Random Walks and Geometry

Random Walks and Geometry

Proceedings of a Workshop
at the Erwin Schrödinger Institute
Vienna
June 18 – July 13, 2001

Editor
Vadim A. Kaimanovich

in collaboration with
Klaus Schmidt and Wolfgang Woess

Walter de Gruyter · Berlin · New York

Editor
Vadim A. Kaimanovich
IRMAR
UMR 6625 du C.N.R.S.
Université de Rennes-1
Campus de Beaulieu
35042 Rennes Cedex
France
e-mail: kaimanov@univ-rennes1.fr

Mathematics Subject Classification 2000:
22D40, 37H15, 43A05, 58J65, 60B99, 60J45, 82B20

Keywords:
random walks on groups and graphs, Markov processes, random matrices, Lyapunov exponents, harmonic functions, stochastic Loewner evolution, expander graphs, quantum chaos, spectral theory, cellular automata, random number generators

♾ Printed on acid-free paper which falls within the guidelines of the ANSI to ensure permanence and durability.

Library of Congress Cataloging-in-Publication Data

Random walks and geometry : proceedings of a workshop at the Erwin Schrödinger Institute, Vienna, June 18−July 13, 2001 / editor, Vadim A. Kaimanovich, in collaboration with Klaus Schmidt and Wolfgang Woess.
p. cm.
English, with one contribution in French.
ISBN 3-11-017237-2 (acid-free paper)
1. Random walks (Mathematics) − Congresses. 2. Geometry − Congresses. I. Kaimanovich, Vadim A. II. Schmidt, Klaus, 1943− III. Woess, Wolfgang, 1954−
QA274.73.R37 2004
519.2′82−dc22
2004043902

ISBN 3-11-017237-2

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

© Copyright 2004 by Walter de Gruyter GmbH & Co. KG, 10785 Berlin, Germany.
All rights reserved, including those of translation into foreign languages. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording or any information storage and retrieval system, without permission in writing from the publisher.

Printed in Germany.
Cover design: Thomas Bonnie, Hamburg.
Typeset using the authors' T_EX files: I. Zimmermann, Freiburg.
Printing and binding: Hubert & Co. GmbH & Co. KG, Göttingen.

Dedicated to the memory of Martine Babillot

Preface

This volume is an outcome of the special semester *2001 - Random Walks* held at the Schrödinger Institute in Vienna, Austria, from February until July 2001. It was dedicated to various problems related to stochastic processes on geometric and algebraic structures, with an emphasis on their interplay, and also on their interaction with theoretical physics. Some of the focal points were: probability on groups; products of random matrices and the Lyapunov spectrum; boundary behaviour, harmonic functions and other potential theoretic aspects; Brownian motion on manifolds; combinatorial and spectral properties of random walks on graphs; random walks and diffusion on fractals. There were two separate main periods of activity in the first (February/March) and in the second (May/June/July) halves of the semester. The first period started with a two-week workshop with the general theme *Random Walks and Statistical Physics* (February 19–March 2). Towards the end of the second period there was another workshop with the general theme *Random Walks and Geometry* which lasted for almost a month (June 18–July 13).

The papers collected in this volume are (with a couple of exceptions) contributed by the participants of the second workshop and show how the ideas connected with Markov chains on geometric and algebraic structures permeate such different subjects as hyperbolic geometry, Lie groups, geometric group theory, cellular automata, graph theory, random number generators, percolation, and statistical physics. Among these papers are both surveys and original research articles. All of them have been thoroughly refereed and proofread.

Fruitful complementary interaction between the geometry and randomness is a common feature and unifying link between all the contributions, and we are glad to present this panorama of recent work in the rapidly growing area at the crossroads of several mathematical disciplines.

We are grateful to the Erwin Schrödinger International Institute for Mathematical Physics in Vienna for generous financial support and for creating excellent working atmosphere during our special semester.

When the work on this volume was almost finished we learned about the untimely death of Martine Babillot caused by a foudroyant and devastating disease. Martine's bright personality, with her ability to synthesize different points of view and approaches, was very close to the spirit of our program, of which she was an active participant. She finished proofreading her contribution to the Proceedings just several weeks before passing away. We dedicate this volume to her memory.

March 2004 *Vadim A. Kaimanovich, Klaus Schmidt and Wolfgang Woess*

Table of contents

Preface vii

Surveys and longer articles

Abbas Alhakim and Stanislav Molchanov
Some Markov chains on abelian groups with applications 3

Raffaella Burioni, Davide Cassi and Alessandro Vezzani
Random walks and physical models on infinite graphs: an introduction 35

Tullio Ceccherini-Silberstein, Francesca Fiorenzi and Fabio Scarabotti
The Garden of Eden Theorem for cellular automata and for symbolic dynamical systems 73

Alex Gamburd
Expander graphs, random matrices and quantum chaos 109

Rostislav I. Grigorchuk and Andrzej Żuk
The Ihara zeta function of infinite graphs, the KNS spectral measure and integrable maps 141

Yves Guivarc'h et Émile Le Page
Simplicité de spectres de Lyapounov et propriété d'isolation spectrale pour une famille d'opérateurs de transfert sur l'espace projectif 181

Gregory F. Lawler
An introduction to the Stochastic Loewner Evolution 261

George A. Willis
A canonical form for automorphisms of totally disconnected locally compact groups 295

Research communications

Martine Babillot
On the classification of invariant measures for horosphere foliations on nilpotent covers of negatively curved manifolds 319

Martin T. Barlow and Steven N. Evans
Markov processes on vermiculated spaces 337

Laurent Bartholdi
Cactus trees and lower bounds on the spectral radius of vertex-transitive graphs 349

Enrique Bendito, Ángeles Carmona and Andrés M. Encinas
Equilibrium measure, Poisson kernel and effective resistance on networks 363

Sébastien Blachère
Internal diffusion limited aggregation on discrete groups of polynomial growth 377

Massimo Campanino and Dimitri Petritis
On the physical relevance of random walks: an example of random walks on a randomly oriented lattice 393

Tullio Ceccherini-Silberstein and Fabio Scarabotti
Random walks, entropy and hopfianity of free groups 413

Anna Erschler
Growth rates of small cancellation groups 421

Alex Eskin and Gregory Margulis
Recurrence properties of random walks on finite volume homogeneous manifolds 431

Alessandra Iozzi
On the cohomology of foliations with amenable groupoid 445

Anders Karlsson
Linear rate of escape and convergence in direction 459

Anna Maria Mantero and Anna Zappa
Remarks on harmonic functions on affine buildings 473

Tatiana Nagnibeda
Random walks, spectral radii, and Ramanujan graphs 487

Sam Northshield
Cogrowth of arbitrary graphs 501

Laurent Saloff-Coste
Total variation lower bounds for finite Markov chains: Wilson's lemma 515

Surveys and longer articles

Some Markov chains on abelian groups with applications

Abbas Alhakim and Stanislav Molchanov

Abstract. We study limit theorems for the local times of the special Markov chains: "quasi-random walks" on the group W_k of binary words of length k, associated with the Bernoulli schemes. Applications include the asymptotical analysis of the computational complexity (due to recent ideas by Kalman, Pinkus and Singer) and new tests for random number generators.

Contents

1 Introduction 3

2 CLT for Markov chains 7

3 A limit theorem for the approximate entropy 9

4 Analysis of the quadratic form $2^k(B_k\boldsymbol{x}\cdot\boldsymbol{x})$ 12

5 Hierarchical matrices and their diagonalization 18

6 Main result 22

7 The eigenvectors of B_k 28

1 Introduction

Kolmogorov complexity theory [7, 8] gives the logical foundation of probability theory but it can not be applied directly to the testing of "randomness" for specific bit strings. In their recent works [11, 12] Kalman, Pincus and Singer tried to develop "effective" or "computational" complexity concepts in the spirit of Kolmogorov's ideas. We will not discuss here the relationship between the two theories. Our goal is rather to prove several limit theorems based on the notion of "Approximate Entropy", introduced in [11, 12] and to propose new algorithms for the testing of RNG's. These tests are

especially efficient for the physical random number generators (RNG's) where one can expect the presence of short correlations and small deviations from homogeneity [4]. Let

$$S = \{\beta_1, \beta_2, \dots, \beta_n, \dots\} \tag{1.1}$$

be an infinite sequence of independent binary random variables. Moving within S along a frame of width k we can construct a sequence X_t, $t \geq 0$ consisting of overlapping words. Namely put

$$X_0 = (\beta_1, \dots, \beta_k)\,,\ X_1 = (\beta_2, \dots, \beta_{k+1})\,, \dots,\ X_t = (\beta_{t+1}, \dots, \beta_{t+k})\,, \dots\,.$$

The resulting sequence is evidently a homogeneous Markov chain (which we will refer to as a Kalman–Pincus–Singer Markov chain, see below).

We now introduce some notations and definitions. For a fixed integer $k \geq 1$ consider the set W_k of all binary words of length k: $(\beta_1, \dots, \beta_k)$, $\beta_i \in \{0, 1\}$. Obviously $|W_k| = 2^k$, and W_k can be seen as a linear space over the field $\mathbb{Z}_2$.

Recall that the classical symmetric random walk on an abelian group G with unit element e is a Markov chain with invariant transition probabilities (see Saloff-Coste [14]). This means that $P(x, y) = P(gx, gy)$ for any $x, y, g \in G$. In other words $P(x, y) = P(e, x^{-1}y)$ and as a result the transition probabilities depend on the measure $\mu(z) = P(e, z) = P(x_{n+1} = xz | x_n = x)$. The K–P–S Markov chain defined above is not a random walk in this classical sense. However one can see that for all x, y, and $g = (g_1, \dots, g_k) \in W_k$ one has $P(g + x, T(g) + y) = P(x, y)$, (hence $P(x, y) = P(0, y - T(x))$), where $T(g_1, g_2, \dots, g_k) = (g_2, \dots, g_k, g_1)$ is an automorphism of the abelian group W_k. The theory of random walks on groups is well developed especially in the symmetric case where $\mu(g) = \mu(g^{-1})$, i.e., $P = P^*$ (see [14]). As we already mentioned, the K–P–S Markov chain is highly asymmetric. This fact is manifested in the nilpotent property $(P - \Pi)^k = 0$, where Π is the invariant measure of the chain.

In fact, the above viewing of the K–P–S Markov chain as some form of random walk – although not essential for the rest of our discussion in this paper – will provide a good analytical tool as we generalize the current problem to the case where the underlying binary sequence above is replaced with a sequence of uniformly distributed random variables in the interval [0,1]. For the rest of the paper we will identify W_k with the set of integers $\{0, 1, \dots, 2^k - 1\}$ as follows: a word $\xi = \beta_1\beta_2 \dots \beta_k$ is identified with the decimal expansion of the binary number $(\beta_1, \dots, \beta_k)_2$. Note that with this identification, the state space W_k is an abelian group with the operation $\xi_1 \dot{+} \xi_2 := (\xi_1 + \xi_2) \mod 2^k$. This enumeration of the state space allows us to write the transition matrix in a tractable form, see below.

Next, for an arbitrary word $\xi \in W_k$ and a finite sequence of length $t - k + 1$ we define the occupation (local) times as

$$\tau\,(\xi, t) = \#\,\{n \leq t :\ X_n = \xi\}\,,$$

and the corresponding frequencies as

$$\pi(\xi, t) = \frac{\tau(\xi, t)}{t}, \quad \xi \in W_k, \ t \geq 1.$$

The approximate entropy (ApEn) of the finite sequence $X_1, X_2, \ldots, X_t$ (or the bit string $\beta_1, \beta_2, \ldots, \beta_{t+k}$) is given by the formulas

$$\begin{aligned} \Phi_k(t) &= -\sum_{\xi \in W_k} \pi(\xi, t) \log_2 \pi(\xi, t), \\ \mathrm{ApEn}(k, t) &= \Phi_k(t) - \Phi_{k-1}(t), \quad k \geq 2, \\ \mathrm{ApEn}(1, t) &= \Phi_1(t). \end{aligned} \tag{1.2}$$

In the special case when the sequence $\{\beta_t, \ t \geq 1\}$ is a random symmetric Bernoulli scheme, i.e., $\{\beta_t, \ t \geq 1\}$ are i.i.d.r.v. with

$$P(\beta_t = 0) = P(\beta_t = 1) = \frac{1}{2}$$

on some probability space $(\Omega, \mathcal{F}, P)$ we have – due to the strong law of large numbers – that P-a.s.

$$\begin{aligned} \lim_{t \to \infty} \pi(\xi, t) &= 2^{-k} = N^{-1}; \quad \xi \in W_k, \\ \lim_{t \to \infty} \Phi_k(t) &= k; \\ \lim_{t \to \infty} \mathrm{ApEn}(k, t) &= 1, \quad k \geq 1. \end{aligned} \tag{1.3}$$

The assumption that the "random" sequence $\{\beta_t, \ t \geq 1\}$ represents a symmetric Bernoulli scheme will be referred to as the basic hypothesis H_0. The central statistical problem in the study of RNG's is to test H_0 on a given confidence level α using appropriate empirical data. Relations (1.3) are the basis for the following definitions ([12, 11]): the sequence S is asymptotically random if for any $k \geq 1$ the formulas (1.3) are valid. In other words, the computational complexity or randomness due to [12, 11] is equivalent to the "normality" of the sequence S in the classical Borel's sense.

For the practical statistical applications we have to study the Gaussian fluctuations in (1.3) under the hypothesis H_0.

The central point in the further development of our analysis is the straightforward observation that the homogeneous K–P–S Markov chain on W_k has the $2^k \times 2^k$

transition matrix

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

Notation. The following notation will be used throughout this paper.

$$N = 2^k,\ N' = 2^{k-1},\ N'' = 2^{k-2},\ N^{(3)} = 2^{k-3},\ \dots .$$

That is, the natural enumeration which we have introduced earlier is:

$$0 \equiv (0 \dots 0),\ 1 \equiv (0 \dots 1), \dots, N^{'} \equiv (1 \dots 0), \dots, N-1 \equiv (1 \dots 1).$$

This K–P–S chain has several important features which are not typical for a general Markov chain. It is ergodic and aperiodic with a uniform stationary distribution $\pi(\xi) = \frac{1}{N} = 2^{-k}$ (due to the double stochasticity of P). Let Π be the limit of the sequence $P^n,\ n \to \infty$, then

$$P^l \equiv \Pi \equiv \left[\pi_{ij} = \frac{1}{N}\right], \quad l \geq k,\ i, j = 0, \dots, N-1. \tag{1.4}$$

The last fact is a direct consequence of the independence of two non-overlapping k-tuples of the Bernoulli sequence S. In other words, the sequence $X_t,\ t \geq 1$ has a finite radius of correlations $R = k$. Equation (1.4) implies that the matrix $(P - \Pi)$ is nilpotent (i.e., $(P - I)^k = 0$). The stochastic matrix P has the simple eigenvalue $\lambda_1 = 1$, while all other eigenvalues are zero: $\lambda_j = 0;\ 1 < j \leq N$. We stress that rank $P = 2^{k-1} = \frac{N}{2}$, i.e., the Jordan form of P contains Jordan cells.

Remark 1.1. On the space L of all functions f such that $\bar{f} = f \cdot \pi = \frac{1}{N}(f \cdot \mathbf{1}) = 0$, where $\mathbf{1}$ is the constant vector (the vector with all entries being equal to one), the matrix $P - I$ is nonsingular.

So far we have the formulas for P and $P^l \equiv \Pi$, $l \geq k$. For the intermediate powers $1 < s < k$ the structure of P^s is also simple:

$$P^s = \left[\begin{array}{cccccccccc}
2^{-s} & \dots & 2^{-s} & 0 & \dots & 0 & \dots & \dots & 0 & \\
0 & \dots & 0 & 2^{-s} & \dots & 2^{-s} & 0 & \dots & 0 & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
0 & \dots & \dots & \dots & \dots & 0 & 2^{-s} & \dots & 2^{-s} & \\
2^{-s} & \dots & 2^{-s} & 0 & \dots & 0 & \dots & \dots & 0 & \\
0 & \dots & 0 & 2^{-s} & \dots & 2^{-s} & 0 & \dots & 0 & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
0 & \dots & \dots & \dots & \dots & 0 & 2^{-s} & \dots & 2^{-s} & \\
 & & & & \vdots & & \vdots & & & \\
2^{-s} & \dots & 2^{-s} & 0 & \dots & 0 & \dots & \dots & 0 & \\
0 & \dots & 0 & 2^{-s} & \dots & 2^{-s} & 0 & \dots & 0 & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
0 & \dots & \dots & \dots & \dots & 0 & 2^{-s} & \dots & 2^{-s} &
\end{array}\right] \begin{array}{l} \left.\begin{array}{l}\\ \\ \\ \\ \end{array}\right\} N^{(s)} \\ \left.\begin{array}{l}\\ \\ \\ \\ \end{array}\right\} N^{(s)} \\ \\ \left.\begin{array}{l}\\ \\ \\ \\ \end{array}\right\} N^{(s)} \end{array}. \quad (1.5)$$

2 CLT for Markov chains

Let X_t, $t \geq 1$ be an ergodic aperiodic Markov chain on a finite state space $\mathcal{X} = \{x_0, \dots, x_{N-1}\}$ with a transition probability matrix P, a limiting invariant distribution $\pi = \pi P$, and a limiting matrix Π (with all rows equal π). Then for appropriate positive constants γ and c we have:

1) $\left|P^t - \Pi\right| \leq ce^{-\gamma t}$, i.e., $\left|P^t_{ij} - \pi_j\right| \leq ce^{-\gamma t}$, for all $x_i, x_j \in \mathcal{X}$

2) For any function $f : \mathcal{X} \to R$ and an arbitrary initial distribution we have P-a.s.

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} f(x_i) = (f \cdot \pi) = \bar{f}. \quad (2.1)$$

Moreover, if $\bar{f} = (f \cdot \pi) = 0$, then

$$\frac{1}{\sqrt{t}} \sum_{s=1}^{t} f(x_s) \overset{\text{law}}{\Longrightarrow} N\big(0, \sigma^2(f)\big). \quad (2.2)$$

The first two results go back to Markov. The local form of CLT for the occupation times (see below) in maximal generality was proven by Kolmogorov [6]. Any book on finite Markov chains contains these theorems with historical comments (for instance, see Kemeny and Snell [5]).

An elegant proof (rather than the most refined proof), with a formula for the limiting variance $\sigma^2(f)$ is based on the martingale difference approach (Billingsley, see [2]):

If $\bar{f} = 0$ one can solve the homological equation

$$g - Pg = f. \tag{2.3}$$

The solution is unique in the class of the functions $\{g : \bar{g} = (\pi \cdot g) = 0\}$ and given by the formal inversion

$$g = f + Pf + P^2 f + P^3 f + \cdots \tag{2.4}$$

(This formal series in fact converges exponentially fast). Then

$$\begin{aligned} \mathcal{F}_0^t &= f(X_0) + \cdots + f(X_t) = g(X_0) - Pg(X_0) + \cdots + g(X_t) - Pg(X_t) \\ &= g(X_0) - Pg(X_t) + [g(X_1) - Pg(X_0)] + \cdots \\ &\quad + [g(X_t) - Pg(X_{t-1})]. \end{aligned} \tag{2.5}$$

Then the sequence $z_i = g(X_{i+1}) - Pg(X_i)$, $i = 1, 2, \ldots, n-1$, is a bounded square integrable martingale difference. Now equation (2.5) can be written as

$$\mathcal{F}_0^t = g(X_0) - Pg(X_t) + \sum_{i=0}^{t-1} z_i.$$

For $\sigma^2(f) = \lim_{t\to\infty} \operatorname{Var} z_i = \lim_{t\to\infty} \frac{\operatorname{Var} \mathcal{F}_0^t}{t}$ we get

$$\begin{aligned} \sigma^2(f) &= (g \cdot g)_\pi - (Pg \cdot Pg)_\pi \\ &= [(g - Pg) \cdot (g + Pg)]_\pi \\ &= \left[f \cdot (f + 2Pf + 2P^2 f + \cdots)\right]_\pi. \end{aligned} \tag{2.6}$$

When the transition matrix P is doubly stochastic (as in our specific case) we have

$$\pi(x) = \frac{1}{N},$$

and

$$\begin{aligned} \sigma^2(f) &= \frac{1}{N}\left(f \cdot \left(f + \left(P + P^*\right) f + \cdots + \left(P^n + \left(P^*\right)^n\right) f + \cdots\right)\right) \\ &= \frac{1}{N}\left(f \cdot \begin{bmatrix} (I - \Pi) + ((P - \Pi) + (P^* - \Pi)) + \cdots \\ \quad + \left((P^n - \Pi) + \left((P^*)^n - \Pi\right)\right) + \cdots \end{bmatrix} f\right) \\ &= (B_k f \cdot f), \end{aligned} \tag{2.7}$$

where

$$B_k = \frac{1}{N}\left[(I - \Pi) + (P + P^* - 2\Pi) + (P^2 + (P^*)^2 - 2\Pi) + \cdots\right].$$

This expression gives the limiting covariance matrix for any function f (i.e., without the centralization condition $\bar{f} = 0$).

For the K–P–S Markov chain, due to the nilpotency property $P^k = \Pi$, the covariance matrix is a finite sum:

$$B_k = 2^{-k}\big[I + (P + P^*) + \cdots + (P^{k-1} + (P^*)^{k-1}) - (2k-1)\Pi\big]. \tag{2.8}$$

Kolmogorov [6] proposed the following interpretation of the CLT. Let

$$\tau(\xi, t) = \sum_{s=1}^{t} I_\xi(x_s), \quad \xi \in W_k = \mathcal{X} \tag{2.9}$$

be the system of occupation times for the chain X (in the sequel we will be using the terms occupation times and local times interchangeably). Then P-a.s.

$$\frac{1}{t}\tau(\xi, t) \to 2^{-k}, \quad t \to \infty,$$

$$\vec{\tau} = \Big\{\tau^*\big(\xi, t\big) = \frac{1}{\sqrt{t}}\big(\tau\big(\xi, t\big) - t \cdot 2^{-k}\big),\ \xi \in W_k\Big\} \overset{\text{law}}{\Longrightarrow} N\big(\mathbf{0}, B_k\big), \tag{2.10}$$

i.e., B_k is the covariance matrix of the limiting (normal) distribution for the system of local times with standard centralization and scaling. For any ergodic chain the matrix B_k is degenerated [6] because $B_k \cdot \mathbf{1} = \mathbf{0}$, and so for a "typical" Markov chain one can expect Rank $B_k = N - 1$, i.e., that there is only one nontrivial relation between the elements of the vector $\{\tau(\xi, t) : \xi \in \mathcal{X}\}$, namely, $\sum_{\xi \in \chi} \tau(\xi, t) = t$ (see the discussion in [6]). A fundamental feature of the K–P–S chain however is the high degeneracy of the covariance matrix B_k, $k \geq 2$ (see Section 4).

3 A limit theorem for the approximate entropy

In this section we apply the CLT to get an estimate for the approximate entropy introduced in Section 1. We have

$$\pi(\xi, t) = \frac{\tau(\xi, t)}{t} = 2^{-k} + \frac{\tau^*(\xi, t)}{\sqrt{t}}, \tag{3.1}$$

where $\{\tau^*(\xi, t)\} \overset{\text{law}}{\Longrightarrow} N(\mathbf{0}, B_k)$. Now we expand the "empirical entropy" (which was defined earlier) as:

$$\begin{aligned}
\Phi(k, t) &= -\sum_{\xi \in W_k} \pi(\xi, t) \log_2 \pi(\xi, t) \\
&= -\sum_{\xi \in W_k} \left(2^{-k} + \frac{\tau^*(\xi, t)}{\sqrt{t}}\right) \log_2 \left(2^{-k} + \frac{\tau^*(\xi, t)}{\sqrt{t}}\right) \\
&= k \sum_{\xi \in W_k} \left(2^{-k} + \frac{\tau^*(\xi, t)}{\sqrt{t}}\right) -
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{\xi \in W_k} \left(2^{-k} + \frac{\tau^*(\xi,t)}{\sqrt{t}}\right) \frac{1}{\ln 2} \ln\left(1 + 2^k \frac{\tau^*(\xi,t)}{\sqrt{t}}\right) \\
&= k - \frac{1}{\ln 2} \sum_{\xi \in W_k} \left(2^{-k} + \frac{\tau^*(\xi,t)}{\sqrt{t}}\right) \\
&\quad \times \left(2^k \frac{\tau^*(\xi,t)}{\sqrt{t}} + 2^{2k-1} \frac{(\tau^*(\xi,t))^2}{t} + O\left(\frac{1}{t^{3/2}}\right)\right) \\
&= k - \frac{1}{\ln 2} \left[2^k \sum_{\xi \in W_k} \frac{(\tau^*(\xi,t))^2}{t} + 2^{k-1} \sum_{\xi \in W_k} \frac{(\tau^*(\xi,t))^2}{t}\right] + O\left(\frac{1}{t^{3/2}}\right) \\
&= k - \frac{3 \cdot 2^{k-1}}{\ln 2} \times \frac{1}{t} \sum_{\xi \in W_k} \left(\tau^*(\xi,t)\right)^2 + O\left(\frac{1}{t^{3/2}}\right). \qquad (3.2)
\end{aligned}
$$

Note that we used the equality $\sum_{\xi \in W_k} \frac{\tau^*(\xi,t)}{\sqrt{t}} = 0$. It means that

$$
t\left(\Phi(k,t) - k\right) \overset{\text{law}}{\Longrightarrow} c_k \sum_{\xi \in W_k} \left(\tau^*(\xi,t)\right)^2, \qquad (3.3)
$$

where $c_k = -\frac{3}{\ln 2} 2^{k-1}$. Asymptotically we have

$$
\vec{\tau} = \left(\tau^*(\xi,t)\right) = \sqrt{B_k}\vec{\theta},
$$

where $\vec{\theta} = \{\theta(\xi),\ \xi \in W_k\}$ is a vector of i.i.d. $N(0,1)$ random variables. Now

$$
\begin{aligned}
c_k\left(\vec{\tau} \cdot \vec{\tau}\right) &= c_k\left(\sqrt{B_k}\vec{\theta} \cdot \sqrt{B_k}\vec{\theta}\right) = c_k\left(B_k\vec{\theta} \cdot \vec{\theta}\right) \\
&\overset{\text{law}}{=} c_k\left(B_k O\,\vec{\theta} \cdot O\,\vec{\theta}\right) = c_k\left(O^* B_k\, O\,\vec{\theta} \cdot \vec{\theta}\right)
\end{aligned}
$$

(for an arbitrary non-random orthogonal matrix O). For an appropriate orthogonal matrix O, $\Lambda_k = O^* B_k\, O$ is a diagonal matrix of the form

$$
\Lambda_k = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{bmatrix}, \quad \lambda_1 \le \lambda_2 \le \cdots \le \lambda_N,
$$

that is, Λ_k is the diagonalization of B_k. Finally, we obtain

Theorem 3.1.

$$
t\left(\Phi(k,t) - k\right) \overset{\text{law}}{\Longrightarrow} -\frac{3 \cdot 2^{k-1}}{\ln 2} \sum_{i=1}^{N} \lambda_i\, \theta_i^2,
$$

where θ_i are i.i.d. $N(0,1)$ random variables, and λ_i $(i = 1, 2, \ldots, N)$ are the eigenvalues of the covariance matrix B_k. That is, the centralized and normalized empirical entropy converges to the so-called generalized χ^2-distribution.

Unfortunately, if λ_i are not identically equal, the quantiles of this distribution corresponding to a given probability α can not be evaluated precisely. It is necessary to use some statistical simulation method or numerical evaluation of the tails of the distribution for the convolution of the densities $p_i(\eta_i)$ of the random variables $\eta_i = \lambda_i \theta_i^2$. Of course,

$$p_i(\eta_i) = \frac{1}{2\sqrt{2\pi}} \cdot \frac{1}{\sqrt{\eta_i \lambda_i}} \cdot \exp\left\{ -\frac{\eta_i}{2\lambda_i} \right\}, \quad \eta_i > 0.$$

An immediate consequence of Theorem 3.1 is the following Limit Theorem for ApEn.

Corollary 3.2. *For a random binary sequence of size t, a word size k and a constant $c_0 = -\frac{3}{\ln 2}$, the statistic $\frac{t(\mathrm{ApEn}(k,t)-1)}{c_0 2^{k-2}}$ converges to a χ^2-distribution with 2^{k-1} degrees of freedom as t goes to infinity.*

Note that the above calculations can be done in a similar way for the case of an m-ary sequence. In fact, an attempt to prove this result in the slightly more general m-ary case was given in [13], however a wrong formula for the covariance matrix B_k was used in the proof. It is worth noting that this result was already known – in principle – to Marsaglia before the notion of ApEn was introduced (see [9, p. 6]).

For the practical testing of PRNG's we recommend the following different statistics: Let $\psi_i(x)\,;\ i = 1, 2, \ldots, N$ be the normalized eigenvectors of B_k with the corresponding eigenvalues λ_i such that

$$\sigma^2(\psi_i) = (B_k \psi_i \cdot \psi_i) = \lambda_i$$

and

$$\mathrm{Cov}(\psi_i, \psi_j) = \left(B_k \psi_i \cdot \psi_j\right) = 0, \quad i \neq j.$$

For $\lambda_i > 0$ put

$$S_i^*(t) = \frac{1}{\sqrt{t\lambda_i}} \sum_{s=0}^{t} \psi_i(X_s)\,. \tag{3.4}$$

Then for $i \geq i_0$ (for some i_0 such that $\lambda_i > 0$ whenever $i \geq i_0$) we have

Theorem 3.3.

$$\sum_{j=i_0}^{i} \left(S_j^*(t)\right)^2 \overset{\text{law}}{\Longrightarrow} \sum_{j=i_0}^{i} \theta_j^2 \simeq \chi^2(i - i_0 + 1)\,,$$

and $\{\theta_i : i = 1, 2, \ldots, N\}$ are again i.i.d. $N(0,1)$ random variables.

To develop a test based on this limit theorem we have to know:

- the spectrum of B_k, i.e., the eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$.
- the corresponding eigenvectors ψ_i, $i = 1, 2, \ldots, N$.

- an evaluation of the remainder term in the CLT for the effective construction of the confidence interval (for a given r and a given confidence level $\alpha \simeq 1$).

Note that the first two requirements can be loosened to knowing only a subset of the spectrum along with the corresponding eigenvectors. As for the third one we will content ourselves with stating the following result without proof.

Theorem 3.4. *Let $(X_t)_{t\geq 0}$ be a Markov chain defined on a* [*countable*] *state space $\mathcal{X}$, with an initial probability distribution $\mu(x),\ x \in \mathcal{X}$ and a transition matrix P that satisfies the following Doeblin condition:*

$$P^k(x, y) \geq \rho\mu(y) > 0 \tag{3.5}$$

for some integer k and real number $\rho \in (0, 1)$ and any $y \in \mathcal{X}$; then for all numbers z and for any function f defined on $\mathcal{X}$ with covariance $\sigma(f)$ (see below) we have

$$P_\pi\left(\left|\frac{1}{\sqrt{t}}\sum_{s=0}^{t-1} f(x_s)\right| > z\sigma(f)\right) \leq c_1\cdot\frac{k^3}{\sqrt{t}}\cdot\gamma^t + 2\left(1+c_2\frac{k^2}{t}\right)\cdot\exp\left(-\frac{z^2}{2}+c_3\frac{k^2}{\sqrt{t}}\right),$$

where γ, c_1, c_2, and c_3 depend on f, z, k and ρ, and $0 < \gamma < 1$.

Theorem 3.4 will be proven in a paper which is yet to appear. For the special case of the K–P–S Markov chain with word size k – in which case $\rho = 1$ – the constants in Theorem 3.4 can be considerably reduced (see [1]).

4 Analysis of the quadratic form $2^k(B_k\boldsymbol{x}\cdot\boldsymbol{x})$

The matrix B_k given by Formula (2.8) is not simple, mainly because it contains not only powers of P but also powers of its conjugate matrix P^* (which does not commute with P). The quadratic form $f_k(\boldsymbol{x}) = 2^k(B_k\boldsymbol{x}\cdot\boldsymbol{x})$ under the condition $(\boldsymbol{x}\cdot\mathbf{1}) = 0 \Rightarrow \sum_{i=0}^{N-1} x_i = 0$ looks slightly simpler (note that we multiply by 2^k only to reduce fractions).

$$\begin{aligned} 2^k(B_k\boldsymbol{x}\cdot\boldsymbol{x}) = &\sum_{i=0}^{N-1} x_i^2 + \sum_{i=0}^{N'-1}(x_{2i}+x_{2i+1})\left(x_i + x_{i+N'}\right) \\ &+ \frac{1}{2}\sum_{i=0}^{N''-1}(x_{4i}+x_{4i+1}+x_{4i+2}+x_{4i+3})\left(x_i + x_{i+N''} + x_{i+2N''} + x_{i+3N^{(3)}}\right) \\ &+ \cdots \end{aligned} \tag{4.1}$$

$$+\frac{1}{2^{l-1}}\sum_{i=0}^{N^{(l)}-1}\left(x_{2^l i}+x_{2^l i+1}+\cdots+x_{2^l i+2^l-1}\right)$$
$$\cdot\left(x_i+x_{i+N^{(l)}}+x_{i+2N^{(l)}}+\cdots+x_{i+(2^l-1)N^{(l)}}\right)$$
$$+\cdots$$
$$+\frac{1}{2^{k-2}}\left[\begin{array}{l}\left(x_0+x_1+\cdots+x_{N'-1}\right)\left(x_0+x_2+\cdots+x_{N-2}\right)\\ +\left(x_{N'}+x_{N'+1}+\cdots+x_{N-1}\right)\left(x_1+x_3+\cdots+x_{N-1}\right)\end{array}\right].$$

It is convenient at this time to remember that the spectral analysis of the symmetric matrix B_k is equivalent to the study of the quadratic form $(B_k\boldsymbol{x}\cdot\boldsymbol{x})$ on the sphere $(\boldsymbol{x}\cdot\boldsymbol{x})=1$. In fact, this is true even in a more general setting. Assuming that A is a symmetric and positive definite matrix on $\mathbb{R}^N$, one can introduce the associated dot product:

$$(x\cdot y)_A=(Ax\cdot y),$$

then the roots of the equation

$$\det\left(B_k-\lambda A\right)=0$$

coincide with the extreme points of the form $f=(B_k\boldsymbol{x}\cdot\boldsymbol{x})$ under the condition $(\boldsymbol{x}\cdot\boldsymbol{x})_A=1$.

In fact, if we let

$$F_\lambda(\boldsymbol{x})=(B_k\boldsymbol{x}\cdot\boldsymbol{x})-\lambda(A\boldsymbol{x}\cdot\boldsymbol{x})$$

be the Lagrangian, then since $x\neq 0$, $\nabla F_\lambda(x)=0\Rightarrow 2\left(B\boldsymbol{x}-\lambda A\boldsymbol{x}\right)=0$, i.e.,

$$\det\left(B_k-\lambda A\right)=0.$$

We will use the following simple particular case and we will refer to it as

Proposition 4.1. *If* $(A\boldsymbol{x}\cdot\boldsymbol{x})=\sum_{i=0}^{N-1}a_i x_i^2$, $a_i>0$ *and* $(B\boldsymbol{x}\cdot\boldsymbol{x})=\sum_{i=0}^{N-1}b_i x_i^2$, *then the extreme values of* $(B\boldsymbol{x}\cdot\boldsymbol{x})$ *under the condition* $(A\boldsymbol{x}\cdot\boldsymbol{x})=1$ *are equal to*

$$\lambda_i=\frac{b_i}{a_i},\quad i=0,1,\ldots,N-1.$$

The corresponding eigenvectors are also simple:

$$\psi_i=\frac{1}{\sqrt{a_i}}\hat{\boldsymbol{e}}_{i+1},\quad i=0,1,\ldots,N-1,$$

where $\{\hat{\boldsymbol{e}}_i : i=1,\ldots,N\}$ *is the canonical basis of* $\mathbb{R}^N$.

An important key in the analysis of the matrix B_k is the possibility to find two simple pairs of invariant subspaces in $\mathbb{R}^N$. The first pair – which will not be used directly but is important ideologically – is based on the perfect symmetry between the 0's and the 1's. We say that a vector is even if $x_i=x_{N-1-i}$ and odd if $x_i=-x_{N-1-i}$, $i=0,1,\ldots,N-1$.

Proposition 4.2. *The subspaces of the even and odd vectors, L^N_{odd} and L^N_{even} are invariant under the operator B_k. Moreover, we have the spectral decomposition*

$$\mathbb{R}^N = L^N_{\text{odd}} \oplus L^N_{\text{even}}.$$

Proof. It is simple. Let us consider the $N \times N$ matrix

$$\tilde{I}_N = [\delta_{i,N-1-i}], \quad i = 0, 1, \ldots, N-1.$$

Simple calculations give $\tilde{I}_N P = P\tilde{I}_N$, then $\tilde{I}_N P^* = P^*\tilde{I}_N$, i.e., $\tilde{I}_N$ commutes with any power of P or P^*. As a result, it commutes with B_k, and invariant subspaces of $\tilde{I}_N$ are B_k-invariant (and vice versa). But L^N_{odd} and L^N_{even} are invariant subspaces of the idempotent matrix $\tilde{I}_N$ both of dimension N' corresponding to the eigenvalues $\lambda_0 = 1$ and $\lambda_1 = -1$ respectively. □

This symmetry is useful in practical calculations. It reduces the dimension of the state space (we can study the even and odd spectral components independently).

More important is the zero-subspace of B_k

$$L^N_0 = \{\boldsymbol{x} : B_k \boldsymbol{x} = 0\},$$

i.e., the eigenspace with eigenvalue $\lambda_0 = 0$. It is nontrivial for any Markov chain because $\mathbf{1} \in L^N_0$. For a generic Markov chain $L^N_0 = \left\{\boldsymbol{x} = c\mathbf{1} : c \in \mathbb{R}^1\right\}$.

Lemma 4.3. *For the K–P–S chain*

$$\dim\, L_0 = 2^{k-1} = N',$$

and it is generated by the system of vectors $\vec{\boldsymbol{e}}_{N'} = \mathbf{1}$ and $\vec{\boldsymbol{e}}_i$ (with $i = 0, \ldots, N' - 1$) defined as

$$\vec{\boldsymbol{e}}_{ij} = \begin{cases} 1, & j = 2i, 2i+1, \\ -1, & j = i, i+N', \\ 0, & \textit{otherwise}. \end{cases} \tag{4.2}$$

Remark 4.4. Note that the vectors $\vec{\boldsymbol{e}}_i$, $i = 0, 1, \ldots, N' - 1$ are linearly dependent because they have one (and exactly one) linear relation; namely $\sum_{i=0}^{N'-1} \vec{\boldsymbol{e}}_i = 0$, but together with the vector $\mathbf{1}$ – which is orthogonal to each of the $\vec{\boldsymbol{e}}_i$ – they form a basis in L^N_0, therefore $\dim L^N_0 = N'$.

Proof. For brevity we will denote L^N_0 simply by L_0. Let us remember (see equation (2.6)) that $\sigma^2(f) = (g \cdot g)_\pi - (Pg \cdot Pg)_\pi$, $g - Pg = f$, and $(f \cdot \mathbf{1}) = (g \cdot \mathbf{1}) = 0$. In our case, $\pi = \frac{1}{N}$, i.e., $\sigma^2(f) = 0 \Leftrightarrow (g \cdot g) = (g \cdot P^*Pg)$, where P^* is the usual conjugation: $\left(P_{ij}\right)^* = \left(P_{ji}\right)$. The matrix P^*P is stochastic and symmetric, therefore its maximum eigenvalue is equal to 1. This immediately implies that

$$\{f : \sigma^2(f) = 0\} = \text{span}\{f = g - Pg \text{ such that } P^*Pg = g\}.$$

Direct calculation gives

$$P^*P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & & & & \\ \frac{1}{2} & \frac{1}{2} & 0 & \dots & 0 & \\ & & \frac{1}{2} & \frac{1}{2} & & \\ 0 & & \frac{1}{2} & \frac{1}{2} & \dots & 0 \\ & & & & & \\ \vdots & & \vdots & & \ddots & \vdots \\ & & & & \frac{1}{2} & \frac{1}{2} \\ 0 & & 0 & \dots & \frac{1}{2} & \frac{1}{2} \end{bmatrix}. \tag{4.3}$$

The eigenspace of P^*P with eigenvalue 1 has the form

$$g = \begin{bmatrix} a_0 \\ a_0 \\ a_1 \\ a_1 \\ \vdots \\ a_{N'-1} \\ a_{N'-1} \end{bmatrix}$$

$$\Rightarrow f = g - Pg = \begin{bmatrix} a_0 \\ a_0 \\ a_1 \\ a_1 \\ \vdots \\ a_{N'-1} \\ a_{N'-1} \end{bmatrix} - \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{N'-1} \\ a_0 \\ \vdots \\ a_{N'-1} \end{bmatrix} = \begin{bmatrix} 0 \\ a_0 - a_1 \\ a_1 - a_2 \\ \vdots \\ \vdots \\ a_{N'-1} - a_{N'-2} \\ 0 \end{bmatrix}$$

$$= \sum_{i=0}^{N'-1} a_i \, \vec{e}_i .$$

□

Using the formulas for the above basis of L_0 we can present the orthogonal complement $L_0^{\perp}$ by the system of equations (using the new variables x_i)

$$\begin{cases} x_{2i} + x_{2i+1} = x_i + x_{i+N'}, & i = 0, 1, \dots, N' - 1, \\ \sum_{i=0}^{N'-1} x_i = 0. \end{cases} \tag{4.4}$$

The first N' equations are linearly dependent (the total sum of both parts gives $\sum_{i=0}^{N'-1} x_i$), but $N' - 1$ among them are independent. Together with the last equation on the second line of (4.4) they provide

$$\dim L_0^{\perp} = N' = \dim L_0. \tag{4.5}$$

Example 4.5. The exact description of the two pairs of invariant subspaces gives a possibility to reduce the volume of calculations, especially for small k. In this example we will present the spectral analysis of B_2. It will be the basis of the further inductive procedure. Here we will use direct calculations with matrices. Our main method though will be based on the variational interpretation of the spectra and quadratic forms. Formula (2.8) gives

$$B_2 = \frac{1}{4}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} + \frac{1}{4}\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$
$$+ \frac{1}{4}\begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} - \frac{3}{4}\begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$
$$= \frac{1}{16}\begin{bmatrix} 5 & -1 & -1 & -3 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ -3 & -1 & -1 & 5 \end{bmatrix}.$$

Due to Lemma 4.3 the eigenvectors $\vec{e}_1 = [0, -1, 1, 0]^*$, $\vec{e}_2 = [1, 1, 1, 1]^*$ generate L_0 and correspond to the eigenvalues $\lambda_1 = \lambda_2 = 0$ (where $*$ stands for vector transposition). By Proposition 4.2, there exists an even vector $\vec{e}_3 = [a, b, b, a]^*$ such that $\vec{e}_3 \cdot \vec{e}_1 = \vec{e}_3 \cdot \vec{e}_2 = 0$ and $a + b = 0$. Choosing $a = 1$ we get $\vec{e}_3 = [1, -1, -1, 1]^*$. Now $B_2\vec{e}_3 = \frac{1}{4}\vec{e}_3$, that is $\lambda_3 = 1/4$. Since the remaining eigenvector is odd, it is simple to see that it is given by $\vec{e}_4 = [1, 0, 0, -1]^*$ and corresponds to $\lambda = 1/2$. Table 1 contains a list of the eigenvalues along with the corresponding (orthonormal) eigenvectors for B_2.

Table 1. The full spectrum for $k = 2$

$\lambda_1 = \lambda_2 = 0$		$\lambda_3 = \frac{1}{4}$	$\lambda_4 = \frac{1}{2}$
$\vec{v}_1 = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$,	$v_2 = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}$	$\vec{v}_3 = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$	$\vec{v}_4 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ 0 \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$

Note that this theory is also applicable to the case $k = 1$, where

$$B_1 = I - P = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

with the only nontrivial eigenvector $[1, -1]^*$ corresponding to the eigenvalue $\lambda = 1$. If we let τ_0 and τ_1 be the frequencies of 0's and 1's in a random binary sequence of size t, then the theory reduces to

$$\frac{\tau_0 - \tau_1}{\sqrt{2t}} \to N(0, 1). \tag{4.6}$$

We return now to the analysis of the general case. Using (4.4) we can rewrite the form f_k as

$$\begin{aligned} f_k(\boldsymbol{x}) = 2^k (B_k \boldsymbol{x} \cdot \boldsymbol{x}) = &\sum_{i=0}^{N-1} x_i^2 + \sum_{i=0}^{N'-1} (x_{2i} + x_{2i+1})^2 \\ &+ \frac{1}{2} \sum_{i=0}^{N''-1} (x_{4i} + x_{4i+1} + x_{4i+2} + x_{4i+3})^2 \\ &\vdots \\ &+ \frac{1}{2^{l-1}} \sum_{i=0}^{N^{(l)}-1} \left(x_{2^l i} + x_{2^l i+1} + \cdots + x_{2^l i+2^l-1}\right)^2 \\ &\vdots \\ &+ \frac{1}{2^{k-2}} \left[\begin{array}{l} \left(x_0 + x_1 + \cdots + x_{N'-1}\right)^2 \\ \quad + \left(x_{N'} + x_{N'+1} + \cdots + x_{N-1}\right)^2 \end{array} \right]. \end{aligned} \tag{4.7}$$

In fact, by repeatedly applying equation (4.4) we prove that in $L_0^{\perp}$ the quadratic form (4.1) can be represented as (4.7). For instance,

$$\begin{aligned} x_i + x_{i+N''} + x_{i+2N''} + x_{i+3N''} &= \left(x_i + x_{i+N'}\right) + \left(x_{i+N''} + x_{i+N''+N'}\right) \\ &= (x_{2i} + x_{2i+1}) + \left(x_{2i+N'} + x_{2i+1+N'}\right) \\ &= \left(x_{2i} + x_{2i+N'}\right) + \left(x_{2i+1} + x_{2i+1+N'}\right) \\ &= (x_{4i} + x_{4i+1} + x_{4i+2} + x_{4i+3}); \qquad i = 0, 1, \ldots, N'' - 1. \end{aligned}$$

We simply re-arranged terms and used equation (4.4) twice.

The representation (4.7) is the starting point of all further analysis. It has a visible hierarchical structure. We will recall a few simple facts from the theory of hierarchical matrices and forms. The central ideas here are due to F. Dyson [3]. Hierar-

chical models with random potentials (hierarchical Anderson model) were studied by Molchanov [10].

5 Hierarchical matrices and their diagonalization

Hierarchical matrices (operators) depend on a system of scalars of ranks which form the following generalized geometric progression $\rho_0 = 1$, $\rho_1 = n_1$, $\rho_2 = n_1 n_2, \ldots, \rho_l = \prod_{i=1}^{l} n_i, \ldots$, where $n_l \in Z^+$, $n_l \geq 2$ and $l = 1, 2, \ldots$. We will discuss only the simplest case here, namely, when we have a geometric progression, in which case $n_l = 2, l = 1, 2, \ldots, k$; and will use the notation of Section 1. For a more thorough discussion of general hierarchical models see [3, 10]. The hierarchical model depends on the scale parameter $\nu \geq 2$ (in our case $\nu = 2$) and a system of weights (here ρ_0, $\rho_1, \rho_2, \ldots$). Let us consider the one-dimensional lattice $\mathbb{Z}_+^1$ and the following family of embedded partitions $T_0 \supset T_1 \supset T_2 \supset \cdots$: T_0 is the point partition. Its elements, the points $x \in \mathbb{Z}_+^1$, will be called the "cubes" of rank $r = 0$. The second partition T_1 consists of the non-overlapping cubes $Q_i^{(1)}$, $i = 0, 1, 2 \ldots$ of rank 1, $|Q_i^{(1)}| = 2$, and each $Q_i^{(1)}$ contains two consecutive cubes of rank 0. Partition T_2 consists of the non-overlapping unions of every two consecutive cubes of rank 1: $Q_i^{(2)} = Q_{2i}^{(1)} \cup Q_{2i+1}^{(1)}$, $|Q_i^{(2)}| = 2^2 = 4$, and so on.

Any point $x \in \mathbb{Z}_+^1$ belongs to exactly one cube in T_r which we denote by $Q^{(r)}(x)$, for any $r \geq 0$. The system of partitions $T_r, r \geq 0$ gives hierarchical (self-similar) structure on $\mathbb{Z}_+^1$. The set of one-to-one mappings on $\mathbb{Z}_+^1$, that preserves this structure forms the hierarchical (renormalization) group G_h. It is generated by the local permutations, which include the following transformations: permutations of the elements inside a cube $Q_i^{(1)}$ (with the identity mapping outside this cube), permutations of rank 1 cubes inside a fixed cube $Q_i^{(2)}$ of rank 2 etc.

Definition 5.1. The (Dyson) hierarchical distance on $\mathbb{Z}_+^1$ is given by

$$d_h(x, y) = \min\{r : \exists i,\, Q_i^{(r)} \ni x, y\}.$$

It is often convenient to consider the following related distance:

$$\tilde{d}_h(x, y) = 2^{d_h(x,y)}.$$

This second distance gives a better approximation to the Euclidian metric on $\mathbb{Z}_+^1$. Now, the hierarchical Laplacian is given by

$$\Delta_h \psi(x) = \sum_{r=1}^{\infty} \rho_r \cdot \frac{\sum_{x' \in Q^{(r)}(x)} \psi(x')}{|Q^{(r)}|}, \quad \sum_{r=1}^{\infty} \rho_r = 1.$$

Both objects, d_h and Δ_h are invariant with respect to the hierarchical group. The truncated Laplacian has the form:

$$\Delta_h^{(k)}\psi(x) = \sum_{r=1}^{k} \rho_r \cdot \frac{\sum_{x' \in Q^{(r)}(x)} \psi(x')}{|Q^{(r)}|}.$$

It is clear that the operator Δ_h is stochastic and symmetric ($\Delta_h = \Delta_h^*$). The corresponding Markov chain (the "hierarchical" random walk on $\mathbb{Z}_+^1$) has the following simple structure: if $x_t = x$, then we have to select the rank r of the next jump with probability ρ_r, $r \geq 1$, and then x_{t+1} will be uniformly distributed inside the cube $Q^{(r)}(x)$.

Definition 5.2. The quadratic hierarchical form corresponding to the truncated Laplacian $\Delta_h^{(k)}$ is given by the expression

$$\begin{aligned}\left(\Delta_h^{(k)}\boldsymbol{x}\cdot\boldsymbol{x}\right) = \rho_0 \sum_{i=0}^{N-1} x_i^2 &+ \frac{\rho_1}{2} \sum_{i=0}^{N'-1} (x_{2i} + x_{2i+1})^2 \\ &+ \frac{\rho_2}{4} \sum_{i=0}^{N''-1} (x_{4i} + x_{4i+1} + x_{4i+2} + x_{4i+3})^2 \\ &\vdots \\ &+ \frac{\rho_{k-1}}{2^{k-1}} \sum_{i=0}^{1} (x_{N'i} + x_{N'i+1} + \cdots + x_{N'i+N'-1})^2 \\ &+ \frac{\rho_k}{2^k}\left(\sum_{i=0}^{N-1} x_i\right)^2,\end{aligned} \tag{5.1}$$

where $N = 2^k$. Note that we added the diagonal term with the coefficient ρ_0.

The spectral analysis of $\Delta_h^{(k)}$ is simple. The smallest eigenvalue, i.e., $\lambda_0 = \min_{(\boldsymbol{x}\cdot\boldsymbol{x})=1}(\Delta_h^{(k)}\boldsymbol{x}\cdot\boldsymbol{x})$ is equal to $\lambda_0 = \rho_0$, and the corresponding invariant subspace L_0 is given by the equations

$$x_{2i} + x_{2i+1} = 0, \quad i = 0, 1, \ldots, N' - 1.$$

In this case all terms in (5.1) vanish except for $\rho_0 \sum_{i=0}^{N-1} x_i^2 = \rho_0$. Evidently, $\dim L_0 = N'$. The natural basis of L_0 consists of the vectors $\{\check{\boldsymbol{e}}_i : i = 0, 1, \ldots, N' - 1\}$, where (with $i = 0, 1, \ldots, N' - 1$)

$$\check{\boldsymbol{e}}_{ij} = \begin{cases} 1, & j = 2i, \\ -1, & j = 2i + 1, \\ 0, & \text{otherwise.} \end{cases}$$

The orthogonal complement, $L_0^{\perp}$, of L_0 is given by the dual equations

$$x_{2i} - x_{2i+1} = 0, \quad i = 0, 1, \ldots, N^{'} - 1.$$

Now let Q_l (with $l = 0, \ldots, k$) be the $N \times N$ block diagonal matrix where each diagonal block is a $2^l \times 2^l$ matrix consisting of 1's. Note that $Q_0 = I$ and $(Q_k)_{ij} \equiv 1$. Then $(\Delta_h^{(k)} \boldsymbol{x} \cdot \boldsymbol{x}) = \sum_{l=0}^{k} \frac{\rho_l}{2^l} (Q_l \boldsymbol{x} \cdot \boldsymbol{x})$. Define

$$L_1 = \{\boldsymbol{x} \in L_0^{\perp} : \ x_{4i} + x_{4i+1} + x_{4i+2} + x_{4i+3} = 0, \ i = 0, 1, \ldots, N^{''} - 1\}.$$

It is easy to see that $\frac{1}{2} Q_1$ acts on $L_0^{\perp}$ as the identity operator I. That is, for $\boldsymbol{x} \in L_1$

$$\Big(\rho_0 I + \frac{\rho_1}{2} Q_1\Big)\boldsymbol{x} = (\rho_0 + \rho_1)\, \boldsymbol{x}.$$

But $Q_l \boldsymbol{x} = 0$ for all $l \geq 2$. Therefore, $\lambda_1 = \rho_0 + \rho_1$ with multiplicity $\dim L_1 = N^{''}$.

The orthogonal complement to $(L_0 \oplus L_1)$ consists of the vectors with the conditions

$$x_{4i} = x_{4i+1} = x_{4i+2} = x_{4i+3}, \quad i = 0, 1, \ldots, N^{''} - 1,$$

and one can continue the same kind of analysis. The complete spectrum is displayed in Table 2.

Table 2. The spectrum of a hierarchical quadratic form

i	0	1	...	$k-1$	k
λ_i	ρ_0	$\rho_0 + \rho_1$	...	$\rho_0 + \cdots + \rho_{k-1}$	$\rho_0 + \cdots + \rho_k$
m_i	N'	N''	...	1	1

Here, of course, $m_i = \dim L_i$ is the multiplicity of the eigenvalue λ_i. The same formula works in the limit as $k \to \infty$, i.e., for the full Laplacian Δ_h, which then has the eigenvalues $\lambda_i = \sum_{j=0}^{i} \rho_j$; $i = 1, 2, \ldots$ each with infinite multiplicity. We can also construct a hierarchical random walk directly on the group G_h. Compare with the construction of R. Grigorchuk (see the corresponding publication in this volume). The fundamental difference is related to the structures of the groups: in the latter example the group has finitely many generators while the group G_h has infinitely many generators (and is locally finite).

It is fairly simple to give a (hierarchical) description of the orthogonal matrix $\tilde{O}_k$ which provides the diagonal form $\Delta_h^{(k)} = \tilde{O}_k^* \Lambda_k \tilde{O}_k$, where Λ_k is the matrix

$$\begin{bmatrix}
\rho_0 & & & & & & & & \\
& \ddots & & \Big\} N' & & 0 & \dots & 0 & \\
& & \rho_0 & & & & & & \\
& & & & \rho_0+\rho_1 & & & & \\
& & 0 & & & \ddots & \dots & 0 & \\
& & & & & & \Big\} N'' & & \\
& & & & & \rho_0+\rho_1 & & & \\
& & \vdots & & & & \ddots & & \\
& & 0 & & & & & \rho_0+\rho_1+\cdots+\rho_k &
\end{bmatrix}$$

As a first step we will make the orthogonal transformation (for $i = 0, 1, \dots, N' - 1$)

$$\begin{cases} \dfrac{x_{2i} + x_{2i+1}}{\sqrt{2}} = a_i, \\ \dfrac{x_{2i} - x_{2i+1}}{\sqrt{2}} = a_{i+N'}. \end{cases} \tag{5.2}$$

In the new coordinates, the form (5.1) can now be presented as

$$(\Delta_h^{(k)} \boldsymbol{a} \cdot \boldsymbol{a}) = \rho_0 \sum_{i=0}^{N-1} a_i^2 + \rho_1 \sum_{i=0}^{N'-1} a_i^2 + \frac{\rho_2}{2} \sum_{i=0}^{N''-1} (a_{2i} + a_{2i+1})^2 + \cdots . \tag{5.3}$$

On the next step we define (for $i = 0, 1, \dots, N'' - 1$ and $j = N', \dots, N - 1$)

$$\begin{cases} \dfrac{a_{2i} + a_{2i+1}}{\sqrt{2}} = b_i, \\ \dfrac{a_{2i} - a_{2i+1}}{\sqrt{2}} = b_{i+N''}, \\ a_j = b_j. \end{cases}$$

Now (5.3) becomes

$$(\Delta_h^{(k)} \boldsymbol{b} \cdot \boldsymbol{b}) = \rho_0 \sum_{i=0}^{N-1} b_i^2 + \rho_1 \sum_{i=0}^{N'-1} b_i^2 + \rho_2 \sum_{i=0}^{N''-1} b_i^2 + \frac{\rho_3}{2} \sum_{i=0}^{N^{(3)}-1} (b_{2i} + b_{2i+1})^2 + \cdots .$$

After k such substitutions we obtain

$$\begin{aligned}(\Delta_h^{(k)}\boldsymbol{z}\cdot\boldsymbol{z})&\\ &= \rho_0\sum_{i=0}^{N-1} z_i^2 + \rho_1\sum_{i=0}^{N'-1} z_i^2 + \rho_2\sum_{i=0}^{N''-1} z_i^2 + \cdots + \rho_l\sum_{i=0}^{N^{(l)}-1} z_i^2 + \cdots + \rho_k z_0^2\\ &= (\rho_0+\cdots+\rho_k)\,z_0^2 + (\rho_0+\cdots+\rho_{k-1})\,z_1^2 + (\rho_0+\cdots+\rho_{k-2})\,(z_2^2+z_3^2)\\ &\quad + \cdots + \rho_0(z_{N'}^2+\cdots+z_{N-1}^2).\end{aligned}$$

The condition $(\boldsymbol{x}\cdot\boldsymbol{x}) = \sum_{i=0}^{N-1} x_i^2 = 1$, due to orthogonality, now has the form $(\boldsymbol{z}\cdot\boldsymbol{z}) = \sum_{i=0}^{N-1} z_i^2 = 1$. Using Proposition 4.1 it can be immediately seen that the extremes of $(\Delta_h^{(k)}\boldsymbol{x}\cdot\boldsymbol{x})$, i.e., the eigenvalues of $\Delta_h^{(k)}$ are exactly those given in Table 2. The same proposition can also be used to provide the corresponding eigenvectors.

6 Main result

Our goal now is to find the complete spectrum for $2^k B_k$. In fact, while the diagonalization of hierarchical matrices is rather straightforward, diagonalizing the matrix $2^k B_k$ is not as simple. The reason is that, although it assumes a hierarchical structure, it only does so on a proper subspace of $\mathbb{R}^N$. The following lemma provides a strong tool to overcome this difficulty.

Lemma 6.1. *For any $k \geq 2$ there exists an orthogonal transformation O (namely, the product of successive hierarchical and orthogonal transformations similar to those from the previous section) that maps the initial variables $\{x_i\}$ into new variables $\{z_i\}$ (where $i = 0, 1, \ldots, N-1$) such that*

i) *The quadratic form $f_k(\boldsymbol{x}) := 2^k\,(B_k\boldsymbol{x}\cdot\boldsymbol{x})$, which has the hierarchical representation* (4.7) *on $L_0^{\perp}$, can be mapped to*

$$f_k(\boldsymbol{z}) = \sum_{i=0}^{N-1} z_i^2 + 2\sum_{i=0}^{N'-1} z_i^2 + 2\sum_{i=0}^{N''-1} z_i^2 + \cdots + 2(z_0^2+z_1^2); \tag{6.1}$$

ii) *The normalization condition*

$$(\boldsymbol{x}\cdot\boldsymbol{x}) = \sum_{i=0}^{N-1} x_i^2 = 1$$

becomes

$$(\boldsymbol{z}\cdot\boldsymbol{z}) = \sum_{i=0}^{N-1} z_i^2 = 1;$$

iii) *The equations* (4.4) *for* $L_0^{\perp}$ *– represented in the new variables – are*

$$z_{2^{l-2}+i} = z_{2^{l-1}+i} \quad i = 0, \ldots, 2^{l-2} - 1,\ l = k, k-1, \ldots, 2;$$

iv) *The total number of independent equations is* N'.

Proof. It will be done by induction and split into two parts:

a) The case $k = 2$. This is not only the base of induction but it also illustrates the main idea. Of course we already know the spectral structure of B_2, see Example 4.5.

The quadratic form (4.7) is simply

$$2^2 (B_2 \boldsymbol{x} \cdot \boldsymbol{x}) = \sum_{i=0}^{3} x_i^2 + \sum_{i=0}^{1} (x_{2i} + x_{2i+1})^2$$

under the condition $(\boldsymbol{x} \cdot \boldsymbol{x}) = \sum_{i=0}^{3} x_i^2 = 1$. The equations (4.4) for $L_0^{\perp}$ become

$$\begin{cases} \sum_{i=0}^{3} x_i = 0, \\ x_0 + x_1 = x_0 + x_2, \\ x_2 + x_3 = x_1 + x_3. \end{cases} \tag{6.2}$$

The first orthogonal transformation is

$$\left.\begin{array}{l} \dfrac{x_0 + x_1}{\sqrt{2}} = y_0 \\ \dfrac{x_0 - x_1}{\sqrt{2}} = y_2 \end{array}\right\} \Longrightarrow \left\{\begin{array}{l} x_0 = \dfrac{y_0 + y_2}{\sqrt{2}} \\ x_1 = \dfrac{y_0 - y_2}{\sqrt{2}} \end{array}\right. \tag{6.3}$$

$$\left.\begin{array}{l} \dfrac{x_2 + x_3}{\sqrt{2}} = y_1 \\ \dfrac{x_2 - x_3}{\sqrt{2}} = y_3 \end{array}\right\} \Longrightarrow \left\{\begin{array}{l} x_2 = \dfrac{y_1 + y_3}{\sqrt{2}} \\ x_3 = \dfrac{y_1 - y_3}{\sqrt{2}} \end{array}\right.$$

The quadratic form now becomes

$$f_k (\boldsymbol{y}) = \sum_{i=0}^{3} y_i^2 + \sum_{i=0}^{1} y_i^2$$

with the condition $\sum_{i=0}^{3} y_i^2 = 1$. Equations (6.2) read

$$\begin{cases} \sum_{i=0}^{3} y_i = 0, \\ y_0 + y_1 = y_0 + y_1 \text{ (trivial)}, \\ y_0 - y_1 = y_2 + y_3. \end{cases} \tag{6.4}$$

The second transformation is:

$$\begin{aligned}
\frac{y_0 + y_1}{\sqrt{2}} &= w_0, \\
\frac{y_0 - y_1}{\sqrt{2}} &= w_1, \\
\frac{y_2 + y_3}{\sqrt{2}} &= w_2, \\
\frac{y_2 - y_3}{\sqrt{2}} &= w_3.
\end{aligned}$$

After this step the form is the same as before:

$$f_k(\boldsymbol{w}) = \sum_{i=0}^{3} w_i^2 + \sum_{i=0}^{1} w_i^2, \quad \text{subject to } \sum_{i=0}^{3} w_i^2 = 1.$$

However, $L_0^{\perp}$ is characterized by the simple equations

$$w_0 = 0, \quad w_1 = w_2$$

(here w_1 and w_3 are independent variables). The final shape of the variational problem is

$$2^2 (B_2 \boldsymbol{w} \cdot \boldsymbol{w}) = 4w_1^2 + w_3^2$$

subject to

$$(\boldsymbol{w} \cdot \boldsymbol{w}) = 2w_1^2 + w_3^2 = 1.$$

Applying Proposition 4.1 we re-obtain the eigenvalues

$$\lambda_1 = \frac{1}{1} = 1, \quad \lambda_2 = \frac{4}{2} = 2$$

for the form $2^2 B_2 = 4B_2$, i.e., the eigenvalues $\frac{1}{4}$ and $\frac{1}{2}$ for the original form B_2, compare with Example 4.5.

Now we will apply the inductive multi-step approach in the general case.

b) The analysis here is similar to that of hierarchical matrices. However, the quadratic form in hand assumes the hierarchical shape only on a proper subspace of $\mathbb{R}^N$, namely on $L_0^{\perp}$. A remedy to this problem is to apply a sequence of hierarchical transformations, in spirit of the case $k = 2$. Those transformations not only diagonalize the quadratic form f_k but also simplify the equations (4.4) to those given in Lemma 6.1(iii). We will start here with the form (4.7) and equations (4.4), apply a sequence of orthogonal transformations and update the quadratic form and the $L_0^{\perp}$ equations after each transformation.

The first orthogonal transformation is a generalization of Equations (6.3), namely, for $i = 0, 1, \ldots, N^{'} - 1$

$$\left.\begin{array}{l} \dfrac{x_{2i} + x_{2i+1}}{\sqrt{2}} = y_i \\ \dfrac{x_{2i} - x_{2i+1}}{\sqrt{2}} = y_{i+N'} \end{array}\right\} \Longrightarrow \left\{\begin{array}{l} x_{2i} = \dfrac{y_i + y_{i+N'}}{\sqrt{2}} \\ x_{2i+1} = \dfrac{y_i - y_{i+N'}}{\sqrt{2}}, \end{array}\right.$$

so that now we are ready to write the quadratic form in the $\boldsymbol{y}$-variables:

$$\begin{aligned} 2^k \left(B_k \boldsymbol{y} \cdot \boldsymbol{y}\right) = \sum_{i=0}^{N-1} y_i^2 + 2 \sum_{i=0}^{N^{'}-1} y_i^2 + \sum_{i=0}^{N^{''}-1} (y_{2i} + y_{2i+1})^2 \\ + \frac{1}{2} \sum_{i=0}^{N^{(3)}-1} (y_{4i} + y_{4i+1} + y_{4i+2} + y_{4i+3})^2 + \cdots \\ + \frac{1}{2^{k-3}} \left[\left(y_0 + \cdots + y_{N^{''}-1}\right)^2 + \left(y_{N^{''}} + \cdots + y_{N^{'}-1}\right)^2 \right] \end{aligned}$$

(subject to the condition $\sum_{i=0}^{N-1} y_i^2 = 1$). Let us calculate the $L_0^{\perp}$ equations in terms of the new variables. For even i, $i \in \{0, 1, \ldots, N^{'} - 1\}$

$$x_{2i} + x_{2i+1} = x_i + x_{i+N'} \implies \sqrt{2} y_i = \frac{y_{\frac{i}{2}} + y_{\frac{i}{2}+N'}}{\sqrt{2}} + \frac{y_{\frac{i}{2}+N^{''}} + y_{\frac{i}{2}+3N^{''}}}{\sqrt{2}}$$

and

$$x_{2i+2} + x_{2i+3} = x_{i+1} + x_{i+1+N'} \implies \sqrt{2} y_{i+1} = \frac{y_{\frac{i}{2}} - y_{\frac{i}{2}+N'}}{\sqrt{2}} + \frac{y_{\frac{i}{2}+N^{''}} - y_{\frac{i}{2}+3N^{''}}}{\sqrt{2}}.$$

Then

$$\begin{aligned} 2y_i &= y_{\frac{i}{2}} + y_{\frac{i}{2}+N'} + y_{\frac{i}{2}+N^{''}} + y_{\frac{i}{2}+3N^{''}}, \\ 2y_{i+1} &= y_{\frac{i}{2}} + y_{\frac{i}{2}+N^{''}} - y_{\frac{i}{2}+2N^{''}} - y_{\frac{i}{2}+3N^{''}}. \end{aligned}$$

Adding then subtracting these two equations, and letting $i = 2l$ for $l = 0, 1, \ldots, N^{''} - 1$, we have

$$\begin{aligned} y_{2l} + y_{2l+1} &= y_l + y_{l+N^{''}}, \\ y_{2l} - y_{2l+1} &= y_{l+2N^{''}} + y_{l+3N^{''}}. \end{aligned}$$

The second equation in (4.4) easily becomes

$$\sum_{i=0}^{N^{'}-1} y_i = 0.$$

We will split the above equations into two sets. The first set is

$$\begin{cases} y_{2l} + y_{2l+1} = y_l + y_{l+N''}, & l = 0, 1, \ldots, N^{''} - 1, \\ \sum_{l=0}^{N^{'}-1} y_l = 0, \end{cases} \tag{6.5}$$

while the second is

$$y_{2l} - y_{2l+1} = y_{l+2N''} + y_{l+3N''}; \quad l = 0, 1, \ldots, N^{''} - 1. \tag{6.6}$$

It is worth noting that the first set of equations is equivalent to equations (4.4) for the operator B_{k-1}.

The second change of variables: This includes the same change as before for $\{y_0, \ldots, y_{N'-1}\}$ and additional change connecting the two subsets of variables $\{y_i : i \le N^{'} - 1\}$ and $\{y_j : j \ge N^{'}\}$. For $i = 0, 1, \ldots, N^{''} - 1$ put

$$\left.\begin{array}{l} \dfrac{y_{2i} + y_{2i+1}}{\sqrt{2}} = w_i \\ \dfrac{y_{2i} - y_{2i+1}}{\sqrt{2}} = w_{i+N''} \end{array}\right\} \Longrightarrow \begin{cases} y_{2i} = \dfrac{w_i + w_{i+N''}}{\sqrt{2}} \\ y_{2i+1} = \dfrac{w_i - w_{i+N''}}{\sqrt{2}}. \end{cases}$$

The total number of the above variables is $2N^{''} = N^{'}$. The other $N^{'}$ ones are as follows:

$$\begin{cases} \dfrac{y_{i+2N''} + y_{l+3N''}}{\sqrt{2}} = w_{i+2N''}, \\ \dfrac{y_{i+2N''} - y_{i+3N''}}{\sqrt{2}} = w_{i+3N''}, \end{cases} \qquad i = 0, 1, \ldots, N^{''} - 1.$$

The $L_0^{\perp}$ equations (6.5) and (6.6) at this stage have the form

$$\begin{cases} w_{2i} + w_{2i+1} = w_i + w_{i+N^{(3)}}, & i = 0, 1, \ldots, N^{(3)} - 1, \\ \sum_{i=0}^{N^{''}-1} w_i = 0, \end{cases}$$

and

$$w_{i+N^{''}} = w_{i+N'}, \quad i = 0, 1, \ldots, N^{''} - 1,$$

while the quadratic form is

$$f_k(\boldsymbol{w}) = \sum_{i=0}^{N-1} w_i^2 + 2\sum_{i=0}^{N^{'}-1} w_i^2 + 2\sum_{i=0}^{N^{''}-1} w_i^2 + \sum_{i=0}^{N^{(3)}-1} (w_{2i} + w_{2i+1})^2 + \cdots + \frac{1}{2^{k-4}}\left[\left(w_0 + \cdots + w_{N^{(3)}-1}\right)^2 + \left(w_{N^{(3)}} + \cdots + w_{N''-1}\right)^2\right]$$

subject to $\sum_{i=0}^{N-1} w_i^2 = 1$.

To see how the induction goes, we should write down another transformation. Introduce a new vector of variables $\boldsymbol{t} = (t_1, \ldots, t_{N-1})$: For $i = 0, \ldots, N^{(3)} - 1$ let

$$t_i = \frac{w_{2i} + w_{2i+1}}{\sqrt{2}},$$
$$t_{i+N^{(3)}} = \frac{w_{2i} - w_{2i+1}}{\sqrt{2}},$$
$$t_{i+2N^{(3)}} = \frac{w_{i+2N^{(3)}} + w_{i+3N^{(3)}}}{\sqrt{2}},$$
$$t_{i+3N^{(3)}} = \frac{w_{i+2N^{(3)}} - w_{i+3N^{(3)}}}{\sqrt{2}},$$

and for $i = N', \ldots, N-1$, let $t_i = w_i$. Clearly, the transformation $(\boldsymbol{w} \to \boldsymbol{t})$ acts on the first half of the state space $0, 1, \ldots, N' - 1$ in the same way as the first transformation $(\boldsymbol{x} \to \boldsymbol{y})$ acts on $0, 1, \ldots, N-1$. The next transformation (which we omit) will therefore yield

$$t_{i+N^{(3)}} = t_{i+N''}; \quad i = 0, \ldots, N^{(3)} - 1.$$

Proceeding in this fashion we see that the $L_0^{\perp}$ equations will be hierarchically transformed into those displayed in Lemma 6.1(iii). Furthermore, the quadratic form becomes

$$f_k(\boldsymbol{z}) = \sum_{i=0}^{N-1} z_i^2 + 2\sum_{i=0}^{N'-1} z_i^2 + 2\sum_{i=0}^{N''-1} z_i^2 + \cdots + 2\sum_{i=0}^{N^{(l)}-1} z_i^2 + \cdots + 2(z_0^2 + z_1^2) \quad (6.7)$$

under the condition

$$\sum_{i=0}^{N-1} z_i^2 = 1.$$

□

Now the main result is formulated as a simple statement:

Theorem 6.2 (Main Theorem). *The full spectrum of the limiting covariance matrix B_k of the system of local times $\{\tau^*(\xi)\}$ is given, with the corresponding multiplicities M_λ, by the following table* (*where the eigenvalues are multiplied by* 2^k)

λ	0	1	...	$k-2$	$k-1$	k
M_λ	2^{k-1}	2^{k-2}	...	2	1	1

Proof. It is an immediate consequence of Lemma 6.1. In fact, Lemma 6.1(iii) shows that the quadratic form $2^k(B_k\boldsymbol{x} \cdot \boldsymbol{x})$ after applying the hierarchical chain of orthogonal transformations (described in Lemma 6.1) has the independent variables

$$z_1;\ z_3;\ z_6,\ z_7;\ z_{12},\ z_{13},\ z_{14},\ z_{15};\ \ldots;\ z_{3N^{(3)}}, \ldots, z_{N'-1};\ z_{3N''}, \ldots, z_{N'-1}$$

whose number is $1 + 1 + 2 + 4 + \cdots + 2^{k-2} = 2^{k-1} = N'$.

In these variables the normalizing form $(z \cdot z) = 1$ can be presented as

$$kz_1^2 + (k-1)z_3^2 + (k-2)(z_6^2 + z_7^2) + \cdots + 2(z_{3N^{(3)}}^2 + \cdots + z_{N'-1}^2) + 1 \cdot (z_{3N''}^2 + \cdots + z_{N-1}^2),$$

while the form $f_k(z)$ given in equation (6.7) reduces to

$$\begin{aligned}
&kz_1^2 + (k-1)z_3^2 + (k-2)(z_6^2 + z_7^2) + \cdots \\
&\quad + 2(z_{3N^{(3)}}^2 + \cdots + z_{N''-1}^2) + 1 \cdot (z_{3N''}^2 + \cdots + z_{N-1}^2) \\
&\quad + 2\big[(k-1)z_1^2 + (k-2)z_3^2 + (k-3)(z_6^2 + z_7^2) + \cdots + 1 \cdot (z_{3N^{(3)}}^2 + \cdots + z_{N''-1}^2)\big] \\
&\quad + 2\big[(k-2)z_1^2 + (k-3)z_3^2 + \cdots + 1 \cdot (z_{3N^{(4)}}^2 + \cdots + z_{N^{(3)}-1}^2)\big] + \cdots + 2z_1^2 \\
&= k^2 z_1^2 + (k-1)^2 z_3^2 + (k-2)^2 \left(z_6^2 + z_7^2\right) + \cdots \\
&\quad + 2^2 \left(z_{3N^{(3)}}^2 + \cdots + z_{N'-1}^2\right) + 1 \cdot (z_{3N''}^2 + \cdots + z_{N-1}^2).
\end{aligned}$$

The last equality follows from the simple identity

$$l + 2(l-1) + 2(l-2) + \cdots + 2 = l + 2 \cdot \frac{l(l-1)}{2} = l^2,$$

applied for $l = 1, \ldots, k$. The result now follows using Proposition 4.1. □

7 The eigenvectors of B_k

For the practical statistical applications of the CLT to K–P–S chains on W_k it remains to know the eigenvectors of B_k. So far we know the spectrum of B_k for any given k. To construct the corresponding eigenvectors one can use a computer software (e.g., MAPLE) for small k. Table 3 (p. 32) displays the two top eigenvectors for the cases $k = 3, \ldots, 6$. These vectors were evaluated using MAPLE. As a matter of fact, the use of MAPLE helped us to formulate the main result and discover the hierarchical structure of B_k discussed in Section 4. However, the use of computer gets more difficult for higher values of k, (e.g., for $k = 10$ the number of entries in B_k exceeds 10^6). In this section we will present efficient recursive algorithms to evaluate the two top eigenvectors corresponding to the simple eigenvalues $\lambda_{k-1} = k - 1$ and $\lambda_k = k$. We will also prove an important identity relating the l_2 and l_∞ norms of these eigenvectors to their corresponding eigenvalues. We first make the remark that the N'' eigenvectors that correspond to the eigenvalue $\frac{1}{N}$ are given by the following very simple formula:

$$v_{ij} = \begin{cases} 1, & \text{if } j = 2i,\ 2i + 1 + N', \\ -1, & \text{if } j = 2i + 1,\ 2i + N', \\ 0, & \text{otherwise.} \end{cases} \tag{7.1}$$

These vectors can be replaced by even and odd vectors in a fashion similar to the one described after Remark 4.4. We next characterize the two top eigenvectors by formulating the following recursive algorithms:

Algorithm 1. Let $\nu^{(k-1)} = [y_0, y_1, \ldots, y_{N'-1}]^*$ be the top eigenvector for the operator $2^{k-1}B_{k-1}$ with minimal integer representation (the entries of $\nu^{(k-1)}$ are relatively prime integers). Expand $\nu^{(k-1)}$ to an N-dimensional vector as follows:

$$\nu^{(k)*} = \begin{cases} [\nu^{(k-1)} : \mathbf{1}_{N'}]^*, & \text{if } k \text{ is even,} \\ [\nu^{(k-1)} : \frac{1}{2}\mathbf{1}_{N'}]^*, & \text{if } k \text{ is odd,} \end{cases}$$

then the top eigenvector with minimal integer representation $\nu^{(k)}$ of B_k is given by

$$\nu^{(k)} = \begin{cases} \frac{1}{2}T\left(\nu^{(k)*}\right), & \text{if } k \text{ is even,} \\ 2T\left(\nu^{(k)*}\right), & \text{if } k \text{ is odd,} \end{cases}$$

where $T([y_0, \ldots, y_{N-1}]^*) = [x_0, \ldots, x_{N-1}]^*$ is given by

$$\begin{cases} x_{2i} = y_i + y_{i+N'} \\ x_{2i+1} = y_i - y_{i+N'} \end{cases}$$

for $i = 0, \ldots, N' - 1$.

Algorithm 2. Let $\eta^{(k-1)}$ be the second top eigenvector for the operator $2^{k-1}B_{k-1}$ with minimal integer representation. In Algorithm 1, replace $\nu^{(k)}$ and $\nu^{(k)*}$ with $\eta^{(k)}$ and $\eta^{(k)*}$ where

$$\eta^{(k)*} = \begin{cases} [\eta^{(k-1)} : \boldsymbol{\xi}_{N'}]^*, & \text{if } k \text{ is even,} \\ [\eta^{(k-1)} : \frac{1}{2}\boldsymbol{\xi}_{N'}]^*, & \text{if } k \text{ is odd,} \end{cases}$$

and $\boldsymbol{\xi}_{N'} = [\underbrace{1, -1, \ldots, 1, -1}_{N'}]^*$. It should be noted that the remaining eigenvectors also admit simple recursive structures similar to the ones given above. The following proposition and its proof will justify Algorithm 1.

Proposition 7.1. *The top eigenvector $\nu^{(k)} = [x_0, \ldots, x_{2^k-1}]^*$ of $2^k B_k$ belongs to the space $L^{\perp}_{\text{odd}}$, and it can be generated recursively by Algorithm* 1. *Furthermore, if we choose $\nu^{(k)}$ with a minimal integer representation, and $x_0 > 0$, then the l_∞ norm of $\nu^{(k)}$ is*

$$\left\|\nu^{(k)}\right\|_\infty = x_0 = \begin{cases} \frac{k}{2}, & \text{if } k \text{ is even,} \\ k, & \text{if } k \text{ is odd.} \end{cases}$$

Also, $x_{2i} - x_{2i+1}$ is constantly equal to 1 *when k is odd and* 2 *when k is even.*

Proof. It is done by induction. Looking at Table 3, we see that the base case is satisfied. Let k be an even integer at first. We need to verify the following:

(i) $\nu^{(k)}$ is odd;

(ii) $x_{2i} - x_{2i+1} = 1$;

(iii) $\left\|\nu^{(k)}\right\|_\infty = \frac{k}{2}$;

(iv) $2^k\,(B_k \boldsymbol{x} \cdot \boldsymbol{x}) = k \cdot 2^k$.

To prove (i) we first let $i = 2l$, then

$$x_{N-i-1} = x_{(2N'-l-1)+1} = \frac{y_{N'-l-1} - 1}{2} = \frac{-y_l - 1}{2},$$

but $x_i = \frac{y_l+1}{2}$. If $i = 2l + 1$, then

$$x_{N-i-1} = x_{2\left(N'-l-1\right)} = \frac{y_{\left(N'-l-1\right)+1}}{2} = \frac{-y_l + 1}{2} = -\frac{y_l - 1}{2} = -x_i.$$

Statement (ii) is obvious. In order to check (iii), look at

$$\begin{aligned}\left\|\nu^{(k)}\right\|_\infty = \max_{0\le i<N} |x_i| &= \max_{0\le i<N'} \left\{\left|\frac{y_i + 1}{2}\right|, \left|\frac{y_i - 1}{2}\right|\right\} \\ &= \max_{0\le i<N} \left|\frac{y_i + 1}{2}\right| = \frac{y_0 + 1}{2} = \frac{k}{2},\end{aligned}$$

where we used the inductive assumption $\left\|\nu^{(k-1)}\right\|_\infty = y_0 = k - 1$. The following remark will be used to finish the proof.

Remark 7.2. The L_2 norm of $\nu^{(k)}$ is given by

$$\left\|\nu^{(k)}\right\|_2^2 = \begin{cases} k \cdot 2^{k-2}, & \text{if } k \text{ is even,} \\ k \cdot 2^k, & \text{if } k \text{ is odd.}\end{cases}$$

This remark can be proved by a simple induction argument.

We now verify statement (iv), that is we verify that $\nu^{(k)}$ is indeed the top eigenvector for $2^k B_k$. Using the above Remark together with the normalization condition $(\boldsymbol{x} \cdot \boldsymbol{x}) = 1$ (see Section 4), we need to check that $2^k \left(B_k \frac{\nu^{(k)}}{\left\|\nu^{(k)}\right\|_2} \cdot \frac{\nu^{(k)}}{\left\|\nu^{(k)}\right\|_2}\right) = k$, or:

$$2^k \left(B_k \nu^{(k)} \cdot \nu^{(k)}\right) = \begin{cases} k^2 \cdot 2^{k-2}, & \text{if } k \text{ is even,} \\ k^2 \cdot 2^k, & \text{if } k \text{ is odd.}\end{cases}$$

In fact, for an even k (4.7) and Algorithm 1 imply (using the induction hypothesis)

$$\begin{aligned}2^k(B_k \nu^{(k)} \cdot \nu^{(k)}) &= \left\|\nu^{(k)}\right\|_2^2 + \frac{1}{2}\left\|\nu^{(k-1)}\right\|_2^2 + \frac{1}{2}\left[2^{k-1}(B_{k-1}\nu^{(k-1)} \cdot \nu^{(k-1)})\right] \\ &= [k + (k-1) + (k-1)^2]2^{k-2} = k^2 \cdot 2^{k-2}.\end{aligned}$$

For an odd k (4.7) reads

$$\begin{aligned} 2^k \left(B_k \nu^{(k)} \cdot \nu^{(k)} \right) &= \left\| \nu^{(k)} \right\|_2^2 + 8 \left\| \nu^{(k-1)} \right\|_2^2 + 8 \left[2^{k-1} \left(B_{k-1} \nu^{(k-1)} \cdot \nu^{(k-1)} \right) \right] \\ &= k \cdot 2^k + 8\,(k-1) \cdot 2^{k-3} + 8\,(k-1)^2 \cdot 2^{k-3} \\ &= [k + (k-1) + (k-1)^2] 2^k = k \cdot 2^k. \end{aligned}$$

Now in order to prove the claim that $\nu^{(k)}$ has minimal integer representation, it is enough to show that for every k, $\nu^{(k)}$ has 1 as an entry. This is the case in fact for $k = 2$ and $k = 3$. In general, suppose $y_l = 1$ is an entry in $\nu^{(k-1)}$. If k is even, $\frac{y_l+1}{2} = 1$ belongs to $\nu^{(k)}$, otherwise we look at $2y_l - 1 = 1$. This proves the claim. Now to finish the proof of Proposition 7.1 one can imitate the above argument for an odd k. □

As a direct consequence of Proposition 7.1 we obtain the following useful fact that connects two different norms of $\nu^{(k)}$ to the corresponding eigenvalue of $\nu^{(k)}$.

Corollary 7.3. *The ratio of the l_∞ norm of $\nu^{(k)}$ to its l_2 norm is equal to the square root of the top eigenvalue, i.e.,*

$$\frac{\left\| \nu^{(k)} \right\|_\infty}{\left\| \nu^{(k)} \right\|_2} = \sqrt{\lambda_{\max}} = \sqrt{\frac{k}{2^k}}.$$

Proof. Using Remark 7.2 and Proposition 7.1 we see that

$$\frac{\left\| \nu^{(k)} \right\|_\infty^2}{\left\| \nu^{(k)} \right\|_2^2} = \frac{\left(\frac{k}{2}\right)^2}{\left(k \cdot 2^{k-2}\right)} = \frac{\frac{k^2}{4}}{\left(k \cdot 2^{k-2}\right)} = \frac{k}{2^k},$$

if k is even, and that the ratio above is equal to $k^2/(k \cdot 2^k) = k/2^k$, if k is odd. This establishes the corollary. □

The justification of Algorithm 2 is a direct imitation of the proof of Proposition 7.1. It can also be proven that the relation in Corollary 7.3 still holds for the second top eigenvector. That is,

Corollary 7.4. *The ratio of the l_∞ norm of $\eta^{(k)}$ to its l_2 norm is equal to the square root of the second top eigenvalue, i.e.,*

$$\frac{\left\| \eta^{(k)} \right\|_\infty}{\left\| \eta^{(k)} \right\|_2} = \sqrt{\frac{k-1}{2^k}}.$$

For each $k = 3, 4, 5, 6$, the two top eigenvectors $\eta^{(k)}$, $\nu^{(k)}$ are displayed, the top eigenvector being always on the right. For $k = 6$ only half of the entries are displayed. The other halves are even and odd continuations respectively.

Table 3. Top eigenvectors

3	4	5	6
$\begin{bmatrix}1\\0\\-1\\0\\0\\-1\\0\\1\end{bmatrix}$ $\begin{bmatrix}3\\1\\1\\-1\\1\\-1\\-1\\-3\end{bmatrix}$	$\begin{bmatrix}3\\1\\-1\\1\\-1\\-3\\-1\\1\\1\\-1\\-3\\-1\\1\\-1\\3\end{bmatrix}$ $\begin{bmatrix}2\\1\\1\\0\\1\\0\\0\\-1\\1\\0\\0\\-1\\0\\-1\\-2\end{bmatrix}$	$\begin{bmatrix}2\\1\\0\\1\\0\\-1\\0\\1\\0\\-1\\-2\\-1\\0\\1\\0\\-1\\0\\-1\\0\\1\\0\\-1\\0\\1\\2\end{bmatrix}$ $\begin{bmatrix}5\\3\\3\\1\\3\\1\\-1\\3\\1\\-1\\-1\\1\\-1\\0\\-1\\1\\-1\\-3\\3\\1\\1\\-1\\1\\-3\\-1\\-1\\0\\-1\\-3\\-5\end{bmatrix}$	$\begin{bmatrix}5\\3\\1\\3\\1\\-1\\1\\3\\1\\-1\\-3\\-1\\1\\-1\\1\\3\\-1\\2\\-1\\-3\\-1\\-3\\-5\\-3\\-1\\1\\-1\\-3\\-1\\1\\-1\\1\\3\\\vdots\end{bmatrix}$ $\begin{bmatrix}3\\2\\2\\1\\2\\1\\1\\0\\2\\1\\1\\0\\1\\0\\0\\-1\\2\\1\\1\\0\\1\\0\\0\\-1\\1\\0\\0\\-1\\-1\\-2\\\vdots\end{bmatrix}$

Acknowledgement. This paper was finished while the first author was visiting the University of Delaware. He is thankful to the faculty and staff at the department of Mathematical Sciences for their continual support and encouragement during his two years of visit.

References

[1] A. Alhakim, On a joint distribution for long runs, a limit theorem for approximate entropy with applications to the testing of random number generators, PhD Thesis, 2001.

[2] P. Billingsley, Convergence of Probability Measures, second edition, Wiley Ser. Probab. Stat., Wiley, New York 1999.

[3] F. J. Dyson, Existence of a phase-transition in a one-dimensional Ising ferromagnet, Comm. Math. Phys. 12 (1969), 91–107.

[4] A. Figotin, A. Gordon, S. Molchanov, J. Quinn and N. Stavrakas, Occupancy numbers in testing random number generators, SIAM J. Appl. Math. 62 (2002), 1980–2011 (electronic).

[5] J. G. Kemeny and J. L. Snell, Finite Markov Chains, The University Series in Undergraduate Mathematics, Van Nostrand, Princeton, NJ, 1960.

[6] A. N. Kolmogorov, A local limit theorem for Markov chains, in: Select. Transl. Math. Statist. and Probability, Vol. 2, Amer. Math. Soc., Providence, RI, 1962, 109–129.

[7] A. N. Kolmogorov and V. A. Uspensky, Algorithms and randomness, Theory Probab. Appl. 32 (1987), 389–412.

[8] A. N. Kolmogorov, Information Theory and the Theory of Algorithms (Russian), Nauka, Moscow 1987.

