-t)}$, |
| $\frac{(1+t^2)^2}{1-t}$, | $\frac{1+t^6}{(1-t)(1+t^2)}$ | $\frac{(1-t^3)(1+t)}{1-t}$, | $\frac{(1-t^3)(1+t)^2}{(1-t)^2}$, |
| $\frac{(1+t^3)(1-t)}{1+t}$, | $1+t^3$, | $\frac{(1+t^3)(1+t)}{1-t}$, | $1-t^3$, |
| $(1+t^2)(1-t)$, | $(1+t^2)(1+t)$, | $\frac{(1+t^2)(1+t)^2}{1-t}$, | $(1-t)^3$, |
| $(1+t)(1-t)^2$, | $(1+t)^2(1-t)$, | $(1+t)^3$, | $\frac{(1+t)^4}{1-t}$. |

## 4. The minimal set of Lefschetz periods $\mathrm{MPer}_L(f)$

In this section we assume that $X$ is a $C^1$ compact manifold. and let $f : X \to X$ be a $C^1$ map. Let $x$ be a hyperbolic periodic point of period $p$ of $f$ and $E^u_x$ its unstable space,

i.e. the subspace of the tangent space $T_xX$ generated by the eigenvectors of $Df^p(x)$ of modulus larger than 1. Let $\gamma$ be the orbit of $x$, the *index* $u$ of $\gamma$ is the dimension of $E^u_x$. We define the *orientation* type $\Delta$ of $\gamma$ as $+1$ if $Df^p(x) : E^u_x \to E^u_x$ preserves orientation and $-1$ if reverses the orientation. The collection of the triples $(p, u, \Delta)$ belonging to all periodic orbits of $f$ is called the *periodic data* of $f$. The same triple can appear more than once if it corresponds to different periodic orbits.

**Theorem 6** (Franks [10])**.** *Let $f$ be a $C^1$ a map on a closed manifold having finitely many periodic points all of them hyperbolic, and let $\Sigma$ be the periodic data of $f$. Then the Lefschetz zeta function $\zeta_f(t)$ of $f$ satisfies*

$$\zeta_f(t) = \prod_{(p,u,\Delta)\in\Sigma} (1 - \Delta t^p)^{(-1)^{u+1}}. \tag{6}$$

Clearly the Morse–Smale diffeomorphisms on orientable and non–orientable closed manifolds satisfy the hypotheses of this theorem.

We remark, this theorem is also true when $X$ is a $C^1$ compact manifold with boundary and $f : X \to X$ a $C^1$ map such that it does not have periodic points on the boundary of $X$, see [10].

Theorem 6 allows to define the minimal set of Lefschetz periods of a $C^1$ map on a compact manifold having finitely many periodic points all of them hyperbolic. Such a map has a Lefschetz zeta function of the form (6). Note that in general the expression of one of these Lefschetz zeta functions is not unique as product of elements of the form $(1 \pm t^p)^{\pm 1}$. For instance the following possible Lefschetz zeta function can be written in four different ways in the form given by (6):

$$\zeta_f(t) = \frac{(1-t^3)(1+t^3)}{(1-t)^2(1+t)} = \frac{1-t^6}{(1-t)^2(1+t)} = \frac{1-t^6}{(1-t)(1-t^2)} = \frac{(1-t^3)(1+t^3)}{(1-t)(1-t^2)}.$$

According with Theorem 6, the first expression will provide the periods $\{1, 3\}$ for $f$, the second the periods $\{1, 6\}$, the third the period $\{1, 2, 6\}$, and finally the fourth the periods $\{1, 2, 3\}$. Then for this Lefschetz zeta function $\zeta_f(t)$ we will define its *minimal set of Lefschetz periods* as

$$\mathrm{MPer}_L(f) = \{1, 3\} \cap \{1, 6\} \cap \{1, 2, 6\} \cap \{1, 2, 3\} = \{1\}.$$

If $\zeta_f(t) \neq 1$ then it can be written as

$$\zeta_f(t) = \prod_{i=1}^{N_\zeta} (1 + \Delta_i t^{r_i})^{m_i}, \tag{7}$$

where $\Delta_i = \pm 1$, the $r_i$'s are positive integers, $m_i$'s are nonzero integers and $N_\zeta$ is a positive integer depending on $f$.

If $\zeta_f(t) \neq 1$ the *minimal set of Lefschetz periods* of $f$ is defined as

$$\mathrm{MPer}_L(f) := \bigcap \{r_1, \ldots, r_{N_\zeta}\},$$

where the intersection is considered over all the possible expressions (7) of $\zeta_f(t)$. If $\zeta_f(t) = 1$, then we define $\mathrm{MPer}_L(f) := \emptyset$. Roughly speaking the minimal set of Lefschetz periods of $f$ is the intersection of all the sets of periods forced by the finitely many different representations of $\zeta_f(t)$ as product and quotient of elements of the form $(1 \pm t^p)^{\pm 1}$. Clearly

$$\mathrm{MPer}_L(f) \subseteq \mathrm{Per}(f).$$

If $M$ is a closed $C^1$ manifold we denote by $\mathcal{F}(M)$ the set of all $C^1$ self-maps on $M$ having finitely many periodic points all of them hyperbolic, and such that the induced maps on homology $f_{*k}$ are quasi-unipotent. By Proposition 2 the Morse-Smale diffeomorphisms on $M$ belong to $\mathcal{F}(M)$. It can be checked that there are elements in $\mathcal{F}(M)$, which are not Morse-Smale diffeormorphisms.

The results of this article apply to maps $f$ in $\mathcal{F}(M)$. Moreover the techniques used in the proofs of the main theorems of the present paper, can be used when the map $f$ satisfies the hypothesis of Theorem 3, i.e. $f$ has finitely many periodic points all of them hyperbolic. Therefore it is possible to obtain similar results with this weaker hypothesis.

**Proposition 7** ([21, 27])**.** *There are no even numbers in $MPer_L(f)$.*

*Proof.* If the number $2d$ is in $\mathrm{MPer}_L(f)$ then $(1 \pm t^{2d})^m$ is a factor of the Lefschetz zeta function $\zeta_f(t)$, for some $m \neq 0$. So if the factor is $(1-t^{2d})^m$ it can be written as $(1-t^d)^m(1+t^d)^m$. Then due to the fact that the intersection of the exponents is taken over all possible expressions (7) of $\zeta_f(t)$, the number $2d$ is not in $\mathrm{MPer}_L(f)$.

If the factor is $(1+t^{2d})^m$, then it can be written as

$$(1+t^{2d})^m = \frac{(1+t^{2d})^m(1-t^{2d})^m}{(1-t^{2d})^m} = \frac{(1-t^{4d})^m}{(1-t^d)^m(1+t^d)^m}.$$

Therefore, again $2d \notin \mathrm{MPer}_L(f)$. □

## 5. $\mathrm{MPer}_L(f)$ for maps in $\mathcal{F}(M_g)$

In this section we summarize the results related to the minimal set of Lefschetz periods for maps on orientable closed surfaces.

**Theorem 8** ([26])**.** *Let $f \in \mathcal{F}(M_g)$.*

(a) *If $g = 0$ then $MPer_L(f) = \{1\}$ if $f$ preserves orientation, and $MPer_L(f) = \emptyset$ if $f$ reverses orientation.*
(b) *If $g = 1$ then $MPer_L(f) = \emptyset$ if $f$ reverses orientation, and*

$$MPer_L(f) = \begin{cases} \emptyset & \text{if} \quad \zeta_f(t) = 1, \\ \{1\} & \text{if} \quad \zeta_f(t) = \dfrac{(1+t)^2}{(1-t)^2}, \ \text{or } \zeta_f(t) = \dfrac{1+t^2}{(1-t)^2}, \\ \{1,3\} & \text{if} \quad \zeta_f(t) = \dfrac{1-t^3}{(1-t)^3}, \ \text{or } \dfrac{1+t^3}{(1-t)(1-t^2)}. \end{cases}$$

*if $f$ preserves orientation.*
(c) *If $g = 2$ and $f$ is orientation reversing then*

$$MPer_L(f) = \begin{cases} \emptyset & \text{if} \quad \zeta_f(t) = 1+t^2, \\ \{1\} & \text{if} \quad \zeta_f(t) = (1-t)^2, \ \text{or } \zeta_f(t) = (1+t)^2, \\ \{1,3\} & \text{if} \quad \zeta_f(t) = \dfrac{1-t^3}{(1-t)}, \ \text{or } \dfrac{1+t^3}{(1+t)}. \end{cases}$$

(d) *If $g = 2$ and $f$ is orientation preserving then its $MPer_L(f)$ is one of the following ones:*

$\{1\}$ *if* $\zeta_f(t)$ *is* $\dfrac{(1-t^3)(1+t^3)}{(1-t)^3(1+t)}$, $\dfrac{(1+t)^4}{(1-t)^2}$, $(1+t)^2$, $(1-t)^2$, $\dfrac{(1+t^2)(1+t)^2}{(1-t)^2}$, $\dfrac{(1+t^2)^2}{(1-t)^2}$, $\dfrac{1+t^4}{(1-t)^2}$ *or* $\dfrac{1+t^6}{(1+t^2)(1-t)^2}$;

$\{3\}$ *if* $\zeta_f(t)$ *is* $\dfrac{(1+t^3)^2}{(1+t)^2(1-t)^2}$;

$\{1,3\}$ *if* $\zeta_f(t)$ *is* $\dfrac{(1-t^3)^2}{(1-t)^4}$, $\dfrac{(1+t^3)(1+t)}{(1-t)^2}$, $\dfrac{1-t^3}{1-t}$, $\dfrac{1+t^3}{1+t}$, $\dfrac{(1-t^3)(1+t)^2}{(1-t)^2}$, $\dfrac{(1+t^2)(1-t^3)}{(1-t)^3}$ *or* $\dfrac{(1+t^2)(1+t^3)}{(1-t)^2(1+t)}$;

$\{1,5\}$ *if* $\zeta_f(t)$ *is* $\dfrac{1-t^5}{(1-t)^3}$ *or* $\dfrac{1+t^5}{(1+t)(1-t)^2}$;

$\emptyset$ *if* $\zeta_f(t)$ *is* $1+t^2$.

(e) *If $g = 3$ and $f$ is orientation reversing then $MPer_L(f)$ is $\emptyset$, $\{1\}$, or $\{1,3\}$.*

(f) *If $g = 3$ and $f$ is orientation preserving then $MPer_L(f)$ is $\emptyset$, $\{1\}$, $\{1,3\}$, $\{1,5\}$, $\{1,7\}$, $\{3\}$, $\{3,5\}$, $\{1,3,5\}$ or $\{1,3,9\}$.*

The proof of Theorem 8 follows from the calculation of all possible Lefschetz zeta functions, as it was shown in Section 3, and after using the definition of the Lefschetz periods.

From the properties of the cyclotomic polynomials we obtain Theorems 9, 10, 11 and 12.

**Theorem 9** ([26])**.** *The following statements hold.*

(a) *If $2g-1$ is an odd prime, then $\{1, 2g-1\}$ is a possible $MPer_L(f)$ for some orientation preserving map $f \in \mathcal{F}(M_g)$.*

(b) *If $g = p^{\alpha-1}(p-1)/2 + 1$ with $p$ odd prime then $p^\alpha$ and $p^{\alpha-1}$ are contained in the $MPer_L(f)$ for some orientation preserving map $f \in \mathcal{F}(M_g)$. If $p = 2$ then $MPer_L(f) = \emptyset$.*

(c) *If $g$ is an odd prime number, then $MPer_L(f) = \emptyset$ for some orientation preserving map $f \in \mathcal{F}(M_g)$.*

We would like to know what kind of subsets of the positive odd integers can be realized as $\mathrm{MPer}_L(f)$ for some maps $f \in \mathcal{F}(M_g)$, for some $g$. The following result gives a partial answer to this question.

**Theorem 10** ([21])**.** *Let $p_1, \dots, p_k$ be distinct odd primes and $\alpha_{i,j}$ integers, such that $\alpha_{i,j} > \alpha_{i,j+1}$ for $1 \le i \le k$, $1 \le j \le l_i$. Let $S$ be one of the following sets*

*(1)* $S = \{p_1, \dots, p_k\}$,

*(2)* $S = \left\{p_i^{\alpha_{i,1}}, \dots, p_i^{\alpha_{i,l_i}}\right\}$, *for* $1 \le i \le k$,

*(3)* $S = \left\{p_1^{\alpha_{1,1}}, \dots, p_1^{\alpha_{1,l_1}}, \dots, p_k^{\alpha_{k,1}}, \dots, p_k^{\alpha_{k,l_k}}\right\}$,

*(4)* $S = \{1, p_1, \dots, p_k\}$,

*(5)* $S = \left\{1, p_i^{\alpha_{i,1}}, \ldots, p_i^{\alpha_{i,l_i}}\right\}$, *for* $1 \leq i \leq k$, *or*

*(6)* $S = \left\{1, p_1^{\alpha_{1,1}}, \ldots, p_1^{\alpha_{1,l_1}}, \ldots, p_k^{\alpha_{k,1}}, \ldots, p_k^{\alpha_{k,l_k}}\right\}$.

*Then there is a possible Lefschetz zeta function* $\zeta_f(t)$, *with* $f \in \mathcal{F}(M_g)$ *for some* $g$, *such that* $MPer_L(f) = S$.

Theorems 11 and 12 give a complete characterization when the number 1 belongs to the minimal set of Lefschetz periods for a map $f \in \mathcal{F}(M_g)$. This characterization is given in terms of the arithmetic properties of the cyclotomic polynomials which are factors of the characteristic polynomial of the induced map on the first homology space.

**Theorem 11** ([21])**.** *Let* $f \in \mathcal{F}(M_g)$ *an orientation reversing map. Let* $q(t)$ *be the characteristic polynomial of* $f_{*1}$. *Then* $q(t)$ *can be written as*

$$q(t) = (c_{n_1}(t))^{\gamma_1} \cdots (c_{n_k}(t))^{\gamma_k} (c_{2^{\alpha_1} m_1}(t))^{\beta_1} \cdots (c_{2^{\alpha_l} m_l}(t))^{\beta_l} (c_1(t))^{2n+1} (c_2(t))^{2m+1},$$

*where* $n_1, \ldots, n_k, m_1, \ldots, m_l$ *are positive odd integers, greater than* 1, *and* $\alpha_i \geq 1$, $\beta_j \geq 0$, $\gamma_i \geq 0$, $n, m \geq 0$, *such that*

$$\left(\sum_{i=1}^{k} \varphi(n_i)\gamma_i\right) + \left(\sum_{i=1}^{l} \varphi(m_i) 2^{\alpha_i - 1} \beta_i\right) + 2n + 2m + 2 = 2g,$$

*with* $\varphi(m)$ *the Euler function of* $m$. *Moreover* $1 \notin MPer_L(f)$ *if and only if*

$$\sum_{\{j\,:\,\alpha_j = 1\}} \mu(m_j)\beta_j + 2m = \sum_{i=1}^{k} \mu(n_i)\gamma_i + 2n,$$

*where* $\mu(m)$ *is the Möbius function of* $m$

**Theorem 12** ([21])**.** *Let* $f \in \mathcal{F}(M_g)$ *an orientation reversing map. Let* $q(t)$ *be the characteristic polynomial of* $f_{*1}$. *Then* $q(t)$ *can be written as Then*

$$q(t) = (c_{n_1}(t))^{\gamma_1} \cdots (c_{n_k}(t))^{\gamma_k} (c_{2^{\alpha_1} m_1}(t))^{\beta_1} \cdots (c_{2^{\alpha_l} m_l}(t))^{\beta_l} (c_1(t))^{2n} (c_2(t))^{2m},$$

*where* $n_1, \ldots, n_k, m_1, \ldots, m_l$ *are positive odd integers greater than* 1, *and* $\alpha_i \geq 1$, $\beta_j \geq 0$, $\gamma_i \geq 0$, $n, m \geq 0$, *such that*

$$\left(\sum_{i=1}^{k} \varphi(n_i)\gamma_i\right) + \left(\sum_{i=1}^{l} \varphi(m_i) 2^{\alpha_i - 1} \beta_i\right) + 2n + 2m = 2g,$$

*with* $\varphi$ *the Euler function. Furthermore* $1 \notin MPer_L(f)$ *if and only if*

$$\sum_{\{j\,:\,\alpha_j = 1\}} \mu(m_j)\beta_j + 2m = \sum_{i=1}^{k} \mu(n_i)\gamma_i + 2(n-1).$$

## 6. $\mathrm{MPer}_L(f)$ for maps in $\mathcal{F}(N_g)$

In this section we summarize the results related to the minimal set of Lefschetz periods for maps on non-orientable closed surfaces.

**Theorem 13** ([28])**.** *Let* $f$ *be a Morse–Smale diffeomorphism on* $N_g$, *or more generally a map belonging to* $\mathcal{F}(N_g)$.

(a) *If* $g = 1$ *then* $MPer_L(f)$ *is* $\{1\}$.
(b) *If* $g = 2$ *then* $MPer_L(f)$ *is* $\emptyset$ *or* $\{1\}$.
(c) *If* $g = 3$ *then* $MPer_L(f)$ *is* $\{1\}$, $\{3\}$ *or* $\{1, 3\}$.
(d) *If* $g = 4$ *then* $MPer_L(f)$ *is* $\emptyset$, $\{1\}$ *or* $\{1, 3\}$.

(e) *If $g = 5$ then $MPer_L(f)$ is $\{1\}$, $\{3\}$, $\{5\}$, $\{1,3\}$ or $\{1,5\}$.*
(f) *If $g = 6$ then $MPer_L(f)$ is $\emptyset$, $\{1\}$, $\{3\}$, $\{1,3\}$ or $\{1,5\}$.*
(g) *If $g = 7$ then $MPer_L(f)$ is $\{1\}$, $\{3\}$, $\{5\}$, $\{7\}$, $\{1,3\}$, $\{1,5\}$ $\{1,7\}$, $\{1,3,5\}$ or $\{1,3,9\}$.*
(h) *If $g = 8$ then $MPer_L(f)$ is $\emptyset$, $\{1\}$, $\{3\}$, $\{1,3\}$, $\{1,5\}$, $\{1,7\}$, $\{3,5\}$, $\{3,9\}$, $\{1,3,5\}$ or $\{1,3,9\}$.*
(i) *If $g = 9$ then $MPer_L(f)$ is $\{1\}$, $\{3\}$, $\{5\}$, $\{7\}$,$\{9\}$, $\{1,3\}$, $\{1,5\}$ $\{1,7\}$, $\{1,9\}$, $\{3,9\}$, $\{1,3,5\}$, $\{1,3,7\}$, $\{1,3,9\}$, $\{3,5,15\}$ or $\{1,3,5,15\}$.*

**Theorem 14** ([28])**.** *The following statement hold.*

(a) *If $g$ is an odd prime, then the sets $\{1,g\}$ and $\{g\}$ are possible $MPer_L(f)$ for some $f \in \mathcal{F}(N_g)$.*
(b) *If $g = p^{\alpha-1}(p-1)+1$ with $p$ an odd prime and $\alpha > 1$, then the set $\{1, p^{\alpha-1}, p^\alpha\}$ is a possible $MPer_L(f)$ for some $f \in \mathcal{F}(N_g)$.*
(c) *If $g$ is even, then the empty set is a possible $MPer_L(f)$ for some $f \in \mathcal{F}(N_g)$.*
(d) *If $g$ is odd, then $MPer_L(f) \neq \emptyset$ for all $f \in \mathcal{F}(N_g)$.*
(e) *For every positive integer $g$, $\{1\}$ is a possible $MPer_L(f)$ for some $f \in \mathcal{F}(N_g)$.*
(f) *If $g$ is odd and $p$ is an odd prime such that $1 < p \leq g$, then $\{p\}$ is a possible $MPer_L(f)$ for some $f \in \mathcal{F}(N_g)$.*
(g) *Let $p_1, \ldots, p_k$ be different odd primes larger than 1, and let $\alpha_1, \ldots, \alpha_k$ be positive integers.*
  (g.1) *Then the set $\{p_1, \ldots, p_k\}$ is a possible $MPer_L(f)$ for some $f \in \mathcal{F}(N_g)$, where $g = \left(\sum_{i+1}^{k} p_i\right) - (k-1)$ if $k$ is odd, and $g = \left(\sum_{i+1}^{k} p_i\right) - (k-2)$ if $k$ is even.*
  (g.2) *Then the set $\{p_1^{\alpha_1}, \ldots, p_k^{\alpha_k}\}$ is a possible $MPer_L(f)$ for some $f \in \mathcal{F}(N_g)$ and for some convenient genus $g$.*
  (g.3) *Let $p$ be an odd prime number. Then the set $\{p^{\alpha_1}, \ldots, p^{\alpha_k}\}$ is a possible $MPer_L(f)$ for some $f \in \mathcal{F}(N_g)$ and for some convenient genus $g$.*
  (g.4) *Let $\alpha_{i,j}$ positive integers such that $\alpha_{i,j} > \alpha_{i,j+1}$ for $1 \leq i \leq k$ and $1 \leq j \leq l_i$. Then the set $\{p_1^{\alpha_{1,1}}, \ldots, p_1^{\alpha_{1,l_1}}, \ldots, p_k^{\alpha_{k,1}}, \ldots, p_k^{\alpha_{k,l_k}}\}$ is a possible $MPer_L(f)$ for some $f \in \mathcal{F}(N_g)$ and for some convenient genus $g$.*

Statement (d) of Theorem 14 shows an important difference in the periodic structure between the $C^1$ Morse–Smale diffeomorphisms on orientable and non–orientable compact surfaces without boundary. Statement (c) of Theorem 9 states that for all orientable compact surfaces without boundary there are $C^1$ Morse–Smale diffeomorphisms having their minimal set of Lefschetz periods empty, and statement (d) of Theorem 14 shows that this is never the case for $C^1$ Morse–Smale diffeomorphisms on the non–orientable closed surfaces.

The results of statements (f) and (g) of Theorem 14 also for $C^1$ Morse–Smale diffeomorphisms on orientable closed surfaces, and their proof are similar to the proof of Theorem 10.

The description of the $\mathrm{MPer}_L(f)$ for Morse-Smale diffeomorphisms on $\mathbb{D}_g$ (and maps in $\mathcal{F}(\mathbb{D}_g)$), which does not have periodic points on the boundary is the same of the $\mathrm{MPer}_L(f)$ of Morse-Smale diffeomorphisms on $N_{g+1}$ (maps in $\mathcal{F}(N_{g+1})$), since the possible Lefschetz zeta functions are the same, see [16].

**Open question**. *Can any finite set of odd positive integers be the minimal set of Lefschetz periods for a $C^1$ Morse–Smale diffeomorphism on some orientable/non–orientable compact surface without boundary with a convenient genus?*

We think that the answer to this open question is positive. In [21] this open question was stated as a conjecture for the $C^1$ Morse–Smale diffeomorphisms on orientable compact surface without boundary.

## 7. $\mathrm{MPER}_L(f)$ FOR MAPS IN $\mathcal{F}(\mathbb{T}^n)$

In this section we summarize the results related to the minimal set of Lefschetz periods for maps on the $n$-dimensional torus.

**Theorem 15** ([14]). *Let $f \in \mathcal{F}(\mathbb{T}^n)$.*

(a) *If $n = 3$ and $f$ is orientation preserving then $MPer_L(f) = \emptyset$ .*
(b) *If $n = 3$ and $f$ is orientation reversing then $MPer_L(f) = \emptyset$, $\{1\}$, or $\{1, 3\}$.*
(c) *If $n = 4$ and $f$ is orientation reversing then $MPer_L(f) = \emptyset$ .*
(d) *If $n = 4$ and $f$ is orientation preserving then $MPer_L(f) = \emptyset$, $\{1\}$,$\{1, 3\}$ or $\{1, 5\}$.*

**Theorem 16** ([14]). *If $n$ is even then $MPer_L(f) = \emptyset$, for $f \in \mathcal{F}(\mathbb{T}^n)$ and orientation reversing. If $n$ is odd then $MPer_L(f) = \emptyset$, for $f \in \mathcal{F}(\mathbb{T}^n)$ and orientation preserving.*

**Theorem 17** ([6]). *Let $f$ be an orientation preserving map in $\mathcal{F}(\mathbb{T}^n)$, with $n = p - 1$ and $p$ and odd prime, such that the characteristic polynomial of $f_{*1}$ is $c_p(t)$, Then $MPer_L(f) = \{1, p\}$.*

## References


[1] Ll. Alsedà, J. Llibre and M. Misiurewicz, *Combinatorial Dynamics and Entropy in dimension one* (Second Edition), Advanced Series in Nonlinear Dynamics, vol **5**, World Scientific, Singapore 2000.

[2] I.K. Babenko and S.A. Bagatyi, The behavior of the index of periodic points under iterations of a mapping, *Math. USSR Izvestiya* **38** (1992), 1–26.

[3] S. Batterson, The dynamics of Morse–Smale diffeomorphisms on the torus, *Trans. Amer. Math. Soc.* **256** (1979), 395–403.

[4] S. Batterson, Orientation reversing Morse–Smale diffeomorphisms on the torus. *Trans. Amer. Math. Soc.* **264** (1981), 29–37.

[5] S. Batterson, M. Handel and C. Narasimhan, Orientation reversing Morse–Smale diffeomorphisms of $\mathbb{S}^2$. *Invent. Math.* **64** (1981), 345–356.

[6] P. Berrizbeitia and V.F. Sirvent, On the Lefschetz zeta function for quasi–unipotent maps on the $n$-dimensional torus, pre-print.

[7] R.F. Brown, *The Lefschetz fixed point theorem*, Scott, Foresman and Company, Glenview, IL, 1971.

[8] J. Casasayas, J. Llibre and A. Nunes, Periods and Lefschetz zeta functions, *Pacific Journal of Mathematics* **165** (1994), 51–66.

[9] N. Fagella and J. Llibre, Periodic points of holomorphic maps via Lefschetz numbers, *Trans. Amer. Math. Soc.* **352** (2000), 4711–4730.

[10] J. Franks, Some smooth maps with infinitely many hyperbolic points, *Trans. Amer. Math. Soc.* **226** (1977), 175–179,

[11] J. Franks, *Homology and dynamical systems*, CBSM Regional Conf. Ser. in Math. **49**, Amer. Math. Soc., Providence, R.I. 1982.

[12] J. Franks and C. Narasimhan, The periodic behaviour of Morse–Smale diffeomorphisms, *Invent. Math.* **48** (1978), 279–292.

[13] J.L. García Guirao and J. Llibre, Periods for the Morse–Smale diffeomorphisms on $\mathbb{S}^2$, Colloquium Mathematicum **200** (2007), 477–483.