[9] G. A. Marsaglia, A current view of random number generators, in: Computer Science and Statistics: The Interface, Elsevier, Amsterdam 1985, 3–10.

[10] S. Molchanov, Hierarchical random matrices and operators. Application to Anderson model, in: Multidimensional Statistical Analysis and Theory of Random Matrices (Bowling Green, OH, 1996), VSP, Utrecht 1996, 179–194.

[11] S. Pincus and R. E. Kalman, Not all (possibly) "random" sequences are created equal. Proc. Nat. Acad. Sci. U.S.A. 94 (1997), 3513–3518.

[12] S. Pincus and B. H. Singer, Randomness and degrees of irregularity, Proc. Nat. Acad. Sci. U.S.A. 93 (1996), 2083–2088.

[13] A. L. Rukhin, Approximate entropy for testing randomness, J. Appl. Probab. 37 (2000), 88–100.

[14] L. Saloff-Coste, Lectures on finite Markov chains, in: Lectures on Probability Theory and Statistics (Saint-Flour, 1996), Lecture Notes in Math. 1665, Springer-Verlag, Berlin 1997, 301–413.

Abbas Alhakim, Division of Mathematics and Computer Science, Clarkson University, Science Center Box 5815, Potsdam, NY 13699, USA

E-mail: aalhakim@clarkson.edu

Stanislav Molchanov, Department of Applied Mathematics, University of North Carolina at Charlotte, 9201 University City Blvd. Charlotte, NC 28223, USA

E-mail: smolchan@uncc.edu

Random walks and physical models on infinite graphs: an introduction

Raffaella Burioni, Davide Cassi and Alessandro Vezzani

Abstract. This paper is a review of some basic mathematical ideas and results, concerning the relations between random walks and physical models on infinite graphs from the physicists point of view. The presentation is mainly focused on statistical models, which are particularly relevant in the physics of matter and in field theory.

Contents

1 Introduction 36
- 1.1 Definitions and notations . . . 37

2 The thermodynamic limit 38
- 2.1 Distance, Van Hove sphere and growth exponent . . . 38
- 2.2 Physical conditions . . . 38
- 2.3 Averages in the thermodynamic limit and sets measure . . . 40

3 Random walks 42
- 3.1 Definitions . . . 42
- 3.2 The local type problem . . . 43
- 3.3 The local spectral dimension . . . 43

4 Thermodynamic averages and random walks 46
- 4.1 Recurrence and transience on the average . . . 46
- 4.2 The average spectral dimension . . . 50
- 4.3 Pure and mixed transience on the average . . . 51
- 4.4 Separability and statistical independence . . . 52

5 Harmonic oscillations 55
- 5.1 The physical model . . . 55
- 5.2 The spectrum of the Laplacian . . . 56

6 The Gaussian model 57
- 6.1 The model . . . 57

6.2 The walk expansion 59
6.3 Gaussian model and spectral dimensions 63
6.4 Universality properties of $\bar{d}$ 64

7 Conclusions 69

1 Introduction

In the last twenty years, theoretical physicists have shown an increasing interest in random walks on infinite graphs, connected with the study of physical properties of inhomogeneous and disordered systems in the thermodynamic limit.

A relevant number of papers appeared on this subject, concerning applications to polymers, glasses, fractals, amorphous solids, disordered magnets, biological matter, electronic states, diffusion and transport phenomena (e.g., see [1, 5, 21, 24, 25, 28, 30, 32]). In the meanwhile, a specific mathematical formalism has been introduced in the physical literature to deal with such kind of problems, and a new language has developed among the researchers involved in this field, often alternative to the usual graph-theoretical one. Moreover, the study of physical problems on infinite graphs led to definition of brand new mathematical concepts and to the proof of theorems concerning them.

Only recently a real collaboration between mathematicians and physicists working on models on infinite graphs has begun, due to the initiative of both (Statistical mechanics and graph theory 2000 ICTP Trieste, Random Walks and Statistical Physics 2001 ESI Wien). To improve the exchange of ideas and expertise of the two communities, a common effort of "translation" of basic concepts and tools of each field is now of primary importance. This paper has to be viewed as a first relevant step in this direction from the physicists-side.

Our main aim is to give a self contained introduction to some basic physical models on infinite graphs, emphasizing several mathematical details, usually skipped in the papers written by physicists. Therefore, we decided to limit ourselves to a restricted class of fundamental ideas and results, which can be rigorously stated and proven. Due to this choice, many interesting topics are not discussed here, such as electrical networks, magnetic models and quantum models; for all of them, we refer the reader to the existing literature for more specific applications.

One of the most difficult problems in our task is undoubtedly the "physical reality" hypothesis implicit in all physical works: by this term we mean a series of unexpressed conditions sufficient to produce a set of behaviours observed in real systems.

Let us give an example: all real physical structures (embedded in three-dimensional space) have been found up to now to exhibit power law behaviour in the low-frequency density of vibrational states and therefore, when considering an infinite graph where a physical model is defined, one always assumes that it satisfies the (often unknown) mathematical conditions sufficient to produce such a behaviour. In our opinion the study of these "physical reality" conditions is now the most promising and interesting

field for a fruitful collaboration between mathematicians and physicists. To this aim, we always explicitly state all the mathematical conditions usually assumed in physical literature, pointing out in the Remarks the still open points or only heuristically "solved" problems.

We hope that this review will be useful to the mathematical community from at least two different point of view: first, it would offer a collection of unsolved mathematical problems, whose solution would be of great importance to physics; second, it should make the interested reader able to understand the language and the ideas which can be found in advanced physical literature concerning infinite graphs.

1.1 Definitions and notations

Let us introduce some definitions and notations that will be useful in the rest of the paper [4, 20, 23].

Definition 1.1. A graph X is a countable set V_X of vertices (or sites) (i) connected pairwise by a set E_X of unoriented edges (or links) $(i, j) = (j, i)$. Two connected vertices are called nearest neighbours. We denote by z_i the connectivity of the site i, i.e. the number of its nearest neighbours.

Definition 1.2. A path in X is a sequence of consecutive edges $\{(i, k)(k, h) \ldots (n, m)(m, j)\}$ and its length is the number of edges in the sequence. A graph is said to be connected if, for any two vertices $i, j \in V_X$, there is always a path joining them.

Definition 1.3. The adjacency matrix A_{ij} is:

$$A_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E_X, \\ 0 & \text{if } (i, j) \notin E_X. \end{cases} \tag{1.1}$$

Definition 1.4. The Laplacian matrix Δ_{ij} is:

$$\Delta_{ij} = z_i \delta_{ij} - A_{ij}. \tag{1.2}$$

Notice that: $z_i = \sum_j A_{ij}$. We define $Z_{ij} = z_i \delta_{ij}$.

A generalization of the Laplacian matrix can be given:

Definition 1.5. The matrix J_{ij} is called a ferromagnetic coupling matrix, if $\exists J_{\max}$, $J_{\min} \in \mathbb{R}^+$:

$$J_{ji} = J_{ij} = \begin{cases} J_{\min} < J_{ij} < J_{\max} & \text{if } (i, j) \in E_X, \\ 0 & \text{if } (i, j) \notin E_X. \end{cases} \tag{1.3}$$

The generalized Laplacian associated to J_{ij} is:

$$L_{ij} = I_i \delta_{ij} - J_{ij}. \tag{1.4}$$

where $I_i = \sum_j J_{ij}$. We also define $I_{ij} = I_i \delta_{ij}$.

2 The thermodynamic limit

2.1 Distance, Van Hove sphere and growth exponent

Unoriented graphs are naturally provided with an intrinsic distance, which in physics is called the chemical distance $r_{i,j}$.

Definition 2.1. $r_{i,j}$ is the length of the shortest path connecting the vertices i and j. The distance between i and the subset $V' \subset V_X$ is $d(i, V') = \inf\{r_{i,k} \in \mathbb{N} |\ k \in V'\}$.

The chemical distance defines on the graph the balls of radius $r \in \mathbb{N}$ and center $o \in V_X$. In the physical literature these subgraphs are called the Van Hove spheres $\mathcal{S}_{o,r}$.

Definition 2.2. $\mathcal{S}_{o,r}$ is the subgraph of X, determined by the set of vertices $V_{o,r} = \{i \in V | r_{i,o} \leq r\}$ and by the set of edges $E_{o,r} = \{(i, j) \in E | i \in V_{o,r}, j \in V_{o,r}\}$. The border of $\mathcal{S}_{o,r}$ is given by the set $\partial V_{o,r} = \{i \in V_{o,r} | \exists j \in V_X, (i, j) \in E_{o,r}, j \notin V_{o,r}\}$. For $V' \subset V_X$, we also define $\tilde{V}_{V',r} = \{i \in V_X |\ d(i, V') \leq r\}$.

In some cases it is useful to introduce sequences of generalized spheres $\mathcal{S}'_{o,r}$, defined by sets $V'_{o,r} \subset V_X$ such that $V'_{o,0} = \{o\}$, $V'_{o,r} \subset V'_{o,r+1}$ and $\bigcup_{r=0}^{\infty} V'_{o,r} = V_X$ and by the sets of edges $E'_{o,r} = \{(i, j) \in E | i \in V'_{o,r}, j \in V'_{o,r}\}$. Here we always use for the sphere Definition 2.2.

Let $|S|$ be the cardinality of a set S. Then $|V_{o,r}|$, as a function of the distance r, describes the growth rate of the graph at the large scale [26]. In particular:

Definition 2.3. A graph is said to have a polynomial growth if $\forall o \in V_X\ \exists c, k$, such that $|V_{o,r}| < c\, r^k$.

Definition 2.4. For a graph satisfying (2.3), we define the upper growth exponent d_g^+ and the lower growth exponent d_g^- as $d_g^+ = \inf\{k |\ |V_{o,r}| < c_1\, r^k, \forall o \in V\}$ and $d_g^- = \sup\{k |\ |V_{o,r}| > c_2\, r^k, \forall o \in V\}$. If $d_g^+ = d_g^-$ we call them the growth exponent d_g, or the classical connectivity dimension.

The connectivity dimension d_g is known for a large class of graphs: on lattices $\mathbb{Z}^d$ it coincides with the usual Euclidean dimension d, and for many fractals it has been exactly evaluated [21].

2.2 Physical conditions

Discrete structures describing real physical systems are characterized by some important properties, which can be translated in mathematical requirements for the graphs we will consider.

p.c.1 We will consider only connected graphs (Definition 1.2), since any physical model on disconnected structures can be reduced to the separate study of the

models defined on each connected component and hence to the case of connected graphs.

p.c.2 Since physical interactions are always bounded, the coordination numbers z_i, representing the number of neighbours interacting with the site i, have to be bounded; i.e. $\exists z_{\max} \mid z_i \leq z_{\max}\ \forall i \in V_X$.

p.c.3 Real systems are always embedded in 3-dimensional space. This constraint requires for the graph $\mathcal{G}$ the conditions:

(a) X has a polynomial growth (Definition 2.3)

(b)

$$\lim_{r\to\infty} \frac{|\partial V_{o,r}|}{|V_{o,r}|} = 0 \tag{2.5}$$

The existence itself of the limit is a physical requirement on $\mathcal{G}$.

Some interesting graphs such as the Bethe lattice do not satisfy (a) and (b). For this kind of structures many results we give in this paper do not apply and one has to introduce different techniques.

Remark 2.5. For a large class of physically interesting graphs we have considered so far, conditions (a) and (b) appears to be equivalent. However for the equivalence of the two conditions a rigorous result is still lacking.

A graph satisfying **p.c.1**, **p.c.2** and **p.c.3** will be called *physical graph* $\mathcal{G}$ and the sets of its vertices and edges will be denoted respectively with V an E. **p.c.1** and **p.c.2** represent strong constraints on $\mathcal{G}$ and, as we will prove in detail, they have very important consequences. For example, **p.c.1** implies a simple but important limitation on the difference of size for spheres of different centers.

Theorem 2.6. *Given a physical graph $\mathcal{G}$, let $\mathcal{S}_{o,r}$ and $\mathcal{S}_{o',r}$ be two spheres of centers o and o', respectively, and radius r. One has:*

$$||V_{o,r}| - |V_{o',r}|| \leq (z_{\max})^{2r_{o,o'}} |\partial V_{o,r}|. \tag{2.6}$$

Proof. Since $V_{o',r} \subset V_{o,r+r_{o,o'}}$,

$$|V_{o',r}| \leq |V_{o,r+r_{o,o'}}| \leq |V_{o,r}| + |V_{o,r+r_{o,o'}} \triangle V_{o,r}|,$$

where $\triangle$ denotes the symmetric difference. Now we have $|V_{o,r} \triangle V_{o,r+r_{o,o'}}| < |\widetilde{V}_{\partial V_{o,r}, r_{o,o'}}|$, where $|\widetilde{V}_{\partial V_{o,r}, r_{o,o'}}|$, as in Definition 2.2, is the number of sites whose distance from $\partial V_{o,r}$ is smaller than $r_{o,o'}$. Form the uniform boundedness of z_i one obtains $|\widetilde{V}_{\partial V_{o,r}, r_{o,o'}}| \leq (z_{\max})^{r_{o,o'}} |\partial V_{o,r}|$, and then:

$$|V_{o',r}| \leq |V_{o,r}| + (z_{\max})^{r_{o,o'}} |\partial V_{o,r}|. \tag{2.7}$$

From the properties of the distance, $r_{i,o} - r_{o,o'} \leq r_{i,o'} \leq r_{i,o} + r_{o,o'}$, hence $\forall i$ such that $r_{i,o} = r$ (i.e $i \in \partial V_{o,r}$) we have $r - r_{o,o'} \leq r_{i,o'} \leq r + r_{o,o'}$ and $i \in \widetilde{V}_{\partial V_{o,r}, r_{o,o'}}$. So again from boundedness of z_i:

$$|\partial V_{o,r}| \leq (z_{\max})^{r_{o,o'}} |\partial V_{o',r}|. \tag{2.8}$$

Inequality (2.6) is a simple consequence of (2.7) and (2.8). □

2.3 Averages in the thermodynamic limit and sets measure

Thermodynamic averages have a crucial role in the study of statistical models on discrete structures. This requires the introduction of infinite graphs and the study of the limit $r \to \infty$ for the Van Hove spheres [13].

Definition 2.7. Given a physical graph $\mathcal{G}$, let $\phi_i : V \to \mathbb{R}$. The average in the thermodynamic limit of ϕ_i is:

$$\overline{\phi} \equiv \lim_{r\to\infty} \frac{\sum_{i\in V_{o,r}} \phi_i}{|V_{o,r}|}. \tag{2.9}$$

The existence itself of limit (2.9) is a physical requirement on the functions ϕ_i.

In [2] more general averages are defined giving to each site a weight $\lambda_{i,r}$. Physical constrains on graph structures, given in Section 2.2, has important consequences for the behaviour of the thermodynamic averages, such as the independence of the limit (2.9) from the choice of the center o.

Theorem 2.8. *Let $\mathcal{G}$ be a physical graph and $\phi_i : V \to \mathbb{R}$ a function bounded from below, i.e. $\phi_i > \phi_{\min}$ $\forall i \in V$. If limit* (2.9) *exists for the Van Hove spheres of center o', then it exists for any possible center o and the result does not depend on o.*

Proof. For any two vertices o and o', we have:

$$\begin{aligned} &\frac{\sum_{i\in V_{o',r-r_{o,o'}}} \phi_i \; + \; \sum_{i\in V_{o,r}\triangle V_{o',r-r_{o,o'}}} \phi_i}{|V_{o,r}|} \\ &\quad = \frac{\sum_{i\in V_{o,r}} \phi_i}{|V_{o,r}|} = \frac{\sum_{i\in V_{o',r+r_{o,o'}}} \phi_i \; - \; \sum_{i\in V_{o',r+r_{o,o'}}\triangle V_{o,r}} \phi_i}{|V_{o,r}|}, \end{aligned} \tag{2.10}$$

where $V_{o,r} \subseteq V_{o',r+r_{o,o'}}$, $V_{o',r-r_{o,o'}} \subseteq V_{o,r}$. From the boundedness of ϕ_i:

$$\begin{aligned} &\frac{\sum_{i\in V_{o',r-r_{o,o'}}} \phi_i - \phi_{\min}|V_{o,r} \,\triangle\, V_{o',r-r_{o,o'}}|}{|V_{o,r}|} \\ &\quad \leq \frac{\sum\limits_{i\in V_{o,r}} \phi_i}{|V_{o,r}|} \leq \frac{\sum_{i\in V_{o',r+r_{o,o'}}} \phi_i + \phi_{\min}|V_{o',r+r_{o,o'}} \,\triangle\, V_{o,r}|}{|V_{o,r}|}. \end{aligned} \tag{2.11}$$

In analogy with (2.6) one proves:

$$\begin{aligned}
&|V_{o,r} \triangle V_{o',r-r_{o,o'}}| \le (z_{\max})^{r_{o,o'}} |\partial V_{o,r}|,\\
&|V_{o,r} \triangle V_{o',r-r_{o,o'}}| \le (z_{\max})^{r_{o,o'}} |\partial V_{o,r-r_{o,o'}}|,\\
&|V_{o,r+r_{o,o'}} \triangle V_{o',r}| \le (z_{\max})^{r_{o,o'}} |\partial V_{o,r+r_{o,o'}}|,\\
&|V_{o,r+r_{o,o'}} \triangle V_{o',r}| \le (z_{\max})^{r_{o,o'}} |\partial V_{o,r}|,
\end{aligned} \tag{2.12}$$

and with property (2.5) we get:

$$\lim_{r\to\infty} \frac{\sum_{i\in V_{o',r-r_{o,o'}}} \phi_i}{|V_{o,r}|} \le \lim_{r\to\infty} \frac{\sum_{i\in V_{o,r}} \phi_i}{|V_{o,r}|} \le \lim_{r\to\infty} \frac{\sum_{i\in V_{o',r+r_{o,o'}}} \phi_i}{|V_{o,r}|},$$

and

$$\begin{aligned}
&\lim_{r\to\infty} \frac{\sum_{i\in V_{o',r-r_{o,o'}}} \phi_i}{|V_{o',r-r_{o,o'}}| + |V_{o,r} \triangle V_{o',r-r_{o,o'}}|}\\
&\quad \le \lim_{r\to\infty} \frac{\sum_{i\in V_{o,r}} \phi_i}{|V_{o,r}|} \le \lim_{r\to\infty} \frac{\sum_{i\in V_{o',r+r_{o,o'}}} \phi_i}{|V_{o',r+r_{o,o'}}| - |V_{o,r} \triangle V_{o',r-r_{o,o'}}|}.
\end{aligned} \tag{2.13}$$

Using again property (2.5) and inequalities (2.12) we get:

$$\lim_{r\to\infty} \frac{\sum_{i\in V_{o',r-r_{o,o'}}} \phi_i}{|V_{o',r-r_{o,o'}}|} \le \lim_{r\to\infty} \frac{\sum_{i\in V_{o,r}} \phi_i}{|V_{o,r}|} \le \lim_{r\to\infty} \frac{\sum_{i\in V_{o',r+r_{o,o'}}} \phi_i}{|V_{o',r+r_{o,o'}}|}. \tag{2.14}$$

Therefore, if the limit with the spheres centered in o' exists, it gives the same result using as center any vertex o. □

In what follows we drop the index o when we evaluate thermodynamic averages. Now we can define the measure of the subsets of $V' \subset V$.

Definition 2.9. Given a physical graph $\mathcal{G}$, the measure of a subset $V' \subset V$ is $\|V'\| = \overline{\chi(V')}$, where $\chi_i(V')$ is the characteristic function defined as $\chi_i(V') = 1$ if $i \in V'$ and $\chi_i(V') = 0$ if $i \notin V'$. The measure of a subset of edges $E' \subset E$ is $\lim_{r\to\infty} |E'_r|/|V_r|$, where $E'_r = \{(i, j) \in E' |\ i \in V_r,\ j \in V_r\}$.

Since $\chi_i(V')$ is bounded from below, when the thermodynamic average exists, the value of the measure $\|V'\|$ does not depend on the choice of the center o. Unfortunately in some cases the limit defining the measure does not exist. A typical example is the subset of $\mathbb{Z}$ defined as $\{i \in \mathbb{Z} |\ 2^{2n} \le |i| \le 2^{2n+1}, \forall\ n \in \mathbb{N}\}$. However, these subsets are not very interesting from a physical point of view, for example, they cannot characterize sites with a certain thermodynamic property, since this property should not be additive. Hence we will consider only subsets with a well-defined measure.

3 Random walks

3.1 Definitions

Let us begin by recalling the basic definitions and results concerning (simple) random walks on infinite graphs. A more detailed treatment can be found in the mathematical reviews by Woess [35, 36].

Definition 3.1. The (simple) random walk on a graph X is defined by the jumping probability p_{ij} between nearest neighbours sites i and j:

$$p_{ij} = \frac{A_{ij}}{z_i} = (Z^{-1}A)_{ij}, \tag{3.15}$$

where $Z_{ij} = z_i \delta_{ij}$. The probability of reaching in t steps site j starting from i is:

$$P_{ij}(t) = (p^t)_{ij}. \tag{3.16}$$

We denote by $F_{ij}(t)$ the probability for a walker starting from i of reaching for the first time in t steps the site $j \neq i$, and by $F_{ii}(t)$ the probability of returning to the starting point i for the first time after t steps ($F_{ii}(0) = 0$).

The basic relationship between $P_{ij}(t)$ and $F_{ij}(t)$ is given by:

$$P_{ij}(t) = \sum_{k=0}^{t} F_{ij}(k) P_{jj}(t-k) + \delta_{ij}\delta_{t0}. \tag{3.17}$$

$F_{ij} \equiv \sum_{t=0}^{\infty} F_{ij}(t)$ turns out to be the probability of ever reaching the site j starting from i (or of ever returning to i if $j = i$). Therefore, $0 < F_{ij} \leq 1$.

Definition 3.2. The generating functions $\widetilde{P}_{ij}(\lambda)$ and $\widetilde{F}_{ij}(\lambda)$ are given by

$$\widetilde{P}_{ij}(\lambda) = \sum_{t=0}^{\infty} \lambda^t P_{ij}(t), \quad \widetilde{F}_{ij}(\lambda) = \sum_{t=0}^{\infty} \lambda^t F_{ij}(t), \tag{3.18}$$

where λ is a complex number.

From definition (3.18) by Abel's lemma we have that $\widetilde{F}_{ij}(\lambda)$ and $\widetilde{P}_{ij}(\lambda)$ are $\mathcal{C}^{\infty}$ functions in $[0, 1)$. Furthermore, $\widetilde{F}_{ij}(\lambda)$ is continuous also for $\lambda = 1$, while $\widetilde{P}_{ij}(\lambda)$ can diverge at this point.

Multiplying equations (3.17) by λ^t and then summing over all possible t we get:

$$\widetilde{P}_{ij}(\lambda) = \widetilde{F}_{ij}(\lambda)\widetilde{P}_{jj}(\lambda) + \delta_{ij}. \tag{3.19}$$

Lemma 3.3. *A simple bound on $\widetilde{F}_{ij}(\lambda)$ and on $\widetilde{P}_{ij}(\lambda)$ is given by:*

$$\widetilde{P}_{ij}(\lambda) \leq (1-\lambda)^{-1}, \quad \widetilde{F}_{ij}(\lambda) \leq (1-\lambda)^{-1}. \tag{3.20}$$

Proof. From $P_{ij}(t) < 1$, $F_{ij}(t) < 1$ and (3.18) one immediately obtains (3.20). □

In the following we will use the notations $\widetilde{P}_i(\lambda) \equiv \widetilde{P}_{ii}(\lambda)$ and $\widetilde{F}_i(\lambda) \equiv \widetilde{F}_{ii}(\lambda)$.

3.2 The local type problem

Infinite graphs can be classified by the long time asymptotic behaviour of simple random walks and in particular by the quantities $\widetilde{F}_i(1)$ and $\lim_{\lambda\to 1}\widetilde{P}_i(\lambda)$ [29].

Definition 3.4. A graph X is called *locally recurrent* if

$$\widetilde{F}_i(1) = 1 \quad \text{or, equivalently,} \quad \lim_{\lambda\to 1}\widetilde{P}_i(\lambda) = \infty \quad \forall i \in V_X. \tag{3.21}$$

On the other hand, X is called *locally transient* if:

$$\widetilde{F}_i(1) < 1 \quad \text{or, equivalently,} \quad \lim_{\lambda\to 1}\widetilde{P}_i(\lambda) < \infty \quad \forall i \in V_X. \tag{3.22}$$

The equivalences in the definitions (3.21) and (3.22) are simple consequences of equation (3.19). By using standard properties of Markov chains one can prove that (3.21) and (3.22) are independent of the vertex i [35], and then Definition 3.4 can be considered as a property of the graph itself.

Local transience and local recurrence satisfy important universality properties [35]. Indeed, local transience and recurrence do not change if we replace the jumping probabilities of the random walk (3.15) with the generalized jumping probabilities:

$$p_{ij} = \frac{J_{ij}}{I_i}. \tag{3.23}$$

In [35] the invariance of the local recurrence properties under a wide class of transformations of the graph itself is also proven. Local recurrence and transience are not modified by the addition a finite number of links or the introduction of second neighbour links on the graph. These invariances put into evidence that local recurrence and transience are determined only by the large scale topology of the graph.

3.3 The local spectral dimension

The behaviour of $\widetilde{P}_i(\lambda)$ for $\lambda \to 1^-$ can be used not only to classify the graph as locally transient or recurrent, but also to introduce the local spectral dimension $\widetilde{d}$ which can be considered as a finer invariant of the graph topology. The spectral dimension has been widely studied in physics [1, 5, 32], since it is closely connected with such important phenomena as the anomalous diffusion and the vibrational spectra of harmonic oscillations. In the following we will use the definition given in [24]. Since $\widetilde{P}_i(\lambda)$ for $\lambda < 1$ is a $\mathcal{C}^\infty$ differentiable function one can define the degree of recurrence of a graph.

Definition 3.5. Let $\widetilde{P}_i^{(n)}(\lambda)$ be:

$$\widetilde{P}_i^{(n)}(\lambda) = \left(\frac{d}{d\lambda}\right)^n \widetilde{P}_i(\lambda). \tag{3.24}$$

A graph $\mathcal{G}$ is recurrent of degree N if:

$$\lim_{\lambda\to 1^-} \widetilde{P}_i^{(n)}(\lambda) < \infty,\ \forall n < N, \quad \text{and} \quad \lim_{\lambda\to 1^-} \widetilde{P}_i^{(N)}(\lambda) = \infty. \tag{3.25}$$

Definition 3.6. Let X be recurrent of degree N. If the limit

$$D = \lim_{\lambda\to 1^-} \frac{\log(\widetilde{P}_i^{(N)}(\lambda))}{-\log(1-\lambda)} \tag{3.26}$$

exists, then the spectral dimension is $\widetilde{d} = 2(N - D + 1)$.

Lemma 3.7. *Let X be a recurrent of degree N graph with local spectral dimension $\widetilde{d}$. We have $\widetilde{d} \le 2$ if $N = 0$ and $2N \le \widetilde{d} \le 2(N+1)$ for $N \ge 1$.*

Proof. Since $\widetilde{P}_i^{(N)}(\lambda)) > 1$, we have $D \ge 0$ and the case $N = 0$ is proven. For the case $N \ge 1$ we have to show that $D \le 1$. Let us suppose that $D > 1$. We will prove that $\lim_{\lambda\to 1^-} \widetilde{P}_i^{(N-1)}(\lambda) = \infty$, hence $\mathcal{G}$ is recurrent of degree $N-1$ leading to a contradiction. From (3.26) we have that $\forall\epsilon > 0$, $\exists\lambda'$ such that $\forall\lambda'', \lambda' < \lambda'' < 1$

$$\widetilde{P}_i^{(N)}(\lambda'') > (1-\lambda'')^{-(D-\epsilon)}. \tag{3.27}$$

Integrating (3.27) between λ' and λ, we get:

$$\begin{aligned}\widetilde{P}_i^{(N-1)}(\lambda) > (D-1-\epsilon)^{-1}\big((1-\lambda)^{-(D-1-\epsilon)} - (1-\lambda')^{-(D-1-\epsilon)}\big) \\ + \widetilde{P}_i^{(N-1)}(\lambda').\end{aligned} \tag{3.28}$$

Hence, $\lim_{\lambda\to 1^-} \widetilde{P}_i^{(N-1)}(\lambda) = \infty$. □

In [24] the independence of $\widetilde{d}$ of the choice of site i is proven. Therefore, the local spectral dimension can be considered as a property of the graph. Furthermore, in [24] some invariance properties such as the invariance for the rescaling of the jumping probability (3.23) are also proven.

Locally recurrent graphs are recurrent of degree 0 and have local spectral dimension smaller than 2. On the Euclidean lattices $\mathbb{Z}^d$, $\widetilde{d} = d$ [27], hence $\widetilde{d}$ can be considered as a generalization of the usual notion of dimension for lattices. Moreover, $\widetilde{d}$ has been evaluated for many graphs such as exactly decimable fractals [24, 31] and bundled structures [3, 17, 18, 34] (The Sierpinski gasket in Fig. 1 with $\widetilde{d} = 2\log(3)/\log(5)$ and the comb graph in Fig. 2 with $\widetilde{d} = 1.5$ are two typical examples of exactly decimable fractals and bundled structures).

Definition 3.5 is more general than the usual definition of the local spectral dimension given in physics, i.e. $\widetilde{P}_i(\lambda) \sim (1-\lambda)^{\widetilde{d}/2-1}$, ($\sim$ denotes the singular asymptotic behaviour). A typical example is the Sierpinski gasket (Fig. 1), which has been widely studied in physics [31]. From Definition 3.5, this structure has dimension $\widetilde{d} = 2\log(3)/\log(5)$. However in [22] it is proven that the asymptotic behaviour of $\widetilde{P}_i(\lambda)$ is more complex since it presents also a small oscillatory part (here $a \ll 1$):

$$\widetilde{P}_i(\lambda) \sim (1-\lambda)^{\widetilde{d}/2-1-N}(1 - a\sin(b\log((1-\lambda)) + c)).$$

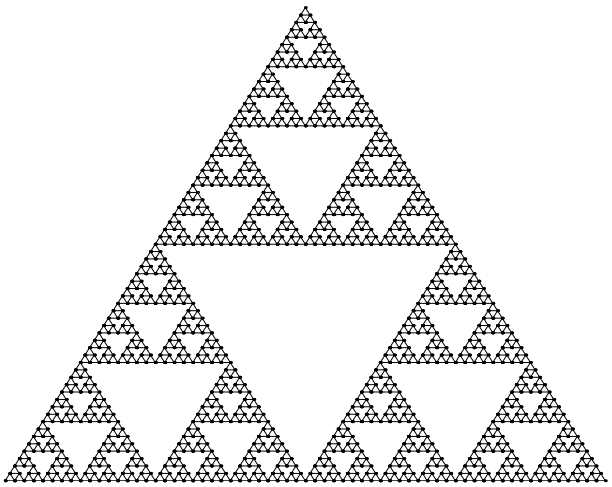


Figure 1. The Sierpinski gasket

Figure 2. The comb graph

Remark 3.8. The existence of the local spectral dimension for any graph recurrent of degree N is an important mathematically open point. Indeed, all the known graphs, for which $\widetilde{d}$ can not be defined (an example is the inhomogeneous Bethe lattice of Fig. 3), are not recurrent of any degree. Moreover, in all the examples we have studied up to now, the local spectral dimension is well-defined for all physical graphs $\mathcal{G}$. Proving that these are indeed sufficient conditions for the existence of $\widetilde{d}$ is another interesting open mathematical problem.

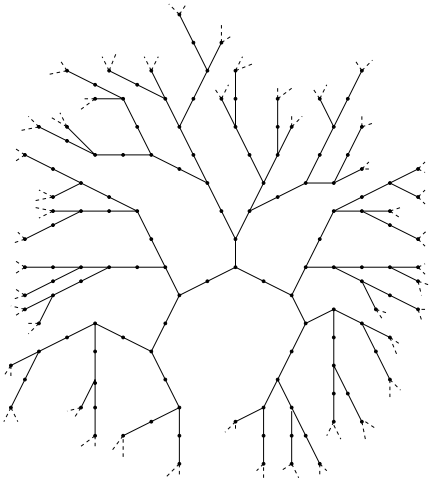


Figure 3. The inhomogeneous Bethe lattice

4 Thermodynamic averages and random walks

4.1 Recurrence and transience on the average

The study of thermodynamic properties of statistical models on infinite graphs requires the introduction of averages of local quantities. The latter are related to random walks by the return probabilities on the average $\overline{P}$ and $\overline{F}$ [13].

Definition 4.1. Given a physical graph $\mathcal{G}$, the return probabilities on the average $\overline{P}$ and $\overline{F}$ are defined by:

$$\overline{P} = \lim_{\lambda \to 1} \overline{\widetilde{P}(\lambda)} \tag{4.29}$$

$$\overline{F} = \lim_{\lambda \to 1} \overline{\widetilde{F}(\lambda)} \tag{4.30}$$

Definition 4.2. $\mathcal{G}$ is called *recurrent on the average* (ROA) if $\overline{F} = 1$, while it is *transient on the average* (TOA) when $\overline{F} < 1$.

Remark 4.3. The main mathematical point in Definitions 4.1 and 4.2 is the existence of the thermodynamic average for the functions $\widetilde{F}_i(\lambda)$ and $\widetilde{P}_i(\lambda)$. The existence of this limit will always be assumed for physical graphs. In [2] an example when $\overline{\widetilde{P}(\lambda)}$ and $\overline{\widetilde{F}(\lambda)}$ are not well-defined is presented. However, the graph of this example does not satisfy **p.c.3**. On the other hand, $\overline{\widetilde{P}(\lambda)}$ and $\overline{\widetilde{F}(\lambda)}$ are well-defined for all physical graphs $\mathcal{G}$ we have studied up to now. A general result in this direction would be an important breakthrough in understanding the average properties of random walks on graphs.

Under the hypothesis of the existence of the thermodynamic averages, the limit $\lambda \to 1^-$ is always well-defined since $\overline{\widetilde{P}(\lambda)}$ and $\overline{\widetilde{F}(\lambda)}$ are increasing functions of λ. Furthermore, the independence of the averages of the center of the spheres is assured by Theorem 2.8 and by the boundedness from below of $\widetilde{F}_i(\lambda)$ and $\widetilde{P}_i(\lambda)$. Hence Definition 4.2 represents a property of the graph.

In [2] a different definition of transience and recurrence on the average is given. There the thermodynamic limit $\lim_{r\to\infty}$ is replaced with $\liminf_{r\to\infty}$ which is always well-defined. Moreover, in [2] the limit $\lambda \to 1^-$ is evaluated before taking the thermodynamic average. This definition leads to another graph classification. For example, the chain of increasing cubes (see Fig. 4) from Definition 4.2 is a TOA graph, whereas it is recurrent on the average according to the definition from [2]. Furthermore, the condition for the limit to be independent of the center of the sphere in this case is weaker than the hypotheses of Section 2.2.

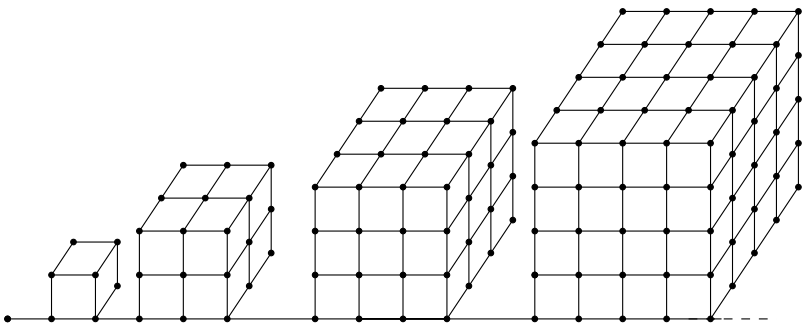


Figure 4. The chain of increasing cubes

Recurrence and transience on the average are in general independent from the corresponding local properties. The first example of this phenomenon occurring on inhomogeneous structures was found in a class of infinite trees called NTD

(Fig. 5) which are locally transient but recurrent on the average [10]. On the other hand, the chain of increasing cubes in Fig. 4 is an example of locally recurrent but transient on the average graph.

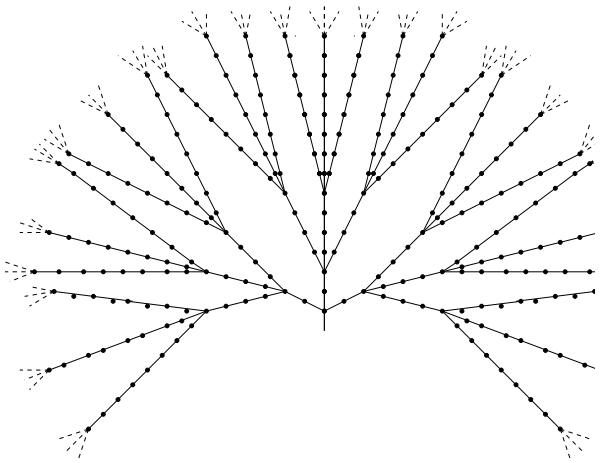


Figure 5. The NTD graph

For (4.30) and (4.29) we cannot prove any simple relation between (4.30) and (4.29) analogous to equation (3.19) for local probabilities. Indeed, averaging (3.19) over all sites i would involve the average of a product, which, due to correlations, is in general different from the product of the averages. Therefore, the equivalence $\widetilde{F}_i(1) = 1 \Leftrightarrow \lim_{\lambda\to 1} \widetilde{P}_i(\lambda) = \infty$ is not true. There are graphs for which $\overline{F} < 1$, but $\overline{P} = \infty$ (an example is shown in Fig. 6), and the study of the relation between $\overline{P}$ and $\overline{F}$ is a non-trivial problem, which will be dealt with in detail in Sections 4.4 and 4.3.

Lemma 4.4. *Let $\mathcal{G}$ be a physical graph such that*

$$\overline{P}(t) = \lim_{r\to\infty} |V_r|^{-1} \sum_{i\in V_r} P_{ii}(t) \quad \text{and} \quad \overline{F}(t) = \lim_{r\to\infty} |V_r|^{-1} \sum_{i\in V_r} F_{ii}(t)$$

are well-defined. For all $\lambda < 1$ we have:

$$\overline{\widetilde{F}(\lambda)} = \sum_{t=0}^{\infty} \lambda^t \overline{F(t)}, \quad \overline{\widetilde{P}(\lambda)} = \sum_{t=0}^{\infty} \lambda^t \overline{P(t)}. \tag{4.31}$$

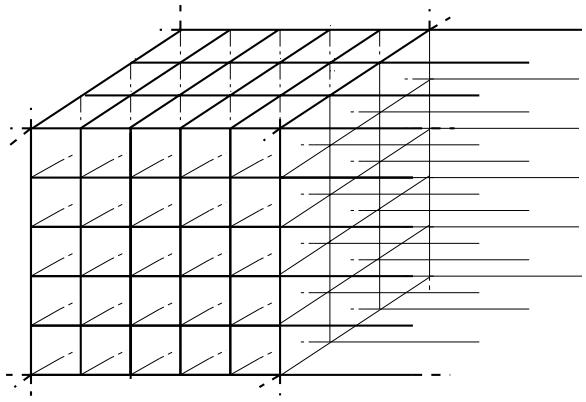


Figure 6. An example of mixed TOA graph

Proof. For all $\lambda < 1$

$$\begin{aligned}\overline{\widetilde{P}(\lambda)} &= \lim_{r\to\infty} \sum_{i\in V_r} |V_r|^{-1} \Big(\sum_{t=0}^{\bar{t}} \lambda^t P_{ii}(t) + \sum_{t=\bar{t}}^{\infty} \lambda^{\bar{t}} P_{ii}(t) \Big) \\ &= \sum_{t=0}^{\bar{t}} \lambda^t \overline{P(t)} + \lim_{r\to\infty} \sum_{i\in V_r} |V_r|^{-1} \sum_{t=\bar{t}}^{\infty} \lambda^t P_{ii}(t).\end{aligned} \tag{4.32}$$

Since $\sum_{i\in S_{o,r}} \|V_r\|^{-1} \sum_{t=\bar{t}}^{\infty} \lambda^{\bar{t}} P_{ii}^{\mathcal{G}'}(t) \le \lambda^{\bar{t}}(1-\lambda)^{-1}$, letting in (4.32) $\bar{t} \to \infty$ we get (4.31). An analogous equation also holds for $F_{ii}(t)$. □

Remark 4.5. In the following we will also assume that $\overline{P(t)}$ and $\overline{F(t)}$ are well-defined on a physical graph. Finding general conditions under which these hypotheses hold is another important mathematical open point in the study of random walks on the average.

Lemma 4.4 shows that the series defining the generating functions and the thermodynamic averages commute. Moreover, from (4.31) and from the Abel lemma we get that $\overline{\widetilde{P}(\lambda)}$ and $\overline{\widetilde{F}(\lambda)}$ are $\mathcal{C}^\infty$ functions in $[0, 1)$. The function $\widetilde{F}_{ij}(\lambda)$ is continuous also at $\lambda = 1$, while $\widetilde{P}_{ij}(\lambda)$ can diverge at this point.

4.2 The average spectral dimension

Since $\overline{\widetilde{P}(\lambda)}$ is $\mathcal{C}^\infty$ in $[0, 1)$, one can define the average spectral dimension $\overline{d}$ in a way analogous to $\widetilde{d}$. Interestingly, the average spectral dimension of real inhomogeneous discrete structures can be experimentally measured [28]. Moreover, $\overline{d}$ has a great influence on the behaviour of the thermodynamic quantities (such as the specific heat) of physical models. Hence $\overline{d}$ is a fundamental quantity in statistical and condensed matter physics.

Definition 4.6. Let $\overline{\widetilde{P}(\lambda)}^{(n)}$ be

$$\overline{\widetilde{P}_i(\lambda)}^{(n)} = \left(\frac{d}{d\lambda}\right)^n \overline{\widetilde{P}(\lambda)}. \tag{4.33}$$

A physical graph $\mathcal{G}$ is recurrent on the average of degree N if

$$\lim_{\lambda\to 1^-} \overline{\widetilde{P}(\lambda)}^{(n)} < \infty, \ \forall n < N \quad \text{and} \quad \lim_{\lambda\to 1^-} \overline{\widetilde{P}(\lambda)}^{(N)} = \infty. \tag{4.34}$$

Definition 4.7. Let $\mathcal{G}$ be recurrent on the average of degree N, and

$$D = \lim_{\lambda\to 1^-} \frac{\log\left(\overline{\widetilde{P}(\lambda)}^{(n)}\right)}{-\log(1-\lambda)}. \tag{4.35}$$

If the limit (3.26) exists, the average spectral dimension is $\overline{d} = 2(N - D + 1)$.

Lemma 4.8. *Let $\mathcal{G}$ be a recurrent of degree N graph with local spectral dimension $\widetilde{d}$. Then $\widetilde{d} \le 2$ if $N = 0$ and $2N \le \widetilde{d} \le 2(N + 1)$ for $N \ge 1$.*

Proof. The proof is completely analogous to that of Lemma 3.7. □

The average spectral dimension has been evaluated for many discrete structures showing that in general, on inhomogeneous graphs, it is different from the local one. We call this phenomenon dynamical dimensional splitting. For example, on the comb graph (Fig. 2) $\widetilde{d} = 1.5$, $\overline{d} = 1$ [19], and on the NTD graph (Fig. 5) $\widetilde{d} = 1 + \log(3)/\log(2)$, $\overline{d} = 1$ [10]. On the other hand, on homogeneous structures, such as the $\mathbb{Z}^d$ lattices, for which all sites are equivalent, we have $\overline{\widetilde{P}(\lambda)} = \widetilde{P}_i(\lambda)$, $\forall i$ and then $\widetilde{d} = \overline{d}$.

Remark 4.9. The behaviour of average quantities $\overline{\widetilde{P}(\lambda)}$ seems to be much more regular than $\widetilde{P}_i(\lambda)$. For example, numerical results for the Sierpinski gasket put into evidence that the oscillations of $\widetilde{P}_i(\lambda)$ [22], which have been described in Section 3.3, disappear in $\overline{\widetilde{P}(\lambda)}$. This is another heuristic result, which requires a rigorous formulation.

Remark 4.10. Even for the average spectral dimension the main open problem from a mathematical point of view is finding general conditions for its existence. As in the

case of $\widetilde{d}$, all the known graphs, for which $\overline{d}$ can not be defined, are not even recurrent on the average of any degree. Furthermore, all these graphs do not satisfy **p.c.2** and **p.c.3**.

4.3 Pure and mixed transience on the average

In this section we study the relation between $\overline{P}$ and $\overline{F}$. This problem, as stated in Section 4.1 is not simple as for the case of local recurrence. In particular we show that a complete picture of the behavior of random walks on graphs can be given by dividing transient on the average graphs into two further classes, which will be called *pure* and *mixed* transient on the average (TOA) [13].

Theorem 4.11. *Let $\mathcal{G}$ be an ROA graph (i.e. $\overline{F} = 1$), then $\overline{P} = \infty$.*

Proof. Since $\overline{F} = 1$, for each $\delta > 0$ it exists ϵ such that $1 - \epsilon \leq \lambda < 1$. Then we have: $1 - \delta \leq \overline{\widetilde{F}(\lambda)} \leq 1$. Let $S = \{ i \in V | \ \widetilde{F}_i(1-\epsilon) < 1 - \sqrt{\delta}\} \subset V$, then

$$\begin{aligned} 1 - \delta \leq \overline{\widetilde{F}(1-\epsilon)} &= \overline{\chi(S)\widetilde{F}(1-\epsilon)} + \overline{\chi(\overline{S})\widetilde{F}(1-\epsilon)} \\ &\leq (1-\sqrt{\delta})\|S\| + \|\overline{S}\| = 1 - \sqrt{\delta}\|S\| \end{aligned} \tag{4.36}$$

(here $\overline{S}$ denotes the complement of S). From (4.36) we get $\|S\| \leq \sqrt{\delta}$, and then $\|\overline{S}\| \geq 1 - \sqrt{\delta}$. Since $\overline{\widetilde{P}(\lambda)}$ is an increasing function of λ, for each $\lambda \geq 1 - \epsilon$ we get:

$$\overline{\widetilde{P}(\lambda)} \geq \overline{\widetilde{P}(1-\epsilon)} \geq \overline{\chi(\overline{S})(1 - \widetilde{F}(1-\epsilon))^{-1}} \geq \|\overline{S}\|\delta^{-1/2} \geq (1-\sqrt{\delta})\delta^{-1/2}. \tag{4.37}$$

In this way we have proved that for an arbitrarily large value of $(1-\sqrt{\delta})\delta^{-1/2}$ (as $\delta \to 0$), it exists ϵ such that for each λ with $1-\epsilon \leq \lambda < 1$ we have $\overline{\widetilde{P}(\lambda)} \geq (1-\sqrt{\delta})\delta^{-1/2}$, and therefore $\overline{P} = \lim_{\lambda\to 1} \overline{\widetilde{P}(\lambda)} = \infty$. □

Theorem 4.11 can be easily generalized

Corollary 4.12. *Let $\mathcal{G}$ be a physical graph with a positive measure subset $V' \subseteq V$ such that $\lim_{\lambda\to 1} \overline{\chi(V')\widetilde{F}(\lambda)} = \|V'\|$. Then*

$$\overline{P} \geq \lim_{\lambda\to 1} \overline{\chi(S')\widetilde{P}(\lambda)} = \infty \quad \forall S' \subseteq V', \|S'\| > 0. \tag{4.38}$$

Hence we proved that $\overline{F} = 1 \Rightarrow \overline{P} = \infty$. Unfortunately, the inverse relation does not hold (an example is given in Fig. 6), and a further classification is needed.

Definition 4.13. We say that a TOA graph is *mixed* if there exists a subset $V' \subset V$ such that $\|V'\| > 0$ and

$$\lim_{\lambda\to 1} \overline{\chi(V')\widetilde{F}(\lambda)} = \|V'\|. \tag{4.39}$$

Definition 4.14. A graph will be called *pure* TOA, if:

$$\lim_{\lambda\to 1} \overline{\chi(V')\widetilde{F}(\lambda)}\|V'\|^{-1} < k < 1 \quad \forall V' \subseteq V,\ \|V\| > 0. \tag{4.40}$$

Remark 4.15. No TOA graph is known which is neither mixed nor pure. In this case $\lim_{\lambda\to 1} \overline{\chi(V')\widetilde{F}(\lambda)}\|V'\|^{-1}$ should be smaller than 1 $\forall V' \subseteq V$ but not smaller of any real number $k < 1$. We will never consider this case and also for this problem a general result is needed.

Theorem 4.16. $\overline{P} < \infty$ *for any pure TOA graph* $\mathcal{G}$.

Proof. For each $0 < \lambda' < 1$ let $S_{\lambda'} \subseteq V$ be defined as $S_{\lambda'} = \{i \in V |\ \widetilde{F}_i(\lambda') > k\}$. Since $\widetilde{F}_i(\lambda)$ is an increasing function, $\forall \lambda > \lambda'$ we get $\overline{\chi(S_{\lambda'})\widetilde{F}_i(\lambda)} > k\|S_{\lambda'}\|$, and then $\lim_{\lambda\to 1} \overline{\chi(S_{\lambda'})\widetilde{F}_i(\lambda)} > k\|S_{\lambda'}\|$. From Definition 4.14, $\|S_{\lambda'}\| = 0$. Then we get:

$$\begin{aligned}\overline{\widetilde{P}(\lambda')} &= \overline{\chi(\overline{S}_{\lambda'})\widetilde{P}(\lambda')} + \overline{\chi(S_{\lambda'})\widetilde{P}(\lambda')} \\ &\leq \overline{\chi(\overline{S}_{\lambda'})(1-\widetilde{F}(\lambda'))^{-1}} + \|S_{\lambda'}\|(1-\lambda')^{-1} \\ &\leq \|\overline{S}_{\lambda'}\|(1-k)^{-1} \leq (1-k)^{-1},\end{aligned} \tag{4.41}$$

where we used Lemma 3.3. Taking the limit $\lambda' \to 1$ in (4.41), we get that for pure TOA graphs $\overline{P}$ is finite. □

Theorem 4.16 can be generalized

Corollary 4.17. *Let* $\mathcal{G}$ *be a physical graph with a subset* $V' \subset V$, $\|V'\| > 0$ *such that for all* $S' \subseteq V'$, $\|S'\| > 0$, $\lim_{\lambda\to 1} \overline{\chi(S')\widetilde{F}(\lambda)} \leq \|S'\|$, *then*

$$\lim_{\lambda\to 1} \overline{\chi(S')\widetilde{P}(\lambda)} < \infty. \quad \forall S' \subseteq V',\ \|S'\| > 0. \tag{4.42}$$

4.4 Separability and statistical independence

Here we prove and discuss an important property characterizing mixed TOA graphs which introduces some simplifications in the study of statistical models on these very inhomogeneous structures. In this case, the graph $\mathcal{G}$ can be always decomposed in a pure TOA subgraph $\mathcal{S}$ and an ROA subgraph $\overline{\mathcal{S}}$ with independent jumping probabilities by cutting out a zero measure set of edges. The separability property implies that the two subgraphs are statistically independent and that their thermodynamic properties can be studied separately. Indeed, the partition functions of magnetic models referring to the two subgraphs factorize [11]. Let us first prove the following lemma.

Lemma 4.18. *The vertices of any mixed TOA graph can be divided into two subsets* $V', \overline{V'} \subset V$ *with* $\|V'\|, \|\overline{V'}\| > 0$ *and* $\|\partial V'\| = \|\partial \overline{V'}\| = 0$, *and such that*

$$\lim_{\lambda\to 1^-} \frac{\overline{\chi(S')\widetilde{F}(\lambda)}}{\|S'\|} < 1 \tag{4.43}$$

for all $S' \subseteq V'$ *with* $\|S'\| > 0$, *and*

$$\lim_{\lambda\to 1^-} \frac{\overline{\chi(S'')\widetilde{F}(\lambda)}}{\|S''\|} = 1 \tag{4.44}$$

for all $S'' \subseteq \overline{V'}$ *with* $\|S''\| > 0$.

Proof. From Definition 4.13 we have that a mixed TOA graph can always be decomposed into two subsets V' and $\overline{V'}$ satisfying (4.43) and (4.44). Now we show that $\|\partial V'\| = \|\partial\overline{V'}\| = 0$. Let us suppose that $\|\partial V'\| > 0$, from the boundedness condition on z_i we have $z_{\max}^{-1}\|\partial V'\| \le \|\partial\overline{V'}\| \le z_{\max}\|\partial V'\|$, and also $\|\partial\overline{V'}\| > 0$. Then from Corollaries 4.12 and 4.17 we get:

$$\lim_{\lambda\to 1} \overline{\chi(\partial V')\widetilde{P}(\lambda)} \le \infty \tag{4.45}$$

and

$$\lim_{\lambda\to 1} \overline{\chi(\partial\overline{V'})\widetilde{P}(\lambda)} = \infty. \tag{4.46}$$

We will show that

$$\lim_{\lambda\to 1^-} \overline{\chi(\partial V')\widetilde{P}(\lambda)} \ge z_{\max}^{-2} \lim_{\lambda\to 1^-} \overline{\chi(\partial\overline{V'})\widetilde{P}(\lambda)}.$$

Then the hypotheses $\|\partial V'\| > 0$ would lead to a contradiction, proving that $\|\partial V'\| = \|\partial\overline{V'}\| = 0$. Let us evaluate $\widetilde{P}_i(\lambda)$ at a site $i \in \partial V'$:

$$\widetilde{P}_i(\lambda) = \sum_t \lambda^t p_{ii}^t = \sum_t \lambda^t \sum_{jk} p_{ik} p_{kj}^{t-2} p_{ji} \ge \sum_t \lambda^t \sum_{j\in\partial\overline{V'}} p_{ij} p_{jj}^{t-2} p_{ji}, \tag{4.47}$$

where in the inequality we do not consider the terms in which $j \ne k$ and $j \notin \partial\overline{V'}$. Exploiting the fact that $p_{ij} \ge 1/z_{\max}$ we get:

$$\widetilde{P}_i(\lambda) \ge \frac{\lambda^2}{z_{\max}^2} \sum_t \lambda^{t-2} \sum_{j\in S_{i,\partial\overline{V'}}} p_{jj}^{t-2} = \frac{\lambda^2}{z_{\max}^2} \sum_{j\in S_{i,\partial\overline{V'}}} \widetilde{P}_j(\lambda), \tag{4.48}$$

where $S_{i,\partial\overline{V'}} = \{j \in \partial\overline{V'} | \exists(i,j) \in E\}$. By averaging over the sites $i \in \partial V'$ we obtain:

$$\begin{aligned} \overline{\chi(\partial V')\widetilde{P}(\lambda)} &\ge \frac{\lambda^2}{z_{\max}^2} \lim_{r\to\infty} \frac{\chi_i(\partial V')}{|V_r|} \sum_{i\in V_r} \sum_{j\in S_{i,\partial\overline{V'}}} \widetilde{P}_j(\lambda) \\ &\ge \frac{\lambda^2}{z_{\max}^2} \lim_{r\to\infty} \frac{\chi_j(\partial\overline{V'})}{|V_r|} \sum_{j\in V_r} \widetilde{P}_j(\lambda) = \frac{\lambda^2}{z_{\max}^2} \overline{\chi(\partial\overline{V'})\widetilde{P}(\lambda)}. \end{aligned} \tag{4.49}$$

Taking the limit $\lambda \to 1$ we have:

$$\lim_{\lambda\to 1} \overline{\chi(\partial V')\widetilde{P}(\lambda)} \geq \frac{1}{z_{\max}^2} \lim_{\lambda\to 1} \overline{\chi(\partial \overline{V'})\widetilde{P}(\lambda)} \tag{4.50}$$

□

Theorem 4.19. *Let $\mathcal{G}$ be a mixed TOA graph. It is always possible to find two subgraphs $\mathcal{G}'$ and $\overline{\mathcal{G}'}$ of $\mathcal{G}$ with the following properties (we denote by $V' \subset V$, $\overline{V'} \subset V$, $E' \subset E$ and $\overline{E'} \subset E$ the positive measure sets of vertices and links defining $\mathcal{G}'$ and $\overline{\mathcal{G}'}$, respectively):*

1. *$\mathcal{G}'$ is pure TOA;*
2. *$\overline{\mathcal{G}'}$ is ROA;*
3. *$\|V'\| > 0$, $\|\overline{V'}\| > 0$;*
4. *$V' \cap \overline{V'} = \emptyset$, $V' \cup \overline{V'} = V$;*
5. *$\|E_{\mathcal{G}',\overline{\mathcal{G}'}}\| = 0$, where $E_{\mathcal{G}',\overline{\mathcal{G}'}} = \{(i,j) \in E |\ (i,j) \notin E', (i,j) \notin \overline{E'}\}$ is the set of links one has to cut for disconnecting the two subgraphs.*

Proof. Let us choose the vertex sets $V' \subset V$ and $\overline{V'} \subset V$ as the two subsets satisfying conditions (4.43) and (4.44) of Lemma 4.18, respectively. We then have $V' \cap \overline{V'} = \emptyset$, $V' \cup \overline{V'} = V$, $\|V'\| > 0$ and $\|\overline{V'}\| > 0$. Defining the sets of edges $E' = \{(i,j) \in E |\ i \in V', j \in V'\}$ and $\underline{E'} = \{(i,j) \in E |\ i \in \overline{V'}, j \in \overline{V'}\}$, one obtains $E_{\mathcal{G}',\overline{\mathcal{G}'}} = \{(i,j) \in E |\ i \in V', j \in \overline{V'}\}$, then from the boundedness of z_i, $\|E_{\mathcal{G}',\overline{\mathcal{G}'}}\| \leq z_{\max}\|\partial V'\| = 0$. Now we have to prove that $\mathcal{G}'$ is a pure TOA graph and $\overline{\mathcal{G}'}$ is an ROA graph with respect to their own transition probabilities $p_{ij}^{\mathcal{G}'}$ and $p_{ij}^{\overline{\mathcal{G}'}}$. We denote the average and the measure evaluated in $\mathcal{G}'$ by adding the superscript $\mathcal{G}'$. Then the following simple relation holds:

$$\overline{\phi'}^{\mathcal{G}'} = \|V'\|^{-1}\overline{\chi(V')\phi}, \tag{4.51}$$

where ϕ is any extension to V of the function $\phi' : V' \to \mathbb{R}$; in particular one has $\|\cdot\|^{\mathcal{G}'} = \|V'\|^{-1}\|\cdot\|$. Putting $\widetilde{V}'_{\partial V',t} = \{i \in V' | d(i,\partial V') \leq t\}$ (Definition 2.2) we get from the boundedness of the coordination number:

$$\|\widetilde{V}'_{\partial V',t}\|^{\mathcal{G}'} = \|V'\|^{-1}\|\widetilde{V}'_{\partial V',t}\| < \|V'\|^{-1}(z_{\max})^t\|\partial V'\| = 0;, \tag{4.52}$$

because $\|\partial V'\| = 0$ (Lemma 4.18). Let S' be any subset of V'. For the average of $F_{ii}^{\mathcal{G}'}(t)$ we have ($\overline{\widetilde{V}}'_{\partial V',t}$ is the complement of $\widetilde{V}'_{\partial V',t}$ in V'):

$$\begin{aligned}\overline{\chi(S')F^{\mathcal{G}'}(t)}^{\mathcal{G}'} &= \overline{\chi(\overline{\widetilde{V}}'_{\partial V',t})\chi(S')F^{\mathcal{G}'}(t)}^{\mathcal{G}'} + \overline{\chi(\widetilde{V}'_{\partial V',t})\chi(S')F^{\mathcal{G}'}(t)}^{\mathcal{G}'} \\ &= \overline{\chi(\overline{\widetilde{V}}'_{\partial V',t})\chi(S')F^{\mathcal{G}'}(t)}^{\mathcal{G}'}.\end{aligned} \tag{4.53}$$

Since $F_{ii}^{\mathcal{G}'}(t) = F_{ii}(t)$ on $\overline{\widetilde{V}}'_{\partial V',t}$, we get

$$\begin{aligned}\overline{\chi(S')F^{\mathcal{G}'}(t)}^{\mathcal{G}'} &= \overline{\chi(\overline{\widetilde{V}}'_{\partial V',t})\chi(S')F(t)}^{\mathcal{G}'} \\ &= \overline{\chi(S')F(t)}^{\mathcal{G}'} = \|V'\|^{-1}\overline{\chi(S')F(t)},\end{aligned} \tag{4.54}$$

where in the second equality we used $\|\overline{\widetilde{V}}'_{\partial V',t}\|^{\mathcal{G}'} = 1$, while in the last one we used (4.51) and the fact that $\chi(S')\chi(V') = \chi(S')$, since $S' \subseteq V'$. From Lemma 4.4 we get $\forall S' \subseteq V'$ and $\forall \lambda < 1$:

$$\begin{aligned}&\left(\|S'\|^{\mathcal{G}'}\right)^{-1}\overline{\chi(S')\widetilde{F}^{\mathcal{G}'}(\lambda)}^{\mathcal{G}'} = \left(\|S'\|\|V'\|^{-1}\right)^{-1}\sum_{t=0}^{\infty}\lambda^t\overline{\chi(S')F^{\mathcal{G}'}(t)}^{\mathcal{G}'} \\ &= \left(\|S'\|\|V'\|^{-1}\right)^{-1}\sum_{t=0}^{\infty}\lambda^t\|V'\|^{-1}\overline{\chi(S')F(t)} = \|S'\|^{-1}\overline{\chi(S')\widetilde{F}(\lambda)}.\end{aligned} \tag{4.55}$$

Taking the limit $\lambda \to 1^-$ from (4.43) we obtain that $\mathcal{G}'$ is a pure TOA graph. In an analogous way one can prove that $\overline{\mathcal{G}'}$ is an ROA graph. □

Remark 4.20. Properties similar to the separability discussed in this section can also be found when considering the average spectral dimension instead of the simple recurrence or transience. This leads to the introduction of the so called *spectral classes* and *spectral subclasses*, which have been studied in details in the physical literature in connection with the problem of critical phenomena on graphs [8, 9]. We refer the reader to [8] for details, where the existence of the spectral dimension for classes and subclasses, as usual in physical papers, is implicitly assumed.

5 Harmonic oscillations

5.1 The physical model

Let us consider a countable system of particles $i \in \mathbb{N}$ interacting pairwise with a harmonic potential. In the simplest case in which all particles have the same mass m and their position can be described by a scalar $x_i(t) \in \mathbb{R}$ ($t \in \mathbb{R}$ is the time), we have that the motion equations for the system are:

$$m\frac{d^2x_i(t)}{dt^2} = \sum_j J_{ji}(x_j(t) - x_i(t)) = -\sum_j L_{ji}x_j(t), \tag{5.56}$$

where J_{ji} (Definition 1.5) represent the elastic constants describing the interaction between particles i and j. When all particles interact with the same strength k, we

get:

$$m\frac{d^2x_i(t)}{dt^2} = \sum_j kA_{ji}(x_j(t) - x_i(t)) = -\sum_j k\Delta_{ji}x_j(t). \tag{5.57}$$

Equation (5.57) has been widely studied in physics to describe the elastic and thermal properties of solids. Solving (5.56) and (5.57) can be reduced to an eigenvalue problem by standard differential equation techniques. Denoting by $\widetilde{x}_i(\omega)$ the Fourier transform of $x_i(t)$ with respect to time t, (5.57) becomes:

$$\omega^2\widetilde{x}_i(\omega) = \frac{k}{m}\sum_j \Delta_{ji}\widetilde{x}_j(\omega) \tag{5.58}$$

Hence the study of the spectral properties of the Laplacian matrix Δ_{ij} plays a fundamental role in understanding the physical properties of harmonic oscillations.

5.2 The spectrum of the Laplacian

For the infinite graph X, the Laplacian Δ can be considered as a linear operator on the Hilbert space $l^2(V_X)$. From this point of view many rigorous results have been proven (see [26] for a review). An important theorem, giving a bound on the harmonic frequencies, is:

Theorem 5.1. *Let* $\mathrm{spec}(\Delta)$ *be the spectrum of the operator* $\Delta : l^2(V_X) \to l^2(V_X)$ *on a graph X with bounded connectivity (satisfying* **p.c.2***). Then* $\mathrm{spec}(\Delta) \subseteq [0, 2z_{\max}]$.

From a physical point of view, it is more useful to explore the graph using the Van Hove spheres $\mathcal{S}_{o,r}$. In particular, we are interested in the study of the low frequency (infrared) spectrum of the Laplacian, since many properties, as low temperature vibrational specific heat, are strictly related to the behaviour of this spectral region.

Definition 5.2. Given a physical graph $\mathcal{G}$, let $\Delta_{ij}^{o,r}$ be the Laplacian matrix relative to $\mathcal{S}_{o,r}$, and let $N_{o,r}(\epsilon)$ be the number of eigenstates of $\Delta_{ij}^{o,r}$ in the interval $[0, \epsilon]$. We define the integrated density of states as:

$$n(\epsilon) = \lim_{r\to\infty} n_{o,r}(\epsilon) = \lim_{r\to\infty} |V_{o,r}|^{-1} N_{o,r}(\epsilon). \tag{5.59}$$

Remark 5.3. The general conditions for the existence and the independence from o of limit (5.59) is another interesting mathematical open problem. In physics the existence and the independence are always assumed.

Definition 5.4. Given a physical graph $\mathcal{G}$, we say that $n(\epsilon)$ has a polynomial infrared behaviour if $\exists c_1, c_2, d_\omega \in \mathbb{R}^+$ such that

$$c_1\epsilon^{d_\omega/2} \le n(\epsilon) \le c_2\epsilon^{d_\omega/2}. \tag{5.60}$$

Remark 5.5. Heuristic results put into evidence that if d_ω is well-defined, then $\overline{d}$ is also well-defined and $d_\omega = \overline{d}$. This point also needs a rigorous mathematical formulation.

6 The Gaussian model

6.1 The model

A simple statistical model describing the average properties of harmonic oscillators is given by the Gaussian model. We first define it for a finite graph and then study the behaviour on an infinite structure using the Van Hove spheres.

Definition 6.1. Given a finite graph X in which each site represents a particle i, let $x_i \in \mathbb{R}$ be the displacement from the particle equilibrium position, and let $k_i \in \mathbb{R}$ ($k_i > 0$) and J_{ij} (Definition 1.5) represent the elastic constants describing the recoil force towards the equilibrium position and the interaction with the nearest neighbour particle j, respectively. The Hamiltonian of the system is

$$H = \frac{1}{4}\sum_{i,j\in X} J_{ij}(x_i - x_j)^2 + \frac{1}{2}\sum_{i\in X} k_i x_i^2 = \frac{1}{2}\sum_{i,j\in X}(L_{ij} + K_{ij})x_i x_j, \qquad (6.61)$$

where $K_{ij} = k_i\delta_{ij}$, and $L_{ij} + K_{ij}$ is called the Hamiltonian matrix. Given a function $f = f(x_1, x_2, \ldots, x_{|V_X|}) : \mathbb{R}^{|V_X|} \to \mathbb{R}$ of the displacements x_i, we define the Boltzmann average of f as

$$\langle f\rangle_X(J, K) = \int f d\mu_X(x), \qquad (6.62)$$

where

$$d\mu(x) = Z^{-1}e^{-H}\prod_{i\in X} dx_i \quad \text{and} \quad Z = \int e^{-H}\prod_{i\in X} dx_i. \qquad (6.63)$$

We denote the Boltzmann averages simply by $\langle f\rangle_X$ (dropping (J, K)), when it is not necessary to evidence the dependence on some specific couplings. Since the Hamiltonian matrix is a positive defined operator, the Boltzmann average is well-defined for all continuous bounded functions. For infinite graphs, we denote by K_{ij} the local coupling matrix $K_{ij} = k_i\delta_{ij}$, $k_i \in \mathbb{R}^+$ ($k_{\max} > k_i > k_{\min}, \forall i \in V, k_{\max}, k_{\min} \in \mathbb{R}^+$).

Definition 6.2. Given a physical graph $\mathcal{G}$, a ferromagnetic coupling matrix J_{ij} and a local coupling matrix K_{ij}, let $\mathcal{S}_r$ be a sequence of Van Hove spheres, and let J^r_{ij} and K^r_{ij} be the matrices on $\mathbb{R}^{|V_r|}$ defined as $J^r_{ij} = J_{ij}$, $K^r_{ij} = K_{ij}$ if $i, j \in V_r$, and $J^r_{ij} = K^r_{ij} = 0$ otherwise. If $f_r : \mathbb{R}^{|V_r|} \to \mathbb{R}$ is a function of the displacements x_i

$i \in V_r$, we denote by $\langle f_r \rangle_r$ the Boltzmann average of f_r (6.62) defined in $\mathcal{S}_r$ by J^r_{ij} and by the constants k_i. The Boltzmann average of f on the graph $\mathcal{G}$ is then

$$\langle f \rangle = \lim_{r\to\infty} \langle f_r \rangle_r. \tag{6.64}$$

In physics the most interesting function f_r is the two point correlation function $x_i x_j$ representing the response of the system in the site j to an excitation in the site i, and the square average displacement:

$$\overline{x^2}^r = |V_r|^{-1} \sum_{i\in V_r} x_i^2 \tag{6.65}$$

Another interesting quantity is the many points correlation function $x_{i_1} x_{i_2} \dots x_{i_n}$. However, for the Gaussian model the average of these functions can be evaluated as:

$$\langle x_{i_1} x_{i_2} \dots x_{i_n} \rangle_r = \begin{cases} 0, & \text{if } n \text{ is odd,} \\ \frac{1}{n!!} \sum_{k_1 \dots k_n = 1}^{n} p^{k_1 \dots k_n} \langle x_{i_{k_1}} x_{i_{k_2}} \rangle_r \dots \langle x_{i_{k_{n-1}}} x_{i_{k_n}} \rangle_r, & \text{if } n \text{ is even,} \end{cases} \tag{6.66}$$

where $p^{k_1 \dots k_n}$ is the tensor of index permutations. Hence $\langle x_{i_1} x_{i_2} \dots x_{i_n} \rangle_r$ can be reduced to the two points correlation function.

Remark 6.3. For Definition 6.2 the main open problem also regards the conditions on the graph $\mathcal{G}$ and on the functions f_r guaranteeing the existence of the limit as $r \to \infty$. In the following we will prove the existence of the limit for the two points correlation functions, and then use (6.66) for $\langle x_{i_1} x_{i_2} \dots x_{i_n} \rangle_r$. On the other hand, a proof for the square average displacement is still lacking. For $\langle x_{i_1} x_{i_2} \dots x_{i_n} \rangle_r$ and $\langle \overline{x^2} \rangle^r$ we will prove also the independence from the center of the spheres o.

Let us introduce an alternative definition for the Gaussian model [24].

Definition 6.4. Given a graph X, a ferromagnetic coupling matrix J_{ij}, and a local coupling matrix K_{ij}, there exists a unique Gaussian probability measure $d\mu_g(x)$ on $l^\infty(V)$ with mean zero and covariance $(L + K)^{-1}$, see [24]. The measure $d\mu_g(x)$ characterizes the Gaussian model, and we will use the notations

$$\langle f(x) \rangle_g = \int f(x) d\mu_g(x), \tag{6.67}$$

and, in particular,

$$\langle x_i x_j \rangle_g = (L + K)^{-1}_{ij}. \tag{6.68}$$

In [24] some interesting properties of the linear operator $(L + K)^{-1}$ on $l^\infty(V)$ are obtained. These properties trivially hold also for the operator $(L^r + K^r)^{-1}$ on $\mathbb{R}^{|V_r|}$, where $L^r = I^r_i \delta_{ij} - J^r_{ij}$, $I^r_i = \sum_j J^r_{ij}$. In particular,

Lemma 6.5. *The operators $(L+K)^{-1}$ on $l^\infty(V)$ and $(L^r+K^r)^{-1}$ on $\mathbb{R}^{|V_r|}$ are positive bounded, and*

$$\|(L+K)^{-1}\| \le k_{\min}^{-1}, \quad \|(L^r+K^r)^{-1}\| \le k_{\min}^{-1}. \tag{6.69}$$

6.2 The walk expansion

Let us first consider the two points correlation functions.

Lemma 6.6. *Let J_{ij} be a ferromagnetic coupling matrix, and K_{ij} be a local coupling matrix on a graph X. The two points correlation function can be evaluated in the following way:*

$$\langle x_i x_j \rangle_g = \left(\left(1 + \frac{z_{\max} J_{\max}}{k_{\min}} \right) k_j \right)^{-1} \widetilde{P}_{ij}^J \left(\left(1 + \frac{k_{\min}}{z_{\max} J_{\max}} \right)^{-1} \right), \tag{6.70}$$

where $\widetilde{P}_{ij}^J(\lambda)$ is the generating function of the random walk defined by the jumping probability p_{ij}^I:

$$p_{ij}^I = \frac{k_{\min}}{z_{\max} J_{\max} k_i} J_{ij} + \left(1 - I_j \frac{k_{\min}}{z_{\max} J_{\max} k_i} \right) \delta_{ij}. \tag{6.71}$$

Proof. Let us represent $\langle x_i x_j \rangle_g$ by a walk expansion:

$$\begin{aligned}
\langle x_i x_j \rangle_g &= (I - J + K)_{ij}^{-1} \\
&= \left(\left(1 + \frac{z_{\max} J_{\max}}{k_{\min}} \right) K \left(\delta - \frac{J + \frac{z_{\max} J_{\max}}{k_{\min}} K - I}{\left(1 + \frac{k_{\min}}{z_{\max} J_{\max}} \right) \frac{z_{\max} J_{\max}}{k_{\min}} K} \right) \right)_{ij}^{-1} \\
&= \left(\left(1 + \frac{z_{\max} J_{\max}}{k_{\min}} \right) k_j \right)^{-1} \left(\delta - \frac{J + \frac{z_{\max} J_{\max}}{k_{\min}} K - I}{\left(1 + \frac{k_{\min}}{z_{\max} J_{\max}} \right) \frac{z_{\max} J_{\max}}{k_{\min}} K} \right)_{ij}^{-1} \\
&= \left(\left(1 + \frac{z_{\max} J_{\max}}{k_{\min}} \right) k_j \right)^{-1} \sum_{t=0}^{\infty} \left(1 + \frac{k_{\min}}{z_{\max} J_{\max}} \right)^{-t} P_{ij}^I(t),
\end{aligned} \tag{6.72}$$

where $P_{ij}^I(t) = (p^I)_{ij}^t$. Notice that p_{ij}^I is a well-defined transition probability with $0 \le p_{ij}^I \le 1$ and $\sum_j p_{ij}^I = 1$, then also $0 \le P_{ij}^I(t) \le 1$. From (6.72) one immediately gets (6.70). □

Corollary 6.7. *The two points correlation function $\langle x_i x_j \rangle_g$ satisfies the inequality*

$$0 \leq \langle x_i x_j \rangle_g \leq k_j^{-1}. \tag{6.73}$$

Proof. Equation (6.73) is a simple consequence of Lemma 6.6 and Lemma 3.3. □

Theorem 6.8. *Let $\mathcal{G}$ be a physical graph , J_{ij} be a ferromagnetic coupling matrix and K_{ij} be a local coupling matrix. For the two points correlation function we have:*

$$\lim_{r\to\infty} \langle x_i x_j \rangle_r = \langle x_i x_j \rangle_g. \tag{6.74}$$

Proof. For the correlation function $\langle x_i x_j \rangle_r$ in analogy to (6.72) we get the equation

$$\begin{aligned} \langle x_i x_j \rangle_r &= (L^r + K^r)^{-1}_{ij} \\ &= \left(\left(1 + \frac{z_{\max} J_{\max}}{k_{\min}} \right) k_j \right)^{-1} \sum_{t=0}^{\infty} \left(1 + \frac{k_{\min}}{z_{\max} J_{\max}} \right)^{-t} P_{ij}^{I_r}(t), \end{aligned} \tag{6.75}$$

where $P_{ij}^{I_r}(t)$ is defined by the jumping probability $p_{ij}^{I_r}$ in $\mathcal{S}_{o,r}$:

$$p_{ij}^{I_r} = \frac{k_{\min}}{z_{\max} J_{\max} k_i} J_{ij}^r + \left(1 - I_j^r \frac{k_{\min}}{z_{\max} J_{\max} k_i} \right) \delta_{ij}. \tag{6.76}$$

Let us now choose spheres of radius $r(T) = r_{i,o} + T + 1$ ($T \in \mathbb{N}$), so that we can get the thermodynamic limit letting $T \to \infty$. We have:

$$\begin{aligned} \langle x_i x_j \rangle_{r(T)} = &\left(\left(1 + \frac{z_{\max} J_{\max}}{k_{\min}} \right) k_j \right)^{-1} \\ &\times \left(\sum_{t=0}^{T} \left(1 + \frac{k_{\min}}{z_{\max} J_{\max}} \right)^{-t} P_{ij}^{I}(t) \right. \\ &\left. + \sum_{t=T+1}^{\infty} \left(1 + \frac{k_{\min}}{z_{\max} J_{\max}} \right)^{-t} P_{ij}^{I_{r(T)}}(t) \right), \end{aligned} \tag{6.77}$$

where we used the property that $P_{ij}^{I_{r(T)}}(t) = P_{ij}^{I}(t)$ in $\mathcal{S}_{o,r(T)}$ for walks starting from i and of length smaller than $T+1$. Let us show that the second term in (6.77) goes to zero if $T \to \infty$, i.e., in the thermodynamic limit. Since $0 \leq P_{ij}^{I_{r(T)}}(t) \leq 1$, one gets:

$$\begin{aligned} 0 \leq &\left(\left(1 + \frac{z_{\max} J_{\max}}{k_{\min}} \right) k_j \right)^{-1} \sum_{t=T+1}^{\infty} \left(1 + \frac{k_{\min}}{z_{\max} J_{\max}} \right)^{-t} P_{ij}^{I_{r(T)}}(t) \\ &\leq k_i^{-1} \left(1 + \frac{k_{\min}}{z_{\max} J_{\max}} \right)^{-T} \longrightarrow 0 \ \text{ for } \ T \to \infty. \end{aligned} \tag{6.78}$$

Letting $T \to \infty$ in (6.77), from (6.78) we have

$$\begin{aligned}\langle x_i x_j \rangle &= \lim_{T\to\infty} \langle x_i x_j \rangle_{r(T)} \\ &= \left(\left(1 + \frac{z_{\max} J_{\max}}{k_{\min}} \right) k_j \right)^{-1} \sum_{t=0}^{\infty} \left(1 + \frac{k_{\min}}{z_{\max} J_{\max}} \right)^{-t} P_{ij}^I(t) \qquad (6.79) \\ &= \langle x_i x_j \rangle_g . \end{aligned}$$

□

Corollary 6.9. *$\langle x_i x_j \rangle = \lim_{r\to\infty} \langle x_i x_j \rangle_r$ exists and is independent of the center of the spheres o, for all graphs X.*

Proof. Since $\langle x_i x_j \rangle_g$ is well-defined and independent of o, one immediately gets the claim from equation (6.79). □

When we deal with the thermodynamic limit we usually restrict ourselves to a physical graph. However, Definition 6.4 and Theorem 6.8 hold even for a graph which does not satisfy **p.c.1**, **p.c.2** and **p.c.3**. Hence the existence of $\langle x_i x_j \rangle$ and its the independence from o is proven for all graphs and not only for physical ones.

Corollary 6.10. *The many points correlation function*

$$\langle x_{i_1} x_{i_2} \dots x_{i_n} \rangle = \lim_{r\to\infty} \langle x_{i_1} \dots x_{i_n} \rangle$$

exists and is independent of the center of the spheres o, for all graphs X. Moreover, $\langle x_{i_1} x_{i_2} \dots x_{i_n} \rangle = \langle x_{i_1} x_{i_2} \dots x_{i_n} \rangle_g$.

Proof. Equations (6.66) hold for the averages $\langle . \rangle_r$ on any finite sphere and for the average $\langle . \rangle_g$ of Definition 6.4, hence from (6.79) we have $\langle x_{i_1} x_{i_2} \dots x_{i_n} \rangle = \langle x_{i_1} x_{i_2} \dots x_{i_n} \rangle_g$. Moreover, since $\langle x_{i_1} x_{i_2} \dots x_{i_n} \rangle_g$ are always well-defined and independent from o, one immediately completes the proof. □

Let us now pass to the study of the average displacement.

Theorem 6.11. *Let $\mathcal{G}$ be a physical graph , let J_{ij} be a ferromagnetic coupling matrix, and let K_{ij} be a local coupling matrix. For the average displacement we have*

$$\overline{\langle x^2 \rangle} = \lim_{r\to\infty} \overline{\langle x^2 \rangle_r} = \overline{\langle x^2 \rangle_g}. \qquad (6.80)$$

Proof. In the hypothesis of the existence of $\overline{P^I(t)} = \lim_{r\to\infty} |V_r|^{-1} \sum_i P_{ii}^I(t)$ from (6.72) and Lemma 4.4 one has

$$\overline{\langle x^2 \rangle_g} = \left(\left(1 + \frac{z_{\max} J_{\max}}{k_{\min}} \right) k_j \right)^{-1} \sum_{t=0}^{\infty} \left(1 + \frac{k_{\min}}{z_{\max} J_{\max}} \right)^{-t} \overline{P^I(t)}. \qquad (6.81)$$

Given a sequence of spheres $\mathcal{S}_r$, let $\widetilde{V}_{\partial V_r, t} = \{ i \in V | d(i, \partial V_r) \le t \}$ (Definition 2.2) and $\widetilde{\widetilde{V}}_{\partial V_r, t}$ its complement, then from **p.c.2** we get

$$|\widetilde{V}_{\partial V_r, t}| \le z_{\max}^T |\partial V_r|, \qquad (6.82)$$

and from **p.c.3**

$$\lim_{r\to\infty}|V_r|^{-1}\sum_{i\in V_r}\chi_i(\widetilde{V}\partial V_r,t)=0,\quad \lim_{r\to\infty}|V_r|^{-1}\sum_{i\in V_r}\chi_i(\overline{\widetilde{V}}_{\partial V_r,t})=1. \tag{6.83}$$

The walk expansion (6.75) gives:

$$\begin{aligned}\lim_{r\to\infty}\langle\overline{x^2}^r\rangle_r = C_j\lim_{r\to\infty}\frac{1}{|V_r|}\sum_{i\in V_r}\Bigg(&\sum_{t=0}^{T}\left(1+\frac{k_{\min}}{z_{\max}J_{\max}}\right)^{-t}\chi_i(\overline{\widetilde{V}}_{\partial V_r,T})P_{ii}^{I}(t)\\ &+\sum_{t=0}^{T}\left(1+\frac{k_{\min}}{z_{\max}J_{\max}}\right)^{-t}\chi_i(\widetilde{V}_{\partial V_r,T})P_{ii}^{I_r}(t)\\ &+\sum_{t=T+1}^{\infty}\left(1+\frac{k_{\min}}{z_{\max}J_{\max}}\right)^{-t}P_{ii}^{I_r}(t)\Bigg),\end{aligned} \tag{6.84}$$

where we used the property that $P_{ii}^{I_r}(t)=P_{ii}^{I}(t)$ on $\overline{\widetilde{V}}_{\partial V_r,T}$ for $t\le T$ and the notation $C_j=(1+z_{\max}J_{\max}/k_{\min})^{-1}k_j^{-1}$. Let us now evaluate the thermodynamic limit. From (6.83) and the boundedness of $P_{ii}^{I_r}(t)$ and $P_{ii}^{I}(t)$ we have

$$\begin{aligned}\lim_{r\to\infty}\langle\overline{x^2}^r\rangle_r = C_j\Bigg(&\sum_{t=0}^{T}\left(1+\frac{k_{\min}}{z_{\max}J_{\max}}\right)^{-t}\overline{P^I(t)}\\ &+\lim_{r\to\infty}|V_r|^{-1}\sum_{i\in V_r}\sum_{t=T+1}^{\infty}\left(1+\frac{k_{\min}}{z_{\max}J_{\max}}\right)^{-t}P_{ii}^{I_r}(t)\Bigg).\end{aligned} \tag{6.85}$$

Since $0\le P_{ij}^{I_{r(T)}}(t)\le 1$, one gets:

$$\begin{aligned}0&\le C_j\lim_{r\to\infty}|V_r|^{-1}\sum_{i\in V_r}\sum_{t=T+1}^{\infty}\left(1+\frac{k_{\min}}{z_{\max}J_{\max}}\right)^{-t}P_{ij}^{I_{r(T)}}(t)\\ &\le k_i^{-1}\left(1+\frac{k_{\min}}{z_{\max}J_{\max}}\right)^{-T}\longrightarrow 0\ \text{ for }\ T\to\infty.\end{aligned} \tag{6.86}$$

Letting $T\to\infty$ in (6.85) from (6.86) we have:

$$\lim_{r\to\infty}\langle\overline{x^2}^r\rangle_r = C_j\sum_{t=0}^{\infty}\left(1+\frac{k_{\min}}{z_{\max}J_{\max}}\right)^{-t}\overline{P^I(t)}=\overline{\langle x^2\rangle_g}. \tag{6.87}$$

□

In the proof of Theorem 6.11 the hypotheses on the graph of satisfying **p.c.1**, **p.c.2**, **p.c.3** and on the existence of the thermodynamic limit are necessary, for example in the inhomogeneous Bethe lattice (Fig. 3) Definitions 6.2 and 6.4 are not equivalent.

Under these hypotheses equation (6.87) proves the independence of $\overline{\langle x^2 \rangle}$ from the choice of the center of the sphere, since $\overline{\langle x^2 \rangle}$ is reduced to the evaluation of averages of positive (bounded from below) functions.

6.3 Gaussian model and spectral dimensions

Lemma 6.6 puts into evidence the deep relation between the Gaussian model and random walks. An even clearer connection can be obtained if we consider the model defined by the Hamiltonian matrix $\Delta + mZ$ ($m \in \mathbb{R}^+$).

Theorem 6.12. *For the two points correlation function*

$$\langle x_i x_j \rangle(A, mZ) = \langle x_i x_j \rangle_g(\Delta, mZ) = z_j^{-1} \widetilde{P}_{ij} \left((1+m)^{-1} \right). \tag{6.88}$$

Proof. The first equation is a simple consequence of Theorem 6.8, while for the second one we have:

$$\begin{aligned} \langle x_i x_j \rangle_g(A, mZ) &= (Z - A + mZ)^{-1}_{ij} = z_j^{-1} (\delta - (1+m)^{-1} Z^{-1} A)^{-1}_{ij} \\ &= z_j^{-1} \sum_{t=0}^{\infty} (1+m)^{-t} P_{ij}(t). \end{aligned} \tag{6.89}$$

□

Equation (6.89) proves that $\langle x_i x_j \rangle_g(A, mZ)$ as a function of m is $\mathcal{C}^\infty$. Theorem 6.12 allows one to recast many graph properties which had been described using random walks in terms of correlation function of the Gaussian model.

Corollary 6.13. *A graph X is locally recurrent if and only if*

$$\lim_{m \to 0^+} \langle x_i x_i \rangle(A, mZ) = \infty.$$

Corollary 6.14. *A graph X is locally recurrent of degree N if and only if*

$$\lim_{m \to 0^+} \langle x_i x_i \rangle^{(n)}(A, mZ) < \infty, \ \forall n < N \quad \text{and} \quad \lim_{m \to 0^+} \langle x_i x_i \rangle^{(N)}(A, mZ) = \infty, \tag{6.90}$$

where $\langle x_i x_i \rangle^{(n)}(A, mZ) = (-1)^n (d^n/dm^n) \langle x_i x_i \rangle(A, mZ)$.

Corollary 6.15. *Let X be a graph recurrent of degree N with local spectral dimension $\widetilde{d}$. Then $\widetilde{d} = 2(N - D + 1)$, where*

$$D = \lim_{m \to 0^+} \frac{\log(\langle x_i x_i \rangle^{(N)}(A, mZ))}{-\log(m)}. \tag{6.91}$$

Proof. Corollaries 6.13, 6.14 and 6.15 are simple consequences of Theorem 6.12 and Definitions 3.4, 3.5 and 3.6. □

In [24] important universality properties are proven. Namely, it is shown that Corollaries 6.13, 6.14 and 6.15 hold even if one replaces Δ and Z with their generalizations L^r and K^r.

Let us now consider the behaviour of the average displacement.

Theorem 6.16. *Let $\mathcal{G}$ a physical graph. Then the average displacement satisfies the following equation*

$$\overline{\langle x^2\rangle}(A, mZ) = \overline{\langle x^2\rangle}(A, mZ) = \overline{z^{-1}\widetilde{P}\left((1+m)^{-1}\right)}. \tag{6.92}$$

Proof. Equation (6.92) is a consequence of Theorems 6.11 and 6.12. □

Corollary 6.17. *A physical graph $\mathcal{G}$ is recurrent on the average of degree N if and only if*

$$\lim_{m\to 0^+} \overline{\langle x^2\rangle}^{(n)}(A, mZ) < \infty, \ \forall n < N \quad \text{and} \quad \lim_{m\to 0^+} \overline{\langle x^2\rangle}^{(N)}(A, mZ) = \infty, \tag{6.93}$$

where $\overline{\langle x^2\rangle}^{(n)}(A, mZ) = (-1)^n (d^n/dm^n)\overline{\langle x^2\rangle}(A, mZ)$.