[14] J.L. García Guirao and J. Llibre, Minimal Lefschetz sets of periods of Morse–Smale diffeomorphisms on the $n$-dimensional torus, *J. of Difference Equations and Applications* **16** (2010), 689–703.
[15] J.L. García Guirao and J. Llibre, The set of periods for the Morse–Smale diffeomorphisms on $\mathbb{T}^2$, Dynamics of Continuous, Discrete and Impulsive Systems Series A: Mathematical Analysis **19** (2012), 471–484.
[16] J.L. García Guirao and J. Llibre, On the set of periods for the Morse-Smale diffeomorphisms on the disc with $n$–holes, to appear in *J. of Difference Equations and Applications.*
[17] A. Gierzkiewicz and K. Wójcik, *Lefschetz sequences and detecting periodic points*, Discrete Contin. Dyn. Syst. Ser. A **32** (2012), 81–100.
[18] J. Guaschi and J. Llibre, *Orders and periods of finite order homology surface maps*, Houston J. of Math. **23** (1997), 449–483.
[19] A. Guillamon, X. Jarque, J. Llibre, J. Ortega, J. Torregosa, Periods for transversal maps via Lefschetz numbers for periodic points, *Trans. Amer. Math. Soc.* **347** (1995), 4779–4806.
[20] A. Hatcher, *Algebraic Topology*, Cambridge University Press, 2002.
[21] B. Iskra and V.F. Sirvent, Cyclotomic Polynomials and Minimal sets of Lefschetz Periods. *J. of Difference Equations and Applications* **18** (2012), 763–783.
[22] J. Jezierski and W. Marzantowicz, *Homotopy methods in topological fixed and periodic points theory*, Topological Fixed Point Theory and Its Applications **3**, Springer, Dordrecht, 2006.
[23] S. Lang, *Algebra*, Addison–Wesley, 1971.
[24] T.Y. Li and J. Yorke, Period three implies chaos, *Amer. Math. Monthly* **82** (1975), 985–992.
[25] J. Llibre, Lefschetz numbers for periodic points, *Contemporary Math.* **152** Amer. Math. Soc., Providence, RI, (1993), 215–227.
[26] J. Llibre and V.F. Sirvent, Minimal sets of periods for Morse–Smale diffeomorphisms on orientable compact surfaces, *Houston J. of Math.* **35** (2009), 835–855.
[27] J. Llibre and V.F. Sirvent, Erratum: Minimal sets of periods for Morse–Smale diffeomorphisms on orientable compact surfaces, *Houston J. of Math.* **36** (2010), 335–336.
[28] J. Llibre and V.F. Sirvent, Minimal sets of periods for Morse-Smale diffeomorphisms on non-orientable compact surfaces without boundary, *J. of Difference Equations and Applications*, 19 (2013), 402-417.
[29] J. Llibre and V.F. Sirvent, $C^1$ Self-maps on closed manifolds with all their periodic points hyperbolic, pre-print.
[30] T. Matsuoka, The number of periodic points of smooth maps, *Erg. Th. & Dyn. Sys.* **9** (1989), 153–163.
[31] C. Narasimhan, The periodic behaviour of Morse–Smale diffeomorphisms on compact surfaces, *Trans. Amer. Soc.* **248** (1979), 145–169.
[32] J. Palis and W. de Melo, *Geometric Theory of Dynamical Systems, An Introduction*, Springer Verlag, New York, 1982.
[33] J. Palis and S. Smale, Structural stability theorems, *Proc. Sympos. Pure Math.* **14**, Amer. Math. Soc., Providence, R.I., 1970, 223–231.
[34] C. Robinson, *Dynamical Systems: Stability, Symbolic Dynamics and Chaos*, Second Edition, CRC Press, Boca Raton, 1999.
[35] M. Shub, Morse–Smale diffeomorphisms are unipotent on homology, *Dynamical systems (Proc. Sympos., Univ. Bahia, Salvador, 1971)*, Academic Press, New York, 1973.
[36] M. Shub and D. Sullivan, Homology theory and dynamical systems, *Topology* **14** (1975), 109–132.
[37] S. Smale, Differentiable dynamical systems, *Bull. Amer. Math. Soc.* **73** (1967), 747–817.
[38] J.W. Vick, *Homology theory. An introduction to algebraic topology*, Springer–Verlag, New York, 1994.

[1] Departament de Matemàtiques. Universitat Autònoma de Barcelona, Bellaterra, 08193 Barcelona, Catalunya, Spain
*E-mail address*: jllibre@mat.uab.cat
*URL*: http://www.gsd.uab.cat/personal/jllibre

[2] Departamento de Matemáticas, Universidad Simón Bolívar, Apartado 89000, Caracas 1086-A, Venezuela
*E-mail address*: vsirvent@usb.ve
*URL*: http://www.ma.usb.ve/~vsirvent

# THE ENTROPY OF AN INVARIANT PROBABILITY FOR THE SHIFT ACTING ON ONE-DIMENSIONAL SPIN LATTICES IS NON-POSITIVE

A. O. LOPES, J. MENGUE, J. MOHR AND R. R. SOUZA

Instituto de Matemática, UFRGS - Porto Alegre, Brasil

We consider a compact metric space $M$ as the state space and we generalize several results of the classical theory of Thermodynamic Formalism. We analyze the shift acting on $M^{\mathbb{N}}$ and consider a fixed probability $\nu$ on $M$ as the a-priori probability. Then, we can define the Transfer (Ruelle) operator and analyze its properties. We study potentials $A$ which can depend on the infinite set of coordinates in $M^{\mathbb{N}}$. We define entropy and by its very nature it is always a nonpositive number. The concepts of entropy and transfer operator are linked. There exist Gibbs states with arbitrary negative value of entropy. Invariant probabilities with support in a fixed point will have entropy equal to minus infinity. The infinite product of $dx$ on $(S^1)^{\mathbb{N}}$ will have zero entropy.

## 1. Introduction

This is a survey paper on the one-dimensional spin lattice model. The proofs of the results presented here appear in [19].

Let $M$ be a metric space, $d_1$ its metric and finally the metric in $M^{\mathbb{N}}$ given by: $d(x,y) = \sum_{n=1}^{\infty} \frac{1}{2^n} d_1(x_n, y_n)$, where $x = (x_1, x_2, ...)$ and $y = (y_1, y_2, ...)$. Note that $\mathcal{B} := M^{\mathbb{N}}$ is compact by Tychonoff´s theorem.

We denote by $H_\alpha$ the set of $\alpha$-Hölder functions $A : \mathcal{B} \to \mathbb{R}$ with the norm $\|A\|_\alpha = \|A\| + |A|_\alpha$, where $\|A\| = \sup_{x \in \mathcal{B}} |A(x)|$ and $|A|_\alpha = \sup_{x \neq y} \frac{|A(x)-A(y)|}{d(x,y)^\alpha}$. $\sigma : \mathcal{B} \to \mathcal{B}$ denotes the shift map which is defined by $\sigma(x_1, x_2, x_3, ...) = (x_2, x_3, x_4, ...)$.

We call the general one-dimensional spin lattice the space $M^{\mathbb{N}}$ and we consider stationary (invariant) probabilities for the shift.

We point out that a Holder potential $A$ defined on $M^{\mathbb{Z}}$ is coboundary with a potential in $M^{\mathbb{N}}$ (same proof as in [26]). In this way the Statistical Mechanics of interactions on $M^{\mathbb{Z}}$ can be understood via the analysis of the similar problem in $M^{\mathbb{N}}$.

We consider a fixed probability $\nu$ on the Borel sigma algebra of $M$. We assume that the support of $\nu$ is *equal* to the set $M$. Note that from our hypothesis if $x_0$ is isolated then $\nu(x_0) > 0$. We stress the crucial point: $\nu$ needs to be a *probability measure*, not only a *measure*.

Let $\mathcal{C}$ be the space of continuous functions from $\mathcal{B}$ to $\mathbb{R}$.

For a fixed potential $A \in H_\alpha$ we define a Transfer Operator (also called Ruelle operator) $\mathcal{L}_A : \mathcal{C} \to \mathcal{C}$ by the rule

$$\mathcal{L}_A(\varphi)(x) = \int_{S^1} e^{A(ax)} \varphi(ax) d\nu(a)\,,$$

*Date*: February, 2013.

where $x \in \mathcal{B}$ and $ax = (a, x_1, x_2, ....)$ denote a pre-image of $x$ with $a \in S^1$.

The so called one-dimensional $XY$ model (see [33],[10]) is considered in several applications to real problems in Physics. It is a particular case when $M = S^1$ and $\nu$ is the Lebesgue probability on $S^1$. The spin in each site of the lattice is described by an angle from $[0, 2\pi)$. In the Physics literature, as far as we know, the potential $A$ depends on two coordinates. A well known example in applications is the potential $A(x) = A(x_0, x_1) = \cos(x_1 - x_0 - \alpha) + \gamma \cos(2x_0)$. We consider here potentials which can depend on the all string $x = (x_1, x_2, ....)$ but which are in the Holder class. Rigorous mathematical results in this topic are considered in [20] and [2].

There are several possible points of view for understanding Gibbs states in Statistical Mechanics (see [31], [28] for interesting discussions). We prefer the transfer operator method because we believe that the eigenfunctions and eigenprobabilities (which can be derived from the theory) allow a more deep understanding of the problem. For example, the information one can get from the main eigenfunction (defined in the whole lattice) is worthwhile, mainly in the limit when temperature goes to zero.

**Examples:**

Now we give a brief description of some other examples that fit in our setting. The last example will be explained in details in section 4.

- If the alphabet is given by $M = \{1, 2, ..., d\}$, and the a-priori measure is given by $\nu = \frac{1}{d}\sum_{i=1}^{d} \delta_i$, then we have the original full shift in a finite set of $d$ symbols and the transfer operator is the classical Ruelle operator associated to a potential $A - \log(d)$ (see for example [26] and [17]). More precisely

$$\mathcal{L}_A(\varphi)(x) = \int_M e^{A(ax)}\varphi(ax)d\nu(a) = \sum_{a\in\{1,2,...,d\}} e^{A(ax)-\log(d)}\varphi(ax).$$

  If we change the a-priori measure to $\nu = \sum_{i=1}^{d} p_i.\delta_i$, where $p_i > 0$, and $\sum_{i=1}^{d} p_i = 1$, then

$$\mathcal{L}_A(\varphi)(x) = \sum_{a\in\{1,2,...,d\}} e^{A(ax)}\varphi(ax)p_a = \sum_{a\in\{1,2,...,d\}} e^{A(ax)+\log(p_a)}\varphi(ax)$$

  is the classical Ruelle operator with potential $A + \log(P)$, where $P(x_1, x_2, ...) = p_{x_1}$.

- If $M_0 = \{z_i, i \in \mathbb{N}\}$ is a countable infinite subset of $S^1$, where each point is isolated, and there is only one accumulating point $z_\infty \in S^1 \backslash M_0$, then $M = M_0 \cup \{z_\infty\}$ is a compact set. In this case $M$ can be identified with $\mathbb{N}$, where a special point $z_\infty$ plays the role of infinity (that is, a one-point compactification). We consider here the restricted distance we get from $S^1$ in $M$. If $\sum_{i\in\mathbb{N}} p_i = 1$ with $p_i \geq 0$ and $\nu = \sum_{i\in\mathbb{N}} p_i\delta_{z_i}$ then $\nu$ is supported on the whole $M$, but $z_\infty$ is not an atom for $\nu$. The Thermodynamic Formalism with state space $\mathbb{N}$, or $\mathbb{Z}$, is

considered for example in [30],[31],[6],[25]. We will analyze in section 4 some of these results on the present setting.

Our main purpose here is to describe a general theory for the Statistical Mechanics of one-dimensional spin lattices. We point out that most of the papers on the subject assume that the potential $A$ depends just on two (or, a finite number of) coordinates (as for instance is the case of [1],[3], [15]). We consider potentials which can depend on the infinite set of coordinates in $M^{\mathbb{N}}$.

In section 3 we consider the entropy, pressure and Variational Principle and its relations with eigenfunctions and eigenprobabilities of the Ruelle operator. This setting, as far as we know, was not considered before. In this case the entropy, by its very nature, is always a nonpositive number. Invariant probabilities with support in a fixed point will have entropy equal to minus infinity. The infinite product of $d\nu$ on $M^{\mathbb{N}}$ will have zero entropy. We point out that, although at first glance, the fact that the entropy we define here is negative may look strange, our definition is the natural extension of the concept of Kolomogorov entropy. In the classical case, the entropy is positive because the a-priori measure is not a probability: is the counting measure. We will explain later carefully this point.

## 2. The Ruelle operator

Let $a^n$ be an element of $M^n$ having coordinates $a^n = (a_n, a_{n-1}, \ldots, a_2, a_1)$, we denote by $a^n x \in \mathcal{B}$ the concatenation of $a^n \in M^n$ with $x \in \mathcal{B}$, i.e., $a^n x = (a_n, \ldots, a_1, x_1, x_2, \ldots)$. In the case of $n = 1$ we will write $a := a^1 \in S^1$, and $ax = (a, x_1, x_2, \ldots)$.

The $n$-th iterate of $\mathcal{L}_A$ has the following expression $\mathcal{L}_A^n(\varphi)(x) = \int_{M^n} e^{S_n A(a^n x)} \varphi(a^n x)(d\nu(a))^n$, where $S_n A(a^n x) = \sum_{k=0}^{n-1} A(\sigma^k(a^n x))$.

**Theorem 1.** *Let us fix $A \in H_\alpha$, then there exists a strictly positive Hölder eigenfunction $\psi_A$ for $\mathcal{L}_A : \mathcal{C} \to \mathcal{C}$ associated to a strictly positive eigenvalue $\lambda_A$. This eigenvalue is simple, which means the eigenfunction is unique (modulo multiplication by constant).*

We say that a potential $B$ is normalized if $\mathcal{L}_B(1) = 1$, which means it satisfies $\int_{S^1} e^{B(ax)} da = 1\,, \forall x \in \mathcal{B}\,.$

Let $A \in H_\alpha$, $\psi_A$ and $\lambda_A$ given by theorem 1, it is easy to see that

$$\int_{S^1} \frac{e^{A(ax)} \psi_A(ax)}{\lambda_A \psi_A(x)}\, d\nu(a) = 1\,,\, \forall x \in \mathcal{B}\,. \tag{1}$$

Therefore we define the normalized potential $\bar{A}$ associated to $A$, as

$$\bar{A} := A + \log \psi_A - \log \psi_A \circ \sigma - \log \lambda_A, \tag{2}$$

where $\sigma : \mathcal{B} \to \mathcal{B}$ is the shift map. As $\psi_A \in H_\alpha$ we have that $\bar{A} \in H_\alpha$.

We say a probability measure $\mu$ is invariant, if for any Borel set $B$, we have that $\mu(B) = \mu(\sigma^{-1}(B))$. We denote by $\mathcal{M}_\sigma$ the set of invariant probability measures.

We note that $\mathcal{B}$ is a compact metric space and by the Riesz Representation Theorem, a probability measure on the Borel sigma-algebra is identified with a positive linear functional $L : \mathcal{C} \to \mathbb{R}$ that sends the constant function 1 to the real number 1. We also note that $\mu \in \mathcal{M}_\sigma$ if and only if, for any $\psi \in \mathcal{C}$ we have $\int_{\mathcal{B}} \psi\, d\mu = \int_{\mathcal{B}} \psi \circ \sigma\, d\mu\,.$

We define the dual operator $\mathcal{L}_A^*$ on the space of Borel measures on $\mathcal{B}$ as the operator that sends a measure $\mu$ to the measure $\mathcal{L}_A^*(\mu)$, defined by $\int_{\mathcal{B}} \psi\, d\mathcal{L}_A^*(\mu) = \int_{\mathcal{B}} \mathcal{L}_A(\psi)\, d\mu\,,$ for any $\psi \in \mathcal{C}$.

**Theorem 2.** *Let $A$ be a Hölder continuous potential, not necessarily normalized, $\psi_A$ and $\lambda_A$ the eigenfunction and eigenvalue given by the Theorem 1. We associate to $A$ the normalized potential $\bar{A} = A + \log \psi_A - \log \psi_A \circ \sigma - \log \lambda_A$. Then*

*(a) there exists an unique fixed point $\mu_A$ for $\mathcal{L}_{\bar{A}}^*$, which is a $\sigma$-invariant probability measure;*

*(b) the measure $\rho_A = \frac{1}{\psi_A}\, \mu_A$ satisfies $\mathcal{L}_A^*(\rho_A) = \lambda_A \rho_A$. Therefore, $\rho_A$ is an eigenmeasure for $\mathcal{L}_A^*$;*

*(c) for any Hölder continuous function $w : \mathcal{B} \to \mathbb{R}$, we have that, in the uniform convergence topology, $\frac{\mathcal{L}_{\bar{A}}^n(w)}{(\lambda_A)^n} \to \psi_A \int_{\mathcal{B}} w\, d\rho_A$ and $\mathcal{L}_{\bar{A}}^n \omega \to \int_{\mathcal{B}} \omega d\mu_A$, where $\mathcal{L}_{\bar{A}}^n$ denotes the $n$-th iterate of the operator $\mathcal{L}_A : H_\alpha \to H_\alpha$.*

We call $\mu_A$ the **Gibbs probability** (or, Gibbs state) for $A$. We will leave the term **equilibrium probability** (or, equilibrium state) for the one which maximizes pressure. As we will see, this invariant probability measure over $\mathcal{B}$ describes the statistics in equilibrium for the interaction described by the potential $A$. The assumption that the potential is Hölder implies that the decay of iteraction is fast.

**Proposition 1.** *The only Holder continuous eigenfunction $\psi$ of $\mathcal{L}_A$ which is totally positive is $\psi_A$.*

**Proposition 2.** *Suppose $\bar{A}$ is normalized, then the eigenvalue $\lambda_{\bar{A}} = 1$ is maximal. Moreover, the remainder of the spectrum of $\mathcal{L}_{\bar{A}} : H_\alpha \to H_\alpha$ is contained in a disk centered at zero with radius strictly smaller than one.*

**Proposition 3.** *If $v, w \in \mathcal{L}^2(\mu_A)$ are such that $w$ is Hölder and $\int w\, d\mu_A = 0$, then, there exists $C > 0$ such that for all $n$ $\int (v \circ \sigma^n)\, w\, d\mu_A \leq C\, (\lambda_{\bar{A}}^1)^n$. In particular $\mu_A$ is mixing and therefore ergodic.*

## 3. Entropy and Variational Principle

In this section (which was taken from [19]) we will introduce a notion of entropy. Initially, this will be done only for Gibbs probabilities, and then we will extend this definition to invariant probabilities. After that we prove that the Gibbs probability obtained in the general setting above satisfies a variational principle.

We point out that any reasonable concept of entropy must satisfy two principles: the entropy of probabilities with support in periodic orbits should be minimal and the entropy of the independent probability should be maximal. This will happen for our definition.

**Definition 1.** *We denote by $\mathcal{G}$ the set of Gibbs measures, which means the set of $\mu \in \mathcal{M}_\sigma$, such that, $\mathcal{L}_B^*(\mu) = \mu$, for some normalized potential $B \in H_\alpha$. We define the entropy of $\mu \in \mathcal{G}$ as $h(\mu) = -\int_{\mathcal{B}} B(x) d\mu(x)$.*

One can show that $-\int B\, d\mu$ is the infimum of $\left\{ -\int A\, d\mu + \log(\lambda_A) \,:\, A \in H_\alpha \right\}$.

The above definition which appears in [19] is different from the one briefly mentioned in section 3 in [2].

**Proposition 4.** *If $\mu \in \mathcal{G}$, then we have $h(\mu) \leq 0$.*

This follow from Jensen's inequality.

It is easy to see that the Gibbs state $(d\nu)^{\mathbb{N}}$ has zero entropy.

Now we state a lemma that is used to prove the main result of this section, namely, the variational principle of Theorem 3. This lemma was shown to be true in the the classical Bernoulli case in [22].

**Lemma 1.** *Let us fix a Hölder continuous potential $A$ and a measure $\mu \in \mathcal{G}$ with associated normalized potential $B$. We call $\mathcal{C}^+$ the space of continuous positive functions on $\mathcal{B}$. Then, we have*

$$h(\mu) + \int_{\mathcal{B}} A(x) d\mu(x) = \inf_{u \in \mathcal{C}^+} \left\{ \int_{\mathcal{B}} \log \left( \frac{\mathcal{L}_A u(x)}{u(x)} \right) d\mu(x) \right\}.$$

**Definition 2.** *Let $\mu$ be an invariant measure. We define the entropy of $\mu$ as*

$$h(\mu) = \inf_{A \in H_\alpha} \left\{ - \int_{\mathcal{B}} A \, d\mu + \log \lambda_A \right\},$$

*where $\lambda_A$ is the maximal eigenvalue of $\mathcal{L}_A$, given by theorem 1.*

This value is non positive and can be $-\infty$ as we will se later.

**Definition 3.** *Given a Hölder potential $A$ we call the pressure of $A$ the value*

$$P(A) = \sup_{\mu \in \mathcal{M}_\sigma} \left\{ h(\mu) + \int_{\mathcal{B}} A(x) d\mu(x) \right\}.$$

*A probability which attains such maximum value is called equilibrium state for $A$.*

**Theorem 3** (Variational Principle)**.** *Let $A \in H_\alpha$ be a Hölder continuous potential and $\lambda_A$ be the maximal eigenvalue of $\mathcal{L}_A$, then*

$$\log \lambda_A = P(A) = \sup_{\mu \in \mathcal{M}_\sigma} \left\{ h(\mu) + \int_{\mathcal{B}} A(x) d\mu(x) \right\}.$$

*Moreover the supremum is attained on the Gibbs measure, i.e. the measure $\mu_A$ that satisfies $\mathcal{L}^*_{\bar{A}}(\mu_A) = \mu_A$.*

Therefore, the Gibbs state and the equilibrium state for $A$ are given by the same measure $\mu_A$, which is the unique fixed point for the dual Ruelle operator associated to the normalized potential $\bar{A}$.

The analyticity of the variation of eigenvalue with respect to the potential can be obtained from results considered in [32].

**Theorem 4** (Pressure as Minimax)**.** *Given a Hölder potential $A$*

$$P(A) = \sup_{\mu \in \mathcal{M}_\sigma} \left[ \inf_{u \in \mathcal{C}^+} \left\{ \int_{\mathcal{B}} \log \left( \frac{\mathcal{L}_A u(x)}{u(x)} \right) d\mu(x) \right\} \right].$$

**Remark:** The entropy of a probability measure supported on periodic orbit can be $-\infty$. Indeed, suppose $M = [0,1]$, and $A_c : [0,1]^{\mathbb{N}} \to \mathbb{R}$ given by $A_c(x) = \log \left( \frac{c}{1-e^{-c}} e^{-cx_1} \right)$. We have that for each $c > 0$, the function $A_c$ is a $C^1$ normalized potential (therefore belongs to $H_\alpha$), which depends only on the first coordinate of $x$. Note that $\mathcal{L}_{A_c}(1) = 1$. Let $\mu$ be the Dirac Measure on $0^\infty$. We have $h(\mu) \leq -\int A_c d\mu = -A_c(0^\infty) = -\log \left( \frac{c}{1-e^{-c}} \right) \to -\infty$ when $c \to \infty$. This shows that $h(\mu) = -\infty$. An easy adaptation of the arguments can be done to prove that, in this setting, invariant measures supported on periodic orbits have entropy $-\infty$.

**Relations with Kolmogorov Entropy:**

Let us consider the construction of the entropy by partitions method, in the case $M$ is finite. We begin by remembering that, by the Kolmogorov-Sinai Theorem, the classical entropy of $\mu$, which we will denote by $H(\mu)$, is given by

$$H(\mu) = \lim_{n\to\infty} -\frac{1}{n} \sum_{i_1,\dots,i_n} \mu([i_1...i_n]) \log\left(\mu([i_1...i_n])\right). \tag{3}$$

**Proposition 5.** *Let $M = \{1, ..., d\}$ and $\nu = \sum_{i=1}^{d} p_i \delta_i$ be the a-priori probability on $M$. For any Gibbs measure $\mu$:*

(a)

$$H(\mu) = h^{\nu}(\mu) - \sum_{i=1}^{d} \log(p_i).\mu([i]),$$

(b)

$$h^{\nu}(\mu) = -\lim_{n\to\infty} \frac{1}{n} \sum_{i_1,\dots,i_n} \mu([i_1...i_n]) \log\left(\frac{\mu([i_1...i_n])}{p_{i_1}...p_{i_n}}\right)$$

*where*

$$[i_1...i_n] = \{x \in M^{\mathbb{N}} : x_1 = i_1, ..., x_n = i_n\}.$$

In particular, it follows from item (a) above that, when $p_i = \frac{1}{d}$, for all $i$, we have

$$h^{\nu}(\mu) = H(\mu) - \log(d).$$

The above proposition can be interpreted in the following way: in the classical definition of Kolmogorov entropy it is considered the a-priori measure $\nu = \sum_{i=1}^{\infty} \delta_i$ on $M$, which is not a probability. The expression above shows that in a consistent way $h^{\nu}(\mu) \leq 0$ (with $\nu = \sum_{i=1}^{d} p_i\, \delta_i$) and $H(\mu) \geq 0$. There is no contradiction, it is just a different point of view. We point out that in the case the state space is not countable, then, it is definition $h^{\nu}(\mu)$ which makes sense.

**Markov Chains with values on $S^1$:**

Now we recall the concept of Markov measures and show that the entropy defined above is an extension of the concept of entropy for Markov measures, as introduced in [20].

Let $K : M^2 \to \mathbb{R}$, $\theta : M \to \mathbb{R}$, satisfying

$$\int_M K(x_1, x_2) d\nu(x_2) = 1,\ \forall x_1 \quad \text{and} \quad \int_M \theta(x_1) K(x_1, x_2) d\nu(x_1) = \theta(x_2)\ , \forall x_2\,. \tag{4}$$

We call $K$ a transition kernel and $\theta$ the stationary measure for $K$. As in [20], we define the absolutely continuous Markov measure associated to $K$ and $\theta$, as

$$\mu(A_1...A_n \times M^{\mathbb{N}}) := \int_{A_1...A_n} \theta(x_1)\, K(x_1, x_2)...K(x_{n-1}, x_n)\, d\nu(x_n)...d\nu(x_1), \tag{5}$$

for any cylinder $A_1...A_n \times M^{\mathbb{N}}$.

The next proposition show us the importance of a.c. Markov measures:

**Proposition 6.** a) *Given a Hölder continuous potential $A(x_1, x_2)$ (not necessarily normalized) depending on two coordinates, there exists a Markov measure that is Gibbs for $A$.*

b) *The converse is also true: given an absolutely continuous Markov measure defined by $K$ and $\theta$, there exists a certain Hölder continuous normalized potential $A(x_1, x_2)$, such that the Markov measure defined by $\theta$ and $K$ is the Gibbs measure for $A$.*

Therefore, any a.c. Markov measure is Gibbs for a potential depending on two variables, and conversely, any potential depending on two variables has a Gibbs measure which is an a.c. Markov Measure.