Corollary 6.18. *Let $\mathcal{G}$ be a physical graph recurrent on the average of degree N with average spectral dimension $\bar{d}$. Then $\bar{d} = 2(N - D + 1)$, where*

$$D = \lim_{m\to 0^+} \frac{\log(\overline{\langle x^2\rangle}^{(N)}(A, mZ))}{-\log(m)}. \tag{6.94}$$

Proof Corollaries 6.17 and 6.18 are simple consequences of Theorem 6.16, Definitions 4.6 and 4.7 and condition **p.c.2**, because in this case

$$z_{\max}^{-1}\overline{\widetilde{P}\left((1+m)^{-1}\right)} < \overline{z^{-1}\widetilde{P}\left((1+m)^{-1}\right)} < \overline{\widetilde{P}\left((1+m)^{-1}\right)}. \qquad \square$$

6.4 Universality properties of $\bar{d}$

In this section we will prove some important universality properties of the average spectral dimension, which have been stated in [6, 7]. Let us begin with

Lemma 6.19. *Let $\mathcal{G}$ be a physical graph. Then:*

$$\left(\frac{\partial^n}{\partial m^n}\right) \lim_{r\to\infty} \overline{\langle x^{2^r}\rangle}_r(J, mK) = \lim_{r\to\infty} \left(\frac{\partial^n}{\partial m^n}\right) \overline{\langle x^{2^r}\rangle}_r(J, mK). \tag{6.95}$$

Proof. From (6.75) one has:

$$\left(\frac{\partial^n}{\partial m^n}\right) \lim_{r\to\infty} \langle \overline{x^2}^r \rangle_r (J, mK)$$
$$= \left(\left(1 + \frac{z_{\max} J_{\max}}{k_{\min}}\right) k_j\right)^{-1} \left(\frac{\partial^n}{\partial m^n}\right) \lim_{r\to\infty} |V_r|^{-1}$$
$$\sum_{i\in V_r} \left(\sum_{t=0}^{T} \left(1 + \frac{mk_{\min}}{z_{\max} J_{\max}}\right)^{-t} P_{ii}^{I_r}(t) + \sum_{t=T+1}^{\infty} \left(1 + \frac{mk_{\min}}{z_{\max} J_{\max}}\right)^{-t} P_{ii}^{I_r}(t) \right), \tag{6.96}$$

where the probabilities $P_{ii}^{I_r}(t)$ obtained from (6.76) are independent of m. In the first term of (6.96), in the hypothesis of the existence of the thermodynamic limit we can exchange the limit and the derivatives, since this term is given by a sum of products of two functions, one independent of m, and the other one independent of r. Furthermore, using the property that $0 < P_{ii}^{I_r}(t) < 1$ one can prove that the second term tends to for $T \to \infty$. Hence, letting $T \to \infty$ we obtain (6.95). □

Theorem 6.20. *Let $\mathcal{G}$ be a physical graph , let J_{ij} be a ferromagnetic coupling matrix, and let K_{ij}, K'_{ij} be two local coupling matrices such that $k'_i \geq k_i, \forall i$. Then*

$$\langle \overline{x^2} \rangle^{(n)} (J, mK) \geq \langle \overline{x^2} \rangle^{(n)} (J, mK'). \tag{6.97}$$

Proof. Let us consider the sequence of spheres $\mathcal{S}_r$. Then

$$\frac{\partial}{\partial k_i} \langle \overline{x^2}^r \rangle_r^{(n)} (J^r, mK^r)$$
$$= \frac{-m}{|V_r|} \left[\frac{1}{L^r + mK^r} \left(K^r \frac{1}{L^r + mK^r} \right)^n \frac{1}{L^r + mK^r} \right]_{ii} \leq 0. \tag{6.98}$$

The square brackets in this formula contain a product of positive defined operators (Lemma 6.5). Hence,

$$\langle \overline{x^2}^r \rangle_r^{(n)} (J^r, mK^r) \geq \langle \overline{x^2}^r \rangle_r^{(n)} (J^r, mK'^r). \tag{6.99}$$

Using Lemma 6.19 and letting $r \to \infty$ we get (6.97). □

Corollary 6.21. *Let $\mathcal{G}$ be a physical graph recurrent of degree N and with average spectral dimension $\overline{d} = 2(N - D + 1)$. Then for any local coupling matrices K_{ij}*

$$\lim_{m\to 0^+} \langle \overline{x^2} \rangle^{(n)} (A, mK) < \infty, \ \forall n < N \quad \text{and} \quad \lim_{m\to 0^+} \langle \overline{x^2} \rangle^{(N)} (A, mK) = \infty, \tag{6.100}$$

and

$$D = \lim_{m\to 0^+} \frac{\log(\langle \overline{x^2} \rangle^{(N)} (A, mK))}{.} - \log(m) \tag{6.101}$$

Proof. From Theorem 6.20 and property **p.c.2** one has:

$$\langle\overline{x^2}\rangle^{(n)}(A, mk_{\max}Z) \leq \langle\overline{x^2}\rangle^{(n)}(A, mK) \leq \langle\overline{x^2}\rangle^{(n)}(A, mk_{\min}z_{\max}^{-1}Z). \tag{6.102}$$

From Corollaries 6.17, 6.18 and inequality (6.102) one gets (6.100) and (6.101). □

We proved the invariance of the average spectral dimension for any bounded rescaling of the local coupling matrix. Let us pass to examine rescalings of the ferromagnetic coupling matrix.

Theorem 6.22. *Let $\mathcal{G}$ be a physical graph , let K_{ij} be a local coupling matrix, and let J_{ij}, J'_{ij} be two ferromagnetic coupling matrices such that $J'_{ij} \geq J_{ij}, \forall (i,j) \in E$. Then*

$$\langle\overline{x^2}\rangle^{(n)}(J, mK) \geq \langle\overline{x^2}\rangle^{(n)}(J', mK). \tag{6.103}$$

Proof. Let us consider the sequence of spheres $\mathcal{S}_r$. Then

$$\begin{aligned}\frac{\partial}{\partial J_{ij}}\langle\overline{x^2}^r\rangle_r^{(n)}(J^r, mK^r) &= -\frac{1}{|V_r|}\Phi_{ii} + \Phi_{ji} - \Phi_{ij} - \Phi_{ji} \\ &= -\frac{1}{|V_r|}v^{(i,j)}\Phi v^{(i,j)} \leq 0,\end{aligned} \tag{6.104}$$

where $v \in \mathbb{R}^{|V_r|}$, $v_h^{(i,j)} = \delta_{ih} - \delta_{jh}$, and

$$\Phi_{hk} = \left[\frac{1}{L^r + mK^r}\left(K^r\frac{1}{L^r + mK^r}\right)^n\frac{1}{L^r + mK^r}\right]_{hk}. \tag{6.105}$$

The inequality in (6.104) holds since Φ (6.105) is a positive defined operator (Lemma 6.5). Therefore,

$$\langle\overline{x^2}^r\rangle_r^{(n)}(J^r, mK^r) \geq \langle\overline{x^2}^r\rangle_r^{(n)}(J'^r, mK^r). \tag{6.106}$$

By using Lemma 6.19 and letting $r \to \infty$, we get (6.103). □

Corollary 6.23. *Let $\mathcal{G}$ be a physical graph recurrent of degree N and with average spectral dimension $\overline{d} = 2(N - D + 1)$. Then for any ferromagnetic coupling matrix J_{ij} and any local coupling matrices K_{ij} we have:*

$$\lim_{m\to 0^+}\langle\overline{x^2}\rangle^{(n)}(J, mK) < \infty,\ \forall n < N \quad \text{and} \quad \lim_{m\to 0^+}\langle\overline{x^2}\rangle^{(N)}(J, mK) = \infty, \tag{6.107}$$

and

$$D = \lim_{m\to 0^+}\frac{\log(\langle\overline{x^2}\rangle^{(N)}(J, mK))}{-\log(m)}. \tag{6.108}$$

Proof. From Theorem 6.22 one has:

$$\begin{aligned}J_{\max}^{-1}\langle\overline{x^2}\rangle^{(n)}(A, mJ_{\max}^{-1}K) &\leq \langle\overline{x^2}\rangle^{(n)}(J, mK) \\ &\leq J_{\min}^{-1}\langle\overline{x^2}\rangle^{(n)}(A, mJ_{\min}^{-1}Z).\end{aligned} \tag{6.109}$$

From Corollaries 6.17, 6.18, 6.21 and inequality (6.109) one gets (6.107) and (6.108). □

So we have proven the invariance of the average spectral dimension for a bounded rescaling of the ferromagnetic couplings. In particular, from a random walks point of view we proved the invariance of $\overline{d}$ for a rescaling of the jumping probabilities given by (3.23). Let us pass to prove the invariance with respect to removing (or adding) a zero measure set of edges.

Theorem 6.24. *Given a physical graph $\mathcal{G}$, a local coupling matrix K_{ij}, a ferromagnetic coupling matrix J_{ij} and $E' \subset E$ such that $\|E'\|$, let $J'_{ij} = 0$ if $(i, j) \in E'$ and $J'_{ij} = J'_{ij}$ otherwise. Then*

$$\langle \overline{x^2} \rangle (J, mK) = \langle \overline{x^2} \rangle (J', mK). \tag{6.110}$$

Proof. Let us define the coupling matrix $J_{ij}(\alpha) = J_{ij}(1 - \alpha \chi_{(i,j)}(E'))$ ($\alpha \in \mathbb{R}$) so that $J_{ij}(0) = J_{ij}$ and $J_{ij}(1) = J'_{ij}$. Consider the sequence of increasing spheres $\mathcal{S}_r$ from (6.104):

$$\frac{\partial}{\partial \alpha} \langle \overline{x^2}^r \rangle_r (J^r(\alpha), mK^r) = \sum_{(i,j) \in E'_r} \frac{1}{|V_r|} J_{ij}(\alpha) v^{(i,j)} \Phi v^{(i,j)} \geq 0, \tag{6.111}$$

where $E'_r = \{(i, j) \in E' |\ i \in V_r, j \in V_r\}$ (Definition 2.9). From the boundedness of Φ ($\|\Phi\| \leq k_{\min}^{-1}$) one has

$$0 \leq \frac{\partial}{\partial \alpha} \langle \overline{x^2}^r \rangle_r (J^r(\alpha), mK^r) \leq 2 J_{\max} k_{\min}^{-1} \frac{|E'_r|}{|V_r|}. \tag{6.112}$$

Integrating (6.112) on $\alpha \in [0, 1]$ we have:

$$0 \leq \langle \overline{x^2}^r \rangle_r (J^r(1), mK^r) - \langle \overline{x^2}^r \rangle_r (J^r(0), mK^r) \leq 2 J_{\max} k_{\min}^{-1} \frac{|E'_r|}{|V_r|}. \tag{6.113}$$

Letting $r \to \infty$ in (6.113), and using the fact that $\|E'\| = 0$, we get

$$\langle \overline{x^2} \rangle (J(1), mK) = \langle \overline{x^2} \rangle (J(0), mK).$$

□

Theorem 6.25. *Given a physical graph $\mathcal{G}$ recurrent of degree N and with spectral dimension $\overline{d} = 2(N - D + 1)$, let $\mathcal{G}'$ be the graph given by $V' = V$ and $E' = \{(i, j) | (i, j) \in E \vee (i, j) | \exists k, (i, k), (k, j) \in E\}$ for all ferromagnetic coupling matrix J'_{ij} on $\mathcal{G}'$ and all local coupling matrices K'_{ij}. Then*

$$\lim_{m \to 0^+} \langle \overline{x^2} \rangle^{(n)} (J', mK') < \infty, \ \forall n < N \quad \text{and} \quad \lim_{m \to 0^+} \langle \overline{x^2} \rangle^{(N)} (J', mK') = \infty, \tag{6.114}$$

and

$$D = \lim_{m\to 0^+} \frac{\log(\langle \overline{x^2}\rangle^{(N)}(J', mK'))}{-\log(m)}. \tag{6.115}$$

Proof. From Corollaries 6.21 and 6.23 we have that it is enough to prove equations (6.114) and (6.115) for two particular matrices J'_{ij} and mK'_{ij}. In particular we can chose $mK'_{ij} = m\delta_{ij}$ and $J'_{ij} = A_{ij}(1 - \alpha(I_i + I_j)) + \alpha \sum_k A_{ik}A_{kj}$, where A_{ij} is a ferromagnetic coupling matrix on $\mathcal{G}$ and $\alpha \le (2z_{\max})^{-1}$. With this condition it is easy to show that J'_{ij} is a well-defined ferromagnetic coupling matrix on $\mathcal{G}'$. Let $\mathcal{S}'_r$ be a sequence of increasing spheres in $\mathcal{G}'$ ($\mathcal{S}'_r$ correspond to spheres of radius $2r$ in $\mathcal{G}$ that will be called $\mathcal{S}_2 r$), we have:

$$\begin{aligned}\langle \overline{x^2}^r\rangle_r^{(n)}(J'^r, K'^r) &= \frac{1}{|V'_r|}\sum_{i\in V'_r}\left(\frac{1}{L'^r + m\delta}\right)^{n+1}_{ii} \\ &= \frac{1}{|V_2 r|}\sum_{i\in V_2 r}\left(\frac{1}{\Delta^{2r} - \alpha(\Delta^{2r})^2 + m\delta}\right)^{n+1}_{ii},\end{aligned} \tag{6.116}$$

where we used the property that the Laplacian corresponding to A'_{ij} is $\Delta^{2r} - \alpha(\Delta^{2r})^2$. If we evaluate expression (6.116) in the base where L is diagonal, denoting by $0 \le l_k^{2r} \le 2z_{\max}$ (Theorem 5.1) the eigenvalues of Δ^{2r} we get

$$\begin{aligned}\frac{1}{|V_2 r|}\sum_k^{|V_2 r|}\left(\frac{1}{l_k^{2r} + m}\right)^{n+1} &\le \frac{1}{|V_2 r|}\sum_k^{|V_2 r|}\left(\frac{1}{l_k^{2r} - \alpha(l_k^{2r})^2 + m}\right)^{n+1} \\ &\le \frac{1}{|V_2 r|}\sum_k^{|V_2 r|}\left(\frac{1}{l_k^{2r}(1 - \alpha 2 z_{\max}) + m}\right)^{n+1}\end{aligned} \tag{6.117}$$

Hence, rewriting equation (6.117) in the base of the sites and letting $r \to \infty$, we get from Lemma 6.19:

$$\begin{aligned}\langle \overline{x^2}\rangle^{(n)}(A, m\delta) &\le \langle \overline{x^2}\rangle^{(n)}(J', K') \\ &\le (1 - \alpha 2 z_{\max})^{-1}\langle \overline{x^2}\rangle^{(n)}(A, m(1 - \alpha 2 z_{\max})^{-1}\delta).\end{aligned} \tag{6.118}$$

Inequality (6.118) proves equations (6.114) and (6.115). □

Applying Theorem 6.25 n times we have that $\overline{d}$ is invariant under addition (or removing) of couplings up to any finite distance n. Let us now introduce a very general transformation on the graph $\mathcal{G}$.

Definition 6.26. Let $\mathcal{G}$ be a physical graph, a topological rescaling of $\mathcal{G}$ is any graph $\mathcal{G}^P$ (defined by V^P and E^P) obtained by the following steps.

- Let $P = \{\mathcal{G}_m, \mathcal{G}_n, \dots\}$ any infinite partition of $\mathcal{G}$ given by the subgraphs $\mathcal{G}_n$ (defined by V_n and E_n) satisfying the properties $\forall n$ $\mathcal{G}_n$ is connected, $\bigcup_n^\infty V_n =$

V, $\forall n, m$ $V_n \cap V_m = \emptyset$, $\exists K \in \mathbb{N}$ such that $|V_n| < K$ $\forall n$ and $E_n = \{(i, j) \in E | i, j \in V_n\}$.

- $V^P = \{n | \exists \mathcal{G}_n \in P\}$.
- $E^P = \{(n, m) | \exists (i, j) \in E, i \in V_n, j \in V_m\}$

Theorem 6.27 ([7]). *Given a physical graph $\mathcal{G}$, its average spectral dimension $\overline{d}$ is invariant with respect to topological rescalings introduced in Definition 6.26.*

Proof. Any topological rescaling can be considered as a finite range transformation, the distance involved in the transformations is bounded by the maximum size of the subgraphs $\mathcal{G}_n$. Hence from Theorem 6.25 one immediately obtains the proof. □

7 Conclusions

The results we presented in the previous sections are the background common to any mathematical-physicist working on physical models and random walks on infinite graphs.

All these topics have direct physical applications in the physics of matter, which are explained in detail in the quoted literature.

Indeed, some more advanced topics, mainly concerning phase transitions and critical phenomena, have not been discussed here. This choice is due to two main reasons: they require specific technical knowledge of the general problem of phase transitions in physics and most of these results are to be considered simply as heuristic investigations.

We refer the interested reader to [33] for a general mathematical introduction to phase transitions and critical phenomena and to [9, 11, 12, 14, 15, 16, 25] for specific results on infinite graphs.

References

[1] S. Alexander and R. Orbach, Density of states on fractals: "fractons", J. Physique Lett. 43 (1982) L625–L631.

[2] D. Bertacchi and F. Zucca, Classification on the average for random walks, J. Statist. Phys. 114 (2004), 947–975.

[3] D. Bertacchi and F. Zucca, Uniform asymptotic estimates of transitions probabilities on combs, J. Austral. Math. Soc. 75 (2003), 325–353.

[4] N. Biggs, Algebraic Graph Theory, Cambridge Tracts in Math. 67, Cambridge University Press, London 1974.

[5] A. Blumen and A. Juriju, Multifractal spectra and the relaxation of polymer networks, J. Chem. Phys. 116 (2002), 2636–2641.

[6] R. Burioni and D. Cassi, Universal properties of spectral dimension, Phys. Rev. Lett. 76 (1996), 1091–1093.

[7] R. Burioni and D. Cassi, Geometrical universality in vibrational dynamics, Mod. Phys. Lett. B 11 (1997), 1095–1101.

[8] R. Burioni, D. Cassi, and C. Destri, Spectral partitions on infinite graphs, J. Phys. A 33 (2000), 3627–3636.

[9] R. Burioni, D. Cassi and C. Destri, $n \to \infty$ limit of $O(n)$ ferromagnetic models on graphs, Phys. Rev. Lett. 85 (2000), 1496–1499.

[10] R. Burioni, D. Cassi and S. Regina, Dynamical dimension splitting on fractals: structures with different diffusive and vibrational spectral dimension, Mod. Phys. Lett. B 10 (1996), 1059–1065.

[11] R. Burioni, D. Cassi and A. Vezzani, Inverse Mermin-Wagner theorem for classical spin models on graphs, Phys. Rev. E 60 (1999), 1500–1502.

[12] R. Burioni, D. Cassi and A. Vezzani, Transience on the average and spontaneous symmetry breaking on graphs, J. Phys. A 32 (1999), 5539–5550.

[13] R. Burioni, D. Cassi and A. Vezzani, The type-problem on the average for random-walks on graphs, Eur. Phys. Jour. B 15 (2000), 665–672.

[14] D. Cassi, Phase transitions and random walks on graphs: a generalization of the Mermin-Wagner theorem to disordered lattices, fractals, and other discrete structures, Phys. Rev. Lett. 68 (1992), 3631–3634.

[15] D. Cassi, Local vs. average behavior on inhomogeneous structures: recurrence on the average and a further extension of Mermin-Wagner theorem on graphs, Phys. Rev. Lett. 76 (1996), 2941–2944.

[16] D. Cassi and L. Fabbian, The spherical model on graphs, J. Phys. A 32 (1999), L93–L98.

[17] D. Cassi and S. Regina, Random walks on d-dimensional comb lattices, Modern Phys. Lett. B 6 (1992), 1397–1403.

[18] D. Cassi and S. Regina, Random walks on bundled structures, Phys. Rev. Lett. 76 (1996), 2914–2917.

[19] D. Cassi and S. Regina, Diffusion and harmonic oscillations on bundled structures: analytical techniques and dynamical dimensional splitting, Mod. Phys. Lett. B 11 (1997), 997–1011.

[20] D. M. Cvetković, M. Doob and H. Sachs, Spectra of Graphs. Theory and Applications, Pure Appl. Math. 87, Academic Press, New York 1980.

[21] Y. Gefen, A. Aharony and B.B. Mandelbrot, Critical phenomena on fractal lattices, Phys. Rev. Lett. 45 (1980), 855–858.

[22] P. J. Grabner and W. Woess, Functional iterations and periodic oscillations for simple random walk on the Sierpiński graph. Stochastic Process. Appl. 69 (1997), 127–138.

[23] F. Harary, Graph Theory, Addison-Wesley, Reading, MA, 1969.

[24] K. Hattori, T. Hattori and H. Watanabe, Gaussian field theories on general networks and the spectral dimensions, Prog. Theor. Phys. Suppl. 92 (1987), 108–143.

[25] F. Merkl and H. Wagner, Recurrent random walks and the absence of continuous symmetry breaking on graphs, J. Statist. Phys. 75 (1994), 153–165.

[26] B. Mohar and W. Woess, A survey on spectra on infinite graphs, Bull. London Math. Soc. 21 (1989), 209–234.

[27] E. Montroll and G. H. Weiss, Random walks on lattices. II, J. Math. Phys. 6 (1965), 167–181.

[28] T. Nakayama, K. Yakubo and R. L. Orbach, Dynamical properties of fractal networks: scaling, numerical simulations, and physical realizations, Rev. Mod. Phys. 66 (1994), 381–443.

[29] G. Polya, Über eine Aufgabe der Wahrscheinlichkeitstheorie betreffend die Irrfahrt im Straßennetz, Math. Ann. 84 (1921), 149–160.

[30] A. Procacci, B. Scoppola and V. Gerasimov, Potts model on infinite graphs and the limit of chromatic polynomials, preprint, cond-mat/0201183 (2002).

[31] R. Rammal, Spectrum of harmonic excitations on fractals, J. Physique 45 (1984), 191–206.

[32] R. Rammal and G. Toulouse, Random walks on fractal structures and percolation clusters, J. Physique Lett. 44 (1983), L13–L22.

[33] D. Ruelle, Statistical mechanics: rigorous results, W.A. Benjamin Inc., New York 1969.

[34] G. Weiss and S. Havlin, Some properties of random walks on the comb lattices, Physica A 134 (1986), 474–484.

[35] W. Woess, Random walks on infinite graphs and groups – a survey on selected topics, Bull. London Math. Soc. 26 (1994), 1–60.

[36] W. Woess, Random walks on infinite graphs and groups, Cambridge Tracts in Math. 138, Cambridge University Press, Cambridge 2000.

Raffaella Burioni, Davide Cassi, Alessandro Vezzani, Istituto Nazionale di Fisica della Materia, Dipartimento di Fisica, Università di Parma, 43100 Parma, Italy

E-mail: burioni@fis.unipr.it, cassi@fis.unipr.it, vezzani@fis.unipr.it

The Garden of Eden Theorem for cellular automata and for symbolic dynamical systems

Tullio Ceccherini-Silberstein[*]*, Francesca Fiorenzi and Fabio Scarabotti*

Dedicated to Benjamin Weiss on his 60th birthday

Abstract. We survey the most recent and general results on Garden of Eden (briefly GOE) type theorems in the setting of Symbolic Dynamical Systems and Cellular Automata. We present a GOE type theorem for (one-dimensional) irreducible sub-shifts of finite type and show that a generalization to sofic shifts does not hold in general. We also present a detailed and self-contained proof of a GOE type theorem of Gromov for maps of bounded propagation, between strongly-irreducible stable spaces of finite type over amenable graphs, admitting a dense pseudogroup of holonomy maps.

Contents

1 Introduction 74

2 GOE theorem for one-dimensional sub-shifts 77
- 2.1 Directed graphs and their entropy . . . 77
- 2.2 Symbolic dynamical systems . . . 79
- 2.3 Sofic shifts . . . 84

3 Gromov's theorem 89
- 3.1 Pseudogroups of partial isometries of a graph . . . 89
- 3.2 Subproduct systems on Δ . . . 90
- 3.3 Irreducibility conditions for subproduct systems . . . 92
- 3.4 Maps of bounded propagation . . . 94
- 3.5 Holonomy maps . . . 94
- 3.6 Entropy . . . 95
- 3.7 Amenability . . . 95
- 3.8 Spliceable spaces . . . 96
- 3.9 Pre-injectivity . . . 96

[*]Partially supported by the Swiss National Science Foundation.

3.10 Strict monotonicity 97
3.11 Pre-injectivity corollary 99
3.12 Surjectivity corollary 100
3.13 Garden of Eden theorem for stable spaces and for amenable shifts 100

4 Appendix 103

1 Introduction

The notion of a *cellular automaton* was introduced by Ulam [35] and von Neumann [32]. In the classical situation [29, 30, 32, 35] the *universe* U is the lattice $\mathbb{Z}^2$ of integer points of Euclidean plane. If S is a finite set (the *set of states* or the *alphabet*), a *configuration* is a map $c : U \to S$. A *transition* (or *local*) *map* is a function $\tau : \mathcal{C} \to \mathcal{C}$ from the set $\mathcal{C}$ of all configurations into itself such that the state $\tau[c](x)$ at a point $x \in U$ only depends on the states $c(y)$ at the *neighbours* y of x. In the literature there are different neighbourhoods (corresponding to different metric distances in the universe $U = \mathbb{Z}^2$), for instance, the Moore neighbourhood [29] and that of von Neumann [32]: to fix the ideas we choose the latter. Thus, denoting by

$$B(x; 1) = \{x, y_1 = x + (1, 0), y_2 = x + (-1, 0), y_3 = x + (0, 1), y_4 = x + (0, -1)\}$$

the ball of radius 1 centered at $x \in U$, there exists a function, called *local rule*, $f : S^{B((0,0);1)} \to S$ such that

$$\tau[c](x) = f[c(x), c(y_1), c(y_2), c(y_3), c(y_4)]. \tag{1.1}$$

As U is countable, $\mathcal{C}$ is a compact *metrizable* space, and one shows (see Proposition 4.4) that $\tau : \mathcal{C} \longmapsto \mathcal{C}$ is a transition map induced by a local rule f (associated with a suitable metric distance on U) if and only if it is (uniformly) continuous and $\mathbb{Z}^2$-*equivariant* (or, equivalently, *commutes* with the *shift*): $\tau[c^g] = \tau[c]^g$, where (using the multiplicative notation), $c^g(h) = c(g^{-1}h)$ for all $c \in \mathcal{C}$ and $g, h \in \mathbb{Z}^2$.

One often speaks of τ as being time dynamics: if c is the configuration at time t, then $\tau[c]$ is the configuration at time $t + 1$. An *initial configuration* is a configuration at time $t = 0$. A configuration c not in the image of τ, namely $c \in \mathcal{C} \setminus \tau[\mathcal{C}]$, is called a *Garden of Eden* (briefly *GOE*) *configuration*, this biblical terminology being motivated by the fact that GOE configurations may only appear as initial configurations.

Given a non-empty finite subset $F \subset U$, a *pattern* of *support* F is a map $p : F \to S$. A pattern p is called GOE if any configuration extending p outside its support F is a GOE configuration: $\tau(c)|_F \neq p$ for all $c \in \mathcal{C}$. Using topological methods, namely a compactness argument, one shows (see Proposition 4.1) that the existence of GOE patterns is equivalent to that of GOE configurations.

Two *distinct* patterns p and p' with a common support F are said to be τ-*mutually erasable* if, for all configurations c and c' extending p and p', respectively, and that are equal outside F, i.e. $c|_{U\setminus F} = c'|_{U\setminus F}$, one has $\tau(c) = \tau(c')$.

Also, τ is *pre-injective* if $\tau[c] \neq \tau[c']$ for all $c, c' \in \mathcal{C}$ such that $c \neq c'$ but they differ only in finitely many points (i.e., there exists a finite subset $F \subseteq U$ such that $c|_{U\setminus F} = c'|_{U\setminus F}$). It is easy to show (see Proposition 4.2) that non-existence of τ-mutually erasable patterns and pre-injectivity of τ are equivalent notions.

Given a finitely generated group $G = \langle A \rangle$ with a finite and symmetric system of generators $A = A^{-1}$ one can define, in perfect analogy with the above setting, the notion of a *cellular automaton* over the universe G [5, 12, 28].

One then says that a group G satisfies the *Myhill property* (resp., the *Moore property*) if given any cellular automaton over G, τ pre-injective $\Rightarrow$ τ surjective (resp., τ surjective $\Rightarrow$ τ pre-injective).

The classical *Garden of Eden Theorem*, due to Moore and Myhill [29, 30], states the equivalence between *surjectivity* and *pre-injectivity* of τ for a cellular automaton over $G = \mathbb{Z}^2$. In other words, both the Moore and the Myhill properties hold for $\mathbb{Z}^2$.

This theorem has been extended to groups of sub-exponential growth in [28] and, more generally, to *amenable* groups [5]. In [5] it is also shown that for groups containing the non-abelian free group $\mathbb{F}_2$ (thus, highly non-amenable) both the Myhill property and the Moore property fail to hold in general: it was posed as a problem whether this failure holds also for all (other) non-amenable groups (e.g., for the free Burnside groups $B(m, n)$ or the Ol'shanskii groups, [4]); clearly, a positive answer to this question would give a new characterization of amenability.

A group G is *surjunctive* [22, 25, 37] if, for any finite alphabet S, any transition map $\tau : S^G \to S^G$ is either surjective or non-injective; equivalently, τ injective $\Rightarrow$ τ surjective, which can be regarded as a sort of *co-hopfianity* condition (see [8] for *hopfianity* and this latter notion). Since injectivity implies pre-injectivity, we have that groups satisfying the Myhill property are surjunctive. We recall that *sofic groups* – a notion due to Gromov and Weiss, generalizing both amenable groups and *residually finite* groups (e.g., the free groups) – are surjunctive.

Section 2 of the present paper is devoted to one-dimensional (i.e., with universe $U = \mathbb{Z}$) Garden of Eden type theorems. Basic notions like *irreducibility* and being of *finite type* are introduced in this setting where, we believe, the reader has a gentler approach than in the multi-dimensional case. Thus, although the main result, namely Theorem A, cannot be recovered as a particular case of Theorem B (i.e., Gromov's theorem, which constitutes the main result of Section 3) because different kinds of irreducibility are involved, this section should be interpreted as a preparation towards the much more articulated following section.

The book by Lind and Marcus [27] is an excellent comprehensive introduction to the theory of symbolic dynamical systems. The theory of *one-dimensional* shifts (or subshifts), i.e., closed shift-invariant subspaces of $S^{\mathbb{Z}}$, is investigated there in full detail focusing the strong connections with other settings (graph theory, theory of formal languages, Perron–Frobenius theory, etc.).

In her PhD dissertation [12], the second named author investigated GOE type theorems for one-dimensional subshifts and for subshifts over amenable groups: clearly the setup is slightly more delicate than in [5], since notions like *irreducibility* and *finite-*

ness (e.g., requirement for a subshift to be of *finite type* or to be *sofic* [27, 36, 37]: see Section 2 for all the details) play a determinant role. Using the notion of *entropy* (e.g., see [25, 27]) one can relax the condition for a transition map to be an *endo*morphism of a single sub-shift, by considering a transition map between two distinct subshifts. The Garden of Eden Theorem in this setting then becomes

Theorem A. *Let X and Y be irreducible (one-dimensional) sub-shifts of finite type and $\tau : X \to Y$ a transition map. Suppose that* $\mathrm{ent}(X) = \mathrm{ent}(Y)$ *(for example, this condition is satisfied if $X = Y$). Then τ is pre-injective if and only if it is surjective.*

Sofic shifts were introduced by B. Weiss [36] as the minimal class containing all the shifts of finite type and closed under *factorizations*: a *factor* is the image under a local map. A natural question then arises: what can one say about GOE type theorems for *sofic* shifts? The answer is known from [12, 13]: the Myhill property holds for irreducible sofic shifts, see Theorem 2.17, but the Moore property in general fails to hold, see Counterexample 2.18.

The last section is devoted to a Garden of Eden type theorem due to Gromov. In [25], using entropic arguments, Gromov generalized the Garden of Eden theorem of [5] in the following way (see our Section 3 for all definitions and notions involved in the statement).

Theorem B. *Let Δ be an amenable graph. Suppose that $X, Y \subseteq S^{\Delta}$ are strongly-irreducible stable spaces of finite type with the same entropy:* $\mathrm{ent}(X) = \mathrm{ent}(Y)$ *(for example, $X = Y$). Then a map of bounded propagation $\tau : X \to Y$ admitting a dense holonomy is surjective if and only if it is pre-injective.*

The remarkable point is that Gromov does not restrict himself to graphs Δ which might occur as Cayley graphs of groups, but to graphs having a sufficient regularity (*dense pseudogroups of isometries* and *dense holonomy*); also he considers, more generally, *subshifts* rather than *full shifts*: the set of configurations $\mathcal{C}$ is now a *closed shift-invariant* subset of S^{Δ}. Thus [25] covers [5] and, in this setting, the generalization can be formulated as follows (Garden of Eden Theorem for strongly-irreducible amenable subshifts):

Corollary C. *Let G be an amenable finitely generated group and $X, Y \subseteq S^{G}$ two strongly-irreducible subshifts of finite type with the same entropy* $\mathrm{ent}(X) = \mathrm{ent}(Y)$ *(for example, $X = Y$). Then a local map $\tau : X \to Y$ is surjective if and only if it is pre-injective.*

We analyze in detail all definitions and terminology considered by Gromov: since these latter slightly differ from the usual ones from the current literature (e.g., from [27]), our purpose is to provide a "dictionary" between these different points of view: the parallelism is not always complete and we shall point out the differences. Also, although the key point in Gromov's proof is essentially the fundamental Lemma 3.18 on strict monotonicity – all other statements, together with several definitions and notions, being scattered, even with more general results, over various preceding sections

of the long paper [25] – our proof should be thought of as a completely self-contained and thus more accessible version of the original proof.

2 GOE theorem for one-dimensional sub-shifts

We start this section by reviewing some known facts about directed graphs and their entropy. The fundamental entropic inequality for subgraphs of irreducible graphs (Theorem 2.2) is presented here in a new version from [34] which avoids the Perron–Frobenius theory, by means of which it is usually proved, but is based on the techniques of Gromov [25] further developed in Section 3.

2.1 Directed graphs and their entropy

A *finite, directed graph* G is given by a finite set $V(G)$ of *vertices*, a finite set $E(G)$ of *edges* and two functions $i, t : E(G) \to V(G)$. If $e \in E(G)$ then $i(e)$ and $t(e)$ are the *initial vertex* and the *terminal vertex* of e, respectively. We say that G is *simple* if $e, e' \in E(V)$ and $e \neq e'$ implies $(i(e), t(e)) \neq (i(e'), t(e'))$. A *path of length* n in G is a finite sequence $\pi = e_1 e_2 \dots e_n$ of edges such that $t(e_k) = i(e_{k+1})$ for $k = 1, 2, \dots, n-1$; π starts at edge $i(\pi) = e_1$ and terminates at edge $t(\pi) = e_n$. A *bi-infinite path* in G is a sequence $\xi = \{e_k\}_{k\in\mathbb{Z}}$ of edges such that $t(e_k) = i(e_{k+1})$ for any $k \in \mathbb{Z}$.

Let G be a finite, simple, directed graph. A *word* in G is a finite sequence $a_1 a_2 \dots a_m$ of vertices such that there is an edge starting in a_i and terminating in a_{i+1} for $i = 1, 2, \dots m-1$. Define $B_m(G)$ as the set of all words in G of length m. The entropy $\mathrm{ent}(G)$ of G is defined as

$$\mathrm{ent}(G) = \lim_{m\to\infty} \frac{\log |B_m(G)|}{m}. \tag{2.1}$$

Such a limit always exists (use the Fekete–Polya lemma, e.g., see Lemma 4.17 and Proposition 4.18 in [27]).

A graph G is *irreducible* when, given any ordered pair of vertices a and b, there exists a path from a to b. A word $a_1 a_2 \dots a_m$ in G is *simple* when all the vertices are distinct; it is a *cycle* when $a_1 = a_m$ and $a_i \neq a_j$ if $\{i, j\} \neq \{1, m\}$.

Given an arbitrary word $w = a_1 a_2 a_3 \dots a_n$, we can form its *decomposition into cycles* as follows. Let i_1 be the largest index such that the vertices $a_1, a_2 \dots a_{i_1-1}$ are all distinct; then $a_{i_1} = a_{j_1}$ for a suitable $j_1 < i_1$, and $c_1 = a_{j_1} a_{j_1+1} \dots a_{i_1}$ is the *first cycle* of the word; also set $r_1 = a_1 a_2 \dots a_{j_1}$. Successively consider the largest index $i_2 > i_1$ such that the vertices $a_{i_1} a_{i_1+1} \dots a_{i_2-1}$ are all distinct; then $a_{i_2} = a_{j_2}$ for a suitable $j_2 \in \{i_1, i_1+1, \dots i_2-1\}$ and $c_2 = a_{j_2} a_{j_2+1} \dots a_{i_2}$ is the *second cycle* of the word; set $r_2 = a_{i_1+1} a_{i_1+2} \dots a_{j_2}$. Continuing this way we obtain a (unique) canonical decomposition $w = r_1 c_1 r_2 c_2 \dots r_k c_k r_{k+1}$, with some overlappings at the extremities,

where $c_1, c_2, \ldots, c_k$ are the cycles and $r_1, r_2, \ldots, r_{k+1}$ are simple (possibly empty) words. With this notation we say that $\overline{w} = a_1 \ldots a_{j_s-1} a_{j_s} a_{i_s+1} \ldots a_n$ is obtained from w by *collapsing* the s-th cycle c_s.

Lemma 2.1. *Let G be an irreducible graph and $e \in E(G)$ an edge in G. Then there exists n such that if $w = a_1 a_2 \ldots a_n$ is a word in G of length n then there exists a word $w' = a_1 a_2' \ldots a_{n-1}' a_n$ containing e.*

Proof. It follows from the irreducibility that there exists a word v starting at a_1, terminating in a_1 and containing e. Also, if n is large enough, in the canonical decomposition $w = r_1 c_1 r_2 c_2 \ldots r_k c_k r_{k+1}$ of w there exists a cycle c repeated many times. If the length of c is ℓ, the length of v is m, and the cycle c is repeated at least m times, we may collapse the first $m-1$ copies of c and add $\ell - 1$ copies of v at the beginning obtaining the desired word. □

The following fundamental entropic inequality is usually proved by means of the Perron–Frobenius theorem (see, e.g., [27]); we present a new proof from [34] based on Gromov's techniques in [25].

Theorem 2.2. *If G is an irreducible graph and H is obtained from G by deleting one edge $e \in E(G)$, then* $\mathrm{ent}(H) < \mathrm{ent}(G)$.

Proof. Let n be the integer from Lemma 2.1. We first show that setting

$$\alpha = \frac{1}{|B_n(G)|}$$

one has, for $k = 1, 2, \ldots$

$$|B_{kn}(H)| \leq (1-\alpha)^k |B_{kn}(G)|. \tag{2.2}$$

Clearly, we have $|B_{n(k-1)}(G)| \geq \alpha \cdot |B_{kn}(G)|$. In what follows a word w of length kn will be represented as the concatenation of words of length n, i.e., in the form $w = w_1 w_2 \ldots w_k$, where w_h is a word of length n for $h = 1, 2, \ldots, k$. Define π_h as the set of all $w \in B_{kn}(G)$ such that w_h contains the edge $i(e)t(e)$. By Lemma 2.1 for any $w' \in B_{(k-1)n}(G)$ there exists a word $vw' \in B_{kn}(G)$ such that v contains $i(e)t(e)$. Then $|\pi_1| \geq |B_{n(k-1)}(G)|$. Therefore,

$$|B_{nk}(G) \setminus \pi_1| \leq (1-\alpha)|B_{nk}(G)|.$$

Then define

$$B_{nk}^h(G) = B_{nk}(G) \setminus \bigcup_{l=1}^{h} \pi_l, \quad C^h = \{w \in B_{nk}^h(G) : w_{h+1} \text{ contains } e\},$$

and D^h as the set of all couples of words (v_1, v_2) such that $v_1 \in B_{hn}(G)$, $v_2 \in B_{(k-h-1)n}(G)$ and there exists $v \in B_n(G)$ such that $v_1 v v_2 \in B_{kn}^h(G)$. Lemma 2.1 implies that for any $(v_1, v_2) \in D^h$ there exists $w = v_1 v' v_2 \in B_{kn}(G)$ such that v'

contains e. Thus $w \in C^h$. Therefore we have again: $|C^h| \geq |D^h| \geq \alpha \cdot |B^h_{kn}(G)|$, and so:

$$c|B_{kn}(G) \setminus \bigcup_{l=1}^{h} \pi_l| = |B^{h-1}_{kn}(G) \setminus \pi_h| = |B^{h-1}_{kn}(G) \setminus C^{h-1}|$$
$$\leq (1-\alpha)|B^{h-1}_{kn}(G)| \leq (1-\alpha)^h |B_{kn}(G)|,$$

where the last inequality follows by an obvious inductive argument on h. Then for $h = k$ we obtain (2.2).

Taking logarithms and dividing by kn in (2.2), we obtain

$$\frac{\log|B_{kn}(H)|}{kn} \leq \frac{\log(1-\alpha)}{n} + \frac{\log|B_{kn}(G)|}{kn},$$

and, therefore, letting $k \to \infty$, we have

$$\mathrm{ent}(H) \leq \frac{\log(1-\alpha)}{n} + \mathrm{ent}(G) < \mathrm{ent}(G). \qquad \square$$

2.2 Symbolic dynamical systems

We start by recalling from [27] some basic facts on symbolic dynamical systems.

Let S be a finite alphabet. A *word w of length n* over S is a finite sequence $w = x_1 x_2 \ldots x_n$ of symbols $x_i \in S, i = 1, 2, \ldots, n$. A *bi-infinite word* x is a bi-infinite sequence of symbols of S: $x = \ldots x_{-2} x_{-1} x_0 x_1 x_2 \ldots$. We say that a word w of length n is contained in the bi-infinite word x if there is an index $i \in \mathbb{Z}$ such that $w = x_i x_{i+1} \ldots x_{i+n-1}$ The set of all bi-infinite words is denoted as usual by $S^{\mathbb{Z}}$ and it is called the *full shift*. It is a compact space if endowed with the product topology (S is a discrete space) and the map $\sigma : S^{\mathbb{Z}} \to S^{\mathbb{Z}}$ defined by $\sigma(x)_i = x_{i+1}$, called the *shift*, is continuous. A *subshift*, or *symbolic dynamical system*, is a subset of $S^{\mathbb{Z}}$ that is closed and shift invariant. If X is a subshift, its *language* $B(X)$ is the set of all words $\{x_i x_{i+1} x_{i+2} \ldots x_{i+n} \mid x \in X, \ i \in \mathbb{Z} \text{ and } n \in \mathbb{N}\}$. A subshift $X \subseteq S^{\mathbb{Z}}$ is always described by means of a set $\mathcal{F}$ of *forbidden words*: there exists a set $\mathcal{F}$ of words over the alphabet S such that X is the set of all $x \in S^{\mathbb{Z}}$ containing no words from $\mathcal{F}$ as subwords; conversely, for any set $\mathcal{F}$, the set of all $x \in S^{\mathbb{Z}}$ not containing the words from $\mathcal{F}$ as subwords is a subshift, denoted $X_{\mathcal{F}}$ (see Proposition 4.3). A subshift $X = X_{\mathcal{F}}$ is *of finite type* if it is possible to describe it in terms of a finite set $\mathcal{F}$ of forbidden words. If this is the case, the maximum M of the lengths of the words from $\mathcal{F}$ is called the *memory* of X. The shifts of finite type are characterized by the following *overlapping property*: X is a shift of memory M if and only if whenever $uv, vw \in B(X)$ and $|v| \geq M - 1$, then $uvw \in B(X)$.

A subshift X is *irreducible* if, for every pair of finite words $u, v \in B(X)$, there exists a word $w \in B(X)$ such that $uwv \in B(X)$.

Let X be a subshift. Define $B_n(X)$ as the set of all words in $B(X)$ of length n. The entropy $\mathrm{ent}(X)$ of X is defined as

$$\mathrm{ent}(X) = \lim_{n\to\infty} \frac{\log |B_n(X)|}{n}, \tag{2.3}$$

and, as for the entropy of a graph (2.1), such a limit always exists.

Let now X be a shift of finite type with memory M. For $m \geq M$ consider the graph $G = G(X, m)$ whose vertices are the allowed words of length m, i.e. $V(G) = B_m(X)$ and the edges are the words of length $m+1$ in $B(X)$: $E(G) = B_{m+1}(X)$ with $i(e) = a_1a_2 \dots a_m$ and $t(e) = a_2 \dots a_m a_{m+1}$ for an edge $e = a_1a_2 \dots a_m a_{m+1}$. This construction comes from [27] where the shift associated to $G(X, m)$ is called the *m-higher block shift*. It is easy to show that if X is irreducible then G is also irreducible. Note also that $|B_n(G)| = |B_{n+m-1}(X)|$, so that $\mathrm{ent}(G) = \mathrm{ent}(X)$. We can now prove the following.

Theorem 2.3. *If X is an irreducible shift of finite type, and Y is a proper subshift of X, then* $\mathrm{ent}(Y) < \mathrm{ent}(X)$.

Proof. Let $X = X_{\mathcal{F}}$ and $Y = X_{\mathcal{F}'}$ with $\mathcal{F}$ finite and contained in $\mathcal{F}'$. Choose $W \in \mathcal{F}' \setminus \mathcal{F}$ and set $m = \max\{M, |W|\}$, where $M = \max\{|v| : v \in \mathcal{F}\}$ is the memory of X. Then if $w \in B_{m+1}(X)$ contains W as a subword one has

$$\mathrm{ent}(Y) \leq \mathrm{ent}(X_{\mathcal{F}\cup\{W\}}) \leq \mathrm{ent}(X_{\mathcal{F}\cup\{w\}}) \leq \mathrm{ent}(X). \tag{2.4}$$

Consider the graph $G = G(X, m)$. Forbidding the word w corresponds to deleting an edge in G, thus Theorem 2.2 applies and the last inequality in (2.4) is strict. □

A map $\tau : S^{\mathbb{Z}} \to S^{\mathbb{Z}}$ is *k-local* (compare with the notion of *transition map* for a cellular automaton from the introduction) if there exists a function $\overline{\tau} : S^{2k+1} \to S$ such that for all $x \in S^{\mathbb{Z}}$ the bi-infinite word $y = \tau(x)$ is defined by:

$$y_n = \overline{\tau}(x_{n-k}, x_{n-k+1}, \dots, x_n, \dots, x_{n+k-1}, x_{n+k})$$

for every $n \in \mathbb{Z}$. We also say that τ has memory k. It is a well-known fact (see Proposition 4.4) that if X and Y are subshifts then a function $\tau : X \to Y$ is k-local for some k if and only if it is continuous and commutes with the shifts.

A local map τ is called *pre-injective* if whenever $x, y \in X$ and the set $\{i \in \mathbb{Z} : x_i \neq y_i\}$ is finite and non-empty, then $\tau(x) \neq \tau(y)$.

If Y is a subset of $S^{\mathbb{Z}}$ which is not necessarily a subshift, we can define

$$\mathrm{ent}(Y) = \liminf_{m\to\infty} \frac{\log |B_{[-m+1,m]}(Y)|}{2m}, \tag{2.5}$$

where $B_{[-m+1,m]}(Y) = \{y_{-m+1}y_{-m+2} \dots y_0 \dots y_{m-1}y_m : y \in Y\}$. Then we have:

Proposition 2.4. *If $Y \subseteq S^{\mathbb{Z}}$ and τ is a local map, then* $\mathrm{ent}(\tau(Y)) \leq \mathrm{ent}(Y)$.

Proof. If τ is k-local, then $|B_{[-m+1,m]}(\tau(Y))| \leq |B_{[-m-k+1,m+k]}(Y)|$; dividing by $2m$ and taking the lim inf we obtain the desired inequality. □

Let G be a finite, directed graph. The *period* of a vertex $v \in V(G)$ is the greatest common divisor of the lengths of the closed paths starting and terminating at v. We recall from Sections 4.4 and 4.5 of [27] the following:

Lemma 2.5. *Let G be an irreducible directed graph. Then:*

(i) *All vertices $v \in V(G)$ have the same period, called the* period *of G.*

(ii) *If m is relatively prime with the period p of G then, for every $u, v \in V(G)$, there exist $k \in \mathbb{N}$ and a path of length km starting at u and terminating at v.*

Corollary 2.6. *If X is an irreducible shift of memory M and the associated graph $G = G(X, M)$ has period p, then, for every $m \geq M$ prime with p, and for every $u, v \in B_m(X)$, there exist $k \in \mathbb{N}$ and a word $w \in B_{km}(X)$ such that $uwv \in B(X)$.*

If X is a shift of memory M and $m \geq M$, the *m-power* shift X^m of X is the 2-memory shift defined by taking $B_m(X)$ as the alphabet and forbidding all the words uv with $u, v \in B_m(X)$ such that $uv \notin B(X)$. It follows from Corollary 2.6 that if X is irreducible and m is prime with the period of the graph associated with X, then X^m is irreducible too. There is a canonical bijection $\psi : X \to X^m$ given by

$$\psi(x)_k = x_{km+1}x_{km+2}\ldots x_{(k+1)m}$$

for any $x \in X$. If $\psi(x)_k = u$ for some $k \in \mathbb{Z}$, we say that x contains u in an *m-position*.

Theorem 2.7. *Let X be an irreducible shift of memory M. Suppose that Y is the subset of X obtained from X by forbidding a word in $B_m(X)$ in any m-position, with m prime to the period of $G(X, M)$. Then* $\mathrm{ent}(Y) < \mathrm{ent}(X)$.

Proof. First note that $|B_k(X^m)| = |B_{mk}(X)|$, so that $\mathrm{ent}(X^m) = m \cdot \mathrm{ent}(X)$. Moreover we may apply Theorem 2.3: Y corresponds under ψ to a proper subshift Y' of X^m, so that $\mathrm{ent}(Y') < \mathrm{ent}(X^m)$. Finally, from $|B_{[-mk+1,mk]}(Y)| = |B_{2k}(Y')|$ and, recalling (2.5), we can deduce that $m \cdot \mathrm{ent}(Y) \leq \mathrm{ent}(Y')$ and the theorem follows. □

We can now prove the following result that stems from the works of Hedlund [19] and of Coven and Paul [10]; also see [25, pp. 268–269]:

Theorem 2.8. *Let X be an irreducible shift of finite type and $\tau : X \to A^{\mathbb{Z}}$ a local map. Then τ is pre-injective if and only if* $\mathrm{ent}(X) = \mathrm{ent}(\tau(X))$.

Proof. Let M denote the maximum among the memory of X and that of τ. Firstly suppose that τ is pre-injective. Let $n \geq 2M$; for each pair of $u, v \in B_M(X)$ and $w \in B_n(\tau(X))$ there is at most one $p \in B_n(X)$ from u to v such that $\tau(upv) = w$. Thus,

$$|B_{n-2M}(\tau(X))| \leq |B_n(X)| \leq |B_n(\tau(X))| \cdot |B_M(X)|^2,$$

and it follows from the definition of entropy that $\mathrm{ent}(X) = \mathrm{ent}(\tau(X))$.

Now suppose that τ is not pre-injective. Then there exist $x, y \in X$ with $x \neq y$ but $x_i = y_i$ for $i \notin \{1, 2, \ldots, k\}$ for some $k \in \mathbb{N}$, and such that $\tau(x) = \tau(y)$. Define $u = x_{-M+1} \ldots x_0 x_1 \ldots x_k \ldots x_{k+t}$ and $v = y_{-M+1} \ldots y_0 y_1 \ldots y_k \ldots y_{k+t}$; where $t \geq M$ is chosen in such a way that the common length $m = k + t + M$ of u and v is prime with the period of the associated graph $G(X, M)$. If now z' is obtained from $z \in X$ by replacing a given occurrence of u by v, then $z' \in X$ and $\tau(z') = \tau(z)$. Therefore, if Y is obtained from X by forbidding u in any m-position, then $\tau(X) = \tau(Y)$ and combining Theorem 2.7 with Proposition 2.4 we get $\mathrm{ent}(\tau(X)) \leq \mathrm{ent}(Y) < \mathrm{ent}(X)$. □

Now we can prove Theorem A from the introduction, namely the Garden of Eden Theorem for irreducible shifts of finite type.

Proof of Theorem A. If τ is pre-injective then, by Theorem 2.8, $\mathrm{ent}(\tau(X)) = \mathrm{ent}(X)$, and Theorem 2.3 applied to $\tau(X) \subseteq Y$ ensures the surjectivity of τ. Conversely, if τ is surjective, then $\mathrm{ent}(\tau(X)) = \mathrm{ent}(Y) = \mathrm{ent}(X)$, so that, again from Theorem 2.8, it follows that τ is pre-injective. □

Counterexample 2.9. Myhill property no longer holds for a 1-dimensional subshift of finite type which is not irreducible.

Proof. Set $X = \{0, 1\}^{\mathbb{Z}}$ and $\overline{X} = X \cup \{\overline{2}\}$, where $\overline{2}$ is the bi-infinite word with constant value 2. Then $\overline{X}$ is a subshift of finite type over the alphabet $\{0, 1, 2\}$ with set of forbidden words $\{02, 20, 12, 21\}$.

Also, $\overline{X}$ is not irreducible; indeed, $1, 2 \in B(\overline{X})$, but for no word $w \in B(\overline{X})$ the word $1w2$ belongs to $B(\overline{X})$. Consider the transition map $\tau : \overline{X} \to \overline{X}$ defined as

$$\tau(c) = \begin{cases} c & \text{if } c \in X, \\ \overline{0} & \text{if } c = \overline{2}. \end{cases}$$

Then it is easy to show that τ is pre-injective but not surjective. □

Counterexample 2.10. Moore-property no longer holds for a shift of finite type which is not irreducible.

Proof. Let X be the shift over the alphabet $\{0, 1, 2\}$ with the set of forbidden words $\{01, 02\}$, so that

$$X = \{1, 2\}^{\mathbb{Z}} \cup \{\overline{0}\} \cup \left\{ w\overline{0}^{r} : w = \ldots w_i w_{i+1} \ldots w_k;\ w_t \in \{1, 2\} \right\},$$

where $\{1, 2\}^{\mathbb{Z}}$ denotes, as usual, the full shift on the letters 1 and 2; $\overline{0}$ is, as above, the bi-infinite sequence of zeroes and, finally, w is a *left*-infinite sequence in $\{1, 2\}^{\mathbb{N}}$, and $\overline{0}^{r} = 00 \cdots \in \{0\}^{\mathbb{N}}$ is a *right*-infinite sequence of zeroes. X is not irreducible, since for no word $u \in B(X)$ the word $0u1$ belongs to $B(X)$. Consider the transition map

$\tau : X \to X$ defined by the local rule:

$$f(a_{-1}, a_0, a_1) = \begin{cases} a_0 & \text{if } a_1 \neq 0, \\ 0 & \text{if } a_1 = 0. \end{cases}$$

The function τ is surjective because we have

$$\begin{aligned} \tau[\overline{0}] &= \overline{0}, \\ \tau[w1\overline{0}^r] &= w0\overline{0}^r \quad \text{for all } w \in \{1,2\}^{\mathbb{N}}, \\ \tau[v] &= v \quad \text{for all } v \in \{1,2\}^{\mathbb{Z}}. \end{aligned}$$

This also shows that τ is not pre-injective; indeed, $\tau[w\mathbf{1}\overline{0}^r] = w0\overline{0}^r = \tau[w\mathbf{2}\overline{0}^r]$. □

In dimension 2, or, more generally for subshifts over finitely generated groups, there is a notion of irreducibility which extends naturally that for one-dimensional sub-shifts: a shift $X \subseteq S^G$ is irreducible if for all pair of patterns p and p' with supports F and F', respectively, there exist $g \in G$ and a configuration $c \in X$ such that $F \cap gF' = \emptyset$, $c|_F = p$ and $c'|_{gF'} = p'$ (here $gF' = \{gf' : f' \in F'\}$ is the (left-)translation of F' by the element g). In topological terms, X is irreducible if and only if for all pairs of open subsets $U, V \subseteq X$, there exists $g \in G$ such that $U \cap V^g \neq \emptyset$ (where $V^g = \{c^g : c \in V\}$ and, as usual, $c^g[h] = c[g^{-1}h]$); indeed, for any pattern p with support F consider the set $U_p = \{c \in X : c|_F = p\}$ consisting of all configurations extending p outside of its support: this is an open set in the topology of X.

It turns out – see the next counterexample – that this is too weak a notion of irreducibility to guarantee a multi-dimensional GOE type theorem. There is a notion of *strong irreducibility* (from [12]: see Definition 3.8 in our Section 3) which ensures a GOE theorem (e.g., see Theorem B and Corollary C).

Counterexample 2.11. The Garden of Eden theorem no longer holds, in general, for two-dimensional irreducible shifts of finite type.

Proof. If X is a subshift of $S^{\mathbb{Z}}$, then the subshift

$$X^2 = \{(s_{ij})_{i,j\in\mathbb{Z}} : (s_{\bar{\imath}j})_{j\in\mathbb{Z}} \in X, \text{ for all } \bar{\imath} \in \mathbb{Z}\}$$

consisting of (independent) horizontal copies of X, is always irreducible and it is of finite type if X is so. A transition map $\tau : X \to X$ may be extended to a transition map $\tau_2 : X^2 \to X^2$ acting independently on each horizontal line. Now, to obtain counterexamples to the Garden of Eden Theorem it suffices to consider the previous one-dimensional counterexamples and apply the above construction. □

2.3 Sofic shifts

Let G be a finite, directed graph. If S is a finite alphabet, a *labelling* of G is a map $L : E(G) \to S$. The label of a path $\pi = e_1 e_2 \dots e_n$ is the word $L(\pi) = L(e_1)L(e_2)\dots L(e_n)$. If $\xi = \{e_k\}_{k\in\mathbb{Z}}$ is a bi-infinite path in G, its label is given by

$$L(\xi) = \dots L(e_{-1})L(e_0)L(e_1)\cdots \in S^{\mathbb{Z}}.$$

The *sofic shift* X *presented* by (G, L) is the set of the labels of the infinite paths in G:

$$X = X(G, L) = \{L(\xi) \mid \xi \text{ is a bi-infinite path in } G\}.$$

If $S = E(G)$, and L is the identity map, then X is called the *edge shift* associated with G; it is of finite type with memory $M = 2$. Every sofic shift is a shift space, and every shift of finite type is sofic. A labelled graph (G, L) is *deterministic* or *right resolving* if for each vertex $v \in V(G)$ the edges starting at v carry different labels, i.e., $e, e' \in E(G)$, $e \neq e'$ and $i(e) = i(e')$ implies $L(e) \neq L(e')$. We recall a basic fact on irreducible sofic shifts (see Section 3.3 in [27] and, in particular, Theorem 3.3.2):

Theorem 2.12. *An irreducible sofic shift may be presented by an irreducible deterministic labelled graph.*

Proposition 2.13. *Let (G, L) be a deterministic labelled graph, Y the edge shift associated to G, and X the sofic shift presented by (G, L). Then* $\mathrm{ent}(Y) = \mathrm{ent}(X)$.

Proof. Clearly, L is a 1-local map from Y onto X. Then, by Proposition 2.4, $\mathrm{ent}(X) \leq \mathrm{ent}(Y)$. Moreover, since (G, L) is deterministic, every word in $B_n(X)$ has at most $|V(G)|$ preimages under L. Therefore, $|B_n(X)| \geq \frac{1}{|V(G)|}|B_n(Y)|$. Taking logarithms, dividing by n and letting $n \to \infty$ we obtain the inverse inequality. □

Theorem 2.14. *If X is an irreducible sofic shift and Y is a proper subshift of X, then* $\mathrm{ent}(Y) < \mathrm{ent}(X)$.

Proof. From Theorem 2.12 there exists an irreducible, deterministic labelled graph (G, L) that presents X. Let X' be the edge shift of G. Then, by Proposition 2.13, $\mathrm{ent}(X) = \mathrm{ent}(X')$. Using the 1-local map L from X' onto X, define $Y' = L^{-1}(Y)$, which is a proper subshift of X'. Then, from Theorem 2.3,

$$\mathrm{ent}(Y) \leq \mathrm{ent}(Y') < \mathrm{ent}(X') = \mathrm{ent}(X).$$ □

Remark 2.15. A *finite-state-automaton* is a (finite) labelled graph with a distinguished *initial state* and a distinguished subset of *terminal states*. A *language* L is a set of words over a finite alphabet. The language associated with a finite-state-automaton is the set of all labels of paths that begin at the initial state and end at a terminal state, and a language is called a *regular language* if it is of this form; there are other equivalent constructive definitions of such languages in terms of *right-* (or equivalently *left-*)

linear grammars, see [18, 20]. As remarked by W. Krieger [26], there is a connection between sofic shifts and regular languages: the languages of sofic shifts are precisely the *factorial* (i.e., closed under subwords) and *prolongable* (i.e., every word in the language can be extended to the left and to the right to obtain a longer word still in the language) regular languages. For further reading on the connections between automata theory (= theory of formal languages) and symbolic dynamics, see the nice survey of Béal and Perrin [2].

Theorem 2.14 is well-known in the setting of the theory of formal languages (e.g., see [6]) and in that of geometric group theory [11, 23], where irreducibility is often called *ergodicity*, the exponential of the entropy is usually called the (*exponential*) *growth rate* and "sofic shift" is replaced by "regular language". This result has been recently extended to *irreducible unambiguous non-linear context-free languages* (a class of languages generalizing the regular languages) in [9]; "unambiguous" corresponds to "deterministic".

Lemma 2.16. *Let (G, L) be an irreducible deterministic labelled graph and denote by $X = X(G, L)$ the corresponding irreducible sofic shift. Let Y be another shift, and $\tau : X \to Y$ a local map. Then $\tau \circ L$ is pre-injective if and only if* $\mathrm{ent}(X) = \mathrm{ent}(\tau(X))$.

Proof. Let X' be the edge shift of G. Then $\tau \circ L : X' \to Y$ is a local map; thus, by Theorem 2.8 applied to the irreducible shift X' of finite type, we have that $\tau \circ L$ is pre-injective if and only if $\mathrm{ent}(X') = \mathrm{ent}(\tau(L(X'))) = \mathrm{ent}(\tau(X))$. By Proposition 2.13, $\mathrm{ent}(X') = \mathrm{ent}(X)$, and the assertion follows. □

Now we can prove the Myhill property for irreducible sofic shifts.

Theorem 2.17. *Let X, Y be irreducible sofic shifts with the same entropy,* $\mathrm{ent}(X) = \mathrm{ent}(Y)$*, and let $\tau : X \to Y$ be a local map. Then τ pre-injective implies τ surjective.*

Proof. Let (G, L) be an irreducible deterministic labelled graph presenting X, and denote by X' the corresponding edge-shift on G. We first show that if τ is pre-injective, then $\tau \circ L$ is pre-injective. Indeed, if $\tau \circ L$ is not pre-injective, then there exist two bi-infinite paths $\xi_1 = \ldots e_{-1}e_0e_1 \ldots e_ne_{n+1} \ldots$ and $\xi_2 = \ldots e_{-1}f_0f_1 \ldots f_ne_{n+1} \ldots$ in G which differ only for finitely many edges (in particular, say $e_0 \neq f_0$) such that $\tau(L(\xi_1)) = \tau(L(\xi_2))$. Setting $a_i := L(e_i)$, $i \in \mathbb{Z}$ and $b_i := L(f_i)$, $i = 0, \ldots, n$, the labelled graph being deterministic, we have $a_0 \neq b_0$, and hence

$$L(\xi_1) = \ldots a_{-2}a_{-1}a_0a_1 \ldots a_{n-1}a_na_{n+1} \ldots$$

and

$$L(\xi_2) = \ldots a_{-2}a_{-1}b_0b_1 \ldots b_{n-1}b_na_{n+1} \ldots$$

are two configurations in X which differ only on a finite (non empty) set and whose images under τ are equal. Therefore, τ is not pre-injective.

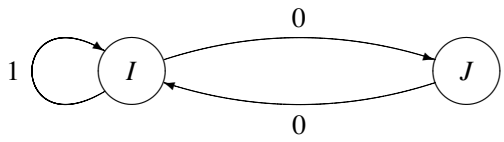
Thus, if τ is pre-injective, the same is true for $\tau \circ L$; by Lemma 2.16 we have $\mathrm{ent}(Y) = \mathrm{ent}(X) = \mathrm{ent}(\tau(X))$. By Theorem 2.14, $\tau(X)$ cannot be a proper subshift of the irreducible shift Y. Hence τ is surjective. □

Counterexample 2.18. There exists an irreducible sofic shift (not of finite type) for which the transition map is surjective but not pre-injective; this yields a counterexample to the Moore property.

Proof. Let X_e denote the *even shift*, that is the subshift of $\{0,1\}^{\mathbb{Z}}$ with forbidden words:

$$\{10^{2n+1}1 \mid n \geq 0\}.$$

The shift X_e is sofic; indeed, it is accepted by the following labelled graph:



Consider the function $f : S^5 \to S$ defined as

$$f(a_1a_2a_3a_4a_5) = \begin{cases} 1 & \text{if } a_1a_2a_3 = 000 \text{ or } a_1a_2a_3 = 111 \text{ or } a_2a_3a_4 = 010\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

and denote by $\tau : X_e \to S^{\mathbb{Z}}$ the induced local map, i.e.,

$$\tau[c](x) = f[c(x-2), c(x-1), c(x), c(x+1), c(x+2)].$$

We want to show that $\tau : X_e \to X_e$ is well-defined and that it is surjective but not pre-injective. We divide the proof into several steps. By abuse of notation, for a word (often called also *block*) $w = a_1a_2 \dots a_k$, with $k \geq 5$ we set

$$\tau(w) = f(a_1a_2a_3a_4a_5)f(a_2a_3a_4a_5a_6)\dots f(a_{k-4}a_{k-3}a_{k-2}a_{k-1}a_k).$$

Step 1. *If a block 0^n1 with $n \geq 3$ has a preimage under τ of length $n+5$ in the language of X_e, say*

a_1	a_2	a_3	a_4	…	a_{n+1}	a_{n+2}	a_{n+3}	a_{n+4}	a_{n+5}
		0	0	…	0	0	1		

,

then this preimage is necessarily of one of the forms

1. (i) $a_1a_2\,xx\,(1-x)(1-x)\dots 11\;00\;11\;000a_{n+4}a_{n+5}$,
   (ii) $a_1a_2\,xx\,(1-x)(1-x)\dots 11\;00\;11\;00100$,

(iii) $a_1a_2\ (1-x)(1-x)\ xx\ldots 00\ 11\ 00\ 111a_{n+4}a_{n+5}$,

when n is even and for a suitable $x \in \{0, 1\}$*;*

2. (i) $a_1a_2\ (1-x)\ xx\ldots 11\ 00\ 11\ 000a_{n+4}a_{n+5}$,

(ii) $a_1a_2\ (1-x)\ xx\ldots 11\ 00\ 11\ 00100$,

(iii) $a_1a_2\ x\ (1-x)(1-x)\ldots 00\ 11\ 00\ 111a_{n+4}a_{n+5}$,

when n is odd and for a suitable $x \in \{0, 1\}$.

Proof of Step 1. The statement may be proved by induction on $n \geq 3$. Suppose $n = 3$. When $\tau(a_1a_2a_3a_4a_5a_6a_7a_8) = 0001$ we have three cases: if $a_4a_5a_6 = 000$ then $a_3 = 1$, if $a_4a_5a_6a_7a_8 = 00100$ then again $a_3 = 1$, and finally if $a_4a_5a_6 = 111$ then necessarily $a_3 = 0$.

Now suppose that the statement is true for n and that $\tau(a_1\ldots a_{n+6}) = 0^{n+1}1$: if n is even, by the inductive hypothesis one has either

$$a_4\ldots a_{n+4} = xx\ (1-x)(1-x)\ldots 11\ 000,$$

or

$$a_4\ldots a_{n+6} = xx\ (1-x)(1-x)\ldots 11\ 00100,$$

or

$$a_4\ldots a_{n+4} = (1-x)(1-x)\ xx\ldots 00\ 111$$

for a suitable $x \in \{0, 1\}$.

In any case we have $a_4 = a_5$. If $a_3 = a_4$, then $f(a_3a_4a_5a_6a_7) = f(a_4a_4a_4a_6a_7) = 1 \neq 0$. Thus $a_3 \neq a_4$. Then the claim follows by the inductive hypothesis. The case n odd may be proved in a similar way. □

Step 2. *The map* τ *is an endomorphism of* X_e*, that is* $\tau(X_e) \subseteq X_e$.

Proof of Step 2. It suffices to prove that no forbidden word 10^n1 with n odd, has a preimage of length $n + 6$ in the language of X_e. First of all one has to check that there is no block $a_1a_2a_3a_4a_5a_6a_7$ of length 7 such that $\tau(a_1a_2a_3a_4a_5a_6a_7) = 101$, distinguishing two cases: $a_3a_4 = 00$ and $a_3a_4a_5 = 111$.

We now prove that no block $a_1\ldots a_{n+6}$ of length $n + 6$ has 10^n1 as image under τ, where $n \in \mathbb{N}$ is odd and strictly greater than 1. If $\tau(a_1\ldots a_{n+6}) = 10^n1$, then by previous step we have $a_4a_5a_6\ldots = x(1-x)(1-x)\ldots$, and being $f(a_1a_2a_3a_4a_5) = 1$, we distinguish two cases:

- $x = 0$. Then $a_3 = 0$ (otherwise we had a forbidden block) and $a_2 = 1$ because $f(a_2a_3a_4a_5a_6) = f(a_20011) = 0$. It follows that $f(a_1a_2a_3a_4a_5) = f(a_11001) = 0 \neq 1$.
- $x = 1$. If $a_3 = 0$ then $a_2 = 0$ and $f(a_2a_3a_4a_5a_6) = f(00100) = 1 \neq 0$. Thus $a_3 = 1$. Then $f(a_2a_3100) = f(a_21100)$ and $f(a_21100) = 0$ implies $a_2 = 0$.

Thus $f(a_1a_2a_310) = f(a_10110) = 0 \neq 1$. Hence 10^n1 has no preimage under τ. □

Step 3. *The transition map $\tau : X_e \to X_e$ is surjective.*

Proof of Step 3. By the compactness argument we alluded to before (also see Proposition 4.1) it suffices to prove the non-existence of Garden of Eden words (=patterns), and in our setting it suffices to prove that each block of the form $10^{n_1}10^{n_2}\dots10^{n_k}1$ (where $n_1, \dots, n_k$ are even integers), has a preimage-block.

Suppose first that $k = 1$:

a_1	a_2	a_3	a_4	a_5	…	a_{n+2}	a_{n+3}	a_{n+4}	a_{n+5}	a_{n+6}
		1	0	0	…	0	0	1		

,

we distinguish the following three cases in which $a_{n+4} \mapsto 1$.

- $n = 0$. Then one of the following 3 possibilities holds:

$$\begin{cases} a_1a_2a_3a_4 = 0000, \\ a_1a_2a_3a_4a_5a_6 = 000100, \\ a_1a_2a_3a_4 = 1111. \end{cases}$$

- $n = 2$. Then one of the following 3 possibilities holds:

$$\begin{cases} a_1a_2a_3a_4a_5a_6 = a_1a_11000, \\ a_1a_2a_3a_4a_5a_6a_7a_8 = a_1a_1100100, \\ a_1a_2a_3a_4a_5a_6 = 000111. \end{cases}$$

- $n \geq 4$. Then, for a suitable $x \in \{0, 1\}$,

$$\tau[(1-x)(1-x)(1-x)xx\dots000a_{n+5}a_{n+6}] = 10^n1.$$

Similarly,

$$\tau[(1-x)(1-x)(1-x)xx\dots00100] = 10^n1,$$

and, finally,

$$\tau[xxx(1-x)(1-x)\dots111a_{n+5}a_{n+6}] = 10^n1.$$

Now, given any word of the form $10^{n_1}10^{n_2}\dots10^{n_k}1$ we can construct a preimage starting from the rightmost block $10n_k1$: over the first on the right 1 we can write arbitrarily $000**$, $111**$ or 00100. In this way we get a word $a_1a_2a_3a_4a_5$ over the second on the left 1, and we can start from this word over 1 to construct a preimage for the second on the right block 10^{n_k-1}, and so on. In each of the possible choices we can find a block whose image under τ is our fixed word. □

Step 4. *The transition map is not pre-injective.*

Proof of Step 4. Consider the configuration c_1:

…	0	0	0	0	0	**1**	**0**	**0**	1	0	0	0	0	0	…

and the configuration c_2:

…	0	0	0	0	0	**0**	**1**	**1**	1	0	0	0	0	0	…

.

These configurations differ only on a finite subset of $\mathbb{Z}$, but they have the same image under τ, namely the configuration

…	1	1	1	1	1	1	0	0	1	0	0	1	1	1	…

,

so that τ is not pre-injective. □

3 Gromov's theorem

This section is devoted to the GOE type theorem of Gromov [25] which we have stated, in a slightly generalized form, as Theorem B in the introduction: all new definitions (with the exception of *strong-irreducibility* and the *finite type* condition) are contained in Gromov's paper [25]. We analyze all these concepts and illustrate them with several examples, counterexamples and remarks (in particular, Lemma 3.11 and the related Counterexample 3.12 are from [12, 14]).

3.1 Pseudogroups of partial isometries of a graph

Let Δ be a simple, infinite countable connected (undirected) graph of bounded valency; i.e. Δ has no loops or multiple edges, each pair of vertices is connected with a path, and there is a positive integer d such that Δ has at most d edges at each vertex. We will not distinguish between the graph and the set of its vertices, that will be denoted Δ. As usual, define the path distance on Δ by setting $\mathrm{dist}(\delta, \delta')$ as the minimal length of a path of edges joining δ and δ'; the ball of radius r and center δ will be denoted $D(\delta, r)$, that is, $D(\delta, r) = \{\delta' \in \Delta \mid \mathrm{dist}(\delta, \delta') \leq r\}$. In general Δ does not have isometries; thus we put our attention on the *partial isometries* of the graph. A partial isometry γ is a bijective map between two subsets Ω and Ω' that preserves the metric dist. Δ has many partial isometries: since it is simple and of bounded valency, for a positive integer r there are at most finitely many isometry classes of balls of radius r. A set Γ of partial isometries of Δ will be called *a pseudogroup of partial isometries acting on* Δ if it satisfies the following four axioms:

(A) Γ contains the identity map $\mathrm{Id}_\Delta : \Delta \to \Delta$,

(B) $\gamma \in \Gamma \Rightarrow \gamma^{-1} \in \Gamma$,

(C) If $\gamma : \Omega \to \Omega'$ and $\gamma' : \Omega' \to \Omega''$ are in Γ, then $\gamma' \circ \gamma : \Omega \to \Omega''$ is in Γ,

(D) For every $\gamma \in \Gamma$, $\gamma : \Omega \to \Omega'$, its restriction $\gamma : \Omega_0 \to \gamma(\Omega_0)$ is also in Γ for all $\Omega_0 \subseteq \Omega$.

Clearly, the set of all partial isometries is a pseudogroup. A pseudogroup Γ is *cofinite* on Δ if, for every $r = 0, 1, 2, \ldots$, there are at most finitely many mutually non-Γ-isometric balls of radius r. Since Δ is simple and of bounded valency, the pseudogroup of all partial isometries is cofinite. We will also say that two points δ and δ' are *r-equivalent* with respect to Γ if the r-balls centered at these points are Γ-isometric. Γ is *dense on* Δ if, for every $r = 0, 1, 2, \ldots$, each *non-empty* class Δ' of r-equivalent points forms a net in Δ, that is there exists an $R = R(\Delta')$ such that Δ' meets every ball of radius R in Δ (equivalently, $\bigcup_{\delta' \in \Delta'} D(\delta', R) = \Delta$).

3.2 Subproduct systems on Δ

Let now S be a finite or countable alphabet. An *S-valued subproduct system on* Δ, denoted by $\{X_\Omega\}$, consists of an assignment to each *finite* subset Ω of Δ of a *finite* set X_Ω of S-valued functions defined on Ω yielding a *projective system* with respect to inclusion of finite subsets; this means that, if $\Sigma \subseteq \Omega$ is an arbitrary finite subset of Ω and $x \in X_\Omega$, then the restriction $x|_\Sigma$ of x to Σ belongs to X_Σ. The term "subproduct" comes form the fact that one might regard each X_Ω as a subset of the cartesian product $\times_{\delta \in \Omega} X_{\{\delta\}}$. *The projective limit* $X = \varprojlim X_\Omega$ of a subproduct system $\{X_\Omega\}$ is the set of all functions $x : \Delta \to S$ such that the restriction $x|_\Omega$ of x to Ω belongs to X_Ω for every finite $\Omega \subset \Delta$. An *exhaustion of Δ by finite subsets* is a sequence $\Omega_1 \subset \Omega_2 \subset \Omega_3 \subset \cdots \subset \Omega_n \subset \cdots$ of finite subsets such that $\bigcup_{n=1}^\infty \Omega_n = \Delta$. It is not difficult to prove that x belongs to the projective limit of $\{X_\Omega\}$ if and only if the restriction of x to Ω_n belongs to X_{Ω_n} for all $n = 1, 2, \ldots$ The projective limits just defined are always non-empty:

Proposition 3.1. *The projective limit of a subproduct system $\{X_\Omega\}$ is non empty.*

Proof. Let $\Omega_1 \subset \Omega_2 \subset \Omega_3 \subset \cdots \subset \Omega_n \subset \cdots$ be an exhaustion of Δ by finite subsets. For $j \geq i$, let X_i^j be the set of the restrictions to Ω_i of the functions in X_{Ω_j}. Then the intersection $X_i^\infty = \bigcap_{j=i}^\infty X_i^j$ is non empty. In fact, this is the intersection of a decreasing ($X_i^j \geq X_i^{j+1}$) sequence of non empty finite subsets. Now denote by $\pi_{i+1,i}$ the projection (restriction of functions) from $X_{\Omega_{i+1}}$ to X_{Ω_i}. We claim that $\pi_{i+1,i}$ is *onto* from X_{i+1}^∞ to X_i^∞. In fact, if $x \in X_i^\infty$ and $j \geq i+1$, then $x \in X_i^j$, that is, there exists $x_j \in X_{\Omega_j}$ such that $x = x_j|_{\Omega_i}$. Setting $x' = x_j|_{\Omega_{i+1}}$ we have $x = x'|_{\Omega_i}$ so that

$x' \in \pi_{i+1,i}^{-1}(x) \cap X_{i+1}^{j} \neq \emptyset$. Then, as before,

$$\pi_{i+1,i}^{-1}(x) \cap X_{i+1}^{\infty} = \bigcap_{j=i}^{\infty} \left(\pi_{i+1,i}^{-1}(x) \cap X_{i+1}^{j}\right) \neq \emptyset.$$

Now we can construct an element of the projective limit of $\{X_\Omega\}$: choose a function of X_1^∞, it may be extended to a function of X_2^∞, that may be extended to a function of X_3^∞ and so on, obtaining a function on all Δ. □

Remark 3.2. Observe that the local finiteness of the projective system $\{X_\Omega\}$ is essential. For instance, if $S = \{0, 1, 2, \dots\}$ and $\Omega_1 \subset \Omega_2 \subset \Omega_3 \subset \cdots \subset \Omega_n \subset \cdots$ is an exhaustion of Δ by finite subsets, setting

$$X_{\Omega_i} = \left\{x : \Omega_i \to \{i, i+1, i+2, \dots\}\right\}$$

we clearly have that $\varprojlim X_{\Omega_i}$ is empty.

Now let X be a set of S-valued functions defined on Δ. X is *locally-finite* if the set $X_{\{\delta\}} = \{x(\delta) : x \in X\}$ is finite for every $\delta \in \Delta$. If, in addition, $\sup\{|X_{\{\delta\}}| : \delta \in \Delta\} < \infty$, then X is *uniformly*-locally-finite. Clearly, if S is finite, every $X \subseteq S^\Delta$ is (uniformly) locally-finite. Now consider a locally-finite set of S-valued functions defined on Δ. For every finite subset Ω of Δ let X_Ω be the space of the restrictions of the functions of X to Ω. Clearly, $\{X_\Omega\}$ is a subproduct system; we call it *the subproduct system generated by* X.

We now present two examples that show what may happen.

Example 3.3. Let $S = \{0, 1\}$, and denote by X the set of S-valued functions on Δ defined by the following condition (finite support): $x : \Delta \to S$ is in X if and only if there exists a finite subset $\Sigma = \Sigma(x) \subset \Delta$ such that $x(\delta) = 0$ for every $\delta \notin \Sigma$. Then if $\{X_\Omega\}$ is the subproduct generated by X, for any Ω, X_Ω is equal to the set of all S-valued functions defined on Ω; consequently the projective limit X' of $\{X_\Omega\}$ is equal to the space of *all* S-valued functions defined on Δ. In particular, $X \subsetneq X'$.

Example 3.4. Let Δ be the usual Cayley graph of the integers (the bi-infinite line graph). If Ω is a (finite) subset of Δ we say that three consecutive integers $i-1, i, i+1$ contained in Ω are in the interior of Ω if $i-2$ and $i+2$ are in Ω as well. Let $S = \{0, 1\}$, and define the subproduct system $\{X_\Omega\}$ by the following condition: $x : \Omega \to S$ is in X_Ω if and only if $x(i-1) = x(i+1)$ and $x(i) \neq x(i-1)$ whenever $i-1$, i, $i+1$ are in the interior of Ω. This means that $x : \Omega \to S$ is in X_Ω if and only it has period 2 except (possibly) at the boundary of Ω. Then the projective limit X' of $\{X_\Omega\}$ consists of two functions with period 2. Now, denoting by $\{X'_\Omega\}$ the subproduct system generated by X', one has that $x : \Omega \to S$ is in X'_Ω if and only if it has period 2 on the *whole* of Ω; this shows that, for all finite $\Omega \subset \Delta$ with non-empty interior, the strict inequality $X'_\Omega \subsetneq X_\Omega$ holds.

These examples suggest the following definition, which is not considered, at least in this form, in [25]:

Definition 3.5. A subproduct system $\{X_\Omega\}$ is *stable* when it is equal to the subproduct system generated by its projective limit. Analogously, we will say that a locally-finite set X of S-valued functions is a *stable space* if it is the projective limit of the subproduct system generated by it.

Remark 3.6. Let X be a set of S-valued functions on Δ, and denote by $\{X_\Omega\}$ the projective system induced by the restrictions. If $X' = \varprojlim X_\Omega$, then $X \subseteq X'$. Thus, $\{X_\Omega\}$ is stable. On the other side, starting from a projective system $\{X_\Omega\}$, and denoting by $\{X'_\Omega\}$ the projective system induced by the projective limit $X' = \varprojlim X_\Omega$, we have $X'_\Omega \subseteq X_\Omega$, and therefore X' is stable. In other words, the projective limit of any projective system or any induced subproduct system are stable.

Our definition has a simple topological interpretation. Endowing the space S^Δ of all S-valued functions with the product topology (S is a discrete countable space), a subspace X of S^Δ is closed (in fact compact, by local finiteness) if and only if it is stable.

Gromov gives local conditions of stability [25], slightly different from ours; our definition seems more natural since it also ensures the following fundamental properties: if $\{X_\Omega\}$ is stable and X is its projective limit, then the projection (restriction) from X to each X_Ω is onto, and, if $\Omega_0 \subseteq \Omega$ then the projection $X_\Omega \to X_{\Omega_0}$ is also onto.

3.3 Irreducibility conditions for subproduct systems

Definition 3.7. A subproduct system $\{X_\Omega\}$ has *propagation* $\leq \ell$ if it satisfies the following condition: $x : \Omega \to S$ belongs to X_Ω if (and only if) the restrictions of x to the intersections $\Omega \cap D(\delta, \ell)$ are contained in $X_{\Omega \cap D(\delta,\ell)}$ for all $\delta \in \Omega$.

For two subsets $\Omega, \Omega' \subseteq \Delta$ set $\mathrm{dist}(\Omega, \Omega') = \min\{\mathrm{dist}(\delta, \delta') : \delta \in \Omega, \delta' \in \Omega'\}$.

Definition 3.8. A subproduct system $\{X_\Omega\}$ and its projective limit X are *M-irreducible* if, for each pair of finite sets $\Omega, \Omega' \subset \Delta$ such that $\mathrm{dist}(\Omega, \Omega') > M$ and for each $x \in X_\Omega$ and $y \in X_{\Omega'}$, the function $z : \Omega \cup \Omega' \to S$ that equals x on Ω and y on Ω' belongs to $X_{\Omega \cup \Omega'}$, equivalently $X_{\Omega \cup \Omega'} = X_\Omega \times X_{\Omega'}$. X is *strongly-irreducible* if it is M-irreducible for some $M > 0$.

Definition 3.9. A stable subproduct system $\{X_\Omega\}$ has *memory* M if it satisfies the following condition: $x : \Delta \to S$ is in the projective limit $X = \varprojlim X_\Omega$ if (and only if) the restriction of x to $D(\delta, M)$ is in $X_{D(\delta,M)}$ for any $\delta \in \Delta$. We say that it is *of finite type* if it has memory M for some $M > 0$.

Remark 3.10. Let X be a (stable) subproduct system. It is obvious that if it has propagation $\leq \ell$ then it also has propagation $\leq \ell + 1$, etc.: one says simply that it has *bounded propagation*. The same holds for strong-irreducibility and for the memory. Thus, in the rest of this section, when thinking of a strongly-irreducible space X of bounded propagation and of finite type, it is not restrictive to suppose that, for a *common* ℓ, the space X is ℓ-irreducible, with propagation $\leq \ell$ and memory ℓ.

Lemma 3.11. *A stable space of bounded propagation is strongly irreducible and of finite type.*

Proof. Let X be a system of propagation $\leq \ell$. Suppose that Ω and Ω' are finite subsets of Δ such that $\mathrm{dist}(\Omega, \Omega') > \ell$. A ball of radius ℓ centered in a point in $\Omega \cup \Omega'$ cannot intersect both Ω and Ω'. From this the ℓ-irreducibility follows immediately. Now suppose that the restriction of $x : \Delta \to S$ to $D(\delta, \ell)$ is in $X_{D(\delta,\ell)}$ for any $\delta \in \Delta$. Then for any finite set $\Omega \subseteq \Delta$ and for any $\delta \in \Omega$ the restriction of x to $\Omega \cap D(\delta, \ell)$ is in $X_{\Omega \cap D(\delta,\ell)}$, so that, by the bounded propagation property, the restriction of x to Ω belongs to X_Ω and, in virtue of the stability condition, $x \in X$. This latter shows that X is of finite type with memory ℓ. □

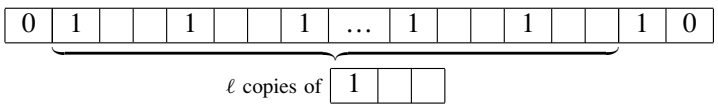
For a partial converse of the above statement in the 1-dimensional setting see [12, Proposition 3.5.11]. In general such a converse does not hold as the following counterexample shows.

Counterexample 3.12. Strong irreducibility and finite type conditions do not imply, in general, bounded propagation.

Proof. Let X be the subshift of $\{0, 1\}^{\mathbb{Z}}$ determined by the following set of forbidden blocks:

$$\{010, 111\}.$$

It is easy to show that X is a strongly-irreducible (in fact 3-irreducible) shift of finite type; for any $\ell \geq 1$ consider the following pattern p whose support F consists only of the boxes containing a digit.



We have $p|_{F \cap D(\alpha,\ell)} \in X_{F \cap D(\alpha,\ell)}$ but $p \notin X_F$: indeed, one can *locally* insert zeroes in the empty boxes yielding admissible words (or patterns); this is no more possible *globally*, i.e., for all the $3\ell + 3$ boxes.

Hence X is not of bounded propagation $\leq \ell$ for each $\ell \geq 1$. □

3.4 Maps of bounded propagation

If Ω is a finite subset of Δ define $\Omega^{-\ell}=\{\delta \in \Delta : D(\delta, \ell)$ is contained in $\Omega\}$ as the *ℓ-interior* of Ω. Let $\{X_\Omega\}$ and $\{Y_\Omega\}$ be S-valued subproduct systems on Δ (if $\{X_\Omega\}$ and $\{Y_\Omega\}$ have different alphabets, say S_X and S_Y, we may take the union $S = S_X \cup S_Y$); a *morphism of bounded propagation* $\leq \ell$ between the two subproduct systems consists of a set of functions $\tau_\Omega : X_\Omega \to Y_{\Omega^{-\ell}}$, denoted by $\{\tau_\Omega\}$, commuting with the (respective) restrictions. Clearly, $\{\tau_\Omega\}$ gives rise to a function τ between the projective limits X of X_Ω and Y of Y_Ω: if $x \in X$ and $\delta \in \Delta$, then $y = \tau(x)$ is defined on δ by: $y(\delta) = \tau_{D(\delta,\ell)}[x|_{D(\delta,\ell)}](\delta)$, where $x|_{D(\delta,\ell)}$ denotes, as usual, the restriction of x to $D(\delta, \ell)$. In other words, for every $\delta \in \Delta$ we have a map from $X_{D(\delta,\ell)}$ to $Y_{\{\delta\}}$ and the set of all these maps determines both $\{\tau_\Omega\}$ and τ.

3.5 Holonomy maps

Let $\{X_\Omega\}$ be a subproduct system. Fix two balls D and D' in Δ. A *holonomy map* h between the projective systems $\{X_\Omega\}_{\Omega\subseteq D}$ and $\{X_{\Omega'}\}_{\Omega'\subseteq D'}$ consists of the following data:

(i) an isometry $\gamma = \gamma_h : D \to D'$ sending the center of D to that of D',

(ii) a set of bijective maps $h_\Omega : X_\Omega \to X_{\gamma(\Omega)}$ for all $\Omega \subset D$ which commute with the restriction maps.

A *holonomy over* Δ is defined as a set H of holonomy maps defined between certain pairs of balls D and D'. The balls admitting holonomies between them are called *equivalent*. According to Gromov [25], we say that H is a *pseudogroup of holonomies* if it satisfies the following axioms:

(1) The identity Id_D, given by the identity map $D \to D$ and the identity map $X_D \to X_D$ is in H.

(2) $h \in H \Rightarrow h^{-1} \in H$.

(3) If h and h' are in H, h is defined between D and D' and h' is between D' and D'', then their composition $h' \circ h$ between D and D'' is also in H.

(4) If a ball D_0 is contained in D then the obvious restriction of each holonomy from D to D_0 belongs to the holonomy over D_0 (where D_0 is not necessarily a concentric ball).

If H is a pseudogroup of holonomies we denote by $\Gamma(H)$ the associated pseudogroup of partial isometries: $\Gamma(H) = \{\gamma_h : h \in H\}$. H is called *cofinite* or *dense* whenever $\Gamma(H)$ is such. Given a morphism $f = \{f_\Omega\}$ of bounded propagation $\leq \ell$ between two projective systems over Δ we shall consider holonomies that *commute*

with $\{f_\Omega\}$. This corresponds to the notion of G-equivariance we alluded to in the introduction when Δ is the Cayley graph of a finitely generated group.

3.6 Entropy

Let X be a set of S-valued functions defined on Δ, and $\Omega_n \subset \Delta$, $n = 1, 2, \ldots$, be a sequence of finite subsets. Denote as usual by X_{Ω_n} the set of the restrictions of the functions of X to Ω_n and by $|A|$ the cardinality of a set A. Then we define *the entropy of X with respect to* $\{\Omega_n\}$ by setting:

$$\mathrm{ent}(X) = \mathrm{ent}(X : \{\Omega_n\}) = \liminf_{n\to\infty} \frac{\log |X_{\Omega_n}|}{|\Omega_n|}. \tag{3.1}$$

Clearly, the entropy is monotone for inclusion: if $X' \subset X$ then $\mathrm{ent}(X') \leq \mathrm{ent}(X)$. Note that if S is finite and X is the space of all S-valued functions on Δ, then $|X_\Omega| = |S|^{|\Omega|}$ and therefore $\mathrm{ent}(X) = \log |S|$.

3.7 Amenability

If Ω is a subset of Δ, its *ℓ-boundary* $\partial_\ell \Omega$ is the set of all $\delta \in \Delta$ such that the ball $D(\delta, \ell)$ intersects both Ω and $\Delta \setminus \Omega$. The 1-boundary will be denoted simply by $\partial\Omega$ and called *the boundary*. We also set $\Omega^{+\ell} = \Omega \cup \partial_\ell \Omega$. The proof of the following proposition is trivial:

Proposition 3.13. *Let Δ a graph and $\Omega \subseteq \Delta$ a (not necessarily finite) subset.*

(i) $\partial_\ell \Omega \subseteq \bigcup_{\delta\in\partial\Omega} D(\delta, \ell)$;

(ii) $\Omega^{+\ell} = \bigcup_{\delta\in\Omega} D(\delta, \ell)$;

(iii) *if $a = \max\{|(D(\delta, \ell)| : \delta \in \Delta\}$ and Ω is finite then $|\partial_\ell \Omega| \leq a \cdot |\Omega|$.* □

A sequence $\Omega_n \subset \Delta$, $n = 1, 2, \ldots$ of finite subsets is called *amenable* if

$$\lim_{n\to\infty} \frac{|\partial\Omega_n|}{|\Omega_n|} = 0.$$

Clearly, by Proposition 3.13, if $\{\Omega_n\}$ is amenable we have also $\lim_{n\to\infty} \frac{|\partial_\ell \Omega_n|}{|\Omega_n|} = 0$ for every positive integer ℓ. (The converse is also true since $\partial_\ell \Omega \supseteq \partial\Omega$).

The graph Δ is *amenable* if it admits an exhaustion by finite subsets $\Omega_1 \subset \Omega_2 \subset \Omega_3 \subset \cdots \subset \Omega_n \subset \ldots$ that is amenable.

From now on, as the graphs Δ we shall deal with, are always amenable, when referring to the entropy of a stable space $X \subseteq S^\Delta$, we shall assume that an exhausting

amenable sequence $\{\Omega_n\} \subseteq \Delta$ has been fixed once for all. Thus we shall simply denote the entropy by $\mathrm{ent}(X)$ instead of $\mathrm{ent}(X : \{\Omega_n\})$.

The following is an analogue of Proposition 2.4.

Proposition 3.14. *Suppose that Δ is an amenable graph and that X and Y are uniformly-locally-finite sets of S-valued functions defined on Δ. If a map of bounded propagation $\tau : X \to Y$ is surjective then* $\mathrm{ent}(Y) \leq \mathrm{ent}(X)$. *Equivalently, one has* $\mathrm{ent}(\tau(X)) \leq \mathrm{ent}(X)$ *for all maps τ of bounded propagation* (*not necessarily surjective*).

Proof. Since $\Omega_n \subseteq (\Omega_n^{+\ell})^{-\ell}$, by the surjectivity of τ, the cardinality of Y_{Ω_n} does not exceed that of $X_{\Omega_n^{+\ell}}$. Setting $b = \sup\{|X_{\{\delta\}}| : \delta \in \Delta\}$, one has $|X_{\Omega_n^{+\ell}}| \leq |X_{\Omega_n}| \cdot b^{|\partial_\ell \Omega_n|}$; thus

$$\frac{\log |Y_{\Omega_n}|}{|\Omega_n|} \leq \frac{\log |X_{\Omega_n^{+\ell}}|}{|\Omega_n|} \leq \frac{\log |X_{\Omega_n}| + \log(b) \cdot |\partial_\ell \Omega_n|}{|\Omega_n|},$$

and taking the lim inf we get the desired inequality. □

3.8 Spliceable spaces

Given $\Omega \subset \Delta$ and two S-valued functions x and x' on Δ, define their *splice over* Ω as the function x which is equal to x on Ω and to x' outside Ω. A space X of S-valued functions on Δ is *ℓ-spliceable* if the conditions $x, x' \in X$ and $x = x'$ on $\partial_\ell \Omega$ imply that their splice x over Ω belongs to X whenever Ω is a finite subset of Δ.

Example 3.15. Let S be any countable set and let Δ be the bi-infinite line graph. Let $X(\ell)$ denote the set of all ℓ-periodic functions. Then $X(\ell)$ is $[\frac{\ell+1}{2}]$-spliceable.

Proposition 3.16. *A stable space X of finite type with memory ℓ is 2ℓ-spliceable.*

Proof. Let Ω be a finite subset of Δ, $x, x' \in X$ such that $x = x'$ on $\partial_{2\ell}\Omega$ and x their splice over Ω. Then for every ball D of radius ℓ, the restriction of x to D is equal either to the restriction of x (when $D \subseteq \Omega^{+2\ell}$) or to the restriction of x' (when $D \cap \Omega = \emptyset$), thus in any case it belongs to X_D and since X has memory ℓ we have that $x \in X$. □

3.9 Pre-injectivity

A map $\tau : X \to Y$ between two spaces X and Y of S-valued functions on Δ is *pre-injective* if whenever $x, x' \in X$ are such that $x \neq x'$ in a finite non-empty subset $\Omega \subset \Delta$ and $x = x'$ outside Ω, then $\tau(x) \neq \tau(x')$.

The following corresponds to one implication in the statement of Theorem 2.8 (the Hedlund–Coven–Paul theorem):

Proposition 3.17. *Let Δ be an amenable graph, and let X be a uniformly-locally-finite irreducible stable space of finite type. If $\tau : X \to S^{\Delta}$ is a pre-injective map of bounded propagation, then* $\mathrm{ent}(\tau(X)) = \mathrm{ent}(X)$.

Proof. Denote by $Y = \tau(X)$ the image of X, and by $\{Y_\Omega\}$ the corresponding sub-product system. Set $b = \sup\{|X_{\{\delta\}}| : \delta \in \Delta\}$ and $c = \sup\{|Y_{\{\delta\}}| : \delta \in \Delta\}$. Since $|Y_{\Omega_n^{+2\ell}}| \le |Y_{\Omega_n}| \cdot c^{|\partial_{2\ell}\Omega_n|}$, $|X_{\Omega_n}| \le |X_{\Omega_n^{-2\ell}}| \cdot b^{|\Omega_n \setminus \Omega_n^{-2\ell}|}$, $|\partial_{2\ell}\Omega_n|/|\Omega_n| \to 0$, and

$$|\Omega_n \setminus \Omega_n^{-2\ell}|/|\Omega_n| \le (|\partial\Omega_n| \cdot \max\{|D(\delta, 2\ell)| : \delta \in \Delta\})/|\Omega_n| \to 0 \quad \text{as } n \to \infty,$$

we have

$$\frac{\log |Y_{\Omega_n^{+2\ell}}|}{\Omega_n} \le \frac{\log |Y_{\Omega_n}|}{\Omega_n} + \frac{\partial_{2\ell}\Omega_n|}{|\Omega_n|} \log c$$

and

$$\frac{\log |X_{\Omega_n}|}{\Omega_n} \le \frac{\log |X_{\Omega_n^{-2\ell}}|}{\Omega_n} + \frac{|\Omega_n \setminus \Omega_n^{-2\ell}|}{|\Omega_n|} \log b.$$

Suppose, by contradiction, that $\mathrm{ent}(Y) < \mathrm{ent}(X)$. Then taking the lim inf in the above formulas we get:

$$\liminf_{n\to\infty} \frac{\log |Y_{\Omega_n^{+2\ell}}|}{|\Omega_n|} < \liminf_{n\to\infty} \frac{\log |X_{\Omega_n^{-2\ell}}|}{|\Omega_n|},$$

and we can find n such that:

$$|Y_{\Omega_n^{+2\ell}}| < |X_{\Omega_n^{-2\ell}}|.$$

Now take $z \in X_{\Omega_n^{+2\ell} \setminus \Omega_n}$; then for every $y \in X_{\Omega_n^{-2\ell}}$ by the strong irreducibility there exists $x' \in X_{\Omega_n^{-2\ell} \cup (\Omega_n^{+2\ell} \setminus \Omega_n)}$ that is equal to z on $\Omega_n^{+2\ell} \setminus \Omega_n$ and to y on $\Omega_n^{-2\ell}$. Using the stability of X, x' may be extended to a function $x \in X_{\Omega_n^{+2\ell}}$. Thus,

$$|\{x \in X_{\Omega_n^{+2\ell}} : x = z \text{ on } \Omega_n^{+2\ell} \setminus \Omega_n\}| \ge |Y_{(\Omega_n^{+2\ell})^{-l}}|.$$

It follows that there exist $x_0, x_1 \in X_{\Omega_n^{+2\ell}}$ such that: $x_0 \ne x_1$, $x_0 = x_1 = z$ on $\Omega_n^{+2\ell} \setminus \Omega_n$ and $\tau_{\Omega_n^{+2\ell}}(x_0) = \tau_{\Omega_n^{+2\ell}}(x_1)$. Then, using the stability of X, we can extend x_0 and x_1 to functions $\tilde{x}_0, \tilde{x}_1$ on all Δ. By Proposition 3.16 X is 2ℓ-spliceable; thus the splice $\tilde{x}$ of $\tilde{x}_0$ and $\tilde{x}_1$ over Ω_n belongs to X. But now $\tilde{x} \ne \tilde{x}_1$ on Ω_n, $\tilde{x} = \tilde{x}_1$ outside Ω_n and $\tau(\tilde{x}) = \tau(\tilde{x}_1)$, since τ has bounded propagation $\le \ell$ and a ball of radius ℓ cannot intersect both Ω_n and $\Delta \setminus \Omega_n^{+2\ell}$. This makes τ *not* pre-injective and proves the proposition. □

3.10 Strict monotonicity

Now we prove the following fundamental analogue of Theorems 2.2, 2.7 and 2.14.

Lemma 3.18. *Let Δ be an amenable graph and let $X \subseteq S^\Delta$ be a uniformly locally-finite irreducible stable space. Suppose a set $\{D_j : j = 1, 2, \dots\}$ of balls of radius ρ constitutes a net in Δ (i.e., some R-neighbourhood of their union equals the whole of Δ). If X' is a subspace of X such that X'_{D_j} is strictly smaller than X_{D_j} on every ball D_j: $|X'_{D_j}| < |X_{D_j}|$, then*

$$\mathrm{ent}(X') < \mathrm{ent}(X).$$

Proof. . We may assume (throwing away some balls if necessary) that the mutual distances between the balls are large, say $\geq 10\ell$. Consider a large Ω_n and suppose that $D_{j_1}, D_{j_2}, \dots, D_{j_N}$, where $N = N(n)$, are *all* the balls that are contained in $\Omega_n^{-2\ell}$. Set

$$\alpha = \left(\max\{|X_{D(\delta,\rho+2\ell)}| : \delta \in \Delta\}\right)^{-1}.$$

We prove by induction that

$$|X'_{\Omega_n}| \leq (1-\alpha)^{N(n)} |X_{\Omega_n}|. \tag{3.2}$$

For $i = 1, 2, \dots N$ choose $x_i \in X_{D_{j_i}} \setminus X'_{D_{j_i}}$ and denote by π_i the projection from X_{Ω_n} to $X_{D_{j_i}}$. Setting $Y = X_{\Omega_n \setminus D_{j_1}^{+2\ell}}$, we have $|Y| \geq \alpha \cdot |X_{\Omega_n}|$ and $|\pi_1^{-1}(x_1)| \geq |Y|$; the last inequality follows from the strong irreducibility of X: the distance between D_{j_1} and $\Omega_n \setminus D_{j_1}^{+2\ell}$ is $> 2\ell$, thus every $y \in Y$ may be extended to a function on Ω_n which is equal to x_1 on D_{j_1}. It follows that $|X_{\Omega_n} \setminus \pi_1^{-1}(x_1)| \leq (1-\alpha)|X_{\Omega_n}|$. Now suppose (inductive hypothesis) that

$$\left| X_{\Omega_n} \setminus \bigcup_{k=1}^{N-1} \pi_k^{-1}(x_k) \right| \leq (1-\alpha)^{N-1} |X_{\Omega_n}|.$$

Define $\overline{X}_{\Omega_n} = X_{\Omega_n} \setminus \bigcup_{k=1}^{N-1} \pi_k^{-1}(x_k)$, $B = \{z \in \overline{X}_{\Omega_n} : z = x_N \text{ on } D_{j_N}\}$ and denote by $\overline{Y}$ the set of the restrictions of the functions of $\overline{X}_{\Omega_n}$ to $\Omega_n \setminus D_{j_N}^{+2\ell}$. By strong irreducibility, given any $y \in \overline{Y}$ we can find $z \in X_{\Omega_n}$ such that $z = y$ on $\Omega_n \setminus D_{j_N}^{+2\ell}$ and $z = x_N$ on D_{j_N}. But for $k = 1, 2, \dots, N-1$, $z = y \neq x_{j_k}$ on D_{j_k}: thus, $z \in \overline{X}_{\Omega_n}$ and so $z \in B$. Therefore, we have again: $|B| \geq |\overline{Y}| \geq \alpha \cdot |\overline{X}_{\Omega_n}|$, so that, being $B \subset \pi_N^{-1}(x_N)$, one has

$$\begin{aligned}\left| X_{\Omega_n} \setminus \bigcup_{k=1}^{N} \pi_k^{-1}(x_k) \right| &= \left| (X_{\Omega_n} \setminus \bigcup_{k=1}^{N-1} \pi_k^{-1}(x_k)) \setminus \pi_N^{-1}(x_N) \right| = |\overline{X}_{\Omega_n} \setminus \pi_N^{-1}(x_N)| \\ &\leq |\overline{X}_{\Omega_n} \setminus B| \leq (1-\alpha)|\overline{X}_{\Omega_n}| \leq (1-\alpha)^N |X_{\Omega_n}|.\end{aligned}$$

Thus (3.2) is proved:

$$|X'_{\Omega_n}| \leq \left| X_{\Omega_n} \setminus \bigcup_{k=1}^{N} \pi_k^{-1}(x_k) \right| \leq (1-\alpha)^N |X_{\Omega_n}|.$$

Set now $\beta = \max\{|D(\delta, R+\rho+\ell)| : \delta \in \Delta\}$. Since all the balls $D_{j_1}, \ldots, D_{j_N}$ are contained in $\Omega_n^{-2\ell}$, we have:

$$\Omega_n \subset \bigcup_{k=1}^{N(n)} D_{j_k}^{+R} \cup (\Omega_n \setminus \Omega_n^{-(R+2\rho+2\ell)}),$$

thus

$$|\Omega_n| \leq N(n) \cdot \beta + |\Omega_n \setminus \Omega_n^{-(R+2\rho+2\ell)}|,$$

and, from the amenability of $\{\Omega_n\}$, it follows that

$$\liminf_{n\to\infty} \frac{N(n)}{|\Omega_n|} \geq \frac{1}{\beta} > 0.$$

Then from (3.2) we have:

$$\frac{\log |X'_{\Omega_n}|}{|\Omega_n|} - \frac{N(n)}{|\Omega_n|} \cdot \log(1-\alpha) \leq \frac{\log |X_{\Omega_n}|}{|\Omega_n|},$$

and taking the lim inf we obtain:

$$\operatorname{ent}(X') < \operatorname{ent}(X') - \frac{1}{\beta} \log(1-\alpha) \leq \operatorname{ent}(X). \qquad \square$$

3.11 Pre-injectivity corollary

The following corresponds to the other implication (see Proposition 3.17 for the first one) of the Hedlund–Coven–Paul theorem (Theorem 2.8).

Proposition 3.19. *Let X be a strongly-irreducible stable space of finite type. Suppose that $\tau : X \to S^{\Delta}$ is a map of bounded propagation admitting a dense pseudogroup of holonomies. Then the equality* $\operatorname{ent}(\tau(X)) = \operatorname{ent}(X)$ *implies that τ is pre-injective.*

Proof. If τ is not pre-injective, there exist x and x' in X and a ball D such that: $x \neq x'$ on D, $x = x'$ outside D and $\tau(x) = \tau(x')$. Using the density of the holonomy pseudogroup we can form a net of balls $\{D_j^{+2\ell}\}$ with corresponding functions x_j and x'_j in $X_{D_j^{+2\ell}}$ such that: $x_j \neq x'_j$ on D_j, $x_j = x'_j$ on $D_j^{+2\ell} \setminus D_j$ and $\tau_{D_j^{+2\ell}}(x_j) = \tau_{D_j^{+2\ell}}(x'_j)$. Then define X' as the set of those functions in X such that no restriction to $D_j^{+2\ell}$ equals x_j.

We claim that $\tau(X') = \tau(X)$. In fact, if $z \in X \setminus X'$ then there exists a non-empty set of integers J such that $z = x_j$ on $D_j^{+2\ell}$ for all $j \in J$. Then define z' as follows:

$z' = z$ outside $\bigcup_{j\in J} D_j^{+2\ell}$, and $z'|_{D_j^{+2\ell}} = x'_j$ for all $j \in J$. Clearly, $z' \in X'$ because X is a stable projective limit of memory ℓ; since τ is of bounded propagation $\leq \ell$, we also have $\tau(z') = \tau(z)$.

From Proposition 3.14 and Lemma 3.18 it then follows that

$$\mathrm{ent}(\tau(X)) = \mathrm{ent}(\tau(X')) \leq \mathrm{ent}(X') < \mathrm{ent}(X). \qquad \square$$

3.12 Surjectivity corollary

Proposition 3.20. *Let $\tau : X \to Y$ be a map of bounded propagation admitting a dense holonomy, and suppose that Y is stable and strongly-irreducible. Then the equality* $\mathrm{ent}(Y) = \mathrm{ent}(\tau(X))$ *implies that τ is surjective.*

Proof. If $Y' = \tau(X) \subset Y$ misses some $y \in Y$, then there exists a ball D such that the restriction of y to D does not belong to Y'_D. Then the dense holonomy carries D densely over Δ, and Lemma 3.18 (with X and X' replaced with Y and Y', respectively) applies, yielding $\mathrm{ent}(\tau(X)) = \mathrm{ent}(Y') < \mathrm{ent}(Y)$. $\square$

3.13 Garden of Eden theorem for stable spaces and for amenable shifts

Proof of Theorem B. If τ is surjective, then $\mathrm{ent}(\tau(X)) = \mathrm{ent}(X)$, and Proposition 3.19 implies that τ is pre-injective.

On the other hand, if τ is pre-injective, then, by Proposition 3.17, $\mathrm{ent}(\tau(X)) = \mathrm{ent}(X)$ and, by Proposition 3.20, τ is surjective. $\square$

Remark 3.21. The original statement of Gromov is slightly weaker: instead of *strong-irreducibility* and *finite type* conditions, he assumes in the hypothesis the *bounded propagation* property for the stable space X, which, as shown in Lemma 3.11 and the relative Counterexample 3.12, is a stronger condition.

Also, one could further generalize the statement by assuming the holonomy to be *big* in the sense of Furstenberg (see [17] and the nice survey on Ergodic Ramsey Theory of Bergelson [3]), rather than dense, to obtain a more general statement.

Finally observe that Gromov's statement holds for *uniformly-locally-finite* spaces. If the alphabet S is finite (as we assumed in the introduction), all subsets $X \subseteq S^\Delta$ are clearly uniformly-locally-finite.

We can now show why Theorem B generalizes the Garden of Eden theorem for amenable subshifts [5, 12].

Proof of Corollary C. Let Δ denote the Cayley graph of G with respect to a suitable finite symmetric generating system A. Then Δ is simple, with the same cardinality

of G, thus at most countable, and of bounded valency: the graph is indeed regular of degree $|A|$.

Because of the strong-regularity of Δ we might consider as holonomy the pseudogroup H_G of partial isometries generated by all the translations $t_g : \Delta \to \Delta$, $g \in G$, where $t_g(\delta) = g\delta$, $\delta \in \Delta$. In other words, we set

$$H_G = \{t_g|_\Omega : \Omega \to g\Omega; \quad g \in G, \quad \Omega \subset \Delta\}.$$

The holonomy H_G is clearly cofinite (indeed, all r-balls are H_G-equivalent) and dense.

A subspace $X \subseteq S^G$ is always *uniformly*-locally-finite and, as we observed in Remark 3.6, it is closed if and only if it is stable.

On the other hand, a map $\tau : X \to Y$ is of bounded propagation ($\leq \ell$) if and only if (tautologically) it is (ℓ-)local. Also, τ admits H_G as a (dense) pseudogroup of holonomies if and only if it is G-equivariant.

Finally, it is well-known, e.g., see [4], that G is amenable (as a group) if and only if Δ is amenable (as a graph). □

Remark 3.22. As with Theorem A, in the statement of Corollary C we can relax the finite type condition for the subshift Y by requiring Y to be sofic, see [12, Lemma 3.5.4]).

Andrzej Żuk asked about examples of *amenable* graphs with *dense pseudogroups of holonomy maps* but admitting *no cocompact isometry groups*.

As remarked above, Cayley graphs of groups have dense pseudogroups of holonomy maps in a trivial way; also, they have cocompact (= cofinite) isometry groups (indeed they have just a single orbit). Thus, Żuk's question refers to *non-trivial* examples.

Example 3.23. A first example is provided by the following 4-regular graph. Let $\Delta = (V, E)$ denote the Cayley graph of the integers with respect to the canonical set of generators $S = \{1, -1\}$. Thus, $V = \mathbb{Z}$ and $E = \{(n, n+1) : n \in \mathbb{Z}\}$. Let now

$$E_0 = \{(2n, 2n+2) : n \in \mathbb{Z}\},$$
$$E_1 = \{(4n+1, 4n+5) : n \in \mathbb{Z}\},$$
$$E_2 = \{(8n-1, 8n+7) : n \in \mathbb{Z}\},$$

and so on, namely, if

$$E_{k-1} = \{(2^k n + a_{k-1}, 2^k(n+1) + a_{k-1}) : n \in \mathbb{Z}\}$$

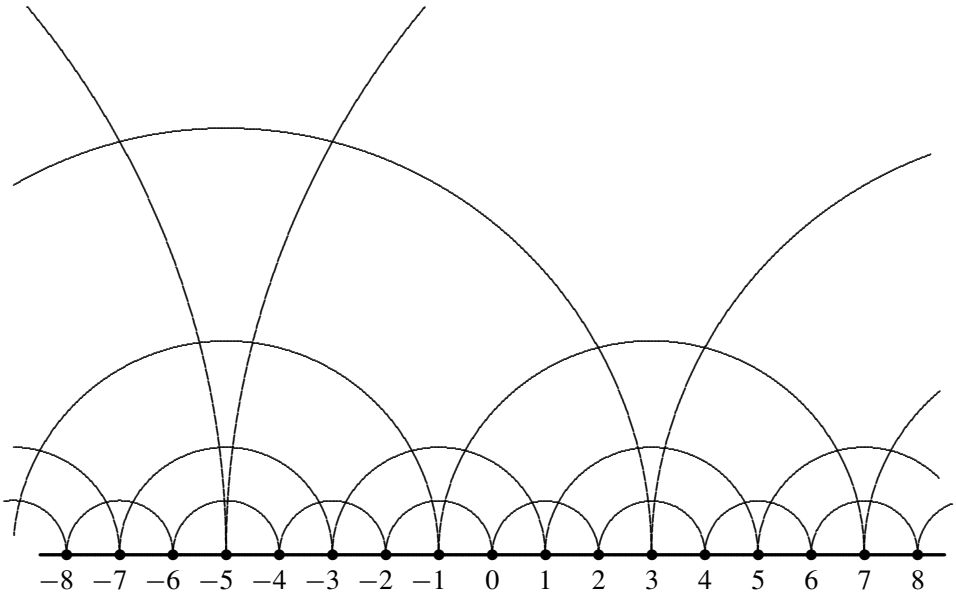
has already been defined, then

$$E_k = \{(2^{k+1} n + a_k, 2^{k+1}(n+1) + a_k) : n \in \mathbb{Z}\},$$

where a_k is such that

$$|a_k| = \min\{|m| : m \notin (2\mathbb{Z}) \cup (4\mathbb{Z}+1) \cup (8\mathbb{Z}-1) \cup \cdots \cup (2^k\mathbb{Z} + a_{k-1})\}.$$

Setting $\overline{E} = E \coprod \left(\coprod_{i=0}^{\infty} E_i \right)$, define $\overline{\Delta} = (V, \overline{E})$. By construction this graph is 4-regular, with a dense pseudogroup of holonomies and with a non-cocompact (= non-cofinite, in this discrete setting) isometry group. Suitable calculations show that $\overline{X}$ is amenable (for more details see [7] where, among other things, $\overline{X}$ is regarded as a Schreier graph associated with the *self-similar action* of a *branch group* on the boundary of the rooted binary tree, [1]).



The 8-neighbourhood of 0 in $\overline{\Delta}$

Example 3.24. Another class of examples is provided by the 1-skeleton of *aperiodic Penrose tilings* of euclidean plane. Around 1975 Penrose discovered a pair of figures (tiles) with the following properties: (i) the plane can be tiled (no gaps, no overlappings) by infinitely many copies of these two, in infinitely many (even continuously many) ways; (ii) none of these tilings is periodic; (iii) any finite part of such a tiling occurs infinitely often in each one of the others. See [21, 33].

We thank Volodia Nekrashevych for suggesting this last example (also see [24, 31] where connections of Penrose tilings with subshifts and Cuntz–Pimsner algebras are investigated).

4 Appendix

Proposition 4.1. *Let $X, Y \subseteq S^G$ be two subshifts over a finitely generated group G and $\tau : X \to Y$ a transition map. Then the following are equivalent:*

(i) *there exist GOE patterns;*

(ii) *there exist GOE configurations (i.e., τ is not surjective).*

Proof. By a *pattern* p in Y, with support say F, we clearly mean a $p = c|_F \in Y_F$ with $c \in Y$.

(i) $\Rightarrow$ (ii). It is clear that the surjectivity of τ implies the non-existence of GOE patterns (in Y).

(ii) $\Rightarrow$ (i). Suppose that for each finite set $F \subseteq G$ and each pattern $p \in Y_F$ there is a configuration $c \in X$ such that $\tau(c)|_F = p$; we prove that τ is surjective. Let $\{\Omega_n\}_n \subseteq G$ be an exhausting increasing sequence of finite subsets of G. If $c' \in Y$, let $c_n \in X$ be such that $\tau(c_n)|_{\Omega_n} = c'|_{\Omega_n}$; hence $\lim_{n\to\infty} \tau(c_n) = c'$. The space X being compact, there is a subsequence $(c_{n_k})_k$ that converges to a configuration $c \in X$, and τ being continuous, we have that $c' = \lim_{k\to\infty} \tau(c_{n_k}) = \tau(c)$. □

Proposition 4.2. *For a transition map $\tau : S^G \to S^G$ the following are equivalent:*

(i) *τ is pre-injective;*

(ii) *there exist no τ-mutually erasable patterns.*

Proof. (i) $\Rightarrow$ (ii). Let p_1 and p_2 be two τ-mutually erasable patterns with support F. Fix $s \in S$ and define $c_1, c_2 \in S^G$ that coincide, respectively, with p_1 and p_2 on F and that are constant, with value s, elsewhere, i.e., $c_1|_{G\setminus F} \equiv s \equiv c_2|_{G\setminus F}$.

Then c_1 and c_2 differ only on a non-empty finite set (since this set is contained in F), and $\tau(c_1) = \tau(c_2)$, so that τ is not pre-injective.

(ii) $\Rightarrow$ (i). Suppose, conversely, that τ is not pre-injective; there exist two configurations c_1 and c_2 such that, for some non empty finite set F, $c_1|_F \neq c_2|_F$, $c_1|_{G\setminus F} = c_2|_{G\setminus F}$, and $\tau(c_1) = \tau(c_2)$.

Set $p_1 := c_1|_{F^{+2M}}$ and $p_2 := c_2|_{F^{+2M}}$, where M is such that τ is M-local. Observe that $p_1 \neq p_2$, and if $\overline{c_1}$, $\overline{c_2}$ are two configurations such that $\overline{c_1}_{|F^{+2M}} = p_1$, $\overline{c_2}_{|F^{+2M}} = p_2$ and $\overline{c_1} = \overline{c_2}$ out of F^{+2M}, then $\tau(\overline{c_1}) = \tau(\overline{c_2})$, as it immediately follows from the M-locality of τ. Thus p_1 and p_2 are τ-mutually erasable. □

The following is a generalization of a classical result for 1-dimensional subshifts (e.g., see [2] or [27, Theorem 6.1.21]). We include both the statement and the proof, in this generalized version, for the sake of completeness and for the convenience of the reader.

Proposition 4.3. *A subset $X \subseteq S^G$ is a subshift (i.e., it is a G-invariant closed subspace) if and only if there exists a subset $\mathcal{F}$ of patterns such that $X = X_{\mathcal{F}}$, where,*

denoting by $F(p)$ the support of a pattern p,

$$X_{\mathcal{F}} = \{c \in S^G : c^g|_{F(p)} \notin \mathcal{F}, \text{ for all } g \in G \text{ and } p \in \mathcal{F}\}.$$

Proof. Suppose that $X \subseteq S^G$ is a shift. The space X being closed, for each $c \notin X$ there exists an integer $n_c > 0$ such that

$$B_{S^G}\left(c, \frac{1}{n_c}\right) \subseteq S^G \setminus X.$$

Let $\mathcal{F}$ be the set

$$\mathcal{F} := \{c_{|D_{n_c}} : c \notin X\}.$$

We prove that $X = X_{\mathcal{F}}$. If $c \notin X_{\mathcal{F}}$, there exists $\overline{c} \notin X$ such that ${c^h}_{|D_{n_{\bar{c}}}} = \overline{c}_{|D_{n_{\bar{c}}}}$ for some $h \in G$. Then $\mathrm{dist}(c^h, \overline{c}) < \frac{1}{n_{\bar{c}}}$ and hence $c^h \in B_{S^G}(\overline{c}, \frac{1}{n_{\bar{c}}}) \subseteq S^G \setminus X$, which implies $c \notin X$, by the G-invariance of X. This proves that $X \subseteq X_{\mathcal{F}}$. For the other inclusion, if $c \notin X$ then, by definition of $\mathcal{F}$, $c_{|D_{n_c}} \in \mathcal{F}$ and hence $c \notin X_{\mathcal{F}}$.

Now, for the converse, we have to prove that a set of type $X_{\mathcal{F}}$ is a shift. Observe that

$$X_{\mathcal{F}} = \bigcap_{p \in \mathcal{F}} X_{\{p\}},$$

and, if the support of p is $F(p) = \{h_1, \ldots, h_N\}$,

$$X_{\{p\}} = \bigcap_{h \in G} \{c \in S^G : {c^h}_{|F(p)} \neq p\} = \bigcap_{h \in G} \Big(\bigcup_{i=1}^{N} \{c \in S^G : c^h(h_i) \neq p(h_i)\} \Big).$$

Thus, in order to prove that $X_{\mathcal{F}}$ is closed, it suffices to prove that for any $i = 1, \ldots, N$ the set

$$\{c \in S^G : c^h(h_i) \neq p(h_i)\} \tag{4.1}$$

is closed. We have

$$(\{c \in S^G : c^h(h_i) \neq p(h_i)\})^h = \{c \in S^G : c(h_i) \neq p(h_i)\},$$

and then the set in (4.1) is closed being the preimage of a closed set under a continuous function. Finally, we have to prove that $X_{\mathcal{F}}$ is G-invariant. If $g \in G$ and $c \in X_{\mathcal{F}}$, for every $h \in G$ and every $p \in \mathcal{F}$ we have

$${c^{gh}}_{|F(p)} \notin \mathcal{F} \Rightarrow {(c^g)^h}_{|F(p)} \notin \mathcal{F},$$

and hence $c^g \in X_{\mathcal{F}}$. □

The following generalizes the well-known Curtis–Lyndon–Hedlund theorem, see [27, Theorem 6.2.9].

Proposition 4.4. *Let $X \subseteq S^G$ be a subshift. A function $\tau : X \to S^G$ is a local map (i.e. induced by a local rule f) if and only if it is G-equivariant (i.e., $\tau[c^g] = \tau[c]^g$ for all $c \in X$ and $g \in G$) and continuous.*

Proof. Suppose that τ is M-local, induced, say, by a local rule f. Then, denoting by $D_M = \{h_1, \ldots, h_m\}$ the ball of radius M centered at the neutral element $1 \in G$ we have that for $g \in G$ and $c \in X$,

$$\tau[c^g](h) = f[c^g(hh_1), c^g(hh_2), \ldots, c^g(hh_m)] = f[c(ghh_1), \ldots, c(ghh_m)],$$

and

$$\tau[c]^g(h) = \tau[c](gh) = f[c(ghh_1), \ldots, c(ghh_m)],$$

so that τ commutes with the G-action.

We prove that τ is continuous. A generic element of a sub-basis of S^G is

$$\xi = \xi(h; s) = \{c \in S^G : c(h) = s\}$$

with $h \in G$ and $s \in S$. It suffices to prove that the set

$$\overline{\xi} := \tau^{-1}(\xi) = \{c \in X \ : \tau[c](h) = s\}$$

is open in X. Actually, if $c \in \overline{\xi}$ and

$$r := \min\{n \in \mathbb{N} : hh_1, \ldots, hh_m \in D_n\}, \tag{4.2}$$

we claim that the ball $B_X(c, \frac{1}{r+1})$ is contained in $\overline{\xi}$. Indeed, if $\overline{c} \in B_X(c, \frac{1}{r+1})$ then

$$\operatorname{dist}(c, \overline{c}) < \frac{1}{r+1},$$

i.e., $c_{|D_r} = \overline{c}_{|D_r}$. Since by (4.2), $hh_i \in D_r$, we have

$$\tau[\overline{c}](h) = f[\overline{c}(hh_1), \ldots, \overline{c}(hh_m)] = f[c(hh_1), \ldots, c(hh_m)] = \tau[c](h) = s,$$

so that $\overline{c} \in \overline{\xi}$.

Conversely, suppose that τ is continuous and commutes with the action of G. Since X is compact, τ is uniformly continuous; fix M in $\mathbb{N}$ such that for every $c, \overline{c} \in X$,

$$\operatorname{dist}(c, \overline{c}) < \frac{1}{M+1} \Rightarrow \operatorname{dist}(\tau[c], \tau[\overline{c}]) < 1.$$

Thus, if c and $\overline{c}$ agree on D_M, then $\tau(c)$ and $\tau(\overline{c})$ agree at the neutral element $1 \in G$: $\tau[c](1) = \tau[\overline{c}](1)$, so that the function $f : S^{D_M} \to S$ defined by $f(s_{h_1}, \ldots, s_{h_M}) = \tau[c](1)$, where c is any configuration such that $c(h_i) = s_{h_i}$, $i = 1, \ldots, M$, is well defined. In addition f serves as a *local rule* for τ; indeed, since τ commutes with the action of G, we have

$$\tau[c](h) = \tau[c]^h(1) = \tau[c^h](1) = f[c(hh_1), \ldots, c(hh_m)] \ ;$$

this shows that τ is M-local and ends the proof. □

It is clear from the above proposition that *the composition of two local functions is still local*. In any case, this can be easily seen by a direct proof. Also, one gets immediately the fact that the inverse of an invertible transition map is still local; in the terminology of cellular automata this can be rephrased as "the inverse of an (invertible) cellular automaton is a cellular automaton" : this is usually proved in a combinatorial way combined with a compactness argument (e.g., see [2, 27]):

Corollary 4.5. *Suppose that a transition map $\tau : X \to Y$ is invertible* (*i.e., surjective and injective*). *Then its inverse $\tau^{-1} : Y \to X$ is also a local map.*

Acknowledgement. We are grateful to Volodia Nekrashevych, Giovanni Stegel and Andrzej Żuk for fruitful discussions and remarks.

References

[1] L. Bartholdi, R. I. Grigorchuk and V. Nekrashevych, From fractal groups to fractal sets, in: Fractals in Graz (Graz, 2001), Birkhäuser, Basel 2003, 25–118.

[2] M.-P. Béal and D. Perrin, Symbolic dynamics and finite automata, in: Handbook of Formal Languages, Vol. 2, Springer-Verlag, Berlin 1997, 463–505.

[3] V. Bergelson, Ergodic Ramsey theory—an update, in: Ergodic theory of Z^d Actions (Warwick, 1993–1994), London Math. Soc. Lecture Note Ser. 228, Cambridge University Press Cambridge, 1996, 1–61.

[4] T. Ceccherini-Silberstein, R. I. Grigorchuk and P. de la Harpe, Amenability and paradoxical decompositions for pseudogroups and discrete metric spaces, Proc. Steklov Inst. Math. 224 (1999), 57–97.

[5] T. G. Ceccherini-Silberstein, A.Machì and F. Scarabotti, Amenable groups and cellular automata, Ann. Inst. Fourier (Grenoble) 49 (1999), 673–685.

[6] T. Ceccherini-Silberstein, A. Machì and F. Scarabotti, On the entropy of regular languages, Theoret. Comput. Sci. (to appear).

[7] T. Ceccherini-Silberstein and V. Nekrashevych, in progress.

[8] T. Ceccherini-Silberstein and F. Scarabotti, Random walks, entropy and hopfianity of free groups, in: Random Walks and Geometry, Proceedings of a Workshop at the Erwin Schrödinger Institute, Walter de Gruyter, Berlin 2004, 413–419.

[9] T. Ceccherini-Silberstein and W. Woess, Growth and ergodicity of context-free languages, Trans. Amer. Math. Soc. 354 (2002), 4597–4625.

[10] E. Coven and M. Paul, Endomorphisms of irreducible shifts of finite Type, Math. Systems Theory 8 (1974/75), 167–175.

[11] D. Epstein, J.Cannon, D. Holt, S. Levy, M. Paterson and W. Thurston, Word Processing in Groups, Jones and Bartlett Publishers, Boston, MA, 1992.

[12] F. Fiorenzi, Cellular Automata and Finitely Generated Groups, PhD thesis, Università di Roma "La Sapienza", 2000.

[13] F. Fiorenzi, The Garden of Eden theorem for sofic shifts, Pure Math. Appl. 11 (2000), 471–484.

[14] F. Fiorenzi, Cellular automata and strongly-irreducible shifts of finite type, Theoret. Comput. Sci. (to appear).

[15] F. Fiorenzi, Semi-strong irreducibility of shifts and Moore-Myhill property, preprint (2001).

[16] E. Følner, On groups with full Banach mean value, Math. Scand. 3 (1955), 243–254.

[17] H. Furstenberg, Recurrence in Ergodic Theory and Combinatorial Number Theory, M. B. Porter Lectures, Princeton University Press, Princeton, NJ, 1981.

[18] M. Harrison, Introduction to Formal Language Theory, Addison-Wesley, Reading, MA, 1978.

[19] G. A. Hedlund, Endomorphisms and automorphisms of the shift dynamical systems, Math. Systems Theory 3 (1969), 320–375.

[20] J. E. Hopcroft and J. D. Ullman, Introduction to Automata Theory, Languages and Computation, Addison-Wesley, Reading, MA, 1979.

[21] M. Gardner, Mathematical games. Extraordinary nonperiodic tiling that enriches the theory of tiles, Scientific American 236 (1977), 110–121.

[22] W. Gottschalk, Some general dynamical notions, in: Recent Advances in Topological Dynamics, Lecture Notes in Math. 318, Springer-Verlag, Berlin 1973, 120–125.

[23] R. I. Grigorchuk and P. de la Harpe, On problems related to growth, entropy and spectrum in group theory, J. Dynam. Control Systems 3 (1997), 51–89.

[24] R. I. Grigorchuk, V. V. Nekrashevich and V. I. Sushchanskiĭ, Automata, dynamical systems, and groups, Proc. Steklov Inst. Math. 231 (2000), 128–203.

[25] M. Gromov, Endomorphisms of symbolic algebraic varieties, J. Eur. Math. Soc. (JEMS) 1 (1999), 109–197.

[26] W. Krieger, On sofic systems, I; II. Israel J. Math. 48 (1984) 305–330; Israel J. Math. 60 (1987), 167–176.

[27] D. Lind and B. Marcus, An Introduction to Symbolic Dynamics and Coding, Cambridge University Press, Cambridge 1995.

[28] A. Machì and F. Mignosi, Garden of Eden configurations for cellular automata on Cayley graphs on groups, SIAM J. Discrete Math. 6 (1993), 44–56.

[29] E. F. Moore, Machine models of self-Reproduction, Proc. Symp. Appl. Math. 14 (1963), 17–34.

[30] J. Myhill, The converse of Moore's Garden of Eden theorem, Proc. Amer. Math. Soc. 14 (1963), 685–686.

[31] V. Nekrashevych, Cuntz-Pimsner algebras of group actions, preprint (2002).

[32] J. von Neumann, The Theory of Self-Reproducing Automata (A. Burks ed.), University of Illinois Press, Urbana and London 1966.

[33] R. Penrose, Pentaplexity: a class of nonperiodic tilings of the plane. Math. Intelligencer 2 (1979/80), 32–37.

[34] F. Scarabotti, On a lemma of Gromov and the entropy of a graph, European J. Combin. (to appear).

[35] S. Ulam, Random processes and transformations, in: Proceedings of the International Congress of Mathematicians, Cambridge, Mass., 1950, Amer. Math. Soc., Providence, RI, 1952, 264–275.

[36] B. Weiss, Subshifts of finite type and sofic systems, Monatsh. Math. 77 (1973), 462–474.

[37] B. Weiss, Sofic groups and dynamical systems, Sankhyā Ser. A 62 (2000), 350–359.

Tullio Ceccherini-Silberstein, Dipartimento di Ingegneria, Università del Sannio, C.so Garibaldi 107, 82100 Benevento, Italy
E-mail: tceccher@mat.uniroma1.it

Francesca Fiorenzi, LIX - École Polytechnique, 91128 Palaiseau Cedex, France
E-mail: fiorenzi@lix.polytechnique.fr

Fabio Scarabotti, Dipartimento MeMoMat, Università degli Studi di Roma “La Sapienza”, Via A. Scarpa 16, 00161 Roma, Italy
E-mail: scarabot@dmmm.uniroma1.it

Expander graphs, random matrices and quantum chaos

Alex Gamburd

Abstract. A basic problem in the theory of expander graphs, formulated by Lubotzky and Weiss, is to what extent being an expander family for a family of Cayley graphs is a property of the groups alone, independent of the choice of generators. While recently Alon, Lubotzky and Wigderson constructed an example demonstrating that expansion is not in general a group property, the problem is open for "natural" families of groups. In particular for $\{\mathrm{SL}_2(\mathbb{F}_p)\}$ numerical experiments indicate that it might be an expander family for "generic" choices of generators (Independence Conjecture).

A basic conjecture in quantum chaos, formulated by Bohigas, Giannoni, and Schmit, asserts that the eigenvalues of a quantized chaotic Hamiltonian behave like the spectrum of a typical member of the appropriate ensemble of random matrices. Both conjectures can be viewed as asserting that a deterministically constructed spectrum "generically" behaves like the spectrum of a large random matrix: "in the bulk" (Quantum Chaos Conjecture) and at the "edge of the spectrum" (Independence Conjecture). After explaining this approach in the context of the spectra of random walks on groups, we will review some recent related results and numerical experiments.

Contents

1 Introduction 110
- 1.1 Spectra of random walks on groups . . . 110
- 1.2 Expander graphs . . . 112
- 1.3 Random matrices . . . 114
- 1.4 Quantum chaos . . . 115

2 Spectral gap: discrete and continuous variations on the expanding theme 116
- 2.1 Independence problem for groups and expanders . . . 116
- 2.2 Generalization of Selberg's theorem . . . 117
- 2.3 Expander graphs and invariant means . . . 119
- 2.4 Random walk on a sphere . . . 120

3 Overview of related results 125
- 3.1 Zig-zag product . . . 125
- 3.2 Example of nonuniform groups . . . 126

3.3 Uniform exponential growth 127
3.4 Uniform diameter bounds 127
3.5 Uniform Kazhdan constant 128
3.6 Unbounded number of generators 129

4 Spacings and discrepancy 131
4.1 Numerical results 131
4.2 Noncommutative diophantine property 133

1 Introduction

1.1 Spectra of random walks on groups

Starting with the foundational paper of Kesten [53], the investigation of the connections between the structure of the group, the geometry of random walks, and the spectrum of the associated convolution operator has developed into a beautiful and rich subject [22, 86, 107]; as the main results of Kesten's paper are pertinent to the discussion below let us begin by briefly recalling them.

Suppose that Γ is a countable group, generated by a finite symmetric set of generators $S = \{g_1, g_1^{-1}, \ldots, g_k, g_k^{-1}\}$. Associate to the pair (Γ, S) the Cayley graph $\mathcal{G}(\Gamma, S)$, which has elements of Γ as vertices and which has an edge from x to y if and only if $x = \sigma y$ for some $\sigma \in S$, and consider the spectrum of the averaging operator T associated with the (uniform) random walk supported on S:

$$T_{g_1,\ldots,g_k} f(x) = \sum_{j=1}^{k} (f(g_j x) + f(g_j^{-1}(x)).$$

Let ν be the spectral measure of T, and let $\|T\| = \max \operatorname{supp}(\nu)$ be the operator norm of T acting on $L^2(\Gamma)$. Starting with a crucial observation

$$\frac{\|T\|}{2k} = \limsup_{n\to\infty} [\text{Probability of return to } e \text{ at the } n\text{th step}]^{\frac{1}{n}}, \tag{1.1}$$

Kesten proves the following fundamental result:

Theorem 1 (Kesten [53]). *Notation being as above we have:*

1. $\|T\| = 2k$ *if and only if* Γ *is amenable.*
2. $\|T\| \geq 2\sqrt{2k-1}$, *with equality if and only if* Γ *is free on* $g_1, \ldots, g_k$. *In the latter case* ν *is supported in* $[-2\sqrt{2k-1}, 2\sqrt{2k-1}]$ *and is given by*

$$d\nu_k(t) = \frac{\sqrt{2k-1-t^2/4}}{2\pi k(1-(t/2k)^2)}\, dt. \tag{1.2}$$

We will refer to measure (1.2) as the *Kesten measure*.

In this paper we will be concerned with the fine structure of the spectra of random walks on groups, and in particular with the way they depend on the choice of generators. The general setting is as follows. Given a group G and a set of elements $\{g_1, \ldots, g_k\}$ in G, let $z_{g_1,\ldots,g_k}$ denote the element in the group ring of G, associated with the symmetric random walk supported on $S = \{g_1,\ g_1^{-1},\ \ldots,\ g_k,\ g_k^{-1}\}$:

$$z_{g_1,\ldots,g_k} = g_1 + g_1^{-1} + \cdots + g_k + g_k^{-1}. \tag{1.3}$$

For such a $z \in \mathbb{R}[G]$, let $\mathrm{supp}(z) = \{g_1, \ldots, g_k\}$, and let Γ_z be the finitely generated subgroup of G generated by $\mathrm{supp}(z)$. Now, given a family of finite-dimensional unitary representations

$$\pi_N : G \to \mathrm{GL}(W_N),$$

with the degree of representations $d_N = \dim(\mathrm{W_N})$ going to infinity, consider $\pi_N(z)$, the image of z under π_N:

$$\pi_N(z) = \pi_N(g_1) + \pi_N(g_1^{-1}) + \cdots + \pi_N(g_k) + \pi_N(g_k^{-1}). \tag{1.4}$$

The $(d_N \times d_N)$ matrix $\pi_N(z)$ is self-adjoint with respect to a suitable inner product on W_N; consequently its spectrum $\mathrm{spec}(\pi_N(z))$ is real and contained in the interval $[-||z||, ||z||] = [-2k, 2k]$.

Let $||\pi_N(z)||$ be the norm of the matrix $\pi_N(z)$ with respect to the above inner product; $||\pi_N(z)||$ is equal to the maximum of $|\lambda_j(\pi_N(z))|$, where

$$\lambda_0(\pi_N(z)) \geq \lambda_1(\pi_N(z)) \geq \cdots \geq \lambda_{d_N}(\pi_N(z)) \tag{1.5}$$

are the eigenvalues of $\pi_N(z)$. The empirical density of eigenvalues (referred to as "density of states" in the physics literature) is described by the sum of point masses

$$\mu_N(z) = \frac{1}{d_N} \sum_{j=0}^{d_N} \delta_{\lambda_j(\pi_N(z))}, \tag{1.6}$$

which is a probability measure supported in $[-2k, 2k]$. As is well-known [87, 95], under quite general conditions (namely, as long as

$$\lim_{N\to\infty} \frac{\chi_N(g)}{d_N} = 0 \quad \text{if} \quad g \neq 1, \tag{1.7}$$

where $\chi_N(g) = \mathrm{trace}(\pi_N(g))$ is a character of π_N), there is a measure $\nu(z)$, such that as $d_N \to \infty$

$$\mu_N(z) \xrightarrow[\text{weak}^*]{} \nu(z). \tag{1.8}$$

The limiting measure depends only on the abstract group Γ_z. For Γ_z free (which as we review in Section 4.2 holds for a generic z), it is precisely the Kesten measure ν_k defined in equation (1.2).

We will be concerned with the following basic problems as the degree of representations goes to infinity:

Problem 1. *Does z have a spectral gap, i.e., is it the case that*

$$\overline{\lim}_{N\to\infty}||\pi_N(z)|| \ < \ ||z|| = 2k? \tag{1.9}$$

We remark that by Kesten's theorem z cannot have a spectral gap if Γ_z is amenable, and that $2\sqrt{2k-1}$, the optimal bound in equation (1.9), can be achieved only for Γ_z free. Following Lubotzky, Phillips, and Sarnak [67, 66], we call an element $z_{g_1,\ldots,g_k}$ satisfying the optimal bound

$$||\pi_N(z)|| \leq 2\sqrt{2k-1} \tag{1.10}$$

for $N \geq 1$, a *Ramanujan element*[1] (see Section 2.4).

Problem 2. *What is the distribution of local spacings between the eigenvalues of $\pi_N(z)$?*

Problem 3. *What is the speed of convergence to spectral measure $\nu(z)$, in particular how does discrepancy $D(\mu_N(z), \nu(z))$ depend on N?*

Here $D(\nu, \mu)$ is the discrepancy between the measures ν and μ, that is

$$D(\nu, \mu) = \sup\{|\nu(I) - \mu(I)| : I = [a, b] \subset \mathbb{R}\}. \tag{1.11}$$

We now briefly explain how some specific questions related to expander graphs, random matrices and quantum chaos fit into this general setting, while indicating at the same time the organization and content of the paper.

1.2 Expander graphs

Expanders are highly-connected sparse graphs widely used in Computer Science, in areas ranging from parallel computation to complexity theory and cryptography; recently they also have found some remarkable applications in pure mathematics, notably in the work of Gromov [36, 37]. There are several ways of making the intuitive notions of connectivity and sparsity precise, the simplest and most widely used is the following:

Definition 1.1. Given an undirected k-regular graph $\mathcal{G}$ and a subset X of V, the *expansion* of X, $c(X)$, is defined to be the ratio $|\partial(X)|/|X|$, where $\partial(X) = \{y \in \mathcal{G} :$

[1] A well-known problem in the theory of operator algebras, "Is $\mathrm{Ext}(C_r^*(F_d))$ a group?", (see [105, 38]) can be reduced to the following question: for a free group $F_d (d \geq 2)$ does there exist a sequence π_N of finite-dimensional unitary representations such that *for every* element in the group ring $\mathbb{C}[F_d]$ we have

$$\overline{\lim}_{N\to\infty}||\pi_N(z)|| \leq ||\rho(z)||,$$

where ρ denotes the regular representation of F_d. The Ramanujan elements of Lubotzky, Phillips and Sarnak provide an affirmative answer for elements z of the form (1.3) (and not just asymptotically!). Very recently U. Haagerup and S. Thorbjornsen [39] gave a complete solution of this problem using Voiculescu's free probability theory and random matrices.

distance$(y, X) = 1\}$. The *expansion coefficient* of a graph $\mathcal{G}$ is given by

$$c(\mathcal{G}) = \inf\left\{c(X) \mid |X| < \frac{1}{2}|\mathcal{G}|\right\}. \tag{1.12}$$

A family of k-regular graphs $\mathcal{G}_{n,k}$ forms a family of C-expanders if there is a fixed positive constant C, such that

$$\liminf_{n\to\infty} c(\mathcal{G}_{n,k}) \geq C\ . \tag{1.13}$$

We note that for a discrete group Γ, Følner's definition of amenability can be stated as follows: for every $\varepsilon > 0$ and every generating set S the Cayley graph $\mathcal{G}(\Gamma, S)$, has a finite subset of vertices X whose boundary $\partial(X)$ satisfies $|\partial(X)| < \varepsilon|X|$, so that the expansion property is precisely the opposite of amenability. The spectral characterization of amenability given by Kesten's theorem is equivalent to Følner's characterization and holds with respect to *any* choice of generators; see [45] for a direct proof.

Recall that the *adjacency matrix* of $\mathcal{G}$, $A(\mathcal{G})$ is the $|\mathcal{G}| \times |\mathcal{G}|$ matrix, with rows and columns indexed by vertices of $\mathcal{G}$, such that the (x, y) entry is 1 if and only if x and y are adjacent and 0 otherwise. For a k-regular graph the adjacency matrix is related to the combinatorial Laplacian Δ by $A = kI - \Delta$ (see [64] for a very clear exposition of this and other topics pertaining to expander graphs). Using the discrete Cheeger–Buser inequality, condition (1.13) can be rewritten in terms of the second largest eigenvalue of the adjacency matrix $A(\mathcal{G})$ as follows:

$$\limsup_{n\to\infty} \lambda_1(A_{n,k}) < k. \tag{1.14}$$

As we (briefly) review in Section 2.1, the explicit constructions of expander graphs (by Margulis [72, 75] and Lubotzky, Phillips and Sarnak [67]) use deep tools (Kazhdan's property (T), Selberg's Theorem, proved Ramanujan conjectures) to construct families of Cayley graphs of finite groups as follows. Starting with an infinite group Γ (e.g., $\mathrm{SL}_2(\mathbb{Z})$) and a finite set of generators S, one considers a family of Cayley graphs $\mathcal{G}_i = \mathcal{G}(G_i, S_i)$, where G_i is an infinite family of finite quotients (e.g., $\mathrm{SL}_2(\mathbb{F}_p)$)) and S_i is an image of S under the natural projection[2].

On the other hand, there are families of groups whose Cayley graphs are not expanders with respect to some *specific* choices of generators, notably S_n with respect to $\{(12), (1, 2, \ldots, n)\}$ (see [32] and references therein). Furthermore, as shown in [68], some families of groups, for example, abelian groups or solvable groups of bounded derived length, cannot be made into families of expanders with respect to *any* choice of generators.

[2]Note that the adjacency matrix of the Cayley graph $\mathcal{G}(G_i, S_i)$, is equal to $\pi_N(z_S)$, where π_N is the regular representation of the group G_i. Keeping in mind equation (1.14), the condition for $\mathcal{G}(G_i, S_i)$ to be an expander family can be expressed exactly as stated in equation (1.9) of Problem 1 with $z = z_S$ and with $|S| = 2k$.

The basic question, posed by Lubotzky and Weiss [68] (see also [64, 65, 69]) is to what extent the expansion property is the property of the family of groups $\{G_i\}$ alone, independent of the choice of generators:

Independence Problem (Lubotzky and Weiss [68]). *Let G_i be a family of finite groups, $\langle S_i \rangle = \langle S_i' \rangle = G_i$ and $|S_i| < k$, $|S_i'| < k$. Does the fact that $\mathcal{G}(G_i, S_i)$ is an expander family imply the same for $\mathcal{G}(G_i, S_i')$?*

In a recent breakthrough paper Reingold, Vadhan and Wigderson [85] introduced a notion of zig-zag product of graphs and gave a purely combinatorial construction of expander graphs (see Section 3.1). Reinterpreting the notion of zig-zag product in the context of semi-direct product of groups, Alon, Lubotzky and Wigderson [2] constructed a family of groups G_i which are expanders with respect to one choice of generators and not with respect to another such choice. The groups G_i are of the form $A_i \rtimes B_i$, where $B_p = SL_2(F_p)$ and $A_p = F_2^{P_1}$, with $P_1 = F_p \cup \{\infty\}$ (see Section 3.2).

On the other hand, the problem is still open for the "natural" families of groups, such as $\{S_n\}$ and $\{\mathrm{SL}(n, \mathbb{F}_p)\}$. In particular, for $\{\mathrm{SL}_2(\mathbb{F}_p)\}$ the affirmative answer is supported by extensive numerical experiments of Lafferty and Rockmore [58, 59, 60], as well as by some recent results which we review in Section 2.

One line of approach to the independence problem is to obtain a more robust proof of results used in the explicit constructions with an eye towards increasing their scope and learning more about their meaning. This approach was pursued in [29] where a generalization of Selberg's theorem for infinite index "congruence" subgroups is proved and consequently a new rich family of expanders was obtained (Section 2.2); and in [30], where a new robust method for establishing the spectral gap property is developed, yielding a new elementary solution of the Ruziewicz problem on the uniqueness of invariant mean on a sphere (Section 2.4).

The intrinsic connection between the expanding property and the uniqueness of invariant mean on the appropriate compact group was emphasized and successfully exploited by Shalom, whose results we review in Section 2.3. In particular, Shalom shows that the affirmative answer to the Independence Problem for the family of groups $\{G_i\}$ has the following equivalent formulation: the Haar measure on G, the profinite completion of $\{G_i\}$, is the unique invariant mean with respect to every finitely generated dense subgroup Γ in G. This in turn can be formulated as the statement that every finitely supported element in the group ring of G (with support generating a dense subgroup) has a spectral gap.

1.3 Random matrices

The subject of Random Matrix Theory (RMT) originated in Wigner's suggestion in the early 50's that the resonance lines of heavy nuclei, their determination by analytic

means being intractable, might be modelled by the spectrum of a large random matrix [106]. He introduced the "canonical" ensembles of random matrices: Gaussian Orthogonal Ensemble (GOE), which is a probability measure on $N \times N$ real symmetric matrices, and its relatives, Gaussian Unitary Ensemble (GUE), and Gaussian Symplectic Ensemble (GSE). As we review in Section 3.6, under quite general assumptions, if we let the number of generators k go to infinity, the measure $\pi_N(z_{g_1,\ldots,g_k})$ converges exactly to GOE measure on the space of real symmetric matrices (or to one of its relatives mentioned above).

The connection with random matrices was implicit (but crucial!) in the early development of expander graphs: the plausibility of explicit construction and the birth of the subject of expander graphs in 1973 stemmed from Pinsker's [84] observation that a random regular graph is a good expander[3]. This corresponds to the following fact about random matrices: random symmetric matrix of size N with k 1's in each row and column and all other entries 0 has the biggest eigenvalue equal to k, but the next eigenvalue will be bounded away from k by a fixed amount independent of N (see [28, 98] for general results about spectral gap for random real symmetric matrices).

Consequently, an affirmative answer to the Independence Problem for "generic" or "random" choices of generators (supported by numerical experiments in cases of $SL_2(\mathbb{F}_p)$ and $SU(2)$) can be viewed as asserting that the spectrum of a generic element in the group ring behaves like the spectrum of a large random matrix "at the edge of the spectrum". That it behaves like the spectrum of a large random matrix "in the bulk of the spectrum" is the content of the basic conjecture in quantum chaos.

1.4 Quantum chaos

The basic conjecture in quantum chaos, formulated by Bohigas, Giannoni, and Schmit in 1984 [14], asserts that the eigenvalues of a quantized chaotic Hamiltonian (after suitable unfolding) behave like the spectrum of a typical member of the appropriate ensemble of random matrices. This conjecture complements (and contrasts with) an earlier conjecture of Berry and Tabor [11], asserting that the eigenvalues of quantized integrable systems follow the Poisson distribution; the Poisson distribution is also expected for *arithmetic* surfaces of constant negative curvature, following the pioneering numerical experiments in [93, 13, 15].

As we review in Section 4.1, one of the simplest examples of this phenomenon is afforded by ergodic actions of groups generated by several linear toral automorphisms

[3]To see how this fits in the framework of Problem 1, note that one of the models([16], [27]) for constructing a random $2k$-regular graph on N vertices consists in picking k random permutation $g_1, \ldots, g_k$ in the symmetric group S_N and looking at the image of $z_{g_1,\ldots,g_k}$ in the N-dimensional permutation representation π_N. Remark that McKay [76] showed that spectral density of random d-regular graphs converges to Kesten's measure and that in [44] it was numerically observed that eigenvalue spacings of random regular graphs follow GOE distribution. In [79] it was numerically observed that the second largest eigenvalue of the adjacency matrix of a random regular graph converges to the Tracy–Widom distribution for GOE matrices in random matrix theory [103]. A dramatic consequence of such behavior would be that the probability of random regular graph being Ramanujan approaches 0.52 as the size of the graph tends to infinity.

– "cat maps". The numerical experiments in [31] indicate that for "generic" choices of cat maps, the unfolded consecutive spacings distribution in the irreducible components of the N-th quantization (given by the N-dimensional Weil representation π_N) approaches the GOE/GSE law of Random Matrix Theory. For certain special "arithmetic" transformations, related to the Ramanujan graphs of Lubotzky, Phillips and Sarnak, the experiments indicate that the unfolded consecutive spacings distribution follows Poisson statistics in analogy to the situation observed for arithmetic manifolds. Similar results hold in SU(2) [30].

One way in which the difference between RMT and the Poisson distribution manifests itself is in the speed of convergence to the density of states: the convergence is much faster in the RMT case [89, 90]. We are thus led to Problem 3. For Ramanujan elements the sharp lower bound of order $\frac{1}{\sqrt{N}}$ is established in [30, 31]: this is the analogue of the lower bounds for the remainder term in Weyl's law for arithmetic hyperbolic surfaces [43, 71]. It turns out that for the "generic" elements the speed of convergence for $\mu_N(z)$ depends on z satisfying the noncommutative *Diophantine condition*. This condition and related results are reviewed in Section 4.2.

2 Spectral gap: discrete and continuous variations on the expanding theme

2.1 Independence problem for groups and expanders

The first explicit construction of expanders was given by Margulis [72], using Kazhdan's property T [48] (see Section 3.5 for the definition of property T). The link between property T and expanders can be stated as follows (following Alon and Milman [3]): if $\Gamma = \langle S \rangle$ has property T, then

$$\{\mathcal{G}(\Gamma/N_i, S_i) \mid N_i \lhd \Gamma, \ [\Gamma : N_i] < \infty\}$$

is a family of expanders. This yields that $\{\mathrm{SL}(n, \mathbb{F}_p)\}$ with $n \geq 3$ is an expanding family with respect to the standard generators of $\mathrm{SL}(n, \mathbb{Z})$.

That $\{\mathrm{SL}_2(\mathbb{F}_p)\}$ is an expanding family follows from Selberg's theorem, whose statement we now recall:

Theorem (Selberg [94]). *Let $\Gamma'(p)$ be any congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$, i.e., subgroup containing principal congruence subgroup $\Gamma(p)$. Let $X'(p) = \Gamma'(p)\backslash\mathbb{H}$. For $p \geq 1$*

$$\lambda_1(X'(p)) \geq \frac{3}{16}. \tag{2.1}$$

Here $\Gamma(p)$ is the principal congruence subgroup of level p:

$$\Gamma(p) = \left\{\gamma \in \mathrm{SL}_2(\mathbb{Z}) \, : \, \gamma \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \mod p\right\}. \tag{2.2}$$

As Lubotzky wrote in [65]:

> What is very frustrating is that all these deep theories [e.g., Selberg's theorem] give some examples with very special sets of generators. A small change of the construction – which seems to be meaningless from the combinatorial point of view – leaves these tools helpless.

Lubotzky illustrated this by the following example. For a prime $p \geq 5$ let us define

$$S_p^1 = \left\{ \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \right\},$$

$$S_p^2 = \left\{ \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \right\},$$

$$S_p^3 = \left\{ \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix} \right\},$$

and for $i = 1, 2, 3$ let $\mathcal{G}_p^i = \mathcal{G}\left(\mathrm{SL}_2(\mathbb{F}_p), S_p^i\right)$, the Cayley graph of $\mathrm{SL}_2(\mathbb{F}_p)$ with respect to S_p^i. The graphs $\mathcal{G}_p^i$ can be viewed as a "discrete approximation" of the hyperbolic manifolds $X_p^i = \Lambda^i(p)\backslash\mathbb{H}$, where Λ^i is the subgroup of $\mathrm{SL}_2(\mathbb{Z})$ generated by $\langle \left(\begin{smallmatrix} 1 & i \\ 0 & 1 \end{smallmatrix}\right), \langle \left(\begin{smallmatrix} 1 & 0 \\ i & 1 \end{smallmatrix}\right) \rangle$. By Selberg's theorem the families X_p^1 and X_p^2 have a spectral gap (form a family of "expander surfaces"), and from this one deduces (using Fell's continuity of induction [26]) that $\mathcal{G}_p^1$ and $\mathcal{G}_p^2$ are families of expander graphs. Given this fact, it is difficult to believe that $\mathcal{G}_p^3$ is not an expander family. Note however, that the group generated by $\langle \left(\begin{smallmatrix} 1 & 3 \\ 0 & 1 \end{smallmatrix}\right), \langle \left(\begin{smallmatrix} 1 & 0 \\ 3 & 1 \end{smallmatrix}\right) \rangle$ has infinite index and thus does not come under the purview of Selberg's theorem. We now turn to results addressing Lubotzky's 1-2-3 question.

2.2 Generalization of Selberg's theorem

We begin by recalling some basic facts about finitely-generated subgroups of $\mathrm{SL}_2(\mathbb{Z})$. In dimension two being finitely generated is equivalent to being geometrically finite, i.e., the fundamental domain $\mathcal{F} = \Lambda\backslash\mathbb{H}$ must have finitely many bounding sides [10]. The limit set of Λ, denoted by $L(\Lambda)$, is a subset of $\mathbb{R} \cup \infty$; it was observed a century ago by Poincaré and Klein that if $\mathcal{F}$ has infinite volume (and Λ is not elementary), then $L(\Lambda)$ is a Cantor-like set; we will denote by $\delta(L(\Lambda))$ its Hausdorff dimension. The spectrum of the Laplacian Δ on $L^2(\mathcal{F})$ will be denoted by $\Omega(\mathcal{F})$.

The spectrum of geometrically finite Fuchsian groups was investigated by Patterson [82]; Sullivan [100], and Lax and Phillips[62] generalized and extended his results in higher dimensions. The main result for $\Omega(\Lambda\backslash\mathbb{H})$ is the following

Theorem ([82, 100, 62]). *Assume that $\delta > \frac{1}{2}$. Then*

(1) *The bottom of the spectrum,* $\lambda_0(\mathcal{F}) = \delta(1-\delta)$*; it is an isolated eigenvalue of multiplicity one.*

(2) *There are finitely many discrete eigenvalues* $\lambda_0(\mathcal{F}), \ldots, \lambda_n(\mathcal{F}) \in [0, 1/4)$.

(3) *If* $\mathrm{vol}(\mathcal{F}) = \infty$*, then the spectrum* $\Omega(\mathcal{F})$ *is continuous in* $[1/4, \infty]$.

We are now ready to state a generalization of Selberg's theorem:

Theorem 2 ([29]). *Let* $\Lambda = \langle A_1, \ldots, A_k \rangle$ *be a finitely generated subgroup of* $\mathrm{SL}_2(\mathbb{Z})$*, let* $\Lambda(p) = \Lambda \cap \Gamma(p)$ *and* $\mathcal{F}(p) = \Lambda(p)\backslash\mathbb{H}$*. For* p *large enough*

$$\Omega(\mathcal{F}(p)) \cap \left[0, \frac{5}{36}\right) = \Omega(\mathcal{F}(1)) \cap \left[0, \frac{5}{36}\right).$$

Corollary. *Suppose that* $\delta > \frac{5}{6}$*. Then* $\Omega(\mathcal{F}(p))$ *has a spectral gap, that is for* p *large*

$$\lambda_1(\mathcal{F}(p)) \geq \min\left(\lambda_1(\mathcal{F}(1)), \frac{5}{36}\right).$$

The proof of Theorem 2, which generalizes the approach of Sarnak and Xue on cocompact arithmetic lattices [91] (also see [19]), stems form the observation that if $\Omega(\mathcal{F}(p))$ has a new eigenvalue λ in $[0, \frac{1}{4})$, it must be of high multiplicity:

$$m(\lambda, \mathcal{F}(p)) > \frac{p-1}{2}.$$

This follows from the result going back to Frobenius, that the smallest dimension of a nontrivial irreducible representation of $\mathrm{SL}_2(\mathbb{F}_p)$ is $\frac{p-1}{2}$, which is large compared to the size of the group[4], which is of order p^3.

The proof of the theorem hinges on bounding the multiplicity from above, more precisely, obtaining the bound

$$m(\lambda, \mathcal{F}(p)) \ll p^{6(1-s)},$$

where $\lambda = s(1-1)$. In order to extend the approach of Sarnak and Xue in obtaining this bound, the main obstacle to overcome is the infinitude of the volume of $\mathcal{F}(p)$, and the key analytic result proved in [29] is the collar lemmas, which state, roughly speaking, that for low-lying eigenvalues (below $\frac{1}{4}$) the L^2 norm of the eigenfunction of the Laplacian in a collar of fixed width, contiguous with a cusp or a flare, is of the same order of magnitude as its L^2 norm in the whole cusp or flare. In a sense, this lemma could be viewed as a generalization of the following fact about the zero eigenvalue and constant eigenfunction (area) in the cusp: hyperbolic area of the collar of width $\ln 2$ is the same as the area of contiguous cusp.

[4]This is in sharp contrast with the situation for (the non-expanding family of) symmetric groups S_n. Vershik and Kerov [104] proved that the ratio of the typical and maximal dimension of the irreducible representation of S_n to $\sqrt{|S_n|}$ decreases at the rate $e^{-c\sqrt{n}}$.

Arithmetic of the problem comes into play in the following estimate on the number of lattice points, (the implied constant is independent of p):

$$N_1(T, \Gamma(p)) \stackrel{\text{def}}{=} \sum_{\substack{\gamma \in \Gamma(p) \\ \|\gamma\| \le T}} 1 \ll \frac{T^{2+\epsilon}}{p^3} + \frac{T^{1+\epsilon}}{p^2} + 1. \tag{2.3}$$

It is only here that the fact that the homomorphism $\Lambda \to \mathrm{SL}_2(\mathbb{F}_p)$ is the reduction modulo p and not an arbitrary one is used; in fact most of the proof goes through with the much weaker assumption $\Lambda(1)/\Lambda(p) \cong \mathrm{SL}_2(\mathbb{F}_p)$.

Using Fell's continuity of induction the following result now follows:

Theorem 3 ([29]). *Let $S = \{A_1, \ldots, A_k\}$ be a symmetric set of generators in $\mathrm{SL}_2(\mathbb{Z})$, let $\Lambda = \langle A_1, \ldots, A_k \rangle$. If the Hausdorff dimension of the limit set, $\delta(L(\Lambda)) > 5/6$, then $\mathcal{G}_p = \mathcal{G}(\mathrm{SL}_2(\mathbb{F}_p), S_p)$ is a family of expanders.*

An example of a group satisfying the condition $\delta(L(\Lambda)) > \frac{5}{6}$ is the following. Let

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}; \qquad S_j = T^j S T^{-j}.$$

Let $\Lambda_n = \langle S, S_1, \ldots, S_n \rangle$. It is proved in [29] that $\delta(L(\Lambda_n)) > \frac{5}{6}$ for $n > 4392$.

We stress that for the group generated by $\langle \left(\begin{smallmatrix} 1 & 3 \\ 0 & 1 \end{smallmatrix}\right), \langle \left(\begin{smallmatrix} 1 & 0 \\ 3 & 1 \end{smallmatrix}\right) \rangle$, whose Hausdorff dimension was found to be 0.753 ± 0.003 by Phillips and Sarnak [83], the question remains open.

2.3 Expander graphs and invariant means

In this section we review the results of Y. Shalom [96] (see also [97]), who successfully exploited the intrinsic connection between the expanding property and the uniqueness of invariant mean on the profinite completion of the family of finite groups. Let $\cdots \xrightarrow{\phi_2} G_2 \xrightarrow{\phi_1} G_1$ be a directed sequence of epimorphisms of finite groups, and let $G = \varprojlim G_i$ be their *profinite completion*. There are natural continuous homomorphisms of G onto each G_i (projections), and if $\Gamma < G$ is any dense subgroup, then Γ is mapped onto each G_i as well. Thus, we have $\Gamma/N_i \cong G_i$, where $N_i \lhd \Gamma$ is a decreasing sequence of finite index subgroups. An equivalent way to define G in terms of Γ and the N_i's, is to let N_i form a base for the neighborhoods of identity in Γ; the completion of Γ with respect to a metric coming from this topology yields $G = \varprojlim \Gamma/N_i \cong \varprojlim G_i$. The group G is a compact topological group and as such supports a unique multiplication invariant probability measure, namely the Haar measure denoted by μ.

The link with expander graphs is given by the following theorem of Shalom:

Theorem 4 (Shalom [96]). *Let $G = \varprojlim G_i$ be as above and $\Gamma < G$ be a countable dense subgroup. The following are equivalent:*

1. *There exist a finite set* $S \subset \Gamma$ *such that* $X(G_i, S_i)$ *is an expander family. Here* S_i *denotes the natural projection of* S *to* G_i.

2. *The Haar measure* μ *is the* unique Γ*-invariant mean on* $L^\infty(G, \mu)$.

Theorem 4 is used as a fundamental tool in the (non-constructive) proof of the following result:

Theorem 5 (Shalom [96]). *Let* $\Gamma < \mathrm{SL}_2(\mathbb{Z})$ *be a subgroup which is* normal *in some congruence subgroup. If* $\lambda_0(\Gamma\backslash\mathbb{H}) < 21/100$, *then there exists* $N < \infty$ *such that the projection of* Γ *to* $\mathrm{SL}_2(\mathbb{F}_p)$ *is onto for* $p > N$, *and there exists a finite set* $S \subset \Gamma$ *such that* $\mathcal{G}(\mathrm{SL}_2(\mathbb{F}_p), S)$ *for* $p > N$ *is a family of expanders.*

The bound $21/100$ comes from the work of Luo, Rudnick and Sarnak [70]; the Selberg $1/4$ conjecture would imply the corresponding strengthening of the result.

Finally, we state the re-formulation of the independence problem mentioned in the introduction:

Proposition 2.1 (Shalom [96]). *In the notation of Theorem* 4 *the following are equivalent:*

1. *For* every k *and choices of any* k*-generators for each* G_i, *the corresponding Cayley graphs form an expander family.*

2. *For* every *finitely-generated dense subgroup* $\Gamma \subset G$, μ *is the unique* Γ*-invariant mean on* $L^\infty(G, \mu)$.

2.4 Random walk on a sphere

Locus classicus for the circle of ideas involving the uniqueness of the invariant mean is the problem posed by Ruziewicz in 1921: is Lebesgue measure on the n-sphere the unique *finitely* additive rotation-invariant measure defined on the Lebesgue subsets[5]? In 1923 (in the paper where the proof of the Hahn–Banach Theorem was presented) Banach [9] showed that for $n = 1$ the answer is negative, using essentially the amenability of SO(2). For $n > 3$ the affirmative answer was obtained in 1980/1 by Margulis [73] and Sullivan[99], who used Kazhdan's property (T). In 1984 Drinfeld [23] established the affirmative answer in the most difficult[6] case of $n = 2$ by proving the *existence* of an element in the group ring of SU(2) which has a spectral gap. Drinfeld's method appeals to some sophisticated machinery from the theory of

[5]The relation between these problems is that an invariant mean on $L^\infty(S^n)$ is a finitely-additive measure ν which is moreover absolutely continuous with respect to λ. From Hausdorff–Banach–Tarski [42, 101] paradoxical decomposition of S^n, $n \geq 2$, it follows that any rotationally invariant finitely additive measure on S^n, $n \geq 2$ is absolutely continuous with respect to λ. Hence for $n \geq 2$ the Ruziewicz problem is equivalent to the problem of invariant mean.

[6]As proved by Sarnak [88], the affirmative answer for $n = 2$ implies, via inductive construction, an affirmative answer for $n \geq 2$.

automorphic representations (in particular, Deligne's solution of the Ramanujan conjectures). The explicit and optimal construction, appealing to the above mentioned tools, was obtained by Lubotzky, Philips and Sarnak in [66].

In [30], a new robust method establishing that certain elements z in the group ring of SU(2) have a spectral gap is presented and consequently an elementary analytic solution of Ruziewicz problem is obtained. The method exploits the trace formula (2.11) to show that z has a spectral gap if the random walk associated with z returns to a small neighborhood of identity[7] not much more frequently than expected. In this section we review the method of [30], starting with a quick derivation of the trace formula (2.11) (see [102] for a survey of discrete trace formulae).

As is well-known, the irreducible representations of $G = \mathrm{SU}(2)$ are given by

$$\pi_N = \mathrm{sym}^N V, \quad N \geq 0,$$

where V is the standard two-dimensional representation of G. The dimension of π_N is $N + 1$, and it may be realized concretely by the linear action

$$(x, y) \ \rightarrow \ (\alpha x + \gamma y, \beta x + \delta y), \quad \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \in G, \tag{2.4}$$

on W_{N+1}, the space of homogeneous polynomials in (x, y) of degree N. The character χ_N of π_N at

$$g = \begin{bmatrix} e^{i\alpha} & 0 \\ 0 & e^{-i\alpha} \end{bmatrix} \quad \text{is} \quad \frac{\sin(N+1)\alpha}{\sin\alpha}. \tag{2.5}$$

Let

$$z_{g_1,\ldots,g_k} = g_1 + g_1^{-1} + \cdots + g_k + g_k^{-1},$$

and assume that Γ_z is free; as we review in Section 4.2, this condition is satisfied generically. We clearly have

$$z^n = \sum_{|w|=n} w,$$

where the sum is over all words in $g_1, \ldots, g_k$ of length n.

Denoting by $\lambda_{j,N}$ the eigenvalues of $\pi_N(z)$, and keeping in mind the character formula (2.5), we have that the image of this relation under π_N is

$$\mathrm{tr}\left(\pi_N(z)^n\right) = \sum_{j=0}^{N} \lambda_{j,N}^n = \sum_{|w|=n} \frac{\sin(N+1)r_w}{\sin r_w},$$

[7] We recall that estimating the probability of return to identity was the crucial starting point in Kesten's work (equation (1.1)) and remark that the estimate (2.3) can be viewed in terms of returns to the neighborhood of identity in $\mathrm{SL}_2(\mathbb{Z})$ in congruence topology.

where the element $g_w \in G$ corresponding to the word w is conjugate to the diagonal matrix

$$\begin{bmatrix} e^{ir_w} & 0 \\ 0 & e^{-ir_w} \end{bmatrix}$$

with $0 \leq r_w \leq \pi$ which is determined from

$$\mathrm{tr}(g) = 2\cos r_w. \tag{2.6}$$

It turns out that instead of using powers z^n it is more advantageous to consider n-th Chebyshev polynomial of the second kind, $U_n(\cos\theta) = \sin(n+1)\theta/\sin\theta$; letting $p = 2k-1$ (we do not assume that p is prime) we have the following relation which is easily established inductively [66]:

$$p^{n/2} U_n\left(z/(2\sqrt{p})\right) = \sum_{|\omega|\leq n}{}' \omega, \tag{2.7}$$

where the sum is over all *reduced* words ω in $g_1, g_1^{-1}, g_2, g_2^{-1}, \ldots g_k, g_k^{-1}$ of length $|\omega| = m \leq n$ with $m \equiv n(\mathrm{mod}2)$. The image of this relation under π_N yields

$$p^{n/2} U_n\left(\pi_N(z)/(2\sqrt{p})\right) = \sum_{|\omega|\leq n}{}' \pi_N(w). \tag{2.8}$$

Write the eigenvalues λ_j of $\pi_N(z)$ as

$$\lambda_j = 2\sqrt{p}\cos(\theta_{j,N}), \quad j = 0, 1, \ldots, N, \tag{2.9}$$

where

$$\begin{cases} \theta_{j,n} \in [0,\pi], & if\ |\lambda_j| \leq 2\sqrt{p}, \\ \theta_{j,n} = i\xi_{j,N},\ \xi_{j,N} > 0, & if\ \lambda_j > 2\sqrt{p}, \\ \theta_{j,n} = \pi + i\xi_{j,N},\ \xi_{j,N} > 0, & if\ \lambda_j < -2\sqrt{p}. \end{cases} \tag{2.10}$$

We call the $\theta_{j,N}$'s not in $[0,\pi]$ *exceptional*. Indeed, since Γ_z is free, most of the θ's are not exceptional as $N \to \infty$.

Taking the trace of both sides of (2.8) yields the following *trace formula*:

$$p^{n/2} \sum_{j=0}^{N} \frac{\sin(n+1)\theta_{j,N}}{\sin\theta_{j,N}} = \sum_{|\omega|\leq n}{}' \frac{\sin(N+1)r_\omega}{\sin r_\omega}\,. \tag{2.11}$$

This trace formula was used in [30] to show that z has a spectral gap if the random walk associated with z returns to a small neighborhood of identity not much more frequently than expected. We proceed to briefly outline how this reduction is performed.

Note that if n is even (which we assume henceforth in this section), then for exceptional $\theta_{j,N}$ we have

$$\frac{\sin(n+1)\theta_{j,N}}{\sin\theta_{j,N}} = \frac{\sinh(n+1)\xi_{j,N}}{\sinh\xi_{j,N}} > 0. \tag{2.12}$$

Hence (2.11) becomes

$$\sum_{\xi_{j,N}} \frac{\sinh(n+1)\xi_{j,N}}{\sinh\xi_{j,N}} + O(Nn) = p^{-n/2} \sum_{0<|\omega|\le n}{}' \frac{\sin(N+1)r_\omega}{\sin r_\omega}. \tag{2.13}$$

In order to exploit the cancellation in the sum on the right hand side of (2.13) we sum over N in the range $N \sim N_0$, N_0 large; applying Poisson summation formula we get:

$$\sum_{N\sim N_0} \sum_{\xi_{j,N}} \frac{\sinh(n+1)\xi_{j,N}}{\sinh\xi_{j,N}} \ll N_0^2 n + p^{-n/2} \#\left\{0 < |\omega| < n : r_\omega \in \left[0, \frac{1}{N_0}\right]\right\}. \tag{2.14}$$

Thus, estimating the left hand side above is reduced to estimating the number of words ω', $|\omega'| < n$ with small rotation (i.e., of size $\frac{1}{N_0}$). To estimate the expected number of such r_ω's we go back to (2.11) and note that for N *fixed*, if we let $n \to \infty$ and use the fact that $-2k < \lambda_{j,N} < 2k$ for $N \ge 1$, we get

$$\lim_{n\to\infty} p^{-n} \sum_{|\omega|\le n}{}' \frac{\sin(N+1)r_\omega}{\sin r_\omega} = 0. \tag{2.15}$$

The number of $|\omega| \le n$ is $(p+1)p^{n-1}$, so (2.15) asserts that the rotations r_ω, $|\omega| \le n$ become equidistributed with respect to $\sin^2\theta\, d\theta/\pi$, that is with respect to the Weyl measure which is the image of the Haar measure on G onto the maximal torus. We might expect at least generically for $(g_1, \ldots, g_k) \in G^{(k)}$ that this continues to hold approximately for small intervals. That is, that

$$\#\{\omega | r_\omega \in [0, N_0^{-1}]\} \ll \frac{\#\{\omega : |\omega| \le n\}}{N_0^3} \ll \frac{p^n}{N_0^3}. \tag{2.16}$$

Granting (2.16), it is not difficult to deduce (using (2.12) and (2.14)) that for sufficiently large N_0 and any j we have:

$$|\lambda_{j,N_0}| \le p^{1/2}(p^{1/3} + p^{-1/3}). \tag{2.17}$$

So the explicit spectral gap (2.17) is what we expect to hold for the generic set $(g_1, g_2, \ldots, g_k)$.

While we don't know how to establish (2.16) for the generic $(g_1, \ldots, g_k)$, we can prove it for some special choices of g_i coming from Hamilton quaternions, whose definition we now recall.

Let $H(\mathbb{Z})$ denote the ring of Hamilton quaternions $\alpha = x_0 + x_1 i + x_2 j + x_3 k$, $x_j \in \mathbb{Z}$. Let $\bar{\alpha} = x_0 - x_1 i - x_2 j - x_3 k$ and $N(\alpha) = \alpha\bar{\alpha}$. For $q \geq 3$ a prime number let $\tilde{g}_1, \tilde{g}_2, \ldots, \tilde{g}_k$ be a subset of $S = \{\alpha \in H(\mathbb{Z}) | N(\alpha) = q\}$ (it is well-known [41] that the latter has $8(q+1)$ elements) satisfying

(I) $\tilde{g}_{j_1} \neq \varepsilon \tilde{g}_{j_2}$ for $j_1 \neq j_2$ and $\varepsilon \in \{\pm 1, \pm i, \pm j, \pm k\}$ a unit.

(II) $\tilde{g}_{j_1} \neq \varepsilon \overline{\tilde{g}_{j_2}}$ for any j_1, j_2 and ε a unit.

The homomorphism of $H(\mathbb{R})$ into SU(2)

$$\alpha \rightarrow \frac{1}{N(\alpha)} \begin{bmatrix} x_0 + x_1 i & x_2 + x_3 i \\ -x_2 + x_3 i & x_0 - x_1 i \end{bmatrix}$$

gives us the corresponding elements $g_1, g_2, \ldots, g_k \in G$. As is well-known (see [66] for example) different reduced words $\omega = R(\tilde{g}_1, \overline{\tilde{g}_1}, \ldots \overline{\tilde{g}_k})$ in $\tilde{g}_1, \ldots, \overline{\tilde{g}_k}$ of length $m \geq 1$ give different quaternions of norm q^m. In particular, to each such word corresponds a unique solution of

$$x_0^2 + x_1^2 + x_2^2 + x_3^2 = q^m, \quad x_0^2 \neq q^m. \tag{2.18}$$

This allows us to estimate the geometric side in (2.16) using elementary arithmetic to obtain:

Theorem 6 ([30]). *Let $q \geq 3, g_1, g_2, \ldots, g_k \in G$ be as above. If $p = 2k-1 > q^{4/5}$, then $z = g_1 + g_1^{-1} + \cdots + g_k + g_k^{-1}$ has a gap, in fact*

$$\overline{\lim}_{N\to\infty} ||\pi_N(z)|| \leq p^{1/2} \left(\frac{q^{2/3}}{p^{1/3}} + \frac{p^{1/3}}{q^{2/3}} \right) < 2k.$$

For example, if $q = 7$,

$$\tilde{g}_1 = 2 - i + j + k, \quad \tilde{g}_2 = 2 - i - j + k, \quad \tilde{g}_3 = 2 + i - j + k$$

satisfy the hypotheses and denoting the corresponding z by $\tilde{z}_7$ we have

$$\overline{\lim}_{N\to\infty} ||\pi_N(z)|| \leq 5.83 \ (< 6).$$

If $q \equiv 1 (\mathrm{mod}\ 4)$, and we choose a maximal such subset $\tilde{g}_1, \ldots, \tilde{g}_k$ of S above, then $2k - 1 = q$, and we get an element z (denoted by z_q), which was shown in [66] to be Ramanujan (as defined following the statement of Problem 1 in the introduction), using automorphic forms and, in particular, Deligne's proof of the Ramanujan Conjectures. For example, for $q = 5$, $\mathrm{supp}(z_5)$ is given by the following matrices:

$$g_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 1+2i & 0 \\ 0 & 1-2i \end{pmatrix}, \quad g_2 = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix}, \quad g_3 = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & 2i \\ 2i & 1 \end{pmatrix}. \tag{2.19}$$

We remark that recently L. Clozel [18] constructed the analogue of Ramanujan elements for higher-dimensional spheres of odd dimensions.

3 Overview of related results

3.1 Zig-zag product

In a recent breakthrough paper Reingold, Vadhan and Wigderson [85] introduced a notion of zig-zag product of graphs and gave a purely combinatorial construction of expander graphs.

Let $\mathcal{G}$ be a d-regular graph, and M its adjacency matrix. We denote by $\lambda(\mathcal{G})$ the second largest (in absolute value) eigenvalue of M/d. Following [85], we say that a graph $\mathcal{G}$ is an $[n, d, \lambda]$-graph if it is a d-regular graph on n vertices, and $\lambda(\mathcal{G}) \leq \lambda$.

Assume that the edges of our d-regular graphs are arbitrarily labelled (colored) by $[d]$ in a one-to-one fashion. Fixing a color $i \in [d]$ and a vertex v, let $v[i]$ be the neighbor of v along the edge colored i. The zig-zag product of two graphs $\mathcal{G}_1 \textcircled{z} \mathcal{G}_2$, with $\mathcal{G}_1$ a $[n_1, d_1, \lambda_1]$-graph and $\mathcal{G}_2$ a $[d_1, d_2, \lambda_2]$-graph is now defined as follows:

Definition 3.1 ([85]). *Let $\mathcal{G}_1$ be a d_1-regular graph on $[n_1]$ and $\mathcal{G}_2$ a d_2-regular graph on $[d_1]$. Then $\mathcal{G}_1 \textcircled{z} \mathcal{G}_2$ is a d_2^2-regular graph on $[n_1] \times [d_1]$ defined as follows: For all $v \in [n_1]$, $k \in [d_1]$, $i, j \in [d_2]$, the edge (i, j) connects the vertex (v, k) to the vertex $(v[k[i]], k[i][j])$.*

Intuitively, there is a "cloud" of vertices of $\mathcal{G}_2$ around every original vertex of $\mathcal{G}_1$. Two vertices (v, k) and (u, l) are adjacent, if we can travel between them in a "zig-zag" path of length 3: one step on $\mathcal{G}_2$ in the v cloud, then switching to the u cloud according to an edge and labeling of $\mathcal{G}_1$, and a final step on $\mathcal{G}_2$ in the u cloud.