In other words, if we restrict our analysis to potentials that depend just on the first two coordinates, we have that the set of a.c. Markov Measures coincides with the set of Gibbs measures.

**Proof:**

(a) Given a potential $A(x_1, x_2)$, non-normalized, as in [20] define $\theta_A : M \to \mathbb{R}$ by

$$\theta_A(x_1) := \frac{\psi_A(x_1)\ \bar{\psi}_A(x_1)}{\pi_A}, \tag{6}$$

and a transition $K_A : M^2 \to \mathbb{R}$ by

$$K_A(x_1, x_2) := \frac{e^{A(x_1,x_2)}\ \bar{\psi}_A(x_2)}{\bar{\psi}_A(x_1)\,\lambda_A}\,, \tag{7}$$

where $\psi_A$ and $\bar{\psi}_A$ are the eigenfunctions associated to the maximal eigenvalue $\lambda_A$ of the operators

$$L_A\psi(x_2) = \int_M e^{A(x_1,x_2)}\,\psi(x_1)d\nu(x_1) \quad \text{and} \quad \bar{L}_A\psi(x_1) = \int_M e^{A(x_1,x_2)}\,\psi(x_2)d\nu(x_2)$$

and $\pi_A = \int_M \psi_A(x_1)\bar{\psi}_A(x_1)d\nu(x_1)$.

Then, by the same arguments used to prove theorem 16 of [2], we obtain that the Markov measure $\mu_A$ defined by (5) (considering $K_A$ and $\theta_A$) is Gibbs for $A$, i.e. a fixed point for the dual Ruelle operator $\mathcal{L}^*_{\bar{A}}$, where $\bar{A} = A + \log\psi_A(x_1) - \log\psi_A(x_2) - \log\lambda_A$.

(b) Let $K$ and $\theta$ satisfying (4), and define $A = \log K$, we have $\bar{L}_A(1) = 1$ which implies $\lambda_A = 1$ and $\bar{\psi}_A = 1$. Let $\psi_A$ be maximal eigenfunction for $L_A$.

Using (7), we get $K_A(x_1, x_2) = e^{A(x_1,x_2)} = K(x_1, x_2)$. Define $\theta_A = \frac{\psi_A}{\pi_A}$. We have that $\theta_A$ is an invariant density for $K$, therefore $\theta_A = \theta$. Then, also by theorem 16 page of [2], we have that the Markov measure defined by $K$ and $\theta$ is Gibbs for $A$. □

Next proposition shows that the concept of entropy introduced in 2 is a generalization of the concept of entropy defined in [20], which could only be applied to a.c. Markov measures:

**Proposition 7.** *Let $\mu$ be the Markov measure defined by a transition kernel $K$ and a stationary measure $\theta$, given in* (5)*. The definition of entropy given in* [20]*:*

$$S(\theta K) = -\int_{M^2} \theta(x_1)K(x_1, x_2)\log(K(x_1, x_2))d\nu(x_1)d\nu(x_2) \leq 0$$

*coincides with the present definition 2.*

## 4. An application to the non-compact case

An interesting example of application of the above theory is the following: consider $M_0 = \{z_i, i \in \mathbb{N}\}$ an increasing infinite sequence of points in $[0, 1)$ and suppose that $z_\infty := 1 = \lim_{i\to\infty} z_i$. We will also suppose $z_1 = 0$. Therefore, each point of $M_0$ is isolated, and there is only one accumulating point $z_\infty = 1$. We consider the induced euclidean metric then $M = M_0 \cup \{1\}$ is a compact set. The state space $M_0$ can be identified with $\mathbb{N}$, and $M$ has a special point $z_\infty = 1$ playing the role of the infinity. Let $\mathcal{B}_0 = M_0^{\mathbb{N}}$ and $\mathcal{B} = M^{\mathbb{N}}$. Note that $\mathcal{B}_0$ is not compact.

Some results in Thermodynamic Formalism for the shift with countable symbols (see [30] [6]) can be recovered from our previous results as we will see.

The main point here is that we can take advantage of the previous results on $\mathcal{B} = M^{\mathbb{N}}$, but in the end, the measures we get need to have support on $\mathcal{B}_0 = M_0^{\mathbb{N}}$.

**Lemma 2.** *Suppose that $A : \mathcal{B}_0 \to \mathbb{R}$ is a Hölder continuous potential. Then it can be extended as a Hölder continuous function $A : \mathcal{B} \to \mathbb{R}$.*

Now let us fix an a-priori measure $\nu := \sum_{i\in\mathbb{N}} p_i \delta_{z_i}$ on $M$ (or $M_0$), where $p_i > 0$ and $\sum_{i\in\mathbb{N}} p_i = 1$. In fact, we have that $z_\infty = 1$ belongs to the support of $\mu$, but is not an atom of $\mu$. All other points of $M$ (i.e. the points of $M_0$) are atoms for $\nu$. On this way for each Hölder continuous potential $A : \mathcal{B}_0 \to \mathbb{R}$ we can consider the following Transfer Operator on $\mathcal{C}(\mathcal{B}_0)$:

$$\mathcal{L}_A(w)(x) := \int_M e^{A(ax)} w(ax) d\nu(a) = \sum_{i\in\mathbb{N}} e^{A(z_i x)} w(z_i x) p_i.$$

Using last lemma and the results of previous sections one can show:

**Proposition 8.** *Let $A : \mathcal{B}_0 \to \mathbb{R}$ be a Hölder potential. Then*

*(a) there exists a positive number $\lambda_A$ and a positive Hölder function $\psi_A : \mathcal{B}_0 \to \mathbb{R}$, such that, $\mathcal{L}_A \psi_A = \lambda_A \psi_A$.*

*If we consider the normalized potential $\bar{A} = A + \log \psi_A - \log \psi_A \circ \sigma - \log \lambda_A$, then*

*(b) there exists an unique fixed point $\mu_A$ for $\mathcal{L}^*_{\bar{A}}$, which is a $\sigma$-invariant probability measure on $\mathcal{B}_0$.*

*(c) the measure*

$$\rho_A = \frac{1}{\psi_A}\, \mu_A$$

*satisfies $\mathcal{L}^*_A(\rho_A) = \lambda_A \rho_A$. Therefore, $\rho_A$ is an eigen-measure for $\mathcal{L}^*_A$.*

*(d) for any Hölder function $w : \mathcal{B}_0 \to \mathbb{R}$, we have that, in the uniform convergence topology,*

$$\frac{\mathcal{L}^n_A(w)}{(\lambda_A)^n} \to \psi_A \int_{\mathcal{B}_0} w \, d\rho_A,$$

*and*

$$\mathcal{L}^n_{\bar{A}} \omega \to \int_{\mathcal{B}_0} \omega d\mu_A \, .$$

Now let us compare this setting with some results contained in [30]. The operator $\mathcal{L}_A$ can be written as

$$\mathcal{L}_A(w)(x) = \sum_i e^{A(z_i x)} w(z_i x) p_i = \sum_i e^{A(z_i x) + \log(p_i)} w(z_i x),$$

that is, the Classical Ruelle Operator with potential $B := A+\log(P)$, where $P(y_1, y_2, y_3, ...) = P(y_1) = p_i$, if, $y_1 = z_i$. We denote this operator by $L_B$, or, $L_{A+\log(P)}$.

Clearly $(A + \log(P))(z_i, y_2, y_3, ...) \to -\infty$, when $i \to +\infty$, because $p_i \to 0$, when, $i \to +\infty$. Furthermore, if we define

$$Var_n(B) = \sup\{|B(x) - B(y)| : x_1 = y_1, ..., x_n = y_n\},$$

then, there exists $C > 0$, such that, $Var_n(B) \leq C\frac{1}{2^{n\alpha}}$, for any $n \geq 1$. This means that $B$ is **locally Hölder continuous** (see [30]).

Define

$$Z_n(B, a) := \sum_{\substack{\sigma^n(y) = y \\ y_1 = a}} e^{S_n B(y)}.$$

**Proposition 9.** *Fix $a \in M_0$, then, there exists a constant $M_a$ and an integer $N_a$, such that, for any $n > N_a$:*

$$\frac{Z_n(B, a)}{(\lambda_A)^n} \in [M_a^{-1}, M_a]$$

In this way, we can say that $B = A+\log(P)$ is **positive recurrent** (see [30] Definition 2). Following [30] Theorem 4 we get a Ruelle-Perron-Frobenius Theorem (as in Theorem 8 above). It follows from the above proposition that $\lambda_A$ is the Gurevic pressure of $B$ (see [30] definition 1).

We would like to point out some differences on the topology considered in our setting with the classical one used in the theory of Thermodynamic Formalism with state space $\mathbb{N}$. The set $M_0^{\mathbb{N}}$ can be identified with $\mathbb{N}^{\mathbb{N}}$, but the metric space $M_0^{\mathbb{N}}$ is different from the metric space $\mathbb{N}^{\mathbb{N}}$ with the discrete product topology. Here, we consider a distance (induced in the subset $M_0 \cup \{z_\infty\} \subset [0, 1]$), such that, for any two points $x = (x_1, x_2, ...)$, $y = (y_1, y_2, ...) \in M_0^{\mathbb{N}}$

$$d(x, y) = \sum_{n\in\mathbb{N}} \frac{1}{2^n} d_{[0,1]}(x_n, y_n).$$

On the other hand, the metric considered in [30] is of the form: for two points $x, y \in \mathbb{N}^{\mathbb{N}}$

$$\widetilde{d}(x, y) = \frac{1}{2^n}, \quad \text{if} \quad x_1 = y_1, ..., x_{n-1} = y_{n-1}, \ x_n \neq y_n.$$

Using that the diameter of $[0, 1]$ is one, it follows that $d(x, y) \leq \widetilde{d}(x, y)$. In particular, any convergent sequence on the metric $\widetilde{d}$ is a convergent sequence on the metric $d$, and any continuous/Hölder function $A$ for the metric $d$ is a continuous/Hölder function for the metric $\widetilde{d}$. But the same is not true in the opposite direction. This is a subtle question. Results in [30] and here are obtained under slight different hypothesis. Anyway, in physical applications this is probably a not very important point.

Considering the dual space, it follows from the relation $d(x, y) \leq \widetilde{d}(x, y)$ that any open set for the metric $d$ is an open set for the metric $\widetilde{d}$. Then, the Borel sigma-algebra generated by $d$ is contained in the Borel sigma-algebra generated by $\widetilde{d}$. on the order hand, the cylinder sets [30] are closed sets for the metric $d$, therefore, they belong to the sigma-algebra generated by $d$. In this way, the Borel sigma-algebra generated by $d$, or, by $\widetilde{d}$, is the same.

## References


[1] M. Asaoka, T. Fukaya, K. Mitsui and M. Tsukamoto, Growth of critical points in one-dimensional lattice systems, preprint Arxiv 2012.

[2] A. T. Baraviera, L. Cioletti, A. O. Lopes, J. Mohr and R. R. Souza, On the general one-dimensional XY model: positive and zero temperature, selection and non-selection. *Rev.Math. Phis.* 23 (2011),no. 10, 1063-1113, 82Bxx.

[3] M. Bertelson, Topological invariant for discrete group actions, Lett. Math. Phys. 62 (2004), 147-156

[4] R. Bissacot and E. Garibaldi, Weak KAM methods and ergodic optimal problems for countable Markov shifts, *Bull. Braz. Math. Soc.* 41, N 3, 321-338, 210.

[5] R. Bissacot and R. Freire Jr., On the existence of maximizing measures for irreducible countable Markov shifts: a dynamical proof, to appear in Erg. Theo. and Dyn. Syst.

[6] Y. Daon, Bernoullicity of equilibrium measures on countable Markov shifts, preprint Arxiv 2012.

[7] A. C. D. van Enter, R. Fernandez and A. D. Sokal, Regularity properties and pathologies of position-space renormalization-group transformations: Scope and limitations of Gibbsian theory, *Journ. of Stat. Phys.* V.72, N 5/6 879-1187, 1993.

[8] R. Ellis, Entropy, Large Deviations, and Statistical Mechanics, Springer Verlag, 2005

[9] A. C. D. van Enter and W. M. Ruszel, Chaotic Temperature Dependence at Zero Temperature, *Journal of Statistical Physics*, Vol. 127, No. 3, 567-573, 2007.

[10] Y. Fukui and M. Horiguchi, One-dimensional Chiral $XY$ Model at finite temperature, Interdisciplinary Information Sciences, Vol 1, 133-149, N. 2 (1995)

[11] G. Gallavotti, Statistical Mechanics: A Short Treatise, Springer Verlag, (2010)

[12] H.-O. Georgii, Gibbs Measures and Phase Transitions. de Gruyter, Berlin, (1988).

[13] D. A. Gomes, A. O. Lopes and J. Mohr, The Mather measure and a large deviation principle for the entropy penalized method. Commun. Contemp. Math. 13 (2011), no.2, 235-268.

[14] D. A. Gomes and E. Valdinoci, Entropy Penalization Methods for Hamilton-Jacobi Equations, Adv. Math. (2007) 215, No. 1, 94-152.

[15] M. Gromov, Singularities, expanders and topology of maps. Part 2: From combinatorics to topology via algebraic isoperimetry. Geom. Funct. Anal. 20 (2010), no. 2, 416-526.

[16] R.B. Israel, Convexity in the theory of lattice gases, Princeton University Press, 1979.

[17] G. Keller, Gibbs States in Ergodic Theory, Cambrige Press, 1998.

[18] O. Lanford, Entropy and Equilibrium States in Classical Statistical Mechanics. Statistical mechanics and mathematical problems. Battelle Rencontres, Seattle, Wash., 1971. Lecture Notes in Physics, 20. Springer-Verlag, Berlin-New York, 1973.

[19] A. O. Lopes, J. K. Mengue, J. Mohr and R. R. Souza, Entropy and Variational Principle for one-dimensional Lattice Systems with a general a-priori measure: positive and zero temperature, Arxiv (2012)

[20] A. O. Lopes, J. Mohr, R. Souza and Ph. Thieullen, Negative entropy, zero temperature and stationary Markov chains on the interval, *Bulletin of the Brazilian Mathematical Society* **40**, 1-52, 2009.

[21] A.O. Lopes and E. Oliveira, Entropy and variational principles for holonomic probabilities of IFS, *Disc. and Cont. Dyn. Systems* series A, vol 23, N 3, 937-955, 2009.

[22] A.O. Lopes, An analogy of charge distribution on Julia sets with the Brownian motion , *J. Math. Phys.* 30 9, 2120-2124, 1989.

[23] R. Mañé, The Hausdorff dimension of invariant probabilities of rational maps. *Lecture Notes in Math.* vol.1331, 86-117, 1988.

[24] D. H. Mayer, *The Ruelle-Araki transfer operator in classical statistical mechanics*, LNP 123, Springer Verlag 1980

[25] I. D. Morris, Entropy for Zero-Temperature Limits of Gibbs-Equilibrium States for Countable-Alphabet Subshifts of Finite Type, Journ. of Statis. Physics, Volume 126, Number 2 (2007), 315-324,

[26] W. Parry and M. Pollicott. *Zeta functions and the periodic orbit structure of hyperbolic dynamics*, Astérisque Vol 187-188 1990

[27] M. Pollicott and M. Yuri, Dynamical systems and Ergodic Theory, Cambrige Press, 1998

[28] D. Ruelle, Thermodynamic Formalism, second edition, Cambridge, 2004.

[29] B. Simon, The Statistical Mechanics of Lattice Gases, Princeton Univ Press, 1993

[30] O. Sarig, Thermodynamic formalism for countable Markov shifts, *Ergodic Theory and Dynamical Systems* 19, 1565-1593, 1999

[31] O. Sarig, Lecture notes on thermodynamic formalism for topological Markov shifts, *Penn State*, 2009.

[32] E. A. da Silva, R. R. da Silva and R. R. Souza The Analyticity of a Generalized Ruelle's Operator, to appear Bull. Braz. Math. Soc.
[33] C. Thompson. *Infinite-Spin Ising Model in one dimension*. Journal of Mathematical Physics. (9): N.2 241-245, 1968.

# HOMOCLINIC TANGENCIES FROM SPIRALLING PERIODIC POINTS

C. A. MORALES

Abstract. In this short note we prove that every surface diffeomorphism with infinitely many periodic points with nonreal eigenvalues can be approximated by ones with homoclinic tangencies. This provides a converse of the result [3].

## 1. Introduction

A *surface diffeomorphisms* is a diffeomorphism of class $C^1$ of a compact connected two-dimensional Riemannian manifold. A *periodic point* of a surface diffeomorphism $f$ is a point $p$ in the surface for which there is $n \in \mathbb{N}^+$ satisfying $f^n(p) = p$. The minimal of such $n$ is the so-called *period* of $p$ denoted by $n_p$. The eigenvalues of $p$ will be those of the linear isomorphism $Df^{n_p}(p)$. We will say that $p$ is a *saddle* if it has eigenvalues of modulus less and bigger than 1 simultaneously. The invariant manifold theory [1] asserts that every saddle $p$ comes equipped with a *stable* and an *unstable* manifold formed by those points whose positive and negative orbits converge to that of $p$ respectively. A *homoclinic tangency* is a point where such manifolds have a nontransverse intersection. We will work with the so-called $C^1$ *topology* in the space of $C^1$ diffeomorphisms measuring the distance between diffeomorphisms and their corresponding derivatives.

It was recently proved that every surface diffeomorphism with homoclinic tangencies can be approximated by diffeomorphisms exhibiting periodic points with purely imaginary eigenvalues [3]. This together with the classical work by Newhouse [2] suggests that every surface diffeomorphism with homoclinic tangencies can be approximated by diffeomorphisms exhibiting infinitely many periodic points with nonreal eigenvalues. What we shall prove here is just the converse assertion, namely, that every surface diffeomorphism exhibiting infinitely many periodic points with nonreal eigenvalues can be approximated by diffeomorphisms with homoclinic tangencies. We can also compare our result with [5] proving that every $C^2$ surface diffeomorphism with infinitely many sinks or sources with unbounded periods can be $C^1$ approximated by ones with homoclinic points.

## 2. Statements and proofs

Let $f$ be a surface diffeomorphism. We define the 1-*preperiodic set* of $f$, $P_1^*(f)$, as the set of points $x$ for which there are sequences $f_n \to f$ and $x_n \to x$ such that $x_n$ is a saddle of $f_n$, $\forall n$ (c.f. [8]).

To prove our result we will need the following lemma.

**Lemma 2.1.** *The set of accumulation points of the periodic points with nonreal eigenvalues of a surface diffeomorphism $f$ is contained in $P_1^*(f)$.*

*Proof.* Take any of such accumulation points $x$. Then, there is an infinity sequence $x_n$ of periodic points with nonreal eigenvalues of $f$ with $x_n \to x$. We have three cases to

consider, namely, every $x_n$ has eigenvalues of modulus 1 or either every $x_n$ is a *sink* (i.e. its eigenvalues have modulus less than 1), or every $x_n$ is a *source* (i.e. a sink for $f^{-1}$).

In the first case, as remarked in p. 976 of [7], we can find a diffeomorphism' s sequence $g_n \to f$ and a sequence $y_n \to x$ such that $y_n$ is a saddle of $g_n$ for all $n$. Therefore, $x \in P_1^*(f)$.

In the second case, if the periods $n_{x_n}$ are bounded, we obtain that $x$ itself is a periodic point. In such a case, if $x$ has no eigenvalues of modulus 1, then $x$ must be a saddle thus $x \in P_1^*(f)$. Otherwise, $x$ has an eigenvalue of modulus 1 and then the aforementioned remark in p. 976 in [7] yields a subsequence $n_k$ of positive integers and a sequence $f_k \to f$ so that $x_{n_k}$ is a saddle of $f_k$ for all $k$. It then follows from the definition that $x \in P_1^*(f)$. Therefore, we can assume that $n_{x_n} \to \infty$. In such a case a result by Pliss (Theorem 3.1 in [6]) yields at once a sequence $n_k \in \mathbb{N}$ and a sequence $f_k \to f$ so that $x_{n_k}$ is a saddle of $f_k$. It then follows from the definition that $x \in P_1^*(f)$.

In the third case we proceed as in the second but with $f^{-1}$ instead of $f$ to obtain $x \in P_1^*(f)$. The lemma follows. □

We also need the following lemma whose proof is contained in Pujals and Sambarino [7] (see also [8]).

**Lemma 2.2.** *For every surface diffeomorphism $f$ which cannot be approximated by diffeomorphisms with homoclinic tangencies there are a neighborhood $U$ of $P_1^*(f)$ and a tangent bundle splitting $T_U M = E_U \oplus F_U$ over $U$ with $dim(E) = 1$ such that $Df(x)(E_x) = E_{f(x)}$ and $Df(x)(F_x) = F_{f(x)}$ for all $x \in U \cap f^{-1}(U)$.*

Now we can state our main result.

**Theorem 2.3.** *Every surface diffeomorphism exhibiting infinitely many periodic points with nonreal eigenvalues can be approximated by diffeomorphisms with homoclinic tangencies.*

*Proof.* Let $f$ be a surface diffeomorphism. By Lemma 2.1 we have that the set of accumulation points of the periodic points with nonreal eigenvalues is contained in $P_1^*(f)$. If $f$ cannot be approximated by diffeomorphisms with homoclinic tangencies, we obtain the splitting in the neighborhood $U$ of $P_1^*(f)$ in Lemma 2.2. Since such a splitting clearly prevents the existence of an infinite sequence of periodic points with nonreal eigenvalues converging to some point of $P_1^*(f)$, we conclude that there are no such accumulation points, and so, there exists only a finite number of such periodic points. This ends the proof. □

## References


[1] Hirsch, M., Pugh, C., Shub, M., *Invariant manifolds*, Lec. Not. in Math. 583 (1977), Springer-Verlag.

[2] Newhouse, S., Diffeomorphisms with infinitely many sinks, *Topology* 13 (1974), 9–18.

[3] Morales, C., A., Purely imaginary eigenvalues from homoclinic tangencies, *Appl. Math. Lett.* 25 (2012), no. 11, 2005–2008.

[4] Palis, J., Takens, F., *Hyperbolicity and sensitive chaotic dynamics at homoclinic bifurcations.* Fractal dimensions and infinitely many attractors. Cambridge Studies in Advanced Mathematics, 35. Cambridge University Press, Cambridge, 1993.

[5] Pujals, E., R., Sambarino, M., On the dynamics of dominated splittings, *Ann. of Math.* (2) 169 (2009), 675–740

[6] Pujals, E., R., Sambarino, M., Density of hyperbolicity and tangencies in sectional dissipative regions, *Ann. Inst. H. Poincar Anal. Non Linéaire* 26 (2009), no. 5, 1971–2000.

[7] Pujals, E., R., Sambarino, M., Homoclinic tangencies and hyperbolicity for surface diffeomorphisms, *Ann. of Math.* (2) 151 (2000), 961–1023.
[8] Wen, L., Homoclinic tangencies and dominated splittings, *Nonlinearity* 15 (2002), no. 5, 1445–1469.

Instituto de Matemática, Universidade Federal do Rio de Janeiro, P. O. Box 68530, 21945-970, Rio de Janeiro, Brazil.,
*E-mail address*: morales@impa.br

# ON THE DISCRETE BICYCLE TRANSFORMATION

S. TABACHNIKOV AND E. TSUKERMAN

## 1. Introduction

The motivation for this paper comes from the study of a simple model of bicycle motion. The bicycle is modeled as an oriented segment in the plane of fixed length $\ell$, the wheelbase of the bicycle. The motion is constrained so that the segment is always tangent to the path of the rear wheel; this non-holonomic constraint is due to the fact that the rear wheel is fixed on the frame, whereas the front wheel can steer. See [8, 10, 13, 15] and the references therein.

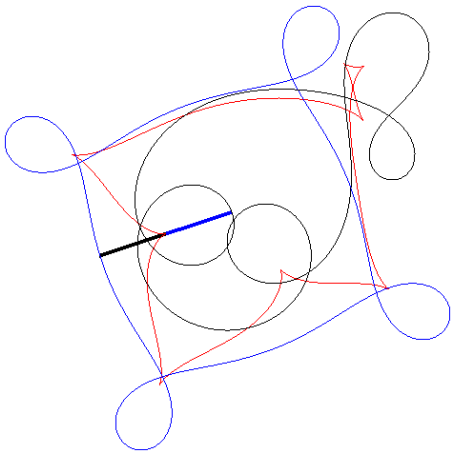

FIGURE 1. Bicycle correspondence. The cusped curve is the rear track, the two smooth curves are front tracks in the bicycle correspondence (figure courtesy of R. Perline).

If the rear wheel path $\gamma$ is prescribed, and the direction of motion is chosen, the front wheel path $\Gamma$ is constructed by drawing the tangent segments of length $\ell$ to $\gamma$. Note that the rear track may have cusp: they occur when the steering angle equals $90°$. Changing the direction of motion to the opposite yields another front track, say, $\Gamma'$. We say that the curves $\Gamma$ and $\Gamma'$ are in the *bicycle correspondence*.[1] See Figure 1.

If the front wheel path $\Gamma$ is prescribed then the rear wheel follows a constant-distance pursuit curve, and its trajectory is uniquely determined, once the initial position of the bicycle is chosen. A monodromy map $M_{\Gamma,\ell}$ arises that assigns to every initial position of the bicycle its terminal position. If $\Gamma$ is a closed curve then $M_{\Gamma,\ell}$ is a self-map of a

[1]One can also call this Darboux or Bäcklund transformation, but we shall use the "bicycle" terminology.

circle of radius $\ell$, uniquely defined up to conjugation. The bicycle monodromy $M_{\Gamma,\ell}$ is a Möbius transformation [9, 10, 13].

All of the above can be extended to the motion of a segment in higher dimensional Euclidean spaces and even Riemannian manifolds (see [12] for elliptic and hyperbolic planes). In the forthcoming paper [16], we shall discuss Liouville integrability of the bicycle transformation in dimensions 2 and 3.

In this paper, following [11, 14], we study a discrete version of the bicycle correspondence. Let $V = (V_1, V_2, \dots)$ be a polygon in $\mathbb{R}^n$, and let $V_1W_1$ be a seed segment of length $\ell$ (so now $\ell$ is twice the lenght of the bicycle frame). The next point $W_2$ is constructed in the plane spanned by $V_1, V_2, W_1$ as follows: one parallel translates point $W_1$ along the vector $V_1V_2$ to point $U$, and then reflects point $U$ in the line $W_1V_2$ to obtain a new point $W_2$. In other words, the plane quadrilateral $V_1V_2W_1W_2$ is an isosceles trapezoid with $|V_1V_2| = |W_1W_2|$ and $|V_1W_1| = |V_2W_2| = \ell$, see Figure 2. Once the point $W_2$ is constructed, one continues the process, shifting the index by one, etc.[2]

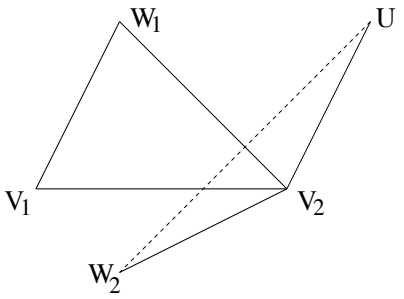


FIGURE 2. Discrete bicycle correspondence

We call the above described correspondence between polygons $V$ and $W$ the *discrete bicycle correspondence* and denote it by $\mathcal{B}_\ell(V, W)$. In the continuous limit, the polygons $V$ and $W$ become the front tire tracks $\Gamma$ and $\Gamma'$, and the discrete bicycle correspondence becomes the above described bicycle correspondence between smooth curves.