The basic property of zig-zag product proved in [85] is the following eigenvalue bound:

Theorem 7 ([85]). *With the notation as above we have*

$$[n_1, d_1, \lambda_1] \textcircled{z} [d_1, d_2, \lambda_2] \to [n_1 \cdot d_1, d_2^2, \lambda_1 + \lambda_2 + \lambda_2^2].$$

Using this result, Reingold, Vadhan, and Wigderson gave for the first time an entirely combinatorial explicit construction of constant-degree expander graphs. For the purposes of discussion below, the main implication of Theorem 7 is that whenever $\{\mathcal{G}_1\}$ and $\{\mathcal{G}_2\}$ are families of expanders, so is $\{\mathcal{G}_1 \textcircled{z} \mathcal{G}_2\}$; conversely, if either $\{\mathcal{G}_1\}$ or $\{\mathcal{G}_2\}$ fails to be an expander family, so does $\{\mathcal{G}_1 \textcircled{z} \mathcal{G}_2\}$.

Definition 3.1 was slightly extended in [2] to include the case of $\mathcal{G}_2$ being an $[m, d, \mu]$-graph, and to let $\mathcal{G}_1$ being an $[n_1, cd_1, \lambda_1]$-graph, which is the (edge) disjoint union of c copies of d_1-regular graphs on the same set of vertices; in the slightly more general definition, the middle step in the "zigzag" is stochastic with c possibilities, whereas it was deterministic in the original definition of [85]. The basic eigenvalue bound of Theorem 7 carries over without change to this more general situation.

3.2 Example of nonuniform groups

Reinterpreting the notion of zig-zag product in the context of semi-direct product of groups, Alon, Lubotzky and Wigderson [2] constructed a family of groups G_i which are expanders with respect to one choice of generators and not with respect to another such choice.

Let A and B be finite groups. Assume that B acts on A, namely there is a homomorphism from B to the automorphism group of A. For elements $a \in A, b \in B$ we use a^b to denote the action of b on a. We also use a^B to denote the orbit of a under this action. Recall that the *semi-direct product* of A and B, denoted $A \rtimes B$, is the group whose elements are the ordered pairs $\{(a, b) : a \in A, b \in B\}$, with the group operation defined by

$$(a_1, b_1)(a, b) = (a_1 a^{b_1^{-1}}, b_1 b).$$

Let α be a generating (multi)set for A. We will work only with symmetric generating (multi)sets, namely the number of occurrences of a and a^{-1} in α is the same for every $a \in A$. The edges of a Cayley graph of a group A with a (multi)set of generators α are naturally labeled as follows: the label of (x, xa) near x is a, and its label near xa is a^{-1}. Note that the graph is $|\alpha|$-regular. Assume that B acts on A as above. Let α, β be sets of generators for A, B respectively, and further assume that α is a (disjoint) union of B-orbits, namely $\alpha = \bigcup_{i=1}^{c} a_i^B$. Define the following set of generators for $A \rtimes B$:

$$\gamma = \{(1, b)(a_i, 1)(1, b') : b, b' \in \beta,\ i \in [c]\}. \tag{3.1}$$

Note that $|\gamma| = c|\beta|^2$, and that $\mathcal{G} = X(A, \alpha)$ is the edge-disjoint union of the c graphs $\mathcal{G}_i = \mathcal{G}(A, a_i^B)$, where the edges of $\mathcal{G}_i$ around every vertex are labelled by the elements of B in the obvious way. We can thus define the zigzag product $\mathcal{G}(A, \alpha) ⓩ \mathcal{G}(B, \beta)$ and notice, using the definitions of semi-direct product and zig-zag product, that following a generator of γ from an element of $A \rtimes B$ leads us to its "zig-zag" neighbor in that group. This gives the connection between the two types of product:

Theorem 8 (Alon, Lubotzky, Wigderson [2]). *Notation being as above, we have*

$$C(A \rtimes B, \gamma) = C(A, \alpha) ⓩ C(B, \beta).$$

Thus, if $\mathcal{G}(A, \alpha)$ and $\mathcal{G}(B, \beta)$ are expanders, and $|\beta|$, c are constants, then *regardless* of the size of α, the graph $\mathcal{G}(A \rtimes B, \gamma)$ is a constant degree expander.

Theorem 8 is applied in [2] to show that the family of groups G_i of the form $A_i \rtimes B_i$, where $B_p = SL_2(F_p)$ and $A_p = F_2^{P_1}$ with $P_1 = F_p \cup \{\infty\}$, is not uniformly expanding, i.e., it is expanding with respect to one set of generators S_1 but not with respect to another such set S_2. In the notation of equation (3.1), the generators for B_p can be chosen as $b = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, b' = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ for both S_1 and S_2. For the non-expanding generating set S_2, $a(p)$ can be chosen to be any fixed unit vector in A_p. The (nonconstructive) choice of $a_1(p)$ and $a_2(p)$ for the expanding set S_1 is based on a counting argument; we refer to [2, 69] for details.

3.3 Uniform exponential growth

If the family $G_i = \Gamma/N_i$ is a uniform expanding family, then one can deduce that Γ is of uniform exponential growth, at least with respect to generating sets of size at most k for some k. In a recent work Eskin, Mozes and Oh [25] showed that in characteristic 0 every linear group of exponential growth is of uniform exponential growth, thus giving some support to $\{\mathrm{SL}_2(\mathbb{F}_p)\}$ being a uniform expanding family.

We remark that uniform exponential growth for hyperbolic groups was established by M. Koubi [56] and for solvable groups with exponential growth by D. Osin [80]. The results in [25, 56, 80] are based on the observation that the groups under consideration contain uniformly a free sub semi-group; see [35] for a general discussion of this and related questions.

3.4 Uniform diameter bounds

The diameter of a graph $\mathcal{G}$ is the length of the longest path between two vertices of $\mathcal{G}$. Given a finite group G with a symmetric set of generators $\Sigma = \{g_1, g_1^{-1}, \ldots, g_k, g_k^{-1}\}$, the Cayley graph $\mathcal{G}(G, \Sigma)$, is a graph which has elements of G as vertices and which has an edge from x to y if and only if $x = \sigma y$ for some $\sigma \in S$. The diameter of $\mathcal{G}(G, \Sigma)$, or, equivalently, the diameter of group G with respect to the set of generators Σ, is the maximum over $g \in G$ over the shortest word in Σ representing g. This quantity is of great interest in connection with efficient communications networks as well as in the study of various random walks on groups; see [6, 7, 63, 61] and references therein.

A simple count of words shows that $\mathrm{diam}(\mathcal{G}(G, \Sigma)) \geq \log_\Sigma(|G|)$; so for Σ bounded it follows that the optimal diameter bound is $O(\log|G|)$. It was proved by Babai, Kantor and Lubotzky [6] that there is a constant C of order 10^{10}) such that every nonabelian finite simple group G has a set Σ of at most 14 generators for which the diameter $\mathrm{diam}(\mathcal{G}(G, \Sigma))$ is at most $C \log|G|$.

If the family of groups $\{G_i\}$ is expanding with respect to the set of generators Σ_i then $\mathrm{diam}(\mathcal{G}(G_i, \Sigma_i)) = O(\log|G_i|)$ (see for example [17].) In [33] it is shown that the family of groups $\mathrm{PSL}_2(\mathbb{Z}/p^n\mathbb{Z})$ has uniform polylog diameter with respect to the set of generators Σ which forms a dense subgroup of $\mathrm{PSL}_2(\mathbb{Z})$; this condition is satisfied in particular if Σ generates a free subgroup of $\mathrm{PSL}_2(\mathbb{Z})$ (so for example

$$\Sigma = \left\{\left(\begin{smallmatrix}1 & 3\\ 0 & 1\end{smallmatrix}\right), \left(\begin{smallmatrix}1 & 0\\ 3 & 1\end{smallmatrix}\right)\right\}$$

is covered.)

The proof in [33] follows the approach in the Solovay–Kitaev theorem [78] and is easily generalized to groups $\mathrm{PSL}_m(\mathbb{Z}/p^n\mathbb{Z})$ for $m \geq 3$; it also gives explicit short word representation for elements in $\mathrm{PSL}_2(\mathbb{Z}/p^n\mathbb{Z})$.

3.5 Uniform Kazhdan constant

Related to Independence Problem is a question about uniform Kazhdan constant posed by Lubotzky in [64]. Recall that a discrete group Γ generated by a finite set S has Kazhdan property (T) if there exists a positive constant $\varepsilon(S)$ such that for every unitary representation $(\pi, \mathcal{H})$ of Γ with no invariant vectors and for any $u \in \mathcal{H}$ there exists $s \in S$ such that $\|\pi(s)u - u\| \geq \varepsilon(S)\|u\|$. Such a constant is called a Kazhdan constant with respect to the set S. That a given group has property (T) does not depend on the choice of generators; the question posed by Lubotzky is whether for a group with property (T) there exists a uniform positive Kazhdan constant independent of S.

For a unitary representation $(\pi, \mathcal{H})$ define the uniform Kazhdan constant $K(\pi, \Gamma)$ as follows:

$$K(\pi, \Gamma) = \inf_{S} \inf_{0 \neq u \in \mathcal{H}} \max_{s \in S} \frac{\|\pi(s)u - u\|}{\|u\|}, \tag{3.2}$$

where the infimum is taken over all finite generating sets S. We say that Γ has *uniform property T* if $\inf_\pi K(\pi, \Gamma) > 0$, where the infimum is taken over all unitary representations with no invariant vectors.

In a recent paper [34] Gelander and Żuk gave examples of Kazhdan groups without uniform Kazhdan constant:

Theorem 9 ([34]). *Let Γ be a Kazhdan group densely embedded in a connected topological group G. Assume that there exist a continuous unitary representation $(\pi_G, \mathcal{H})$ without invariant vectors. Then Γ does not have uniform property T.*

An example of G and Γ that satisfy the assumptions of the theorem is $G = \mathrm{SO}_n(R)$ and $\Gamma = \mathrm{SO}_n(\mathbb{Z}[1/5])$ for $n \geq 5$ (see [73]) or $\Gamma = \mathrm{SO}(Q) \cap \mathrm{SL}_n(\mathbb{Z}[\sqrt{2}])$, where Q is the quadratic form $Q(x) = x_1^2 + \cdots + x_{n-2}^2 - \sqrt{2}(x_{n-1}^2 + x_n^2)$ ([99]). These are precisely the groups used by Margulis and Sullivan to affirmatively answer Banach–Ruziewicz problem for higher-dimensional spheres.

It was pointed out by the referee that a larger class of Kazhdan groups without uniform property (T) was provided by D. Osin in [81], where it is proved that any hyperbolic group with property (T) does not have it uniformly, and furthermore the infimum is zero even if we take it only over finite generating sets with fixed number of generators. To the best of our knowledge, the question of the existence of groups with uniform property (T) is open.

3.6 Unbounded number of generators

The problems become much more tractable if we let the number of generators go to infinity. Concerning expansion property the following result was proved by Alon and Roichman.

Theorem 10 (Alon and Roichman [4]). *For every $1 > \delta > 0$ there is a $c(\delta) > 0$ such that the following holds. Let G be a group of order n, and let S be a set of $c(\delta) \log_2 n$*

random elements of G. Denote by λ_1^* the second largest eigenvalue of the normalized adjacency matrix. Then

$$E(|\lambda_1^*[X(G,S)]|) < 1 - \delta.$$

Corollary. *For every $1 > \epsilon > 0$ there exists a $c(\epsilon)$ such that the following holds. Let G be a group of order n, and let S be a random set of $c(\epsilon)\log_2(n)$ elements of G, then the Cayley graph $X(G,S)$ is an ϵ-expander almost surely. That is to say, the probability it is such an expander tends to 1 as n tends to infinity.*

Finally, by means of transition to the numerical results about spacings distribution presented in the next section, we mention the following result which holds in great generality and admits a short simple proof, which we present in the case of $G = \mathrm{SU}(2)$ and N even.

Proposition 3.1 ([30]). *Let $\nu_{n,k}$ be the direct image of $dg_1 dg_2 \dots dg_k$ on $G^{(k)}$ under the map*

$$(g_1, g_2, \dots, g_k) \to \frac{1}{\sqrt{k}}\pi_N(g_1 + g_1^{-1} + \dots + g_k + g_k^{-1}).$$

Thus, $\nu_{n,k}$ is a probability measure on $\mathcal{H}_{N+1}$, the space of $(N+1)\times(N+1)$ real symmetric matrices. As $k \to \infty$, $\nu_{n,k}$ converges in measure to the standard GOE measure on $\mathcal{H}_{N+1}$.

Proof. The measure $\nu_{n,k}$ on the real vector space $\mathcal{H}_{N+1}$ is a sum of i.i.d. random variables. We may therefore appeal to the general vector valued central limit theorem. The distribution of the individual summand is $H(g) = \pi_N(g) + \pi_N(g^{-1})$, $g \in G$. Thus, once we show that $\int_G H(g)dg = 0$ we know that the limit of $\nu_{n,k}$ as $k \to \infty$ is Gaussian. The issue is to identify this limit and its support.

When N is even, the matrix $H(g) = (h_{ij}(g))$ is real symmetric, so we can consider the $(N+1)(N+2)/2$ dimensional space h_{ij}, $1 \le i \le j \le N+1$, i.e., $\mathcal{H}_{N+1}$. We assert that

$$\int_G h_{ij}(g)\, dg = 0 \tag{3.3}$$

for $1 \le i \le j \le N+1$, and

$$\int_G h_{ij}(g)h_{rs}(g)\, dg = \frac{2}{N+1}\left(\delta_{is}\delta_{jr} + \delta_{ir}\delta_{js}\right). \tag{3.4}$$

To see this recall that since π_N is irreducible we have from Schur's Lemma (assuming $N \ge 1$) that

$$\int_G 1 \cdot \pi_N(i,j)(g)dg = 0, \tag{3.5}$$

and

$$\int_G \pi_N(i,j)(g)\pi_N(m,n)(g^{-1})dg = \frac{\delta_{in}\delta_{jm}}{N+1} \tag{3.6}$$

for any $1 \le i, j, m, n \le N+1$.

Thus, (3.3) follows from (3.5), while (3.4) follows from (3.6) together with the fact that $\pi_N(g)$ is orthogonal, and hence

$$\int_G \pi_N(i,j)(g)\pi_N(r,s)(g)dg = \int_G \pi_N(i,j)(g)\pi_N(s,r)(g^{-1})dg = \frac{\delta_{ir}\delta_{js}}{N+1}.$$

The equalities (3.3) and (3.4) identify the covariance-matrix for the limiting Gaussian law as:

$$C_N e^{-\frac{N+1}{4}(h_{11}^2+\cdots+h_{N+1,N+1}^2)-\frac{N+1}{2}(\sum_{1\le i<j\le N+1} h_{ij}^2)} \prod_{1\le i\le j\le N+1} dh_{ij}$$
$$= C_N e^{-\frac{N+1}{4}\mathrm{tr}(H^2)} \prod_{i\le j} dh_{ij}.$$

This Gaussian law is $\mathrm{O}(N+1)$ invariant on $\mathcal{H}_{N+1}$ and is exactly the GOE measure on $\mathcal{H}_{N+1}$, see [77, p. 39]. □

4 Spacings and discrepancy

4.1 Numerical results

An important model for understanding quantization of classical chaotic systems is afforded by symplectic maps [108]. The simplest of these are the linear area-preserving transformations of the torus $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$; that is, transformations $\binom{x_1}{x_2} \mapsto A\binom{x_1}{x_2}$ with $A \in \mathrm{SL}(2,\mathbb{Z})$. These transformations, which have received a lot of attention in the physics and mathematics literature, go by the name "cat maps," which derives from the pictures in [5] which show a cartoon cat face and its images under a few iterates of A, displaying the chaotic features of $x \mapsto Ax$.

The quantization of such a linear transformation can be carried out by periodizing any one of the standard quantization procedures in $\mathbb{R}^2$. This has been carried out first by Hannay and Berry in [40] and has since been studied by many authors, see [8, 12, 20, 21, 51, 55] and references therein. We will adopt the quantization procedure given by Kurlberg and Rudnick in [57]. It yields for each integer $N \ge 1$ ("$N = 1/\hbar$") a unitary matrix $\pi_N(A)$ acting on $L^2(\mathbb{Z}/N\mathbb{Z})$. As detailed in [57], $\pi_N(A)$ is essentially the Weil or metaplectic representation of A reduced modulo N (first considered by Kloosterman [54]).

The behavior of the eigenstates of $\pi_N(A)$ has been the subject of intensive investigations in papers cited above, with important recent breakthroughs by Kurlberg and Rudnick [57]. The distribution of the eigenvalues of $\pi_N(A)$ is degenerate, and not what is expected for the quantization of a generic chaotic system, as shown by Keating [49]. Recently, two ways of recovering the predicted random matrix distribution for modified cat maps have been proposed. One, considered by Keating and Mezzadri

in [50], is to perturb a cat map by nonlinear shears; another, considered by Keppeler, Marklof, and Mezzadri in [52], is to couple a cat map with a two-spinor processing in a magnetic field.

In [31] it was shown how to recover the RMT predictions while staying within the framework of linear maps and representation theory. The basic idea, following [30], is to consider the ergodic action of the group generated by *several* linear toral automorphisms (i.e., "several maps of a cat").

More precisely, let $\Gamma_{A_1,\dots,A_k}$ be the group generated by the transformations $A_1, \dots, A_k$ with $A_i \in \mathrm{SL}(2,\mathbb{Z})$. Recall that the action of the group Γ is *strongly ergodic* if the associated element in the group ring

$$z_{A_1,\dots,A_k} = A_1 + A_1^{-1} + \cdots + A_k + A_k^{-1}$$

has a spectral gap [92]. We consider the quantizations $\pi_q(z)$,

$$\pi_q(z) = \pi_q(A_1) + \pi_q(A_1^{-1}) + \cdots + \pi_q(A_k) + \pi_q(A_k^{-1}) \ .$$

For technical reasons we restrict ourselves to primes $q \equiv 1 \bmod 4$. The representation π_q is not irreducible but decomposes into two irreducible components π_q^- and π_q^+ of dimensions $\frac{1}{2}(q-1)$ and $\frac{1}{2}(q+1)$ respectively. With a suitable choice of basis, $\pi_q^+(z)$ lies in the space of real symmetric matrices, while $\pi_q^-(z)$ lies in the linear space of matrices H satisfying

$$H^* = H, \quad J^t H J = H^t, \quad J = \begin{pmatrix} E & 0 & \dots & 0 \\ 0 & E & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & E \end{pmatrix}, \quad E = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \quad (4.1)$$

In particular, in the latter case the eigenvalues of $\pi_q^-(z)$ are of the form $\lambda_1, \lambda_2, \dots, \lambda_M$, where $M = (q-1)/4$, and each λ_j occurs with multiplicity 2. The numerical experiments in [31] indicate that the unfolded consecutive spacing distribution for "generic" z follows the GSE law of Random Matrix Theory [77] for $\pi_q^-(z)$ and GOE law of Random Matrix Theory for $\pi_q^+(z)$ (see Figure 1); similar results hold for SU(2) [30].

We also consider Ramanujan elements related to Ramanujan graphs of Lubotzky, Phillips and Sarnak (see [19]); for them the unfolded consecutive spacing distributions of $\pi_q^-(z_p)$ and $\pi_q^+(z_p)$ follow the Poisson distribution (see Figure 2), in exact analogy to the arithmetic manifolds [93, 13, 15].

One way in which the difference between RMT and the Poisson distribution manifests itself is in the speed of convergence to the Kesten measure: the convergence is much faster in the RMT case [89, 90]. Numerical experiments in [31] indicate that for a fixed z all irreducible representations of $\mathrm{SL}_2(\mathbb{F}_q)$ behave in way similar to π_q^- and π_q^+ (depending on parity) with respect to spacings distributions; this is consistent with the observations made in [58, 59, 60] regarding the similarity of spectral properties of various irreducible representation of $\mathrm{SL}_2(\mathbb{F}_q)$. As a result, the different behavior

with respect to the speed of convergence to Kesten's measure is also apparent when we consider the image of z in the regular representation of $\mathrm{SL}_2(\mathbb{F}_q)$.

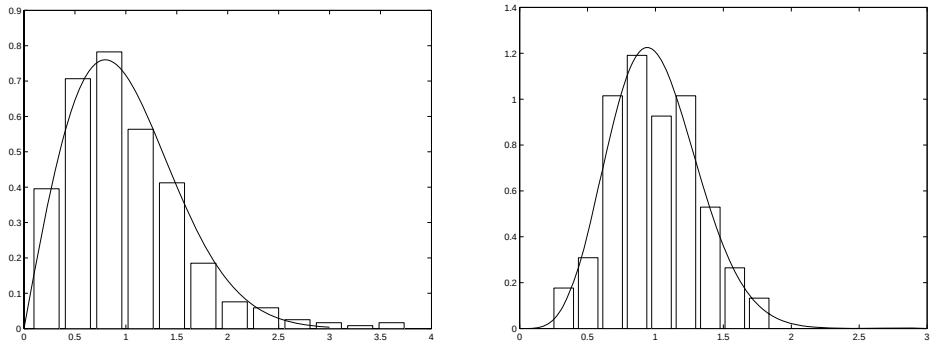
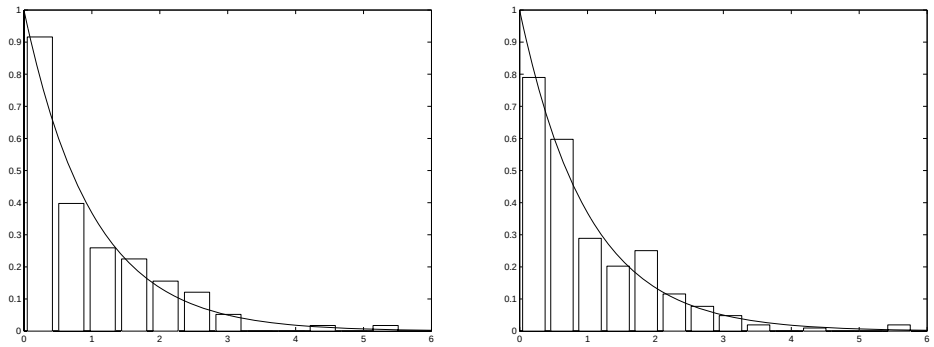


Figure 1. Spacing distribution for a random z with $|\operatorname{supp}(z)| = 3$ vs. the Wigner surmise for the GOE (left) for the irreducible component π^- of the Weil representation π_{509} coming from the principal series representation, and the spacing distribution vs. the Wigner surmise for the GSE (right) for the irreducible component of the Weil representation π^+, coming from the discrete series representation.



Fibure 2. Spacing distribution for the Ramanujan element z_5 vs. the exponential distribution for the irreducible components of the Weil representation π_{509} coming from the principal series representation (π^-), and from the discrete series representation π^+.

Figure 3 shows the "density of states" versus the Kesten measure for random elements and for Ramanujan (Lubotzky–Phillips–Sarnak) elements for the regular representation, together with the distribution at an individual irreducible representation. For "generic" elements in $\mathrm{SU}(2)$ and $\mathrm{SL}_2(\mathbb{Z})$ satisfying Diophantine condition[8]

[8] For elements z_S in $\mathrm{SL}_2(\mathbb{Z})$ the analogue of the Diophantine condition (in congruence topology) amounts to the logarithmic girth bound for the associated Cayley graphs; this is established for Γ_z free in [31].

(we use this terminology to draw an analogy with Diophantine approximation; see next section), the upper bound of order $\frac{1}{\log N}$ is established in [30] and [31]. This bound is probably very far from the true upper bound of order $\frac{\log N}{N}$ predicted by Random Matrix Theory. For Ramanujan elements the sharp lower bound of order $\frac{1}{\sqrt{N}}$ is established (using trace formulae) in [30, 31]: this is the analogue of the lower bounds for the remainder term in Weyl's law for arithmetic hyperbolic surfaces [43, 71].

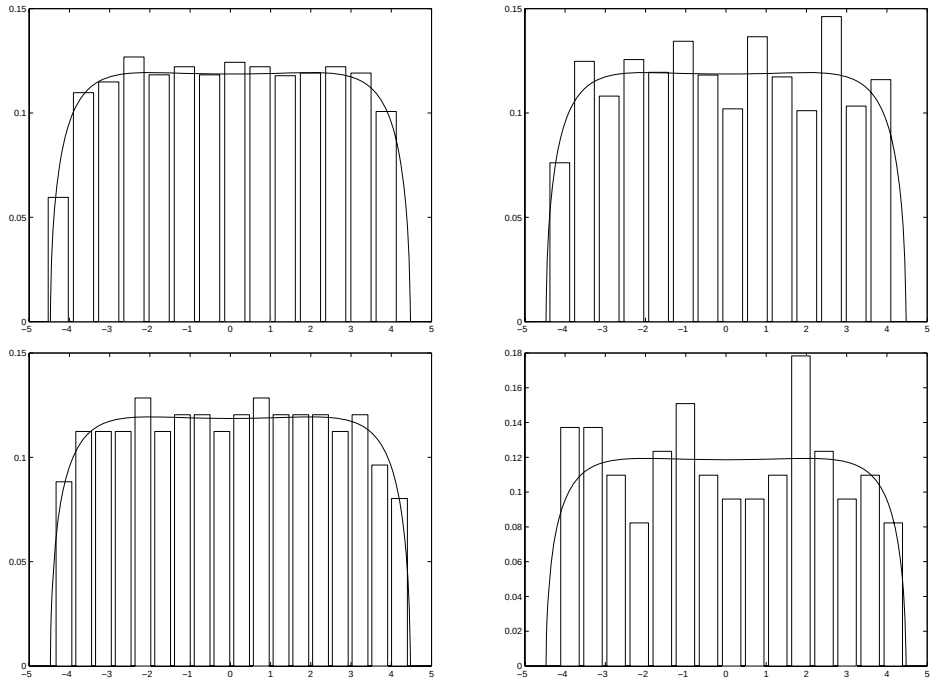


Figure 3. Eigenvalue density plots for empirical distributions versus the Kesten measure. The top row shows the distribution of the full spectrum of a random 6-regular Cayley graph (left) and the Ramanujan element z_5 for $\mathrm{SL}_2(\mathbb{F}_{61})$ (right), corresponding to the spectrum of the regular representation. The bottom row shows the distribution of the spectrum for random and Ramanujan elements at an individual irreducible representation for $\mathrm{SL}_2(\mathbb{F}_{509})$.

4.2 Noncommutative diophantine property

As mentioned above, the speed of convergence in (1.8) turns out to be related to the Diophantine properties of $g_1, \ldots, g_k \in G$:

Definition 4.1 ([30]). *For $k \geq 2$, we say that $g_1, g_2, \ldots, g_k \in G$ are Diophantine (or satisfy a Diophantine condition) if there is $B = B(g_1, \ldots, g_k) > 0$ such that for any $m \geq 1$ and a word R_m in $g_1, g_2, \ldots, g_k$ of length m with $R_m \neq \pm e$ we have*

$$||R_m \pm e|| \ \geq \ B^{-m}.$$

where

$$\left|\left|\left[\begin{array}{cc} a & b \\ c & d \end{array}\right]\right|\right|^2 = |a|^2 + |b|^2 + |c|^2 + |d|^2.$$

Note that for $g \in G$ we have

$$||g \pm e||^2 = 2|\mathrm{trace}(g) \mp 2|. \tag{4.2}$$

For example, if $\Gamma_{(g_1,\ldots,g_k)}$ is finite then $g_1, \ldots, g_k$ are Diophantine, however we are mainly interested in the case when $\Gamma_{(g_1,\ldots,g_k)}$ is free. In this case it follows by a pigeon hole argument similar to Dirichlet's that for any $m \geq 1$ there always is a word $R \neq \pm e$ in $g_1, g_1^{-1}, \ldots, g_k, g_k^{-1}$ of length at most m satisfying

$$||R - e|| \leq \frac{10}{(2k-1)^{m/6}} \tag{4.3}$$

(here and elsewhere we assume that $k \geq 2$). This shows that the exponential behavior in the definition of Diophantine is the appropriate one.

As was first exploited by Hausdorff [42], for $G = \mathrm{SU}(2)$ the relation

$$R_m(g_1, g_1^{-1}, \ldots, g_k, g_k^{-1}) = e,$$

where R is a *reduced* word of length $m \geq 1$, is not satisfied identically in $G^{(k)}$. Hence the sets

$$V(R_m) \ := \ \{(g_1, \ldots, g_k) | R_m(g) = e\}$$

are of codimension at least one in $G^{(k)}$. It follows that $\bigcup_{m\geq 1} V(R_m)$ is of zero measure in $G^{(k)}$, and also it is of the first Baire category in $G^{(k)}$. Thus, the generic $(g_1, \ldots, g_k) \in G^{(k)}$ (in both senses) generates the free group. This holds quite generally as proved by D.B.A. Epstein:

Theorem 11 (D. B. A. Epstein [24]). *Let G be a connected, finite-dimensional non-solvable Lie group. Then for each $k > 0$, and for almost all k-tuples $(g_1, \ldots, g_k)$ of elements of G, the group generated by $g_1, \ldots, g_k$ is free on these k elements.*

Now, the set of $(g_1, \ldots, g_k) \in G^{(k)}$ for which $\langle g_1, \ldots, g_k \rangle$ is not free is clearly dense in $G^{(k)}$, so it follows easily that the set of $(g_1, \ldots, g_k) \in G^{(k)}$ which are *not* Diophantine is of the second (Baire) category in $G^{(k)}$. That is to say the topologically generic $(g_1, \ldots, g_k)$ is free but not Diophantine. On the other hand, it was proved in [30] that the elements with algebraic number entries are Diophantine and for generic in measure elements Kaloshin and Rodnianski established the following result:

Theorem 12 ([46]). *For almost every pair* $(A, B) \in \mathrm{SO}(3) \times \mathrm{SO}(3)$ *there is a constant* $D > 0$ *such that for any* n *and any word* $W_n(A, B)$ *of length* n *in* A *and* B *we have the following weak Diophantine property*

$$\| W_n(A, B) \pm \mathrm{Id} \| \geq D^{-n^2}$$

We expect that generic in the measure sense $(g_1, \ldots, g_k)$ is Diophantine; this question remains open.

Acknowledgement. It is a pleasure to thank Professors Kaimanovich and Woess for organizing an extremely stimulating workshop. The author was supported in part by the NSF Postdoctoral Fellowship.

References

[1] N. Alon, Eigenvalues and expanders, Combinatorica 6 (1986), 83–96.

[2] N. Alon, A. Lubotzky and A. Wigderson, Semi-direct product in groups and zig-zag product in graphs: connections and applications (extended abstract), in: 42nd IEEE Symposium on Foundations of Computer Science (Las Vegas, NV, 2001), IEEE Computer Soc., Los Alamitos, CA, 2001, 630–637.

[3] N. Alon and V. Milman, λ_1, isoperimetric inequalities for graphs, and superconcentrators, J. Combin. Theory Ser. B 38 (1985), 73–88.

[4] N. Alon and Y. Roichman, Random Cayley graphs and expanders, Random Structures Algorithms 5 (1994), 271–284.

[5] A. Arnold and A. Avez, Ergodic Problems of Classical Mechanics, Benjamin, New York 1968.

[6] L. Babai, W. M. Kantor and A. Lubotzky, Small-diameter Cayley graphs for finite simple groups, Europ. J. Combin. 10 (1989), 507–522.

[7] L. Babai, G. Hetyei, W. M. Kantor, A. Lubotzky, and A. Seress, On the diameter of finite groups, in: Proc. 31st IEEE Symp. on Foundations of Computer Science, St. Louis, MO, 857–865.

[8] R. Balian and C. Itzykson, Observations sur la mécanique quantique finie, C. R. Acad. Sci. Paris Sér. I Math. 303 (1986), 773–778.

[9] S. Banach, Sur le probleme de mesure, Fund. Math. 4 (1923), 7–33.

[10] A. F. Beardon, The Geometry of Discrete Groups, Springer-Verlag, New York 1983.

[11] M. V. Berry and M. Tabor, Level clustering in the regular spectrum, Proc. Roy. Soc. London Ser. A 356 (1977), 375–394.

[12] S. De Bièvre, M. Degli Esposti and R. Giachetti, Quantization of a class of piecewise affine transformations on the torus, Comm. Math. Phys. 176 (1996), 73–94.

[13] E. B. Bogomolny, B. Georgeot, M.-J. Giannoni and C. Schmit, Chaotic billiards generated by arithmetic groups, Phys. Rev. Lett. 69 (1992), 1477–1480.

[14] O. Bohigas, M. Giannoni and C. Schmit, Characterization of chaotic quantum spectra and universality of level fluctuation laws, Phys. Rev. Lett. 52 (1984), 1–4.

[15] J. Bolte, G. Steil and F. Steiner, Arithmetical chaos and violation of universality in energy level statistics, Phys. Rev. Lett. 69 (1992), 2188–2191.

[16] A. Broder and E. Shamir, On the second eigenvalue of random regular graphs, 28th Annual Symp. on Found. of Comp. Sci., 1987, 286-294.

[17] F. R. K. Chung, Spectral Graph Theory, Amer. Math. Soc., Providence, RI, 1994.

[18] L. Clozel, Automorphic forms and the distribution of points on odd-dimensional spheres, Israel J. Math. 132 (2002), 175–187.

[19] G. Davidoff, P. Sarnak and A. Valette, Elementary Number Theory, Group Theory, and Ramanujan Graphs, Cambridge University Press, Cambridge 2003.

[20] M. Degli Esposti, Quantization of the orientation preserving automorphism of the torus, Ann. Inst. H. Poincaré Phys. Théor. 58 (1993), 323–341.

[21] M. Degli Esposti, S. Graffi and S. Isola, Classical limit of the quantized hyperbolic toral automorphisms, Comm. Math. Phys. 167 (1995), 471–507.

[22] P. Diaconis, Random walks on groups: characters and geometry (to appear).

[23] V. Drinfeld, Finitely-additive measures on S^2 and S^3, invariant with respect to rotations, Funct. Anal. Appl. 18 (1984), 245–246.

[24] D. B. A. Epstein, Almost all subgroups of a Lie group are free, J. Algebra 19 (1971), 261–262.

[25] A. Eskin, S. Mozes and H. Oh, Uniform exponential growth for linear groups, Internat. Math. Res. Notices 31 (2002), 1675–1683.

[26] J. M. G. Fell, Weak containment and induced representations of groups, Canad. J. Math. 14 (1962), 237–268.

[27] J. Friedman, On the second eigenvalue and random walks in random d-regular graphs, Combinatorica 11 (1991), 331–362.

[28] Z. Füredi and J. Komlós, The eigenvalues of random symmetric matrices, Combinatorica 1 (1981), 233–241.

[29] A. Gamburd, On the spectral gap for infinite index "congruence" subgroups of $\mathrm{SL}_2(\mathbb{Z})$, Israel J. Math. 127 (2002), 157–200.

[30] A. Gamburd, D. Jakobson and P. Sarnak, Spectra of elements in the group ring of SU(2), J. Eur. Math. Soc. (JEMS) 1 (1999), 51–85.

[31] A. Gamburd, J. Lafferty and Rockmore, Eigenvalue spacings for quantized cat maps, J. Phys. A 36 (2003), 3487-3499.

[32] A. Gamburd and I. Pak, Expansion of product-replacement graphs, Combinatorica (to appear).

[33] A. Gamburd and M. Shahshahani, Uniform diameter bounds for some families of Cayley graphs, preprint, 2003.

[34] T. Gelander and A. Żuk, Dependence of Kazhdan constants on generating subsets, Israel J. Math., 129 (2002), 93-98.

[35] R. I. Grigorchuk and P. de la Harpe, On problems related to growth, entropy and spectrum in group theory, J. Dynam. Control Systems 3 (1997), 51–89.

[36] M. Gromov, Spaces and questions, in: GAFA 2000 (Tel Aviv, 1999), Geom. Funct. Anal. 2000, Special Volume, Part I, 118–161.

[37] M. Gromov, Random walk on random groups, preprint, IHES (2002).

[38] U. Haagerup, Random matrices, free probability and the invariant subspace problem relative to a von Neumann algebra, Proceedings ICM 2002 (to appear).

[39] U. Haagerup and S. Thorbjornsen, A new application of random matrices: $\mathrm{Ext}(C_r^*(F_2))$ is not a group, Centre for Mathematical Physics and Stochastics, Research Report no. 45 (2002).

[40] J. Hannay and M. Berry, Quantization of linear maps on a torus – Fresnel diffraction by a periodic grating, Phys. D. 1 (1980), 267–290.

[41] G. Hardy and E. Wright, An Introduction to the Theory of Numbers, fifth ed., The Clarendon Press, Oxford University Press, New York 1979.

[42] F. Hausdorff, Grundzüge der Mengenlehre, von Veit, Leipzig 1914.

[43] D. A. Hejhal, The Selberg Trace Formula for $PSL_2(\mathbb{R})$, Vol. I, Lecture Notes in Math. 548, Springer-Verlag, Berlin 1976; Vol. 2, Lecture Notes in Math. 1001, Springer-Verlag, Berlin 1983.

[44] D. Jacobson, S. Miller, I. Rivin and Z. Rudnick, Eigenvalue spacings for regular graphs, in: Emerging Applications of Number Theory (Minneapolis, MN, 1996), IMA Vol. Math. Appl. 109, Springer-Verlag, New York 1999, 317–327.

[45] V. A. Kaimanovich and A. M. Vershik, Random walks on discete groups: boundary and entropy, Ann. Probab. 11 (1983), 457–490.

[46] V. Kaloshin and I. Rodnianski, Diophantine properties of elements of $SO(3)$, Geom. Funct. Anal. 11 (2001), 953–970.

[47] N. M. Katz and P. Sarnak, Random Matrices, Frobenius Eigenvalues, and Monodromy, American Mathematical Society Colloquium Publications, 45, Amer. Math. Soc., Providence, RI, 1999.

[48] D. A. Kazhdan, Connection of the dual space of a group with the structure of its closed subgroups, Funct. Anal. Appl. 1 (1967), 63–65.

[49] J. P. Keating, The cat map: quantum mechanics and classical motion, Nonlinearity 4 (1991), 309–341.

[50] J. P. Keating and F. Mezzadri, Pseudo-symmetries of Anosov maps and spectral statistics, Nonlinearity 13 (2000), 747–775.

[51] J. P. Keating, F. Mezzadri and J. M. Robbins, Quantum boundary conditions for torus maps, Nonlinearity 12 (1999), 579–591.

[52] S. Keppeler, J. Marklof, F. Mezzadri Quantum Cat Maps with spin $1/2$, Nonlinearity 14 (2001), 719–738.

[53] H. Kesten, Symmetric random walks on groups, Trans. Amer. Math. Soc. 92 (1959), 336–354.

[54] H. D. Kloosterman, The behavior of general theta functions under the modular group and the characters of binary modular congruence groups. I, Ann. of Math. (2) 47 (1946), 317–375.

[55] S. Knabe On the quantization of Arnold's cat, J. Phys. A 23 (1990), 2013–2025.

[56] M. Koubi, Croissance uniforme dans les groupes hyperboliques, Ann. Inst. Fourier (Grenoble) 48 (1998), 1441–1453.

[57] P. Kurlberg and Z. Rudnick, Hecke theory and equidistribution for the quantization of linear maps of the torus, Duke Math. J. 103 (2000), 47–77.

[58] J. D. Lafferty and D. Rockmore, Fast Fourier analysis for SL_2 over a finite field and related numerical experiments, Experiment. Math., 1 (1992), 115–139.

[59] J. D. Lafferty and D. Rockmore, Numerical investigation of the spectrum for certain families of Cayley graphs, in: Expanding Graphs (Princeton, NJ, 1992), DIMACS Ser. Discrete Math. Theoret. Comput. Sci., 10, Amer. Math. Soc., Providence, RI, 1993, 63–73.

[60] J. D. Lafferty and D. Rockmore, Level spacings for Cayley graphs, in: Emerging Applications of Number Theory (Minneapolis, MN, 1996), IMA Vol. Math. Appl., 109, Springer-Verlag, New York 1999, 373–386.

[61] M. Larsen, Navigating the Cayley graph of $\mathrm{SL}_2(\mathbb{F}_p)$, Internat. Math. Res. Notices 27 (2003), 1465–1471.

[62] P. D. Lax and R. S. Phillips, The asymptotic distribution of lattice points in Euclidean and non-Euclidean spaces, J. Funct. Anal. 46 (1982), 280–350.

[63] M. Liebeck and A. Shalev, Diameters of finite simple groups: sharp bounds and applications, Ann. of Math. 154 (2001), 383–406.

[64] A. Lubotzky, Discrete Groups, Expanding Graphs and Invariant Measures, Progr. Math. 125, Birkhäuser, Basel 1994.

[65] A. Lubotzky, Cayley graphs: eigenvalues, expanders and random walks, in: Surveys in Combinatorics, 1995 (Stirling), London Math. Soc. Lecture Note Ser. 218, Cambridge University Press, Cambridge 1995, 155–189.

[66] A. Lubotzky, R. Phillips and P. Sarnak, Hecke operators and distributing points on S^2. I, Comm. Pure Appl. Math. 39 (S) (1986), S149–S186; II, Comm. Pure Appl. Math. 40 (1987), 401–420.

[67] A. Lubotzky, R. Phillips and P. Sarnak, Ramanujan Graphs, Combinatorica 8 (1988), 261–277.

[68] A. Lubotzky and B. Weiss, Groups and expanders, in: Expanding Graphs (Princeton, NJ, 1992), DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 10, Amer. Math. Soc., Providence, RI, 1993, 95–109.

[69] A. Lubotzky and A. Żuk, On property (τ), preprint (2002).

[70] W. Luo, Z. Rudnick and P. Sarnak, On Selberg's eigenvalue conjecture, Ceom. Funct. Anal. 5 (1995), 387–401.

[71] W. Luo and P. Sarnak, Number variance for arithmetic hyperbolic surfaces, Comm. Math. Phys. 161 (1994), 419–432.

[72] G. A. Margulis, Explicit constructions of concentrators, Problems Inform. Transmission 10 (1975), 325–332.

[73] G. A. Margulis, Some remarks on invariant means, Monatsh. Math. (90) 1980, 233–235.

[74] G. A. Margulis, Explicit construction of graphs without short cycles and low density codes, Combinatorica 2 (1982), 71–78.

[75] G. A. Margulis, Explicit group-theoretic constructions of combinatorial schemes and their applications in the construction of expanders and concentrators, Problems Inform. Transmission 24 (1988), 39–46.

[76] B. McKay, The expected eigenvalue distribution of a large regular graph, Linear Algebra Appl. 40 (1981), 203–216.

[77] M. Mehta, Random Matrices, 2nd ed., Academic Press, Boston 1991.

[78] M. A. Nielsen and I. L. Chuang, Quantum computation and quantum information, Cambridge University Press, Cambridge 2000.

[79] T. Novikoff, Asymptotic behavior of the random 3-regular bipartite graph, preprint.

[80] D. V. Osin, The entropy of solvable groups, Ergodic Theory Dynam. Systems 23 (2003), 907–918.

[81] D. V. Osin, Kazhdan constants of hyperbolic groups, Funktsional. Anal. i Prilozhen. 36 (4) (2002), 46–54; English translation in Funct. Anal. Appl. 36 (2002), 290–297:

[82] S. J. Patterson, The limit set of a Fuchsian group, Acta. Math. 136 (1975), 241–273.

[83] R. S. Phillips and P. Sarnak, On the spectrum of the Hecke groups, Duke Math. J. 52 (1985), 211–221.

[84] M. Pinsker, On the complexity of concentrator, in: 7th Annual Teletrafic Conference, Stockholm 1973, 318/1–318/4.

[85] O. Reingold, S. Vadhan and A. Wigderson, Entropy waves, the zig-zag graph product, and new constant-degree expanders, Ann. of Math. (2) 155 (2002), 157–187.

[86] L. Saloff-Coste, Probability on groups: random walks and invariant diffusions, Notices Amer. Math. Soc. 48 (2001), 968–977.

[87] P. Sarnak, Statistical properties of eigenvalues of the Hecke operators, in: Analytic Number Theory and Diophantine Problems (Stillwater, OK, 1984), Progr. Math. 70, Birkhäuser, Boston, MA, 1987, 321–331.

[88] P. Sarnak, Some Applications of Modular Forms, Cambridge Tracts in Math. 99, Cambridge University Press, Cambridge 1990.

[89] P. Sarnak, Arithmetic quantum chaos, in: The Schur Lectures (1992) (Tel Aviv), Israel Math. Conf. Proc., 8, Bar-Ilan Univ., Ramat Gan 1995, 183–236.

[90] P. Sarnak, Spectra and eigenfunctions of Laplacians, in: Partial Differential Equations and their Applications (Toronto, ON, 1995), CRM Proc. Lecture Notes 12, Amer. Math. Soc., Providence, RI, 1997, 261–276.

[91] P. Sarnak and X. X. Xue, Bounds for multiplicities of automorphic representations, Duke Math. J. 64 (1991), 207–227.

[92] K. Schmidt, Amenability, Kazhdan's property T, strong ergodicity and invariant means for ergodic group-actions, Ergodic Theory Dynam. Systems 1 (1981), 223–236.

[93] C. Schmit, Quantum and classical properties of some billiards on the hyperbolic plane, in: Chaos et Physique Quantique (Les Houches, 1989), 331–370, North-Holland, Amsterdam 1991, 331–370.

[94] A. Selberg, On the estimation of Fourier coefficients of modular forms, Proc. Symp. Pure Math. VII, Amer. Math. Soc., Providence RI 1965, 1–15.

[95] J-P. Serre, Répartition asymptotique des valeurs propres de l'opérateur de Hecke T_p, J. Amer. Math. Soc. 10 (1997), 75–102.

[96] Y. Shalom, Expanding graphs and invariant means, Combinatorica 17 (1997), 555–575.

[97] Y. Shalom, Expander graphs and amenable quotients, in: Emerging Applications of Number Theory (Minneapolis, MN, 1996), IMA Vol. Math. Appl. 109, Springer-Verlag, New York 1999, 571–581.

[98] Ya. G. Sinai and A. B. Soshnikov, A refinement of Wigner's semicircle law in a neighborhood of the spectrum edge for random symmetric matrices, Funct. Anal. Appl. 32 (1998), 114–131.

[99] D. Sullivan, For $n > 3$ there is only one finitely additive rotationally invariant measure on the n-sphere defined on all Lebesgue measurable subsets, Bull. Amer. Math. Soc. (N.S.) 4 (1981), 121–123.

[100] D.Sullivan, Discrete conformal groups and measurable dynamics, Bull. Amer. Math. Soc. (N.S.) 6 (1982), 57–73.

[101] A. Tarski, Algebraische Fassung des Mass Problems, Fund. Math. 31 (1938), 47–66.

[102] A. Terras, Fourier Analysis on Finite Groups and Applications, London Math. Soc. Stud. Texts 43, Cambridge University Press, Cambridge 1999.

[103] C. Tracy and H. Widom, Distribution Functions for Largest Eigenvalues and Their Applications, in: Proceedings of the International Congress of Mathematicians, Beijing 2002, Vol. I, ed. LI Tatsien, Higher Education Press, Beijing 2002, 587–596.

[104] A. M. Vershik and S. V. Kerov, Asymptotic of the largest and typical dimensions of irreducible representations of a symmetric group, Funct. Anal. Appl. 19 (1985), 21–31.

[105] D. Voiculescu, Around quasidiagonal operators, Integral Equations Operator Theory 17 (1993), 137–149.

[106] E. Wigner, Random matrices in physics, SIAM Review 9 (1967), 1–123.

[107] W. Woess, Random Walks on Infinite Graphs and Groups, Cambridge Tracts in Math. 138, Cambridge University Press, Cambridge 2000.

[108] S. Zelditch, Index and dynamics of quantized contact transformations, Ann. Inst. Fourier (Grenoble) 47 (1997), 305–363.

Alex Gamburd, Department of Mathematics, Stanford University, Stanford, CA 94305
E-mail: agamburd@math.stanford.edu

The Ihara zeta function of infinite graphs, the KNS spectral measure and integrable maps

Rostislav I. Grigorchuk and Andrzej Żuk

Abstract. We define the Ihara zeta function for Cayley graphs of infinite finitely generated groups. We extend the definition of the Ihara zeta function to infinite graphs which are limits of sequences $\{X_n\}_{n=1}^{\infty}$ of finite k-regular graphs such that X_{n+1} covers X_n. We associate to such a graph a measure μ with support in $[-1, 1]$ called the Kesten–von Neumann–Serre spectral measure. We present a few examples of computation of the zeta function and the measure μ for Schreier graphs of some fractal groups generated by finite automata. These computations are closely related to the integrability of some 2-dimensional mappings which are also in focus of our considerations.

Contents

1 Introduction 142

2 The Markov operator and the spectral measure 143

3 Convergence of graphs and coverings 144

4 The definition of the Ihara zeta function for infinite graphs 145

5 The integral presentation and the problem of moments 148

6 The Ihara zeta function for groups 149

7 Schreier graphs and subgroup separability 151

8 The spectral problem and integrability 152

9 The examples with absolutely continuous KNS spectral measure 154

10 The examples with totally discontinuous spectrum 161

11 Zeta function related to the lamplighter group 167

12 An example of iterated monodromy group 172

13 Related topics and concluding remarks 175

1 Introduction

Recently new results about spectral properties of infinite regular graphs were obtained. Namely in [1] an example of a graph with a spectrum which is a Cantor set was constructed, while in [17] an example of an infinite Cayley graph with a pure point spectrum of the Laplace operator was treated. The last result was used to answer a question of Atiyah about L^2 Betti numbers [12]. Some other results about spectra of Schreier graphs and amenability were obtained in [18, 19].

The idea of this paper is to use the techniques and results from the cited papers to introduce the Ihara zeta function using the residual finiteness of groups and graphs related to automata.

We also use the von Neumann trace to define the Ihara zeta function for Cayley graphs of infinite finitely generated groups.

The Ihara zeta function $\zeta_X(t)$ of a finite graph X is one of the analogues of the Riemann's zeta function

$$\zeta(s) = \sum_{n\geq 1} n^{-s} = \prod_{p \text{ prime}} \left(1 - p^{-s}\right)^{-1} \quad \text{for } \operatorname{Re} s > 1.$$

It was introduced in [23] (and studied in [3, 25, 26, 34, 35, 36, 39, 40, 41] and many other papers) by the relation

$$\zeta_X(t) = \prod_{[C]} \left(1 - t^{|C|}\right)^{-1},$$

where the product is over all equivalence classes $[C]$ of primitive closed paths C in X, and $|C|$ denotes the length of C.

The Ihara zeta function of a finite graph plays an important role in number theory, graph theory and mathematical physics [26, 40, 41, 43]. It is useful in the study of spectral properties of graphs and leads to the definition of Ramanujan graphs as graphs for which the analogue of the Riemann hypothesis about the zeros of zeta functions holds [23, 27, 37].

We show that the Ihara zeta function can be easily defined for some infinite graphs, first of all for Cayley graphs and for Schreier graphs of finitely generated groups. In Section 6 using von Neumann trace we define the Ihara zeta function for Cayley graphs of infinite finitely generated groups. Then, using a result of Serre [38], we extend the definition of the Ihara zeta function to infinite graphs which are limits of sequences $\{X_n\}_{n=1}^{\infty}$ of finite graphs such that X_{n+1} covers X_n. This definition works, for instance, for graphs which are Schreier graphs of a pair (G, H) where H is a subgroup of a group G which is an intersection of subgroups of finite index in G.

Also, we associate to such a graph a measure μ with support in $[-1, 1]$ which plays a role analogue to the Kesten spectral measure defined in [24] for Cayley graphs, and is called by us the Kesten–von Neumann–Serre (KNS) spectral measure. This measure is used to get an integral presentation for the zeta function.

Finally we present a few examples of computation of the zeta function for Schreier graphs of some fractal groups considered in [1, 17, 18, 19].

The possibility of such a computation is based on a trick of introducing extra parameters in the spectral problem and reducing it to the study of multidimensional rational mappings of the real Euclidean space $\mathbb{R}^d$ for $d \geq 2$. In the situation when this map is integrable, i.e., there are functions $\psi : \mathbb{R}^d \to \mathbb{R}$ and $\alpha : \mathbb{R} \to \mathbb{R}$ such that $\psi F = \alpha\psi$, the spectral problem usually can be completely treated, and this is the case in most of our examples. The examples of integrable mappings of $\mathbb{R}^2$ obtained in [1, 17, 19] come from finite automata and look mysterious. We hope that new examples of integrable mappings associated to finite automata and groups generated by them will be discovered, and this will give a new impulse both to the spectral theory of graphs and groups and to the theory of dynamical systems.

2 The Markov operator and the spectral measure

The graphs $X = (V_X, E_X)$ which we consider are connected, locally finite, and their edges are not oriented. The graphs are allowed to have loops and multiple edges (such graphs are often called multigraphs). For a vertex v we denote by $\deg(v)$ the degree of v, i.e., the number of edges containing v (a loop counts twice). For more on such a definition of graphs see [4]. Let $l^2(V_X, \deg)$ denote the space of real functions on the set of vertices which are square summable with the weight deg.

On such a graph X one can consider a random walk operator M acting on functions $f \in l^2(V_X, \deg)$ as follows:

$$Mf(v) = \frac{1}{\deg(v)} \sum_{(w,v)} f(w),$$

where the sum is taken over all edges (w, v) containing the vertex v. If there are k loops at the vertex v, $f(v)$ will appear $2k$ times in the sum, and if between the distinct vertices v and w there are n edges, $f(w)$ will appear n times in the sum.

This is a self-adjoint operator on $l^2(X, \deg)$ with the spectrum $\mathrm{Sp}(M) \subset [-1, 1]$.

The motivation for such a definition becomes clear if we consider the Hecke type operator

$$\mathcal{H} = \frac{1}{2|S|} \sum_{s \in S} \left(\pi(s) + \pi(s^{-1})\right)$$

acting on $l^2(G/H)$, where $H < G$ is a subgroup, S is a set of generators, and π is a quasi-regular representation $\lambda_{G/H}$. The realization of $\mathcal{H}$ as a Markov operator for a simple random walk on the Schreier graph $\mathcal{S}(G, H, S)$ (about Schreier graphs see Section 7) leads to the above definition of M.

As the operator M is bounded ($||M|| \leq 1$) and self-adjoint, it has the spectral decomposition

$$M = \int_{-1}^{1} \lambda dE(\lambda),$$

where E is the spectral measure. This spectral measure is defined on Borel subsets of the interval $[-1, 1]$ and takes values in the set of projections on the Hilbert space $l^2(X, \deg)$. The matrix μ^X of measures μ^X_{xy}, $x, y \in V_X$ can be associated with E as follows:

$$\mu^X_{xy}(B) = \langle E(B)\delta_x, \delta_y \rangle,$$

where B is a Borel subset of $[-1, 1]$, and δ_x is the function which equals 1 at x and 0 elsewhere.

In general, $\lambda \in \mathrm{Sp}(M)$ if and only if for every $\varepsilon > 0$ there exists μ^X_{xy} such that $|\mu^X_{xy}((\lambda - \varepsilon, \lambda + \varepsilon))| > 0$. But in fact (see [24]) $\lambda \in \mathrm{Sp}(M)$ if and only if for every $\varepsilon > 0$ there exists $x \in X$ such that $\mu^X_{xx}((\lambda - \varepsilon, \lambda + \varepsilon)) > 0$.

Indeed, let us show that if, for $B = (\lambda - \varepsilon, \lambda + \varepsilon)$, we have $|\mu^X_{xy}(B)| > 0$ then $\mu^X_{xx}(B) > 0$. As $E(B)$ is a projection, one has

$$\begin{aligned} 0 < (\mu^X_{xy}(B))^2 &= \langle E(B)\delta_x, \delta_y \rangle^2 \leq \langle E(B)\delta_x, E(B)\delta_x \rangle \langle \delta_y, \delta_y \rangle \\ &= \langle E(B)\delta_x, \delta_x \rangle \deg(y) = \mu^X_{xx}(B) \deg(y), \end{aligned}$$

as needed.

We call the measures $\mu^X_x = \mu^X_{xx}$ the Kesten spectral measures.

If the graph X is vertex transitive the measures μ^X_x are independent of x. In particular this is the case when X is a Cayley graph of a group. In the latter case the measure μ can be alternatively defined via von Neumann's approach as $\mu(B) = \mathrm{tr}\, E(B)$, where tr is the trace in the von Neumann algebra of the group generated by the left regular representation [29].

In Section 4 we will define an analogue of the Kesten–von Neumann spectral measure for not vertex transitive graphs, which would be the average of the Kesten spectral measures over the set of vertices in the situation when the graph is the limit of a covering sequence of finite graphs in the sense of the next section (more details in Section 4).

3 Convergence of graphs and coverings

In this section we recall few well known facts about convergence of graphs. For details see [16].

The distance between two vertices $v_a \neq v_b \in V_X$ is the minimal number of edges needed to connect them, i.e.,

$$\begin{aligned}\operatorname{dist}(v_a, v_b) = &\min\{n : \exists\, v_0, \ldots, v_n \in V_X, v_0 = v_a, v_n = v_b, \\ &(v_i, v_{i+1}) \in E_X \text{ for } 0 \leq i \leq n-1\}.\end{aligned}$$

Let us consider a family $\{(X_n, v_n)\}$ of marked graphs, i.e., graphs with chosen vertices $v_n \in V_{X_n}$.

On the space of marked graphs there is a metric Dist defined as follows

$$\operatorname{Dist}((X_1, v_1), (X_2, v_2)) = \inf\left\{\frac{1}{n+1} : \ B_{X_1}(v_1, n) \text{ is isometric to } B_{X_2}(v_2, n)\right\},$$

where $B_X(v, n)$ is the ball of radius n in X centered on v.

For a sequence of marked graphs (X_n, v_n) we say that (X, v) is the limit graph if

$$\lim_{n\to\infty} \operatorname{Dist}((X, v), (X_n, v_n)) = 0.$$

The limit graph is unique up to an isometry.

For any sequence $\{X_n, v_n\}$ of finite marked graphs of uniformly bounded degree there exists an infinite marked graph (X, v) which is the limit of a subsequence of $\{(X_n, v_n)\}$ (see [16, Theorem 3]).

The graphs we consider are naturally equipped with a structure of 1-dimensional CW complexes, and we can use a general theory of coverings (see for instance [32]).

Definition 3.1. The sequence of finite graphs $\{X_n\}$ is a covering sequence if X_{n+1} covers X_n.

Any covering sequence has a limit $\lim X_n$ in the sense of the above definition ([16]).

In Section 7 we will consider special cases of covering sequences of algebraic origin.

4 The definition of the Ihara zeta function for infinite graphs

The Ihara zeta function $\zeta_X(t)$ for a finite regular graph X satisfies the relation

$$\zeta_X(t) = \exp\Big(\sum_{r=1}^{\infty} c_r t^r / r\Big), \tag{4.1}$$

where c_r is the number of closed, non-oriented loops of length r in the graph X [23].

Indeed

$$\ln \zeta_X(t) = -\sum_{[C]} \ln(1 - t^{|C|}) = \sum_{[C]} \sum_{j \geq 1} \frac{1}{j} t^{|C| j} = \sum_{j \geq 1} \sum_{d \geq 1} \sum_{[C], |C| = d} \frac{1}{j} t^{dj}$$
$$= \sum_{j \geq 1} \sum_{d \geq 1} \sum_{C, |C| = d} \frac{1}{dj} t^{dj} = \sum_{r \geq 1} \frac{c_r}{r} t^r,$$

because there are d representatives in $[C]$ if the length of C is d, and

$$\sum_{d|r} \sum_{C, |C| = d} = c_r$$

as C is a primitive loop. Therefore,

$$\ln \zeta_X(t) = \sum_{r=1}^{\infty} \frac{c_r}{r} t^r.$$

In [23] it was also shown that

Theorem 4.1 ([23]). *The Ihara zeta function $\zeta_X(t)$ for a finite regular graph X of degree k satisfies:*

$$\zeta_X(t) = \left(1 - t^2\right)^{-\frac{1}{2}(k-2)|X|} \det\left(1 - tkM + (k-1)t^2\right)^{-1}, \tag{4.2}$$

where M is the Markov operator on X.

We are going to use these two basic facts and the following result of Serre to define the zeta function for infinite graphs, which are limits of sequences of finite k-regular graphs.

Following Serre [38], we say that the eigenvalues $\lambda_i(X_n)$ of the Markov operators M_n on the sequence of finite graphs X_n are equidistributed with respect to a measure μ which has support in $[-1, 1]$ if the sequence of measures μ_n

$$\mu_n = \sum \delta_{\lambda_i(X_n)} / |X_n|$$

which we call counting measures, and where δ_x is the Dirac measure at x, converges weakly to the measure μ.

Theorem 4.2 ([38]). *Let X_n be a family of finite graphs of degree k. The following conditions are equivalent:*

1) *The eigenvalues of $M(X_n)$ are equidistributed with respect to some measure μ.*

2) *The formal power series*

$$\zeta_{X_n}(t)^{1/|X_n|}$$

has a limit in $\mathbb{R}[[t]]$.

3) *For every $r \geq 1$, the sequence*

$$c_r(X_n)/|X_n|, \tag{4.3}$$

where c_r is the number of closed loops of length r, has a limit denoted by $\widetilde{c_r}$.

As a corollary of Theorem 4.2 we obtain

Corollary 4.3. *Let $\{X_n\ , v_n\}$ be a covering sequence of finite k-regular graphs and M_n the sequence of the corresponding Markov operators. Then the eigenvalues of M_n are equidistributed with respect to some measure μ.*

Proof. The ratio $c_r(X_n)/|X_n|$ decreases with n. Thus the limit $\widetilde{c_r}$ of $c_r(X_n)/|X_n|$ exists, and therefore all statements of Theorem 4.2 hold, in particular the eigenvalues of M_n are equidistributed with respect to some measure μ. □

Definition 4.4. The above limit measure μ will be called the Kesten–von Neumann–Serre (KNS) spectral measure of X and $\{X_n\}$, where X is the limit of the sequence $\{X_n, v_n\}$.

The measure μ is an average of Kesten spectral measures $\mu_{v,v}$ over the set of vertices v of the graph, using the approximating sequence $\{X_n\}$.

Let $\{X_n\}$ be a sequence of graphs converging to X. Using Theorem 4.1 we have

$$\frac{1}{|X_n|}\ln\zeta_{X_n}(t) = -\frac{k-2}{2}\ln\left(1-t^2\right) - \frac{1}{|X_n|}\ln\det\left(1 - tkM_n + (k-1)t^2\right) \tag{4.4}$$

$$= \sum_{r=1}^{\infty} \frac{c_r(X_n)}{|X_n|} t^r/r. \tag{4.5}$$

Passing to the limit we come to the following

Definition 4.5. Let $X = \lim_{n\to\infty} X_n$ where X_n is a sequence of k-regular marked graphs such that the limit of (4.3) exists when $n \to \infty$. The zeta function $\zeta_X(t)$ of the graph X, with respect to the sequence $\{X_n\}$, is defined by

$$\ln\zeta_X(t) = \lim_{n\to\infty} \frac{1}{|X_n|}\ln\zeta_{X_n}(t) = \sum_{r=1}^{\infty} \widetilde{c_r} t^r/r. \tag{4.6}$$

The series on the right-hand side of (4.6) has a nontrivial interval of convergence around the point 0. Namely, an easy estimate gives $\widetilde{c_r} \leq k^r$, from which it follows that the radius of convergence is at least $\frac{1}{k}$. Thus function $\zeta_X(t)$ is well-defined and analytic in a neighborhood of 0. In Section 5 we give a better estimate on the radius of convergence.

The definition of $\zeta_X(t)$ depends on the approximating sequence X_n, but we will usually omit writing $\{X_n\}$.

Remark 4.6. It follows from formula (4.6) that for any tree X which admits a finite quotient (for instance, a homogenous tree) the zeta function $\zeta_X(t)$ is equal to 1. Indeed,

for such X one can construct a sequence of finite graphs X_n such that $X = \lim_{n\to\infty} X_n$ and such that the girths of X_n tend to infinity. In this case $\widetilde{c_r} = 0$.

5 The integral presentation and the problem of moments

In this section we show that the moments of the KNS spectral measure are determined by the zeta function. We prove

Theorem 5.1. *Let $X = \lim X_n$ under assumption that the limit* (4.3) *exists, and let μ be as in Definition* 4.4. *The radius of convergence of the series* (4.6) *is at least $\frac{1}{k-1}$, and for any t such that $|t| < \frac{1}{k-1}$ the following relation holds*

$$\ln \zeta_X(t) = -\frac{k-2}{2}\ln\left(1-t^2\right) - \int_{-1}^{1} \ln\left(1 - tk\lambda + (k-1)t^2\right) d\mu(\lambda). \tag{5.1}$$

The KNS spectral measure μ is uniquely determined by the zeta function ζ_X.

Proof. Passing to the limit ($n \to \infty$) in (4.4) and (4.5) and using the notation from Theorem 4.2 we get

$$\begin{aligned}
\ln \zeta_X(t) &= -\frac{k-2}{2}\ln\left(1-t^2\right) - \lim_{n\to\infty} \frac{1}{|X_n|} \ln\det\left(-ktM_n + 1 + (k-1)t^2\right) \\
&= -\frac{k-2}{2}\ln\left(1-t^2\right) - \lim_{n\to\infty} \int_{-1}^{1} \ln\left(1 - tk\lambda + (k-1)t^2\right) d\mu_n(\lambda) \\
&= -\frac{k-2}{2}\ln\left(1-t^2\right) - \int_{-1}^{1} \ln\left(1 - tk\lambda + (k-1)t^2\right) d\mu(\lambda)
\end{aligned}$$

as $1 - tk\lambda + (k-1)t^2 > 0$ on $[-1, 1]$, for $|t| < \frac{1}{k-1}$.

We will show that the function $\zeta_X(t)$ given by the formula (5.1) determines all moments of the measure μ.

We have

$$1 - tk\lambda + (k-1)t^2 = (1 - a(\lambda)t)(1 - b(\lambda)t),$$

where

$$\begin{aligned}
a(\lambda) &= \frac{k\lambda - \sqrt{k^2\lambda^2 - 4k + 4}}{2}, \\
b(\lambda) &= \frac{k\lambda + \sqrt{k^2\lambda^2 - 4k + 4}}{2}.
\end{aligned}$$

Thus for $|t| < \frac{1}{k-1}$ we get

$$\begin{aligned}\ln \zeta_X(t) + \frac{k-2}{2} \ln\left(1-t^2\right) &= -\int_{-1}^{1} \ln\left(1 - tk\lambda + (k-1)t^2\right) d\mu(\lambda) \\ &= -\int_{-1}^{1} \ln(1-a(\lambda)t)(1-b(\lambda)t)\, d\mu(\lambda) \\ &= -\int_{-1}^{1} (\ln(1-a(\lambda)t) + \ln(1-b(\lambda)t))\, d\mu(\lambda) \\ &= -\int_{-1}^{1} \sum_{n=1}^{\infty} \frac{(a^n(\lambda)+b^n(\lambda))t^n}{n}\, d\mu(\lambda) \\ &= -\sum_{n=1}^{\infty} \frac{t^n}{n} \int_{-1}^{1} (a^n(\lambda)+b^n(\lambda))\, d\mu(\lambda) \qquad (5.2)\end{aligned}$$

as the above integrals are non-singular for $|t| < \frac{1}{k-1}$.

From the expression of $a(\lambda)$, $b(\lambda)$ we get for $|\lambda| \le 1$

$$\max\{|a(\lambda)|, |b(\lambda)|\} \le \frac{k + \sqrt{k^2-4k+4}}{2} = k-1.$$

Thus from (5.2) it follows that the radius of convergence of $\ln \zeta_X(t)$ is at least $\frac{1}{k-1}$.

We remark that

$$a^n(\lambda) + b^n(\lambda)$$

is a polynomial in λ of degree n as can be easily checked.

For $\ln \zeta_X(t) + \frac{k-2}{2} \ln(1-t^2) = \sum_{n=1}^{\infty} d_n t^n$ we have

$$\int_{-1}^{1} (a^n(\lambda)+b^n(\lambda))\, d\mu(\lambda) = -n d_n,$$

which shows that for any n, the n-th moment of μ is determined by $\ln \zeta_X(t)$. One can also express the n-th moment using Chebyshev polynomials. □

6 The Ihara zeta function for groups

Let G be a finitely generated group with a finite set of generators S, and let $\mathcal{S} = \mathcal{S}(G, S)$ be the corresponding Cayley graph. We are going to define the Ihara zeta function for $\mathcal{S}$ with no restrictions on G.

Let us use the relation

$$\frac{1}{|X|} \ln \zeta_X(t) = -\frac{k-2}{2} \ln\left(1-t^2\right) - \frac{1}{|X|} \ln\det\left(1 - tkM + (k-1)t^2\right),$$

which holds for finite k-regular graphs X, and which can be rewritten as

$$\frac{1}{|X|}\ln\zeta_X(t) = -\frac{k-2}{2}\ln(1-t^2) - \operatorname{tr}\ln\left(1 - tkM + (k-1)t^2\right),$$

where tr is a normalized trace. Recall that the von Neumann algebra $\mathcal{N}(G)$ generated by the left regular representation in $l^2(G)$ is of finite type, i.e., there is a canonical trace tr defined by the relation

$$\operatorname{tr} a = \langle a\delta_1, \delta_1\rangle,$$

where $\delta_1 \in l^2(G)$ is the delta function at the group identity.

The last relation can be used to define

$$\ln\zeta_G(t) = -(m-1)\ln\left(1-t^2\right) - \operatorname{tr}\ln\left(1 - 2tmM + (2m-1)t^2\right), \tag{6.1}$$

where m is the number of generators (thus $k = 2m$), and $M \in \mathcal{N}(G)$ is the Markov operator on $\mathcal{S}$.

Let G be a residually finite group, with a finite set of generators S, $|S| = m$, and let $H_n \lhd G$, $n = 1, \ldots$, be a descending sequence of normal subgroups with the trivial intersection. Let $\mathcal{S}$ and $\mathcal{S}_n$ by the Cayley graphs of the groups G and $G_n = G/H_n$ with respect to the system of generators S and its images S_n in G_n, respectively. Then $\mathcal{S}_n$, $n = 1, \ldots$, constitute the covering system of finite graphs, and $\mathcal{S} = \lim_n \mathcal{S}_n$.

Proposition 6.1. *The Ihara zeta function of $\mathcal{S}$ defined by the relation* (6.1) *coincides with the Ihara zeta function of $\mathcal{S}$ with respect to the approximating sequence given by Definition* 4.5.

Moreover, the function given by (6.1) *coincides with the function*

$$-(m-1)\ln\left(1-t^2\right) - \int_{-1}^{1}\ln\left(1 - 2tm\lambda + (2m-1)t^2\right)d\mu(\lambda),$$

where μ is the Kesten spectral measure associated to the random walk on $\mathcal{S}$.

Corollary 6.2. *Definition* 4.5 *of the Ihara function on $\mathcal{S}$ does not depend on the choice of the above sequence H_n of normal subgroups.*

Proof. It is known that $\mu = \lim_n \mu_n$, where μ_n are uniform measures on S_n [16]. Thus the integral presentation leads to the relation for the Ihara zeta function for $\mathcal{S}$ and $\mathcal{S}_n$. In particular we obtain independence from $\{H_n\}$.

By definition of the spectral measure μ,

$$\operatorname{tr}\ln\left(1 - 2tmM + (2m-1)t^2\right) = \int_{-1}^{1}\ln\left(1 - 2tm\lambda + (2m-1)t^2\right)d\mu(\lambda). \quad \square$$

For $\mathbb{Z}$, and more generally for $\mathbb{Z}^d$, the spectral measure with respect to the standard generators is well known [33]. This leads to the following formula for the corresponding Ihara zeta function for $\mathbb{Z}^d$:

$$\ln \zeta_{\mathbb{Z}^d}(t) = -d \ln\left(1-t^2\right) - \int_{-1}^{1} \ln\left(1-2dt\lambda+(2d-1)t^2\right) d\mu^{*d}(\lambda),$$

where μ^{*d} is the d-fold convolution of the measure $d\mu(\lambda) = \frac{1}{\pi\sqrt{1-\lambda^2}} d\lambda$.

In particular, for $\mathbb{Z}$ we get

$$\ln \zeta_{\mathbb{Z}}(t) = -\int_{-1}^{1} \ln(1-2t\lambda+t^2) \frac{1}{\pi\sqrt{1-\lambda^2}} d\lambda = 0.$$

The above result can also be obtained from formula (4.6). Indeed, the remark at the end of Section 4 shows that zeta functions of trees which admit finite quotients (for instance, homogenous trees) are equal to 1.

The spectral measure with respect to the standard generators was also computed for the free group F_m of rank m (see [24]). This leads to the following formula for the corresponding Ihara zeta function

$$\ln \zeta_{F_m}(t) = -m \ln\left(1-t^2\right) - \int_{-\frac{\sqrt{2m-1}}{m}}^{\frac{\sqrt{2m-1}}{m}} \ln\left(1-2tm\lambda+(2m-1)t^2\right) \frac{\sqrt{2m-1-m^2\lambda^2}}{\pi(1-\lambda^2)} d\lambda.$$

7 Schreier graphs and subgroup separability

The important examples of graphs we consider are the following. Let G be a finitely generated group and $H < G$ be a subgroup of infinite index. H is called separable if there exists a sequence H_n of finite index subgroups of G such that $\bigcap H_n = H$. The group G is called subgroup separable if all finitely generated subgroups of G are separable. The examples of such groups include free groups [21], polycyclic groups, some fundamental groups of 3-manifolds, and the group from [10] as was shown recently in [15].

The classical construction due to Schreier associates to the triple (G, H, S), where S is a system of generators of G, a graph $\mathcal{S}(G, H, S)$ often denoted as $\mathcal{S}(G/H)$ as follows. The set of vertices for $\mathcal{S}$ is the set G/H of left cosets, and two cosets gH, hH are joined by an edge, labelled by a generator $s \in S$, if $sgH = hH$. The Cayley graph of a group is a particular case of the Schreier graph, corresponding to the case $H = 1$. We consider $\mathcal{S}(G/H)$ as a marked graph with H as a distinguished vertex (forgetting about the labelling).

Proposition 7.1. *Let H be a separable subgroup of G. Then the Schreier graphs $X_n = \mathcal{S}(G/H_n, S)$ converge to the Schreier graph $X = \mathcal{S}(G/H, S)$, and the counting measures μ_n converge to the measure μ.*

Proof. Consider more general situation when the sequence of marked graphs (X_n, v_n) converges to the marked graph (X, v). Then the measures $\mu_{v_n v_n}^{X_n}$ converge weakly to the measure μ_{vv}^{X} (see [16]). Indeed, in this case the moments of the measures $\mu_{v_n v_n}^{X_n}$ converge to the moments of the measure μ_{vv}^{X}. As it is easy to see, in our situation this implies weak convergence of corresponding measures. Indeed, the l-th moment of the measure μ_{yy}^{Y} for a graph Y and $y \in Y$ is given by

$$(\mu_{yy}^{Y})^{(l)} = \int_{-1}^{1} \lambda^l \mu_{yy}^{Y}(\lambda) = \int_{-1}^{1} \lambda^l \langle E(\lambda)\delta_y, \delta_y \rangle = \langle M^l \delta_y, \delta_y \rangle.$$

Thus the l-th moment of the measure μ_{yy}^{Y} is equal to the probability of going from y to y in l steps. But for n sufficiently large, the balls $B_{X_n}(v_n, l)$ and $B_X(v, l)$ are isometric, and the l-th moment of the measure $\mu_{v_n v_n}^{X_n}$ is the same as the l-th moment of the measure μ_{vv}^{X}. □

Thus if H is a separable subgroup of G, and $H = \bigcap H_n$, then to this data and the set of generators S we can associate a graph $\mathcal{S}(G/H, S)$ with a KNS spectral measure.

In fact, up to a cover of degree 2 any regular graph can be realized as a Schreier graph [27].

In the latter sections we shall consider examples of triples (G, H, S) which come from finite automata.

8 The spectral problem and integrability

In the following sections we will describe several examples of computation of the Ihara function of Cayley graphs and Schreier graphs related to some fractal groups. These examples come from [1, 17, 19] and correspond to groups generated by finite automata. In order to make our computations more clear for the reader we would like to recall a few facts on which the computation is based.

We omit the definition of the group $G = G(A)$, generated by a finite non-initial automaton A and basic facts about such groups. These can be found in [13, 17, 18, 19]. Let us just remind that G naturally acts on a d-regular rooted tree T by automorphisms, where d is the cardinality of the alphabet, and there is a natural embedding $G \to G \wr_D S_d$ in the permutational wreath product, where the symmetric group S_d acts on the alphabet D of the automaton A, and $d = |D|$.

To any point $\xi \in \partial T$ one associates a parabolic subgroup $P = \mathrm{St}_G(\xi)$ and a Schreier graph $\mathcal{S} = \mathcal{S}(G, P, S)$, where S is a system of generators of G. If $P = \{1\}$ then $\mathcal{S}$ is the Cayley graph of G. The graph $\mathcal{S}$ is the limit of finite Schreier graphs $\mathcal{S}_n = \mathcal{S}(G, P_n, S)$, where $P_n = \mathrm{St}_n(u_n)$, and u_n is the vertex on the path ξ on the level n. The spectrum of the Markov operator M corresponding to $\mathcal{S}$ is approximated by

spectra of Markov operators M_n corresponding to the graphs $\mathcal{S}_n$. The KNS spectral measure μ on $\mathcal{S}$ is the limit of counting measures μ_n of the spectrum of M_n.

Let $S = \{a_1, \ldots, a_m\}$, and

$$M = \frac{1}{2m} \sum_{i=1}^{m} \big(\rho(a_i) + \rho(a_i^{-1})\big)$$

be the Markov operator in $l^2(G/P)$ corresponding to the simple random walk on G/P, where $\rho = \lambda_{G/P}$ is a quasi-regular representation. This operator is approximated by the operators

$$M_n = \frac{1}{2m} \sum_{i=1}^{m} \big(\rho_n(a_i) + \rho_n(a_i^{-1})\big),$$

where $\rho_n = \lambda_{G/P_n}$.

The idea used in [1, 17] for the computation of the spectrum is to consider the multi-parametric family of operators

$$M(\eta_1, \ldots, \eta_m, \lambda) = \frac{1}{2m} \sum_{i=1}^{m} \eta_i \big(\rho(a_i) + \rho(a_i^{-1})\big) - \lambda I,$$

$$M_n(\eta_1, \ldots, \eta_m, \lambda) = \frac{1}{2m} \sum_{i=1}^{m} \eta_i \big(\rho_n(a_i) + \rho_n(a_i^{-1})\big) - \lambda I_n,$$

where $\eta_1, \ldots, \eta_m, \lambda$ are parameters taking values in $\mathbb{R}$ (the number of parameters can be reduced by 1), and I, I_n are the unit operators in the spaces $l^2(G/P)$, $l^2(G/P_n)$. It may happen, and this is the case in our examples, that for the determinant

$$Q_n(\eta_1, \ldots, \eta_m, \lambda) = \det(M_n(\eta_1, \ldots, \eta_m, \lambda))$$

the following recursion holds

$$Q_{n+1}(\eta_1, \ldots, \eta_m, \lambda) = f(\eta_1, \ldots, \eta_m, \lambda) g^{d^n}(\eta_1, \ldots, \eta_m, \lambda) Q_n(F(\eta_1, \ldots, \eta_m, \lambda)),$$

where $f, g : \mathbb{R}^{m+1} \to \mathbb{R}$ are some functions, and F is a rational map $\mathbb{R}^{m+1} \to \mathbb{R}^{m+1}$. Thus

$$Q_{n+1}(\overline{\eta}, \lambda)$$

depends on

$$Q_1(F^{n-1}(\overline{\eta}, \lambda)),$$

where $\overline{\eta} = (\eta_1, \ldots, \eta_m)$. Therefore the problem of the determination of the spectrum

$$\mathrm{Sp}(M_n(\eta_1, \ldots, \eta_m, \lambda)) = \{(\eta_1, \ldots, \eta_m, \lambda) : Q_n(\eta_1, \ldots, \eta_m, \lambda) = 0\}$$

(the spectrum of M_n is $\mathrm{Sp}(M_n(1,\dots,1,\lambda))$), is closely related to the properties of the dynamical system generated by the mapping F. The situation is much simpler in the case when the dynamical system is integrable, i.e., satisfies the following

Definition 8.1. We call a rational map $F : \mathbb{R}^{m+1} \to \mathbb{R}^{m+1}$ integrable if there exist a rational map $\psi : \mathbb{R}^{m+1} \to \mathbb{R}$ and a polynomial $\alpha : \mathbb{R} \to \mathbb{R}$ such that

$$\alpha\psi = \psi F. \tag{8.1}$$

In this case F is semi-conjugated to a one-dimensional map α. In such a situation Q_n can be factorized into factors (the nature of which depends on ψ), and the problem of the computation of $\zeta_X(t)$ can be solved in two ways.

First, by computation of the KNS spectral measure and use of formula (5.1), and second by use of the factorization of Q_n (see the examples below).

The idea of integrability based on the relation (8.1) is not new in dynamical systems and goes back perhaps to Poincaré (we heard this from E. Ghys). In [42] a stronger version of the definition, based on the relation of commutation

$$\alpha\beta = \beta\alpha$$

is discussed. For the case when F is a meromorphic function and ψ and α are rational mappings of $\mathbb{C}$, the equation (8.1) is completely treated in [6]. It would be nice to develop theory of integrable multidimensional rational mappings of $\mathbb{R}^d$, in particular for dimension 2.

9 The examples with absolutely continuous KNS spectral measure

In this section we will consider two examples of groups generated by finite automata whose Schreier graphs have absolutely continuous KNS spectral measure.

Note that often in our examples below we use non-normalized Markov operators in computations, statements and figures.

The groups $G = \langle a, b, c, d\rangle$ and $\widetilde{G} = \langle a, \widetilde{b}, \widetilde{c}, \widetilde{d}\rangle$ are generated by automata given in Figures 1 and 2. They have intermediate growth between polynomial and exponential. The group G embeds in $\widetilde{G}$ via the map $a \to a$, $b \to \widetilde{b}\widetilde{d}$, $c \to \widetilde{c}\widetilde{b}$, $d \to \widetilde{d}\widetilde{c}$.

Let $P = \mathrm{St}_G(1^\infty)$ and $\widetilde{P} = \mathrm{St}_{\widetilde{G}}(1^\infty)$ be parabolic subgroups, where 1^∞ corresponds to the infinite path $1, 1, 1, \dots$. Let $\mathcal{S} = \mathcal{S}(G, P, S)$ and $\widetilde{\mathcal{S}} = \mathcal{S}(\widetilde{G}, \widetilde{P}, \widetilde{S})$ denote the Schreier graphs, where $S = \{a, b, c, d\}$ and $\widetilde{S} = \{a, \widetilde{b}, \widetilde{c}, \widetilde{d}\}$. The Schreier graphs $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are described in [13].

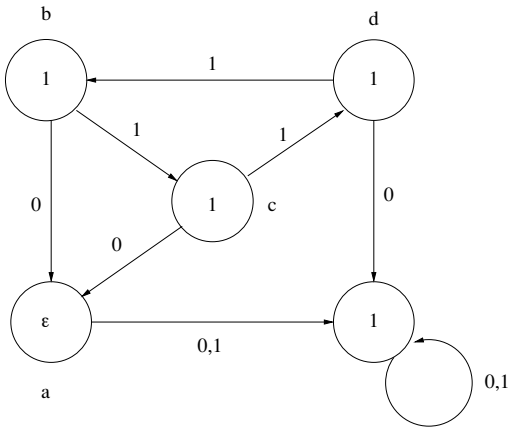


Figure 1. The automaton generating the group G

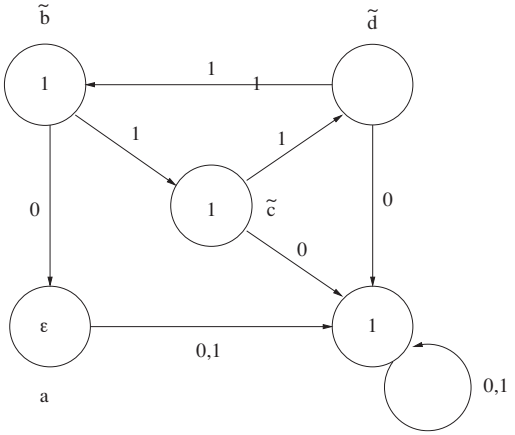


Figure 2. The automaton generating the group $\widetilde{G}$

Theorem 9.1 ([1]). *The spectrum of the graph $\mathcal{X}$ coincides with the set*

$$\left[-\frac{1}{2}, 0\right] \cup \left[\frac{1}{2}, 1\right],$$

and the KNS spectral measure is equal to

$$\frac{|1-4x|}{2\pi\sqrt{x(2x-1)(2x+1)(1-x)}}\,dx.$$

The spectrum of the graph $\widetilde{\mathcal{X}}$ coincides with the set

$$[0, 1],$$

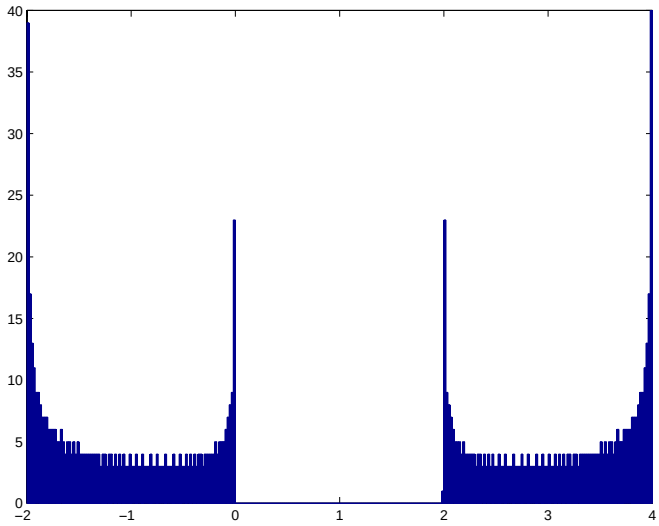


Figure 3. The histogram of the spectrum in the case of the group G

and the KNS spectral measure is equal to

$$\frac{1}{\pi\sqrt{x-x^2}}dx.$$

Corollary 9.2. *The Ihara zeta functions for the Schreier graphs $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ of the groups G and $\widetilde{G}$ are equal to*

$$\ln\zeta_{\mathcal{S}}(t) = -3\ln\left(1-t^2\right) - \int_{-\frac{1}{2}}^{0} \frac{\ln(1-t8x+7t^2)|1-4x|}{2\pi\sqrt{x(2x-1)(2x+1)(1-x)}}\,dx$$
$$-\int_{\frac{1}{2}}^{1} \frac{\ln(1-t8x+7t^2)|1-4x|}{2\pi\sqrt{x(2x-1)(2x+1)(1-x)}}\,dx,$$

and

$$\ln\zeta_{\widetilde{\mathcal{S}}}(t) = -3\ln(1-t^2) - \int_0^1 \frac{\ln(1-8xt+7t^2)}{\pi\sqrt{x^2-x}}\,dx.$$

Now we are going to recall some steps of the computation of the measure μ for these examples, using the integrability of the corresponding map F. We will also get an alternative integral representation for $\ln\zeta(t)$.

Let

$$M_n = \rho_n(a) + \rho_n(b) + \rho_n(c) + \rho_n(d)$$

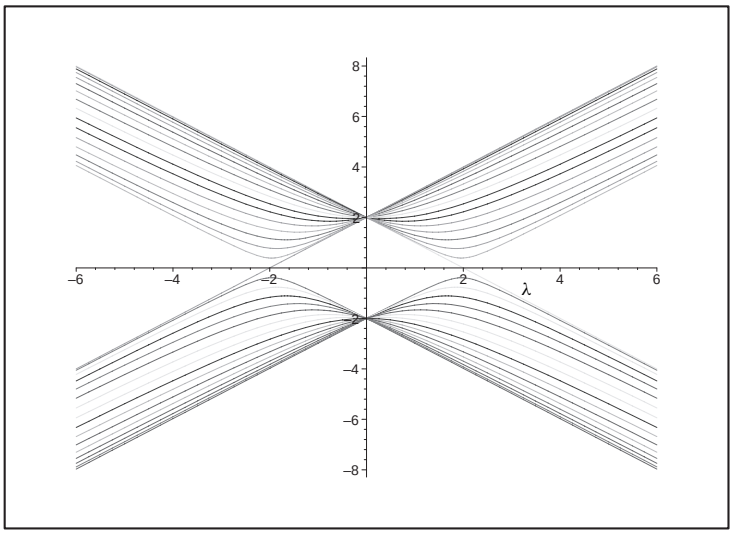


Figure 4. The zeros of Q_6 associated to the group G

be an $2^n \times 2^n$ matrix representing the normalized Markov operator on the Schreier graph $\mathcal{S} = (G, P_n, S)$ ($P_n = \mathrm{St}_G(1^n)$). As $a^2 = b^2 = c^2 = d^2 = 1$, we do not need inverses. Let

$$M_n(\lambda, \eta) = M_n - (\lambda + 1)\rho(a) - (\eta + 1)I_n,$$

$$Q_n(\lambda, \eta) = \det M_n(\lambda, \eta).$$

For $n \geq 2$, as was shown in [1], we have

$$Q_n(\lambda, \eta) = (4 - \eta^2)^{2^{n-2}} Q_{n-1}(F(\lambda, \eta)), \tag{9.1}$$

where

$$F(\lambda, \eta) = \left(\frac{2\lambda^2}{4 - \eta^2}, \eta + \frac{\eta\lambda^2}{4 - \eta^2} \right).$$

The map F admits the following inverse

$$F^{-1}(\lambda, \eta) = \left(\pm \sqrt{\frac{\lambda}{2}\left(4 - \left(\frac{2\eta}{2 + \lambda}\right)^2\right)}, \frac{2\eta}{2 + \lambda} \right).$$

The set of F-preimages of a point drawn with the help of a computer looks like the set given in Figure 5. It is natural to conjecture that the map F has this set as an attractor, but this is only our guess.