Our ultimate goal is to establish Liouville integrability of the discrete bicycle correspondence and to describe its dynamics in detail. In this paper, we make steps in this direction. Let us list basic properties of the discrete bicycle correspondence.

Let $V$ be a closed $k$-gon in $\mathbb{R}^n$ (that is, $V_{i+k} = V_i$ for all $i$). The polygon $W$ is not necessarily closed, and the discrete bicycle monodromy $M_{V,\ell}$ arises, similarly to the continuous case.

**Theorem 1.** *The monodromy $M_{V,\ell} : S^{n-1} \to S^{n-1}$ is a Möbius transformation of the sphere of radius $\ell$.*

Thus, fixed points of the monodromy $M_{V,\ell}$ correspond to closed polygons $W$ in the discrete bicycle correspondence with $V$.

**Theorem 2.** *Let $V$ and $W$ be closed polygons in $\mathbb{R}^n$ in the discrete bicycle correspondence. Then, for every $\lambda$, the monodromies $M_{V,\lambda}$ and $M_{W,\lambda}$ are conjugated to each other.*

[2]The definition in [11, 14], given in 3-dimensional case, involves another, twist, parameter.

Theorem 2 implies that the invariants of the conjugacy class of the monodromy, viewed as functions of the "spectral parameter" $\lambda$, are integrals of the discrete bicycle correspondence. We refer to them as the monodromy integrals.

The next theorem states that the discrete bicycle correspondences with different length parameters commute with each other ("Bianchi permutability"). Recall that we write $\mathcal{B}_\ell(V, W)$ to indicate that polygons $V$ and $W$ are in the discrete bicycle correspondence with the length parameter $\ell$.

**Theorem 3.** *Let $V, W, S$ be closed $k$-gons in $\mathbb{R}^n$ such that $\mathcal{B}_\ell(V, W)$ and $\mathcal{B}_\lambda(V, S)$ hold. Then there exists a closed polygon $T$ such that $\mathcal{B}_\ell(S, T)$ and $\mathcal{B}_\lambda(W, T)$ hold.*

In the case of 3-dimensional space, Theorems 1-3 are not new: in [14], they are proved using quaternions. We give different proofs in Section 2.

V. Adler [1, 2] studied complete integrability of a correspondence on the space of polygons in Euclidean space called the *recutting of polygons*. The recutting $R_i$ of polygon $V$ at $i$th vertex is the reflection of $V_i$ in the perpendicular bisector hyperplane of the segment $V_{i-1}V_{i+1}$ . Recuttings of $k$-gons form a group with generators $R_i,\ i = 1, \dots, k$ and the relations

$$R_i^2 = 1,\ R_iR_j = R_jR_i \text{ for } |i - j| \geq 2, \text{ and } R_iR_{i+1}R_i = R_{i+1}R_iR_{i+1},$$

where the indices are understood cyclically.

The recutting is closely related to the discrete bicycle correspondence. In Section 3, we show that certain integrals of the recutting, discovered by Adler, are integrals of the discrete bicycle correspondence. To do so, we construct a discrete analog of the rear track trajectory, a chain of mutually tangent spheres.

We also have the following result relating the discrete bicycle correspondence and the recutting.

**Theorem 4.** *1) The monodromy is preserved by the recutting. In particular, the monodromy integrals are also integrals of the recutting.*
*2) The discrete bicycle correspondence commutes with the recutting.*

To illustrate the first claim of Theorem 4, a parallelogram and the corresponding kite have the same monodromy, see Figure 3.

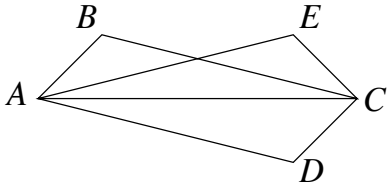


FIGURE 3. The parallelogram $ABCD$ and the kite $AECD$ have the same monodromy

Theorems 1-4 are proved in Section 2.

Consider the low-dimensional situation. If the dimension equals 2 then the discrete bicycle monodromy belongs to $SL(2, \mathbb{R})$. Then one has the trichotomy: $M_{V,\ell}$ may be elliptic, parabolic, and hyperbolic. In the last case, $M_{V,\ell}$ has two fixed points, and one can choose one (say, the attracting one) to construct a closed polygon $W$ in the discrete bicycle correspondence with $V$ (with length parameter $\ell$). According to Theorem 2, $M_{W,\ell}$ is again hyperbolic, and one may iterate the construction by choosing

the other fixed point of $M_{W,\ell}$ (otherwise, one gets back to $V$). Thus, the discrete bicycle correspondence becomes a map on polygons, and we write $\mathcal{T}_\ell(V) = W$.

In dimension three, the discrete bicycle monodromy belongs to $SL(2,\mathbb{C})$. If the monodromy is not the identity, it has two fixed points (perhaps, coinciding), and once again, one can consider the discrete bicycle correspondence as a mapping of the space of polygons in $\mathbb{R}^3$.

In Sections 4 and 5, we study the discrete bicycle transformation on plane polygons. We prove that the discrete bicycle transformation is defined on convex cyclic polygons only if the length parameter does not exceed the diameter of the circumcircle, and in this case, the transformation is a rotation about the circumcenter. We also compute the eigenvalues of the discrete bicycle monodromy and derive a criterion for the monodromy to be parabolic.

In Section 5, we give a complete description of the dynamics of the discrete bicycle transformation on plane quadrilaterals. As an application, we classify the so-called bicycle $(4k,k)$-gons (see Section 5 for definition).

## 2. Proofs of basic properties

**Proof of Theorem 1.** Recall that the Mobïus group $O(n,1)$ consists of linear isometries of the pseudo-Euclidean space $\mathbb{R}^{n,1}$, and it acts projectively on $S^{n-1}$, the spherization of the null cone; it is also the group of isometries of $n$-dimensional hyperbolic space (in the hyperboloid model).

Let $M$ be the monodromy along segment $V_1V_2$ in Figure 2. We need to show that $M \in O(n,1)$.

Let $u, v$ and $x$ be the unit vectors along $V_1W_1, V_2W_2$ and $V_1V_2$, respectively, and let $|V_1V_2| = a$. The reflection of vector $u$ in vector $\xi$ is given by the formula

$$v = \frac{2u\cdot\xi}{|\xi|^2}\xi - u.$$

Applying this to $\xi = ax - \ell u$, we obtain

$$v = \frac{u + \frac{2a^2(x\cdot u)}{\ell^2-a^2}x - \frac{2a\ell}{\ell^2-a^2}x}{\frac{\ell^2+a^2}{\ell^2-a^2} - \frac{2a\ell(x\cdot u)}{\ell^2-a^2}}. \tag{1}$$

On the other hand, a matrix from $O(n,1)$ has the form

$$\begin{pmatrix} A & \xi \\ \eta^t & \lambda \end{pmatrix}$$

where $A$ is an $n\times n$ matrix, $\xi$ and $\eta$ are $n$-vectors, and the following relations hold:

$$A^tA = E + \eta\otimes\eta^t,\ A^t(\xi) = \lambda\eta,\ \xi\cdot\xi = \lambda^2 - 1,$$

where $E$ is the unit matrix, and $\eta\otimes\eta^t$ is the rank one matrix obtained by multiplying a column and a row vectors. The projective action of such a matrix is given by the formula:

$$u \mapsto \frac{A(u)+\xi}{\eta\cdot u + \lambda}. \tag{2}$$

We observe that (1) has the form (2) with

$$A = E + \frac{2a^2}{\ell^2-a^2}x\otimes x,\ \xi = \eta = -\frac{2a\ell}{\ell^2-a^2}x,\ \lambda = \frac{\ell^2+a^2}{\ell^2-a^2},$$

which completes the proof. □

In dimension two, one identifies the unit circle with the real projective line via stereographich projection from point $(-1, 0)$. Then Möbius transformations become fractional-linear. If $\alpha$ is the angular coordinate on $S^1$ then $x = \tan(\alpha/2)$ is the respective affine coordinate on $\mathbb{RP}^1$. In Figure 2, assume that $V_1V_2$ is horizontal, the direction of $V_1W_1$ is $\alpha$ and that of $V_2W_2$ is $\beta$. If $x = \tan(\alpha/2)$ and $y = \tan(\beta/2)$ then the monodromy is given by the formula

$$y = \frac{\ell + a}{\ell - a}x,$$

or

$$M_\ell = \begin{pmatrix} \ell + a & 0 \\ 0 & \ell - a \end{pmatrix}.$$

In general, if the direction of $V_1V_2$ is $\phi$ then

$$M_\ell = \begin{pmatrix} \ell + a\cos\phi & -a\sin\phi \\ -a\sin\phi & \ell - a\cos\phi \end{pmatrix}. \tag{3}$$

Now we prove a property of isosceles trapezoids that is fundamental for what follows. Let $ABCD$ be a plane isosceles trapezoid, see Figure 4. We call the closed quadrilateral $ABDC$, made of the lateral sides and diagonals of a trapezoid, a *Darboux butterfly*.

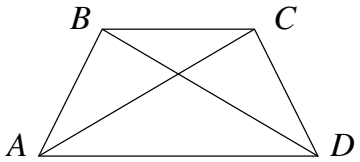


FIGURE 4. A Darboux butterfly

**Lemma 2.1** (Butterfly Lemma)**.** *The monodromy (with any length parameter $\ell$) along a Darboux butterfly is the identity. Conversely, if the monodromy along a closed quadrilateral is the identity for some value of $\ell$ then the quadrilateral is a Darboux butterfly.*

*Proof.* The first statement of the lemma is 3-dimensional: if $w$ is a test vector at vertex $A$ then the respective vectors at all other vertices (the "transports" of $w$ along the quadrilateral) belong to the 3-dimensional space, spanned by the plane of the trapezoid and the vector $w$.

In fact, it suffices to consider the case when $w$ is in the plane of the trapezoid. Indeed, in dimension three, the monodromy is considered as an orientation preserving isometry of hyperbolic space acting on the sphere at infinity. If such an isometry has more than two fixed points then it is the identity.

In dimension two, we shall prove that the monodromy along the polygonal path $ABD$ equals the monodromy along the path $ACD$ if and only if $ABDC$ is a Darboux butterfly. Without loss of generality, assume that $AD$ is horizontal. Let $a, b, c, d$ be the length of the segments $AB, BD, AC, CD$, and let $\alpha, \beta, \gamma, \delta$ be the angles made with the positive horizontal axis. Let $|AD| = g$.

The product of the matrices from equation (3) is

$$\begin{pmatrix} \ell - b\cos\beta & -b\sin\beta \\ -b\sin\beta & \ell + b\cos\beta \end{pmatrix} \begin{pmatrix} \ell - a\cos\alpha & -a\sin\alpha \\ -a\sin\alpha & \ell + a\cos\alpha \end{pmatrix},$$

so we have the monodromy

$$M(a,b,\alpha,\beta)=\begin{pmatrix} \ell^2-lg+ab\cos(\alpha-\beta) & -ab\sin(\alpha-\beta) \\ ab\sin(\alpha-\beta) & \ell^2+lg+ab\cos(\alpha-\beta)\end{pmatrix}. \tag{4}$$

For equality to hold, we must have

$$M(a,b,\alpha,\beta)=k(\ell)M(c,d,\gamma,\delta)$$

for some constant $k(\ell)$ dependent only on $\ell$. Therefore

$$\frac{\ell^2-\ell g+ab\cos(\alpha-\beta)}{\ell^2-\ell g+cd\cos(\gamma-\delta)}=\frac{ab\sin(\alpha-\beta)}{cd\sin(\gamma-\delta)}=\frac{\ell^2+\ell g+ab\cos(\alpha-\beta)}{\ell^2+\ell g+cd\cos(\gamma-\delta)}.$$

Set $X=\ell^2+ab\cos(\alpha-\beta)$ and $Y=\ell^2+cd\cos(\gamma-\delta)$. Then

$$\frac{X-\ell g}{Y-\ell g}=\frac{X+\ell g}{Y+\ell g},$$

hence $X=Y$ and

$$ab\cos(\alpha-\beta)=cd\cos(\gamma-\delta),\quad ab\sin(\alpha-\beta)=cd\sin(\gamma-\delta). \tag{5}$$

The second equation (5) implies that the signed area of triangle $ABD$ is equal to that of triangle $ACD$, so that the quadrilateral $ABDC$ has a total signed area of zero. It also follows that $\tan(\alpha-\beta)=\tan(\gamma-\delta)$, so that $\alpha-\beta=\gamma-\delta$ or $\alpha-\beta=\gamma-\delta\pm\pi$. Since the signed areas are equal, the angles must be equal, and it follows that the quadrilateral is cyclic, and thus a Darboux butterfly.

Note that if the equality holds for one (non-zero) value of $\ell$ then it holds for all values of $\ell$.

Finally, consider a non-planar quadrilateral $ABDC$ with the trivial monodromy (for some value of $\ell$). Assume that the monodromy along $ABD$ and $ACD$ are equal. Denote this monodromy by $M$. Then $M$ preserves the segments that lie in the plane $ABD$ and in the plane $ACD$, and hence, in their intersection, the line $AD$. In the plane $ABD$, the monodromy $M$ is given by formula (4). If the horizontal axis is an eigendirection then $ab\sin(\alpha-\beta)=0$. This implies that the segments $AB$ and $BD$ are collinear, a contradiction. □

As a consequence of Butterfly Lemma, for every $n$, we can construct a family of $2n$-gons with identity monodromy for all values of $\ell$. These polygons are obtained by attaching Darboux butterflies to each other along the common sides, see Figure 5.

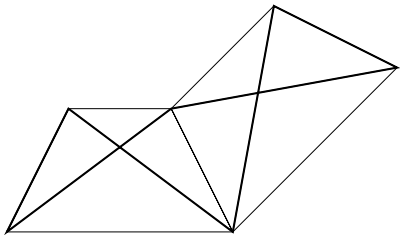

FIGURE 5. Constructing polygons with identity monodromy

Now we are in a position to prove the rest of the theorems.

**Proof of Theorem 2.** It follows from the Butterfly Lemma that, in Figure 2, one has:

$$M_{W_1W_2,\lambda}=M_{V_2W_2,\lambda}M_{V_1V_2,\lambda}M^{-1}_{V_1W_1,\lambda}.$$

Taking the composition over the closed polygon $V$ yields the result. □

**Proof of Theorem 3.** Consider the points $V_1, W_1, S_1$, and let $T_1$ be the point such that $V_1W_1T_1S_1$ is a Darboux butterfly. Consider the discrete bicycle transformation of the segment $V_1V_2$ along the Darboux butterfly $V_1W_1T_1S_1$. According to the Butterfly Lemma, the resulting quadrilateral, say, $Q$, is closed and, according to Theorem 2, it has the trivial monodromy (for any length parameter). Hence, by the Butterfly Lemma again, $Q$ is a Darboux butterfly as well.

It is clear from Figure 2 that the discrete bicycle transformation of the segment $V_1W_1$ along $V_1V_2$ is the same as the discrete bicycle transformation of the segment $V_1V_2$ along $V_1W_1$. It follows that three of the vertices of $Q$ are $V_2, W_2$ and $S_2$. Denote the fourth vertex by $T_2$.

A continuation of this process yields a closed polygon $T$ satisfying the assertion of the theorem. □

**Proof of Theorem 4.** An equivalent description of recutting $V_i \mapsto V_i'$ is that the quadrilateral $V_{i-1}V_iV_{i+1}V_i'$ is a Darboux butterfly.

To prove the first statement, we use Butterfly Lemma:

$$M_{V_{i-1}V_i'V_{i+1},\lambda} = M_{V_{i-1}V_iV_{i+1}V_i'V_{i+1},\lambda} = M_{V_{i-1}V_iV_{i+1},\lambda}.$$

For the second statement, let $W$ be be a polygon in the discrete bicycle correspondence with $V$. Let $V_i'W_i'$ be the discrete bicycle transformation of the segment $V_{i-1}W_{i-1}$ along the segment $V_{i-1}V_i'$. Since $V_{i-1}V_iV_{i+1}V_i'$ is a Darboux butterfly, the discrete bicycle transformation takes $V_i'W_i'$ to $V_{i+1}W_{i+1}$. Thus the polygon $\dots W_{i-1}W_i'W_{i+1}\dots$ is in the discrete bicycle correspondence with $\dots V_{i-1}V_i'V_{i+1}\dots$

We want to show that the recutting of $W$ on $i$th vertex yields $W_i'$ or, equivalently, that $W_{i-1}W_iW_{i+1}W_i'$ is a Darboux butterfly. According to Butterfly Lemma, we need to show that the monodromy along the closed polygon $W_{i-1}W_iW_{i+1}W_i'$ is the identity. Indeed, using that the monodromy of each Darboux butterfly is trivial, we obtain:

$$\begin{aligned} M_{W_{i-1}W_iW_{i+1}W_i'W_{i-1},\lambda} &= M_{W_{i-1}V_{i-1}V_iW_iV_iV_{i+1}W_{i+1}V_{i+1}V_i'W_i'V_i'V_{i-1}W_{i-1},\lambda} \\ &= M_{W_{i-1}V_{i-1}V_iV_{i+1}V_i'V_{i-1}W_{i-1},\lambda} = Id, \end{aligned}$$

and we are done. □

## 3. Integrals

As we mentioned earlier, the discrete bicycle transformation preserves the conjugacy equivalence class of the monodromy $M_\lambda$, thus yielding the monodromy integrals. These integrals do not change if a polygon is acted upon by an isometry of the ambient space. We plan to study the monodromy integrals in a forthcoming paper. In this section, we study the integrals introduced in [1, 2] as integrals of the recutting. One of these integrals, $J(V)$, is not preserved by isometries. The other integral, $A(V)$, was described, in the 3-dimensional case, in [14].

Given a closed polygon $V$, consider the vector $J$ and the bivector $A$ given by the formulas

$$\begin{aligned} J(V) &= \sum_i (|V_{i+1}|^2 - |V_{i-1}|^2)V_i = \sum_i |V_i|^2(V_{i-1} - V_{i+1}), \\ A(V) &= \sum_i V_i \wedge V_{i+1}, \end{aligned} \tag{6}$$

where the sums are cyclic. In dimension 2, $A(V)$ is the signed area of the polygon $V$.

**Theorem 5.** *Both $A$ and $J$ are integrals of the discrete bicycle transformation.*

As a preparation to the proof, we describe a discrete counterpart to the rear bicycle track (the middle curve with cusps in Figure 1).

We shall consider collections of spheres such that the first one is tangent to the second, the second to the third, ... , and the last one is tangent to the first. We call such a collection a *chain*. The radii of the spheres are signed. By convention, if two spheres have an exterior tangency then their radii have the same sign, and if the tangency is interior then the radii have the opposite signs. We allow infinite radii, that is, we consider hyperplanes as spheres as well. An infinite radius has no sign (equivalently, one may consider the curvatures, not excluding zero curvature form consideration). A chain is called *oriented* if one can choose the signs of the radii consistent with the sign convention. That is, a chain is oriented if and only if the number of interior tangencies is even, see Figure 6.

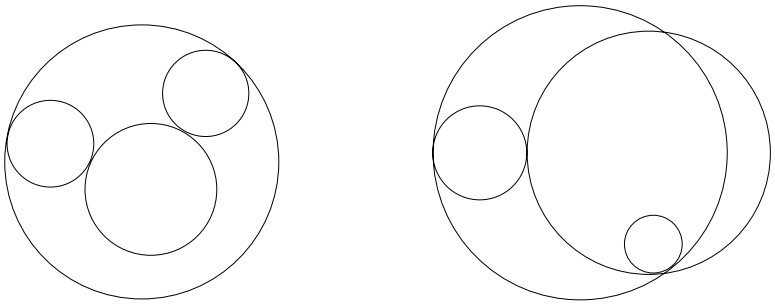

FIGURE 6. An oriented and a non-oriented chain of four circles

In what follows, we use half-integers as the indices for the centers of the spheres and of their radii. Consider an oriented chain of spheres with centers $P_j$ and signed radii $r_j$. Denote by $Q_i$ the tangency point of the adjacent spheres with centers $P_{i-\frac{1}{2}}$ and $P_{i+\frac{1}{2}}$. Let $V_i$ and $W_i$ be the two points on the line $P_{i-\frac{1}{2}}P_{i+\frac{1}{2}}$ located at distance $\ell$ from $Q_i$. The choice of labels is consistent for all $i$: if the segments $V_iW_i$ and $Q_iP_{i+\frac{1}{2}}$ have the same orientations then the segments $V_{i+1}W_{i+1}$ and $Q_{i+1}P_{i+\frac{1}{2}}$ have the opposite orientations, and vice versa.

**Lemma 3.1.** *The polygons $V$ and $W$ are in the discrete bicycle correspondence. Conversely, given polygons $V$ and $W$ in the discrete bicycle correspondence, let $P_{i+\frac{1}{2}}$ be the intersection point of the lines $V_{i+1}W_{i+1}$ and $V_iW_i$. Then there exists an oriented chain of spheres centered at points $P_j$, such that the tangency points $Q_i$ are the midpoints of the segments $V_iW_i$.*

The construction is illustrated in Figure 7.

*Proof.* By construction, a homothety centered at $P_{i+\frac{1}{2}}$ takes $V_i$ to $W_i$ and $W_{i+1}$ to $V_{i+1}$. For example, in Figure 7, the homothety with the coefficient

$$-\frac{\ell - r_{\frac{3}{2}}}{\ell + r_{\frac{3}{2}}},$$

centered at $P_{\frac{3}{2}}$, takes $V_1W_2$ to $W_1V_2$. Since $|V_iW_i| = |V_{i+1}W_{i+1}| = 2\ell$, the quadrilateral $V_iW_iV_{i+1}W_{i+1}$ is a Darboux butterfly.

Conversely, by construction,

$$|P_{i+\frac{1}{2}}W_i| = |P_{i+\frac{1}{2}}V_{i+1}|, \quad |P_{i+\frac{1}{2}}V_i| = |P_{i+\frac{1}{2}}W_{i+1}|,$$

hence $|P_{i+\frac{1}{2}}Q_i| = |P_{i+\frac{1}{2}}Q_{i+1}| := r_{i+\frac{1}{2}}$ where $Q_i$ is the midpoint of the segment $V_iW_i$. The sphere with this radius passes through points $Q_i$ and $Q_{i+1}$ and is orthogonal to the lines $P_{i+\frac{1}{2}}Q_i$ and $P_{i+\frac{1}{2}}Q_{i+1}$. Thus one obtains a chain of spheres, and this chain is oriented. □

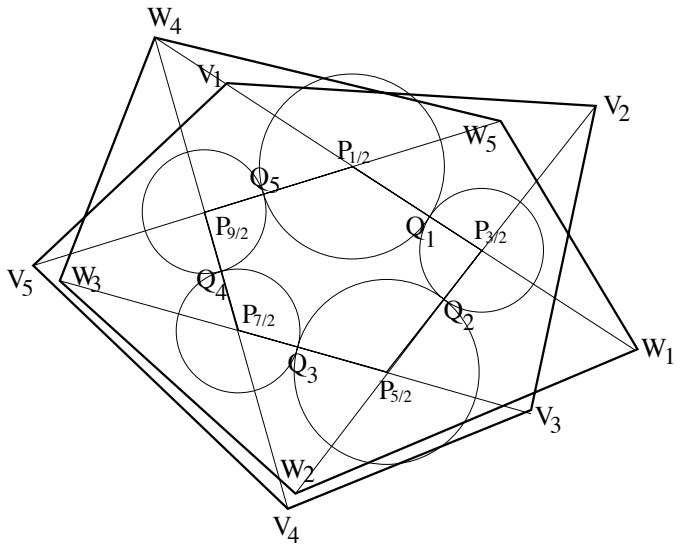


FIGURE 7. Polygons $V$ and $W$ are in the discrete bicycle correspondence

The polygon $Q$ is the discrete rear bicycle track. We apply Lemma 3.1 to prove Theorem 5.

**Proof of Theorem 5.** Given an oriented chain with centers at points $P_j$ and signed radii $r_j$ (where $j$ is half-integer), the tangency points $Q_i$ have the following coordinates:

$$Q_i = \frac{r_{i+\frac{1}{2}}P_{i-\frac{1}{2}} + r_{i-\frac{1}{2}}P_{i+\frac{1}{2}}}{r_{i-\frac{1}{2}} + r_{i+\frac{1}{2}}}$$

(note that this formula does not change if all radii are negated). Then the points $V_i$ and $W_i$ are given by the formula

$$\frac{(r_{i+\frac{1}{2}} - \ell)P_{i-\frac{1}{2}} + (r_{i-\frac{1}{2}} + \ell)P_{i+\frac{1}{2}}}{r_{i-\frac{1}{2}} + r_{i+\frac{1}{2}}}, \tag{7}$$

where the positive $\ell$ gives $V_i$ and the negative $\ell$ gives $W_i$.

To prove the invariance of $A$, we need to show that $A$ is an even function of $\ell$. Indeed, using formula (7), we find that the odd (linear in $\ell$) part of $A$ is

$$\sum \frac{(P_{i+\frac{1}{2}} - P_{i-\frac{1}{2}}) \times (r_{i+\frac{3}{2}} P_{i+\frac{1}{2}} + r_{i+\frac{1}{2}} P_{i+\frac{3}{2}})}{(r_{i-\frac{1}{2}} + r_{i+\frac{1}{2}})(r_{i+\frac{1}{2}} + r_{i+\frac{3}{2}})} + \frac{(r_{i+\frac{1}{2}} P_{i-\frac{1}{2}} + r_{i-\frac{1}{2}} P_{i+\frac{1}{2}}) \times (P_{i+\frac{3}{2}} - P_{i+\frac{1}{2}})}{(r_{i-\frac{1}{2}} + r_{i+\frac{1}{2}})(r_{i+\frac{1}{2}} + r_{i+\frac{3}{2}})} = \sum \frac{P_{i+\frac{1}{2}} \times P_{i+\frac{3}{2}}}{r_{i+\frac{1}{2}} + r_{i+\frac{3}{2}}} - \sum \frac{P_{i-\frac{1}{2}} \times P_{i+\frac{1}{2}}}{r_{i-\frac{1}{2}} + r_{i+\frac{1}{2}}} = 0,$$

as needed.

To prove that $J$ is invariant, one makes a similar computation. Let $e_i$ be the unit vector from $Q_i$ to $P_{i+\frac{1}{2}}$.

One has $V_i = Q_i + \ell e_i$ and $W_i = Q_i - \ell e_i$. Hence

$$|V_i|^2 = \ell^2 + 2\ell \; Q_i \cdot e_i + |Q_i|^2.$$

It follows that

$$|V_{i+1}|^2 - |V_{i-1}|^2 = |Q_{i+1}|^2 - |Q_{i-1}|^2 + 2\ell \; (Q_{i+1} \cdot e_{i+1} - Q_{i-1} \cdot e_{i-1}),$$

and the odd (linear in $\ell$) part of $J$ is

$$\sum (|Q_{i+1}|^2 - |Q_{i-1}|^2) \; e_i + 2(Q_{i+1} \cdot e_{i+1} - Q_{i-1} \cdot e_{i-1}) \; Q_i. \tag{8}$$

Rewrite negative (8) as

$$\begin{aligned} \sum |Q_i|^2 (e_{i+1} - e_{i-1}) + 2Q_i \cdot e_i (Q_{i+1} - Q_{i-1}) = \\ \sum |Q_i|^2 ((e_{i+1} + e_i) - (e_i + e_{i-1})) + 2Q_i \cdot e_i (Q_{i+1} - Q_{i-1}). \end{aligned} \tag{9}$$

Using the formulas

$$Q_{i-1} = Q_i - r_{i-\frac{1}{2}}(e_i + e_{i-1}), \; Q_{i+1} = Q_i + r_{i+\frac{1}{2}}(e_i + e_{i+1}),$$
$$P_{i-\frac{1}{2}} = Q_i - r_{i-\frac{1}{2}} e_i, \; P_{i+\frac{1}{2}} = Q_i + r_{i+\frac{1}{2}} e_i,$$

rewrite (9) as

$$\sum (|Q_i|^2 + 2r_{i+\frac{1}{2}} Q_i \cdot e_i)(e_{i+1} + e_i) - (|Q_i|^2 - 2r_{i-\frac{1}{2}} Q_i \cdot e_i)(e_{i-1} + e_i) = \sum (|P_{i+\frac{1}{2}}|^2 - r_{i+\frac{1}{2}}^2)(e_{i+1} + e_i) - \sum (|P_{i-\frac{1}{2}}|^2 - r_{i-\frac{1}{2}}^2)(e_{i-1} + e_i) = 0,$$

as needed. □

**Remark 3.2.** One has the following relation between the integrals $A$ and $J$:

$$D_\xi(J)(V) = -2A(V) \cdot \xi = 2 \sum_i (V_i \cdot \xi) \; (V_{i-1} - V_{i+1}), \tag{10}$$

where $D_\xi$ is the directional derivative along a vector $\xi$ and where dot is the Euclidean pairing of 2-vectors and vectors. Of course, (10) is also an integral for every vector $\xi$.

**Remark 3.3.** The integral $A$ is invariant under parallel translations, but $J$ is neither invariant under parallel translations nor commutes with them. In dimension two, we adjust the integral $J$ so that it commutes with parallel translations and thus becomes

a "center", associated with a polygon. Namely, rotate $J(V)$ through $90^\circ$ and divide by four times the area:

$$\frac{1}{4A(V)}\big(\sum(y_i^2y_{i+1}-y_iy_{i+1}^2+x_i^2y_{i+1}-x_{i+1}^2y_i),\ \sum(x_ix_{i+1}^2-x_i^2x_{i+1}+x_iy_{i+1}^2-x_{i+1}y_i^2)\big),$$

where $V_i=(x_i,y_i)$ and the sums are cyclic. We call this point the *circumcenter of mass* of the polygon $V$ and denote it by $CCM(V)$.

A justification of this terminology is as follows. Consider a triangulation of the polygon $V$, and let $O_i$ be the circumcenter of $i$th triangle. Then $CCM(V)$ is the center of mass of the points $O_i$, taken with the weight equal to the (oriented) area of $i$th triangle. The result does not depend on triangulation. This construction is mentioned in [1]; we plan to study it in detail in a forthcoming paper [17].

Let us mention, without proof, two properties of $CCM(V)$. First, if $V$ is an equilateral polygon then the circumcenter of mass coincides with the center of mass. This agrees with the observation, made in [4] that, in our terminology, the discrete bicycle transformation of an equilateral polygon preserves its center of mass.

Second, in the continuous limit, as $V$ becomes a curve $\gamma$, the circumcenter of mass of $V$ tends to the center of mass of the homogeneous lamina bounded by $\gamma$. As a consequence, the continuous bicycle transformation preserves the center of mass.

We plan to study the monodromy integrals in a separate paper. We comment on these integrals in dimension two in the next section.

## 4. In the plane

In this section, we consider the discrete bicycle transformation in the plane. We start with a simple observation: for an inscribed polygon, a rotation about the circumcenter is a discrete bicycle transformation, see Figure 8.

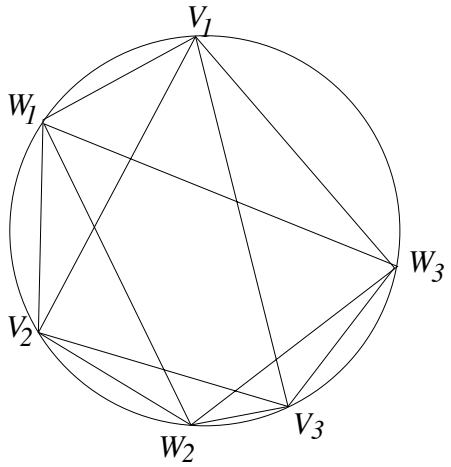


FIGURE 8. Triangles $V$ and $W$ are in the discrete bicycle correspondence

Our first result concerns convex inscribed polygons.

**Theorem 6.** *Let $V$ be a convex inscribed polygon, and let $d$ be the diameter of the circumcircle. The discrete bicycle monodromy $M_{V,\ell}$ is elliptic for $\ell > d$, parabolic for*

*$\ell = d$, and hyperbolic for $\ell \in (0, d)$. In the last case, the discrete bicycle transformation is a rotation about the circumcenter.*

*Proof.* As we mentioned, if $\ell \in (0, d)$ then a rotation about the circumcenter is a discrete bicycle transformation.

Let $W = T_\ell(V)$. Then $W$ has the same perimeter and the same oriented area as $V$, see Theorem 5. It is known that, among polygons with given side lengths, there exists a unique area maximizing one, and this is an inscribed convex polygon. It follows that $W$ is inscribed, and hence congruent to $W$. It follows from Theorem 5 that the circumcenter of $W$ coincides with that of $V$, see Remark 3.3. It follows that $W$ is a rotation of $V$ about the circumcenter. □

In particular, Theorem 6 completely described the discrete bicycle transformation on triangles.

Next we consider a $2k$-gon whose sides lie, in an alternating fashion, on two concentric circles. In the limiting case, the two concentric circles may become two parallel lines.

**Proposition 4.1.** *Let $C_1$ and $C_2$ be concentric circles with the center $O$ (or parallel lines). Let the odd vertices of a $2k$-gon lie on $C_1$ and the even ones on $C_2$. Let $W_1$ be a point of $C_2$. Then the discrete bicycle transformation of $V$ with the initial segment $V_1W_1$ is a closed $2k$-gon whose odd vertices lie on $C_2$ and the even ones on $C_1$. The second iteration of this discrete bicycle transformation sends $V$ to an isometric polygon.*

*Proof.* Reflect $V_1$ in the perpendicular bisector of the segment $W_1V_2$ to obtain $W_2$, and continue in the same way, see Figure 9. Let the lower case letters denote the angular coordinates of the respective points. Then

$$w_2 = w_1 + v_2 - v_1,\ w_3 = w_2 + v_3 - v_2 = w_1 + v_3 - v_1,$$

etc. It follows that $w_{2k+1} = w_1 + v_{2k+1} - v_1 = w_1$, hence the polygon $W$ is closed.

We see that the discrete bicycle transformation $\mathcal{T}$ is the composition of two commuting transformations: the rotation through the angle $w_1 - v_1$, and the involution that interchanges the points of $C_1$ and $C_2$ on the same radial ray. Hence $\mathcal{T}^2$ is a rotation.

The argument for parallel lines is analogous, and the resulting polygon $W$ is obtained from $V$ by a glide reflection. In this case, the orbit of the polygon is unbounded. □

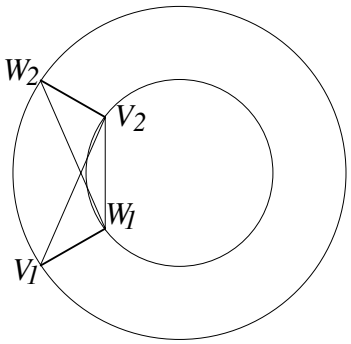


FIGURE 9. $W_2$ is the reflection of $V_1$ in the perpendicular bisector of $W_1V_2$

Note that, in this construction, the polygon $Q$, whose vertices are the midpoints of the segments $V_iW_i$ (see Lemma 3.1), is inscribed in a circle with the center $O$. We also have the following consequence of the proof.

**Corollary 4.2.** *If the polygon in the preceding Proposition is a rhombus then its image under the bicycle transformation is a congruent rhombus.*

Now we discuss the monodromy integrals for plane polygons. The monodromy along a side is given by formula (3); the full monodromy $M$ is the product of these monodromies over the consecutive sides of the polygon. The monodromy is defined only up to a multiplicative factor, and the invariant quantity is

$$\frac{\mathrm{Tr}^2(M)}{\det(M)},$$

considered as a function of $\ell$. Note that the determinant of the matrix (3) equals $\ell^2 - a^2$, that is, is also an integral. Thus $\mathrm{Tr}(M)$ is an integral.

**Proposition 4.3.** *Consider a $k$-gon whose sides have the lengths $a_1, \dots, a_k$ and the directions $\alpha_1, \dots, \alpha_k$. Then*

$$\mathrm{Tr}(M) = 2(\ell^k + c_1\ell^{k-1} + c_2\ell^{k-2} + \dots + c_k)$$

*with all odd coefficients $c_1, c_3, \dots$ equal to zero. If $k$ is even then the free term $c_k$ equals*

$$a_1 \dots a_k \cos(\alpha_1 - \alpha_2 + \dots - \alpha_k).$$

*One also has:*

$$c_2 = -\frac{1}{2}\sum a_i^2.$$

*Proof.* One has

$$M = \prod_{i=1}^{k} (\ell E + a_i A(\alpha_i))$$

where

$$A(\alpha) = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ -\sin\alpha & -\cos\alpha \end{pmatrix}.$$

Therefore

$$\mathrm{Tr}(M) = \sum_{j=0}^{k} \ell^{k-j} a_{i_1} \dots a_{i_j} \mathrm{Tr}(A(\alpha_{i_1}) \dots A(\alpha_{i_j})).$$

Notice that

$$A(\alpha)A(\beta) = \begin{pmatrix} \cos(\alpha-\beta) & \sin(\alpha-\beta) \\ -\sin(\alpha-\beta) & \cos(\alpha-\beta) \end{pmatrix}, \tag{11}$$

a rotation matrix. More generally, the product of an odd number of the matrices $A(\alpha_i)$ is traceless, and the product of an even number is a rotation through the alternating sum of the respective angles. This implies the first two claims.

For the last claim, let $u_1, \dots, u_k$ be the vectors of the sides of the polygon. Using (11), we find that

$$c_2 = \sum_{i<j} u_i \cdot u_j.$$

One has: $\sum u_i = 0$. Taking dot with itself yields:

$$0 = \sum u_i \cdot u_i + 2\sum_{i<j} u_i \cdot u_j.$$

Thus

$$c_2 = -\frac{1}{2}\sum a_i^2,$$

as claimed. □

**Corollary 4.4.** *The quantity* $\cos(\alpha_1 - \alpha_2 + \cdots - \alpha_k)$ *is an integral of the discrete bicycle transformation on even-gons.*

Let polygons $V$ and $W$ be in the discrete bicycle correspondence. Let

$$\alpha_i = \angle V_{i-1}V_iW_i = \angle V_{i-1}W_{i-1}W_i,$$

see Figure 10. If one knows the cyclic sequence of angles $\alpha_i$ then one can construct $W$ from $V$: indeed, the lengths of all the segments $V_iW_i$ are equal to $2\ell$.

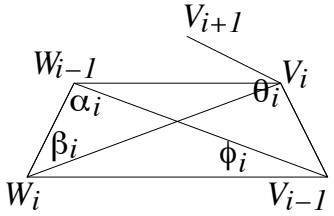


FIGURE 10. Notations for Proposition 4.5

The angles $\alpha_i$ satisfy a first order nonlinear difference equation with periodic coefficients. Let $\theta_i = \angle V_{i-1}V_iV_{i+1}$ and $c_i = |V_{i-1}V_i|$.

**Proposition 4.5.** *One has*

$$2\ell \cos\left(\frac{\alpha_i - \alpha_{i-1} + \theta_{i-1}}{2}\right) = c_i \cos\left(\frac{\alpha_i + \alpha_{i-1} - \theta_{i-1}}{2}\right). \tag{12}$$

*Proof.* Let

$$\beta_i = \angle W_{i-1}V_{i-1}V_i = \angle W_{i-1}W_iV_i, \quad \phi_i = \angle W_{i-1}V_{i-1}W_i = \angle V_iW_iV_{i-1}.$$

Then $2\phi_i = \pi - \alpha_i - \beta_i$. Since $\angle W_iV_iV_{i+1} = \beta_{i+1}$, one has $\beta_{i+1} = \theta_i - \alpha_i$. Therefore

$$\phi_i = \frac{\pi}{2} - \frac{\alpha_i - \alpha_{i-1} + \theta_{i-1}}{2}, \ \beta_i + \phi_i = \frac{\pi}{2} - \frac{\alpha_{i-1} + \alpha_i - \theta_{i-1}}{2}. \tag{13}$$

By Sine Rule in triangle $V_{i-1}V_iW_i$,

$$\frac{2\ell}{\sin(\beta_i + \phi_i)} = \frac{c_i}{\sin\phi_i},$$

or

$$2\ell \cos\left(\frac{\alpha_i - \alpha_{i-1} + \theta_{i-1}}{2}\right) = c_i \cos\left(\frac{\alpha_i + \alpha_{i-1} - \theta_{i-1}}{2}\right),$$

as claimed. □

As an application of Proposition 4.5, we compute the eigenvalue of the fixed point of the monodromy map of the polygon $V$ corresponding to the pair of polygons $V, W$ in the discrete bicycle correspondence. Since the monodromy is a Möbius transformation, the eigenvalues of its two fixed points are reciprocals of each other.

**Theorem 7.** *The eigenvalue in question equals*

$$\prod_{i=1}^{n} \frac{|V_{i-1}W_i|}{|V_iW_{i-1}|} = \prod_{j=1/2}^{n+1/2} \frac{|\ell + r_j|}{|\ell - r_j|}.$$

*In particular, the monodromy is parabolic if and only if*

$$\prod_{i=1}^{n} |V_{i-1}W_i| = \prod_{i=1}^{n} |V_iW_{i-1}| \quad \text{or} \quad \prod_{j=1/2}^{n+1/2} |\ell + r_j| = \prod_{j=1/2}^{n+1/2} |\ell - r_j|.$$

*Proof.* To compute the eigenvalue, one linearizes equation (12): if $u_i$ is a variation of $\alpha_i$ then the linearization is as follows:

$$2\ell(u_i - u_{i-1}) \sin\left(\frac{\alpha_i - \alpha_{i-1} + \theta_{i-1}}{2}\right) = c_i(u_i + u_{i-1}) \sin\left(\frac{\alpha_i + \alpha_{i-1} - \theta_{i-1}}{2}\right),$$

and hence

$$u_i \left[2\ell \sin\left(\frac{\alpha_i - \alpha_{i-1} + \theta_{i-1}}{2}\right) - c_i \sin\left(\frac{\alpha_i + \alpha_{i-1} - \theta_{i-1}}{2}\right)\right] =$$
$$u_{i-1} \left[2\ell \sin\left(\frac{\alpha_i - \alpha_{i-1} + \theta_{i-1}}{2}\right) + c_i \sin\left(\frac{\alpha_i + \alpha_{i-1} - \theta_{i-1}}{2}\right)\right].$$

By elementary geometry of the trapezoid in Figure 4.5 and formulas (13), one has:

$$2\ell \sin\left(\frac{\alpha_i - \alpha_{i-1} + \theta_{i-1}}{2}\right) = \frac{1}{2}(|V_{i-1}W_i| + |V_iW_{i-1}|),$$
$$c_i \sin\left(\frac{\alpha_i + \alpha_{i-1} - \theta_{i-1}}{2}\right) = \frac{1}{2}(|V_{i-1}W_i| - |V_iW_{i-1}|).$$

Therefore

$$u_i|V_iW_{i-1}| = u_{i-1}|V_{i-1}W_i|,$$

which implies the first formula for the eigenvalue.

For the second formula, note that a homothety centered at point $P_{i+1/2}$ takes segment $V_iW_{i+1}$ to segment $V_{i+1}W_i$, see Figure 7. The coefficient of this homothety is $|\ell + r_{i+1/2}|/|\ell - r_{i+1/2}|$, and we obtain the second formula for the eigenvalue.

It remains to notice that the monodromy is parabolic if and only if the two reciprocal eigenvalues coincide. □

**Remark 4.6.** The continuous analogs of Proposition 4.5 and Theorem 7 are contained in [13]. Namely, the continuos version of (12) is the differential equation

$$\frac{d\alpha}{dx} + \frac{\sin\alpha}{\ell} = \kappa(x)$$

where $\alpha(x)$ is the angle made by the bicycle frame with the front wheel trajectory, $x$ is the arc length parameter along this trajectory, and $\kappa(x)$ is the curvature of this curve. The endpoint of the segment of length $\ell$ describes the rear wheel trajectory.

The continuos version of Theorem 7 states that the eigenvalues of the bicycle monodromy are $e^{\pm length(\gamma)}$ where $\gamma$ is the rear wheel trajectory, and the length is algebraic: the sign changes after one traverses a cusp. In particular, the monodromy is parabolic if and only if the rear track has zero length.

## 5. Case study: plane quadrilaterals

In this section, we describe the dynamics of the discrete bicycle transformation on plane quadrilaterals.

We have a trichotomy according to the position of the circumcenter of mass, see Remark 3.3. Consider a quadrilateral $ABCD$. The first case is when the diagonals $AC$ and $BD$ are not parallel. Let $O$ be the intersection point of the perpendicular bisectors of these diagonals, see Figure 11 on the left.

**Lemma 5.1.** *$O$ is the circumcenter of mass of the quadrilateral $ABCD$.*

*Proof.* The circumcenters of the triangles $ABD$ and $BCD$ lie on the perpendicular bisector of the segment $BD$, and the circumcenters of the triangles $ABC$ and $ACD$ lie on the perpendicular bisector of the segment $AC$. Hence $O = CCM(ABCD)$. □

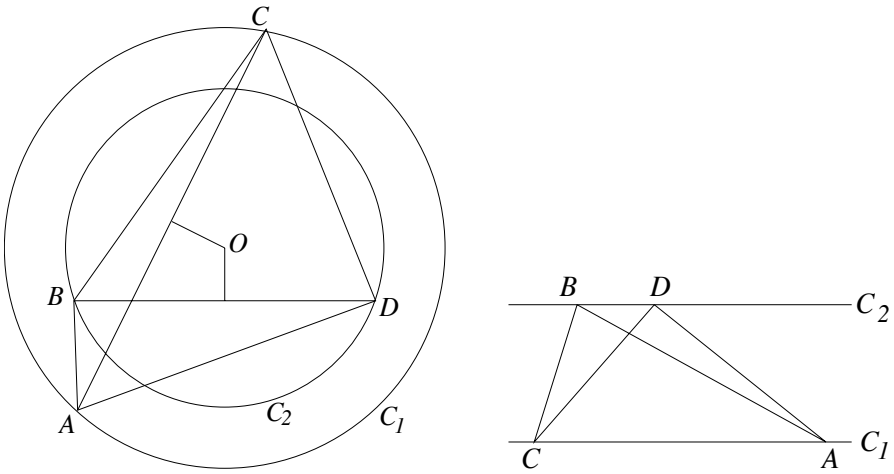


Figure 11. Two types of quadrilaterals: the circumcenter is finite or infinite

In the first case, $A$ and $C$ lie on one circle, say, $C_1$, and $B$ and $D$ on another circle, $C_2$, centered at $O$. Denote their radii by $r_1$ and $r_2$, and assume that $r_1 \geq r_2$.

The second case is when the diagonals are parallel but the quadrilateral is not a Darboux butterfly, see Figure 11 on the right. In this case, the two concentric circles are replaced by two parallel lines, and the center $O$ is at infinity. Although both radii are infinite, their difference $r_1 - r_2$ is still defined and equals the distance between the parallel lines. Note that, in this case, the quadrilateral $ABCD$ has zero area.

The third case is when the quadrilateral is a Darboux butterfly. In this case, there exists an infinite family of pairs of concentric circles $C_1, C_2$ such that $A, C \in C_1$ and $B, D \in C_2$. The centers of these circles lie on the common perpendicular bisector of the segments $AC$ and $BD$, including the point at infinity, when the circles become parallel lines.

**Theorem 8.** *Let $ABCD$ be a quadrilateral. If $ABCD$ is not a Darboux butterfly then the discrete bicycle monodromy about the quadrilateral is elliptic for $\ell \in (0, r_1 - r_2) \cup (r_1 + r_2, \infty)$, hyperbolic for $\ell \in (r_1 - r_2, r_1 + r_2)$, and parabolic for $\ell = r_1 \pm r_2$. For $\ell$ in the hyperbolic or parabolic range, the discrete bicycle correspondence is induced by a*

*point $A' \in C_2$, as described in Proposition 4.1. If $ABCD$ is a Darboux butterfly then the monodromy is the identity. For every starting point $A'$, there exists a circle (or straight line) $C_2$ that passes through $A'$, and the discrete bicycle correspondence is again described by Proposition 4.1.*

*Proof.* If $\ell \in [r_1 - r_2, r_1 + r_2]$ then there exist two points $A' \in C_2$ such that $|AA'| = \ell$ (these two points coincide for $\ell = r_1 \pm r_2$), and Proposition 4.1 describes the discrete bicycle transformation.

Conversely, assume that $A'B'C'D'$ is a discrete bicycle transformation of $ABCD$. Let $l_1, l_2, l_3$ and $l_4$ be the perpendicular bisectors of the segments $A'B, B'C, C'D$ and $D'A$, respectively. Let $R_i$ be the reflection in the line $l_i$, $i = 1, 2, 3, 4$. By definition of the bicycle monodromy,

$$B' = R_1(A),\ C' = R_2(B),\ D' = R_3(C),\ A' = R_4(D),$$

see Figure 9. Note also that

$$B = R_1(A'),\ C = R_2(B'),\ D = R_3(C'),\ A = R_4(D').$$

We claim that the lines $l_1, l_2, l_3, l_4$ are concurrent (as a particular case, the four lines may be parallel).

Consider the composition $F = R_3 \circ R_2 \circ R_1$: it is either a reflection or a glide reflection. We claim that the former is the case. Two given congruent line segments $AA'$,$D'D$ are related by just one odd isometry. Since $AA'D'D$ is an isosceles trapezoid, this isometry is a reflection.

Since $R_3 \circ R_2 \circ R_1$ is a reflection, the lines $l_1, l_2$ and $l_3$ are concurrent. Applying the same argument to $l_2, l_3, l_4$, we conclude that all four lines are concurrent.

To fix ideas, let us assume that the intersection point of the lines $l_1, l_2, l_3, l_4$ is finite (the case of parallel lines is similar). Denote this point by $Q$. We claim that $Q = O$, the circumcenter of the quadrilateral $ABCD$.

Indeed, $R_2 \circ R_1(A) = C$, hence $Q$ lies on the perpendicular bisector of the diagonal $AC$. Likewise, $R_3 \circ R_2(B) = D$, hence $Q$ lies on the perpendicular bisector of the diagonal $BD$. Thus $Q = O$.

Since $A' = R_1(B)$, it follows that $A' \in C_2$, and we are in the situation of Proposition 4.1.

It remains to consider the case of a Darboux butterfly. For any starting point $A'$, we can find a circle $C_2$ through $A', B$ and $D$ with the center $O$ on the perpendicular bisector of the segments $AC$ and $BD$. Then another circle $C_1$, centered at $O$, passes through $A$ and $C$, and we are in the situation described in Proposition 4.1. □

**Remark 5.2.** The preceding argument provides an alternative proof of the fact that the monodromy of a Darboux butterfly is the identity for all $\ell$.

We now discuss an application of Theorem 8 to the following problem in "bicycle mathematics". Suppose one is given two closed curves, the front and rear bicycle tracks. Can one always determine in which direction the bicycle went? Usually, one can, but sometimes one cannot: consider, for example, two concentric circles.

Describing such pairs of "ambiguous" bicycle tracks is an interesting and difficult problem, and only partial results are available. This problem is equivalent to Ulam's problem of describing uniform (2-dimensional) bodies that float in equilibrium in all positions. We refer to [3, 4, 8, 18, 19] for the literature on this intriguing topic.

A discrete version of this problem was introduced in [15]. Define a *bicycle* $(n,k)$*-gon* as an equilateral $n$-gons whose $k$-diagonals have equal length. More precisely, if the polygon is $V_1V_2\dots V_n$ then we require that $V_iV_{i+1}V_{i+k+1}V_{i+k}$ be a Darboux butterfly for all $i$ (as usual, the indices are understood cyclically). The problem is to describe bicycle $(n,k)$-gons, in particular, to determine for which pairs $(n,k)$ such a polygon must be regular. See also [6, 7].

For example, it is shown in [15] that bicycle $(n,2)$-gons, $(2k+1,k)$-gons, and $(3k,k)$-gons are regular. On the other hand, an example of a non-regular bicycle polygon is shown in Figure 12. This construction generalizes to all pairs $(n,k)$ and yields 1-parameter families of bicycle $(n,k)$-gons with even $n$ and odd $k$. Note that the even and the odd vertices of a polygon in Figure 12 lie on two concentric circles and that the polygons have dihedral symmetry.

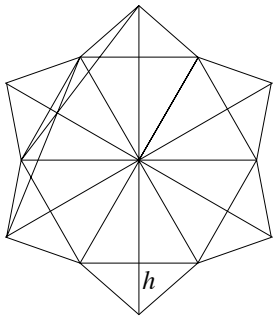


FIGURE 12. A bicycle $(12,3)$-gon: $h$ is a parameter of the construction

Let $\ell$ be the length of the $k$-diagonal of a bicycle $(n,k)$-gon, and let $S$ be the cyclic relabeling of the vertices: $V_i \mapsto V_{i+1}$. One can restate the definition in terms of the discrete bicycle transformation $\mathcal{T}_\ell$: $V$ is a bicycle $(n,k)$-gon if $\mathcal{T}_\ell(V) = S^k(V)$.