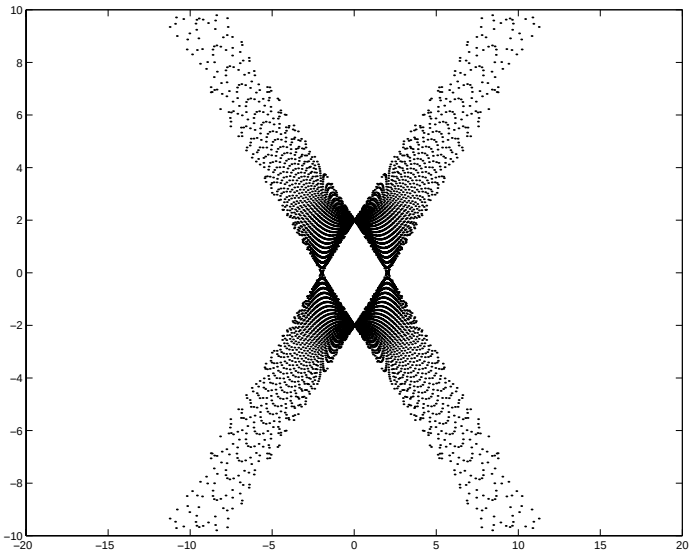


Figure 5. The attractor associated to the group G

Proposition 9.3. *The map F is integrable with*

$$\psi(\lambda, \eta) = \frac{4 - \eta^2 + \lambda^2}{4\lambda},$$
$$\alpha(x) = 2x^2 - 1.$$

Proof. Indeed, the direct computation gives

$$\alpha\psi = \frac{1}{8}(-4 + \eta^2 - \lambda^2)^2/\lambda^2 - 1 = \psi F. \qquad \square$$

The mapping $x \to 2x^2 - 1$ is known as the von Neumann–Ulam map [31]. It is conjugated to the map $x \to x^2 - 2$. Outside the interval the dynamics of this map is simple: any point x, $|x| > 1$ goes to infinity. On the other hand, in the interval $[-1, 1]$ the dynamics is chaotic. Indeed, the map α is conjugated to the expanding map of S^1, namely $z \to z^2$ and to the one-sided Bernoulli shift on two symbols. The measure $\frac{dx}{\pi\sqrt{1-x^2}}$ is invariant with respect to α and is the measure of the maximal entropy, because it corresponds to the $\left(\frac{1}{2}, \frac{1}{2}\right)$ Bernoulli shift.

In [1] the following factorization for $Q_n(\lambda, \eta)$ was proven:

$$Q_n(\lambda, \eta) = (2 - \eta - \lambda)(2 - \eta + \lambda) \prod_{\substack{\phi = \frac{2\pi j}{2^n} \\ j = 1, \ldots, 2^{n-1} - 1}} H_{\cos(\phi)}(\lambda, \eta),$$

where

$$H_\theta(\lambda, \eta) = 4 - \eta^2 + \lambda^2 + 4\lambda\theta.$$

It can be alternatively written as

$$Q_n(\lambda, \eta) = (2 - \eta - \lambda)(2 - \eta + \lambda) \prod_{\theta \in \bigcup_{i=0}^{n-2} \alpha^{-i}(0)} H_\theta(\lambda, \eta).$$

Indeed, from the integrability of F it follows that

$$H_\theta(F(\lambda, \eta)) = H_{\theta_1}(\lambda, \eta) H_{\theta_2}(\lambda, \eta),$$

where θ_1, θ_2 are the preimages of θ under α, and in order to get a factorization of Q_n one can use relation (9.1).

The map F maps the hyperbola $H_\theta(\lambda, \eta) = 0$ onto the hyperbola $H_{\alpha(\theta)}(\lambda, \eta) = 0$. The spectrum of the 2-parametric family $M(\lambda, \eta)$ is a union of hyperbolas.

Proposition 9.4. *The following relation holds for the graph associated to the group G:*

$$\ln \zeta(t) = -3 \ln(1 - t^2) - \frac{1}{2} \int_{-1}^{1} \ln(1 - 4t - 2t^2 - 28t^3 + 49t^4 + 16t^2 x)\, d\nu(x),$$

where

$$d\nu(x) = \frac{dx}{\pi \sqrt{1 - x^2}}.$$

Proof. Using the factorization of Q_n we get

$$\begin{aligned}
\zeta_n(t) &= \left(1 - t^2\right)^{-3 \cdot 2^n} \det\left(1 - t2M_n + 7t^2\right)^{-1} \\
&= \left(1 - t^2\right)^{-3 \cdot 2^n} (2t)^{-2^n} \det\left(M_n + \frac{1 + 7t^2}{-2t}\right)^{-1} \\
&= \left(1 - t^2\right)^{-3 \cdot 2^n} (2t)^{-2^n} Q_n^{-1}\left(-1, \frac{1 + 7t^2}{2t} - 1\right) \\
&= \left(1 - t^2\right)^{-3 \cdot 2^n} (2t)^{-2^n} \left(8t - 1 - 7t^2\right)^{-1} \left(4t - 1 - 7t^2\right)^{-1} \times
\end{aligned}$$

$$\times \prod_{\theta \in \bigcup_{i=0}^{n-2} \alpha^{-i}(0)} \left(-1, \frac{1+7t^2}{2t} - 1\right)^{-1}$$
$$= \left(1-t^2\right)^{-3\cdot 2^n} \left(8t - 1 - 7t^2\right)^{-1} \left(4t - 1 - 7t^2\right)^{-1}$$
$$\times \prod_{\theta \in \bigcup_{i=0}^{n-2} \alpha^{-i}(0)} \left(5(2t)^2 - (1+7t^2-2t)^2 - 4(2t)^2\theta\right)^{-1}.$$

Thus, we have

$$\lim_{n\to\infty} \frac{1}{2^n} \ln \zeta_n(t) = -3 \ln\left(1-t^2\right)$$
$$- \lim_{n\to\infty} \frac{1}{2^n} \sum_{\theta \in \bigcup_{i=0}^{n-2} \alpha^{-i}(0)} \ln(1 - 4t - 2t^2 - 28t^3 + 49t^4 + 16t^2\theta)$$
$$= -3 \ln\left(1-t^2\right)$$
$$- \frac{1}{2} \lim_{n\to\infty} \int_{-1}^{1} \ln(1 - 4t - 2t^2 - 28t^3 + 49t^4 + 16t^2 x)\, d\nu_n(x),$$

where ν_n is the discrete measure uniformly distributed on the set $\bigcup_{i=0}^{n-2} \alpha^{-i}(0)$ of $2^{n-1} - 1$ points. The measures ν_n converge to the measure of the maximal entropy of α which we call ν. This measure corresponds to the $\left(\frac{1}{2}, \frac{1}{2}\right)$ Bernoulli measure in the realization of α as the Bernoulli shift and is the projection on $[-1, 1]$ via the map $x \to \cos x$ of the Lebesgue measure on S^1. This ends the proof. □

For the group $\widetilde{G}$ we use the same notations but with tilde. We have the relation

$$\frac{1}{2} Q_n(\lambda, \eta) = \widetilde{Q}_n(\lambda, \eta).$$

In this case

$$F(\lambda, \eta) = \left(\frac{\lambda^2}{1-\eta^2}, \eta + \frac{\mu\lambda^2}{1-\eta^2}\right),$$

and the following proposition holds:

Proposition 9.5. *For*

$$\psi(\lambda, \eta) = \frac{2(1 - \eta^2 + \lambda^2)}{\lambda}$$

and

$$\alpha(x) = 2x^2 - 1$$

we have (8.1).

Proof. This follows from Proposition 9.3. Indeed, the relation $\frac{1}{2} Q_n = \widetilde{Q}_n$ implies that the maps F for G and $\widetilde{G}$ are conjugated by the map $(\lambda, \eta) \to (2\lambda, 2\eta)$. □

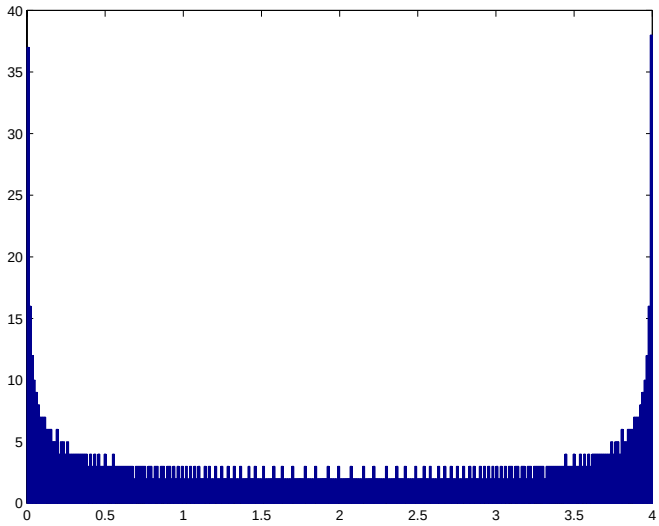


Figure 6. The histogram of the spectrum of the group $\widetilde{G}$

In the same way as for G one can also compute the zeta function for the group $\widetilde{G}$ using the factorization of Q_n.

10 The examples with totally discontinuous spectrum

In this and in the following sections we will consider the examples of fractal groups for which the associated Schreier graphs have discrete spectral measure. These are the Fabrikowski–Gupta group from Figure 7, the Gupta–Sidki 3-group from Figure 8 and the group generated by the automaton from Figure 9 studied in [1]. Although the number of different groups is 3, the number of different spectra, KNS measures and Ihara zeta functions is only 2 because the last two cases lead to the same Schreier graphs (if we forget the labelling of the edges).

Let $P = \mathrm{St}_\Gamma(2^\infty)$, $\overline{P} = \mathrm{St}_{\overline{\Gamma}}(2^\infty)$ and $\overline{\overline{P}} = \mathrm{St}_{\overline{\overline{\Gamma}}}(2^\infty)$ be parabolic subgroups, where 2^∞ corresponds to the infinite path $2, 2, 2, \ldots$. Let $\mathcal{S}$, $\overline{\mathcal{S}}$, $\overline{\overline{\mathcal{S}}}$ be the corresponding Schreier graphs.

All groups Γ, $\overline{\Gamma}$ and $\overline{\overline{\Gamma}}$ are fractal, the groups Γ and $\overline{\overline{\Gamma}}$ are branch, the group $\overline{\Gamma}$ is weakly branch, the groups Γ and $\overline{\Gamma}$ are virtually torsion free, and the group $\overline{\Gamma}$ is a torsion 3-group.

Let us first consider the group $\Gamma = \langle a, s\rangle$. The group Γ embeds into permutational wreath product $\Gamma \wr_D S_3$ via the map $a \to (1, 1, 1)\varepsilon$, $s \to (a, 1, s)1$, where $\varepsilon \in S_3$ is

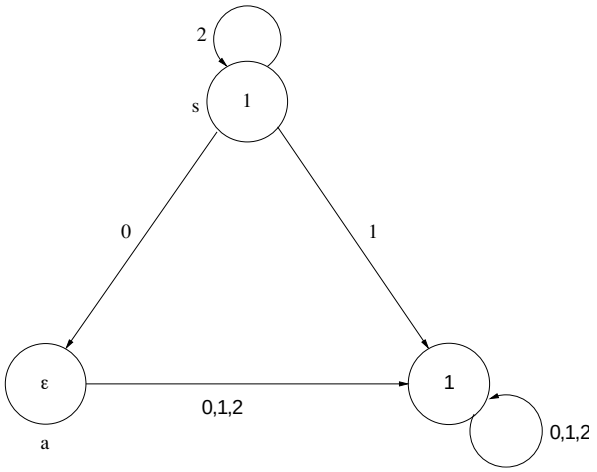


Figure 7. The automaton generating the group Γ

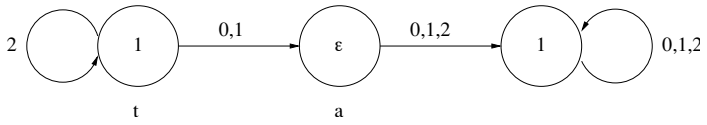


Figure 8. The automaton generating the group $\overline{\Gamma}$

the cyclic permutation. Let

$$M = \rho(a) + \rho(a^{-1}) + \rho(s) + \rho(s^{-1})$$

be the Markov operator on $l^2(G/P)$, and

$$M_n = \rho_n(a) + \rho_n(a^{-1}) + \rho_n(s) + \rho_n(s^{-1})$$

be the approximation on the n-th level.

Theorem 10.1 ([1]). *The spectrum of the graph $\mathcal{S}$ is the closure of the set of points*

$$1 \pm \underbrace{\sqrt{6 \pm \sqrt{6 \pm \cdots \pm \sqrt{6}}}}_{n \text{ times}}$$

for $n = 0, 1, \ldots$.

The above set is a union of a Cantor set of null Lebesgue measure and a countable collection of isolated points supporting the KNS spectral measure μ which is discrete and which has value $\frac{1}{3^{n+1}}$ at the point whose definition involves n radicals.

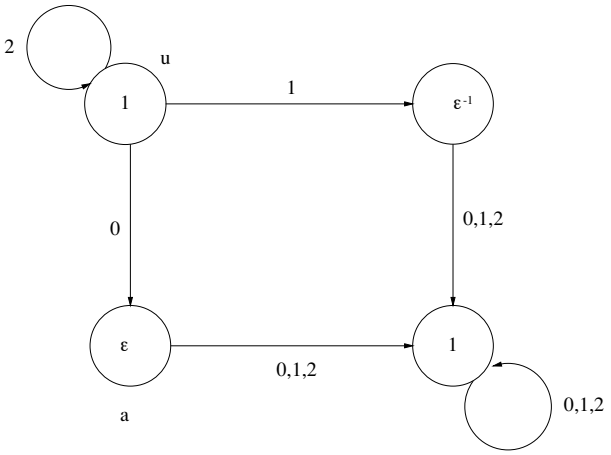


Figure 9. The automaton generating the group $\overline{\overline{\Gamma}}$

From Theorem 5.1 we obtain the following

Corollary 10.2. *The Ihara zeta functions for the above graph is equal to*

$$\ln \zeta_{\mathcal{S}}(t) = -\ln\left(1-t^2\right) - \frac{1}{3}\ln\left(1-t-3t^2\right) - \sum_{j=1}^{\infty} \ln\left(1 - t\left(1 \pm \underbrace{\sqrt{6 \pm \sqrt{6 \pm \cdots \pm \sqrt{6}}}}_{j \text{ times}}\right) + 3t^2\right) \frac{1}{3^{j+1}},$$

where all possible combinations of the sign $\pm$ *appear in the above sum.*

Now we are going to recall some steps of the computation of the spectrum and the KNS spectral measure μ in order to obtain an alternative description of $\zeta_{\mathcal{S}}(t)$.

Let

$$Q_n(\lambda,\eta) = \det\left(\lambda(\rho_n(a)+\rho_n(a^{-1})) + \rho_n(b) + \rho_n(b^{-1}) - \eta I_n\right).$$

For $\alpha = 2-\eta+\lambda$, $\beta = 2-\eta-\lambda$, $\gamma = 1+\eta+\lambda$, $\delta = 1+\eta-\lambda$ we have for $n \geq 2$

$$Q_n(\lambda,\eta) = (\alpha\beta\gamma^2)^{3^{n-2}} Q_{n-1}(F(\lambda,\eta)),$$

where

$$F(\lambda,\eta) = \left(\frac{\lambda^2(2-\eta-\lambda)}{(2-\eta+\lambda)(\eta^2-\lambda^2-\eta-2)}, \eta + 2\frac{\lambda^2(\eta^2-\lambda^2-\eta-2)}{(2-\eta+\lambda)(\eta^2-\lambda^2-2\eta-\lambda)}\right).$$

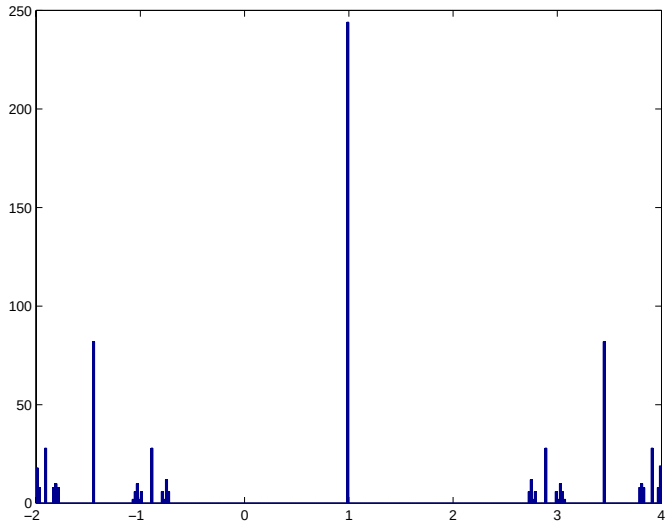


Figure 10. The histogram of the spectrum in case of the group Γ

Proposition 10.3. *The map F is integrable with*

$$\psi(\lambda,\eta) = -\frac{\eta^2 - \lambda\eta - 2\lambda^2 - 2 - \eta}{\lambda},$$
$$\alpha(x) = 4 - 2x - x^2.$$

Proof. Indeed, the direct computation shows

$$\alpha\psi = -(4\lambda^4 + 4\lambda^3 + 4\eta\lambda^3 - 3\eta^2\lambda^2 + 4\lambda^2 + 6\eta\lambda^2 - 2\lambda\eta^3 + 4\lambda + 6\lambda\eta + \eta^4 - 2\eta^3 - 3\eta^2 + 4 + 4\eta)/\lambda^2 = \psi F. \qquad \square$$

Let

$$H_\theta(\lambda,\eta) = \eta^2 - \lambda\eta - 2\lambda^2 - 2 - \eta + \theta\lambda.$$

The above proposition is related to the fact that

$$H_\theta(F(\lambda,\eta)) = \frac{\beta}{\alpha\gamma} H_{\theta_1}(\lambda,\eta) H_{\theta_2}(\lambda,\eta)$$

where θ_1, θ_2 are the preimages of θ under α.

This leads to the following factorization

$$Q_n(\lambda,\eta) = (2 + 2\lambda - \eta)(2 - \lambda - \eta)^{3^{n-1}+1} \prod_{\substack{2\le m\le n \\ \theta\in\alpha^{-m}(-1)}} H_\theta^{3^{n-m}+1}(\lambda,\eta).$$

In the same way as it was done for the group G one can prove

Proposition 10.4. *The following relation holds for the graph associated to the group* Γ

$$\begin{aligned}\ln \zeta(t) = &- \ln\left(1-t^2\right) - \frac{1}{3}\ln\left(1-t+3t^2\right) \\ &- \sum_{j=0}^{\infty}\left(\ln\left(1-2t+2t^2-6t^3+9t^4\right.\right. \\ &\qquad \left.\left. + \left(1 \pm \underbrace{\sqrt{6 \pm \sqrt{6 \pm \cdots \pm \sqrt{6}}}}_{j \text{ times}}\right) t^2\right) \frac{1}{3^{j+2}}\right)\end{aligned}$$

for small $|t|$*, where all possible combinations of the sign* $\pm$ *appear in the above sum.*

Proof. Using the factorization for Q_n we have

$$\begin{aligned}\zeta_n(t) &= \left(1-t^2\right)^{-3^n} \det(1 - tM_n + 3t^2)^{-1} \\ &= \left(1-t^2\right)^{-3^n} t^{-3^n} Q_n^{-1}\left(1, \frac{1+3t^2}{t}\right) \\ &= \left(1-t^2\right)^{-3^n} t^{-3^n}\left(4 - \frac{1+3t^2}{t}\right)^{-1}\left(1 - \frac{1+3t^2}{t}\right)^{-3^{n-1}-1} \\ &\quad \times \prod_{\substack{2 \le m \le n \\ \theta \in \alpha^{-m}(-1)}} H_\theta^{-3^{n-m}-1}\left(1, \frac{1+3t^2}{t}\right).\end{aligned}$$

Thus, because of the definition of the measure μ from Theorem 10.1, we get

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{3^n}\ln \zeta_n(t) = &- \ln\left(1-t^2\right) - \frac{1}{3}\ln\left(1-t+3t^2\right) \\ &- \frac{1}{3}\sum_{j=1}^{\infty}\ln\left(1-2t+2t^2-6t^3+9t^4\right. \\ &\qquad \left. + \left(1 \pm \underbrace{\sqrt{6 \pm \sqrt{6 \pm \cdots \pm \sqrt{6}}}}_{j \text{ times}}\right) t^2\right) \frac{1}{3^{j+1}}.\end{aligned}$$

□

Let us check that the zeta functions from Corollary 10.2 and Proposition 10.4 coincide. This follows from the following:

$$\ln\left(1-t\left(1+\underbrace{\sqrt{6\pm\sqrt{6\pm\cdots\pm\sqrt{6}}}}_{j\text{ times}}\right)+3t^2\right)\frac{1}{3^{j+1}}$$
$$+\ln\left(1-t\left(1-\underbrace{\sqrt{6\pm\sqrt{6\pm\cdots\pm\sqrt{6}}}}_{j\text{ times}}\right)+3t^2\right)\frac{1}{3^{j+1}}$$
$$=\frac{1}{3^{j+1}}\ln\left(1-2t+2t^2-6t^3+9t^4+\left(1\pm\underbrace{\sqrt{6\pm\sqrt{6\pm\cdots\pm\sqrt{6}}}}_{j-1\text{ times}}\right)t^2\right).$$

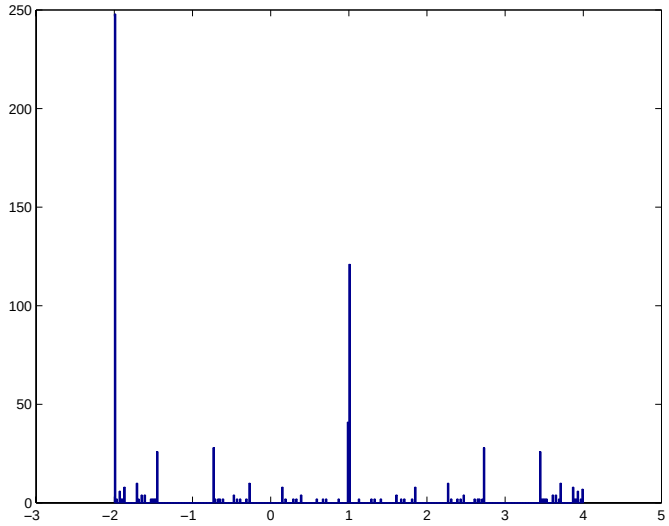


Figure 11. The histogram of the spectrum in case of the group $\overline{\Gamma}$

The computation of the spectrum and spectral measures for the graphs $\overline{\mathcal{S}}$ and $\overline{\overline{\mathcal{S}}}$ is similar to the computation for the graph $\mathcal{S}$.

Theorem 10.5 ([1]). *The spectra of the graphs $\overline{\mathcal{S}}$ and $\overline{\overline{\mathcal{S}}}$ coincide and are the closure of the set of points*

$$4,\ -2,\ 1,\ 1 \pm \frac{1}{\sqrt{2}} \sqrt{9 \pm \underbrace{\sqrt{45 \pm 4\sqrt{45 \pm \cdots \pm 4\sqrt{45 \pm 4 \cdot 3}}}}_{n \text{ times}}}$$

for $n = 0, 1, \ldots$.

It is a Cantor set of null Lebesgue measure. The KNS spectral measure μ is discrete and has values 0, $\frac{1}{3}$, $\frac{2}{9}$ for the first three points and $\frac{2}{3^{n+2}}$ at the points whose definition involves n radicals.

Corollary 10.6. *The Ihara zeta functions for the above graph is equal to*

$$\zeta_{\overline{\mathcal{S}}}(t) = \left(1 - t^2\right)^{-1} \left(1 + 16t + 7t^2\right)^{\frac{1}{3}} \left(1 - 8t + 7t^2\right)^{\frac{2}{3}} \prod_{j=0}^{\infty} \left(1 - t\left(1 \pm \frac{1}{\sqrt{2}} \sqrt{9 \pm \sqrt{45 \pm 4\sqrt{45 \pm \cdots \pm 4 \cdot 3}}}\right) + 4t^2\right)^{\frac{2}{3^j}}.$$

For these groups one has the following relation for $n \geq 2$

$$Q_n(\lambda, \eta) = (\gamma\delta)^{2 \cdot 3^{n-2}} (\alpha\beta)^{3^{n-2}} Q_{n-1}(F(\lambda, \eta)),$$

where $\alpha = 2 - \eta + \lambda$, $\beta = 2 - \eta - \lambda$, $\gamma = 1 + \eta + \lambda$, $\delta = 1 + \eta - \lambda$, and

$$F(\lambda, \eta) = \left(-2\frac{\lambda^2}{(2 - \eta + \lambda)(1 + \eta - \lambda)}, \eta + 2\frac{\lambda^2(\eta - \lambda - 1)}{(2 - \eta + \lambda)(1 + \eta - \lambda)}\right).$$

Proposition 10.7. *The map F is integrable with*

$$\psi(\lambda, \eta) = \frac{\eta^2 - \lambda\eta - 2\lambda^2 - 2 - \eta}{-\lambda},$$
$$\alpha(x) = 6 + \frac{1}{2}x - \frac{1}{2}x^2.$$

Proof. Indeed, the direct computation shows

$$\alpha\psi = 6 - \frac{\eta^2 - \lambda\eta - 2\lambda^2 - 2 - \eta}{2\lambda} - \frac{\eta^2 - \lambda\eta - 2\lambda^2 - 2 - \eta)^2}{2\lambda^2} = \psi F. \qquad \square$$

11 Zeta function related to the lamplighter group

The difference between the example considered here and previous cases is that although the Markov operator has pure point spectrum, the spectrum considered just as a set is the interval $[-1, 1]$ and is not totally disconnected as in the previous section.

Let $L = (\mathbb{Z}/2\mathbb{Z}) \wr \mathbb{Z} = \left(\bigoplus_{\mathbb{Z}} \mathbb{Z}/2\mathbb{Z}\right) \rtimes \mathbb{Z}$ be the lamplighter group [17]. It is a metabelian, non-finitely presented group of exponential growth generated by the elements a, v, where v is a generator of the group $\mathbb{Z}$, and

$$a = (\dots 010 \dots) \in \oplus_{\mathbb{Z}} \mathbb{Z}/2\mathbb{Z},$$

where 1 corresponds to the zero position. This group has a presentation

$$L = \langle a, v | a^2 = [a^{v^n}, a] = 1, n = 1, 2, \dots \rangle,$$

where $x^y = y^{-1}xy$ denotes the conjugation.

Consider the system of generators v, w where $w = av^{-1}$. Then L is isomorphic to the group $G(A) = \langle \mathcal{A}_v, \mathcal{A}_w \rangle$ generated by the automaton A from Figure 12 as was shown in [17], and the isomorphism $L \simeq G(A)$ holds via the map $v \to \mathcal{A}_v$, $w \to \mathcal{A}_w$.

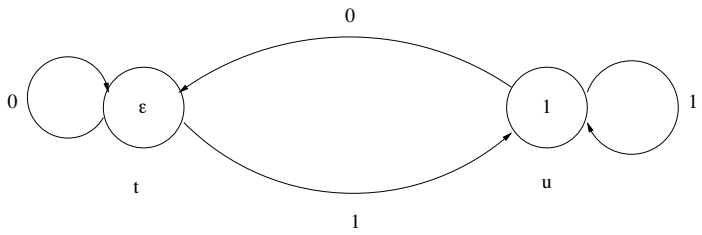


Figure 12. The automaton A generating the lamplighter group

Let $P = \mathrm{St}_L(\xi)$, $\xi \in \partial T$, where T is the binary tree. Then P is either cyclic of trivial (for the set of points of ∂T of the full measure) [17, Proposition 5]. Let $\mathcal{S}$ be the Cayley graph $\mathcal{S}(L, \{v, w\})$.

Theorem 11.1 ([17]). *The Markov operator on the graph $\mathcal{S}$ has a pure point spectrum concentrated on the points* $\cos \frac{p}{q}\pi$, $q = 2, 3, \dots$, $p \in \mathbb{Z}$. *The KNS spectral measure is discrete, concentrated on the above points with values*

$$\mu\left(\cos \frac{p}{q}\pi\right) = \frac{1}{2^q - 1} \quad \text{if} \quad (p, q) = 1.$$

Corollary 11.2. *For the lamplighter group the zeta function is equal to*

$$\zeta_{\mathcal{S}}(t) = \left(1 - t^2\right)^{-1} \prod_{q=2}^{\infty} \prod_{(p,q)=1} \left(1 - t4\cos \frac{p}{q}\pi + 3t^2\right)^{\frac{1}{2^q - 1}}.$$

Consider the operators

$$M(\lambda, \eta) = \rho(v) + \rho(v^{-1}) + \rho(w) + \rho(w^{-1}) - \lambda I - \eta\rho(a)$$

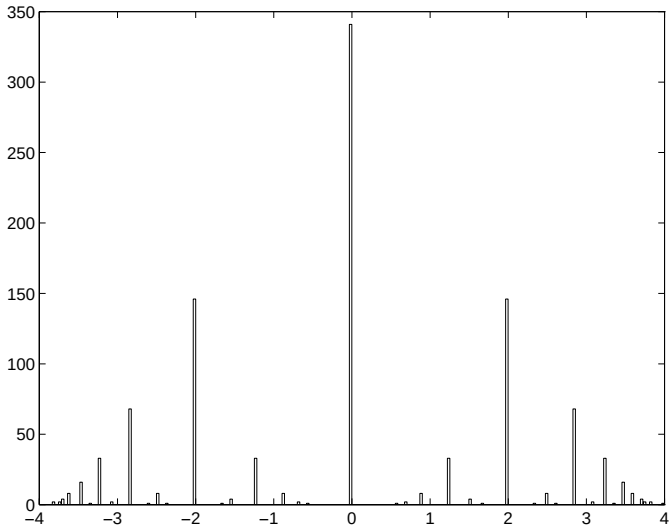


Figure 13. The histogram of the spectrum of M_n for the lamplighter group for $n = 10$

in $l^2(L/P)$ and

$$M_n(\lambda, \eta) = \rho_n(v) + \rho_n\left(v^{-1}\right) + \rho_n(w) + \rho_n\left(w^{-1}\right) - \lambda I_n - \eta\rho_n(a)$$

in $l^2(L/P_n)$. The Markov operators on $\mathcal{S}(G/P)$ and $\mathcal{S}(G/P_n)$ corresponds to the above operators when $\eta = 0$. Let $Q_n = \det M_n$. Then

$$Q_0 = 4 - \lambda - \eta,$$
$$Q_1 = (\eta - \lambda)(4 - \lambda - \eta),$$

and for $n \geq 2$ the following recursion holds:

$$Q_n(\lambda, \eta) = (\eta - \lambda)^{2^{n-1}} Q_{n-1}(F(\lambda, \eta)),$$

where

$$F(\lambda, \eta) = \left(\frac{\lambda^2 - \eta^2 - 2}{\lambda - \eta}, \frac{2}{\lambda - \eta}\right).$$

The map F admits the following inverse

$$F^{-1}(\lambda, \eta) = \left(\frac{\lambda}{2} + \frac{1}{\eta} + \frac{\eta}{2}, \frac{\lambda}{2} - \frac{1}{\eta} - \frac{\eta}{2}\right),$$

which is defined for $\eta \neq 0$.

Proposition 11.3. *The map F is integrable with $\psi(\lambda, \eta) = \lambda + \eta$ and $\alpha(x) = x$.*

Proof. The direct computation shows

$$\psi F = \lambda + \eta = \alpha\psi. \qquad \square$$

We see that in the case of the lamplighter group the dynamics of F is very simple. It preserves the lines l_c given by the equation $\lambda+\eta = c$. Depending on c, the dynamics of F on l_c is either a south-north dynamics or F is conjugated to a rotation of the circle. Indeed, we have

Proposition 11.4. *On the line l_c parameterized by the parameter λ, the map F is the Möbius map*

$$\lambda \to \frac{c\lambda - \left(\frac{c^2}{2}+1\right)}{\lambda - \frac{c}{2}},$$

and the corresponding matrix

$$C = \begin{pmatrix} c & -\frac{c^2}{2}-1 \\ 1 & -\frac{c}{2} \end{pmatrix} \in \mathrm{SL}(2,\mathbb{R})$$

is elliptic if $|c| < 4$, parabolic if $|c| = 4$ and hyperbolic if $|c| > 4$.

Proof. The second statement follows from $\mathrm{tr}(c) = \frac{c}{2}$ and the well known fact that a Möbius map c is elliptic if and only if $|\,\mathrm{tr}(c)| < 2$. The corresponding regions are shown in Figure 14. $\square$

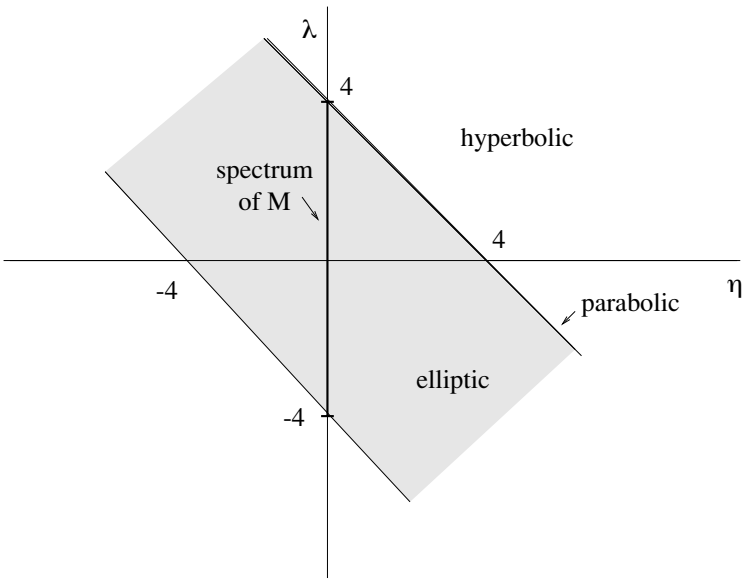


Figure 14. The dynamical system related to the automaton generating the lamplighter group

The map F has two fixed points

$$\lambda = \frac{3c \pm \sqrt{c^2 - 4}}{4}$$

on the line l_c in the case when it is hyperbolic.

The computation done in [17] shows that for $\lambda + \eta = 4\cos z$

$$Q_n(\lambda, \eta) = (4 - 4\cos z)2^n \frac{1}{(\sin z)^{2^{n-1}}} \prod_{k=2}^{n} (\sin kz)^{2^{n-k}} \sin(z(n+1)). \qquad (11.1)$$

The factor $\sin kz$ is related to the Chebyshev polynomial T_k by the relation

$$\sin kz = \sqrt{1 - T_k^2\left(\frac{\lambda + \eta}{4}\right)}.$$

In contrast to previous examples, the factors from the decomposition of Q_n are parameterized by the preimages of the Chebyshev polynomials (rather than by the preimages of the map α). The family $\{T_k\}_{k=1}^{\infty}$ is well known in dynamics. It consists of commuting transformations preserving the measure

$$\frac{dx}{\pi\sqrt{1 - x^2}}$$

on $[-1, 1]$, and T_k is conjugated to $z \to z^k$.

It is not clear how this measure can be used in our situation, but the alternative presentation for the Ihara zeta function can be easily obtained from the relation (11.1). Namely, since

$$\begin{aligned} \zeta_n(t) &= \left(1 - t^2\right)^{-2^n} \det(1 - tM_n + 3t^2)^{-1} \\ &= \left(1 - t^2\right)^{-2^n} t^{-2^n} Q_n^{-1}\left(\frac{1 + 3t^2}{-t}, 0\right), \end{aligned}$$

using the factorization (11.1) we get

$$\begin{aligned} \ln \zeta(t) = &-\frac{1}{4}\ln\left(1 - t^2\right) - \frac{1}{4}\ln\left(1 - T_1^2\left(\frac{1 + 3t^2}{-4t}\right)\right) \\ &- \sum_{k=2}^{\infty} \frac{1}{2^{k+1}} \ln\left(1 - T_k^2\left(\frac{1 + 3t^2}{-4t}\right)\right), \end{aligned}$$

and $|T_k(x)| < \sqrt{\frac{\pi}{2}}\frac{1}{2^k}$ if $x \in [-1, 1]$.

By the results from Section 5 the above series converges for $0 < |t| < \frac{1}{3}$ as then $\left|\frac{1+3t^2}{-4t}\right| < 1$.

12 An example of iterated monodromy group

Let H be the group generated by the automaton B from Figure 15 on three states. As the state c corresponds to identity map, the group H is indeed 2-generated by the elements a and b corresponding to the states of the automaton. The group H embeds into $H \wr S_2$ via the maps $a \mapsto (1, b)$ and $b \mapsto (1, a)\varepsilon$.

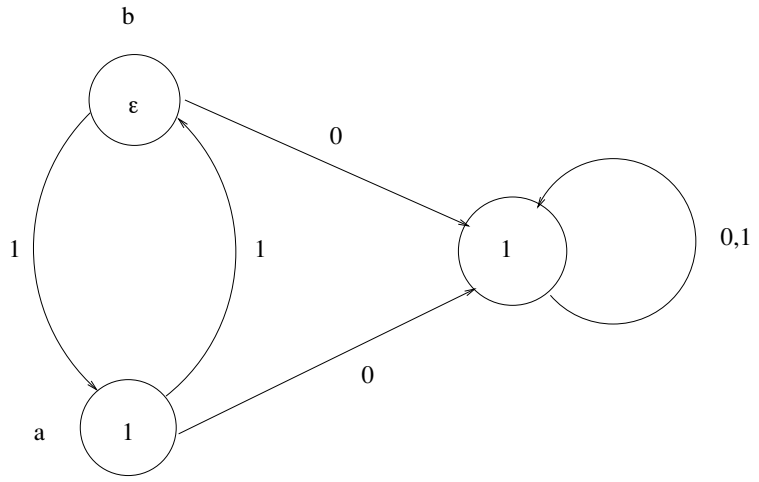


Figure 15. The automaton B

This group was studied in [18, 19] where many interesting facts about its algebraic and algorithmic properties were proven. In [2] it was shown that H is the iterated monodromy group of the map $z \to z^2 - 1$ and that the geometry of the Schreier graphs of this group is closely related to geometry of the Julia set of the map. An open question about H related to some famous problems is whether this group is amenable or not [19].

Now we are going to restate some results from [19] about spectral properties of H and to discuss them in connection with the subject of this paper.

We consider the quasi-regular representations ρ and ρ_n of H in the spaces $l^2(H/P)$ and $l^2(H/P_n)$. Let us consider the operators

$$M(\lambda, \eta) = \rho(a) + \rho(a^{-1}) + \lambda\big(\rho(b) + \rho(b^{-1})\big) - \eta I$$

and

$$M_n(\lambda, \eta) = \rho_n(a) + \rho_n(a^{-1}) + \lambda(\rho_n(b) + \rho_n(b^{-1})) - \eta I_n.$$

The operators M_n are given by $2^n \times 2^n$ matrices. Let $Q_n = \det M_n$.

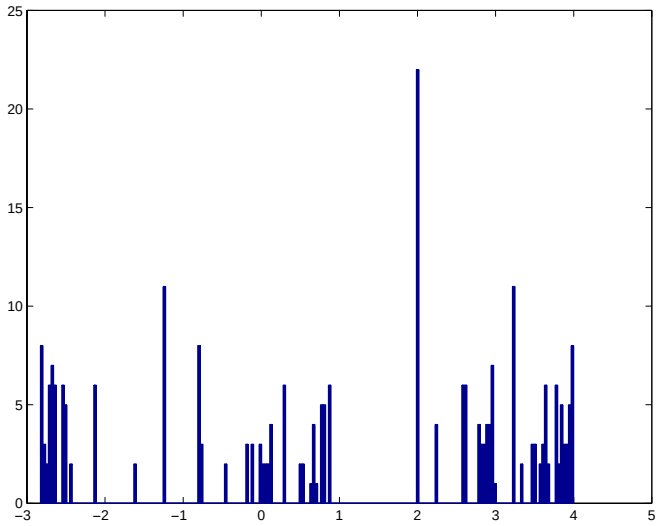


Figure 16. The histogram of the spectrum in case of the group H

For instance,

$$
\begin{aligned}
Q_1(\lambda,\eta) &= 2\eta + 2 - \lambda,\\
Q_2(\lambda,\eta) &= -(2\eta - 2 + \lambda)(2\eta + 2 - \lambda),\\
Q_3(\lambda,\eta) &= (2\eta + 2 - \lambda)(2\eta - 2 + \lambda)(4\eta^2 - \lambda^2 + 4)\\
Q_4(\lambda,\eta) &= -(-2 + \lambda)(2\eta + 2 - \lambda)(2\eta - 2 + \lambda)\\
&\quad (2\lambda^2 - 8 - \lambda^3 + 4\lambda + 4\eta^2\lambda)(4\eta^2 - \lambda^2 + 4)\\
Q_5(\lambda,\eta) &= (-2 + \lambda)^2(2\eta + 2 - \lambda)(2\eta - 2 + \lambda)(4\eta^2 + 2\lambda - \lambda^2)\\
&\quad (4\eta^2 + 4 - \lambda^2)(4\eta^2\lambda - 8 + 4\lambda + 2\lambda^2 - \lambda^3)\\
&\quad (16\eta^4\lambda - 8\eta^2\lambda^3 + 16\eta^2\lambda^2 + 8\eta^2\lambda - 16\eta^2\\
&\quad + \lambda^5 - 4\lambda^4 + 16\lambda^2 - 16\lambda)\\
Q_6(\lambda,\eta) &= -(-2 + \lambda)^3(2\eta - 2 + \lambda)(2\eta + 2 - \lambda)\\
&\quad (2\lambda^2 - 8 - \lambda^3 + 4\lambda + 4\eta^2\lambda)(4\eta^2 - \lambda^2 + 4)\\
&\quad (16\eta^4\lambda - 16\eta^2 + 8\eta^2\lambda - 16\lambda + 16\eta^2\lambda^2 + 16\lambda^2 - 8\eta^2\lambda^3\\
&\quad - 4\lambda^4 + \lambda^5)(8\eta^4 - 16 + 12\eta^2\lambda + 16\lambda - 6\eta^2\lambda^2 - 4\lambda^3 + \lambda^4)\\
&\quad (-8\lambda^8 + 16\lambda^7 - 384\eta^2\lambda - 160\eta^6\lambda^3 - 320\eta^2\lambda^2 + 72\eta^4\lambda^5\\
&\quad - 14\eta^2\lambda^7 - 512\lambda^2 + 256\eta^2 + 256\lambda - 608\eta^4\lambda + 608\eta^4\lambda^2\\
&\quad + 128\lambda^4 + 256\lambda^3 + 128\eta^8\lambda + 736\eta^2\lambda^3 + 320\eta^6\lambda^2 + 96\eta^6\lambda -
\end{aligned}
$$

$$-272\eta^2\lambda^4 - 160\lambda^5 + 136\eta^4\lambda^3 - 288\eta^4\lambda^4 + 32\lambda^6 + \lambda^9$$
$$-104\eta^2\lambda^5 + 84\eta^2\lambda^6 - 192\eta^6)(4\eta^2 + 2\lambda - \lambda^2)^2,$$

and we see that the degrees of factors grow. We have

Theorem 12.1 ([19]).

a) *If* $n \geq 1$ *then*

$$Q_{n+1}(\lambda, \eta) = \lambda^{2^{n+1}} Q_n(F(\lambda, \eta)),$$

where

$$F : \begin{cases} \lambda & \mapsto & -2 - \frac{\lambda(2-\lambda)}{\eta^2} \\ \eta & \mapsto & \frac{\lambda-2}{\eta^2} \end{cases}.$$

b) *The spectrum* Σ *of* $M(\lambda, \eta)$*, i.e., the set of pairs* (λ, η) (*including multiplicities*) *for which the operator* $M(\lambda, \eta)$ *is not invertible, is invariant with respect to the map* $F : \mathbb{R}^2 \to \mathbb{R}^2$*, i.e.,* $F^{-1}(\Sigma) = \Sigma$.

The spectrum Σ_n of M_n is given in Figure 16 while the set of accumulation points of the set $\bigcup_{n=1}^{\infty} F^{-n}(x_0, y_o)$ is shown in Figure 18, suggesting that the map F should have an attractor topology which is very different from the topology of the attractor shown in Figure 5.

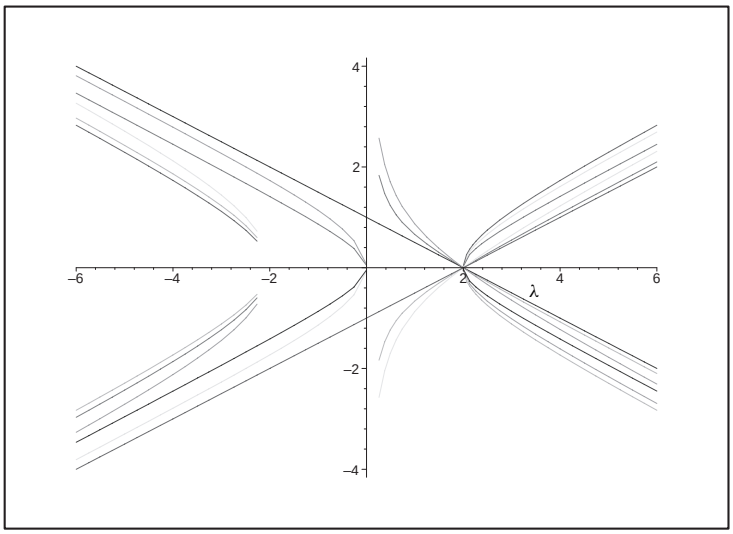


Figure 17. Zeros of $Q_n(\lambda, \eta)$

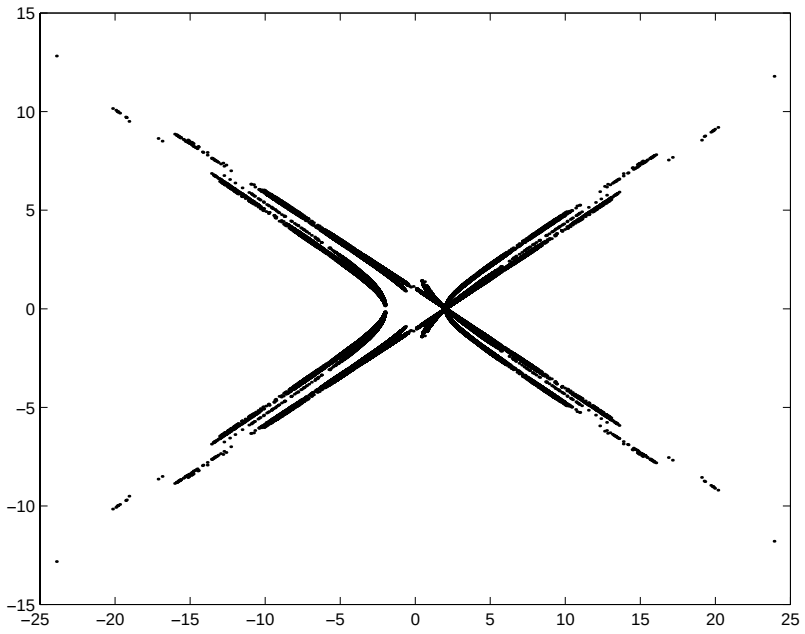


Figure 18. The attractor associated to the automaton from Figure 15

We see that the situation for this example is more complicated than in the previous examples. The degree of the curves into which Σ_n splits is growing while in the previous examples it was constant and equal to 2. At the moment we are not able to solve the problem of complete factorization of $Q_n(\lambda, \eta)$ and we do not know if the map F is integrable in the sense of (8.1). Looking at the picture of the histogram of the distribution of the spectrum of the operators

$$M_n(1, \eta) = \rho_n\big(a + a^{-1} + b + b^{-1}\big)$$

we can guess that the spectrum of the Markov operator

$$M(1, 0) = \rho\big(a + a^{-1} + b + b^{-1}t\big)$$

is a totally disconnected set, but this is only a conjecture.

13 Related topics and concluding remarks

There are three ways of averaging in the case of a group action. First of all one can use Følner averaging which works for amenable groups and was used in dynamical

systems (for instance in ergodic theorems) and in statistical physics (when defining averaged statistical sum) [5].

The second way is to use the von Neumann trace (this was used in [30]).

The third way (which works for residually finite groups) is to use approximation by finite quotients (see [14] for the definition of the averaged free energy on infinite residually finite groups and [7, 28] for computation of L^2 Betti numbers).

The second approach was used in this paper to define Ihara zeta function for residually finite graphs. One may ask if the first method works for amenable graphs. Our answer is yes at least in the situation of Cayley graphs. Here are our arguments.

Let G be an amenable group generated by a finite set S. Thus there exists a sequence of finite sets $A_n \subset G$, such that for any $r \geq 1$

$$\lim_{n\to\infty} \frac{|\partial_r A_n|}{|A_n|} = 0, \tag{13.1}$$

where $\partial_r A_n$ denotes the elements in $G \setminus A_n$ which are at distance at most r from A_n (by a distance we mean the distance on the Cayley graphs (G, S)).

Our idea is to define the zeta function for the graph using relation (4.5). Therefore we need to show that for any r the limit

$$\frac{c_r(A_n)}{|A_n|}$$

exists when n tends to infinity and A_n is considered as a subgraph of $S(G, S)$.

Proposition 13.1. *Let G be an amenable group and let $A_n \subset G$ be a Følner sequence. Then the limit*

$$\frac{c_r(A_n)}{|A_n|}$$

exists.

Proof. For any m, n we have

$$\frac{c_r(A_n)}{|A_n \cup \partial_r A_n|} \leq \frac{c_r(A_m \cup \partial_r A_m)}{|A_m|} \leq \frac{c_r(A_m) + |S|^r |\partial_r A_m|}{|A_m|}.$$

This implies that when $n, m \to \infty$

$$\left| \frac{c_r(A_n)}{|A_n|} - \frac{c_r(A_m)}{|A_m|} \right|$$

tends to zero. Thus the sequence $c_r(A_n)/|A_n|$ indeed converges to some limit. □

We can use the relation (4.6) to define $\widetilde{\zeta}_X(t)$.

Our definition of the Ihara function of an infinite graph X depends on the approximating sequence $\{X_n\}$ of finite graphs, but we do not have an example of two approximating sequences $\{X_n\}$ and $\{Y_n\}$ of the same graph, which lead to different zeta

functions. In particular, this question is open for Schreier graphs related to automata groups.

It is not clear how our definition of the Ihara zeta function of Cayley graphs of amenable groups agrees with other definitions. For simple examples like $\mathbb{Z}^d$ this can be checked by a direct computation.

It is also an open problem for which finite automata one can associate a multi-dimensional map F relating the n-th and $n+1$-th determinants and whether such a map is integrable. New examples in this direction would be welcome. Such examples could provide new examples of spectra of Schreier graphs and Cayley graphs related to automata groups.

In all examples from this paper with an integrable map F, the spectrum of the Markov operator is an intersection of the line $\{\lambda = const\}$ with the part of $\mathrm{Im}(F)$ which corresponds to the chaotic region of behavior of the map α (this is the case for the examples from Sections 10 and 11, in which $\alpha = z^2 - 1, 4 - z - z^2, 6 + z - z^2$) or the chaotic behavior of the map F on the integral lines (which is the case for the lamplighter group, when F on l_c acts as a rotation of a circle). It would be interesting to understand better this phenomenon with irrational angle for most values of $|c| < 4$.

In analogy with the finite case, one can define an infinite k-regular Ramanujan graph by the condition $\mathrm{Sp}\, M \subset \left[-\frac{2\sqrt{k-1}}{k}, \frac{2\sqrt{k-1}}{k} \right]$. For a residually finite graph this corresponds to the situation when all zeros are on $\mathrm{Re}\, z = \frac{1}{2}$. The characterization of such graphs in terms of the cogrowth follows from the formula for the spectral radius found in [9] (namely as graphs with cogrowth at most $\sqrt{2m-1}$). Approximating an infinite Ramanujan graph X (which is not necessary a tree) by finite ones could possibly produce new examples of finite Ramanujan graphs or at least of expanders.

Acknowledgment. We would like to express our greatest thanks to Etienne Ghys for several discussions and ideas. We would also like to thank Bruno Sevennec for his valuable help with the computers. The first author was supported by the NSF grant 0308985.

References

[1] L. Bartholdi and R. I. Grigorchuk, On the spectrum of Hecke type operators related to some fractal groups, Proc. Steklov Inst. Math. 231 (2000), 1–41.

[2] L. Bartholdi, R. I. Grigorchuk and V. Nekrashevych, From fractal groups to fractal sets, in: Fractals in Graz (Graz, 2001), Birkhäuser, Basel 2003, 25–118.

[3] H. Bass, The Ihara-Selberg zeta function of a tree lattice, Internat. J. Math. 3 (1992), 717–797.

[4] D. M. Cvetković, M. Doob and H. Sachs, Spectra of Graphs. Theory and Applications. Third edition, Johann Ambrosius Barth, Heidelberg 1995.

[5] J. Dodziuk and V. Mathai, Approximating L^2 invariants of amenable covering spaces: a combinatorial approach, J. Funct. Anal. 154 (1998), 359–378.

[6] A. E. Eremenko, Some functional equations connected with the iteration of rational functions, Leningrad Math. J. 1 (1990), 905–919.

[7] M. Farber, Geometry of growth: approximation theorems for L^2 invariants, Math. Ann. 311 (1998), 335–375.

[8] F. Gecseg and I. Peak, Algebraic Theory of Automata, Disquisitiones Mathematicae Hungaricae, 2, Akademiai Kiado, Budapest 1972.

[9] R. I. Grigorchuk, Symmetrical random walks on discrete groups, in: Multicomponent Random Systems, Adv. Probab. Related Topics 6, Dekker, New York 1980, 285–325.

[10] R. I. Grigorchuk, On Burnside's problem on periodic groups, Funct. Anal. Appl. 14 (1980), 41–43.

[11] R. I. Grigorchuk, On the Milnor problem of group growth, Soviet Math. Dokl. 28 (1983), 23–26.

[12] R. I. Grigorchuk, P. Linnell, T. Schick and A. Żuk, On a question of Atiyah, C. R. Acad. Sci. Paris Sér. I Math. 331 (2000), 663–668.

[13] R. I. Grigorchuk, V. V. Nekrashevych and V. I. Sushchansky, Automata, dynamical systems and groups, Proc. Steklov Inst. Math. 231 (2000), 128–203.

[14] R. I. Grigorchuk and A. M. Stepin, Gibbs states on countable groups, Theory Probab. Appl. 29 (1984), 359–362.

[15] R. I. Grigorchuk and J. S. Wilson, The conjugacy problem for certain branch groups, Proc. Steklov Inst. Math. 231 (2000), 204–219

[16] R. I. Grigorchuk and A. Żuk, On the asymptotic spectrum of random walks on infinite families of graphs, in: Random Walks and Discrete Potential Theory (Cortona, 1997), Sympos. Math., XXXIX, Cambridge University Press, Cambridge 1999, 188–204.

[17] R. I. Grigorchuk and A. Żuk, The lamplighter group as a group generated by a 2-state automaton and its spectrum, Geom. Dedicata 87 (2001), 209–244.

[18] R. I. Grigorchuk and A. Żuk, On a torsion-free weakly branch group defined by a three state automaton, Internat. J. Algebra Comput. 12 (2002), 1–24.

[19] R. I. Grigorchuk and A. Żuk, Spectral properties of a torsion-free weakly branch group defined by a three state automaton, in: Computational and Statistical Group Theory (Las Vegas, NV/Hoboken, NJ, 2001), Contemp. Math. 298, Amer. Math. Soc., Providence, RI, 2002, 57–82.

[20] R. I. Grigorchuk and A. Żuk, Self-similar C^* algebras and spectra of Markov elements, in preparation.

[21] M. Hall, Coset representation in free groups, Trans. Amer. Math. Soc. 67 (1949), 421–432.

[22] K. Hashimoto, Zeta functions of finite graphs and representations of p-adic groups, in: Automorphic Forms and Geometry of Arithmetic Varieties, Adv. Stud. Pure Math. 15, Academic Press, Boston, MA, 1989, 211–280.

[23] Y. Ihara, On discrete subgroups of the two by two projective linear group over $\mathfrak{p}$-adic fields, J. Math. Soc. Japan 18 (1966), 219–235.

[24] H. Kesten, Symmetric random walks on groups, Trans. Amer. Math. Soc. 92 (1959), 336–354.

[25] M. Kotani, T. Sunada, Zeta functions of finite graphs, J. Math. Sci. Univ. Tokyo 7 (2000), 7–25.

[26] W. Li, Geometry, graph theory and number theory, in: Algebra and Geometry (Taipei, 1995), Lect. Algebra Geom. 2, Internat. Press, Cambridge, MA, 1998, 83–102.

[27] A. Lubotzky, Discrete Groups, Expanding Graphs and Invariant Measures. With an appendix by Jonathan D. Rogawski, Progr. Math. 125, Birkhäuser, Basel 1994.

[28] W. Lück, Approximating L^2-invariants by their finite-dimensional analogues, Geom. Funct. Anal. 4 (1994), 455–481.

[29] W. Lück, L^2-Invariants: Theory and Applications to Geometry and K-theory, Ergeb. Math. Grenzgeb. (3) 44, Springer-Verlag, Berlin 2002.

[30] F. Lund, M. Rasetti and T. Regge, The Ising model and dimer model on the Lobachevskii plane, Theoret. and Math. Phys. 33 (1977), 246–271.

[31] M. Lyubich, The quadratic family as a qualitatively solvable model of chaos, Notices Amer. Math. Soc. 47 (2000), 1042–1052.

[32] W. S. Massey, A Basic Course in Algebraic Topology, Grad Texts in Math. 127, Springer-Verlag, New York 1991.

[33] B. Mohar, W. Woess, A survey on spectra of infinite graphs, Bull. London Math. Soc. 21 (1989), 209–234.

[34] H. Nagoshi, On arithmetic infinite graphs, Proc. Japan Acad. Ser. A Math. Sci. 76 (2000), 22–25.

[35] S. Northshield, Two proofs of Ihara's theorem, in: Emerging Applications of Number Theory (Minneapolis, MN, 1996), IMA Vol. Math. Appl. 109, Springer-Verlag, New York 1999, 469–478.

[36] S. Northshield, A note on the zeta function of a graph, J. Combin. Theory Ser. B 74 (1998), 408–410.

[37] P. Sarnak, Some Applications of Modular Forms, Cambridge Tracts Math. 99, Cambridge University Press, Cambridge 1990.

[38] J.-P. Serre, Répartition asymptotique des valeurs propres de l'opérateur de Hecke T_p, J. Amer. Math. Soc. 10 (1997), 75–102.

[39] H. M. Stark, Multipath zeta functions of graphs, in: Emerging Applications of Number Theory (Minneapolis, MN, 1996), IMA Vol. Math. Appl. 109, Springer-Verlag, New York 1999, 601–615.

[40] H. M. Stark, A. A. Terras, Zeta functions of finite graphs and coverings. II, Adv. Math. 154 (2000), 132–195.

[41] T. Sunada, Fundamental groups and Laplacians, in: Geometry and Analysis on Manifolds (Katata/Kyoto, 1987), Lecture Notes in Math. 1339, Springer-Verlag, Berlin 1988,

248–277.

[42] A. P. Veselov, Integrable mappings, Russian Math. Surveys 46 (5)(1991), 1–51.

[43] M. Waldschmidt, P. Moussa, J. M. Luck and C. Itzykson (eds.), From Number Theory to Physics. Papers from the Meeting on Number Theory and Physics held in Les Houches, March 7–16, 1989, Springer-Verlag, Berlin 1992.

[44] W. Woess, Random Walks on Infinite Graphs and Groups, Cambridge Tracts in Math. 138, Cambridge University Press, Cambridge 2000.

Rostislav I. Grigorchuk, Steklov Mathematical Institute, Gubkina Str. 8 Moscow, 119991, Russia
E-mail: grigorch@mi.ras.ru

Andrzej Żuk, CNRS, Ecole Normale Supérieure de Lyon, Unité de Mathématiques Pures et Appliquées, 46, Allée d'Italie, F-69364 Lyon Cedex 07, France
E-mail: azuk@umpa.ens-lyon.fr

Simplicité de spectres de Lyapounov et propriété d'isolation spectrale pour une famille d'opérateurs de transfert sur l'espace projectif

Yves Guivarc'h et Émile Le Page

Abstract. We consider a finite dimensional real vector space and a non-degenerate probability measure μ on the corresponding linear group. We study an operator-valued Laplace transform of μ and prove a spectral gap theorem for the corresponding transfer operators on the projective space. We introduce a general class of Markov systems and prove a general theorem of simplicity of the Lyapunov spectrum for these systems, which implies the spectral gap property.

Abstract. Nous considérons un espace vectoriel réel de dimension finie, une mesure de probabilité μ non dégénérée sur le groupe linéaire correspondant. Nous étudions les opérateurs transformés de Laplace de μ sur l'espace projectif et nous prouvons un théorème d'isolation spectrale pour ces opérateurs. Nous sommes amenés à introduire une classe générale de systèmes Markoviens, à prouver un théorème général de simplicité du spectre de Lyapounov pour ces systèmes et la propriété d'isolation spectrale en résulte.

Contents

1 Introduction 182

2 Résultats principaux 184

3 Proximalité, μ-proximalité et propriétés algébriques des sous-groupes de GL(V) 191
- 3.1 Proximalité sur l'espace projectif . . . 191
- 3.2 Proximalité sur l'espace des drapeaux . . . 196

4 Les opérateurs de transfert P^s et leur normalisation 200
- 4.1 Construction d'une fonction propre pour l'opérateur P^s . . . 200
- 4.2 Convexité logarithmique et comportement asymptotique du rayon spectral de P^s . . . 206
- 4.3 Les opérateurs Markoviens Q^s et la condition de continuité en variation des probabilités Q^s_v . . . 211
- 4.4 Extension des théorèmes précédents à l'espace des drapeaux . . . 216

5 Frontières et noyaux harmoniques 220
5.1 Notations . . . 220
5.2 Construction de noyaux harmoniques . . . 222
5.3 Unicité du noyau harmonique . . . 223
5.4 Ergodicité du système Markovien $(X, q \otimes \mu)$ et du décalage correspondant . . 227
5.5 Cas où l'application équivariante $z(x, \omega)$ est indépendante de x . . . 228

6 Le noyau harmonique et les propriétés de contraction dans l'espace projectif 229
6.1 Les convergences en direction . . . 229
6.2 La continuité absolue des probabilités Q_v^s . . . 235

7 Les exposants caractéristiques de $S_n(\omega)$ 238
7.1 Quelques lemmes préparatoires . . . 238
7.2 Le plus grand exposant est de multiplicité 1 . . . 243

8 Propriétés de contraction et d'isolation spectrale 246
8.1 Opérateurs Markoviens contractants en moyenne . . . 246
8.2 Opérateurs de transfert sur l'espace projectif . . . 251

1 Introduction

Soit μ une probabilité sur le groupe linéaire $\mathrm{GL}(V)$, où V est un espace vectoriel réel de dimension finie d. On fixe une norme (notée $\|\cdot\|$) sur V, et l'espace $\mathrm{End}\, V$ est donc naturellement normé. On note S_μ le support de μ, $\Omega = S_\mu^{\mathbb{N}}$ l'espace produit, $\mu^{\mathbb{N}}$ la probabilité produit sur Ω de facteurs μ, $g_k(\omega)$ la $k^{\text{ième}}$ coordonnée de $\omega \in \Omega$, et l'on considère le produit de matrices indépendantes $S_n(\omega) = g_n(\omega) \dots g_1(\omega)$. Les théorèmes connus donnant le comportement asymptotique de $S_n(\omega)$ (cf. par exemple [36, 24, 38, 6, 19]) montrent que, sous de larges conditions, la suite de variables aléatoires $\log \|S_n(\omega)v\|$ $(v \in V)$ se comporte essentiellement comme une somme de variables aléatoires indépendantes identiquement distribuées [14, 7, 45]; cette dernière situation correspond d'ailleurs au cas spécial $d = 1$. On développe ici ce parallèle dans la direction de la transformée de Laplace. Ici la mesure μ vérifie seulement la condition de non dégénérescence notée (i–p) plus bas, et peut donc être purement atomique.

Pour $z = s + it \in \mathbb{C}$ avec $(s, t) \in \mathbb{R}^2$, la transformée de Laplace de la variable aléatoire réelle $\log \|S_n(\omega)v\|$ est donnée par

$$\mathbf{E}_\mu\left(e^{z \log \|S_n(\omega)v\|}\right) = \mathbf{E}_\mu\left(\|S_n(\omega)v\|^z\right) = \int \|gv\|^z \, d\mu^n(g),$$

où μ^n désigne la $n^{\text{ième}}$ convolée de μ. Elle est donc bien définie si μ satisfait $\int \|g\|^s \, d\mu(g) < +\infty$ pour tout $s \in \mathbb{R}_+$. On supposera cette condition vérifiée dans cette introduction et l'on se limitera aussi à $z = s \in \mathbb{R}_+$. On note $P(V)$ l'espace projectif de V et $g.v \in P(V)$ le transformé de $v \in P(V)$ par $g \in \mathrm{GL}(V)$. Introduisons

suivant [17] l'opérateur P^s sur $P(V)$ par

$$P^s\varphi(v) = \int \varphi(g.v)\|gv\|^s \, d\mu(g),$$

où $\varphi \in C[P(V)]$. On a alors $\int \|gv\|^s \, d\mu^n(g) = \big((P^s)^n 1\big)(v)$, ce qui montre que l'étude asymptotique de $\mathbf{E}_\mu(\|S_n(\omega)v\|^s)$ se réduit à celle de l'itération de l'opérateur P^s. Cet opérateur correspond à la transformée de Mellin si $V = \mathbb{R}$, End $V = \mathbb{R}_*$, et l'on étudie donc ici la structure de P^s, si dim $V = d > 1$. Si l'on considère le réel $k(s)$ défini par $k(s) = \lim_n[\int \|g\|^s \, d\mu^n(g)]^{1/n}$, on montre que, sous la condition de non dégénérescence notée (i–p), l'itération de P^s $(s \geq 0)$ est contrôlée de manière précise par les puissances de $k(s)$. La fonction $k(s)$ est analytique, possède plusieurs des propriétés de la transformée de Laplace d'une probabilité portée par $\mathbb{R}$, et une partie de nos résultats est consacrée à préciser ce fait.

Cette étude est motivée par des questions relatives aux milieux désordonnés, aux relations de récurrence à coefficients aléatoires, aux théorèmes de renouvellement pour les produits de matrices indépendantes [33, 35] et aux estimations précises de grandes déviations pour les produits de matrices aléatoires.

La famille des opérateurs P^s intervient en particulier dans l'étude de la localisation pour l'équation de Schrödinger stationnaire avec un potentiel aléatoire V_n de loi μ sur $\mathbb{R}$ ou $\mathbb{Z}$ [21, 10]. Sur $\mathbb{Z}$, cette équation s'écrit

$$-(u_{n+1} + u_{n-1}) + V_n u_n = \lambda u_n,$$

où $(u_n)_{n\in\mathbb{Z}}$ est la fonction inconnue. Alors, μ est une probabilité sur $\mathrm{GL}(2, \mathbb{R})$ concentrée sur les matrices de la forme $\left(\begin{smallmatrix} x-\lambda & -1 \\ 1 & 0 \end{smallmatrix}\right)$, où x est un réel aléatoire de répartition donnée décrivant le potentiel V, et λ est un réel fixé qui a le sens d'une valeur propre de l'opérateur de $\ell^2(\mathbb{Z})$ défini par le premier membre de l'équation précédente.

La famille des opérateurs P^s joue aussi un rôle essentiel dans l'étude des relations de récurrence à coefficients aléatoires $X_{n+1} = A_{n+1}X_n + B_{n+1}$ où $(A_n, B_n) \in \mathrm{GL}(d, \mathbb{R}) \times \mathbb{R}^d$ sont des variables aléatoires i.i.d. Cette étude est menée en [33] sous l'hypothèse de positivité des A_n ou bien celle d'existence d'une densité pour μ, et les opérateurs P^s permettent alors d'étudier les fluctuations des variables $\log \|S_n(\omega)v\|$. Plus précisément, si l'on pose $M(\omega, v) = \sup\{\log \|S_n(\omega)v\| : n \geq 1\}$, l'étude de l'allure à l'infini de la loi stationnaire de X_n se ramène à celle d'un équivalent exponentiel de $\mathbf{P}_\mu\{M(\omega, v) > x\}$ si x tend vers l'infini (Théorème 2.5), étude qui nécessite de développer la théorie du renouvellement pour les produits de matrices aléatoires. Les applications de la transformée de Laplace aux relations de récurrence $X_{n+1} = A_{n+1}X_n + B_{n+1}$ et à leurs mesures stationnaires seront données dans un autre article basé sur ce travail (cf. [37] pour un exposé des résultats principaux et [28] pour un énoncé du théorème de renouvellement dans le cadre plus général de la frontière de Martin [1]). Dans ce travail, on établit les propriétés d'isolation spectrale des opérateurs P^s qui sont nécessaires dans l'étude des situations précédentes.

La restriction $s \geq 0$ fait jouer un rôle essentiel aux propriétés de contraction des applications projectives et conduit à un formalisme parallèle au formalisme thermody-

namique unidimensionnel, qui n'en est cependant pas un cas particulier. Les aspects qualitatifs de ces propriétés de contraction s'expriment commodément dans le langage de la théorie des frontières de H. Furstenberg; on en développe ici seulement quelques propriétés et l'on renvoie, pour des informations plus détaillées, aux exposés [18] et [39, Chapter 6]. Dans une direction plus quantitative, la quasi-compacité de $P^s/k(s)$ dans un espace de fonctions Höldériennes convenable sur $P(V)$ est établie ainsi que la validité d'une condition "de Doeblin–Fortet" pour les opérateurs Markoviens associés (cf. Théorème 2.2) en montrant d'abord une propriété d'absolue continuité des lois conditionnelles analogue à celle de la dynamique hyperbolique. Les résultats obtenus s'étendent, en renforçant les hypothèses d'irréductibilité, au cas multidimensionnel des actions sur les frontières, le paramètre $s \in \mathbb{R}_+$ étant remplacé par un vecteur (s) de $\mathbb{R}_+^d$.

Le résultat principal de ce travail est donc le Théorème 2.2; cependant nous soulignons le rôle technique essentiel du Théorème 2.8 relatif au spectre de Lyapounov, et de sa preuve basée sur [25] et [9]. Notons qu'un énoncé proche du Théorème 2.2 a été utilisé en [38], en relation avec les estimations précises de grandes déviations. La preuve donnée dans ce dernier travail est incomplète, dans le cas des opérateurs précédents, mais le Théorème 2.8 et le Théorème 4.20 permettent de justifier l'énoncé correspondant. Notons aussi que la preuve du Théorème 2.8 a été reprise en [5], dans le cadre de la dynamique hyperbolique, où un énoncé plus général que le Théorème 2.8 est obtenu. Enfin on pourra consulter [30, 8] pour des études d'opérateurs analogues aux opérateurs P^s, des théorèmes limites correspondants et des applications en géométrie et en dynamique. Dans le cas où s est imaginaire pur, la propriété d'isolation spectrale a été prouvée en [36], par une méthode qui est un cas particulier de celle développée ici.

Une partie de nos résultats a été décrite en [37] et [24] (cf. aussi [23]). Les preuves sont développées ici à partir de [37] en tenant compte des informations géométriques supplémentaires obtenues depuis en [20, 24, 22, 19, 3, 4, 43, 39]. On donne d'abord un aperçu des résultats principaux en Section 2, puis on développe des techniques préliminaires en Sections 3,4,5 dont on tire les conséquences en Section 6, et les résultats principaux sont obtenus en Section 7 et Section 8.

2 Résultats principaux

Notons $\Gamma_\mu \subset \mathrm{GL}(V)$ le semi-groupe fermé engendré par le support S_μ de μ, et $\check{\mu}$ l'image de μ par la symétrie $g \mapsto g^{-1}$.

Notre étude se place essentiellement dans le cadre de la définition suivante:

Définition 2.1. On dira que le semi-groupe $\Gamma \subset \mathrm{GL}(V)$ vérifie la condition (i–p) (pour irréductibilité et proximalité) si

a) Le semi-groupe agit d'une manière fortement irréductible sur V, c'est-à-dire il n'existe pas de réunion finie de sous-espaces propres $V_i \subset V$, $(i = 1, 2, ...r)$ telle que $g\big(\bigcup_{i=1}^r V_i\big) = \bigcup_{i=1}^r V_i$ pour tout $g \in \Gamma$;

b) Γ contient au moins un élément g dont les modules des valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_d$ satisfont $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_d|$.

Considérons les fonctions log-convexes k^+, k définies par

$$k^+(s) = \lim_n \left[\sup_{\|v\|=1} \int \|gv\|^s d\mu^n(g) \right]^{1/n}, \quad k(s) = \lim_n \left[\int \|g\|^s \, d\mu^n(g) \right]^{1/n},$$

et, pour un fermé $F \subset \mathrm{GL}(V)$, introduisons l'exposant de croissance de F

$$\gamma^\infty(F) = \lim_n \frac{1}{n} \sup\big\{\log \|g_n \ldots g_1\| : g_k \in F,\ 1 \leq k \leq n\big\}.$$

Notons S^{d-1} la sphère unité de V, observons que $P(V)$ s'identifie au quotient de S^{d-1} par la symétrie $x \to -x$ et notons $v \to \bar{v}$ la projection de S^{d-1} sur $P(V)$. On définit la distance δ sur $P(V)$ par

$$\delta(x, x') = \inf\big\{\|v - v'\| : \|v\| = \|v'\| = 1, \bar{v} = x, \bar{v}' = x'\big\}.$$

Désignons par $H_\varepsilon(P(V))$ l'espace de Banach des fonctions ε-Höldériennes sur $P(V)$ normé par $\|f\|_\varepsilon = |f|_\infty + [f]_\varepsilon$, où

$$|f|_\infty = \sup_{x \in P(V)} |f(x)|, \quad [f]_\varepsilon = \sup_{x, x' \in P(V)} \frac{|f(x) - f(x')|}{\delta^\varepsilon(x, x')}.$$

Notons enfin γ_1 le premier exposant caractéristique du produit $S_n(\omega)$:

$$\gamma_1 = \lim_n \frac{1}{n} \int \log \|g\| \, d\mu^n(g).$$

Théorème 2.2. *Supposons que Γ_μ vérifie la condition* (i–p) *et que pour $\sigma > 0$ on ait*

$$\int \|g\|^\sigma \, d\mu(g) < +\infty, \quad \int \|g\|^\sigma \log |\det g| \, d\mu(g) > -\infty.$$

Alors, pour $s \in [0, \sigma[$ on a $k^+(s) = k(s)$, la fonction $k(s)$ est analytique sur $[0, \sigma[$, et sa dérivée en 0 vaut γ_1. Il existe une fonction continue positive $c(s, v) > 0$ définie sur $\mathbb{R}_+ \times P(V)$ telle que l'on ait

$$\mathbf{E}_\mu(\|S_n(\omega)v\|^s) = k^n(s)[c(s, v) + \varepsilon_n(s, v)]$$

avec $\limsup_n |\varepsilon_n(s, v)|^{1/n} < 1$ *uniformément en $v \in P(V)$.*

Sur $C[P(V)]$, l'opérateur P^s a pour rayon spectral $k(s)$. Il admet, pour $s \in [0, \sigma[$, une unique fonction propre continue strictement positive normalisée e_s, une unique mesure propre positive normalisée ν^s, et les valeurs propres correspondantes coincident avec $k(s)$:

$$P^s e_s = k(s) e_s, \quad P^s \nu^s = k(s) \nu^s, \quad |e_s|_\infty = 1, \quad \nu^s(e_s) = 1.$$

Si $\sigma = \infty$, on a $\lim_{s\to\infty} \log k(s)/s = \gamma^\infty(S_\mu)$.

Pour tout $s \in [0, \sigma[$ et pour ε assez petit, $k(s)$ est valeur spectrale isolée de P^s sur $H_\varepsilon(P(V))$. Plus précisément, P^s s'écrit sous forme de somme directe

$$P^s = k(s)(\nu^s \otimes e_s + U^s),$$

où U^s commute avec le projecteur $\nu^s \otimes e_s$, c'est-à-dire

$$U^s(\nu^s \otimes e_s) = (\nu^s \otimes e_s)U^s = 0,$$

et admet un rayon spectral strictement inférieur à 1.

Il découle de ce théorème qu'on peut normaliser l'opérateur P^s et définir l'opérateur Markovien Q^s par

$$Q^s\varphi(v) = \frac{1}{k(s)e_s(v)} P^s[\varphi e_s](v) = \int q^s(v, g)\varphi(g.v)\, d\mu(g),$$

où

$$q^s(v, g) = \frac{e_s(g.v)}{k(s)e_s(v)} \|gv\|^s.$$

On a alors en itérant

$$(Q^s)^n\varphi(v) = \int q_n^s(v, g)\varphi(g.v)\, d\mu^n(g),$$

où

$$q_n^s(v, g) = \frac{1}{k^n(s)} \frac{e_s(g.v)\|gv\|^s}{e_s(v)}.$$

Définition 2.3. Soit Q un opérateur linéaire borné d'un espace de Banach H et r un réel positif. On dira que Q est r-quasi-compact si la résolvante $(zI - Q)^{-1}$ de Q est méromorphe pour $|z| > r$ et si les sous-espaces spectraux correspondants aux pôles sont de dimension finie.

Si $r(Q)$ est le rayon spectral de Q, alors Q est $r(Q)$-quasi-compact, mais la notion n'est intéressante que pour $r < r(Q)$. On pourra se reporter à [31] et [29] pour l'étude de conditions suffisantes de r-quasi-compacité, en particulier dans le cadre d'un opérateur Markovien Q. Les opérateurs de transfert du formalisme thermodynamique uni-dimensionnel (voir plus bas) satisfont une telle propriété (cf. [42], [8]).

Corollaire 2.4. *Avec les notations précédentes, posons pour $\varepsilon > 0$, $s \in [0, \sigma[$*

$$\rho_n^s(\varepsilon) = \sup_{v,v'} \int \frac{\delta^\varepsilon(g.v, g.v')}{\delta^\varepsilon(v, v')} q_n^s(v, g)\, d\mu^n(g), \quad \rho^s(\varepsilon) = \lim_n \left[\rho_n^s(\varepsilon)\right]^{1/n},$$

et considérons pour $z = s + it \in \mathbb{C}$, $s = \Re z \in [0, \sigma[$, l'opérateur Q^z défini par $Q^z\varphi = \dfrac{1}{k(s)e_s} P^z(e_s\varphi)$. Alors, pour ε assez petit, on a $\rho^s(\varepsilon) < 1$, et pour tout $n \in \mathbb{N}$

il existe une constante $C_{z,n}$ telle que, pour toute $\varphi \in H_\varepsilon[P(V)]$

$$[(Q^z)^n\varphi]_\varepsilon \leq \rho_n^s(\varepsilon)[\varphi]_\varepsilon + C_{z,n}|\varphi|_\infty.$$

L'opérateur Q^z est $\rho^s(\varepsilon)$-quasi-compact dans $H_\varepsilon[P(V)]$, et la résolvante de Q^z admet un pôle simple en $\lambda = 1$, si $t = 0$. Si $t \neq 0$, le rayon spectral de Q^z sur $H_\varepsilon[P(V)]$ est strictement inférieur à 1.

On montre que, en général, $k^+(s) \neq k(s)$ si $s < -d$, et que les propriétés spectrales précédentes de P^s ne sont plus, en général, valides pour $s < -d$.

On peut illustrer ces théorèmes par une application (cf. [37]) qui correspond, si $d = 1$, à l'estimation de Cramér, bien connue en théorie des fluctuations [14, p. 414], [45, p. 222].

Théorème 2.5. *Supposons que Γ_μ vérifie la condition* (i–p) *, que $k(s)$ soit définie sur $\mathbb{R}_+$, que $\gamma^\infty(S_\mu) > 0$ et $\gamma_1 < 0$. Soit χ le réel positif défini par $k(\chi) = 1$ et $\chi > 0$. Alors pour tout $v \in V$ avec $\|v\| = 1$ la famille de fonctions*

$$t^\chi \mathbf{P}_\mu\{\sup_n \|S_n(\omega)v\| > t\}$$

converge si $t \to +\infty$ vers une fonction continue strictement positive proportionnelle à $e_\chi(v)$.

Ce théorème, ainsi que les théorèmes de renouvellement correspondants, seront justifiés dans un prochain travail. On peut observer que, pour $d = 1$, l'énoncé correspondant comporte la condition supplémentaire que le sous-groupe engendré par le support de μ soit dense dans $\mathbb{R}$ (cf. [14, p. 391]), condition dont l'analogue pour $d \geq 2$ est réalisée à cause de la non-commutativité de Γ_μ.

Comme souligné plus haut, le Théorème 2.2 est basé sur l'étude d'un spectre de Lyapounov donné par le Théorème 2.8 que nous décrivons ci-dessous. On considère le cadre général suivant.

Soit (X, d) un espace métrique compact, μ une mesure de Radon positive sur $C(X, X)$, de support S_μ, $q(x, a)$ une fonction continue positive sur $X \times S_\mu$ avec $\int q(x, a)\, d\mu(a) = 1$ pour tout $x \in X$. On considère l'espace produit $\Omega = S_\mu^{\mathbb{N}}$ et la mesure produit $\mu^{\mathbb{N}}$ sur Ω. On note $S_\mu^{\otimes k}$ les espaces produits partiels et $\mu^{\otimes k}$ la mesure produit sur $S_\mu^{\otimes k}$, projection de $\mu^{\mathbb{N}}$. Pour $\omega = (a_k)_{k\in\mathbb{N}} \in \Omega$ on pose

$$s_n = s_n(\omega) = a_n \dots a_1 \in C(X, X), \quad q_n(x, \omega) = \prod_{k=0}^{n-1} q(s_k x, a_{k+1}).$$

On peut alors définir les probabilités de Markov Q_x sur Ω par

$$Q_x(A_1 \times \cdots \times A_n) = \int_{A_1\times\cdots\times A_n} q_n(x, \omega)\, d\mu^{\otimes n}(\omega).$$

On prendra garde au fait que les Q_x ne sont pas les probabilités canoniques, désignées par la même notation, dans la théorie des chaînes de Markov [7, 33]. Ici, les probabilités

Q_x peuvent être comparées aisément lorsque x varie. Pour fixer le cadre de travail, on donne alors les définitions suivantes

Définition 2.6. On appellera système Markovien un couple $(X, q \otimes \mu)$ où X est un espace métrique compact, μ une mesure de Radon positive sur $C(X, X)$, et q une fonction continue positive sur $X \times S_\mu$ tels que

a) $\int q(x, a)\, d\mu(a) = 1$ pour tout $x \in X$ et $\hat{q}(a) = \inf_{x \in X} q(x, a) > 0$ sur S_μ;

b) Q_x dépend continûment de $x \in X$ en variation totale, c'est-à-dire

$$\lim_{x' \to x} \|Q_{x'} - Q_x\| = 0.$$

Définition 2.7. On dira que le système Markovien $(X, q \otimes \mu)$ est transitif si les seules solutions continues de l'équation $h(x) = Qh(x) = \int h(ax) q(x, a)\, d\mu(a)$ sont les constantes.

Si le système $(X, q \otimes \mu)$ est transitif, on voit facilement que, si π est une probabilité Q-invariante sur X, la mesure $Q_\pi = \int Q_x\, d\pi(x)$ est invariante par le décalage θ sur Ω, indépendante de π et ergodique.

Un exemple bien connu de système Markovien transitif est donné par la construction d'une mesure de Gibbs sur un espace produit $A^{\mathbb{Z}}$ où A est un alphabet fini. On note alors θ le décalage sur $A^{\mathbb{Z}}$, et l'on se donne une fonction Höldérienne $f(\omega)$ sur $A^{\mathbb{Z}}$ (le potentiel). D'après [44], f s'écrit de manière unique sous la forme

$$f = h + \varphi \circ \theta - \varphi + c,$$

où $c \in \mathbb{R}$, les fonctions φ et h sont Höldériennes, h dépend seulement des coordonnées négatives de $\omega \in A^{\mathbb{Z}}$ et satisfait la condition de normalisation $\sum_{a \in A} e^{h(xa)} = 1$ où $x \in A^{[0,-\infty]} = X$ et xa désigne l'élément de $A^{[0,-\infty]}$ obtenu par juxtaposition de la "lettre" $a \in A$ au "mot infini" $x \in A^{[0,-\infty]}$. On note Q l'opérateur de transfert sur X défini par

$$Q\varphi(x) = \sum_{a \in A} e^{h(xa)} \varphi(xa).$$

Alors Q admet une unique mesure invariante π, et la mesure de Gibbs sur $A^{\mathbb{Z}}$ est l'unique mesure θ-invariante de projection π sur $A^{[0,-\infty]}$. Dans ces conditions, Γ est le semi-groupe de $C(X, X)$ engendré par les transformations $x \to xa$, μ est la mesure de dénombrement sur $A \subset C(X, X)$, $S_\mu = A$, et la fonction $q(x, a)$ vaut $e^{h(xa)}$. Alors, avec les notations précédentes, Q_x est une probabilité sur $\Omega = A^{[1,\infty]}$ qui s'interprète comme la loi conditionnelle de $\omega \in A^{[1,\infty]}$, sachant $x \in A^{[0,-\infty]}$ (cf. [12]). En particulier, les probabilités Q_x et Q_y sont équivalentes, la dérivée de Radon–Nikodym correspondante satisfait l'équation

$$\log \frac{dQ_x}{dQ_y}(\omega) = \sum_{k=0}^{\infty} \big[h(s_k x) - h(s_k y)\big].$$

Elle est donc Höldérienne, ce qui montre que la condition de continuité en variation totale de la Définition 2.6 est vérifiée par les probabilités Q_x $(x \in X)$.

Le cas particulier de la situation générale qui motive ce travail est donné par $X = P(V)$, Γ un semi-groupe d'application projectives, μ est une probabilité sur le groupe projectif de $P(V)$, et $q(x, g)$, où $g \in \mathrm{GL}(V)$ est défini par la normalisation de la fonction $(g, x) \mapsto \|gx\|^s$ considérée dans le Théorème 2.2. Le cas où le corps réel serait remplacé par un corps local ultramétrique conduirait à une situation formellement proche du modèle symbolique précédent, mais la dynamique de $\Gamma_{\alpha(\mu)}$ sur $P(V)$ mettrait en jeu, comme dans le cas réel, les propriétés d'attraction et de répulsion du semi-groupe $\Gamma_{\alpha(\mu)} \subset \mathrm{PGL}(V)$.

La chaîne de Markov Q d'espace d'états X jouera seulement un rôle auxiliaire dans les théorèmes généraux d'exposants, tandis que le système (Ω, θ, Q_π) et le semi-groupe $\Gamma_{\alpha(\mu)}$ défini ci-dessous interviendront de manière essentielle. Dans cette situation générale, le Théorème 2.8 ci-dessous et son Corollaire précisent le spectre de Lyapounov d'un produit de matrices stationnaires gouverné par une mesure du type de Gibbs.

Étant donné un système Markovien transitif $(X, q \otimes \mu)$, soit V un espace vectoriel normé de dimension finie et α une application Borélienne de S_μ dans le groupe linéaire $\mathrm{GL}(V)$. On considère alors le produit de matrices Q_π-stationnaire $S_n(\omega) = \alpha(a_n) \ldots \alpha(a_1)$ et l'on suppose que les deux intégrales $\int \log \|\alpha(a)\| q(a)\, d\mu(a)$ et $\int \log \|\alpha(a)^{-1}\| q(a)\, d\mu(a)$ sont finies, où $q(a) = \sup_x q(x, a)$. Alors les deux premiers exposants caractéristiques de $S_n(\omega)$ notés γ_1^q, γ_2^q sont des constantes bien définies par $q \otimes \mu$ et α à cause de la condition d'intégrabilité et du fait que Q_π est indépendante de π. Rappelons que d'après le théorème ergodique multiplicatif (cf. [41])

$$\gamma_1^q = \lim_n \frac{1}{n} \int \log \|S_n(\omega)\|\, dQ_\pi(\omega),$$

$$\gamma_1^q + \gamma_2^q = \lim_n \frac{1}{n} \int \log \|S_n(\omega) \wedge S_n(\omega)\|\, dQ_\pi(\omega).$$

On notera $\Gamma_{\alpha(\mu)}$ le semi-groupe de $\mathrm{GL}(V)$ engendré par $S_{\alpha(\mu)}$ et δ la distance sur $P(V)$ associée à la norme.

Théorème 2.8. *Soit $(X, q \otimes \mu)$ un système Markovien transitif, α une application Borélienne de X dans le groupe linéaire* $\mathrm{GL}(V)$ *telle que le semi-groupe $\Gamma_{\alpha(\mu)}$ vérifie la condition* (i–p) *et que les intégrales*

$$\int \log \|\alpha(a)\| q(a)\, d\mu(a), \quad \int \log \|\alpha(a)^{-1}\| q(a)\, d\mu(a)$$

soient finies. Alors les exposants caractéristiques γ_1^q et γ_2^q du produit

$$S_n = \alpha(a_n) \ldots \alpha(a_1)$$

satisfont

$$\gamma_2^q - \gamma_1^q = \lim_n \frac{1}{n} \sup_{\substack{x\in X \\ v,v'\in P(V)}} \int \log \frac{\delta[S_n(\omega).v,\, S_n(\omega).v']}{\delta(v,\, v')}\, dQ_x(\omega) < 0.$$

Dans le cas particulier des mesures de Gibbs, le Théorème 2.8 s'énonce ainsi:

Corollaire 2.9. *Supposons que A soit un alphabet fini, que π soit une mesure de Gibbs sur $A^{\mathbb{N}}$ définie par un potentiel Höldérien, α une application de A dans* $\mathrm{GL}(d, \mathbb{R})$*, et que le semi-groupe de* $\mathrm{GL}(V)$ *engendré par les matrices $\alpha(a)$ $(a \in A)$ soit fortement irréductible et contienne une matrice possédant une valeur propre dominante simple, unique en module. Alors les exposants caractéristiques γ_1^π et γ_2^π du produit $S_n(\omega)$ satisfont l'inégalité $\gamma_1^\pi > \gamma_2^\pi$.*

Pour une version plus générale de cet énoncé on pourra consulter [5].