The next result is a further step toward classification of bicycle polygons.

**Theorem 9.** *If $k$ is even then a bicycle $(4k,k)$-gon is regular. If $k$ is odd then the even vertices of a bicycle $(4k,k)$-gon are equally spaced on a circle and its odd vertices are equally spaced on a concentric circle, that is, the polygon is obtained from a regular $2k$-gon by the construction depicted in Figure 12.*

*Proof.* Given a bicycle $(4k,k)$-gon $V$, consider the rhombus $V_0V_kV_{2k}V_{3k}$. The discrete bicycle transformation with the length parameter $V_0V_1$ takes this rhombus to $V_1V_{k+1}V_{2k+1}V_{3k+1}$, to $V_2V_{k+2}V_{2k+2}V_{3k+2}$, and so on.

Let $O$ be the center of the rhombus $V_0V_kV_{2k}V_{3k}$, and let $C_1$ and $C_2$ be the concentric circles centered at $O$ such that $V_0, V_{2k} \in C_1$ and $V_k, V_{3k} \in C_2$. By Corollary 4.2, all the consecutive rhombi are congruent, and $V_1 \in C_2, V_2 \in C_1, V_3 \in C_2, V_4 \in C_1$, etc.

Therefore, if $k$ is even, then $V_k \in C_1$, and hence $C_1 = C_2$. It follows that the rhombus is a square and $V$ is a regular $4k$-gon. If $k$ is odd then the even vertices of $V$ form a regular $2k$-gon inscribed into $C_1$, and the odd ones form a regular $2k$-gon inscribed into $C_2$. Thus $V$ is obtained from a regular $2k$-gon by the construction in Figure 12. □

**Acknowledgments**. We have discussed the discrete bicycle transformation with many a mathematician, and we are grateful to them all. In particular, it is a pleasure to

acknowledge interesting discussions with I. Alevi, A. Bobenko, T. Hoffmann, U. Pinkall, B. Springborn, Yu. Suris, and A. Veselov. This project originated during the program Summer@ICERM 2012; we are grateful to ICERM for support and hospitality. S. T. was partially supported by the NSF grant DMS-1105442.

## References

[1] V. Adler, *Cutting of polygons*. Funct. Anal. Appl. **27** (1993), 141–143.
[2] V. Adler, *Integrable deformations of a polygon.* Phys. D **87** (1995), 52–57.
[3] J. Bracho, L. Montejano, D. Oliveros, *A classification theorem for Zindler carrousels.* J. Dynam. Control Systems **7** (2001), 367–384.
[4] J. Bracho, L. Montejano, D. Oliveros, *Carousels, Zindler curves and the floating body problem.* Period. Math. Hungar. **49** (2004), 9–23.
[5] A. Calini, T. Ivey, *Bäcklund transformations and knots of constant torsion.* J. Knot Theory Ramifications **7** (1998), 719–746.
[6] R. Connelly, B. Csikós, *Classification of first-order flexible regular bicycle polygons.* Studia Sci. Math. Hungar. **46** (2009), 37–46.
[7] B. Csikós, *On the rigidity of regular bicycle $(n, k)$-gons.* Contrib. Discrete Math. **2** (2007), 93–106.
[8] D. Finn, *Which way did you say that bicycle went?* Math. Mag. **77** (2004), 357–367.
[9] R. Foote, *Geometry of the Prytz planimeter*. Reports Math. Physics **42** (1998), 249–271.
[10] R. Foote, M. Levi, S. Tabachnikov, *Tractrices, bicycle tire tracks, hatchet planimeters, and a 100-year-old conjecture*. Amer. Math. Monthly, in print.
[11] T. Hoffmann, *Discrete Hashimoto surfaces and a doubly discrete smoke-ring flow*. Discrete differential geometry, 95–115, Oberwolfach Semin., 38, Birkhäuser, Basel, 2008.
[12] S. Howe, M. Pancia, V. Zakharevich, *Isoperimetric inequalities for wave fronts and a generalization of Menzin's conjecture for bicycle monodromy on surfaces of constant curvature*. Adv. Geom. **11** (2011), 273–292.
[13] M. Levi, S. Tabachnikov, *On bicycle tire tracks geometry, hatchet planimeter, Menzin's conjecture, and oscillation of unicycle tracks*. Experiment. Math. **18** (2009), 173–186.
[14] U. Pinkall, B. Springborn, S. Weissmann, *A new doubly discrete analogue of smoke ring flow and the real time simulation of fluid flow*. J. Phys. A **40** (2007), 12563–12576.
[15] S. Tabachnikov, *Tire track geometry: variations on a theme*. Israel J. Math. **151** (2006), 1–28.
[16] S. Tabachnikov, *Remarks on the bicycle transformation and the filament equation*, in preparation.
[17] S. Tabachnikov, E. Tsukerman, *Circumcenter of mass*, in preparation.
[18] F. Wegner, *Floating bodies of equilibrium.* Stud. Appl. Math. **111** (2003), 167–183.
[19] F. Wegner, *Three problems – one solution.* `http://www.tphys.uni-heidelberg.de/~wegner/Fl2mvs/Movies.html#float`

Department of Mathematics, Pennsylvania State University, University Park, PA 16802, USA and Stanford University
*E-mail address*: `tabachni@math.psu.edu`
*E-mail address*: `emantsuk@stanford.edu`

# STABILITY ANALYSIS FOR VIRUS SPREADING IN COMPLEX NETWORKS WITH QUARANTINE

ROBERTO BERNAL JAQUEZ[A], ALEXANDER SCHAUM[A], LUIS ALARCÓN[A], CARLOS RODRÍGUEZ LUCATERO[B]
[A] DEPARTAMENTO DE MATEMATICAS APLICADAS Y SISTEMAS
[B] DEPARTAMENTO DE TECNOLOGIAS DE LA INFORMACION
UNIVERSIDAD AUTONOMA METROPOLITANA-CUAJIMALPA

Abstract. The stability of a discrete-time complex network-based Markov process model for virus spreading with quarantine is studied on the basis of a ($S \to I \to Q \to S$) state automaton. Size independent spectral properties of the underlying nonlinear dynamics are identified, and conditions for extinction are derived in dependence of quarantine rates, infection probability, recovery and interaction rate. Numerical simulations are presented to illustrate the underlying basic bifurcation behavior, whose understanding is the first step towards the development of adequately tailored control strategies for these kind of problems.

## 1. Introduction

One of the main concerns when having to face the spread of infectious diseases is how to control their propagation. Historically, many methods have been proposed for this end, as for instance, the isolation of the infected individuals, or in other words introducing quarantine [1],[2],[3],[4],[5].

Modeling the spread of infectious diseases among individuals has been a relevant problem during many years.. In their classical work Kermack and McKendrick [6] gave birth to the so called SIR model that divides the population in three different compartments or groups: Susceptible individuals, i.e. healthy individuals that are capable of contracting the disease, Infective individuals that have contracted the disease and are the agents to transmit and spread the disease, and Removed (or Recovered) individuals that were infected but become immune or died. As many infections do not confer any immunity, a simplified version of the SIR model was created, the so-called SIS model. In this model, the total population $N$ is divided in two groups or compartments, the group of susceptible individuals S and the the group I of infectious individuals.

In the field of computer virus propagation, Kephart and White proposed one of the first models [7, 8], the so-called homogeneous model in which communications among individual computers (nodes) were modeled by directed graphs (symbolizing connection). Using a rate of infection and a death rate they were able to calculate the infection threshold. One of the main flaws in this theory was supposing the network homogeneity. Experimental evidence shows hat real world computer networks are not homogeneous and follow a power law structure instead [9, 10]. That means that the number of connections (the degree k of the node) the different nodes may have, follow a distribution of the form $P(k) \propto k^{-\gamma}$ where $\gamma$ is known as the power-law exponent. In this kind of networks, there exist nodes with very high connectivity, but the majority of the nodes have low connectivity.

Authors like Pastor-Satorras and others have extensively studied infection spread in these kind of networks [11, 12, 13, 14, 15], using the model developed by Barabasi and Albert [16]. Nevertheless, their results are rather concerned with the case of $\gamma = 3$ what does not hold for many networks.

Using also the Mean Field Approximation Model (MFA), that consider nodes with the same degree as dynamically equivalent , they were able to calculate the epidemic threshold . Although this approach is quite interesting, it is not applicable in many realistic cases because nodes with the same degree not necessarily behave the same way .

The use of complex and dynamical networks on the other hand provided a new impulse in modeling that give new insights in understanding the dynamics of infectious diseases [17, 18].

The notion of epidemics can be defined as an outbreak of a disease over a short time period. The disease is called endemic if it persists in a population over a long period of time. In [1] six modifications of standard SIS and SIR endemic models are proposed that include a class $Q$ of quarantined and isolated individuals that do not infect the susceptible individuals in class $S$. When the members of $S$ become infected they pass to the class $I$, and after a period of time some of them return to $S$ while others are transferred to the class $Q$ and remain there until they are no longer infectious. This model is called $SIQS$. In another type of the model proposed in [1] permanent immunity is conferred by transferring the infected individuals to a class $R$ of recovered individuals. Additionally births, natural deaths and immigration effects are considered that produce a growth of the set $S$ of susceptible individuals. These modifications of the model yields a variable size of the total population. Under these new conditions the thresholds as well as the associated equilibria and their stability are studied.

In [3] models of infectious disease are developed that incorporate the movement of individuals over a range of spatial scales. A general model for a disease that can be transmitted between different species and multiple patches is proposed, and the behavior of the system is studied for the case that the spatial component consists of a ring of patches. The influence of various parameters on the spatial and temporal spread of the disease is numerically analyzed, and focus is spent on quarantine in the form of travel restrictions. Furthermore it was remarked that very often mathematical models of disease spread tend to ignore spatial dynamics , unless the spatio-temporal spread of epidemic diseases was noticed early on since Athens in 430 B.C. It was highlighted that at a local scale, contacts between infective and susceptible individuals lead to the propagation of the epidemics, but in a larger scale, it is through contacts between individuals living in distinct regions that a disease becomes spatially mobile and then the problem becomes more complex. To model mathematically the spread of a disease under such circumstances partial differential equations are frequently used and the problem is approached as a diffusion phenomena. Typical assumptions are that (i) there is diffusion of infection-bearing individuals in a susceptible population, and (ii) the receptors of the disease are fixed and the infectious agent diffuses among them.

The diffusion mathematical approach is not well adapted to some situations as can be those concerning the spread of disease through sparsely populated regions, with several species having distinct rates of mobility and different migration patterns. One example of this last situation is the propagation of the bubonic plague or the avian influenza. For more details of their mathematical model see [3].

In [19] it is mentioned that in absence of vaccines or medicines a good method for trying to control the propagation of a disease is the quarantine by isolation. One epidemic of this kind is the SARS(Severe acute respiratory syndrome). Their paper discusses the application of the optimal quarantine and isolation strategies for SARS outbreak control via the Pontryagins Maximum Principle. They construct a multigroup SARS transmission model for traveling population and introduce pairs of control variables in terms of the quarantine and isolation strategies. A number of infected individuals and a linear cost and a quadratic cost on the controls are imposed. Based on that the SARS disease spreading is simulated and the results obtained enablethe illustration of the importance of the early quarantine and isolation strategies, and the necessity of the observation and quarantine of travelers to control the outbreaks of epidemics. They claim that the early quarantine and isolation strategies, as well as the observation and quarantine of travelers, are critically important to contain the epidemics.

On the other side many research efforts has been done during the last ten years concerning the application of these mathematical models in the behavior of other complex networks as for instance the internet. Recently with the constant augmentation in the number of internet users as well as the growth in the complexity of such networks [16], [20], [9],[21], [22] new security problems have appeared [23], [12], [13], [14], [24] in the scene and there is a lack of adequate security methods for facing attacks under this new setting. These new environments are for instance the P2P networks, sensor networks, social nets or wireless networks, where information is to be stored, generated and retrieved. So under this new environments it can be very important to study and model how the information is spread or how to keep the spreading of a virus under control in such a way that the information remains being useful under these vulnerable circumstances. In [18], [25] it is studied the problem of information survival threshold in sensor and P2P networks, modeling the problem as a non-linear dynamical system and using fixed point stability theorems [26]. A closed form solution is obtained that depends on an additional parameter, the largest eigenvalue of the dynamical system matrix. In sensor networks for instance, the nodes can loss their communication links and the nodes can stop working because of a system failure produced by a virus infection and quarantine process or a system maintenance procedure. Under such conditions they try to answer the question *under what conditions a datum can survive in a network.* Given that the nodes as well as the links can fail with some probability the obvious model can be a Markov chain, but such a model can grow in complexity very quickly because the number of possible states becomes $3^N$ where $N$ is the number of states. To avoid this mathematical problem, one alternative is to model the system as a non-linear dynamical system.

Concerning the modalities of contact that we are going to take into account in our mathematical model and simulations we can mention the following ones [17]:

- Contact process , i.e. each node intends once to contact each neighbor (and hence there is a contact probability)
- Reactive process , i.e. each node intends infinitely many times to contact each neighbor (and hence the contact probability is equal to one)
- Intermediate types of contact.

In this work, we follow the theoretical framework for contact-based spreading of diseases in complex networks [17, 18] with additional quarantine state. Our formulation is based on probabilistic discrete-time Markov chains and applies to weighted and

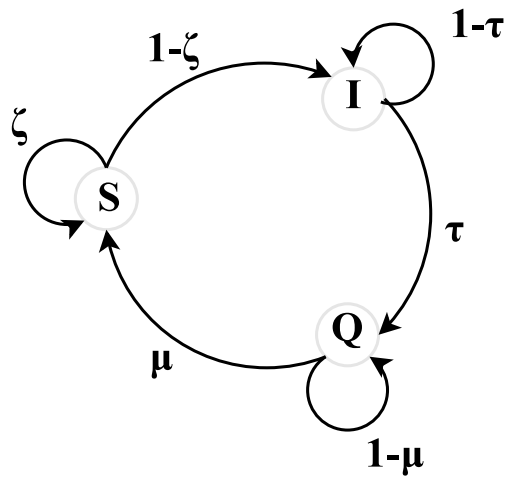


FIGURE 1. State transition diagram associated to the SIQS mechanism.

unweighted complex networks. Within this context it is possible to quantify the microscopic probabilistic dynamics at the individual level. Global virus extinction conditions are derived for the three-state Markov process including susceptible-infected-quarantine states, thus generalizing the ones reported in previous studies on SIS dynamics [17, 18]. The theoretical findings are illustrated and corroborated with numerical simulations. The presented results may constitute the basis for future development of immunization and vaccination policies.

The paper is organized as follows: In Section 2, the mathematical model for the SIQ process is introduced and the associated fixed points are identified. In Section 3, global extinction conditions are derived and discussed in the light of previous ones reported for the SIS process. In Section 4, numerical simulation studies are presented that illustrate the theoretical results developped in Section 3. In Section 5, conlusions are presented.

## 2. The SIQ Model

In this section the SIQ model is presented and the associated fixed points are identified.

2.1. **Model equations.** Considering the classic SIS model for virus spread in complex networks [17, 18] with intermediate quarantine state $Q$ as illustrated in the state automata depicted in Figure 1, the following mathematical model is obtained as an adaptation from [17, 18]:

$$\begin{aligned} s_i(t+1) &= \zeta_i(t)s_i(t) + \mu q_i(t) \\ p_i(t+1) &= (1-\tau)p_i(t) + (1-\zeta_i(t))s_i(t) \ , \quad i = 1, \ldots, N \\ q_i(t+1) &= \tau p_i(t) + (1-\mu)q_i(t) \end{aligned} \tag{1}$$

$$s_i(t) + p_i(t) + q_i(t) = 1 \tag{2}$$

where $s_i(t)$ is the probability of being susceptible at time $t$,

$$\zeta_i(t) = \prod_{j \neq i} [1 - \beta r_{ij} p_i(t)] \tag{3}$$

is the probability of not being infected by any neighbor, $\beta$ is the probability of infection during a single contact,

$$r_{ij} = a_{ij}(t)\left[1 - \left(1 - \frac{1}{\sum_{i=1}^{N} A_{ij}}\right)^{\lambda}\right] \tag{4}$$

[17] is the connection probability which depends on the number of contact intents (or interaction rate) $\lambda$, with $A \in \mathbb{R}^{N\times N}$ being the time-varying adjacency matrix representing interconnections between the $N$ nodes of the network. $p_i$ is the probability of being infected, $\tau$ the internment probability associated to quarantine, $q_i$ is the probability of being in quarantine, and $\mu$ is the recovery rate.

The interconnection type model (4) is a generalization of the two classical cases: (i) the contact process (with $\lambda = 1$) for which in each time step exactly one contact is established between connected nodes, and (ii) the reactive process (with $\lambda \to \infty$) where any connected nodes are in continuous contact. The maximum value $r_{ij} = 1$ is attained for the reactive process. Obviously, the spectral properties of the associated matrix

$$R = \{r_{ij}\}_{i,j} \tag{5}$$

will depend on the interaction rate $\lambda$, and consequently the dynamic behavior of the solutions of (1) shall depend on $\lambda$ as well.

Given the conservation-like property (2), it can be easily seen that, for any node $i$, all the trajectories of the dynamics (1) are constraint to the triangle set (see Figure 2)

$$T = \{x \in [0,1]^3 : 1 + x_1 + x_2 + x_3 = 0\}. \tag{6}$$

Obviously the same is true for the mean probabilities

$$\rho_s = \frac{1}{N}\sum_{i=1}^{N} s_i, \quad \rho_p = \frac{1}{N}\sum_{i=1}^{N} p_i, \quad , \rho_q = \frac{1}{N}\sum_{i=1}^{N} q_i. \tag{7}$$

Solving the equation (2) for $s_i$ followed by substitution into the dynamics equations (1), the following effective dynamics is obtained

$$\begin{array}{l} p_i(t+1) = (1-\tau)p_i(t) + (1-\zeta_i(t))(1-p_i(t)-q_i(t)) \\ q_i(t+1) = (1-\mu)q_i(t) + \tau p_i(t) \\ s_i(t) = 1 - p_i(t) - q_i(t). \end{array} \quad , \quad i = 1, \ldots, N \tag{8}$$

2.2. **Fixed points.** Subtitute the fixed-point relation

$$p_i(t+1) = p_i(t) = p_i, \quad q_i(t+1) = q_i(t) = q_i, \quad s_i(t+1) = s_i(t) = 1 - p_i - q_i \tag{9}$$

into the dynamics equations (8) to obtain

$$\begin{aligned} 0 &= -\tau p_i + (1-\zeta_i)(1-p_i-q_i) \\ 0 &= -\mu q_i + \tau p_i, \\ s_i &= 1 - p_i - q_i, \\ \zeta_i &= \Pi_{j\neq i}[1 - r_{ij}\beta p_j]. \end{aligned}$$

Solving the second equation for $q_i$ and substituting the result into the first equation yields

$$q_i = \frac{\tau}{\mu}p_i, \quad p_i = \frac{1-\zeta_i}{\tau + (1+\tau/\mu)(1-\zeta_i)}$$

The solution for $s_i$ follows directly from the above equations.

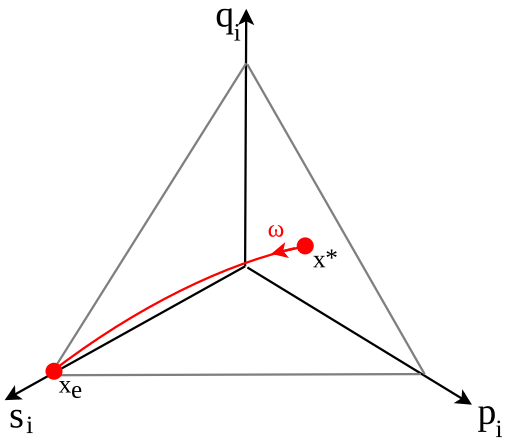


FIGURE 2. Triangle set $T$ (6) containing all trajectories possible for any node $i$ with extinction state $x_e$ (11) (lower left vertex), and the second fixed point which moves with decreasing $\omega$ towards $x_e$.

Introducing the factor

$$\omega_i = (1 - \zeta_i) \tag{10}$$

the fixed-points are given by

$$x_i = [p_i, q_i, s_i]' = \left[\frac{\mu\omega_i}{\mu\tau + (\mu+\tau)\omega_i}, \frac{\tau\omega_i}{\mu\tau + (\mu+\tau)\omega_i}, \frac{\mu\tau}{\mu\tau + (\mu+\tau)\omega_i}\right]$$

where $\omega_i = 1 - \zeta_i$. From this equation, we can identify two solutions, the extinction state

$$x_e = [0, 0, 1]' \tag{11}$$

for $\omega_i = 1 - \zeta_i = 0$, and a survival state $x_s \neq 0$ for $0 < \omega_i = 1 - \zeta_i \leq 1$. Obviously, the maximum value for $\omega_i$ is 1, with associated fixed point

$$x^* = (p^*, q^*, s^*) = \left(\frac{\mu}{\mu+\tau+\mu\tau}, \frac{\tau}{\mu+\tau+\mu\tau}, \frac{\tau\mu}{\mu+\tau+\mu\tau}\right) \tag{12}$$

This fixed point is the same for any node $i$ and moves towards $x_e$ for decreasing $\omega_i$ (see Figure 2), with $x_{si}$ not necessarily being equal to $x_{sj}$ for $i \neq j$ when $0 < \omega_i < 1$.

## 3. System dynamics

In this section, the basic dynamic properties of system (1) are analyzed and sufficient conditions for the global asymptotic stability of the extinction state are derived.

As an illustrative step, consider the dynamics of trajectories born close to the extinction state $x_e$, meaning that

$$p_i(t) \approx \epsilon_i(t), \quad q_i(t) \approx \theta_i(t), \quad 0 \leq \epsilon_i, \theta_i \ll 1. \tag{13}$$

Substituting (13) into the dynamics (8) and neglecting higher order terms, the following local approximation is obtained

$$\begin{array}{l} \epsilon_i(t+1) \approx \beta \sum_{j \neq i} r_{ij} \epsilon_i(t) + (1-\tau) p_i(t) \\ \theta_i(t+1) \approx \tau \epsilon_i(t) + (1-\mu)\theta_i(t) \end{array} \tag{14}$$

or equivalently in vector form

$$\left[ \begin{array}{c} \epsilon(t+1) \\ \theta(t+1) \end{array} \right] \approx \left[ \begin{array}{cc} (1-\tau)\mathbf{I}_N + \beta \mathbf{R} & 0 \\ \tau \mathbf{I}_N & (1-\mu)\mathbf{I}_N \end{array} \right] \left[ \begin{array}{c} \epsilon(t) \\ \theta(t) \end{array} \right] \tag{15}$$

where

$$\epsilon(t) = (\epsilon_1(t), \ldots, \epsilon_N(t))', \quad \theta(t) = (\theta_1(t), \ldots, \theta_N(t))'. \tag{16}$$

This linear dynamics are asymptotically stable if and only if the eigenvalues of the associated matrix

$$\mathbf{A} = \left[ \begin{array}{cc} (1-\tau)\mathbf{I}_N + \beta \mathbf{R} & 0 \\ \tau \mathbf{I}_N & (1-\mu)\mathbf{I}_N \end{array} \right] \tag{17}$$

are contained in the open unit circle, i.e.

$$\mathrm{eig}(\mathbf{A}) \in C_1 = \{\omega \in \mathbb{C} : \, |\omega| < 1\}. \tag{18}$$

In terms of the quadratic form associated to the matrix $A$, this condition is equivalent to

$$\langle x, \mathbf{A}x \rangle < \langle x, \mathbf{I}_N x \rangle \,, \tag{19}$$

or equivalently,

$$\langle x, (\mathbf{I} - \mathbf{A})x \rangle > 0 \tag{20}$$

Given the triangular structure of the matrix $\mathbf{A}$ (17), this last condition holds if

$$\tau > \beta \max |\mathrm{eig}(R)|, \quad \mu > 0. \tag{21}$$

Note that this condition is necessary and sufficient for local stability. Actually, given that for small values of $\varepsilon$ the dynamical behavior is dominated by the linear dynamics (15), if the eigenvalues of the matriz $A$ are positive, the linear, and hence the nonlinear dynamics are unstable.

As a preliminary step towards the derivation of global stability conditions, consider the following Lemma.

**Lemma 1.** *The following bound holds for the probability of being infected* (3)

$$1 - \zeta_i(t) \leq \beta \sum_{j \neq i} r_{ij} p_j(t). \tag{22}$$

**Proof:** Given that $0 \leq \beta, r_{ij}, p_j \leq 1$ the following holds:

$$\begin{aligned} \zeta_i(t) =& \Pi_{j \neq i}[1 - \beta r_{ij} p_j(t)] = 1 - \beta \sum_{j \neq i} r_{ij} p_j(t) + \mathcal{O}^2(p) - \mathcal{O}^3(p) \\ \geq& 1 - \beta \sum_{j \neq i} r_{ij} p_j(t). \end{aligned}$$

**QED.**

The next theorem generalizes the preceding local result for the extinction fixed point stability, and states sufficient conditions for the global extinction.