Afin d'énoncer un théorème de simplicité de spectre de Lyapounov, notons $\mathcal{F}$ l'espace des drapeaux complets de V, $e_1, e_2, \ldots, e_d$ une base de V, $\widehat{V}$ le sous-espace de $V \otimes \wedge^2 V \otimes \cdots \otimes \wedge^{d-1} V$ engendré par l'orbite sous $\mathrm{GL}(V)$ de

$$\hat{e} = e_1 \otimes (e_1 \wedge e_2) \otimes \cdots \otimes (e_1 \wedge \cdots \wedge e_{d-1}).$$

Nous identifions $\mathcal{F}$ à l'orbite de $\hat{e}$ sous $\mathrm{GL}(V)$ dans $P(\widehat{V})$ et nous notons δ la distance naturelle sur $\mathcal{F} \subset P(\widehat{V})$. Un élément de $\mathrm{GL}(V)$ est dit $\mathcal{F}$-proximal si ses valeurs propres sont distinctes en module. On dira ainsi qu'un semi-groupe $\Gamma \subset \mathrm{GL}(V)$ satisfait la condition ($\mathcal{F}$–i–p) s'il contient un élément $\mathcal{F}$-proximal et s'il agit de manière fortement irréductible sur $\widehat{V}$. On sait (cf. [20]) que si l'adhérence de Zariski de Γ contient $\mathrm{SL}(V)$ cette condition est réalisée. On a alors

Théorème 2.10. *Soit $X, q \otimes \mu)$ un système Markovien transitif, α une application Borélienne de X dans* $\mathrm{GL}(V)$ *telle que le semi-groupe $\Gamma_{\alpha(\mu)}$ vérifie la condition* ($\mathcal{F}$–i–p) *et que les intégrales*

$$\int \log \|\alpha(a)\| q(a)\, d\mu(a), \quad \int \log \|\alpha(a)^{-1}\| q(a)\, d\mu(a)$$

soient finies. Alors les exposants caractéristiques $\gamma_i^q (\gamma_1^q \geq \gamma_2^q \geq \cdots \geq \gamma_d^q)$ du produit $S_n(\omega)$ sont tous différents, et l'on a

$$\lim_n \frac{1}{n} \sup_{x,b,b'} \int \log \frac{\delta[S_n(\omega).b,\, S_n(\omega).b']}{\delta(b,\, b')}\, dQ_x(\omega) = \sup_{1\leq i<d} (\gamma_{i+1} - \gamma_i) < 0.$$

Corollaire 2.11. *Avec les notations et hypothèses du Corollaire* 2.9*, supposons que le semi-groupe $\Gamma_{\alpha(\mu)}$ vérifie la condition* ($\mathcal{F}$–i–p)*. Alors les exposants caractéristiques γ_i^q $(1 \leq i \leq q)$ du produit $S_n(\omega)$ satisfont $\gamma_1^q > \gamma_2^q > \cdots > \gamma_d^q$.*

Les Théorèmes 2.8, 2.10 et leurs Corollaires découlent directement du Théorème 7.6 et de ses Corollaires. Il en est de même du Théorème 2.2 ci-dessus qui découle du Théorème 8.8. L'extension au cas d'un paramètre multidimensionnel est donnée en Section 8 (Corollaires 8.10, 8.11).

Remarque 2.12. A l'exception de l'holomorphie de $(\lambda I - P^z)^{-1}$ sur le cercle $|\lambda| = k(s)$ pour $z = s + it$ fixé, $t \neq 0$, donnée par le Corollaire du Théorème 2.2, toutes les propriétés décrites dans ce paragraphe restent valables si V est un espace vectoriel sur $\mathbb{C}$ ou un corps local général, en remplaçant $P(V)$ par l'espace projectif correspondant au corps considéré. La méromorphie de $(\lambda I - P^z)^{-1}$ pour $|\lambda| > \rho^s(\varepsilon)$ et en particulier pour $|\lambda| = k(s)$ reste valable en général, mais $(\lambda I - P^z)^{-1}$ peut avoir des pôles simples sur $|\lambda| = k(s)$, si $t \neq 0$, ce qui correspond à l'existence de solutions non-triviales pour certaines équations cohomologiques sur $\Gamma \times P(V)$ (voir [11, 19, 28]).

La propriété du corps des réels qui joue un rôle ici (cf. Théorème 3.19) est le fait que les éléments de norme 1 de ce corps forment un ensemble fini, donc algébrique, ce qui n'est pas le cas des corps locaux autres que $\mathbb{R}$. La simplicité du spectre de Lyapounov de μ sous la condition que l'adhérence de Zariski de $\Gamma_\mu \subset \mathrm{GL}(d, \mathbb{R})$ contienne $\mathrm{SL}(d, \mathbb{R})$ dépend aussi de cette propriété à travers la condition ($\mathcal{F}$–i–p) [20]. Pour les énoncés relatifs aux exposants caractéristiques dans le cas d'un corps local général et d'un groupe semi-simple G, on pourra consulter [27].

3 Proximalité, μ-proximalité et propriétés algébriques des sous-groupes de GL(V)

On rappelle ci-dessous, en les complétant, quelques propriétés utiles pour l'étude de l'action d'un semi-groupe $\Gamma \subset \mathrm{GL}(V)$ sur $P(V)$ en se basant sur [22, 19, 11]. Pour un exposé d'ensemble récent de ce sujet on pourra se reporter à [15].

3.1 Proximalité sur l'espace projectif

Définition 3.1. On dit que le semi-groupe Γ agit de manière proximale sur l'espace métrique (X, d) si pour tous $x, y \in X$ il existe une suite $\gamma_n \in \Gamma$ telle que $\lim_n \delta(\gamma_n.x, \gamma_n.y) = 0$.

Définition 3.2. On dit que le semi-groupe Γ agit de manière fortement proximale sur l'espace métrique (X, d) si pour toute probabilité ν sur X, il existe une suite $\gamma_n \in V$ avec $\lim_n \gamma_n.\nu = \delta_z$, où $z \in X$.

D'après [39, p. 196], la proximalité de Γ sur X et l'existence, pour tout $x \in X$, d'un voisinage de x pouvant être contracté par Γ en un point, impliquent la proximalité forte de Γ sur X. Pour $X = P(V)$ et $\Gamma \subset \mathrm{GL}(V)$, une propriété plus précise sera montrée plus loin (Proposition 3.13).

Définition 3.3. Un semi-groupe Γ de $\mathrm{GL}(V)$ sera dit irréductible (resp., fortement irréductible) s'il ne laisse pas de sous-espace (resp., de réunion finie de sous-espaces) invariant.

Définition 3.4. On note $\Gamma_1 \subset \operatorname{End} V$ l'ensemble des endomorphismes u tels qu'il existe une suite $\gamma_n \in \Gamma$ avec $u = \lim_n \gamma_n / \|\gamma_n\|$ et $\operatorname{rang} u = 1$.

Définition 3.5. Un élément $g \in \mathrm{GL}(V)$ est dit proximal si ses valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_d$ satisfont l'inégalité $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_d|$.

Pour un élément proximal g, la suite $g^{2k}/\|g^{2k}\|$ converge vers $u \in \operatorname{End} V$ avec $\operatorname{rang} u = 1$ et $u^2 \neq 0$. On note $\Gamma_0 \subset \Gamma$ l'ensemble des éléments proximaux de Γ. Pour $u \in \Gamma_1$, on note $p^+(u)$ le point de $P(V)$ défini par la droite $\operatorname{Im} u$. Pour $g \in \Gamma_0$ on note aussi $p^+(g)$ le point de $P(V)$ défini par le vecteur propre dominant de g, et $p^-(g) \in P(V^*)$ l'hyperplan défini par le vecteur propre dominant de la transposée g^t opérant dans V^*.

Le théorème suivant découle alors des Théorèmes 2.3 et 2.9 de [19].

Théorème 3.6. *Soit* Γ *un semi-groupe de* $\mathrm{GL}(V)$. *Alors les conditions suivantes sont équivalentes:*

a) Γ *est irréductible et proximal sur* $P(V)$*;*

b) Γ *est fortement irréductible, et* $\Gamma_1 \neq \phi$*;*

c) Γ *est fortement irréductible, et* $\Gamma_0 \neq \phi$.

Définition 3.7. On dira dans la suite que Γ vérifie la condition (i–p) s'il vérifie l'une des conditions équivalentes du Théorème 3.6.

On verra plus loin (Corollaire 3.15) que ces conditions entraînent la proximalité forte de Γ sur $P(V)$.

Définition 3.8. On note $\mathrm{Zc}(\Gamma)$ l'adhérence algébrique de Γ, c'est-à-dire l'ensemble des zéros communs à tous les polynômes à coefficients réels, nuls sur Γ (cf. [40]).

Le théorème suivant découle du Corollaire III-4 de [26] (cf. aussi [20, 43]). Il donne une vaste classe de semi-groupes satisfaisant la condition (i–p). Il suffit en effet que $\mathrm{Zc}(\Gamma) \supset \mathrm{SL}(V)$, et c'est "génériquement" le cas si Γ est engendré par deux éléments. Ici "génériquement" signifie presque partout par rapport à la mesure de Haar sur $\mathrm{GL}(V) \times \mathrm{GL}(V)$.

Théorème 3.9. Γ *vérifie l'une des conditions du Théorème* 3.6 *si et seulement si* $\mathrm{Zc}(\Gamma)$ *la vérifie*.

Observons en particulier que, sous la condition d'irréductibilité forte, l'existence d'un élément proximal dans le semi-groupe Γ est assurée par l'existence d'un élément proximal dans le groupe engendré par Γ.

Définition 3.10 ([19, 22]). Supposons que Γ vérifie (i–p). On appelle ensemble limite de Γ, la partie $L(\Gamma)$ de $P(V)$ formée des points correspondants aux droites $\operatorname{Im} u$ de V, avec $u \in \Gamma_1$.

Proposition 3.11 ([19]). *Supposons que* Γ *vérifie* (i–p) . *Alors l'ensemble* $L(\Gamma)$ *est le plus petit fermé* Γ*-invariant de* $P(V)$*, et l'action de* Γ *sur* $L(\Gamma)$ *est donc minimale. De plus, on a* $\overline{p^+(\Gamma_0)} = L(\Gamma)$*, et* $L(\Gamma)$ *n'est pas contenu dans une réunion finie de sous-espaces projectifs.*

Il en découle que si Γ vérifie la condition (i–p) , un fermé invariant contient toujours $L(\Gamma)$.

Le lemme suivant montre le rôle de la propriété d'irréductibilité dans la condition (i–p) .

Lemme 3.12. *Supposons que* Γ *vérifie la condition* (i–p) *et soient* $\Delta_1, \Delta_2, \ldots, \Delta_r$ *des sous-espaces de* $P(V)$*. Alors il existe* $g \in \Gamma_0$ *tel que l'hyperplan* $p^-(g)$ *ne contienne aucun des sous-espaces* Δ_i*.*

Preuve. Le semi-groupe $\Gamma^t \subset \mathrm{GL}(V^*)$ satisfait aussi la condition (i–p) et l'on peut donc considérer son ensemble limite $L(\Gamma^t) \subset P(V^*)$. D'après la Proposition 3.11, on a $\overline{p^-(\Gamma_0^t)} = L(\Gamma^t)$. La condition qu'un hyperplan contienne Δ_i définit un sous-espace de V^*. Si $p^-(g)$ contient l'un des Δ_i, pour tout $g \in \Gamma_0$, alors par densité, tout élément de $L(\Gamma^t)$ contiendrait l'un des Δ_i. Mais l'invariance de $L(\Gamma^t)$ sous Γ^t implique alors que Γ^t laisse invariante une réunion finie de sous-espaces de $P(V^*)$, ce qui contredit la condition (i–p) . □

Proposition 3.13. *Supposons que* Γ *vérifie la condition* (i–p) . *Alors on peut trouver des éléments* $a_1, a_2, \ldots, a_d$ *de* Γ_0 *et une sous-suite d'entiers* $N' \subset \mathbb{N}$ *tels que, pour tout* y *de* $P(V)$*, la suite* $a_d^{n'} \ldots a_1^{n'}.y$ *converge vers* $p^+(a_d)$ *lorsque* $n' \in N'$ *tend vers l'infini.*

La preuve découle du lemme suivant appliqué pour $k = d = \dim V$.

Lemme 3.14. *Soit* $k \in \{1, 2, \ldots d\}$*. Alors il existe* $a_1, a_2, \ldots, a_k \in \Gamma_0$*, des sous-espaces* Δ_j *de codimension* k *au moins* $(1 \leq j \leq 2^{k-1})$ *et une suite d'entiers* $n' \in N_k \subset \mathbb{N}$ *tel que*

$$\lim_{n'} a_k^{n'} \ldots a_1^{n'}.y = p^+(a_k) \quad \forall\, y \notin \bigcup_{1 \leq j \leq 2^{k-1}} \Delta_j .$$

Preuve. On procède par récurrence sur k. Pour $k = 1$, on peut trouver $a_1 \in \Gamma_0$ tel que si $y \notin \Delta_1 = p^-(a_1)$, on a $\lim_n a_1^n.y = p^+(a_1)$. C'est en effet le contenu de la condition c du Théorème 3.6.

Supposons l'assertion du lemme vérifiée au rang k et considérons les restrictions $\gamma_{n'}^j$ de $\gamma_{n'} = a_k^{n'} \ldots a_1^{n'}$ aux Δ_j de $P(V_j)$ $(1 \leq j \leq 2^{k-1})$. On note encore $\gamma_{n'}^j$ une application linéaire de V_j dans V correspondant à $\gamma_{n'}^j$ et normalisée par $\|\gamma_{n'}^j\| = 1$. En extrayant une sous-suite de N_k on obtient une partie $N_{k+1} \subset N_k$ telle que $\gamma_{n'}^j$ converge, si $n' \in N_{k+1}$, vers une application linéaire u_j de V_j dans V. Posons $V_j' = \mathrm{Ker}\, u_j \neq V_j$, $\Delta_j' = P(V_j')$. Alors $u_j(V_j \backslash V_j') = u_j(V_j) \backslash \{0\}$, et la suite d'applications

projectives $\gamma_{n'}^j$ converge vers l'application quasi-projective associée à u_j, en dehors de Δ'_j. Cette application sera encore notée u_j. De plus Δ'_j est de codimension 1 au moins dans Δ_j, et $u_j(\Delta_j \setminus \Delta'_j)$ est un sous-espace projectif $D_j \subset P(V)$. Puisque Γ vérifie (i–p), le Lemme 3.12 permet de trouver $a_{k+1} \in \Gamma_0$ tel que $p^-(a_{k+1})$ ne contienne aucun des sous-espaces D_j et $p^+(a_k)$. Alors $D_j \cap p^-(a_{k+1})$ est de codimension 1 dans D_j, et $u_j^{-1}(D_j \cap p^-(a_{k+1})) = \Delta''_j$ est aussi de codimension 1 dans Δ_j. Considérons alors la suite $a_{k+1}^{n'} \dots a_1^{n'}.y$ pour $n' \in N_{k+1}$ et $y \notin \bigcup_{1 \le j \le 2^{k-1}} (\Delta'_j \cup \Delta''_j)$. Alors $\gamma_{n'}^j.y$ converge vers un point n'appartenant pas à $p^-(a_{k+1})$. Il en découle $\lim_n a_{k+1}^{n'}.y = p^+(a_{k+1})$. Les nouveaux Δ_j pour $1 \le j \le 2^k$ sont les Δ'_j et Δ''_j pour $1 \le j \le 2^{k-1}$, puisque, par construction, ils sont de codimension $k+1$, au moins. □

Corollaire 3.15. *Si Γ vérifie la condition* (i–p)*, alors Γ est fortement proximal sur $P(V)$.*

Preuve. On utilise la suite $g_{n'} = a_d^{n'} \dots a_1^{n'}$ donnée par la proposition. Soit ν une probabilité sur $P(V)$, et $\varphi \in C([P(V)])$. Alors $g_{n'}.\nu(\varphi) = \int \varphi(g_{n'}.x)\, d\nu(x)$. Puisque pour tout $x \in P(V)$, $\lim_{n'} g_{n'}.x = p^+(a_d)$, on a, par convergence dominée, $\lim_{n'} g_{n'}.\nu(\varphi) = \varphi[p^+(a_d)]$, et $\lim_{n'} g_{n'}.\nu = \delta_{p^+(a_d)}$. □

Remarque 3.16. Pour une preuve différente du corollaire on pourra consulter [18] ou [39, p. 205]. Ici la suite $g_{n'}$ vaut simultanément pour toute les probabilités.

Rappelons [18] que si μ est une probabilité sur Γ avec $\Gamma_\mu = \Gamma$, et si Y est un Γ-espace compact, alors Y est dit μ-proximal si pour toute probabilité ν sur Y avec $\mu * \nu = \nu$ la suite $\gamma_1 \dots \gamma_n.\nu$ converge presque partout vers une mesure de Dirac. Dans le cas $Y = P(V)$, la condition (i–p) implique la μ-proximalité et donc l'unicité de ν. Ici, la propriété suivante interviendra plus loin :

Proposition 3.17 ([24]). *Soit μ une probabilité sur* $\mathrm{GL}(V)$*. On suppose que Γ_μ satisfait la condition* (i–p)*. Alors l'équation $\mu * \nu = \nu$, où ν est une probabilité sur $P(V)$, a une unique solution. De plus le support de ν est égal à $L(\Gamma_\mu)$, la mesure ν ne charge pas de sous-espace projectif, et, pour tout $\varphi \in C[P(V)]$, la suite $(\check{\mu}^n * \varphi)(x) = \int \varphi(g.x)\, d\mu^n(g)$ converge uniformément vers $\nu(\varphi)$.*

Lemme 3.18. *Si le semi-groupe Γ vérifie la condition* (i–p)*, alors $L(\Gamma)$ n'est pas contenu dans une réunion dénombrable de sous-espaces projectifs. De plus Γ opère de manière fortement irréductible sur l'espace vectoriel complexifié $V \otimes \mathbb{C}$.*

Preuve. Soit ν une probabilité sur Γ avec $\mu * \nu = \nu$, et supposons en défaut la première assertion. Alors puisque le support de ν est égal à $L(\Gamma)$, ν charge au moins l'un des sous-espaces considérés, ce qui contredit la Proposition 3.17. Si Γ n'est pas fortement irréductible sur $V \otimes \mathbb{C}$, on peut trouver un sous-groupe $\Gamma' \subset \mathrm{Zc}(\Gamma)$ d'indice fini, et un sous-espace complexe $W \subset V \otimes \mathbb{C}$ qui est Γ'-invariant. Soit $\overline{W}$ le sous-espace conjugué de W; alors, puisque Γ' vérifie aussi (i–p), on a $W \cap \overline{W} = \{0\}$, $W + \overline{W} =$

$V \otimes \mathbb{C}$, et donc $V \otimes \mathbb{C} = W \oplus \overline{W}$. Soit alors $g \in \Gamma'$ proximal et $v \in V$ dominant avec $gv = \lambda v$, $\lambda \in \mathbb{R}$. Ecrivant $v = w + \bar{w}$ avec $w \in W$, $\bar{w} \in \overline{W}$, on en déduit $gw = \lambda w$, $g\bar{w} = \lambda \bar{w}$, d'où w colinéaire à $\bar{w}$, ce qui est impossible. D'où la deuxième assertion. □

Pour être complet, donnons ici la preuve du résultat élémentaire suivant de [24, Proposition 3], qui suffira pour résoudre plusieurs questions d'apériodicité en Section 8.

Théorème 3.19. *Soit S un fermé de* $\mathrm{GL}(V)$ *qui engendre le semi-groupe* Γ. *On suppose que* Γ *vérifie la condition* (i–p), *que φ est une fonction continue sur $L(\Gamma)$, que $t, \alpha \in \mathbb{R}$, et que*

$$e^{i\alpha}\varphi(x) = \|gx\|^{it}\varphi(g.x) \quad \forall\, g \in S,\ x \in L(\Gamma).$$

Alors $t = 0$, $e^{i\alpha} = 1$, $\varphi = cte$ sur $L(\Gamma)$.

Preuve. Soit $\widetilde{L}(\Gamma)$ l'ensemble des points v de V dont la projection $\bar{v} \in P(V)$ appartient à $L(\Gamma)$. Considérons la fonction $\psi(v)$ définie sur $\widetilde{L}(\Gamma)$ par $\psi(v) = \|v\|^{it}\varphi(\bar{v})$. Alors ψ est continue, et la relation satisfaite par φ s'écrit

$$\psi(gv) = e^{i\alpha}\psi(v) \quad \forall\, g \in S,\ x \in \widetilde{L}(\Gamma).$$

Supposons $t \neq 0$ et notons $\rho = e^{\frac{2\pi}{|t|}}$. Alors $\psi(\rho^k v) = \psi(v)$, et la condition $\psi(\lambda v) = \psi(v)$ pour v fixé et $\lambda \in \mathbb{R}_+$ équivaut à $\lambda \in \rho^{\mathbb{Z}}$. Soit c une valeur de ψ, et $L_c = \psi^{-1}(c) \subset \widetilde{L}(\Gamma)$. Alors L_c est invariant par le groupe, noté $\rho^{\mathbb{Z}}$, des homothéties de rapport ρ^k ($k \in \mathbb{Z}$). De plus $gL_c = L_{ce^{i\alpha}}$ si $g \in S$. Enfin, si $x \in L_c$, la droite $\mathbb{R}x$ coupe L_c suivant l'ensemble dénombrable $\pm\rho^{\mathbb{Z}}a$ avec $a \in L_c$.

Soit $g_k \in \Gamma$ et $n_k \in \mathbb{Z}$ avec $\rho^{n_k} < \|g_k\| \leq \rho^{n_k+1}$. Alors $1 < \left\| \dfrac{g_k}{\rho^{n_k}} \right\| \leq \rho$, et l'on peut donc supposer en extrayant une sous-suite, que $\dfrac{g_k}{\rho^{n_k}}$ converge vers $u \in \mathrm{End}\, V$ avec $1 \leq \|u\| \leq \rho$, et donc $u \neq 0$. Alors u envoie V sur le sous-espace $\mathrm{Im}\, u$, et l'on a $u(\mathrm{Ker}\, u) = \{0\}$. D'après la condition (i–p) on peut supposer $\dim \mathrm{Im}\, u = 1$, $\dim \mathrm{Ker}\, u = d - 1$. Si $g_k \in S^{p_k}$ avec $p_k \in \mathbb{N}$, on a $\psi(g_k x) = e^{ip_k\alpha}\psi(x)$, donc $u(L_c) \subset \{0\} \cup L_{e^{i\theta}c}$ avec $e^{i\theta} = \lim_k e^{ip_k\alpha}$. On a donc

$$u(L_c) = [\mathrm{Im}\, u \cap (L_{e^{i\theta}c})] \cup \{0\}) = \pm\rho^{\mathbb{Z}}a \cup \{0\},$$

avec $a \neq 0$, d'après ce qui précède. Donc

$$L_c \cup \{0\} \subset (\pm\rho^{\mathbb{Z}})b \cup \{0\} + \mathrm{Ker}\, u,$$

avec $u(b) = a$ et $b \in L_c$. En d'autres termes, $L_c \cup \{0\}$ se projette dans l'espace $V/\mathrm{Ker}\, u$ sur un ensemble dénombrable invariant par le groupe $\rho^{\mathbb{Z}}$.

Si H est plus généralement un sous-espace vectoriel, montrons que cette propriété de V/H est stable par intersection. Une intersection de sous-espaces est toujours une intersection finie et l'on peut donc supposer $H = \bigcap_{i=1}^{r} H_i$ où $L_c \cup \{0\}$ se projette dans

V/H_i suivant un ensemble dénombrable stable par $\rho^{\mathbb{Z}}$. Comme l'espace quotient de V par $\bigcap_{i=1}^r H_i$ s'identifie à la diagonale du produit $\prod_{i=1}^r V/H_i$, la projection de L_c dans V/H est contenue dans le produit de r ensembles dénombrables ; elle est de plus $\rho^{\mathbb{Z}}$-invariante. Soit alors H le plus petit sous-espace de V possédant la propriété pour L_c. Par définition de L_c on a $\lambda L_c = L_{|\lambda|^{it}c}$ pour tous $t \in \mathbb{R}$ et $\lambda \in \mathbb{R}$. Le sous-espace H est donc aussi le plus petit sous-espace possédant la propriété pour $L_{e^{i\alpha}c}$. D'où $g(H) = H$ pour tout g de S, et H est Γ-invariant. On ne peut avoir $H = \{0\}$ car alors L_c, donc $L(\Gamma)$, serait dénombrable, ce qui est exclu d'après le Lemme 3.18. On ne peut avoir $H \neq \{0\}$ car cela contredirait l'irréductibilité de Γ. On a donc $t = 0, \varphi(g.x) = e^{i\alpha}\varphi(x)$ pour tous $x \in L(\Gamma),\ g \in S$. Avec les notations de la Proposition 3.17 on a alors

$$\check{\mu}^n * \varphi = e^{in\alpha}\varphi, \quad \nu(\varphi) = \lim_n \check{\mu}^n * \varphi = \lim_n e^{in\alpha}\varphi,$$

donc $e^{i\alpha} = 1$.

Alors $\varphi(g.x) = \varphi(x)$ pour tout $g \in \Gamma$ et $x \in L(\Gamma)$. La Proposition 3.11 donne finalement $\varphi = cte$ par minimalité de $L(\Gamma)$. □

3.2 Proximalité sur l'espace des drapeaux

On va maintenant étendre les résultats précédents au cas d'une action sur les drapeaux.

On notera $(z) = (z_1, \ldots, z_d)$ un élément de $\mathbb{C}^d$.

Définition 3.20. Supposons qu'un groupe G opère continûment sur un espace métrique compact X. On dira qu'une fonction continue $\sigma(g, x)$ à valeurs complexes sur $G \times X$ est un cocycle si l'on a

$$\sigma(gh, x) = \sigma(g, h.x)\sigma(h, x) \quad \forall g, h \in G,\ x \in X.$$

Rappelons qu'un drapeau complet de $P(V)$ est une suite emboîtée de sous-espaces stricts $V_1 \subset \cdots \subset V_{d-1}$ telle que $\dim V_i = i$. Un tel drapeau est repéré par une suite de multivecteurs

$$b = (v_1, v_1 \wedge v_2, \ldots, v_1 \wedge \cdots \wedge v_{d-1})$$

que l'on peut toujours supposés normés ($\|v_1 \wedge \cdots \wedge v_k\| = 1$). On note $\mathcal{F}$ l'espace des drapeaux complets. On peut réaliser $\mathcal{F}$ comme l'unique orbite compacte de $G = \mathrm{GL}(V)$ dans $P(\widehat{V})$ où $\widehat{V}$ est l'espace de la représentation irréductible suivante (cf. [40, pp. 83, 84], [43]). Si $e_1, e_2, \ldots, , e_d$ est une base de V, $\widehat{V}$ est le sous-espace G-invariant de $V \otimes \wedge^2 V \otimes \cdots \otimes \wedge^{d-1} V$ engendré par

$$\hat{e} = e_1 \otimes (e_1 \wedge e_2) \otimes \cdots \otimes (e_1 \wedge \cdots \wedge e_{d-1}).$$

Si l'on désigne par $\wedge^k$ (resp., $\hat{g}$) l'extension naturelle de $g \in G$ à $\wedge^k V$ (resp., $\widehat{V}$), $\mathcal{F}$ s'identifie à $\widehat{G}.\hat{e}$. Ayant identifié $\wedge^d V$ à $\mathbb{R}$, on note pour $b \in \widehat{V}, b' \in \widehat{V}$,

$b_k \in \wedge^k V, b'_{d-k} \in \wedge^{d-k} V$

$$\langle b_k, b'_{d-k}\rangle = b_k \wedge b'_{d-k}, \quad \langle b, b'\rangle = \prod_1^{d-1} \langle b_k, b'_{d-k}\rangle.$$

Ainsi $\widehat{V}$ est en dualité avec lui-même, et pour $g \in G$, $\hat{g}^t$ peut être considéré comme une application de $\widehat{V}$ dans lui-même induisant sur $\mathcal{F}$ une application projective encore notée $\hat{g}^t$. Si l'on note δ_k la distance sur $P(\wedge^k V)$ associée à une norme sur $\wedge^k V$, on peut considérer la distance δ sur $\mathcal{F}$ donnée par $\delta(b, b') = \sum_{k=1}^{d-1} \delta_k(b_k, b'_k)$, où $b = (b_1, \ldots, b_{d-1}), b' = (b'_1, \ldots, b'_{d-1})$, et les b_k, b'_k correspondent à des vecteurs unitaires de $\wedge^k V$. Donnons alors, comme ci-dessus, les définitions:

Définition 3.21. Un élément $g \in \mathrm{GL}(V)$ sera dit $\mathcal{F}$-proximal si ses valeurs propres satisfont $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_d|$.

Définition 3.22. On dira qu'un semi-groupe $\Gamma \subset \mathrm{GL}(V)$ satisfait la condition ($\mathcal{F}$–i–p) si $\widehat{\Gamma} \subset \mathrm{GL}(\widehat{V})$ est irréductible et si $\widehat{\Gamma}$ opère de manière proximale sur $\mathcal{F} \subset P(\widehat{V})$.

On note $\widehat{\Gamma}_0$ l'ensemble des éléments $\mathcal{F}$-proximaux de Γ, et l'on observe que si $g \in \widehat{\Gamma}_0$, alors $p^+(\hat{g}) \in \mathcal{F}$ et $p^-(\hat{g}) \in \mathcal{F}$. De plus, si $g \in \widehat{\Gamma}_0$, alors pour tout $x \in \mathcal{F}$ tel que $\langle x, p^-(\hat{g})\rangle \neq 0$, on a

$$\lim_{n\to+\infty} \hat{g}^n.x = p^+(\hat{g}).$$

On sait d'après [20, 43], que si $\mathrm{Zc}(\Gamma) \supset \mathrm{SL}(V)$, alors $\widehat{\Gamma}_0 \neq \phi$, donc Γ satisfait la condition ($\mathcal{F}$–i–p). Aussi, comme en [22], la condition ($\mathcal{F}$–i–p) équivaut à $\widehat{\Gamma}_0 \neq \phi$ et $\widehat{\Gamma}$ fortement irréductible. On peut alors définir l'ensemble limite $L_{\mathcal{F}}(\Gamma)$ de Γ sur $\mathcal{F}$, comme plus haut, et l'on a

Proposition 3.23 ([19]). *Supposons que Γ vérifie la condition ($\mathcal{F}$–i–p). Alors $L_{\mathcal{F}}(\Gamma) = L(\widehat{\Gamma}) \subset \mathcal{F}$ est un fermé Γ-invariant non vide. De plus*

a) $\widehat{\Gamma}_0 \neq \phi$;

b) *Pour tout $b \in \mathcal{F}$, $\overline{\Gamma.b} \supset L_{\mathcal{F}}(\Gamma)$;*

c) $\overline{p^+(\widehat{\Gamma}_0)} = L_{\mathcal{F}}(\Gamma)$;

d) *Soit μ une probabilité sur $\mathrm{GL}(V)$ dont le support engendre Γ. Si la probabilité ν sur $\mathcal{F}$ satisfait $\mu * \nu = \nu$, alors ν est unique et de support égal à $L_{\mathcal{F}}(\Gamma)$.*

Remarque 3.24. A cause de l'irréductibilité dans $\widehat{V}$, la condition ($\mathcal{F}$–i–p) est plus forte que la condition plus analytique suivante considérée en [25, Théorème 2.19]: Γ *est totalement irréductible et contient une suite contractante vis-à-vis de $\mathcal{F}$.*

Les cocycles sur $\mathrm{GL}(V) \times \mathcal{F}$ s'expriment à l'aide des cocycles fondamentaux

$$\sigma_k(g, b) = \|g(v_1 \wedge \cdots \wedge v_k)\| \quad (1 \leq k < d), \quad \sigma_d(g) = |\det g|,$$

où $b = (v_1, v_1 \wedge v_2, \ldots, v_1 \wedge \cdots \wedge v_{d-1})$. On note $\sigma^{(z)}(g, b) = \prod_{k=1}^{d} \sigma_k^{(z)}(g, b_k)$ pour $(z) = (z_1, \ldots, z_d) \in \mathbb{C}^d$. Plus précisément, on a, par transitivité de $\mathrm{GL}(V)$ sur $\mathcal{F}$ (cf. [16]).

Proposition 3.25. *Les fonctions $\sigma_k(g, b)$ $(1 \leq k \leq d)$ sont des cocycles positifs. Pour tout cocycle positif σ, il existe $(s) = (s_1, s_2, \ldots, s_d) \in \mathbb{R}^d$ et φ continue sur $\mathcal{F}$ avec*

$$\sigma(g, b) = \sigma_1^{s_1}(g, b)\sigma_2^{s_2}(g, b) \ldots \sigma_d^{s_d}(g, b)\frac{\varphi(g.b)}{\varphi(b)}.$$

Les cocycles $\mathrm{SO}(d)$-invariants (i.e., $\sigma(k, b) = 1, \ \forall k \in \mathrm{SO}(d)$) correspondent à $\varphi = 1$.

L'étude des propriétés spectrales des opérateurs $P^{(s)}$ (Section 8) introduit des équations cohomologiques sur $\mathrm{GL}(V) \times \mathcal{F}$. Le théorème suivant découle des arguments de la preuve du théorème principal de [4] et permet de les résoudre. Il sera utilisé dans les preuves des Corollaires 8.10 et 8.11.

Théorème 3.26. *Supposons que $S \subset \mathrm{GL}(V)$ engendre le semi-groupe Γ. Supposons que $\mathrm{Zc}(\Gamma) \supset \mathrm{SL}(V)$. Alors la relation*

$$\sigma^{(s)}(g, b) = k\frac{\varphi(g.b)}{\varphi(b)},$$

où $k \in \mathbb{C}^*$, $(s) \in \mathbb{C}^d$, $\varphi \in C(\mathcal{F})$ sont fixés et (g, b) décrit $S \times L_{\mathcal{F}}(\Gamma)$, équivaut à $|\det g|^{\Re s_d} = |k|$ sur S, $s_1 = \cdots = s_{d-1} = 0$ et $\varphi = cte$ sur $L_{\mathcal{F}}(\Gamma)$.

Proposition 3.27. *Supposons que le semi-groupe Γ de $\mathrm{GL}(V)$ vérifie la condition $(\mathcal{F}$–i–p$)$. Alors il existe $r = \dim \widehat{V}$ éléments a_k de Γ et une partie N' de $\mathbb{N}$ tels que pour tout b de $\mathcal{F}$, la suite $a_r^{n'} \ldots a_1^{n'}.b$ converge vers $p^+(a_r) \in \mathcal{F}$ lorsque n' tend vers l'infini. En particulier, Γ est fortement proximal sur $\mathcal{F}$.*

Preuve. On utilise la Proposition 3.17 en remplaçant V par $\widehat{V}$. □

Remarque 3.28. En reprenant la preuve de la Proposition 3.13, on peut voir que ici $N' = \mathbb{N}$. En utilisant les techniques de [25], on peut obtenir $r = 2$ mais ce fait ne sera pas utilisé ici.

La proposition suivante mentionnée en [24] permet de ramener l'estimation des normes d'endomorphismes à des calculs d'algèbre linéaire.

Proposition 3.29. *Soit m_p la mesure uniforme sur l'ensemble V_p des p-multivecteurs décomposables unitaires. La différence*

$$\int_{V_p} \log \|(\wedge^p u)x\| \, dm_p(x) - \log \| \wedge^p u\|$$

est bornée inférieurement sur End V.

Preuve. On peut décomposer $u \in \text{End } V$ sous la forme $u = kak'$ avec $k, k' \in \text{SO}(d)$ et $a = \text{diag}(a_1, a_2, \ldots, a_d)$ avec $a_1 \geq a_2 \geq \cdots \geq a_d \geq 0$.

On peut aussi écrire $x = k''E$ avec $E = e_1 \wedge e_2 \wedge \cdots \wedge e_p$ et $k'' \in \text{SO}(d)$, de sorte que

$$\begin{aligned}\int_{V_p} \log \| \wedge^p ux \| \, dm_p(x) &= \int_{\text{SO}(d)} \log \| \wedge^p ak'k''E \| \, dm(k'') \\ &= \int_{\text{SO}(d)} \log \| \wedge^p akE \| \, dm(k),\end{aligned}$$

où m est la mesure de Haar normalisée sur $\text{SO}(d)$. De plus, on a

$$\| \wedge^p akE \| \geq |\langle \wedge^p akE, E \rangle| \geq \| \wedge^p a \| \cdot |\langle kE, E \rangle|.$$

On a alors

$$\begin{aligned}\int_{V_p} \log \| \wedge^p ux \| \, dm_p(x) &\geq \log \| \wedge^p u \| + \int_{\text{SO}(d)} \log |\langle kE, E \rangle| \, dm(k) \\ &= \log \| \wedge^p u \| + \int_{V_p} \log |\langle x, E \rangle| \, dm_p(x).\end{aligned}$$

Il suffit donc de voir que l'intégrale au second membre est finie.

Or V_p est une sous-variété algébrique de la sphère unité de $\wedge^p V$ de dimension r, et m_p en est la mesure r-dimensionnelle normalisée. Comme l'application $x \to \langle x, E \rangle^2$ est polynômiale il existe un entier $N > 0$ et $c > 0$ tels que l'on ait

$$m_p \left\{ x : |\langle x, E \rangle|^2 \leq t \right\} \leq ct^N.$$

La mesure image σ de m_p par cette application est donc une probabilité ayant une densité f concentrée sur $[0, 1]$ et satisfaisant $f(t) \leq ct^{N/2}/t$.

On a donc

$$\int \log |\langle x, E \rangle| \, dm_p(x) = \int_0^1 \log t \, d\sigma(t) > -\infty$$

car l'intégrale $\displaystyle\int_0^1 t^{N/2} (\log t) \frac{dt}{t}$ est convergente pour $N > 0$. □

4 Les opérateurs de transfert P^s et leur normalisation

Soit μ une probabilité sur $G = \mathrm{GL}(V)$ et $s \geq 0$ tel que $\int \|g\|^s \, d\mu(g) < +\infty$. On note $k(s)$ le réel défini par

$$k(s) = \inf_n \left(\int \|g\|^s \, d\mu^n(g) \right)^{1/n} = \lim_n \left(\int \|g\|^s \, d\mu^n(g) \right)^{1/n}.$$

Le semi-groupe fermé Γ_μ engendré par le support de μ sera éventuellement noté Γ afin d'alléger les notations. On notera aussi $S_\mu^{\otimes n}$ l'espace produit de n copies de S_μ et $\mu^{\otimes n}$ la mesure produit sur $S_\mu^{\otimes n}$ associée à μ.

Pour $s \geq 0$, la sous-multiplicativité de $g \mapsto \|g\|$ entraîne que la suite $u_n(s) = \int \|g\|^s \, d\mu^n(g)$ satisfait $u_{m+n}(s) \leq u_m(s) u_n(s)$. Donc la limite de $u_n^{1/n}(s)$ existe et est égale à sa borne inférieure. L'existence de la limite implique que $\log k(s)$ est une fonction convexe comme $\dfrac{1}{n} \log u_n(s)$, dans l'intervalle où l'intégrale $\int \|g\|^s \, d\mu(g)$ est finie. On montre ici, en s'inspirant de [33], que la normalisation de P^s défini par $P^s\varphi(x) = \int \|gx\|^s \varphi(g.x) \, d\mu(g)$ à l'aide du facteur $k(s)$ conduit à des opérateurs Markoviens ayant de fortes propriétés d'ergodicité et vérifiant en particulier une condition d'absolue continuité à la Doeblin–Fortet (cf. [12]) dès que Γ_μ vérifie la condition (i–p) .

4.1 Construction d'une fonction propre pour l'opérateur P^s

Théorème 4.1. *Supposons que la probabilité μ sur $\mathrm{GL}(V)$ vérifie $\int \|g\|^s \, d\mu(g) < +\infty$ pour un $s > 0$, que $\Gamma = \Gamma_\mu$ soit irréductible et proximal sur $P(V)$ (condition (i–p)).*

Alors l'équation $P^s\varphi = k\varphi$ avec $k \in \mathbb{R}_+$ a une unique solution strictement positive normalisée $\varphi = e_s$. On a $k = k(s)$, et e_s est Höldérienne d'ordre $\bar{s} = \min\{1, s\}$.

Si une fonction complexe continue ψ vérifie $P^s\psi = k\psi$, alors on a $|k| \leq k(s)$, et si $|k| = k(s)$, ψ est proportionnelle à e_s et $k = k(s)$. Si, ψ continue positive, vérifie $P^s\psi \leq k(s)\psi$, et ne s'annule pas identiquement sur $L(\Gamma)$, alors ψ est proportionnelle à e_s.

Si une mesure de probabilité ν satisfait $P^s\nu = k\nu$, alors on a $k = k(s)$ et le support de ν contient $L(\Gamma)$.

Enfin, il existe $c > 0$ tel que l'on ait pour tout n et tout $x \in P(V)$

$$\frac{1}{c} k^n(s) \leq \int \|gx\|^s \, d\mu^n(g) \leq c k^n(s).$$

Remarque 4.2. a) On verra plus loin (Théorème 6.8), grâce à une méthode de martingale introduite en [18], que la mesure de probabilité $\nu = \nu^s$ est unique et de support égal à $L(\Gamma)$. Pour $s = 0$, ce fait est contenu dans la Proposition 3.17. Pour $s \neq 0$, on donnera une justification directe à la fin de ce paragraphe. Pour l'extension à $(s) \in \mathbb{R}_+^d$, on utilisera en Sections 5, 6 une méthode de martingale généralisant [18].

L'existence et l'unicité de φ et ν ne semblent cependant pas être des propriétés suffisantes pour l'étude des exposants caractéristiques associés et donc pour une étude approfondie des opérateurs P^s. Cette étude sera poursuivie en Section 7.

b) D'autre part une construction différente de φ a été donnée en [25] et reprise en [38]. Son utilisation montre que φ est au moins de classe $C^{s+\varepsilon}$ pour un certain $\varepsilon > 0$.

c) En général il existe des fonctions continues φ positives, égales à zéro sur $L(\Gamma)$ et des nombres k tels que $P^s\varphi = k\varphi$ et $0 < k < k(s)$. Cette situation ne se produit pas si $L(\Gamma) = P(V)$ comme le montre le Lemme 4.10 ci-dessous.

d) Un fait essentiel est l'uniformité en $n \in \mathbb{N}$, $x \in P(V)$ donnée par la dernière assertion. Une telle uniformité ne peut être valable sur $P(V)$, pour $s \leq 0$, en général comme on le montre plus loin dans la situation où $\check{\mu}$ est concentré sur les matrices positives de déterminant 1. On a alors, en effet, pour certains μ

$$\lim_n \inf_{\|x\|=1} \left[\int \|gx\|^{-d}\, d\mu^n(g)\right]^{1/n} < 1, \quad \lim_n \sup_{\|x\|=1} \left[\int \|gx\|^{-d}\, d\mu^n(g)\right]^{1/n} \geq 1.$$

La preuve du Théorème 4.1 découle des lemmes qui suivent.

Lemme 4.3. *Soit ν une probabilité sur $P(V)$ non portée par un sous-espace projectif. Alors il existe une constante $\varepsilon(s, \nu) > 0$ telle que pour tout g appartenant à* End V

$$\|g\|^s \leq \varepsilon(s, \nu) \int \|gv\|^s\, d\nu(v).$$

Preuve. Par homogénéité on peut supposer $\|g\| = 1$. Il suffit alors de voir que la fonction de g donnée par l'intégrale $\int \|gv\|^s\, d\nu(v)$ est bornée inférieurement par $\varepsilon > 0$ sur la sphère de End V définie par $\|g\| = 1$. Puisque s ≥ 0, la fonction $g \mapsto \int \|gv\|^s\, d\nu(v)$ est continue sur la sphère unité ; elle atteint donc sa borne inférieure en $A \neq 0$. Si l'on avait $\int \|Av\|^s\, d\nu(v) = 0$, on aurait $Av = 0$ pour tout v du support de ν. Alors ν serait portée par le sous-espace projectif associé au noyau de A, ce qui est exclu par hypothèse. Donc

$$\inf_{\|g\|=1} \int \|gv\|^s\, d\nu(v) = \int \|A\, v\|^s\, d\nu(v) > 0. \qquad \square$$

Lemme 4.4. *Soit ν une mesure de probabilité sur $P(V)$ et $k \in \mathbb{R}_+$ tels que $P^s\nu = k\nu$. Alors le support de ν est Γ_μ-invariant. De plus, ν n'est pas portée par un sous-espace projectif et l'on a*

$$\varepsilon(s, \nu)^{-1} \int \|g\|^s\, d\mu^n(g) \leq k^n \leq \int \|g\|^s\, d\mu^n(g).$$

En particulier, si ν existe on a $k = k(s)$.

Preuve. D'après l'équation $P^s\nu = k\nu$, le support S_ν de ν est Γ_μ-invariant. Il en est de même du sous-espace projectif engendré par S_ν. D'après l'irréductibilité de Γ_μ,

ce sous-espace est égal à $P(V)$. On peut alors appliquer le Lemme 4.3 au calcul de $\langle (P^s)^n \nu, 1\rangle$:

$$k^n \nu(1) = \langle (P^s)^n \nu, 1\rangle = \int\int \|gv\|^s \, d\nu(v)\, d\mu^n(g) \geq \varepsilon^{-1}(s,\nu) \int \|g\|^s \, d\mu^n(g).$$

De plus on a par définition de $\|g\|$

$$\int \|gv\|^s \, d\nu(v)\, d\mu^n(g) \leq \int \|g\|^s \, d\mu^n(g),$$

d'où l'inégalité annoncée. □

Lemme 4.5. *Il existe une mesure de probabilité ν sur $P(V)$ telle que $P^s\nu = k(s)\nu$, et l'on a l'inégalité suivante, avec $\varepsilon(s,\nu)$ donné au Lemme 4.3,*

$$\varepsilon(s,\nu)^{-1} \int \|g\|^s \, d\mu^n(g) \leq k^n(s) \leq \int \|g\|^s \, d\mu^n(g).$$

Preuve. Par la formule $\bar{P}^s\nu = P^s\nu / P^s\nu(1)$, le théorème de Schauder–Tychonoff appliqué à $\bar{P}^s$ et au convexe compact des probabilités sur $P(V)$ entraîne l'existence de $k > 0$ avec $P^s\nu = k\nu$. Le lemme précédent fournit $k = k(s)$ et l'inégalité annoncée. □

Lemme 4.6. *Soit $s \geq 0$ et $\bar{s} = \min\{s, 1\}$. Pour tous v, v' de $P(V)$, on a les inégalités*

$$\big|\|gv\|^s - \|gv'\|^s\big| \leq (s+1)\|g\|^s \delta^{\bar{s}}(v, v'),$$
$$\delta(g.v, g.v') \leq \frac{2}{\|gv\|}\|g(v - v')\|.$$

Preuve. Notons que pour deux réels $a, b \geq 0$, on a $|a^s - b^s| \leq |a-b|^s$ si $s \leq 1$, et $|a^s - b^s| \leq s|a-b|\max\{a,b\}^{s-1}$ si $s \geq 1$.

La première inégalité découle donc de

$$\big|\|gv\| - \|gv'\|\big| \leq \|g(v-v')\| \leq \|g\| \cdot \|v - v'\|.$$

On choisit des représentants de v, v' de norme 1, et l'on écrit

$$\delta(g.v, g.v') \leq \left\| \frac{gv}{\|gv\|} - \frac{gv'}{\|gv'\|} \right\|,$$
$$\|gv\| \cdot \|gv'\| \left(\frac{gv}{\|gv\|} - \frac{gv'}{\|gv'\|} \right) = gv(\|gv'\| - \|gv\|) + \|gv\|(gv - gv'),$$
$$\|gv\| \cdot \|gv'\| \delta(g.v, g.v') \leq 2\|gv\| \cdot \|g(v-v')\|,$$
$$\delta(g.v, g.v') \leq \frac{2}{\|gv'\|}\|g(v-v')\|.$$ □

Lemme 4.7. *La suite* $(P^s)^n 1(v)$ *satisfait, pour tous* v *et* v' *de* $P(V)$

$$\left|(P^s)^n 1(v) - (P^s)^n 1(v')\right| \leq (s+1)\left[\int \|g\|^s \, d\mu^n(g)\right] \delta(v, v')^{\bar{s}}$$

avec $\bar{s} = \min\{s, 1\}$.

Preuve. D'après la première inégalité du Lemme 4.6

$$\left| \|gv\|^s - \|gv'\|^s \right| \leq (s+1)\delta^{\bar{s}}(v, v')\|g\|^s.$$

Puisque $(P^s)^n 1(v) = \int \|gv\|^s \, d\mu^n(g)$, la relation voulue en découle par intégration. □

Lemme 4.8. *Il existe une fonction strictement positive* φ, *Höldérienne d'ordre* $\bar{s} = \min\{1, s\}$ *telle que* $P^s\varphi = k(s)\varphi$. *On a de plus*

$$|\varphi(v) - \varphi(v')| \leq c_s \delta^{\bar{s}}(v, v'),$$

où c_s *est une constante*.

Preuve. D'après les Lemmes 4.6 et 4.7, la suite de fonctions $(P^s)^n 1(v)/k^n(s)$ est équicontinue. Elle est bornée supérieurement car, d'après le Lemme 4.5

$$(P^s)^n 1(v) = \int \|gv\|^s \, d\mu^n(g) \leq \frac{1}{\varepsilon} k^n(s).$$

Le module de continuité commun est égal à $(s+1)\delta^{\bar{s}}(v, v')\varepsilon(s, \nu)$, d'après le Lemme 4.7. On considère alors la suite de fonctions $f_n(v) = \dfrac{1}{n}\sum_1^n \dfrac{(P^s)^m 1(v)}{k^m(s)}$. Le théorème d'Ascoli permet d'extraire de f_n une sous-suite convergente vers $\varphi \geq 0$ avec

$$|\varphi(v) - \varphi(v')| \leq (s+1)\delta^{\bar{s}}(v, v')\varepsilon(s, \nu).$$

A la limite, on a de plus $P^s\varphi = k(s)\varphi$ puisque $\dfrac{1}{n}\dfrac{(P^s)^{n+1}1(v)}{k^{n+1}(s)}$ tend vers 0 d'après le Lemme 4.5.

Supposons que φ s'annule en v_0. Alors l'équation $P^s\varphi = k(s)\varphi$ entraîne $\varphi\left(\overline{\Gamma_\mu . v_0}\right) = 0$. Si η est une probabilité de support égal à $\overline{\Gamma_\mu . v_0}$, on a alors $\eta(\varphi) = 0$. Puisque $\overline{\Gamma_\mu . v_0}$ contient $L(\Gamma)$ et donc (cf. Proposition 3.11) n'est pas contenu dans un sous-espace projectif propre, on a d'après le Lemme 4.3

$$\int \|gv\|^s \, d\eta(v) \geq \varepsilon(s, \eta)\|g\|^s.$$

Alors

$$\eta\left(\frac{(P^s)^m 1}{k^m(s)}\right) \geq \varepsilon(s, \eta)\frac{\int \|g\|^s \, d\mu^m(s)}{k^m(s)} \geq \varepsilon(s, \eta),$$

d'où $\eta(f_n) \geq \varepsilon(s,\eta)$ et $\eta(\varphi) \geq \varepsilon(s,\eta)$, ce qui contredit la relation $\eta(\varphi) = 0$. D'où $\varphi > 0$. □

Lemme 4.9. *Il existe $\varepsilon > 0$ tel que*

$$\int \|gv\|^s \, d\mu^n(g) \geq \varepsilon k^n(s) \quad \forall\, v \in P(V).$$

Preuve. D'après le Lemme 4.8, en fixant ε' tel que $\varepsilon' \leq \varphi \leq \dfrac{1}{\varepsilon'}$ on a

$$\varepsilon' k^n(s) \leq k^n(s)\varphi(v) = \int \|gv\|^s \varphi(g.v)\, d\mu^n(g) \leq \frac{1}{\varepsilon'} \int \|gv\|^s \, d\mu^n(g),$$

d'où la relation annoncée avec $\varepsilon = (\varepsilon')^2$. □

Lemme 4.10. *Soit ψ continue positive et non identiquement nulle sur $L(\Gamma)$ et $k > 0$ vérifiant $P^s\psi = k\psi$. Alors $k = k(s)$, et ψ est proportionnelle à la fonction φ du Lemme* 4.7 *qui est donc unique à un coefficient près. Si ψ est continue positive et satisfait $P^s\psi \leq k(s)\psi$, alors ψ est proportionnelle à φ sur $L(\Gamma)$. Si ψ est continue et satisfait $P^s\psi = k(s)\psi$, alors ψ est proportionnelle à φ.*

Preuve. L'ensemble Z des zéros de ψ est un fermé Γ-invariant d'après l'équation $P^s\psi = k\psi$. D'après la Proposition 3.11, si $Z \neq \phi$, on a $Z \supset L(\Gamma)$, ce qui contredit l'hypothèse sur ψ. Donc $\psi > 0$ dès que ψ n'est pas identiquement nulle sur $L(\Gamma)$. On écrit alors

$$\varepsilon \leq \psi \leq \frac{1}{\varepsilon} \quad \text{et} \quad k^n\psi(v) = \int \|gv\|^s \psi(g.v)\, d\mu^n(g),$$

$$\varepsilon \int \|gv\|^s \, d\mu^n(g) \leq k^n\psi(v) \leq \frac{1}{\varepsilon} \int \|gv\|^s \, d\mu^n(g).$$

D'après le lemme précédent, on a donc $k = k(s)$.

Considérons le nouvel opérateur Markovien Q^s sur $P(V)$ défini par $Q^s u = \dfrac{1}{k(s)\varphi} P^s[u\varphi]$. L'équation $P^s\psi = k(s)\psi$ s'écrit alors $Q^s\left(\dfrac{\psi}{\varphi}\right) = \dfrac{\psi}{\varphi}$. L'ensemble Z' des points où $\dfrac{\psi}{\varphi}$ atteint son minimum m est un fermé Γ-invariant. Comme plus haut, on a $Z' \supset L_\Gamma$, en particulier $\psi = m\varphi$ sur $L(\Gamma)$. Le même argument utilisant le maximum m' de $\dfrac{\psi}{\varphi}$ donne $\psi = m'\varphi$ sur $L(\Gamma)$. Puisque φ n'est pas identiquement nulle, on en conclut $m = m'$ et $\psi = m\varphi$ sur $P(V)$, d'où les deux premières assertions du lemme.

Si ψ est continue et satisfait $P^s\psi \leq k(s)\psi$, la première partie de l'argument précédent donne encore $\psi = m\varphi$ sur $L(\Gamma)$. Si $k = k(s)$, il donne aussi $\psi = m\varphi$ sur $P(V)$. □

Lemme 4.11. *Soit ψ une fonction continue complexe et $k \in \mathbb{C}$ tels que $P^s\psi = k\psi$. Alors $|k| \leq k(s)$. Dans le cas $|k| = k(s)$, on a $k = k(s)$, et ψ est proportionnelle à la fonction φ du Lemme* 4.7.

Preuve. On utilise encore l'opérateur Q^s du lemme précédent défini à l'aide de la fonction q:

$$Q^s f(v) = \int f(g.v)q(v,g)\,d\mu(g)$$

avec $q(v,g) = \dfrac{\varphi(g.v)}{k(s)\varphi(v)}\|gv\|^s$ et $\varphi = e_s$. Alors $Q^s 1 = 1$, et $\psi' = \dfrac{\psi}{\varphi}$ satisfait

$$Q^s\psi' = \frac{k}{k(s)}\psi', \quad Q^s|\psi'| \geq \left|\frac{k}{k(s)}\right||\psi'|,$$

d'où $|\psi'|_\infty \geq \left|\dfrac{k}{k(s)}\right||\psi'|_\infty$ et $|k| \leq k(s)$.

Dans le cas où $k = k(s)e^{i\theta}$ avec $|e^{i\theta}| = 1$, on obtient $Q^s|\psi'| \geq |\psi'|$. Donc, dans ce cas, l'ensemble des points où $|\psi'|$ atteint son maximum est un fermé Γ-invariant noté M. D'après la Proposition 3.17 on a $M \supset L(\Gamma)$, donc $|\psi'| = |\psi'|_\infty > 0$. On a de plus $Q^s\psi' = e^{i\theta}\psi'$, donc $\psi'(g.v) = e^{i\theta}\psi'(v)$ sur $L(\Gamma)$ pour tout $g \in \Gamma$. On en déduit $\check{\mu}^n * \psi' = e^{in\theta}\psi'$ sur $L(\Gamma)$. Désignons par ν la probabilité stationnaire sur $L(\Gamma)$, qui satisfait donc l'équation $\mu * \nu = \nu$.

La Proposition 3.17 donne alors, en passant à la limite, $\lim_n e^{in\theta}\psi'(v) = \nu(\psi')$ pour tout v de $L(\Gamma)$. Donc $\psi' = cte$ sur $L(\Gamma)$. Comme $|\psi'| > 0$ sur $L(\Gamma)$, on en déduit que $\psi'(v) = \nu(\psi')$ pour tout $v \in L(\Gamma)$, $e^{i\theta} = 1$ et $k = k(s)$. On peut donc trouver $c \neq 0$ tel que $\psi' - c = 0$ sur $L(\Gamma)$. On a donc $Q^s\psi' = \psi'$ et le lemme précédent donne, puisque $\psi' = c$ sur $L(\Gamma)$, $\psi' = c$ sur $P(V)$, et donc $\psi = c\varphi$. □

Preuve du Théorème 4.1. L'existence de φ Höldérienne d'ordre $\bar{s}$, avec $P^s\varphi = k(s)\varphi$ découle du Lemme 4.8. Le Lemme 4.10 montre que si ψ est positive et satisfait $P^s\psi = k\psi$, on a $k = k(s)$, $\psi = c\varphi$, d'où les deux premières assertions du théorème.

Les deux assertions suivantes découlent des Lemmes 4.10 et 4.11.

Si la probabilité ν satisfait $P^s\nu = k\nu$, les Lemmes 4.3, 4.4, 4.5 montrent que $k = k(s)$. De plus le support de ν est un fermé Γ-invariant. D'après la Proposition 3.17, il contient $L(\Gamma)$, d'où l'avant dernière assertion.

La dernière assertion découle du Lemme 4.9. □

4.2 Convexité logarithmique et comportement asymptotique du rayon spectral de P^s

Les lemmes élémentaires suivants permettront de comparer l'exposant de croissance d'un fermé $F \subset \mathrm{GL}(V)$ défini par

$$\gamma^\infty(F) = \lim_n \frac{1}{n} \sup \left\{ \log \|g\| : g \in F^n \right\}$$

aux quantités $\lim_{s\to+\infty}(\log k_\mu(s))/s$ associées aux probabilités μ portées par F, dans le cas où $k_\mu(s) = k(s)$ est définie sur $\mathbb{R}_+$ (cf. Théorème 4.17 plus bas).

Lemme 4.12. *Supposons que le fermé $F \subset \mathrm{GL}(V)$ engendre un semi-groupe vérifiant la condition* (i–p) *. Alors on a l'égalité*

$$\gamma^\infty(F) = \limsup_n \frac{1}{n} \sup \left\{ \log r(g) : g \in F^n \right\},$$

où $r(g)$ est le rayon spectral de g.

Preuve. Clairement, on a $r(g) \le \|g\|$ pour $g \in F^n$, d'où l'inégalité

$$\gamma^\infty(F) \ge \limsup_n \frac{1}{n} \sup\{\log r(g) : g \in F^n\}.$$

Pour montrer l'inégalité inverse notons que si $\{h_i : i = 1, \dots, d^2\}$ est une base de $\mathrm{End}\, V$, on a avec une certaine constante C

$$\|g\| \le C \sup_{1\le i\le d^2} r(gh_i) \quad \forall\, g \in \mathrm{End}\, V.$$

En effet, notant Tr la forme trace, les d^2 formes linéaires $g \mapsto \mathrm{Tr}(gh_i)$ forment une base du dual de $\mathrm{End}\, V$, et donc, d'après l'équivalence des normes sur $\mathrm{End}\, V$, il existe une constante C_1 telle que

$$\|g\| \le C_1 \sup_{1\le i\le d^2} |\mathrm{Tr}(gh_i)|.$$

Mais on a clairement $|\mathrm{Tr}(gh_i)| \le d\, r(gh_i)$, et donc $\|g\| \le d\, C_1 \sup_{1\le i\le d^2} r(gh_i)$.

Puisque d'après le Lemme 3.18, le semi-groupe Γ engendré par F est absolument irréductible, le théorème de densité de Burnside implique l'existence d'une base h_i $(1 \le i \le d^2)$ de $\mathrm{End}\, V$, avec $h_i \in \Gamma$, chacun des h_i étant donc de la forme $h_i = g_1 \dots g_{p_i}$ où $g_1, \dots, g_{p_i}$ appartiennent à F, et les p_i sont des entiers inférieurs à q. On a donc, pour tout $g \in F^n$

$$\|g\| \le C \sup \left\{ r(g') : g' \in \bigcup_{0\le i\le q} F^{n+i} \right\}.$$

Posons $c = \limsup_n \sup \left\{ r(g)^{1/n} : g \in F^n \right\}$, fixons $\varepsilon > 0$ et notons que, pour n assez grand $(n > N)$ on a

$$r(g') \le (c+\varepsilon)^n \quad \forall\, g' \in F^n.$$

D'après l'inégalité précédente

$$\|g\|^{1/n} \le C^{1/n}\Big[\sup_{1\le k\le q}(c+\varepsilon)^{n+k}\Big]^{1/n} \quad \forall\, g \in F^n,$$

d'où

$$\gamma^\infty(F) \le \lim_n \Big[\sup_{1\le k\le q}\log(c+\varepsilon)^{n+k}\Big]^{1/n} = \log(c+\varepsilon)$$

et $\gamma^\infty(F) \le \log c$, puisque ε est arbitraire. □

Lemme 4.13. *Supposons que le rayon spectral de $g \in \operatorname{End} V$ soit 1 au plus. Alors, pour tout $\varepsilon > 0$, il existe une norme sur V telle que $\|g\| \le (1+\varepsilon)$. En particulier, si les valeurs propres de g sont supérieures à 1 en module, il existe une norme sur V telle que $\|gy\| \ge (1-\varepsilon)\|y\|\ \forall\, y \in V$.*

Preuve. Fixons une norme $x \mapsto |x|$ sur V et posons

$$\|x\| = \sum_{n=0}^{\infty} |g^n x| \left(1 - \frac{\varepsilon}{2}\right)^n.$$

Alors $x \mapsto \|x\|$ est bien définie puisque la série au second membre converge géométriquement. De plus $x \mapsto \|x\|$ est une norme, et si $\varepsilon < 1$

$$\|gx\| \le \frac{1}{1-\varepsilon/2}\|x\| \le (1+\varepsilon)\|x\|.$$

Dans l'hypothèse de la deuxième assertion, le rayon spectral de g^{-1} est au plus 1. Donc, d'après ce qui précède, on peut normer V avec

$$\|g^{-1}x\| \le (1+\varepsilon)\|x\| \quad \forall\, x \in V.$$

Posant $g^{-1}x = y$, on obtient

$$\|gy\| \ge \frac{1}{1+\varepsilon}\|y\| \ge (1-\varepsilon)\|y\| \quad \forall\, y \in V.$$ □

Lemme 4.14. *Soit $g \in \mathrm{GL}(V)$ de rayon spectral 1, et supposons que $V = V_1 \oplus V_r$ $(0 < r < 1)$, où $gV_1 = V_1$, $gV_r = V_r$, et les valeurs propres de g dans V_1 (resp., V_r) sont de module 1 (resp., de module $\le r$). Alors il existe une norme $v \mapsto \|v\|$ sur V telle que si l'on note, pour ε petit,*

$$U_\varepsilon = \{(y,z) \in V_1 \oplus V_r : \varepsilon\|y\| \ge 2(1-\varepsilon)\|z\|\},$$

on a $g(U_\varepsilon) \subset \mathring{U}_\varepsilon$, et $\|gx\| \ge (1-\varepsilon)\|x\|$ pour tout $x \in U_\varepsilon$.

Preuve. Supposons $(1+\varepsilon)r < 1-\frac{\varepsilon}{2}$ et normons V_1, V_r, en vertu du lemme précédent, de façon que

$$\|gy\|_1 \geq \left(1-\frac{\varepsilon}{2}\right)\|y\|_1 \quad \forall\, y \in V_1, \quad \|gz\|_r \leq r(1+\varepsilon)\|z\|_r \quad \forall\, z \in V_r.$$

Pour $x = y + z \in V_1 \oplus V_r$ on pose alors $\|x\| = \|y\|_1 + \|z\|_r$, et l'on a

$$\|gx\| \geq \left(1-\frac{\varepsilon}{2}\right)\|y\| \geq (1-\varepsilon)(\|y\| + \|z\|) = (1-\varepsilon)\|x\|$$

dès que $\varepsilon\|y\| \geq 2(1-\varepsilon)\|z\|$, c'est-à-dire $x \in U_\varepsilon$.

On a aussi, dans ces conditions

$$\|gy\| \geq \left(1-\frac{\varepsilon}{2}\right)\|y\|, \quad \|gz\| \leq (1+\varepsilon)r\|z\|,$$

donc

$$\frac{\|gy\|}{\|gz\|} \geq \frac{\|y\|}{\|z\|}\,\frac{1-\frac{\varepsilon}{2}}{r(1+\varepsilon)},$$

où $1-\frac{\varepsilon}{2} < r(1+\varepsilon)$. Cette dernière inégalité implique $gU_\varepsilon \subset \mathring{U}_\varepsilon$. □

Lemme 4.15. *Supposons que $g \in \mathrm{GL}(V)$ ait pour rayon spectral $r(g)$. Alors, il existe une constante $C > 0$ et une fonction $f(\varepsilon)$, définie pour ε petit et vérifiant $\lim_{\varepsilon\to 0} f(\varepsilon) = 0$, telle que si la suite $a_k \in \mathrm{GL}(V)$ vérifie $\|a_k - I\| \leq f(\varepsilon)$ alors on a*

$$\|ga_1 \dots ga_n\| \geq C\, r(g)^n (1-\varepsilon)^n \quad \forall\, n \in \mathbb{N}.$$

Preuve. On peut supposer $r(g) = 1$ et utiliser la norme définie au lemme précédent afin d'établir la relation de l'énoncé.

Montrons que l'on a alors $\|ga_1 \dots ga_n\| \geq (1-\varepsilon)^n$. Puisque, d'après le lemme précédent, $gU_\varepsilon \subset \mathring{U}_\varepsilon$, on peut trouver $f(\varepsilon) < \varepsilon/2$ tel que la condition $\|a - I\| \leq f(\varepsilon)$ implique $a(gU_\varepsilon) \subset U_\varepsilon$. On a de plus, pour tout $x \in gU_\varepsilon$

$$ax \in U_\varepsilon, \quad \|ax\| \geq \|x\| - \|(a-I)x\| \geq \|x\|[1 - f(\varepsilon)] \geq \|x\|(1-\varepsilon).$$

Soient alors $x \in gU_\varepsilon \subset U_\varepsilon$ et $a_1, a_2, \dots a_n$ vérifiant $\|a_k - I\| \leq f(\varepsilon)$. On a

$$ga_n x \in U_\varepsilon, \quad \|ga_n x\| \geq \|x\|(1-\varepsilon),$$

d'où par récurrence

$$\|ga_1 \dots ga_n x\| \geq \|x\|(1-\varepsilon)^n.$$

En particulier, $\|ga_1 \dots ga_n\| \geq (1-\varepsilon)^n$. □

Lemme 4.16. *Soit $S \subset \mathrm{GL}(V)$ et Γ le semi-groupe engendré par S. On suppose que Γ vérifie la condition* (i–p) *et que* $\dim V > 1$. *Alors le rapport $\|gx\|/\|x\|$ ne peut être égal à une constante k indépendante de la norme sur $S \times L(\Gamma)$.*

Preuve. Soit $v \mapsto |v|$ une autre norme sur V, désignons par $\bar{v}$ le point de $P(V)$ correspondant à v et notons $\|v\| = \varphi(\bar{v})|v|$ où $\varphi \in C[P(V)]$.

On a alors $\dfrac{|gv|}{|v|} = \dfrac{\varphi(\bar{v})}{\varphi(g.\bar{v})}\,\dfrac{\|gv\|}{\|v\|}$, donc, si l'assertion du lemme n'est pas satisfaite, $\varphi(g.\bar{v}) = \varphi(\bar{v})$ sur $S \times L(\Gamma)$. La minimalité de $L(\Gamma)$ implique alors $\varphi = cte$ sur $L(\Gamma)$; comme $L(\Gamma)$ contient au moins deux points différents, on peut trouver la norme $|\cdot|$, de façon que $\|v\|/|v|$ ne soit pas constant sur $L(\Gamma)$, ce qui donne la contradiction voulue. □

Théorème 4.17. *Supposons que la probabilité μ sur* $\mathrm{GL}(V)$ *vérifie* $\int \|g\|^s\, d\mu(g) < +\infty$ *pour tout* $s > 0$ *et que* Γ_μ *vérifie la condition* (i–p). *Alors la fonction* $\log k(s)$ *définie par*

$$\log k(s) = \lim_n \frac{1}{n} \log \int \|g\|^s\, d\mu^n(g)$$

est strictement convexe, sa dérivée à droite en $s = 0$ *vaut* γ_1, *et on a*

$$\lim_{s\to+\infty} \frac{\log k(s)}{s} = \gamma^\infty(S_\mu).$$

Preuve. La convexité de $\log k(s)$ découle de la relation

$$k(s) = \lim_{n\to\infty} \left[\int \|g\|^s\, d\mu^n(g)\right]^{1/n}$$

et de l'inégalité de Hölder ci-dessous, où $s, s', \alpha, \alpha' \geq 0, \alpha + \alpha' = 1$,

$$\int \|g\|^{\alpha s+\alpha' s'}\, d\mu^n(g) \leq \left[\int \|g\|^s\, d\mu^n(g)\right]^\alpha \times \left[\int \|g\|^{s'}\, d\mu^(g)\right]^{\alpha'}.$$

La dérivée en $s = 0$ de $\dfrac{1}{n}\log\left(\int \|g\|^s\, d\mu^n(g)\right)$ vaut $\dfrac{1}{n}\int \log\|g\|\, d\mu^n(g)$. Comme cette dernière suite converge vers γ_1 et que les fonctions de cette suite sont convexes, on a bien

$$(\log k)'(0^+) = \lim_n \frac{1}{n}\int \log\|g\|\, d\mu^n(g) = \gamma_1$$

(voir aussi la Proposition 8.2).

Pour obtenir la stricte convexité de $\log k(s)$ on utilise le Théorème 4.1 et les lemmes correspondants. Les relations $P^s e_s = k(s)e_s$ et $P^{s'} e_{s'} = k(s')e_{s'}$ entraînent, d'après l'inégalité de Hölder,

$$P^{\alpha s+\alpha' s'} e_s^\alpha e_{s'}^{\alpha'} \leq k^\alpha(s) k^{\alpha'}(s') e_s^\alpha e_{s'}^{\alpha'}.$$

Si l'on avait $k(\alpha s + \alpha' s') = k(s)^\alpha k(s')^{\alpha'}$, l'avant dernière assertion du Lemme 4.10 entraînerait $e_s^\alpha e_{s'}^{\alpha'} = cte\, e_{s\alpha+s'\alpha'}$ sur $L(\Gamma)$. Il y aurait donc égalité dans l'inégalité de Hölder précédente, au moins si $x \in L(\Gamma)$. On en déduit

$$\|gx\|^s e_s(g.x) = c\|gx\|^{s'} e_{s'}(g.x)$$

avec $c > 0$ pour tout $g \in S_\mu$ et tout $x \in L(\Gamma)$. En intégrant par rapport à μ,

$$k(s)e_s(x) = ck(s')e_{s'}(x) \quad \text{et} \quad \|gx\|^s k(s) = k(s')\|gx\|^{s'}.$$

Si $s \neq s'$, on en déduit $\|gx\| = k$ sur $S_\mu \times L(\Gamma)$, où la constante k vaut

$$k = \left[\frac{k(s)}{k(s')}\right]^{1/(s-s')}$$

et est donc indépendante de la norme choisie, par définition de $k(s)$. D'après le Lemme 4.16, ceci est impossible, donc $s = s'$.

Pour montrer la dernière assertion on observe que

$$\int \|g\|^s \, d\mu^n(g) \leq \sup\left\{\|g\|^s : g \in S_\mu^n\right\}.$$

Donc $\log k(s) \leq s\gamma^\infty(S_\mu)$ et $\lim_{s\to\infty}(\log k(s))/s \leq \gamma^\infty(S_\mu)$, où la limite au premier membre existe par convexité de $\log k(s)$. D'autre part, soit $g \in S_\mu$ et $\varepsilon > 0$, avec $r(g) = \sup_{h\in S_\mu} r(h) - \varepsilon$, et soit B_g^ε la boule de centre $g \in G$ définie par $\|g^{-1}h - I\| \leq \varepsilon$. Puisque $g \in S_\mu$, on a $\mu(B_g^\varepsilon) > 0$. On a alors

$$\begin{aligned}\int \|g_1 \dots g_n\|^s \, d\mu(g_1) \dots d\mu(g_n) &\geq \int_{(B_g^\varepsilon)^{\otimes n}} \|h_1 \dots h_n\|^s \, d\mu(h_1) \dots \, d\mu(h_n) \\ &\geq c(g) r(g)^{ns} (1-\varepsilon)^{ns} [\mu(B_g^\varepsilon)]^n\end{aligned}$$

d'après le Lemme 4.14. On en déduit

$$\begin{aligned}\lim_n \left[\int \|h\|^s d\mu^n(h)\right]^{1/n} &\geq r(g)^s (1-\varepsilon)^s \mu(B_g^\varepsilon), \\ \frac{\log k(s)}{s} &\geq \log r(g) + \log(1-\varepsilon) + \frac{1}{s}\log \mu(B_g^\varepsilon).\end{aligned}$$

En passant à la limite en s et ε,

$$\lim_{s\to+\infty} \frac{\log k(s)}{s} \geq \sup_{h\in(S_\mu)^n} \log r(h).$$

On a de même, pour tout $n \in \mathbb{N}$

$$\lim_{s\to+\infty} \frac{\log k(s)}{s} \geq \sup_{h\in(S_\mu)^n} \left(\frac{\log r(h)}{n}\right).$$

D'après le Lemme 4.12, on a donc

$$\lim_{s\to+\infty} \frac{\log k(s)}{s} \geq \limsup_n \frac{1}{n} \sup\left\{\log r(h) : h \in S_\mu^n\right\} = \gamma^\infty(S_\mu). \qquad \square$$

Remarque 4.18. L'analyticité de $k(s)$ sera obtenue en Section 8 (Théorème 8.8). Pour obtenir cette propriété il est nécessaire de développer l'étude, pour un s donné, de certains opérateurs Markoviens associés à μ afin de montrer que $k(s)$ est la valeur

spectrale isolée de P^s sur un espace de fonctions Höldériennes (Théorème 8.7). Cette étude fait l'objet des Sections 5, 6 qui suivent.

4.3 Les opérateurs Markoviens Q^s et la condition de continuité en variation des probabilités Q_v^s

On a vu au cours de la preuve du Lemme 4.11, que l'étude des opérateurs P^s se ramenait à celle des opérateurs Markoviens Q^s définis par

$$Q^s\varphi(x) = \frac{1}{k(s)e_s(x)} P^s(\varphi e_s)(x) = \int \varphi(g.x) q^s(x, g)\, d\mu(g),$$

où

$$q^s(x, g) = \frac{\|gx\|^s}{k(s)} \cdot \frac{e_s(g.x)}{e_s(x)}.$$

On note $\Omega = (S_\mu)^{\mathbb{N}}$, et l'on considère la probabilité Q_v^s sur Ω dont les projections sur les facteurs $(S_\mu)^{\otimes n}$ sont données par les probabilités $q_n^s(v, \omega) d\mu^{\otimes n}(\omega)$ où $\omega = (g_k)_{k>0}$ et

$$q_n^s(v, \omega) = q^s(v, g_1) q^s(g_1.v, g_2) \ldots q^s(g_{n-1} \ldots g_1.v, g_n).$$

Alors $v \mapsto Q_v^s$ est un noyau Markovien de $P(V)$ dans Ω, continu en topologie faible. On montre au Théorème 4.20 ci-dessous que cette famille de noyaux satisfait une condition d'absolue continuité à la Doeblin–Fortet (voir [12]), si la condition (i–p) est satisfaite par Γ_μ.

Définition 4.19. Pour $s \geq 0$, on note

$$q^s(v, g) = \frac{\|gv\|^s}{k(s)} \cdot \frac{e_s(g.v)}{e_s(v)},$$

où v appartient à $P(V)$, $g \in S_\mu$, et $k(s)$, $e_s > 0$ sont uniquement définis par la relation $P^s e_s = k(s) e_s$, $|e_s|_\infty = 1$ du Théorème 4.1. L'opérateur Markovien Q^s défini par

$$Q^s(\varphi) = \frac{1}{k(s)e_s} P^s(\varphi e_s)$$

sera dit opérateur Markovien associé à P^s.

Théorème 4.20. *Supposons la condition* (i–p) *satisfaite par* Γ_μ. *Alors, pour tout* $s > 0$, *il existe une constante* c_s *telle que pour tous* v *et* v' *de* $P(V)$

$$\|Q_v^s - Q_{v'}^s\| \leq c_s \delta^{\bar{s}}(v, v'),$$

où $\|\cdot\|$ *désigne la norme en variation des mesures sur* Ω, *et* $\bar{s} = \inf\{1, s\}$.

La preuve est basée sur le lemme d'équicontinuité suivant.

Lemme 4.21. *On note* $\theta^s(v,g) = \|gv\|^s \dfrac{e_s(g.v)}{e_s(v)}$. *Alors, il existe une constante* $K_s > 0$ *telle que* $|\theta^s(v,g) - \theta^s(v',g)| \le K_s \delta^{\bar{s}}(v,v')\|g\|^s$.

Preuve. D'après le théorème précédent, e_s est Höldériennes d'ordre $\bar{s}$, et l'on a donc, d'après le Lemme 4.6

$$|e_s(g.v) - e_s(g.v')| \le c_s \delta^{\bar{s}}(g.v, g.v') \le c'_s \left(\frac{\|g\|}{\|gv\|}\right)^{\bar{s}} \delta^{\bar{s}}(v,v')$$

avec $c'_s = 2^{\bar{s}}(s+1)c_s$, d'où

$$\begin{aligned}\left|\frac{e_s(g.v)}{e_s(v)} - \frac{e_s(g.v')}{e_s(v')}\right| &\le e_s(g.v)\left|\frac{1}{e_s(v)} - \frac{1}{e_s(v')}\right| + \frac{1}{e_s(v)}|e_s(g.v) - e_s(g.v')| \\ &\le K'_s \delta^{\bar{s}}(v,v')\left[1 + \left|\frac{\|g\|}{\|gv\|}\right|^{\bar{s}}\right]\end{aligned}$$

avec une certaine constante K'_s. On en déduit

$$\begin{aligned}|\theta^s(v,g) - \theta^s(v',g)| &\le \|gv\|^s \left|\frac{e_s(g.v)}{e_s(v)} - \frac{e_s(g.v')}{e_s(v')}\right| + \frac{e_s(g.v')}{e_s(v')}\big|\|gv\|^s - \|gv'\|^s\big| \\ &\le K'_s \delta^{\bar{s}}(v,v')\big[\|gv\|^s + \|g\|^s\|gv\|^{s-\bar{s}}\big] + K''_s \delta^{\bar{s}}(v,v')\|g\|^s.\end{aligned}$$

Puisque $s - \bar{s} \ge 0$, on a $\|g\|^{\bar{s}}\|gv\|^{s-\bar{s}} \le \|g\|^s$, et

$$|\theta^s(v,g) - \theta^s(v,g')| \le \big(2K'_s + K''_s\big)\|g\|^s \delta^{\bar{s}}(v,v'),$$

d'où l'inégalité voulue, avec $K_s = 2K'_s + K''_s$. □

Preuve du Théorème 4.20. Notons $g_k(\omega)$ $(k \ge 1)$ les coordonnées de $\omega \in S_\mu^{\otimes\mathbb{N}}$ et $\omega^k = g_k \dots g_1$. Si φ dépend des n premières coordonnées seules, on a, d'après la propriété de cocycle de θ^s

$$Q^s_v(\varphi) = \frac{1}{k(s)} \int \varphi(\omega)\theta^s(v,\omega^n)\, d\mu(g_1)\dots d\mu(g_n)$$

et d'après le lemme précédent

$$\begin{aligned}|Q^s_v(\varphi) - Q^s_{v'}(\varphi)| &\le \frac{1}{k^n(s)} \int |\varphi(\omega)|\,|\theta^s(v,\omega^n) - \theta^s(v',\omega^n)|\, d\mu(g_1)\dots d\mu(g_n) \\ &\le \frac{1}{k^n(s)}|\varphi|_\infty \int |\theta^s(v,g)| - \theta^s(v',g)|\, d\mu^n(g) \\ &\le K_s \frac{\delta^{\bar{s}}(v,v')}{k^n(s)}\,|\varphi|_\infty \int \|g\|^s\, d\mu^n(g).\end{aligned}$$

D'après le Lemme 4.4, on a $\int \|g\|^s\, d\mu^n(g) \le \varepsilon(s,\nu)k^n(s)$, ce qui donne la majoration

$$|Q^s_v(\varphi) - Q^s_{v'}(\varphi)| \le K_s \varepsilon(s,\nu)|\varphi|_\infty \delta^{\bar{s}}(v,v').$$

En faisant varier φ on obtient bien

$$\|Q_v^s - Q_{v'}^s\| \leq c_s \delta^{\bar{s}}(v, v')$$

avec $c_s = K_s \varepsilon(s, \nu)$. □

Nous montrons maintenant, pour compléter, l'unicité de la mesure propre ν^s, à un coefficient près.

Proposition 4.22. *Supposons que $s > 0$ et que Γ_μ vérifie la condition* (i–p)*. Alors, pour toute fonction φ continue sur $P(V)$, la suite $(Q^s)^n\varphi$ est équicontinue.*

Preuve. Il suffit de voir que si φ est ε-Höldériennes (avec $\varepsilon < \bar{s}$), les $(Q^s)^n\varphi$ ont un module de continuité indépendant de n. Or, dans ce cas

$$|(Q^s)^n\varphi(v) - (Q^s)^n\varphi(v')| \leq |\varphi|_\infty \int |q^s(v, g) - q^s(v', g)|\, d\mu^n(g) + [\varphi]_\varepsilon \int q_s(v', g)\delta^\varepsilon(g.v, g.v')\, d\mu^n(g).$$

D'après le lemme précédent, la première intégrale est majorée par

$$\frac{1}{k^n(s)} K_s \delta^{\bar{s}}(v, v') \int \|g\|^s\, d\mu^n(g).$$

D'après le Lemme 4.6, la deuxième intégrale est majorée par

$$K_s'' \frac{1}{k^n(s)} \int \|gv\|^{s-\varepsilon} \|g\|^\varepsilon\, d\mu^n(g),$$

donc par

$$K_s''' \frac{1}{k^n(s)} \int \|g\|^s\, d\mu^n(g),$$

car $\|gv\|^{s-\varepsilon} \leq \|g\|^{s-\varepsilon}$. Puisque, d'après le Lemme 4.5, la quantité

$$\frac{1}{k^n(s)} \int \|g\|^s\, d\mu^n(g)$$

est bornée, indépendamment de n, on a, avec une certaine constante c_ε,

$$\left|(Q^s)^n\varphi(v) - (Q^s)^n\varphi(v')\right| \leq c_\varepsilon \delta^\varepsilon(v, v').$$ □

Corollaire 4.23. *Supposons que Γ_μ satisfait* (i–p) *et que $\int \|g\|^s\, d\mu(g) < +\infty$ avec $s > 0$. Alors l'opérateur Markovien Q^s associé à P^s admet une unique probabilité invariante π^s. En particulier, il existe une unique probabilité ν telle que $P^s\nu = k(s)\nu$, et son support est égal à $L(\Gamma)$.*

Cet énoncé découle immédiatement du lemme élémentaire suivant dont la preuve est laissée au lecteur.

Lemme 4.24. *Soit X un espace métrique compact, Q un opérateur Markovien préservant $C(X)$. On suppose que pour toute $\varphi \in C(X)$, la suite $Q^n\varphi$ est équicontinue et que les seules fonctions Q-invariantes continues sont les constantes. Alors Q admet une unique probabilité Q-invariante.*

Preuve du Corollaire 4.23. Le fait que le support de π^s soit égal à $L(\Gamma)$ découle du théorème de Markov–Kakutani qui fournit une probabilité π portée par $L(\Gamma)$. Comme le support de π est un fermé Γ_μ-invariant contenu dans $L(\Gamma)$, l'assertion découle de la minimalité de $L(\Gamma)$ donnée par la Proposition 3.11. □

Remarque 4.25. Le corollaire précédent montre que l'énoncé du Théorème 4.1 peut être complété : la probabilité ν vérifiant $P^s\nu = k(s)\nu$ est unique et de support $L(\Gamma)$. Si $s = 0$, cela résulte de la Proposition 3.17. Dans la suite, on va examiner le cas où s est remplacé par un vecteur $(s) \in \mathbb{C}^d$. Il est alors commode de prouver d'abord les analogues des Théorèmes 4.1 et 4.20. Dans le cas où l'une des composantes de (s) s'annule, l'unicité de ν résulte de l'étude menée en Sections 5 et 6.

Nous montrons enfin que pour $s > 0$ les quantités

$$\inf_x \int \frac{1}{\|gx\|^s}\, d\check{\mu}^n(g), \quad \sup_x \int \frac{1}{\|gx\|^s}\, d\check{\mu}^n(g)$$

se comportent de manière différente pour certains s, si $\check{\mu}$ est portée par les matrices positives. On peut voir que c'est également le cas si le support de μ est contenu dans un sous-groupe du type de Schottky [11], au moins pour certains μ. Soit $V^+ = \mathbb{R}^d_+$ le cône positif de V, et normons V par

$$\|x\| = \sum_{i=1}^{d} |x_i|, \quad x = \sum_{1}^{d} x_i e_i.$$

Alors, si $\mu \in \operatorname{End} V$, on voit que

$$\|u\| = \sup_{1 \le i \le d} \|ue_i\|.$$

Notons aussi pour $\varepsilon > 0$

$$V_\varepsilon^+ = \{x \in V^+ : x_i \ge \varepsilon\|x\|,\ 1 \le i \le d\},$$
$$E_\varepsilon = \{u \in \operatorname{End} V : ue_i \in V_\varepsilon^+,\ 1 \le i \le d\}.$$

On note que E_ε est un sous-semi-groupe de $\operatorname{End} V$.

Proposition 4.26. *Avec les notations précédentes supposons que $x \in V_\varepsilon^+$ et $u, v \in E_\varepsilon$. Alors $\|ux\| \ge \varepsilon\|u\| \cdot \|x\|$ et $\|uv\| \ge \varepsilon^2\|u\| \cdot \|v\|$. En particulier, si μ est portée par les matrices positives et si $s > 0$, $x \in V_\varepsilon^+$, on a pour tout $n \in \mathbb{N}$*

$$\int \frac{1}{\|gx\|^s}\, d\mu^n(g) \le \varepsilon^{-2ns} \left[\int \frac{1}{\|g\|^s}\, d\mu(g)\right]^n.$$

Preuve. Dans la base e_i on a, pour $x \in V_\varepsilon^+$ et $u \in E_\varepsilon$

$$(ux)_i = \sum_j u_{ij}x_j \geq \varepsilon\|x\| \sum_j u_{ij}, \quad \|ux\| \geq \varepsilon\|x\| \sum_{i,j} u_{ij} \geq \varepsilon\|x\|\|u\|.$$

En particulier, pour $u, v \in E_\varepsilon$

$$\|uvx\| \geq \varepsilon\|vx\| \cdot \|u\| \geq \varepsilon^2\|u\| \cdot \|v\| \cdot \|x\|.$$

Si $S_\mu \subset E_\varepsilon^+$ désigne le support de μ, on a, pour tous $g_1, g_2, \ldots, g_n$ de S_μ et $x \in V_\varepsilon^+$ avec $\|x\| = 1$

$$\|g_1 \ldots g_n x\| \geq \varepsilon^{2n}\|g_1\| \ldots \|g_n\|,$$

d'où la dernière assertion. □

Corollaire 4.27. *Supposons que $\check{\mu}$ soit portée par $E_\varepsilon \cap \mathrm{SL}(V)$. Alors on a les inégalités :*

$$\inf_{x \in P(V)} \int \frac{1}{\|gx\|^d}\, d\check{\mu}^n(g) \leq \varepsilon^{-4n} \left[\int \frac{1}{\|g\|^d}\, d\check{\mu}(g)\right]^n,$$
$$\sup_{x \in P(V)} \int \frac{1}{\|gx\|^d}\, d\check{\mu}^n(g) \geq 1.$$

Preuve. La première inégalité découle de la proposition précédente avec $s = d$. Notons que si m est la mesure invariante par rotation sur $P(V)$, on a

$$\frac{dg^{-1}.m}{dm}(x) = \frac{1}{\|gx\|^d},$$

puisque g préserve la mesure de Lebesgue sur V. On en déduit

$$\int \frac{1}{\|gx\|^d}\, d\check{\mu}^n(g) dm(x) = (\mu^n * m)(1) = 1.$$

Comme $\sup_{x \in P(V)} \int \frac{1}{\|gx\|^d} d\check{\mu}^n(g)$ est minoré par $\int \frac{1}{\|gx\|^d}\, d\check{\mu}^n(g) dm(x)$, on a bien la dernière inégalité du corollaire. □

Nous pouvons maintenant justifier la Remarque 4.2(d). Si μ est comme dans le corollaire on en déduit :

$$\limsup_n \sup_x \left[\int \frac{1}{\|gx\|^d}\, d\check{\mu}^n(g)\right]^{1/n} \geq 1,$$
$$\liminf_n \inf_x \left[\int \frac{1}{\|gx\|^d}\, d\check{\mu}^n(g)\right]^{1/n} \leq \varepsilon^{-4} \int \frac{1}{\|g\|^d}\, d\check{\mu}(g),$$

et il suffit de choisir $\check{\mu}$ sur $E_\varepsilon \cap \mathrm{SL}(V)$ de sorte que l'intégrale $\int \frac{1}{\|g\|^d} d\check{\mu}(g)$ soit au plus $\varepsilon^4/4$, pour obtenir l'assertion voulue. Si le support de $\check{\mu}$ est contenu dans les matrices de norme minorée par $4\varepsilon^{-4/d}$, ceci est réalisé. En utilisant l'inégalité de Hölder, on obtient un résultat analogue pour $s \geq d$.

4.4 Extension des théorèmes précédents à l'espace des drapeaux

On décrit maintenant brièvement l'extension des Théorèmes 4.1, 4.17, 4.20 et de leurs corollaires à l'action de $\Gamma_\mu = \Gamma$ sur $\mathcal{F}$, en reprenant les notations de la fin de Section 3.