**Theorem 1.** *Consider the SIQ model* (1). *The extinction state* $x_e = (1, 0, 0)'$ (11) *is a global attractor for any node* $i$ *in the set* $T$ *if the following conditions hold*

$$\tau > \beta \max |eig(R)|, \quad \mu > 0. \tag{23}$$

*If* $\tau < \beta \max |eig(R)|$ *and* $\mu \geq 0$, *the extinction state state is a repulsor.*

**Proof:** Write the dynamics (8) in vector form

$$\begin{bmatrix} p(t+1) \\ q(t+1) \end{bmatrix} = \begin{bmatrix} (1-\tau)\mathbf{I}_N & 0 \\ \tau\mathbf{I}_N & (1-\mu)\mathbf{I}_N \end{bmatrix} \begin{bmatrix} p(t) \\ q(t) \end{bmatrix} + \begin{bmatrix} \phi(t) \\ 0 \end{bmatrix} \tag{24}$$

where

$$\begin{aligned} p(t) =& (p_1(t), \ldots, p_N(t))', \quad q(t) = (q_1(t), \ldots, q_N(t))', \\ \phi(t) =& [(1 - p_1 - q_1)(1 - \zeta_1(t)), \ldots, (1 - p_N - q_N)(1 - \zeta_N(t))]'. \end{aligned}$$

According to the Lemma 1, the solutions of the global nonlinear $p$-dynamics in (24) are bounded as follows

$$\begin{aligned} p_i(t+1) =& (1-\tau)p_i(t) + (1 - p_i(t) - q_i(t))[1 - \zeta_i(t)] \\ =& (1-\tau)p_i(t) + [1 - \zeta_i(t)] \\ \leq& (1-\tau)p_i(t) + \beta \sum_{j=1}^{N} r_{ij}(t)p_j(t) \end{aligned}$$

implying that the solutions $p_i(t)$ are bounded by solutions of the linear dominating dynamics, i.e.

$$p_i(t) \leq z_i(t), \quad z_i(t) = (1-\tau)z_i(t) + \beta \sum_{j=1}^{N} r_{ij}(t)z_j(t), \quad i = 1, \ldots, N. \tag{25}$$

In vector notation (with $z_{2i} = q_i$), the linear dominating dynamics are given by

$$\begin{bmatrix} z_1(t+1) \\ z_2(t+1) \end{bmatrix} = \begin{bmatrix} (1-\tau)\mathbf{I}_N + \beta\mathbf{R} & 0 \\ \tau\mathbf{I}_N & (1-\mu)\mathbf{I}_N \end{bmatrix} \begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix}. \tag{26}$$

or equivalently,

$$z(t+1) = Az(t), \quad z(t) = (z_1'(t), z_2'(t))' \tag{27}$$

with $A$ given in (17). By virtue of this bound for the right hand side of the nonlinear dynamics (24), the following holds

$$\begin{bmatrix} p(0) \\ q(0) \end{bmatrix} = \begin{bmatrix} z_1(0) \\ z_2(0) \end{bmatrix} \quad \Rightarrow \quad \left\| \begin{bmatrix} p(t) \\ q(t) \end{bmatrix} \right\| \leq \left\| \begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix} \right\| \quad \forall\, t \geq 0. \tag{28}$$

This in turn implies that

$$\lim_{t\to\infty} ||(z_1'(t), z_2'(t))'|| = 0 \quad \Rightarrow \quad \lim_{t\to\infty} ||(p'(t), q'(t))'|| = 0. \tag{29}$$

On the other hand, the linear dynamics (27) are asymptotically stable if the conditions (21) hold. Consequently, the conditions (23) are sufficient for the global asymptotic stability of the extinction state $x_e$ (11).

The second affirmation follows directly from the analysis of the linearization (15). **QED.**

It is noteworthy that the first condition (23) of Theorem 1 is the one corresponding to the extinction condition for the SIS dynamics relating the recovery rate to the infection

rate, reported in several studies (see [17, 18] and references therein), but with one important difference: here, the recovery parameter $\mu$ is involved only in the second condition, while the first one relates the rate of internment of infected nodes $\tau$ and the infection probability (depending on $\beta$ and $R$).

When the first condition of Theorem 1 is satisfied, and $\mu > 0$, the rate of convergence to the extinction state will depend on the difference between $\tau$ and the product $\beta \max \mathrm{eig}(\mathbf{A})$, as well as the value of $\mu$, in the sense that for larger values of $\tau - \beta \max \mathrm{eig}(\mathbf{A})$ and $\mu$ the convergence is faster than for small values. This follows directly from the triangular form of the dynamics matrix $\mathbf{A}$ (17) and the fact that the associated linear dynamics represents introduces an upper bound (27) for the nonlinear dynamics (8).

## 4. Numerical simulations

In order to illustrate the preceding theoretic assessments, in this section numerical simulation results are presented for a population of $N = 10.000$ nodes and four representative cases:

(1) $(\tau, \beta, \mu, \lambda) = (0.6, 0.15, 0.5, 1)$ satisfying the stability conditions (23) and illustrating their sufficiency.
(2) $(\tau, \beta, \mu, \lambda) = (0.6, 0.5, 0.5, 1)$: (i) illustrating the sufficiency of the conditions (23), and (ii) verifying the improved convergence rate.
(3) $(\tau, \beta, \mu, \lambda) = (0.6, 0.15, 0.5, 1.5)$ violating the stability conditions (23) due to an increment in the maximum eigenvalue of the matrix $R$ (5).
(4) $(\tau, \beta, \mu, \lambda) = (0.45, 0.15, 0.5, 1)$ violating the stability condition (23) due to insufficient internment rate $\tau$.

The variables represented in the figures are the mean probabilities of being susceptible ($\rho_s$), infected ($\rho_p$), and in quarantine ($\rho_q$) (7).

### 4.1. **Extinction cases.**

The simulation results corresponding to the parameter set

$$(\tau, \beta, \mu, \lambda) = (0.6, 0.15, 0.5, 1) \tag{30}$$

are shown in Figure 3, illustrating the global convergence to the extinction state $x_e$ (in the origin of the $(\rho_s, \rho_q)$-phase portrait).

For comparison purposes, in Figure (4) are presented the trajectories for the parameters $\tau, \beta, \lambda$ given in (30) and two different values of $\mu$: (i) $\mu_1 = 0.15$ (black line), and (ii) $\mu_2 = 0.5$ (grey line), and initial conditions $p_i(0) = s_i(0) = 0.1, q_i = 0.9\, i = 1, \ldots, N$. The simulations illustrate that the convergence rate becomes about three times faster with increasing recovery rate $\mu_2 \approx 3\mu_1$, in accordance to the theoretical developments presented in section 2.2.

### 4.2. **Non-extinction cases.**

For the purpose of illustrating the necessity of the condition pair (23) first consider the parameter set

$$(\tau, \beta, \mu, \lambda) = (0.6, 0.15, 0.5, 1.5) \tag{31}$$

which is different from the previous one (30) due to a higher interaction rate $\lambda$ in the generalized interaction process mechanism (4). The corresponding simulation results are presented in Figure 5, showing the projection of the trajectories onto the $(\rho_p, \rho_q)$-phase plane. As can be seen in Figure 5, due to the higher interaction rate, the extinction state is not reached and the virus survives in the network. This is due to the fact that

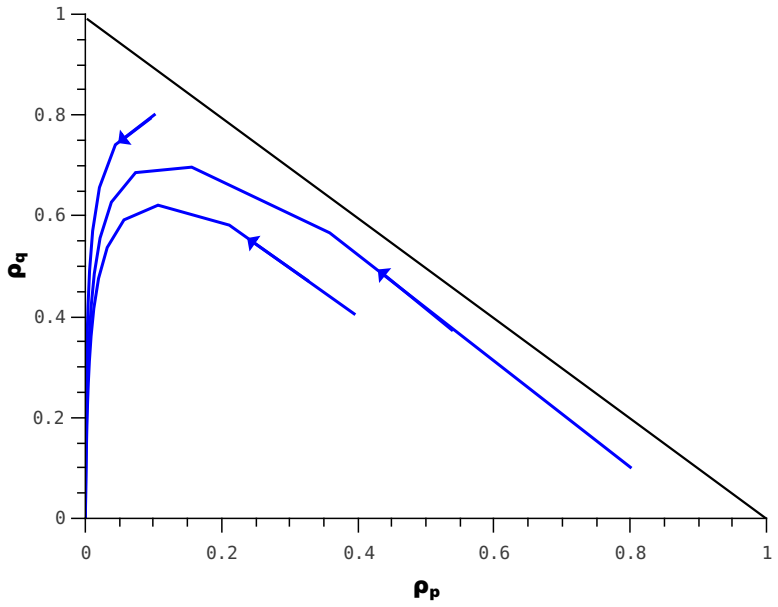


FIGURE 3. Projection onto the $(\rho_p, \rho_q)$-phase plane of the trajectories associated to the parameter set (30).

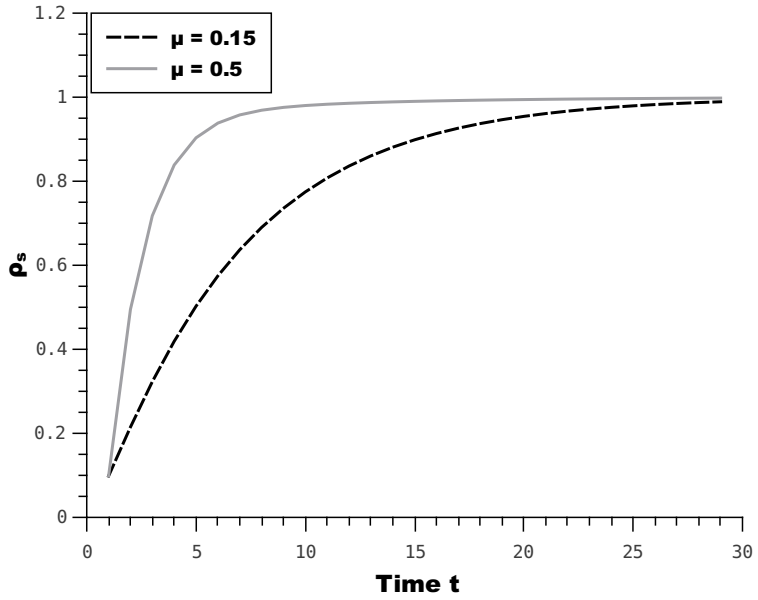


FIGURE 4. Time evolution of the mean probability $\rho_s$ of being susceptible for $(\tau, \beta, \lambda)$ as given in (30) and two different values of the recovery rate; $\mu_1 = 0.15$ (discontinuous black curve), and $\mu_2 = 0.5$ (continuous grey curve).

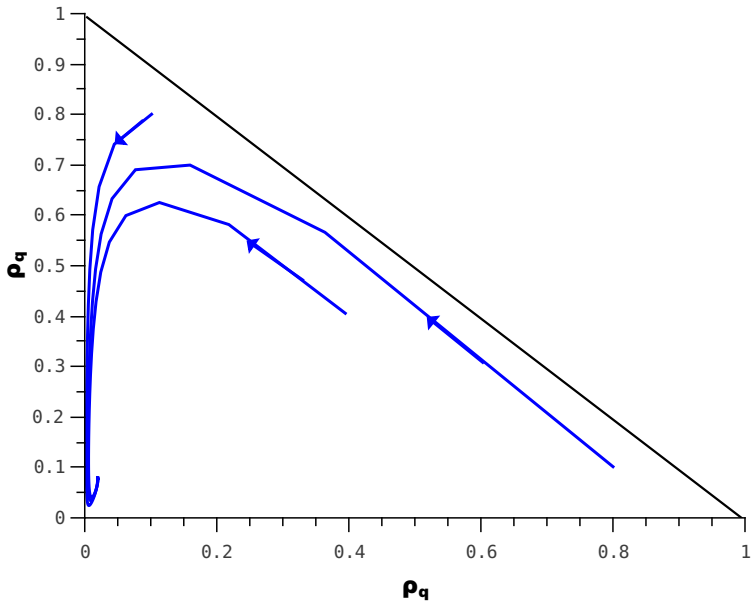


FIGURE 5. Projection onto the $(\rho_p, \rho_q)$-phase plane of the trajectories associated to the parameter set (31).

the eigenvalue of the matrix $R$ (5) increased with the increase in $\lambda$ and the threshold condition (23) is violated.

Next, to illustrate the influence of the relation between $\tau$ and $\beta$ consider the parameter set

$$(\tau, \beta, \mu, \lambda) = (0.45, 0.15, 0.5, 1) \tag{32}$$

which distinguishes from the first one (30) in that $\tau < \beta$. The projection of the corresponding trajectories onto the $(\rho_p, \rho_q)$-phase plane are shown in Figure 6, illustrating that the extinction state $x_e$ is not reached. This behavior is due to the fact that the threshold condition (23) is violated due to an insufficient internment rate $\tau$.

These simulation result illustrate the necessity of the threshold condition (23) and the complex interplay between the system parameters $\tau, \beta, \lambda$.

## 5. Conclusions

The stability of virus transmission in free-scale networks with quarantine mechanism was analyzed. Extinction conditions were derived which are structurally independent of the population size. This analysis provides basic knowledge about the structural dependence of the virus propagation on the internment rate associated to quarantine, and identifies the complex interplay between internment rate, recovery rate, infection probability and interaction frequency (the number of intents made by each node to contact its neighbors). Some basic properties of the associated system spectrum and the basic bifurcation behavior were identified, and numerical simulation studies were presented to illustrate the theoretical findings. The presented analysis is an important preliminary step for the design of control strategies in order to identify optimal quarantine politics.

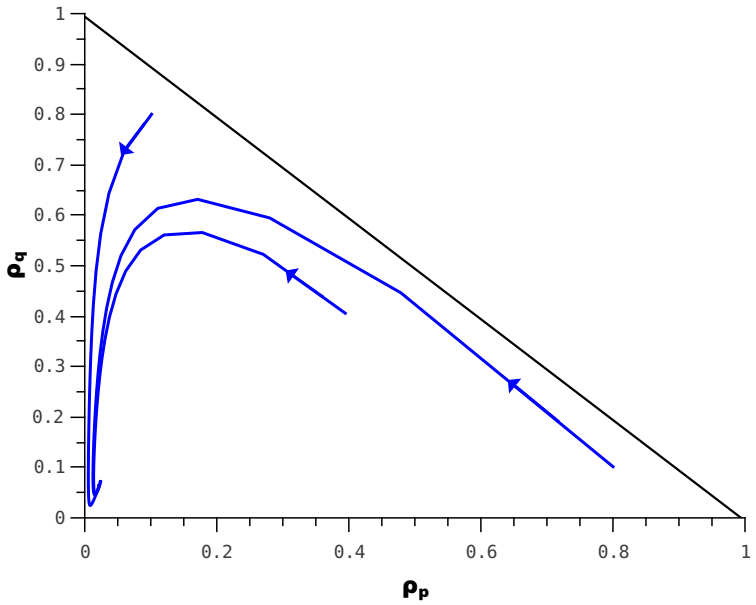


FIGURE 6. Projection onto the $(\rho_p, \rho_q)$-phase plane of the trajectories associated to the parameter set (32).

## REFERENCES

[1] L. Shengbing H. Hethcote, M. Zhien. Effects of quarantine in six endemic models for infectious diseases. *Mathematical Biosciences (Elsevier)*, (180):141–160, 2002.

[2] B. Kumar Mishra and A. Kumar Singh. Two quarantine models on the attack of malicious objects in computer network. *Mathematical Problems in Engineering*, 2012:1–13, 2012.

[3] P. van den Driessche J. Arino, R. Jordan. Quarantine in a multi-species epidemic model with spatial dynamics. *Mathematical Biosciences (Elsevier)*, 2007(206):46–60, 2005.

[4] Z. Feng M. Nuno, C. Castillo-Chavez and M. Martcheva. *Mathematical Models of Influenza: The Role of Cross-Immunity,Quarantine and Age-Structure*. Lectures Notes in Mathematics Springer, chapter 13 edition, 2008.

[5] X. YAN and Y. ZOU. Control of epidemics by quarantine and isolation strategies in highly mobile populations. *International Journal of information and system sciences*, 5(3-4):271–286, 2009.

[6] A. G. McKendrick. Applications of mathematics to medical problems. *In Proceedings of Edin. Math. Society*, vol. 14:98130, 1926.

[7] J. O. Kephart and S. R. White. Directed-graph epidemiological models of computer viruses. *In Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy*, page 343359, 1991.

[8] J. O. Kephart and S. R. White. Measuring and modeling computer virus prevalence. *In Proceedings of the 1993 IEEE Computer Society Symposium on Re- search in Security and Privacy*, page 215, 1993.

[9] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *In Proceedings Sigcomm 1999*, 1999.

[10] M. Ripeanu, I. Foster, and A. Iamnitchi. Mapping the gnutella network: Properties of large- scale peer-to-peer systems and implications for system design. *IEEE Internet Computing Journal*, 6(1), 2002.

[11] Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B*, 26(26):521–529, February 2002.

[12] R. Pastor-Satorras and A. Vespignani. Epidemic dynamics and endemic states in complex networks. *Physical Review E*, (Volume 63, Issue 2):0661171–0661178, 2001.

[13] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, (Volume 86, Number 14):3200–3203, 2001.
[14] R. Pastor-Satorras and A. Vespignani. Epidemic dynamics in finite size scale-free networks. *Physical Review E*, (Volume 65):0351081–0351084, 2002.
[15] R. Pastor-Satorras and A. Vespignani. *Epidemics and immunization in scale- free networks*. 2002.
[16] A.L. Barabási and R. Albert. Emergence of scaling in random graphs. *Science*, 286:509–512, 1999.
[17] S. Gomez, A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno. Discrete-time Markov chain approach to contact-based disease spreading in complex networks. *EPL*, 89(26):38009p1–38009p6, February 2010.
[18] J. Leskovec, D. Chakrabarti, C. Faloutsos, S. Madden, C. Guestrin, and M. Faloutsos. Information survival threshold in sensor and p2p networks. In *IEEE INFOCOM 2007*, 2007.
[19] M. Xie, Z. Jia, Y. Chen, and Q. Deng. Simulating the spreading of two competing public opinion information on complex network. *Applied Mathematics*, (3):1074–1078, 2012.
[20] Q. Chen, H. Chang, S. Jamin R. Govindan, S. Shenker, and W. Willinger. The origins of power-laws in internet topologies revisited. In *IEEE INFOCOM 2002*, 2002.
[21] A.L. Barabási and R. Albert. Edge overload breakdown in evolving networks. *Physical Review E*, 66, 2002.
[22] M. Mihaila, C. Papadimitriou, and A. Saberic. On certain connectivity properties of the internet topology. *Journal of Computer and System Sciences*, 72:239–251, 2006.
[23] C. Borgs, J. Chayes, A. Ganesh, and A. Saberi. How to distribute antidote to control epidemics. *Random Structures & Algorithms John Wiley & Sons, Inc. New York, NY, USA*, (Volume 37, Issue 2):204–222, 2010.
[24] A. Dcmers, D. Greene, C. Hauser, W. Irish, Howard Sturgis Dan Swinehart J. Larson, Scott Shenker, and Doug Terry. Epidemic algorithms for replicated database maintenance. In *In Proceedings of the sixth annual ACM Symposium on Principles of distributed computing*, 1987.
[25] Y. Wan, S. Roy, and A. Saberi. Network design problems for controlling virus spread. In *In Proceedings of the 46th IEEE Conference on Decision and Control*, 2007.
[26] F. Radicchi, J.J. Ramasco, A. Barrat, and S. Fortunato. Complex networks renormalization: Flows and fixed points. *Physical Review Letters*, (Volume 65):1487011–1487014, 2008.

# BREAKING NONLINEARITY WITH PARTIAL BLEACHING SIMPLIFIES THE ANALYSIS OF BIOMOLECULE TRANSPORT OBSERVATIONS IN INTACT CELLS.

EMILIANO PÉREZ IPIÑA AND SILVINA PONCE DAWSON

Abstract. Diffusion is one of the main transport processes that occur inside cells determining the spatial and time distribution of relevant action molecules. In most cases these molecules not only diffuse but also interact with others as they get transported. Under certain conditions the net resulting transport is still approximately diffusive but with an *effective* (concentration-dependent) rather than *free* diffusion coefficient. By fluorescently-labeling the biomolecules it is possible to use optical techniques to infer the rate at which this net transport occurs in intact cells. Interpreting the experimental results is complicated since effective coefficients are not unique. In this paper we discuss how some coefficients are derived from intrinsically linear problems while others are not. We then show how, when probing transport in intact cells with optical techniques, we can obtain one or another type of coefficient by partially bleaching the population of tagged biomolecules. This provides a useful tool to extract more quantitative information from a living system with minimum disruption.

## 1. Introduction

Diffusion is key to many physiologically relevant processes. The transport of information within cells usually involves changes in the concentration of signaling agents. These messengers most likely diffuse inside the cells but also interact with other species as they move. By fluorescently-labeling the messengers it is possible to use optical techniques to infer their transport rate in intact cells. Ideally, one would be willing to follow individual molecules as they get transported. However, this is not possible in many relevant cases. The problem then becomes nonlinear since the biomolecules compete for the same interacting partners. This is reflected in the fact that the net transport over a sufficiently large observation volume is characterized by *effective diffusion coefficients* that depend on *free* diffusion coefficients, reaction rates and *concentrations*. As shown in [4], effective coefficients are not unique. The *single molecule* coefficient, $D_t$, determines the time-dependence of a single (tagged) molecule mean-square displacement. The *collective* coefficient, $D_u$, determines the rate at which concentration inhomogeneities spread out with time. They are both concentration-dependent weighted averages of the free diffusion cofficients of the species involved but they can take on arbitrarily different numerical values depending on the problem. In this paper we discuss how the single molecule coefficient is derived from intrinsically linear problems while the collective one comes from a nonlinear problem that is subsequently linearized. We then show how, when probing transport in intact cells with optical techniques, we can obtain one or the other coefficient by partially bleaching the population of tagged biomolecules.

Two optical techniques that are commonly used in intact cells to estimate diffusion coefficients of biomolecules are Fluorescence Recovery After Photobleaching (FRAP) [1]

and Fluorescence Correlation Spectroscopy (FCS) [2]. In FRAP the fluorescence is *bleached* ("deleted") inside a region and from its subsequent recovery the rate at which marked biomolecules diffuse back into the region can be inferred. FCS monitors fluorescence fluctuations in a small observation volume. Computing the autocorrelation function of the fluctuations around the mean the correlation timescales can be inferred and the diffusion coefficients associated to them can be derived. These techniques give either free or effective coefficients depending on whether the tagged biomolecules diffuse freely or interact with partners. As shown in [3, 4], FRAP gives $D_t$ and as shown in [5], FCS gives $D_u$ and, in certain cases, $D_t$ as well [5]. In what follows we show in which circumstances the transport dynamics is described by either one of these coefficients and how partial photobleaching allows to "highlight" one or the other experimentally. To this end we consider the simplest possible model which still gives an useful insight into a problem of great relevance for the quantitation of biological observations.

## 2. The underlying biophysical model

We consider the simplest possible model with species that diffuse and react, some of which are fluorescent. Namely, we assume that there are free particles, $P_f$, "traps" or binding sites, $S$, and bound particles, $P_b$. Free and bound particles can either be tagged (*i.e.*, fluorescent, indicated with the superscript $t$) or untagged (indicated with the superscript $u$). Both tagged and untagged particles interact with $S$ according to the scheme:

$$\begin{aligned} P_f^t + S &\underset{k_{off}}{\overset{k_{on}}{\longleftrightarrow}} P_b^t, \\ P_f^u + S &\underset{k_{off}}{\overset{k_{on}}{\longleftrightarrow}} P_b^u. \end{aligned} \tag{1}$$

Free particles, bound particles and traps diffuse with free coefficients $D_f$, $D_S$ and $D_S$, respectively. It is implicit in the latter that $S$ is massive enough so that the diffusion rate of a single $S$ molecule or of a bound particle, $P_b$, is the same. The evolution equations for the concentrations, $P_f^t$, $P_b^t$, $P_f^u$, $P_b^u$ and $S$, are then given by:

$$\begin{aligned} \frac{\partial P_f^t}{\partial t} &= D_f \nabla^2 P_f^t - k_{on} P_f^t S + k_{off} P_b^t, \\ \frac{\partial P_b^t}{\partial t} &= D_S \nabla^2 P_b^t + k_{on} P_f^t S - k_{off} P_b^t, \\ \frac{\partial S}{\partial t} &= D_S \nabla^2 S - k_{on}(P_f^t + P_f^u) S + k_{off}(P_b^t + P_b^t), \\ \frac{\partial P_f^u}{\partial t} &= D_f \nabla^2 P_f^u - k_{on} P_f^u S + k_{off} P_b^u, \\ \frac{\partial P_b^u}{\partial t} &= D_S \nabla^2 P_b^u + k_{on} P_f^u S - k_{off} P_b^u. \end{aligned} \tag{2}$$

Both for FRAP and FCS the spatially uniform equilibrium solution, $P_{feq}$, $P_{beq}$, $S_{eq}$, is relevant. It satisfies:

$$\begin{aligned} P_{feq} S_{eq} &= K_D P_{beq}, \\ P_{feq} + P_{beq} &= P_T, \\ S_{eq} + P_{beq} &= S_T, \end{aligned} \tag{3}$$

where $K_D \equiv k_{off}/k_{on}$ and $P_T$ and $S_T$ are the total concentrations of particles and binding sites, respectively. Tagged and untagged equilibrium concentrations are such that $P_{feq} = P^t_{feq} + P^u_{feq}$, $P_{beq} = P^t_{beq} + P^u_{beq}$, $P^t_{feq} S_{eq} = K_D P^t_{beq}$, and $P^u_{feq} S_{eq} = K_D P^u_{beq}$.

2.1. **The equations of FCS.** FCS monitors fluorescence fluctuations in a small observation volume which is determined by how the sample is illuminated. From the analysis of these fluctuations it is possible to estimate the rate at which the fluorescent species and its interacting partners enter and leave the observation volume. In order to quantify these rates it is necessary to have a theoretical model of the main quantity that is obtained with the experiments (the auto-correlation function). In the theory, the intensity distribution of the illumination spot is assumed to be given by:

$$I(\mathbf{r}) = I(0) e^{-\frac{2r^2}{w_r^2}} e^{-\frac{2z^2}{w_z^2}}, \tag{4}$$

where $I(0)$ is the illumination intensity at $\mathbf{r} = 0$, $(r, z)$ are cylindrical coordinates with $z$ the spatial coordinate along the beam propagation direction and $r$ a radial coordinate in the perpendicular plane. $w_z$ and $w_r$ are the sizes of the beam waist along $z$ and $r$ respectively, in general, $w_z > w_r$. The fluorescence collected from the illuminated volume at any given time, $F(t)$, is then related to the number of fluorescent molecules that are inside the volume at that time. To be more specific, in the case of the simple model considered in this paper, $F(t)$ is given by:

$$F(t) = \int Q I(\mathbf{r}) \left( P^t_f(\mathbf{r}, t) + P^t_b(\mathbf{r}, t) \right) d^3 r, \tag{5}$$

if both free and bound particles have the same photophysical properties. In Eq. (5) the concentrations are computed at time, $t$, and spatial point, $\mathbf{r}$ and the parameter, $Q$, takes into account the detection efficiency, the fluorescence quantum yield and the absorption cross-section at the wavelength of excitation of all the fluorescent particles.

Fluctuations around the mean fluorescence, $\langle F(t) \rangle$, are characterized by the time-averaged autocorrelation function (ACF) which is given by:

$$G(\tau) = \frac{\langle \delta F(t) \delta F(t+\tau) \rangle}{\langle F(t) \rangle^2}, \tag{6}$$

where $\delta F(t) = F(t) - \langle F(t) \rangle$. For the model considered here, an expression for $G(\tau)$ can be written in terms of the solutions of Eqs. (2) linearized around the equilibrium solution. These linearized equations read:

$$\frac{\partial \delta P^t_f}{\partial t} = D_f \nabla^2 \delta P^t_f - k_{on}(S_{eq} \delta P^t_f + P^t_{feq} \delta S) + k_{off} \delta P^t_b, \tag{7}$$

$$\frac{\partial \delta P^t_b}{\partial t} = D_S \nabla^2 \delta P^t_b + k_{on}(S_{eq} \delta P^t_f + P^t_{feq} \delta S) - k_{off} \delta P^t_b, \tag{8}$$

$$\frac{\partial \delta S}{\partial t} = D_S \nabla^2 \delta P^t_b - k_{on}(S_{eq} \delta P_f + P_{feq} \delta S) + k_{off} \delta P_b, \tag{9}$$

$$\frac{\partial P^u_f}{\partial t} = D_f \nabla^2 \delta P^u_f - k_{on}(S_{eq} \delta P^u_f + P^u_{feq} \delta S) + k_{off} \delta P^u_b, \tag{10}$$

$$\frac{\partial \delta P^u_b}{\partial t} = D_S \nabla^2 \delta P^u_b + k_{on}(S_{eq} \delta P^u_f + P^u_{feq} \delta S) - k_{off} \delta P^u_b, \tag{11}$$

where $P_f = P^t_f + P^u_f$, $P_b = P^t_b + P^u_b$, $\delta P^{t,u}_{f,b} = P^{t,u}_{f,b} - P^{t,u}_{f,b,eq}$ and $\delta S = S - S_{eq}$. In order to compute $G(\tau)$ the solution of Eqs. (7)–(11) is computed in Fourier space and written in terms of branches of eigenvalues, $\lambda(\mathbf{q})$, and eigenvectors, $\chi(\mathbf{q})$, where $\mathbf{q}$

is the wavenumber vector, *i.e.*, the variable in Fourier space conjugate to the spatial coordinate, **r** [6, 5]. If all particles are fluorescent (*i.e*, if $P_f^u = P_b^u = 0$ and $P_f = P_f^t$, $P_b = P_b^t + P_b^u$), the dynamics is described by Eqs. (7)–(9) with $P_f = P_f^t$ and $P_b = P_b^t$. Fitting the theoretical ACF to the one derived from the experiments it is possible to infer a series of timescales from which some transport rates can be estimated.

2.2. **The equations of FRAP.** In FRAP the system is at equilibrium and, at $t = 0$, the fluorescence in a region is turned off (the particles are *bleached* by means of a very intense laser beam). From the point of view of the system variables this means that $P_f^t(t = 0) = P_b^t(t = 0) = 0$ in the photobleached region. The equilibrium condition, on the other hand, implies that $P_f^u(t = 0) + P_f^t(t = 0) = P_{feq}$ everywhere in space. Thus, $\partial(P_f^u + P_f^t)/\partial t|_{t=0} = \partial(P_b^u + P_b^t)/\partial t|_{t=0} = \partial S/\partial t|_{t=0} = 0$ so that $S = S_{eq}$, $P_f^u + P_f^t = P_{feq}$, and $P_b^u + P_b^t = P_{beq}$ for all time. Therefore, the 5 nonlinear coupled equations describing the evolution of $P_f^u$, $P_f^t$, $P_b^u$, $P_b^t$ and $S$ reduce analytically, as in [3, 4], to the following linear equations:

$$\frac{\partial P_f^t}{\partial t} = D_f \nabla^2 P_f^t - k_{on} P_f^t S_{eq} + k_{off} P_b^t, \tag{12}$$

$$\frac{\partial P_b^t}{\partial t} = D_S \nabla^2 P_b^t + k_{on} P_f^t S_{eq} - k_{off} P_b^t, \tag{13}$$

$$\frac{\partial P_f^u}{\partial t} = D_f \nabla^2 P_f^u - k_{on} P_f^u S_{eq} + k_{off} P_b^u, \tag{14}$$

$$\frac{\partial P_b^u}{\partial t} = D_S \nabla^2 P_b^u + k_{on} P_f^u S_{eq} - k_{off} P_b^u, \tag{15}$$

while $S = S_{eq}$ everywhere in space for all times. In the experiments the fluorescence in a localized region is bleached and is subsequently monitored. From its recovery with time the net transport rate of the fluorescent species is inferred. In order to estimate this rate it is necessary to fit the observations with an analytical expression. The latter can be obtained for the model by computing Eq. (5) over the observed region.

## 3. Effective diffusion coefficients and theories of FCS and FRAP

3.1. **FCS when all particles are fluorescent.** If all particles are fluorescent, the branches of eigenvalues of Eqs. (7)–(9) are:

$$\begin{aligned} \lambda_1 &= -D_S q^2 \\ \lambda_2 &= -\frac{1}{2}\left(k_{off} + k_{on}(P_{feq} + S_{eq}) + (D_S + D_f)\, q^2\right) + \frac{1}{2}\sqrt{\cdot} \\ \lambda_3 &= -\frac{1}{2}\left(k_{off} + k_{on}(P_{feq} + S_{eq}) + (D_S + D_f)\, q^2\right) - \frac{1}{2}\sqrt{\cdot}, \end{aligned} \tag{16}$$

with $\sqrt{\cdot} = \sqrt{(D_f - D_S)^2 q^4 + 2q^2 (D_f - D_S)\, k_{off}\, (a - h) + \nu^2}$, $\nu = k_{off}(a + h)$, $a \equiv S_{eq}/K_D$ and $h \equiv S_T/S_{eq}$. Given Eqs. (3) $h$ satisfies $h = 1 + P_f/K_D$ so that $a + h = 1 + S_{eq}/K_D + P_f/K_D$. The first eigenvalue, $\lambda_1$, corresponds to diffusion with the free diffusion coefficient of the traps, $S$. The second one, in the long time or long wavelength ($q \to 0$) limit is also diffusive:

$$\lambda_2 \approx -D_u q^2 \tag{17}$$

with the effective coefficient that we call *collective* [4]:

$$D_u = \frac{(k_{off} + k_{on}P_{feq})\,D_f + k_{on}S_{eq}D_S}{k_{off} + k_{on}\,(P_{feq} + S_{eq})} = \frac{D_f + \frac{S_{eq}^2}{K_D S_T}D_S}{1 + S_{eq}^2/(K_D S_T)}. \tag{18}$$

This coefficient describes the long time decay dynamics of a small perturbation to the equilibrium solution. The third eigenvalue is not diffusive and describes a rapid exponential decay ($\lambda_3 \approx -(k_{off} + k_{on}S_{eq} + k_{on}P_{feq})$ for $q \to 0$). Under the assumption that the correlation length is much smaller than the distance between fluorescent particles so that fluctuations in the concentrations of the fluorescent species ($\delta C_1 \equiv \delta P_f$, $\delta C_2 \equiv \delta P_b$) satisfy $\langle \delta C_j(\mathbf{r},t)\delta C_k(\mathbf{r},t)\rangle \propto \delta_{jk}\delta(\mathbf{r}-\mathbf{r}')$, $1 \le j,k \le 2$ and that fluctuations in the number of fluorescent particles of a given species follow a Poisson distribution [6], *i.e.*, that $\langle \delta C_j(\mathbf{r},t)\delta C_k(\mathbf{r},t)\rangle = \langle C_j\rangle\delta_{jk}\delta(\mathbf{r}-\mathbf{r}')$, $1 \le j,k \le 2$, $G(\tau)$ can be written as:

$$G(\tau) = G_1(\tau) + G_2(\tau) + G_3(\tau), \tag{19}$$

$$G_1(\tau) = \frac{Go_S}{\left(1 + \frac{\tau}{\tau_S}\right)\sqrt{1 + \frac{\tau}{w^2\tau_S}}} \tag{20}$$

$$G_2(\tau) = \frac{P_{feq}}{4hk_{off}P_T^2}\int \frac{d^3\mathbf{q}}{(2\pi)^3} e^{-W(q)+\lambda_2\tau} \times \left(2\nu + \sqrt{\cdot} + \frac{\nu^2 - (D_S - D_f)^2 q^4}{\sqrt{\cdot}}\right) \tag{21}$$

$$G_3(\tau) = \frac{P_{feq}}{4hk_{off}P_T^2}\int \frac{d^3\mathbf{q}}{(2\pi)^3} e^{-W(q)+\lambda_3\tau} \times \left(2\nu - \sqrt{\cdot} - \frac{\nu^2 - (D_S - D_f)^2 q^4}{\sqrt{\cdot}}\right) \tag{22}$$

where

$$Go_S = \frac{P_{beq}^2}{V_{eff}P_T^2 S_T}, \tag{23}$$

with $V_{eff}$ the effective sampling volume, $\tau_S = w_r^2/(4D_S)$ and $W(\mathbf{q}) \equiv w_r^2 q_r^2/4 + w_z^2 q_z^2/4$ with $q_r$ and $q_z$ the variables in Fourier space that are conjugate to $r$ and $z$, respectively and $q^2 = q_r^2 + q_z^2$, the wavenumber squared. When the observation volume is such that many reactions occur during the typical time it takes for the particles to diffuse out of it, the system is in the "fast reaction limit". In such a case, the ACF can be approximated by [5]:

$$G(\tau) = \frac{Go_S}{\left(1 + \frac{\tau}{\tau_S}\right)\sqrt{1 + \frac{\tau}{w^2\tau_S}}} + \frac{Go_{ef}}{\left(1 + \frac{\tau}{\tau_u}\right)\sqrt{1 + \frac{\tau}{w^2\tau_u}}} \tag{24}$$

where $Go_S$ and $\tau_S$ are the same as before, $\tau_u = w_r^2/(4D_u)$ and $Go_{ef}$ is given by:

$$Go_{ef} = \frac{1}{V_{eff}P_T} - \frac{P_{beq}^2}{V_{eff}P_T^2 S_T}. \tag{25}$$

We see in Eq. (24) that the third component of the ACF is lost in this limit. Thus, the timescales that can be extracted from the ACF (*i.e.*, from the experiments) are diffusive

and associated to the eigenvalues, $\lambda_1$ and $\lambda_2$, from which the free diffusion coefficient of the traps, $D_S$, and the collective diffusion coefficient, $D_u$, can be estimated.

3.2. **FRAP.** FRAP is most commonly used to analyze diffusion in two space dimensions (*e.g.* proteins diffusing on the plasma membrane). In the fast reaction limit and for a circular bleach spot of radius $w$, the fluorescence in this circular region can be written in terms of modified Bessel functions as [3, 7]:

$$F(t) = \exp(-\tau_D/2t)\left(I_o\left(\frac{\tau_D}{2t}\right) + I_1\left(\frac{\tau_D}{2t}\right)\right), \tag{26}$$

where $\tau_D = w^2/D_t$ with

$$D_t = \frac{k_{off}D_f + k_{on}S_{eq}D_S}{k_{off} + k_{on}S_{eq}} = \frac{D_f + \frac{S_{eq}}{K_D}D_S}{1 + S_{eq}/K_D}. \tag{27}$$

If diffusion in three dimensions is considered, instead, the fluorescence can be written as [8]:

$$F(t) = F_0 \sum_{\ell \le 0} \frac{m^{3/2}(-\beta)^\ell}{\ell!} \frac{1}{m + b\ell + (b\ell m t/\tau_D)} \times \frac{1}{\sqrt{m + b\ell + (b\ell m t/(R\tau_D))}} \tag{28}$$

where $\tau_D = w_r^2/(8D_t)$, $R = w_z^2/w_r^2$, $w_r$ and $w_z$ as in (4), $m$ the number of photons required to generate a fluorescence photon, $b$ the number of photons absorbed in a bleaching event, $\beta$ the bleach depth parameter that depends on the bleaching action cross section, the average of the peak intensity at the center of the focal spot and the bleaching pulse duration. In either case, the diffusion coefficient that is recovered is the *single particle* effective coefficient given by Eq. (27) which is also the one that appears in the constant of proportionality between the mean square displacement of a single marked particle and the time elapsed [4].

3.3. **Effective coefficients and optical techniques.** As discussed in [5] and in the previous Section, in the limit of fast reactions, FRAP and FCS (the latter in the case in which all particles are fluorescent) give information on two different effective diffusion coefficients. FRAP allows the estimate of $D_t$ and FCS, when all particles are fluorescent, gives information on $D_u$. $D_t$ and $D_u$ are formally very similar. They both derive from the eigenvalues of a system of equations that are very similar among themselves: Eqs. (7)–(9) with $P_f = P_f^t$ and $P_b = P_b^t$ for FCS and Eqs. (12)–(15) for FRAP. We must first note that the latter is formed by two uncoupled sets: Eqs. (12)–(13) and Eqs. (14)–(15) which solutions only differ in their initial conditions. Thus, for FRAP, there are two relevant branches of eigenvalues which, written in terms of wavenumbers, read:

$$\lambda_{1,2} = -\frac{1}{2}\left(k_{off} + k_{on}S_{eq} + (D_S + D_f)q^2\right) \pm \frac{1}{2}\sqrt{\beta}, \tag{29}$$

with $\beta = \alpha^2 + 2(D_f - D_S)q^2(k_{on}S_{eq} - k_{off}) + (D_f - D_S)^2q^4$ and $\alpha = k_{off} + k_{on}S_{eq}$. Approximations to the eigenvalues in the fast reaction limit can be obtained by taking $q \to 0$. In this limit only $\lambda_1$ is relevant for the recovery of the fluorescence and it is given by:

$$\lambda_1 \approx -\left(\frac{k_{off}D_f + k_{on}S_{eq}D_S}{k_{off} + k_{on}S_{eq}}\right)q^2 = -D_t q^2, \tag{30}$$

with $D_t$ given by Eq. (27). The difference between $D_u$ and $D_t$ or, equivalently, between Eqs. (17) and (30) is due to the intrinsic nonlinearity of the reaction term in Eqs. (2). Then, when these equations are linearized around the equilibrium solution, as done in the theory of FCS, the term $k_{on}S_{eq}\delta P_f + k_{on}P_{feq}\delta S$ arises instead of $k_{on}S_{eq}\delta P_f$, the type of term that appears in the "FRAP" Eqs. (12)–(15) which are linear from the very beginning.

Even though $D_t$ and $D_u$ are formally very similar between themselves and coincide in certain limits, they can also have arbitrarily different numerical values depending on the parameters [4]. In fact, we believe that this potential difference is the reason that underlies the huge discrepancies between the diffusion coefficient of the protein *Bicoid* determined with FRAP [9] and with FCS [10]. Bicoid (Bcd) is one of the most widely studied morphogens. Its distribution is determinant for the organization of the anterior-posterior axis in *Drosophila* embryos [11]. About 80 minutes after egg deposition a stable Bcd gradient is established with larger Bcd concentrations at the anterior pole and an exponential decay towards the posterior end. This exponential distribution is consistent with the so called SDD model in which the protein is synthesized at the anterior end and subsequently diffuses and is degraded throughout the embryo. Within this model the Bcd diffusion coefficient is key to set the timescale over which the Bcd gradient forms and becomes stable. The estimates of this coefficient obtained with FRAP were too small to account for the establishment of the gradient within SDD model and the experimentally observed times [9]. FCS, on the other hand, gave several values one of which was compatible with the SDD model [10]. Our interpretation of these experimental results is based on the theory described in this paper. Namely, we think that both the FRAP and the FCS estimates are correct and that their difference is perfectly understandable in terms of reactions of Bcd with binding sites [12]. This points to the importance of having the right theory to interpret experimental results and quantitate physical parameters of interest.

It is the nonlinearities intrinsic to the reactions that are ultimately responsible for the disparity of the estimates one of which describes the diffusion of individual Bcd molecules (the messengers) and the other one that of their population (the message) [4].

## 4. FCS. Partial bleaching and the recovery of both effective coefficients.

It is clear from the previous Sections that when molecules diffuse and react at least two effective diffusion coefficients characterize their transport dynamics even for the very simple biophysical model given by Eqs. (2). FRAP and FCS, on the other hand, allow the estimate of either one of these coefficients which are weighted (concentration-dependent) averages of the free coefficients of the species involved. The single molecule coefficient, $D_t$, determines the time dependence of the mean square displacement of a single marked particle while the collective one, $D_u$, gives the rate at which a concentration inhomogeneity spreads out with time. If, however, all the particles are distinguishable from the rest (*e.g.* by being fluorescently labeled) the rate at which the inhomogeneity spreads out is determined by $D_t$ instead. We thus see that there is a connection between distinguishability and whether $D_t$ or $D_u$ rules the dynamics. This, in turn, is related to the fact that $D_t$ is derived for intrinsically linear equations while $D_u$ is obtained for nonlinear equations that are linearized around an equilibrium solution. The relationship between linearity and $D_t$ is confirmed by the fact that for a problem in which the

particles simply switch between two states:

$$P_f \underset{k_2}{\overset{k_1}{\rightleftarrows}} P_b \tag{31}$$

and diffuse with $D_f$ while they are in the form $P_f$ and with $D_S$ if they are in the form $P_b$, the diffusive eigenvalue is given by $\lambda = -(k_2 D_f + k_1 D_S)q^2/(k_2 + k_1)$. The connection between linearity and distinguishability suggests that by marking a subset of unmarked particles (or, equivalentely, photobleaching some of the fluorescently labeled particles) one should be able to go from a situation in which $D_u$ rules the dynamics to another one that is ruled by $D_t$. In fact, it was found in [5] that if fluorescent and non-fluorescent particles coexist so that $P^t_{feq} = fP_{feq}$, $P^t_{beq} = fP_{beq}$, in the fast reaction limit, the ACF can be approximated by:

$$\begin{aligned} G(\tau) &= \frac{Go_{coll}}{\left(1+\frac{\tau}{\tau_u}\right)\sqrt{1+\frac{\tau}{w^2\tau_u}}} \\ &+ \frac{Go_{sm}}{\left(1+\frac{\tau}{\tau_t}\right)\sqrt{1+\frac{\tau}{w^2\tau_t}}} \\ &+ \frac{Go_S}{\left(1+\frac{\tau}{\tau_S}\right)\sqrt{1+\frac{\tau}{w^2\tau_S}}}, \end{aligned} \tag{32}$$

where

$$Go_{sm} = \frac{1-f}{V_{ef}P^t_T}, \qquad Go_{coll} = fGo_{ef}, \tag{33}$$

with $Go_{ef}$, $Go_S$, $\tau_S$ and $\tau_u$ defined as before and $\tau_t = w_r^2/(4D_t)$. Thus, having both fluorescent and non-fluorescent particles FCS can give information on both $D_u$ and $D_t$. We now discuss how this can be understood in terms of the (diffusive) eigenvalues of Eqs. (7)–(8).

In order to go on with the discussion it is better to introduce a change of variables. Namely, we will work with $P_f = P^t_f + P^u_f$, $P_b = P^t_b + P^u_b$, $S_T$, $P^t_f$ and $P^t_b$. In these new variables the matrix of the linear problem defined by Eqs. (7)–(8) reads:

(34)
$$\begin{bmatrix} D_f\nabla^2 - k_{on}S_{eq} & k_{off} + k_{on}P_{feq} & -k_{on}P_{feq} & 0 & 0 \\ k_{on}S_{eq} & D_f\nabla^2 - k_{off} - k_{on}P_{feq} & k_{on}P_{feq} & 0 & 0 \\ 0 & 0 & D_S\nabla^2 & 0 & 0 \\ 0 & k_{on}P^t_{feq} & -k_{on}P^t_{feq} & D_f\nabla^2 - k_{on}S_{eq} & k_{off} \\ 0 & -k_{on}P^t_{feq} & k_{on}P^t_{feq} & k_{on}S_{eq} & D_f\nabla^2 - k_{off} \end{bmatrix}.$$

Three blocks determine the branches of eigenvalues:

$$\begin{bmatrix} D_f\nabla^2 - k_{on}S_{eq} & k_{off} + k_{on}P_{feq} \\ k_{on}S_{eq} & D_f\nabla^2 - k_{off} - k_{on}P_{feq} \end{bmatrix}, \tag{35}$$

$$\begin{bmatrix} D_S\nabla^2 & \end{bmatrix}, \tag{36}$$

$$\begin{bmatrix} D_f\nabla^2 - k_{on}S_{eq} & k_{off} \\ k_{on}S_{eq} & D_f\nabla^2 - k_{off} \end{bmatrix}. \tag{37}$$

The one of Eq. (36) gives the eigenvalue associated to $D_S$. The one of Eq. (35) gives the eigenvalue associated to the collective effective coefficient, $D_u$, in the long wavelength limit. The one of Eq. (37) gives the eigenvalue associated to the single molecule effective coefficient, $D_t$, in the long wavelength limit. It is also the matrix that is obtained in FRAP (see Eqs. (12)–(15)). The difference between Eq. (35) and Eq. (37) is solely due to the nonlinearity of the reaction term. Eq. (34) highlights the fact that optical distinguishability breaks up the nonlinearity of the reaction scheme and, in this way, allows the single molecule effective coefficient, $D_t$, to be present in the ACF, as shown in Eq. (32). Furthermore, since the relative weight of the terms of the ACF associated to $D_t$ and $D_u$ depends on the fraction of fluorescent particles, $f$ (see Eq. (32)) photobleaching provides a tool by which the experimentalist can switch between situations in which either $D_u$ or $D_t$ can be extracted from the experiment.

## 5. Discussion and Conclusions

The diffusion of biomolecules plays a relevant role for the transport of information inside cells. Most often, biomolecules do not diffuse freely inside cells but also react with binding sites. This interaction usually introduces nonlinearities in the equations that rule the dynamics of the problem. The net resulting transport that occurs over long times is still approximately diffusive but with *effective* (concentration-dependent) rather than *free* diffusion coefficients. In [4] we showed that two different effective diffusion coefficients, the single particle, $D_t$, and the collective one, $D_u$, can describe this transport dynamics depending on the situation. They are both weighted averages of the free coefficient of the particles, $D_f$, and of the traps, $D_S$, but they can have arbitrarily different numerical values. $D_t$ is the simplest and most intuitive between the two, but which one is obtained from experiments depends on the experimental situation. Diffusion rates in cells can be estimated experimentally using optical techniques and fluorescentely tagged biomolecules. Two widely used such techniques are FRAP and FCS. In [5] we compared which effective coefficients can be estimated with each of these techniques when the biomolecules diffuse and react with non-fluorescent "traps". In FRAP, fluorescent and non-fluorescent versions of the moelcules of interest coexist and the technique estimates the single molecule coefficient, $D_t$. FCS gives the free trap diffusion coefficient, $D_S$, and $D_u$ if only the fluorescent version of the particles is present and it gives $D_t$ as well if non-fluorescent particles are present too. We have discussed in this paper how the coexistence of fluorescent and non-fluorescent molecules of the same species uncouples the variables of the problem and, in this way, the relevant eigenvalues that rule the eventual diffusive dynamics go from depending on the collective effective coefficient to depending on the the single particle one. Roughly speaking, distinguishability breaks the non-linear coupling. Therefore, photobleaching a subset of the fluorescently labeled particles in FCS experiments one can go from a situation in which only $D_u$ can be estimated to another in which mainly information on $D_t$ can be extracted. This provides a useful tool to extract more quantitative information from a living system with minimum disruption. There has been some controversy lately on the rate at which, Bicoid, a key morphogen for the establishment of the dorso-ventral axis in fly embryos, diffuses [9, 10]. In our view [12], a mechanistic underlying theory as the one analyzed in this paper allows to interpret apparently disparate results within a unified framework. This points to the importance of having dynamical biophysical models to interpret and quantitate biological experiments.

**Acknowledgements:** This research has been supported by UBA (UBACyT 20020100100064), ANPCyT (PICT 2010-1481 and PICT 2010-2767), CONICET (PIP 5131). We acknowledge useful conversations with L. Sigaut and A. Colman-Lerner.

## References

[1] Axelrod, D., Ravdin, P., Koppel D.E., Schlessinger J. , Webb, W.W., Elson, E.L. and Podelski, T.R. *Lateral motion of fluorescently labeled acetycholine receptors in membranes of developing muscle-fibers* Proc. Natl. Acad. Sci (USA) **73** (1976) pp. 4594 – 4598

[2] Magde, D., Elson, E. and Webb, W. W. *Thermodynamic Fluctuations in a Reacting System Measurement by Fluorescence Correlation Spectroscopy* Phys. Rev. Lett **29** (1972) pp. 705-708

[3] Sprague, B.L., Pego, R.L., Stavreva, D.A. and McNally, J.G. *Analysis of Binding Reactions by Fluorescence Recovery after Photobleaching*Biophys. J. **86** (2004) pp. 3473-3495

[4] Pando, B., Dawson, S. P., Mak, D. O. D. and Pearson, J. E. *Messages diffuse faster than messengers.* Proc. Natl. Acad. Sci (USA) **103** (2006) pp. 5338-5342

[5] Sigaut, L., Ponce, M. L., Colman-Lerner, A. and Dawson, S. P. *Optical techniques provide information on various effective diffusion coefficients in the presence of traps.* Phys. Rev. E **82** (2010) 051912, 11 pages

[6] Krichevsky, O. and Bonnet, G. *Fluorescence correlation spectroscopy: the technique and its applications* Rep. Prog. Phys. **65** (2002) pp. 251-297

[7] Soumpasis, D.M. *Theoretical analysis of fluorescence photobleaching recovery experiments* Biophys. J. **41** (1983) pp. 95-97

[8] Brown, EB, Wu, ES, Zipfel, W and Webb, WW *Measurement of Molecular Diffusion in Solution by Multiphoton Fluorescence Photobleaching Recovery* Biophys. J. **77** (1999) pp. 2837-2849

[9] Gregor, T., Wieschaus, E. F., McGregor, A. P., Bialek, W. and Tank, D. W. *Stability and nuclear dynamics of the bicoid morphogen gradient.* Cell **130** (2007) pp. 141-152

[10] Abu-Arish, A., Porcher, A., Czerwonka, A., Dostatni, N. and Fradin, C. *High Mobility of Bicoid Captured by Fluorescence Correlation Spectroscopy: Implication for the Rapid Establishment of Its Gradient.* Biophys. J. **99** (2010) pp. L33-L35

[11] Driever, W. and Nussleinvolhard, C. *A Gradient of Bicoid Protein in Drosophila Embryos.* Cell **54** (1988) pp. 83-93

[12] Sigaut, L., Pearson, J.E., Colman-Lerner, A. and Dawson, S. P. *Messages do diffuse faster than messengers: FCS and FRAP yield consistent estimates of Bcd effective diffusion.* (to be published)

Departamento de Física, FCEN-UBA e IFIBA, CONICET-UBA
*E-mail address*: emperipi@df.uba.ar

Departamento de Física, FCEN-UBA e IFIBA, CONICET-UBA
*E-mail address*: silvina@df.uba.ar

Las Publicaciones Matemáticas del Uruguay (PMU) tienen como objetivo reflejar parte de las actividades de investigación matemática que se lleva a cabo en Uruguay. Nuestro interés es publicar artículos de investigación, así como artículos tipo survey, anuncios, y otros trabajos que el comité editorial considere adecuado. Los volúmenes no necesariamente serán arbitrados. Esto se indicará cuidadosamente en cada volumen.

---

Todos los artículos de este volumen han sido arbitrados.

---

The goal of Publicaciones Matemáticas del Uruguay (PMU) is to reflect part of the mathematical research activities that takes place in Uruguay. It is our interest to publish research articles, survey-type articles, research announcements and other papers considered suitable by the Editorial Board. The editorial process may or may not involve a revision by referees. This will be carefully indicated in each volume.

---

All papers in this volume have been revised by referees.

---

## CONTENTS