On note $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$ et $(s) = (s_1, s_2, ..., s_d)$ désigne un point de $\mathbb{R}_+^d$. En particulier, on note $\bar{s} = \inf\{\bar{s}_i : \bar{s}_i > 0\}$. On note aussi

$$\|\wedge g\|^{(s)} = \sup_{b\in\mathcal{F}} \sigma^{(s)}(g,b) = \prod_{k=1}^{d} \|\wedge^k g\|^{s_k},\ k[(s)] = \lim_{n\to\infty} \left[\int \|\wedge g\|^{(s)}\, d\mu^n(g)\right]^{1/n}.$$

Cette limite existe par sous-multiplicativité de $g \mapsto \|\wedge g\|^{(s)}$. Comme au début de cette Section, la fonction $k[(s)]$ est convexe sur la partie convexe de $(\mathbb{R}_+)^d$ où elle est définie.

Théorème 4.28. *Supposons que la probabilité μ sur $\mathrm{GL}(V)$ vérifie la condition $\int \|g\|^{(s)}\, d\mu(g) < +\infty$ pour un $(s) \in \mathbb{R}_+^d$, que Γ_μ satisfasse la condition ($\mathcal{F}$–i–p), et considérons l'opérateur $P^{(s)}$ sur $C(\mathcal{F})$ défini par*

$$P^{(s)}\varphi(b) = \int \sigma^{(s)}(g,b)\varphi(g.b)\, d\mu(g).$$

Alors l'équation $P^{(s)}\varphi = k\varphi$ a une unique solution strictement positive normalisée $\varphi = e_{(s)}$. On a $k = k[(s)]$, et $e_{(s)}$ est Höldériennes d'ordre $\bar{s} = \inf\{s_i, 1 : s_i > 0\}$.

Si une mesure de probabilité ν sur $\mathcal{F}$ satisfait $P^{(s)}\nu = k\nu$, on a $k = k[(s)]$, et le support de ν contient $L_{\mathcal{F}}(\Gamma)$. Si une fonction complexe continue ψ vérifie $P^{(s)}\psi = k\psi$, on a $|k| \leq k[(s)]$. Si $|k| = k[(s)]$, alors ψ est proportionnelle à $e_{(s)}$ et $k = k[(s)]$. Si ψ est positive et ne s'annule pas identiquement sur $L_{\mathcal{F}}(\Gamma)$, alors ψ est proportionnelle à $e_{(s)}$. Enfin, si ψ continue positive vérifie $P^{(s)}\psi \leq k[(s)]\psi$, alors ψ est proportionnelle à $e_{(s)}$ sur $L_{\mathcal{F}}(\Gamma)$.

Remarque 4.29. a) Il découle du Corollaire 6.9 que $\nu = \nu^{(s)}$ est unique de support $L_{\mathcal{F}}(\Gamma)$ et du Théorème 8.8, que $k[(s)]$ est analytique. Pour atteindre ces résultats il est nécessaire de développer des outils généraux supplémentaires aux Sections 5, 6 et 7.

b) Le Théorème 4.28 ne donne pas d'information précise pour $(s) = 0$. En ce cas $e_0 = 1$ appartient bien à la classe de Hölder, et ν est unique [15]. On peut montrer que $e_{(s)}$ est Hölder d'ordre $\varepsilon > 0$ pour (s) voisin de 0, que $k[(s)]$ est analytique au voisinage de 0 et que ses dérivées partielles aux ordres 1 et 2, admettent des

interprétations probabilistes en termes de moyennes et variances, comme dans le cas $d = 1$ (Corollaire 8.10).

Preuve du Théorème 4.28. La modification essentielle par rapport à la preuve du Théorème 4.1 résulte du changement de $\|gb\|^s$ en $\sigma^{(s)}(g, b)$. Par ailleurs, on peut considérer, comme en Section 3, $\mathcal{F}$ comme l'unique orbite compacte de $\mathrm{GL}(V)$ dans $P(\widehat{V})$. Les propriétés d'irréductibilité et de proximalité de Γ dans $P(\widehat{V})$ sont valides par hypothèse. Par exemple, pour établir l'analogue du Lemme 4.3, on raisonne dans $\mathrm{End}\,\widehat{V}$ et on obtient $\| \wedge g\|^{(s)} \leq c(\nu) \int \sigma^{(s)}(g, b)\, d\nu(b)$, dès que ν n'est pas portée par un sous-espace de $P(\widehat{V})$, d'où les analogues des Lemmes 4.3, 4.4 et 4.5.

Pour obtenir l'analogue du Lemme 4.6, on note que si x et y sont des points de $[0, 1]^d$, on a

$$|x_1 \dots x_d - y_1 \dots y_d| \leq \sum_{i=1}^{d} |x_k - y_k|.$$

Le début de la preuve du Lemme 4.6 donne

$$\left| \|(\wedge^k g)b_k\|^{s_k} - \|(\wedge^k g)(b'_k)\|^{s_k} \right| \leq (s_k + 1)\|b_k - b'_k\|^{\bar{s}_k} \| \wedge^k g\|^{s_k}.$$

Prenant alors

$$x_k = \left(\frac{\| \wedge^k gb_k\|}{\| \wedge^k g\|} \right)^{s_k}, \quad y_k = \left(\frac{\| \wedge^k gb'_k\|}{\| \wedge^k g\|} \right)^{s_k},$$

on obtient

$$|\sigma^{(s)}(g, b) - \sigma^{(s)}(g, b')| \leq \| \wedge g\|^{(s)} \sum_{s_k>0} (s_k + 1)\|b_k - b'_k\|^{\bar{s}_k},$$

d'où, en posant $|(s)| = \sup_k s_k$,

$$|\sigma^{(s)}(g, b) - \sigma^{(s)}(g, b')| \leq d(|s| + 1)2^{|s|} \sum_{s_k>0} \|b_k - b'_k\|^{\bar{s}} \| \wedge g\|^{(s)},$$

soit, avec une constante $c'_s > 0$,

$$|\sigma^{(s)}(g, b) - \sigma^{(s)}(g, b')| \leq c'_s \| \wedge g\|^{(s)} \delta^{\bar{s}}(b, b'),$$

d'où l'analogue du Lemme 4.6 et l'équicontinuité de la suite $\dfrac{1}{n} \displaystyle\sum_{j=1}^{n} \frac{(P^s)^j 1}{k^j[(s)]}$. On notera que la fonction propre $e_{(s)}$ ainsi construite ne dépend que des b_k pour lesquels $s_k > 0$. On a en particulier, avec une constante c''_s,

$$\left| e_{(s)}(b) - e_{(s)}(b') \right| \leq c''_s \sum_{s_k>0} \delta_k^{\bar{s}}(b_k, b'_k)$$

et $e_{(s)} = 1$ pour $(s) = 0$.

On en déduit l'existence d'une sous-suite convergente vers la fonction φ strictement positive et Höldériennes d'ordre $\bar{s}$ vérifiant $P^{(s)}\varphi = k[(s)]\varphi$. Les analogues des Lemmes 4.7, 4.8, 4.9 et 4.10 s'en déduisent et le Théorème 4.28 en résulte formellement. □

Pour une partie C de G on note

$$\gamma^{\infty}_{(s)}(C) = \lim_n \sup_{g_i \in C} \frac{1}{n} \log \| \wedge (g_1 \dots g_n)\|^{(s)}$$

et

$$\gamma_k = \lim_n \frac{1}{n} \int \log \| \wedge^k g \| \, d\mu^n(g).$$

On a aussi, comme plus haut

Théorème 4.30. *Supposons que S_μ engendre le semi-groupe Γ_μ, et que Γ_μ vérifie la condition* ($\mathcal{F}$–i–p) *. Alors la fonction*

$$\log k[(s)] = \lim_n \frac{1}{n} \log \int \| \wedge g\|^{(s)} \, d\mu^n(g)$$

est strictement convexe, et sa $k^{ième}$ dérivée partielle à droite en $s = 0$ est égale à $\gamma_1 + \gamma_2 + \dots + \gamma_k$. Si la fonction $k[(s)]$ est définie sur $\mathbb{R}^d_+$, on a pour tout $(s) \in \mathbb{R}^d_+$

$$\lim_{t \to +\infty} \frac{\log k[t(s)]}{t} = \gamma^{\infty}_{(s)}(S_\mu).$$

La preuve est analogue à celle du Théorème 4.17 en utilisant les espaces $(\wedge^k V)$, au lieu de V afin d'obtenir l'analogue du Lemme 4.15 et l'analogue du Lemme 4.16 afin d'obtenir la stricte convexité de $\log k[(s)]$.

On définit des noyaux normalisés $Q^{(s)}$ à l'aide du Théorème 4.28. On pose pour $(s) \in \mathbb{R}^d_+$, $b \in \mathcal{F}$, $g \in S_\mu$, $\varphi \in C(\mathcal{F})$

$$q^{(s)}(b, g) = \frac{\sigma^{(s)}(g, b)}{k[(s)]} \frac{e_{(s)}(g.b)}{e_{(s)}(b)}, \quad Q^{(s)}\varphi(b) = \int q^{(s)}(b, g)\varphi(g.b) \, d\mu(g).$$

On note aussi $Q^{(s)}_b$ les probabilités sur $\Omega = (S_\mu)^{\mathbb{N}}$ dont les projections sur les produits partiels $(S_\mu)^{\otimes n}$ valent $q^{(s)}_n(b, \omega) d\mu^{\otimes n}(\omega)$, où

$$q^{(s)}_n(b, \omega) = q(b, g_1) q^{(s)}(g_1.b, g_2) \dots q^{(s)}(g_{n-1} \dots g_1.b, g_n).$$

Définition 4.31. On pose

$$q^{(s)}(b, g) = \frac{\sigma^{(s)}(g, b)}{k[(s)]} \frac{e_{(s)}(g.b)}{e_{(s)}(b)},$$

où $k[(s)]$ et $e_{(s)}$ sont uniquement définis par le Théorème 4.28. Le noyau Markovien défini par

$$u \mapsto Q^{(s)}u(b) = \int u(g.b)q^{(s)}(b, g)\, d\mu(g)$$

sera appelé opérateur Markovien associé à $P^{(s)}$.

Théorème 4.32. *Il existe une constante $c_{(s)} > 0$ telle que pour tous b et b' de $\mathcal{F}$, on ait $\|Q_b^{(s)} - Q_{b'}^{(s)}\| \leq c_{(s)}\delta^{\bar{s}}(b, b')$.*

On établit le lemme d'équicontinuité suivant, analogue du Lemme 4.21 et la preuve du Théorème est ensuite identique à celle du Théorème 4.20.

Lemme 4.33. *Soient $\theta^{(s)}(g, b) = \sigma^{(s)}(g, b)\dfrac{e_{(s)}(g.b)}{e_{(s)}(b)}$. Alors il existe une constante $c_{(s)} > 0$ telle que*

$$|\theta^{(s)}(g, b) - \theta^{(s)}(g, b')| \leq c_{(s)}\delta^{\bar{s}}(b, b')\| \wedge g\|^{(s)}.$$

Preuve. On établit d'abord la majoration

$$|e_{(s)}(g.b) - e_{(s)}(g.b')| \leq c''_{(s)} \sum_{s_k>0} \left(\frac{\| \wedge^k g\|}{\| \wedge^k gb_k\|}\right)^{\bar{s}} \delta_k^{\bar{s}}(b_k, b'_k).$$

On a en effet, d'après la fin de la preuve du Théorème 4.28

$$|e_{(s)}(g.b) - e_{(s)}(g.b')| \leq c''_s \sum_{s_k>0} \delta_k^{\bar{s}}(g.b_k, g.b'_k).$$

De plus, comme au Lemme 4.6,

$$\delta_k(g.b_k, g.b'_k) \leq \frac{\| \wedge^k g\|}{\| \wedge^k gb_k\|}\delta_k(b_k, b'_k),$$

d'où l'inégalité annoncée. On écrit alors

$$\begin{aligned}|\theta^{(s)}(g, b) - \theta^{(s)}(g, b')| \leq{}& \sigma^{(s)}(g, b)\frac{1}{e_{(s)}(g.b)}\left|e_{(s)}(g.b) - e_{(s)}(g.b')\right| \\ &+ e_{(s)}(g.b')\frac{1}{e_{(s)}(b)}\left|\sigma^{(s)}(g, b) - \sigma^{(s)}(g, b')\right| \\ &+ \sigma^{(s)}(g, b)\left|\frac{1}{e_{(s)}(b)} - \frac{1}{e_{(s)}(b')}\right| e_{(s)}(g.b').\end{aligned}$$

D'après ce qui précède, le premier terme est majoré par la quantité

$$\left|\frac{1}{e_{(s)}}\right|_\infty c''_{(s)}\delta^{\bar{s}}(b, b') \sum_{s_k>0} \| \wedge_k \|^{\bar{s}} \prod_{i\neq k} \| \wedge^i gb_i\|^{s_i}\| \wedge^k gb_k\|^{s_k-\bar{s}},$$

donc, puisque $s_k - \bar{s} \geq 0$, par

$$\left|\frac{1}{e_{(s)}}\right|_\infty c''_{(s)} \delta^{\bar{s}}(b, b') \sum_{s_k>0} \| \wedge_k g\|^{s_k} \prod_{i\neq k} \| \wedge^i g b_i \|^{s_i},$$

d'où la majoration du premier terme par

$$c''_{(s)} d \left|\frac{1}{e_{(s)}}\right|_\infty \delta^{\bar{s}}(b, b') \| \wedge g\|^{(s)}.$$

D'après la fin de la preuve du Théorème 4.1 le deuxième terme est majoré par

$$\left|\frac{1}{e_{(s)}}\right|_\infty |e_{(s)}|_\infty c'_{(s)} \delta^{\bar{s}}(b, b') \| \wedge g\|^{(s)}.$$

Comme le troisième terme admet une majoration analogue, on a bien, avec une nouvelle constante $c_{(s)}$,

$$|\theta^{(s)}(g, b) - \theta^{(s)}(g, b')| \leq c_{(s)} \| \wedge g\|^{(s)} \delta^{\bar{s}}(b, b'). \qquad \square$$

5 Frontières et noyaux harmoniques

5.1 Notations

On reprend ici les notations générales de Section 2 et on considère donc un système Markovien $(X, q \otimes \mu)$.

Définition 5.1. On appellera système Markovien étendu la donnée d'un système Markovien $(X, q \otimes \mu)$ et d'une application Borélienne α de S_μ dans un semi-groupe Γ de $C(Y, Y)$ où Y est un espace métrique compact. On notera par $(X, q \otimes \mu, \alpha)$ un tel système. Cette notation sera précisée en $(X, q \otimes \mu, \alpha, Y)$ ou $(X, q \otimes \mu, \alpha, \Gamma)$ suivant la situation considérée.

On pourra toujours supposer que Γ est fermé en convergence uniforme, on notera $S_{\alpha(\mu)}$ le support de la mesure image $\alpha(\mu)$ et $\Gamma_{\alpha(\mu)}$ le semi-groupe engendré par $S_{\alpha(\mu)}$. On considère un espace métrique compact Y et une application α de S_μ dans $C(Y, Y)$. On notera $M^1(Y)$ le convexe compact, en topologie vague, des probabilités sur Y, et on posera, pour $\omega \in \Omega = S_\mu^{\mathbb{N}}$,

$$\alpha_n(\omega) = \alpha(a_1) \ldots \alpha(a_n) \in C(Y, Y).$$

Étant donné un système Markovien étendu, comme ci-dessus, on peut définir une action de S_μ sur $X \times Y$ ou $X \times \Gamma$ par les formules

$$a(x, y) = [ax, \alpha(a)y], \quad a(x, \gamma) = [ax, \gamma\alpha(a)].$$

On peut alors considérer la chaîne de Markov de noyau $\widetilde{Q}$ sur $X \times \Gamma$ définie par

$$\widetilde{Q}\psi(x, \gamma) = \int \psi[ax, \gamma\alpha(a)]q(x, a)\, d\mu(a).$$

On notera que $\widetilde{Q}$ commute avec les translations à gauche $(x, \gamma) \mapsto (x, \gamma'\gamma)$ sur $X \times \Gamma$. Rappelons qu'une fonction f sur $X \times \Gamma$ est $\widetilde{Q}$-harmonique ($\widetilde{Q}f = f$) si

$$f(x, \gamma) = \int f[ax, \gamma\alpha(a)]q(x, a)\, d\mu(a).$$

On s'intéresse ici aux fonctions harmoniques bornées "transversalement continues" au sens où la fonction $\sup_{\gamma\in\Gamma} |f(x, \gamma) - f(x', \gamma)|$ s'annule continument sur la diagonale de $X \times X$.

Définition 5.2. Soit $(X, q \otimes \mu, \alpha, Y)$ un système Markovien étendu, ν_x un noyau Markovien de X dans Y. On dira que ce noyau est harmonique s'il est continu en variation ($\lim_{x'\to x} \|\nu_{x'} - \nu_x\| = 0$) et s'il satisfait l'équation

$$\nu_x = \int \alpha(a)\nu_{ax}q(x, a)\, d\mu(a) \quad \forall\, x \in X. \tag{5.1}$$

Étant donné un noyau comme ci-dessus, la formule

$$f(x, \gamma) = \gamma\nu_x(\varphi),$$

où $\varphi \in C(X)$, définit des fonctions harmoniques transversalement continues.

On est également amené à considérer le décalage $\widetilde{\theta}$ sur l'espace des trajectoires $X \times \Gamma \times \Omega$ de la chaîne de noyau $\widetilde{Q}$

$$\widetilde{\theta}(x, \gamma, \omega) = [a_1x, \gamma\alpha(a_1), \theta\omega]$$

et les fonctions F, $\widetilde{\theta}$-invariantes,

$$F(x, \gamma, \omega) = F[a_1x, \gamma\alpha(a_1), \theta\omega].$$

La formule $f(x, \gamma) = \int F(x, \gamma, \omega)\, dQ_x(\omega)$ permet de passer d'une fonction $\widetilde{\theta}$-invariante F à une fonction $\widetilde{Q}$-harmonique f.

Si z est une application Borélienne de $X \times \Omega$ dans Y vérifiant l'équation d'équivariance

$$\alpha(a_1)z(a_1x, \theta\omega) = z(x, \omega), \tag{5.2}$$

la formule

$$F(x, \gamma, \omega) = \varphi[\gamma z(x, \omega)],$$

où $\varphi \in C(X)$, fournit des fonctions $\widetilde{\theta}$-invariantes, et la loi ν_x de $z(x, \omega)$ sous Q_x satisfait l'équation (5.1). En particulier, si $z(x, \omega) = z(\omega)$ est indépendante de x, l'équation (5.2) se réduit à

$$\alpha(a_1)z \circ \theta = z.$$

Alors ν_x est continue en variation car Q_x l'est. On va construire ici, en généralisant une importante méthode de H. Furstenberg ([18], [39, Chapter 6]), un noyau harmonique associé à un système Markovien étendu et une application équivariante $z(x, \omega)$ correspondante vérifiant l'équation (5.2). Le même type d'équation a été considéré en [9] dans le cadre d'algorithmes de fractions continues multidimensionnelles et joue un rôle essentiel dans l'étude de la simplicité du spectre de Lyapounov. Ceci va nous permettre de poursuivre l'étude de la situation de Section 4.

Dans ce dernier cas, on a $X = P(V)$, le groupe $\mathrm{GL}(V)$ opère projectivement sur X, et l'on note $g.x$ l'image de $x \in P(V)$ par l'application projective associée à $g \in \mathrm{GL}(V)$. Alors $q(x, a) = q^s(x, g)$ dépend d'un paramètre $s \geq 0$ et est donné par

$$q^s(x, g) = \frac{\|gx\|^s}{k(s)} \frac{e_s(g.x)}{e_s(x)}$$

(cf. Définition 4.19). Dans cette situation on a aussi $\alpha(g) = g$ pour $g \in S_\mu$. D'après les Théorèmes 4.1 et 4.20 on sait que $q^s \otimes \mu$ définit bien un système Markovien transitif.

Pour $s = 0$, on a $q^s(x, g) = 1$ et on peut prendre X réduit à un point. L'équation (5.1) s'écrit alors à l'aide d'une seule probabilité ν sur $Y = P(V)$

$$\nu = \int g.\nu \, d\mu(g) = \mu * \nu.$$

C'est l'équation de mesure stationnaire. Dans ce cas l'existence et l'unicité de ν et $z(\omega)$ sont bien connues sous la condition (i–p) (cf. [18], [39, Chapter 6]). Nos résultats seront appliqués dans la situation de Section 4, ainsi que dans la situation considérée en Section 2, où X est un espace produit $A^{\mathbb{N}}$, $q(x, a)$ est défini par une mesure de Gibbs et α est une fonction à valeurs dans $\mathrm{GL}(V)$ (cf. Sections 6 et 7).

5.2 Construction de noyaux harmoniques

Théorème 5.3. *Soit $(X, q \otimes \mu, \alpha, Y)$ un système Markovien étendu. Alors, avec les notations précédentes, l'équation*

$$\nu_x = \int \alpha(a) \nu_{ax} q(x, a) \, d\mu(a) \tag{5.3}$$

a au moins une solution continue en variation. Plus précisément, cette solution ν_x vérifie $\|\nu_x - \nu_{x'}\| \leq \|Q_x - Q_{x'}\|$.

Preuve. On suit la preuve du théorème de point fixe de Markov–Kakutani. On note $N(X, Y)$ le convexe des noyaux continus en topologie vague de X dans Y.

a) On considère l'opérateur R de $N(X, Y)$ dans lui-même défini par

$$(R\nu)_x = \int \alpha(a) \nu_{ax} q(x, a) \, d\mu(a).$$

Si m est une probabilité sur Y, on définit un noyau constant $\widehat{m}$ de X dans Y par $\widehat{m}_x = m$. Dans ce cas, on a la formule

$$(R^n \widehat{m})_x = \int q_n(x, \omega)\alpha(a_1) \dots \alpha(a_n) m \, d\mu^{\mathbb{N}}(\omega),$$

d'où la majoration

$$\|(R^n \widehat{m})_x - (R^n \widehat{m})_{x'}\| \leq \int |q_n(x, \omega) - q_n(x', \omega)| d\mu^{\mathbb{N}}(\omega) \leq \|Q_x - Q_{x'}\|.$$

b) Considérons sur $N(X, Y)$ la topologie donnée par la convergence vague sur Y et la convergence uniforme sur X. Alors, si $\varepsilon(x, x')$ est une fonction continue sur $X \times X$, nulle sur la diagonale, l'ensemble

$$N^\varepsilon = \{\nu \in N(X, Y) : \|\nu_x - \nu_{x'}\| \leq \varepsilon(x, x')\}$$

est un convexe compact de $N(X, Y)$, d'après le théorème d'Ascoli.

c) Considérons la suite de noyaux

$$\widehat{m}^n = \frac{1}{n} \sum_{k=0}^{n-1} R^k \widehat{m}$$

et prenons $\varepsilon(x, x') = \|Q_x - Q_{x'}\|$, fonction qui satisfait bien la condition de b) en vertu de l'hypothèse du théorème. Le calcul de a) montre que $\widehat{m}^n \in N^\varepsilon$, et on peut donc en extraire une sous-suite convergente vers le noyau ν. Clairement,

$$R\widehat{m}^n = \widehat{m}^n + \frac{1}{n}(-\widehat{m} + R^{n+1}\widehat{m}).$$

A la limite $R\nu = \nu$. Donc l'équation (5.3) est satisfaite par ν.

On a de plus, d'après a)

$$\|(\widehat{m}^n)_x - (\widehat{m}^n)_{x'}\| \leq \|Q_x - Q_{x'}\|,$$
$$\|\nu_x - \nu_{x'}\| \leq \lim_n \|(\widehat{m}^n)_x - (\widehat{m}^n)_{x'}\| \leq \|Q_x - Q_{x'}\|. \qquad \square$$

5.3 Unicité du noyau harmonique

On étudie maintenant l'unicité de la solution de (5.3). La définition générale suivante correspond à des situations géométriques variées, comprenant, en particulier,

- la situation considérée en Section 3 dans le cadre de la condition (i–p) (voir Corollaire 3.15), où Γ est un sous-semigroupe de $\mathrm{GL}(V)$ agissant sur $P(V)$ par transformations projectives [22];

- la situation des bords à l'infini des espaces hyperboliques et d'un groupe d'isométries Γ agissant sur ce bord (cf. [2, 32]);

- la situation des systèmes symboliques considérée en Section 2, Γ étant alors le semi-groupe des mots sur l'alphabet A, agissant sur l'espace produit $X = A^{[0,-\infty]}$ [46].

Définition 5.4. Soit Z un espace métrique compact, $\Gamma \subset C(Z, Z)$ un semi-groupe d'applications continues de Z dans Z, $\mathcal{D}$ une partie de $\mathcal{P}(Z)$ dont les éléments sont des fermés de Z différents de Z. On dit que $(\Gamma, \mathcal{D})$ satisfait la condition (I–P) si les propriétés suivantes sont vérifiés :

a) Pour toute suite $f_n \in \Gamma$, on peut trouver $D \in \mathcal{D}$ et une sous-suite convergeant uniformément sur tout compact de $Z \setminus D$ vers une application f de $Z \setminus D$ dans Z;

b) Γ est fortement proximal sur Z.

c) Pour tout $D \in \mathcal{D}$ et $c \in D$, il existe $g \in \Gamma$ avec $g.c \notin D$.

Théorème 5.5. *Soient $(X, q \otimes \mu, \alpha, Y)$ un système Markovien étendu, ν_x un noyau harmonique de X dans Y. On suppose que $S_{\alpha(\mu)}$ engendre un semi-groupe $\Gamma_{\alpha(\mu)}$ de $C(Y, Y)$ satisfaisant la condition* (I–P). *Alors pour tout $x \in X$ l'ensemble Ω_x des $\omega \in \Omega$, tels que la martingale $\alpha(a_1) \dots \alpha(a_n)\nu_{a_n \dots a_1 x}$ converge au sens vague vers une mesure de Dirac $\delta_{z(x,\omega)}$ est un Borélien vérifiant $Q_x(\Omega_x) = 1$.*

La preuve découle des trois lemmes suivants.

Lemme 5.6. *Avec les notations du théorème, fixons $x \in X$ et posons pour $(\omega, \eta) \in \Omega \times \Omega$ et $r, n \in \mathbb{N}$*

$$f_n(\omega) = \alpha(a_1) \dots \alpha(a_n)\nu_{a_n \dots a_1 x},$$
$$f_n^r(\omega, \eta) = \alpha(a_1) \dots \alpha(a_n)\alpha(\eta_1) \dots \alpha(\eta_r)\nu_{\eta_r \dots \eta_1 a_n \dots a_1 x}$$

et considérons la mesure produit $\mathbf{P}_\mu = \mu^{\mathbb{N}}$ sur Ω. Alors l'ensemble $(\Omega \times \Omega)_x$ des $(\omega, \eta) \in \Omega \times \Omega$, tels que pour tout r les suites $f_n^r(\omega, \eta)$ et $f_n(\omega)$ convergent et aient la même limite, est de $Q_x \otimes \mathbf{P}_\mu$-probabilité 1.

Preuve. Soit $\varphi \in C(Y)$ et notons $f_n^\varphi(\omega) = f_n(\omega)(\varphi)$. Puisque le noyau ν_x est harmonique, la suite $f_n^\varphi(\omega)$ est une Q_x-martingale bornée. Si Ω_x^φ est son ensemble de convergence, on a donc $Q_x(\Omega_x^\varphi) = 1$. Soit $(\Omega_x^{r,\varphi}) \subset \Omega \times S_\mu^{\otimes r}$ l'ensemble des $(\omega, \eta) \in \Omega \times S_\mu^{\otimes r}$, tels que $f_n^{r,\varphi} - f_n^\varphi$ converge vers zéro. Puisque $C(Y)$ est séparable et que $Q_x(\Omega_x^\varphi) = 1$, il suffit, pour obtenir l'assertion de l'énoncé, de voir que $(Q_x \otimes \mu^{\otimes r})(\Omega_x^{r,\varphi}) = 1$, pour tout $r \in \mathbb{N}$.

Notons $\widehat{q}(a) = \inf_{x \in X} q(x, a)$ et observons que la projection de Q_x sur $S_\mu^{\otimes r}$ est minorée par $(\widehat{q}\mu)^{\otimes r}$, qui est une mesure équivalente à $\mu^{\otimes r}$. On a donc

$$\int |f_n^{r,\varphi}(\omega, \eta) - f_n^\varphi(\omega)|^2 \, dQ_x(\omega) \, d(\widehat{q}\mu)^{\otimes r}(\eta) \leq \int |f_{n+r}^\varphi(\omega) - f_n^\varphi(\omega)|^2 \, dQ_x(\omega).$$

Puisque f_m^φ est une martingale, l'intégrale au second membre vaut

$$\int \left(|f_{n+r}^\varphi(\omega)|^2 - |f_n^\varphi(\omega)|^2 \right) dQ_x(\omega).$$

La série $\sum_1^\infty \int (f_{r+n}^\varphi(\omega) - f_n^\varphi(\omega))^2 \, dQ_x(\omega)$ se réduit donc à r termes, et on a la majoration

$$\sum_{n=1}^{\infty} \int |f_n^{r,\varphi}(\omega,\eta) - f_n^\varphi(\omega)|^2 \, dQ_x(\omega) \, d(\widehat{q}\mu)^{\otimes r}(\eta) \le r|\varphi|_\infty^2.$$

L'ensemble $\widetilde{\Omega}_x^{r,\varphi} \subset \Omega \times \Omega$ de convergence de la série précédente satisfait donc

$$[Q_x \otimes (\widehat{q}\mu)^{\otimes r}](\widetilde{\Omega}_x^{r,\varphi}) = 1.$$

Puisque $\widetilde{\Omega}_x^{r,\varphi} \subset \Omega_x^{r,\varphi}$ et que $(\widehat{q}\mu)^{\otimes r}$ est équivalente à $\mu^{\otimes r}$, on en conclut

$$(Q_x \otimes \mu^{\otimes r})(\Omega_x^{r,\varphi}) = 1. \qquad \square$$

Lemme 5.7. *Soit Z un espace métrique compact, ν_n une suite de probabilités sur Z, convergeant vaguement vers ν. Soit f_n une suite d'applications continues de Z dans Z convergeant uniformément vers f. Alors la suite de probabilités $f_n(\nu_n)$ converge vaguement vers $f(\nu)$.*

Preuve. Soit $\varphi \in C(Z)$ et estimons

$$I_n = f_n(\nu_n)(\varphi) - f(\nu)(\varphi) = (\nu - \nu_n)(\varphi \circ f_n) + \nu(\varphi \circ f - \varphi \circ f_n),$$
$$|I_n| \le |\varphi \circ f - \varphi \circ f_n|_\infty + |(\nu - \nu_n)(\varphi \circ f)|.$$

Puisque f_n converge uniformément vers f et que ν_n converge vaguement vers ν, les deux termes au second membre ont pour limite 0, d'où $\lim_n I_n = 0$. $\square$

Lemme 5.8. *Soit $\Omega' \subset \Omega$ un Borélien de P_μ-probabilité 1, M un compact en variation de $M^1(Y)$, $D \in \mathcal{D}$ et $x \in X$. Alors, il existe $c \notin D$ et une suite $\eta^k \in \Omega'$ telle que, en topologie vague*

$$\lim_k \alpha_k(\eta^k)\rho = \delta_c \quad \forall\, \rho \in M.$$

Preuve. Soit $\{\nu_i\}_{i\in\mathbb{N}}$ une suite dense de M, et posons $\nu = \sum_{i=1}^\infty 2^{-i}\nu_i$. D'après la proximalité forte de $\Gamma_{\alpha(\mu)}$ donnée par la condition (I–P), on peut trouver une suite $\gamma_k \in S_{\alpha(\mu)}^k$ telle que la suite $\gamma_k\nu$ converge vaguement vers une mesure de Dirac δ_c, et, d'après la condition (I–P), on peut supposer $c \notin D$. Puisque $\mathbf{P}_\mu(\Omega') = 1$, le fermé $\overline{\alpha_k(\Omega')}$ contient $S_{\alpha(\mu)}^k$, et on peut trouver $\eta^k \in \Omega'$ telle que $\alpha_k(\eta^k)\nu$ soit distant de $1/k$ au plus de $\gamma_k\nu$ en topologie vague. Alors

$$\lim_k \alpha_k(\eta^k)\nu = \delta_c.$$

Puisque $\nu = \sum_{i=1}^{\infty} 2^{-i}\nu_i$, on a aussi

$$\lim_k \alpha_k(\eta^k)\nu_i = \delta_c \quad \forall\, i \in \mathbb{N}.$$

Puisque $\{\nu_i\}_{i\in\mathbb{N}}$ est dense en variation dans M et que $\Gamma_{\alpha(\mu)}$ agit isométriquement sur $M^1(Y)$, pour la variation, on en déduit en topologie vague

$$\lim_k \alpha_k(\eta^k)\rho = \delta_c \quad \forall\, \rho \in M. \qquad \square$$

Preuve du Théorème 5.5. Fixons $x \in X$ et notons, pour $\omega \in \Omega$,

$$\Omega_{x,\omega} = \{\eta \in \Omega : (\omega, \eta) \in (\Omega \times \Omega)_x\},$$
$$\Omega'_x = \{\omega \in \Omega : \mathbf{P}_\mu(\Omega_{x,\omega}) = 1\}.$$

On sait, d'après le Lemme 5.6, que $Q_x(\Omega'_x) = 1$, et on va voir que $\Omega'_x \subset \Omega_x$, ce qui donnera $Q_x(\Omega_x) = 1$. Soit $\omega \in \Omega'_x$, et posons, pour $\eta \in \Omega_{x,\omega}$,

$$\nu^{x,\omega} = \lim_n \alpha_n(\omega)\nu_{s_n(\omega)x} = \lim_n \alpha_n(\omega)\alpha_r(\eta)\nu_{\eta_r\ldots\eta_1 s_n(\omega)x}.$$

D'après la partie a) de (I–P), on peut trouver $D \in \mathcal{D}$ tel que, pour une sous-suite $n_k \in \mathbb{N}$, on ait

$$\lim_k \alpha_{n_k}(\omega) = \tau_\omega$$

uniformément sur tout compact de $Y \setminus D$, où τ_ω est une application continue de $Y \setminus D$ dans Y. Par compacité de Y et par continuité de noyau ν_x, on peut supposer

$$\lim_k s_{n_k}(\omega)x = x^\omega, \quad \lim_k \nu_{\eta_r\ldots\eta_1 s_{n_k}(\omega)x} = \nu_{\eta_r\ldots\eta_1 x^\omega}.$$

Désignons par ρ^D la restriction de ρ à $Y \setminus D$. Alors, d'après ce qui précède et le Lemme 5.7

$$\nu^{x,\omega} \geq \tau_\omega\big[(\alpha_r(\eta)\nu^D_{\eta_r\ldots\eta_1 x^\omega})\big].$$

Puisque l'image de X par le noyau continu $x \to \nu_x$ est un compact en variation de $M^1(Y)$, le Lemme 5.8 permet de trouver $c \notin D$ et $\eta^r \in \Omega_{x,\omega}$ tels que

$$\lim_r \alpha_r(\eta^r)\nu_{s_r(\eta^r)x^\omega} = \delta_c.$$

Alors, prenant $\eta = \eta^r$ et passant à la limite, on obtient

$$\nu^{x,\omega} \geq \tau_\omega(\delta_c), \quad \nu^{x,\omega} = \delta_{\tau_\omega(c)}. \qquad \square$$

Théorème 5.9. *Avec les notations et hypothèses des Théorèmes* 5.3 *et* 5.5*, le noyau harmonique $x \mapsto \nu_x$ est unique. Soit $(X \times \Omega)^z = \bigcup_{x\in X}\{x\} \times \Omega_x$ le Borélien de $X \times \Omega$ définie par le Théorème* 5.5 *et $\widehat{\theta}$ la transformation de $X \times \Omega$ définie par $\widehat{\theta}(x, \omega) = (a_1 x, \theta\omega)$. Alors $(X \times \Omega)^z$ est $\widehat{\theta}$-invariant, satisfait $Q_x(\Omega_x) = 1$ pour tout*

x, et l'application z du Théorème 5.5 *vérifie la relation d'équivariance*

$$\alpha(a_1)z[a_1x, \theta\omega] = z(x, \omega).$$

Réciproquement, soit z' une application définie sur un Borélien de $X \times \Omega$, de la forme $\bigcup_{x\in X}\{x\} \times \Omega'_x$, qui satisfait les propriétés précédentes et telle que la loi sous Q_x de $z'(x, \omega)$ dépende continûment de x en variation. Alors $z' = z$, Q_x-p.p. pour tout $x \in X$.

Preuve. Les propriétés de $(X \times \Omega)^z$, Ω_x, $z(x, \omega)$ découlent des définitions et du Théorème 5.5. Soient ν_x, ν'_x deux noyaux $q \otimes \mu$-harmoniques, continus en variation, z et z' les deux applications définies par le Théorème 5.5, Ω_x et Ω'_x les parties de Ω correspondantes. Le noyau $x \mapsto (\nu_x + \nu'_x)/2$ est aussi $q \otimes \mu$-harmonique et le Théorème 5.5 fournit une application z'' et un ensemble Ω''_x correspondants. Puisque $Q_x(\Omega_x) = Q_x(\Omega'_x) = Q_x(\Omega''_x) = 1$, on a Q_x-presque partout $\Omega_x = \Omega'_x = \Omega''_x$ pour tout $x \in X$. En particulier, $\Omega_x \cap \Omega'_x \cap \Omega''_x \neq \varnothing$ et $Q_x(\Omega_x \cap \Omega'_x \cap \Omega''_x) = 1$. Pour $\omega \in \Omega_x \cap \Omega'_x \cap \Omega''_x$, on a clairement $(\delta_{z(x,\omega)} + \delta_{z'(x,\omega)})/2 = \delta_{z''(x,\omega)}$. On en déduit $z(x, \omega) = z'(x, \omega)$, Q_x-presque partout pour tout x. Comme l'image de Q_x par l'application $\omega \mapsto z(x, \omega)$ est ν_x, on a bien $\nu_x = \nu'_x$.

Si z' est donné comme dans le Théorème, on définit ν'_x comme étant l'image de Q_x par l'application $\omega \to z'(x, \omega)$, et l'hypothèse donne la continuité de ν'_x en variation. Alors la relation

$$\int \alpha(a_1)\nu'_{a_1x} q(x, a_1)\, d\mu(a_1) = \nu'_x$$

est satisfaite et le Théorème 5.5 fournit $\delta_{z'(x,\omega)} = \lim_n \alpha(a_1) \ldots \alpha(a_n)\nu'_{a_n \ldots a_1 x}$. D'après ce qui précède on a $\nu' = \nu$, donc Q_x-p.p. $z' = z$ pour tout $x \in X$. □

5.4 Ergodicité du système Markovien $(X, q \otimes \mu)$ et du décalage correspondant

Définition 5.10. Le décalage θ sur $\Omega = S_\mu^{\mathbb{N}}$ sera dit $q \otimes \mu$-ergodique, si pour toute fonction F Borélienne sur Ω vérifiant $F \circ \theta = F$, il existe une constante $c \in \mathbb{R}$ telle que $F(\omega) = c$ Q_x-p.p. pour tout $x \in X$.

Clairement la $q \otimes \mu$-ergodicité implique l'ergodicité de Q_π pour toute mesure invariante π.

Proposition 5.11. *Supposons le noyau Q_x continu en variation. Alors le décalage θ sur $S_\mu^{\mathbb{N}}$ est $q \otimes \mu$ ergodique, dès que la condition suivante est vérifiée: les seules solutions continues $h \in C(X)$ de l'équation $Qh = h$ sont les constantes. En ce cas, si π est une probabilité Q-invariante sur X, la probabilité Q_π sur $S_\mu^{\mathbb{N}}$ est indépendante de π. Si de plus, pour tout $\varphi \in C(X)$, la suite $Q^n\varphi$ est équicontinue, la probabilité Q-invariante π est unique.*

Preuve. Soit F une fonction bornée sur Ω avec $F \circ \theta = F$ et posons $f(x) = Q_x(F)$. Alors f est Q-harmonique, car

$$Qf(x) = \int f(ax)q(x,a)\,d\mu(a) = \int Q_{ax}(F \circ \theta)q(x,a)\,d\mu(a) = Q_x(F) = f(x).$$

Clairement $|f(x) - f(x')| \leq |F|_\infty \|Q_x - Q'_x\|$, ce qui implique la continuité de f, puisque Q_x est continue en variation. L'hypothèse entraîne alors que $f(x) = c$ pour tout x de X. On a de plus $\mathbf{E}_x(F|a_1, \ldots, a_n) = f(a_n \ldots a_1 x)$, et le théorème de convergence des espérances conditionnelles entraîne

$$F = \lim_n \mathbf{E}_x(F|a_1, \ldots, a_n) = c$$

Q_x-p.p., d'où la première assertion.

Si π est une probabilité Q-invariante sur X, alors la probabilité sur Ω définie par $Q_\pi = \int Q_x\,d\pi(x)$ est θ-invariante et ergodique par définition. En particulier, si π et π' sont Q-invariantes, on a $Q_\pi = Q_{\pi'}$. En effet $(\pi + \pi')/2 = \pi''$ est alors aussi Q-invariante, et la mesure $Q_{\pi''} = (Q_\pi + Q_{\pi'})/2$ est ergodique, ce qui implique $Q_\pi = Q_{\pi'}$.

La dernière assertion découle du lemme général d'unicité (Lemme 4.24). □

5.5 Cas où l'application équivariante $z(x, \omega)$ est indépendante de x

Si la condition (I–P) est satisfaite, on peut espérer, sous de larges conditions, que l'application équivariante $z(x, \omega)$ du Théorème 5.9 soit indépendante de x. Ce sera le cas en Section 6 où ce résultat sera obtenu directement. En général, on a

Proposition 5.12. *Avec les hypothèses du Théorème* 5.9*, supposons de plus que, pour tout $x \in X$, les probabilités Q_x soient équivalentes et que :*

$$\lim_n \int d(a_n \ldots a_1 x, a_n \ldots a_1 x')\,dQ_x(\omega) = 0 \quad \forall\, x, x' \in X.$$

Alors $z(x, \omega) = z(\omega)$ est indépendante de x.

Preuve. Soit ν_x l'unique noyau harmonique donné par le Théorème 5.9. D'après le Théorème 5.5 on a, Q_x-p.p.

$$\delta_{z(x,\omega)} = \lim_n \alpha_n(\omega)\nu_{s_n(\omega)x}.$$

Donc, puisque Q_x et $Q_{x'}$ sont équivalentes,

$$\int \|\delta_{z(x,\omega)} - \delta_{z(x',\omega)}\|\,dQ_x(\omega) \leq \limsup_n \int \|\nu_{s_n(\omega)x} - \nu_{s_n(\omega)x'}\|\,dQ_x(\omega).$$

Posons, pour $x, x' \in X$, $\varepsilon > 0$ fixés

$$\Omega_{n,\varepsilon} = \{\omega \in \Omega : d[s_n(\omega)x, s_n(\omega)x'] \leq \varepsilon\}, \quad \Omega^c_{n,\varepsilon} = \Omega \setminus \Omega_{n,\varepsilon},$$

et observons que l'hypothèse de l'énoncé implique $\lim_n Q_x(\Omega^c_{n,\varepsilon}) = 0$. D'après ce qui précède, on a la majoration suivante, pour tout $\varepsilon > 0$,

$$\int \|\delta_{z(x,\omega)} - \delta_{z(x',\omega)}\| \, dQ_x(\omega) \leq \limsup_n Q_x(\Omega^c_{n,\varepsilon}) + \sup_{d(x,x')\leq\varepsilon} \|\nu_x - \nu_{x'}\|.$$

Comme $x \mapsto \nu_x$ est continu en variation, le deuxième terme au second membre converge vers 0 lorsque ε tend vers 0, d'où

$$\int \|\delta_{z(x,\omega)} - \delta_{z(x',\omega)}\| \, dQ_x(\omega) = 0,$$

et $z(x, \omega) = z(x', \omega)$ pour Q_x-presque tout ω. □

6 Le noyau harmonique et les propriétés de contraction dans l'espace projectif

On précise ici les résultats de la Section 5, dans le cas où X est un espace métrique compact général et $Y = P(V)$, puis dans les cas $X = P(V)$, $Y = P(V^*)$ et $X = P(V)$, $Y = \mathcal{F}$; dans ces cas $(X, q \otimes \mu)$ est toujours un système Markovien. Ici α est une application de S_μ dans le groupe linéaire $\mathrm{GL}(V)$, qui définit une application encore notée α de S_μ dans le groupe projectif $\mathrm{PGL}(V)$. Le semi-groupe $\Gamma_{\alpha(\mu)}$ est supposé satisfaire la condition d'irréductibilité et de proximalité de la Section 3; d'après la Proposition 3.13, la condition (I–P) de la Section 5 est donc satisfaite avec $\mathcal{D}$ égal à l'ensemble des sous-espaces projectifs de $P(V)$. Les théorèmes de la Section 5 sont donc valides ici, et impliquent l'existence et l'unicité du noyau $q \otimes \mu$-harmonique ν_x de X dans $P(V)$, vérifiant l'équation

$$\nu_x = \int q(x, a)\, \alpha(a)\, \nu_{ax}\, d\mu(a). \tag{6.1}$$

On va étudier le noyau ν_x et l'application équivariante $z(x, \omega)$ qui ont été définis en Section 5. Ces informations permettront de préciser, par passage aux transposées, le comportement asymptotique de $S_n(\omega) = \alpha(a_n) \ldots \alpha(a_1) \in \mathrm{GL}(V)$, celui de l'application projective associée $s_n(\omega) \in \mathrm{PGL}(V)$, et d'obtenir ainsi l'extension des résultats de [18] annoncée en Section 5.

6.1 Les convergences en direction

On exploite ici le résultat de convergence de la martingale $\alpha_n(\omega)\nu_{s_n(\omega)x}$ vers $\delta_{z(x,\omega)}$, donné par le Théorème 5.5, et on montre que, dans cette convergence, la mesure variable $\nu_{s_n(\omega)x}$ peut être remplacée par une mesure fixe, ne chargeant pas de sous-espace projectif, par exemple la mesure m, invariante par rotation sur $P(V)$. Pour

cela, on montre d'abord une propriété essentielle : les mesures ν_x ne chargent pas de sous-espace projectif.

Théorème 6.1. *Soit $(X, q \otimes \mu)$ un système Markovien, α une application Borélienne de X dans $\mathrm{GL}(V)$ telle que $\Gamma_{\alpha(\mu)} \subset \mathrm{GL}(V)$ soit fortement irréductible. Alors, pour tout sous-espace projectif propre $H \subset P(V)$ et tout $x \in X$ on a $\nu_x(H) = 0$ pour l'unique noyau harmonique ν_x de X dans $P(V)$*

La preuve va découler de deux lemmes.

Lemme 6.2. *Soit ν une mesure de probabilité sur $P(V)$, $\mathcal{H}_\nu$ l'ensemble des sous-espaces projectifs de dimension minimum tels que $\nu(H) > 0$, et $\lambda = \sup\{\nu(H) : H \in \mathcal{H}_\nu\}$. Alors l'ensemble des éléments H de $\mathcal{H}_\nu$ tels que $\nu(H) = \lambda$ est non vide et fini. De plus il existe $\varepsilon_\nu > 0$ tel que pour tout $H \in \mathcal{H}_\nu$, on ait $\nu(H) = \lambda$ ou $\nu(H) \leq \lambda - \varepsilon_\nu$.*

Preuve. Si H et H' appartiennent à $\mathcal{H}_\nu$ et sont différents, on a $\nu(H \cap H') = 0$ car $\dim(H \cap H') < \dim H$. Il en découle que, pour $\alpha > 0$, l'ensemble des H de $\mathcal{H}_\nu$ tels que $\nu(H) \geq \alpha$, a un cardinal majoré par $1/\alpha$, d'où la première assertion du Lemme. Si la deuxième assertion n'était pas vraie, il existerait une suite H_n de $\mathcal{H}_\nu$ avec $\lambda/2 < \nu(H_n) < \lambda$ et $\lim_n \nu(H_n) = \lambda$. Par extraction éventuelle d'une sous-suite, on peut alors supposer que $\nu(H_n) \neq \nu(H_{n'})$ dès que $n \neq n'$, ce qui contredit le fait que les H_n sont en nombre au plus $2/\lambda$. □

Lemme 6.3. *Avec les notations précédentes, soit r un entier fixé, $\mathcal{H}$ une famille de réunions de r sous-espaces projectifs distincts préservée par l'action de $\Gamma_{\alpha(\mu)}$*

$$\gamma^{-1}(W) \in \mathcal{H} \quad \forall\, W \in \mathcal{H},\ \gamma \in \Gamma_{\alpha(\mu)}.$$

On pose $h(x) = \sup\{\nu_x(W) : W \in \mathcal{H}\}$. Alors h est continue, et l'ensemble des points où elle atteint son maximum est un fermé Γ_μ-invariant.

Preuve. Pour W fixé, la fonction $x \mapsto \nu_x(W)$ est continue, car $|\nu_x(W) - \nu_y(W)| \leq \|\nu_x - \nu_y\|$, d'où $|h(x) - h(y)| \leq \|\nu_x - \nu_y\|$, ce qui prouve la continuité de $h(x)$. On a, pour tout $W \in \mathcal{H}$

$$\nu_x(W) = \int q(x, a)\nu_{ax}[\alpha(a)^{-1}W]\, d\mu(a),$$

d'où $h(x) \leq \int q(x, a)\, h(ax)\, d\mu(a)$ par définition de h.

Soit $\lambda = \sup_{x \in X} h(x)$ et $C = \{x : h(x) = \lambda\}$. Puisque h est continue, C est un fermé non vide, et, si $x \in C$, l'inégalité précédente entraîne $h(ax) = \lambda$ pour tout $a \in S_\mu$, puisque $q(x, a) > 0$. Donc $\Gamma_{\alpha(\mu)} C \subset C$. □

Preuve du Théorème 6.1. Soit $\mathcal{H}_k$ l'ensemble des sous-espaces projectifs de dimension k $(0 \leq k < d)$ et notons

$$d(x) = \inf\{\dim H : H \in \mathcal{H}_k,\ \nu_x(H) > 0\},$$
$$m(x) = \sup\{\nu_x(H) : H \in \mathcal{H}_k,\ \dim H = d(x)\},$$
$$w(x) = \{H \in \mathcal{H}_k : \nu_x(H) = m(x)\}.$$

D'après le Lemme 6.2, $w(x)$ est non vide, de cardinal $n(x) < +\infty$. Posons aussi

$$p = \inf\{d(x) : x \in X\}, h_p(x) = \sup\{\nu_x(H) : H \in \mathcal{H}_p\}.$$

D'après le Lemme 6.3, $h_p(x)$ atteint sa borne supérieure λ sur un fermé $F \subset X$ qui est $\Gamma_{\alpha(\mu)}$-invariant. On peut supposer que F est $\Gamma_{\alpha(\mu)}$-minimal en remplaçant éventuellement F par un fermé plus petit. Alors, par définition de p on a, $h_p(x) = \lambda = m(x)$ sur F, d'où $d(x) = p$ sur F.

La relation $n(x)m(x) \leq 1$ implique $n(x) \leq 1/\lambda$ sur F. Posons alors $r = \sup\{n(x) : x \in F\}$, soit $\mathcal{H}_{p,r}$ l'ensemble des réunions de r éléments de $\mathcal{H}_p$ et notons $h_{p,r}(x) = \sup\{\nu_x(W) : W \in \mathcal{H}_{p,r}\}$. Alors le Lemme 6.3 entraîne $h_{p,r}(x) = r\lambda$ sur F. Puisque $m(x) = \lambda$, cette relation entraîne $n(x) = r$ sur F. Notons $W(x) = \bigcup_{H \in w(x)} H$ et montrons que la fonction $W(x)$ à valeurs dans $\mathcal{H}_{p,r}$ est localement constante sur F.

Le Lemme 6.2 entraîne que l'on a sur F

$$\sup\{\nu_x(H) : H \in \mathcal{H}_p,\ H \not\subset W(x)\} = \lambda(x) < \lambda.$$

Soit $x \in F$ et U voisinage de x tel que

$$\|\nu_y - \nu_x\| < \lambda - \lambda(x) \quad \forall\, y \in U.$$

Soit $H_y \in \mathcal{H}_p$ avec $\nu_y(H_y) = \lambda$. Alors

$$\lambda - \nu_x(H_y) = \nu_y(H_y) - \nu_x(H_y) \leq \|\nu_x - \nu_y\| < \lambda - \lambda(x).$$

Donc $\nu_x(H_y) > \lambda(x)$ et d'après le Lemme 6.2 $H_y \subset W(x)$. Donc $W(y) \subset W(x)$ et puisque $n(y) = n(x) = r$, on a bien $W(y) = W(x)$, pour tout $y \in U$.

Alors, puisque F est compact, $W = \bigcup_{x \in F} W(x)$ est une réunion finie de sous-espaces. D'autre part, les relations :

$$p\lambda = \nu_x[W(x)] = \int q(x,a)\alpha(a)\nu_{ax}[W(x)]\, d\mu(a), \quad p\lambda \geq \alpha(a)\nu_{ax}[W(x)]$$

entraînent que, pour tout x de F, on a μ-p.p. $p\lambda = \alpha(a)\nu_{ax}[W(x)]$.

Par définition de $W(ax)$, on en déduit $\alpha(a)^{-1}W(x) = W(ax)$, μ-presque partout. D'où, par continuité, $\alpha(a)^{-1}W(x) = W(ax)$ pour tout a de S_μ.

La relation

$$\alpha(a)^{-1}W = \bigcup_{x \in F} \alpha(a)^{-1}W(x) = \bigcup_{x \in F} W(ax) \subset W$$

montre que W est $\Gamma_{\alpha(\mu)}$-invariant. La condition d'irréductibilité forte de $\Gamma_{\alpha(\mu)}$ entraîne alors

$$W = P(V),\ r = 1,\ \lambda = 1,\ p = d - 1,\ d(x) = d - 1,\ m(x) = 1$$

pour tout x de X, d'où le théorème. □

Théorème 6.4. *Soit $(X, q \otimes \mu, \alpha)$ un système Markovien étendu, et supposons que $\Gamma_{\alpha(\mu)} \subset \mathrm{GL}(V)$ vérifie la condition* (i–p) *. Alors, la suite de mesures $\alpha(a_1) \dots \alpha(a_n)m$ converge Q_x-p.p., pour tout x vers une mesure de Dirac $\delta_{z(\omega)}$. De plus on a Q_x-p.p., pour tout x*

$$\alpha(a_1)z \circ \theta = z.$$

Cette condition définit uniquement l'application Borélienne z. De plus l'unique noyau ν_x qui est harmonique vérifie la relation $z(Q_x) = \nu_x$.

La preuve repose sur le lemme.

Lemme 6.5. *Soit u_n une suite d'applications projectives de $P(V)$ dans lui-même, ν_n et m des mesures de probabilité vérifiant $\nu_n(H) = m(H) = 0$ pour tout sous-espace projectif $H \subset P(V)$. On suppose la suite ν_n relativement compacte en variation. Alors la convergence vague de $u_n(\nu_n)$ vers une mesure de Dirac δ_z implique aussi celle de $u_n(m)$ vers δ_z.*

Preuve. On peut se ramener au cas où ν_n converge en variation vers ν, en extrayant des sous-suites. Par le même procédé on peut aussi supposer que u_n converge simplement vers une application quasi-projective continue en dehors d'un sous-espace projectif H [18]. On a alors, d'après le Lemme 5.7, $u(\nu) = \delta_z$. Donc $u(y)$ est constante et égale à z, ν-p.p. Puisque $\nu_n(H) = 0$ et que ν_n converge vers ν en variation, on a aussi $\nu(H) = 0$. Comme u est continue en dehors de H, on obtient $u(y) = z$ pour tout $y \notin H$. Comme $m(H) = 0$, on en déduit $u(m) = \delta_z$. □

Preuve du Théorème 6.4. Puisque le noyau ν_x est continu en variation, l'image par ν de l'espace compact X est un compact en variation. En particulier la suite $\nu_{a_n \dots a_1 x}$ est relativement compacte en variation. Soit Ω_x l'ensemble de convergence de la suite $\alpha(a_1) \dots \alpha(a_n)\nu_{a_n \dots a_1 x}$ vers $\delta_{z(x,\omega)}$ (cf. Théorème 5.5). On a donc d'après le Lemme 6.5 $\lim_n \alpha(a_1) \dots \alpha(a_n)m = \delta_{z(x,\omega)}$, et $z(x, \omega) = z(\omega)$ est bien indépendante de x.

Soit Ω° l'ensemble de convergence de la suite $\alpha(a_1) \dots \alpha(a_n)m$ vers une mesure de Dirac. On a donc $\Omega^\circ \supset \Omega_x$ pour tout x et $Q_x(\Omega^\circ) = Q_x(\Omega_x) = 1$. D'après le Théorème 5.9 on a aussi $z(Q_x) = \nu_x$. L'unicité de l'application $z(x, \omega)$ donnée par ce théorème implique ici $z(x, \omega) = z(\omega)$ Q_x-p.p. pour tout x. En particulier l'application z est uniquement définie par l'équation $\alpha(a_1)z \circ \theta = z$. □

Rappelons le lemme géométrique suivant qui repose sur le caractère projectif des applications considérées et dont on trouvera la preuve en [24].

Lemme 6.6. *Soit V un espace vectoriel réel de dimension finie, f_n une suite d'éléments de $\mathrm{GL}(V)$ et f_n^t la suite transposée (qui opère sur $P(V^*)$), δ la distance naturelle sur $P(V)$, m^* la mesure invariante par rotation sur $P(V^*)$, z un point de $P(V^*)$, z^* l'hyperplan de $P(V)$ défini par z. On suppose que $f_n^t(m^*)$ converge vers la mesure*

de Dirac δ_z. Alors pour tout $v, v' \in P(V) \setminus z^*$ on a

$$\lim_n \frac{\delta(f_n v, f_n v')}{\delta(v, v')} = 0,$$

et la convergence est uniforme si v et v' décrivent des compacts de $P(V) \setminus z^*$.

De plus, si v et z correspondent à des vecteurs unitaires de V et V^*, alors

$$\lim_n \frac{\|f_n v\|}{\|f_n\|} = |z(v)|.$$

On va maintenant considérer l'espace $Y = P(V^*)$ et l'application α^* définie par $\alpha^*(a) = \alpha(a)^t \in \mathrm{GL}(V^*)$. Par dualité, la condition (i–p) est satisfaite par le semi-groupe $\Gamma_{\alpha^*(\mu)} \subset \mathrm{PGL}(V^*)$. Les résultats généraux obtenus en Section 5 donnent donc l'existence de $z(\omega) \in P(V^*)$ satisfaisant la relation $\alpha^*(a)z \circ \theta = z$ et la convergence de $S_n^t(\omega).m^*$ vers $\delta_{z(\omega)} \in P(V^*)$. Le Lemme 6.6 va nous permettre d'en déduire les comportements en direction et norme du vecteur $S_n(\omega)v$ pour $v \in V$. On note aussi $z(\omega)$ le point de $P(V^*)$ défini par la relation $\alpha(a)z \circ \theta = z$ en vertu du Théorème 6.4. On identifiera, si besoin est, v et $z(\omega)$ à des vecteurs unitaires de V et V^*, et on utilisera la distance δ sur $P(V)$ définie par $\delta(v, v') = \|v \wedge v'\|$ avec $\|v\| = \|v'\| = 1$.

Théorème 6.7. *On reprend les notations et hypothèses du Théorème* 6.4. *Soit Ω° l'ensemble de convergence de la suite $s_n^t(\omega)m^*$ vers une mesure de Dirac. Alors si v et v' décrivent des compacts du complémentaire de $z^*(\omega)$ dans $P(V)$, la suite*

$$\frac{\delta(s_n(\omega)v, s_n(\omega)v')}{\delta(v, v')}$$

converge uniformément vers 0 *dès que $\omega \in \Omega^\circ$. De plus l'intégrale*

$$\int \delta[s_n(\omega)v, s_n(\omega)v']\, dQ_x(\omega)$$

converge uniformément vers 0, *et, si $\omega \in \Omega^\circ$ et $v \notin z^*(\omega)$ on a*

$$\lim_n \frac{\|S_n(\omega)v\|}{\|S_n(\omega)\|} = |z(\omega)(v)|.$$

Enfin, si f est une fonction Lipschitzienne sur $P(V)$, la suite de fonctions sur $X \times P(V)$ définie par

$$f_n(x, v) = \int f[s_n(\omega)v]\, dQ_x(\omega)$$

est équicontinue.

Preuve. D'après ce qui précède l'énoncé, le Théorème 6.4 donne la convergence de $\alpha(a_1)^t \ldots \alpha(a_n)^t m^*$ vers $\delta_{z(\omega)}$ pour tout $\omega \in \Omega^\circ$. Le Lemme 6.6 donne alors la première convergence annoncée.

Posons

$$\delta_n^x = \sup_{v,v' \in P(V)} \int \delta[s_n(\omega)v, s_n(\omega)v']\, dQ_x(\omega).$$

Comme l'intégrale est fonction continue de (v, v') sur $P(V) \times P(V)$, on a

$$\delta_n^x = \int \delta[s_n(\omega)\ v_n, s_n(\omega)\ v_n']\, dQ_x(\omega)$$

avec $(v_n, v_n') \in P(V) \times P(V)$. On peut supposer que (v_n, v_n') convergent vers (v, v'). Comme ν_x ne charge pas de sous-espace projectif, d'après le Théorème 6.1 on a Q_x-p.p. $v \notin z^*(\omega)$, $v' \notin z^*(\omega)$. On peut également supposer que $v_n \notin z^*(\omega)$, $v_n' \notin z^*(\omega)$ en prenant n assez grand. Alors la première convergence du théorème donne

$$\lim_n \delta[s_n(\omega)\ v_n, s_n(\omega)\ v_n'] = 0 \quad Q_x\text{-p.p.}$$

Par convergence dominée on obtient, pour tout x

$$\lim_n \delta_n^x = \lim_n \int \delta[s_n(\omega)v_n, s_n(\omega)v_n']\, dQ_x(\omega) = 0.$$

Puisque Q_x est continue en variation, la suite δ_n^x est équicontinue et a un module de continuité majoré par $\|Q_x - Q_{x'}\|$. La convergence de δ_n^x vers 0 est donc uniforme, d'où la deuxième assertion.

L'application du Lemme 6.6 donne encore pour $\omega \in \Omega^0$

$$\lim_n \frac{\|\alpha(a_n) \ldots \alpha(a_1)v\|}{\|\alpha(a_n) \ldots \alpha(a_1)\|} = |z(\omega)(v)|.$$

Notons

$$[f]_\delta = \sup_{v,v'} \frac{|f(v) - f(v')|}{\delta(v, v')}, \quad |\delta_n|_\infty = \sup_{x \in X} \delta_n^x.$$

Alors

$$\begin{aligned} |f_n(x, v) - f_n(x', v')| &\leq \int |f(s_n(\omega)v) - f(s_n(\omega)v')|\, dQ_x(\omega) \\ &\quad + \left| \int f(s_n(\omega)v')\, d(Q_x - Q_{x'})(\omega) \right| \\ &\leq [f]_\delta \int \delta(s_n(\omega)v, s_n(\omega)v')\, dQ_x(\omega) + |f|_\infty \|Q_x - Q_{x'}\| \\ &\leq [f]_\delta |\delta_n|_\infty + |f|_\infty \|Q_x - Q_{x'}\|. \end{aligned}$$

L'équicontinuité de la suite $f_n(x, v)$ découle alors des relations $\lim_n |\delta_n|_\infty = 0$ et $\lim_{x' \to x} \|Q_{x'} - Q_x\| = 0$. □

6.2 La continuité absolue des probabilités Q_v^s

On se place maintenant dans le cas $X = P(V)$, $Y = P(V^*)$, $S_\mu \subset \mathrm{GL}(V)$, et pour $g \in S_\mu$, $\alpha(g) = g^t$. On suppose que $\int \|g\|^s \, d\mu(g) < +\infty$, que S_μ satisfasse (i–p), et on considère pour $s \geq 0$ l'opérateur P^s sur $P(V)$ défini par

$$P^s\varphi(v) = \int \|gv\|^s \varphi(g.v)\, d\mu(g)$$

qui a déjà été considéré en Section 4. En reprenant les notations de la Section 4, on note e_s l'unique fonction continue définie par l'équation $P^s e_s = k(s)e_s$, $|e_s|_\infty = 1$. L'équation (6.1) s'écrit ici

$$\nu_x = \int g^t.\nu_{g.x} q^s(x, g)\, d\mu(g),$$

où, comme en Section 4, on pose

$$q^s(v, g) = \frac{\|gv\|^s}{k(s)} \frac{e_s(g.v)}{e_s(v)}.$$

On a vu en Section 4 que le noyau Q_x^s associé à $q^s \otimes \mu$ est continu en variation. Comme, par hypothèse, S_μ satisfait (i–p), les résultats précédents donnent l'existence de $z_s \in P(V^*)$ et du noyau ν_x^s de $P(V)$ dans $P(V^*)$ harmoniques relativement au système Markovien $[P(V), q^s \otimes \mu, \alpha]$. D'où, en vertu du Théorème 6.7, la convergence uniforme vers 0 de

$$\frac{\delta[S_n(\omega).v,\, S_n(\omega).v']}{\delta(v, v')}$$

pour $\omega \in \Omega^{\circ,s}$, ensemble de définition de $z_s(\omega)$, lorsque v et v' parcourent des compacts de $P(V) \setminus z_s^*(\omega)$.

Théorème 6.8. *Avec les notations précédentes, pour tous x, y de $P(V)$, les probabilités Q_x^s et Q_y^s sont équivalentes, et*

$$\frac{dQ_x^s}{dQ_y^s}(\omega) = \left|\frac{z_s(\omega)(x)}{z_s(\omega)(y)}\right|^s \frac{e_s(y)}{e_s(x)}. \tag{6.2}$$

Les probabilités ν_x^s et ν_y^s sur $P(V^*)$ sont aussi équivalentes, et l'on a

$$\frac{d\nu_x^s}{d\nu_y^s}(z) = \left|\frac{z(x)}{z(y)}\right|^s \frac{e_s(y)}{e_s(x)} \tag{6.3}$$

pour tous x, $y \in P(V)$, $z \in P(V^*)$.

Preuve. La première assertion découle du Théorème 6.7 et des observations précédant l'énoncé. Par définition de Q_x et Q_y on a

$$\mathbf{E}_y\left(\frac{dQ_x^s}{dQ_y^s}\middle| g_1, \ldots, g_n\right) = \frac{q_n^s(x, \omega^n)}{q_n^s(y, \omega^n)},$$

où $\omega^n = (g_n \dots g_1)$. D'après la forme de q_n^s on a

$$\frac{q_n^s(x,\omega^n)}{q_n^s(y,\omega^n)} = \frac{\|g_n \dots g_1 x\|^s}{\|g_n \dots g_1 y\|^s} \cdot \frac{e_s(g_n \dots g_1.x)}{e_s(g_n \dots g_1.y)} \cdot \frac{e_s(y)}{e_s(x)}.$$

D'après le théorème de convergence des espérances conditionnelles on a, Q_y^s-p.p.,

$$\lim_n \frac{q_n^s(x,\omega^n)}{q_n^s(y,\omega^n)} = \frac{dQ_x^s}{dQ_y^s}(\omega).$$

D'autre part, d'après la dernière assertion du Théorème 6.7, on a, Q_y^s-p.p.,

$$\lim_n \delta(g_n \dots g_1.y, g_n \dots g_1.x) = 0.$$

Ce théorème implique aussi que, Q_y^s-p.p.

$$\lim_n \frac{\|g_n \dots g_1 x\|}{\|g_n \dots g_1\|} = |z_s(\omega)(y)|, \quad \lim_n \frac{\|g_n \dots g_1 y\|}{\|g_n \dots g_1\|} = |z_s(\omega)(y)|.$$

On en déduit (6.2). Comme $\nu_x^s = z_s(Q_x^s)$ et $\nu_y^s = z_s(Q_y^s)$, on obtient également (6.3). □

Corollaire 6.9. *Avec les notations précédentes, la chaîne de Markov de noyau Q^s sur $P(V)$ admet une unique probabilité invariante π^s, et on a aussi $\pi^s = \int \nu_x^s \, d\pi^s(x)$. De plus, pour tout $x \in P(V)$ les probabilités ν_x^s sont équivalentes à π^s. Enfin, l'équation de mesure propre $P^s \nu^s = k(s)\nu^s$ admet une unique solution, à un coefficient près.*

Preuve. D'après le Théorème 6.7, si f est Lipschitzienne, la suite de fonctions $(Q^s)^n f$ est équicontinue. La même propriété reste valide pour f continue. Comme d'autre part, d'après le Théorème 4.1, les seules solutions continues de l'équation $Q^s h = h$ sont les constantes, on peut utiliser le Lemme 4.24. Celui-ci donne donc l'unicité de la probabilité Q^s-invariante π^s. L'équation (6.1) satisfaite par le noyau ν_x^s implique que si la probabilité π est Q^s-invariante, il en est de même de $\pi' = \int \nu_x^s \, d\pi(x)$. D'après l'unicité de π^s, on en déduit $\pi^s = \int \nu_x^s \, d\pi^s(x)$. Comme, d'après le Théorème 6.8, les mesures ν_x^s sont toutes équivalentes, cette formule montre que ν_x^s est équivalente à π^s.

L'unicité de la mesure propre pour P^s résulte de l'unicité de la probabilité Q^s-invariante. □

Remarque 6.10. D'après les Théorèmes 6.7 et 6.8 les hypothèses de la Proposition 5.12 sont satisfaites et donc l'indépendance de $z(x,\omega)$ par rapport à x en découle, mais ces théorèmes sont eux-mêmes basés sur cette propriété, qui ici a du être prouvée a priori. D'autre part, l'unicité de ν^s a déjà été prouvée au Corollaire 4.23 pour $s > 0$. La preuve précédente va s'étendre au cas d'un paramètre multi-dimensionnel (Corollaire 6.12.

Considérons maintenant, comme en Sections 3 et 4, le cocycle $\sigma^{(s)}(g,b) = \|(\wedge g)b\|^{(s)}$ sur $G \times \mathcal{F}$ et l'opérateur $P^{(s)}$ correspondant, sous la condition ($\mathcal{F}$–i–p).

On considère comme en Section 3, $\mathcal{F}$ plongé dans $P(\widehat{V})$. Le système Markovien $(\mathcal{F}, q^{(s)} \otimes \mu)$ a été défini en Section 4 par :

$$q^{(s)}(b, g) = \frac{1}{k[(s)]}\sigma^{(s)}(g, b)\frac{e_{(s)}(g.b)}{e_{(s)}(b)},$$

et on considère le système Markovien étendu $(\mathcal{F}, q^{(s)} \otimes \mu, \alpha, \mathcal{F})$ avec $\alpha(g) = \hat{g}^t \in \mathrm{GL}(\widehat{V})$. L'équation (6.1) pour le noyau $b \mapsto \nu_b^{(s)}$ de $\mathcal{F}$ dans $\mathcal{F}$ s'écrit ici

$$\nu_b^{(s)} = \int \hat{g}^t.\nu_{g.b}q^{(s)}(b, g)\, d\mu(g).$$

On sait, d'après le Théorème 6.7, que si $\hat{m}$ est la probabilité invariante par rotation sur $\mathcal{F}$, la suite $\hat{g}_1^t \dots \hat{g}_n^t.m$ converge $Q_b^{(s)}$-p.p. vers la mesure de Dirac concentrée au point $\hat{z}_{(s)}(\omega) \in \mathcal{F}$. Alors $\nu_b^{(s)}$ est la loi de $\hat{z}_{(s)}(\omega)$ sous $Q_b^{(s)}$.

Pour $\xi, b \in \mathcal{F}$ unitaires avec $\xi = (\xi_1, \xi_2, \dots, \xi_{d-1})$, $b = (b_1, b_2, \dots, b_{d-1})$ on définit enfin, avec les notations de la Section 3

$$|\langle \xi, b\rangle|^{(s)} = \prod_{k=1}^{d-1} |\langle b_k, \xi_{d-k}\rangle|^{s_k}.$$

On peut alors énoncer l'analogue du Théorème 6.8. La preuve est analogue à celle du Théorème 6.8.

Théorème 6.11. *Avec les notations précédentes supposons que Γ_μ satisfait la condition ($\mathcal{F}$–i–p) et soit $(s) \in \mathbb{R}_+^d$. Alors, pour tous $b, b' \in \mathcal{F}$ les probabilités $Q_b^{(s)}$ et $Q_{b'}^{(s)}$ sont équivalentes, et l'on a*

$$\frac{dQ_b^{(s)}}{dQ_{b'}^{(s)}}(\omega) = \frac{|\langle \hat{z}_{(s)}(\omega), b\rangle|^{(s)}}{|\langle \hat{z}_{(s)}(\omega), b'\rangle|^{(s)}} \cdot \frac{e_{(s)}(b')}{e_{(s)}(b)}.$$

Les probabilités $\nu_b^{(s)}$, $\nu_{b'}^{(s)}$ sur $\mathcal{F}$ sont aussi équivalentes, et pour tout $\xi \in \mathcal{F}$

$$\frac{d\nu_b^{(s)}}{d\nu_{b'}^{(s)}}(\xi) = \frac{|\langle \xi, b\rangle|^{(s)}}{|\langle \xi, b'\rangle|^{(s)}} \cdot \frac{e_{(s)}(b')}{e_{(s)}(b)}.$$

Comme précédemment, on en déduit

Corollaire 6.12. *Avec les notations du Théorème* 6.11*, la chaîne de Markov de noyau $Q^{(s)}$ sur $\mathcal{F}$ admet une unique probabilité invariante $\pi^{(s)}$ et l'on a $\pi^{(s)} = \int \nu_b^{(s)} d\pi^{(s)}(b)$. De plus, pour tout $b \in \mathcal{F}$, les probabilités $\nu_b^{(s)}$ sont équivalentes à $\pi^{(s)}$. Enfin l'équation $P^{(s)}\nu^{(s)} = k[(s)]\nu^{(s)}$ admet une unique solution, à un coefficient près.*

7 Les exposants caractéristiques de $S_n(\omega)$

On considère un système Markovien étendu $(X, q \otimes \mu, \alpha)$ où α est à valeurs dans $\mathrm{GL}(V)$, et on reprend les notations générales de la Section 2. On suppose ici la finitude des intégrales $\int \log \|\alpha(a)\| q(a)\, d\mu(a)$ et $\int \log \|\alpha(a)^{-1}\| q(a)\, d\mu(a)$.

Le décalage sur $\Omega = S_\mu^{\mathbb{N}}$ est supposé $q \otimes \mu$-ergodique, ce qui assure que la probabilité Q_π sur Ω est bien définie. Les exposants γ_1^q et γ_2^q sont alors des constantes définies par

$$\gamma_1^q = \lim_n \frac{1}{n} \int \log \|S_n(\omega)\|\, dQ_\pi(\omega),$$

$$\gamma_2^q + \gamma_1^q = \lim_n \frac{1}{n} \int \log \| \wedge^2 S_n(\omega)\|\, dQ_\pi(\omega).$$

en raison de l'ergodicité de Q_π et des conditions d'intégrabilité (cf. Section 2).

On montre ici que, sous des conditions naturelles, on a $\gamma_2^q < \gamma_1^q$, ce qui généralise un résultat de [24, 25] et permet de prouver le Théorème 2.8 et son corollaire.

7.1 Quelques lemmes préparatoires

Afin d'obtenir des expressions intégrales de γ_1^q, γ_2^q on note $P^2(V) \subset P(\wedge^2 V)$ l'ensemble des 2-plans de V, et $P^{1,2}(V) \subset P(V) \times P^2(V)$ l'ensemble des couples ξ formés d'un point v de $P(V)$ et d'une droite passant par ce point. Un tel élément de contact ξ est représenté par son origine $v \in P(V)$, qui correspond à un vecteur unitaire v de V et par un bivecteur $v \wedge v'$ de $\wedge^2(V)$, qui définit la droite de $P(V)$, et est également supposé unitaire.

On note σ_1, σ_2 les cocycles sur $G \times P(V)$ ou $G \times P^2(V)$ définis par

$$\sigma_1(g, v) = \|gv\|, \quad \sigma_2(g, v \wedge v') = \|g(v \wedge v')\|.$$

De plus on note

$$\sigma(g, \xi) = \frac{\|g(v \wedge v')\|}{\|gv\|^2},$$

où ξ est repéré par $v \in P(V)$ et $(v \wedge v') \in P^2(V)$.

On note $\widetilde{Q}^1$ l'extension à $X \times P(V)$ de la chaîne de Markov de noyau $q \otimes \mu$

$$\widetilde{Q}^1 \varphi(x, v) = \int \varphi[ax, \alpha(a)v]\, q(x, a)\, d\mu(a).$$

De même $\widetilde{Q}^2$ et $\widetilde{Q}^{1,2}$ désigneront les extensions analogues de Q à $X \times P^2(V)$ et $X \times P^{1,2}(V)$. On remplacera éventuellement $\widetilde{Q}^1$, $\widetilde{Q}^2$, $\widetilde{Q}^{1,2}$ par la notation générique $\widetilde{Q}$, l'indice étant alors donné par le contexte. On note aussi M_V^1, M_V^2, $M_V^{1,2}$ les convexes formés des probabilités $\widetilde{Q}$-invariantes.

Pour $\eta_1 \in M_V^1$ (avec $\widetilde{Q}^1\eta_1 = \eta_1$) on note

$$I^1_{\eta_1} = \int \log \sigma_1(\alpha(a), v)\, d\eta_1(x, v)\, q(x, a)\, d\mu(a),$$

de même pour $\eta_2 \in M_V^{1,2}$

$$I^2_{\eta_2} = \int \log \sigma_2(\alpha(a), w)\, d\eta_2(x, w)\, q(x, a)\, d\mu(a),$$

et si $\eta \in M_V^{1,2}$ se projette sur η_1 et η_2

$$I^{1,2}_{\eta} = \int \log \sigma(\alpha(a), \xi)\, d\eta(x, \xi)\, q(x, a)\, d\mu(a) = I^2_{\eta_2} - 2I^1_{\eta_1}.$$

On énonce d'abord quelques lemmes généraux sur les exposants γ_1^q, γ_2^q.

Lemme 7.1. *Supposons que le décalage sur $\Omega = S_\mu^{\mathbb{N}}$ soit $q \otimes \mu$-ergodique, et que*

$$\int \log \|\alpha(a)\| q(a)\, d\mu(a) < +\infty, \quad \int \log \|\alpha(a)^{-1}\| q(a)\, d\mu(a) < +\infty.$$

Alors on a

$$\gamma_1^q = \sup_{\eta \in M_V^1} I^1_\eta, \quad \gamma_1^q + \gamma_2^q = \sup_{\eta \in M_V^2} I^2_\eta.$$

Preuve. Détaillons la preuve de la première formule et indiquons les modifications qui permettent d'obtenir la deuxième. On applique le théorème de Birkhoff à la chaîne de Markov $\widetilde{Q}^1$ sur $X \times P(V)$, pour la loi initiale $\eta \in M_V^1$ et la fonction $f(x, v, \omega) = \log \|\alpha(a_1)v\|$. Le décalage $\widetilde{\theta}$ sur $X \times P(V) \times \Omega$ (l'espace des trajectoires de la chaîne de noyau $\widetilde{Q}^1$) est donné par

$$\widetilde{\theta}(x, v, \omega) = (a_1 x, \alpha(a_1)v, \theta\omega).$$

L'identité de cocycle pour σ^1 donne

$$\sum_{k=0}^{n-1} f \circ \widetilde{\theta}^k(x, v, \omega) = \log \|S_n(\omega)v\|. \tag{7.1}$$

Par hypothèse la fonction f est intégrable, et on peut supposer η ergodique, afin d'obtenir la première formule. Le théorème de Birkhoff donne alors, puisque η est $\widetilde{Q}^1$-invariante,

$$\begin{aligned} I^1_\eta &= \lim_n \frac{1}{n} \int \log \|S_n(\omega)v\|\, dQ_x(\omega)\, d\eta(x, v) \\ &\le \lim_n \frac{1}{n} \int \log \|S_n(\omega)\|\, dQ_x(\omega)\, d\pi(x) \\ &= \lim_n \frac{1}{n} \int \log \|S_n(\omega)\|\, dQ_\pi(\omega), \end{aligned}$$

d'où $I_\eta^1 \le \gamma_1^q$, par définition de γ_1^q.

Montrons l'existence de $\eta \in M_V^1$ telle que $\gamma_1^q = I_\eta^1$. D'après la Proposition 3.29, il existe une constante $c > 0$ telle que pour tout $g \in \mathrm{GL}(V)$

$$0 \le \log \|g\| - \int \log \|gv\| \, dm(v) \le c,$$

d'où, en intégrant,

$$0 \le \int dm(v) \int \big(\log \|S_n(\omega)\| - \log \|S_n(\omega)v\|\big) \, dQ_\pi(\omega) \le c.$$

La suite de fonctions $h_n(v)$ définie par

$$h_n(v) = \frac{1}{n} \int \big(\log \|S_n(\omega)\| - \log \|S_n(\omega)v\|\big) \, dQ_\pi(\omega)$$

est positive, majorée par c et satisfait la relation

$$\lim_n \int h_n(v) \, dm(v) \le \lim_n \frac{c}{n} = 0.$$

Donc, il existe une sous-suite $h_{n_j}(v)$ convergeant m-p.p. vers 0 et telle que

$$\gamma_1^q = \lim_j \frac{1}{n_j} \int \log \|S_{n_j}(\omega)v\| \, dQ_\pi(\omega).$$

Considérons alors la fonction continue $\bar{f}$ de $(x, v) \in X \times P(V)$

$$\bar{f}(x, v) = \int f(x, v, \omega) \, dQ_x(\omega) = \int \log \|\alpha(a_1)v\| q(x, a_1) \, d\mu(a_1).$$

En intégrant la relation (7.1) par rapport à Q_π, on obtient

$$\sum_{k=0}^{n-1} \widetilde{Q}^k(\pi \times \delta_v)(\bar{f}) = \int \log \|S_n(\omega)v\| \, dQ_\pi(\omega).$$

Considérons la suite de probabilités η_n sur $X \times P(V)$ donnée par

$$\eta_j = \frac{1}{n_j} \sum_{k=1}^{n_j} \widetilde{Q}^k(\pi \times \delta_v).$$

On peut en extraire une sous-suite convergente vers η, et l'on a, d'après la continuité de $\bar{f}$,

$$\gamma_1^q = \lim_j \frac{1}{n_j} \int \log \|S_{n_j}(\omega)v\| \, dQ_\pi(\omega) = \eta(\bar{f}).$$

Par construction de η, on a $\widetilde{Q}\eta = \eta$, donc $\eta \in M_V^1$, et, puisque $\sup_{\eta \in M_V^1} I_\eta^1 \le \gamma_1^q$,

$$\gamma_1^q = \eta(\bar{f}) = I_\eta^1 = \sup_{n \in M_V^1} I_\eta^1.$$

La preuve de l'inégalité $I_\eta^1 \le \gamma_1^q$ reste valide en remplaçant $P^1(V)$ par $P^2(V)$, et fournit $I_\eta^2 \le \gamma_1^q + \gamma_2^q$ pour tout $\eta \in M_V^2$.

Pour obtenir l'inégalité inverse, on raisonne comme plus haut, en observant que si m_2 est la mesure invariante par rotation sur $P^2(V) \subset P(\wedge^2 V)$, on a aussi, pour tout $g \in \mathrm{GL}(V)$, comme plus haut,

$$\log \| \wedge^2 g\| \le \int \log \|(\wedge^2 g)\ (v \wedge v')\|\, dm_2(v \wedge v') + c',$$

où c' est une nouvelle constante. On peut donc trouver comme ci-dessus $v \wedge v' \in P^2(V)$ tel que

$$\gamma_1^q + \gamma_2^q = \lim_j \frac{1}{n_j} \int \log \| \wedge^2 S_{n_j}(\omega) v \wedge v'\|\, dQ_\pi(\omega).$$

D'où, comme ci-dessus, l'existence d'une mesure $\eta \in M_V^2$ telle que $I_\eta^2 = \gamma_1^q + \gamma_2^q$. □

Lemme 7.2. *Soit E un espace métrique compact, P un noyau Markovien continu sur E, et préservant $C(E)$, f une fonction continue sur E, M le convexe des probabilités P-invariantes. Alors*

$$\sup_{x \in E} \frac{1}{n} \sum_{k=0}^{n-1} P^k f(x) \to \sup_{\eta \in M} \int f\, d\eta.$$

Preuve. Soit J l'ensemble des valeurs d'adhérence des suites

$$\frac{1}{n} \sum_{k=0}^{n-1} P^k f(x_n)$$

avec $x_n \in E$. On va montrer que l'enveloppe convexe de J est égale à l'ensemble $\{\eta(f) : \eta \in M\}$.

Clairement J est fermé. Si la suite $\frac{1}{n_k} \sum_{i=0}^{n_k-1} P^i f(x_{n_k})$ converge vers $\ell \in \mathbb{R}$, on peut considérer la suite de probabilités $\eta_k = \frac{1}{n_k} \sum_{i=0}^{n_k-1} P^i \delta_{x_{n_k}}$ et en extraire une sous-suite convergente vers $\eta \in M$. Alors par continuité de f, on a

$$\eta(f) = \lim_k \frac{1}{n_k} \sum_{i=0}^{n_k-1} (P^i f)(x_{n_k}) = \ell.$$

Inversement, si $\eta \in M$ est donnée et ergodique, en appliquant le théorème de Birkhoff à la suite $\frac{1}{n} \sum_{i=0}^{n-1} P^i f(x)$, pour η-presque tout x on a

$$\lim_n \frac{1}{n} \sum_{i=0}^{n-1} P^i f(x) = \eta(f),$$

d'où l'existence de x tel que $\eta(f)$ soit limite de la suite $\frac{1}{n}\sum_{i=0}^{n-1} P^i f(x)$. Si η n'est pas ergodique, elle est barycentre de probabilités ergodiques, et $\eta(f)$ est barycentre de valeur d'adhérences des suites $\frac{1}{n}\sum_{i=0}^{n-1} P^i f(x_n)$. Donc $\eta(f)$ appartient à l'enveloppe convexe de J. D'après ce qui précède, tout point de l'enveloppe convexe de J appartient aussi à M, d'où l'assertion voulue.

Puisque J est fermé, l'enveloppe convexe de J est un intervalle fermé $[a, b]$. On a donc :

$$b = \lim_n \frac{1}{n} \sup_{x\in E} \sum_{i=0}^{n-1} P^i f(x) = \sup_{\eta\in M} \int f\, d\eta. \qquad \square$$

Lemme 7.3. *Avec les hypothèses du Lemme* 7.1, *si* $\Gamma_{\alpha(\mu)}$ *est irréductible et proximal sur* $P(V)$, *on a* $\gamma_1^q = I_\eta^1$ *pour tout* $\eta \in M_V^1$.

Preuve. On peut supposer η ergodique, afin de prouver l'énoncé. Le théorème de Birkhoff appliqué à la transformation $\widetilde{\theta}$ sur $X \times P(V) \times \Omega$, à la fonction $f(x, v, \omega) = \log \|a_1(\omega)v\|$ et à la probabilité $\widetilde{\eta} = \int \delta_{(x,v)} \times Q_x\, d\eta(x, v)$ donne l'existence de $v \in P(V)$ tel que, sur un ensemble de Q_π-probabilité positive, on ait

$$I_\eta^1 = \lim_n \frac{1}{n} \log \|S_n(\omega)v\|.$$

Le Théorème 6.7 montre que, sous les conditions du Lemme, on a Q_π-p.p.

$$\lim_n \frac{1}{n} \log \frac{\|S_n(\omega)v\|}{\|S_n(\omega)\|} = \lim_n \frac{1}{n} \log |z(\omega)(v)| = 0,$$

puisque $z(\omega)(v) \neq 0$ Q_π-p.p. Donc

$$I_\eta^1 = \lim_n \frac{1}{n} \int \log \|S_n(\omega)\|\, dQ_x(\omega) = \gamma_1^q. \qquad \square$$

Rappelons le lemme suivant, prouvé en [34] (cf. aussi [25]).

Lemme 7.4. *Soit* (E, T, λ) *un système dynamique avec* λ *probabilité* T*-invariante,* f *une fonction* λ*-intégrable. Si l'on a* λ*-p.p.* $\lim_n \sum_{k=0}^n f \circ T^k = -\infty$, *alors on a* $\int f\, d\lambda < 0$.

Lemme 7.5. *On considère le cocycle* σ_1 *sur* $\mathrm{GL}(V) \times P(V)$ (*resp.,* σ_2 *sur* $\mathrm{GL}(V) \times P^2(V)$). *Alors on a*

$$\gamma_1^q = \lim_n \frac{1}{n} \sup_{x,v} \int \log \sigma_1[S_n(\omega), v]\, dQ_x(\omega),$$

$$\gamma_1^q + \gamma_2^q = \lim_n \frac{1}{n} \sup_{x, v\wedge v'} \int \log \sigma_2[S_n(\omega), v \wedge v']\, dQ_x(\omega).$$

La suite de fonctions $\frac{1}{n}\int \log \sigma_1[S_n(\omega), v]\, dQ_x(\omega)$ *converge uniformément sur* $X \times P(V)$ *vers* γ_1^q *dès que* $\Gamma_{\alpha(\mu)}$ *est irréductible et proximal sur* $P(V)$.

Preuve. On considère la chaîne de Markov $\widetilde{Q} = \widetilde{Q}^1$ sur $X \times P(V)$ et la fonction $f(x, v) = \int \log \sigma_1(a, v) q(x, a)\, d\mu(a)$. Alors on a comme plus haut

$$\sum_{k=0}^{n-1} \widetilde{Q}^k f(x, v) = \int \log \sigma_1(S_n(\omega), v)\, dQ_x(\omega).$$

Il suffit donc d'appliquer le Lemme 7.2 pour obtenir la première partie de l'énoncé. Les deux assertions finales découlent des Lemmes 7.1, 7.2 et 7.3 par une argumentation classique (cf. [16]). □

7.2 Le plus grand exposant est de multiplicité 1

Théorème 7.6. *Soit $(X, q \otimes \mu)$ un système Markovien transitif, α une application Borélienne de X dans* $\mathrm{GL}(V)$ *telle que les intégrales*

$$\int \log \|\alpha(a)\| q(a)\, d\mu(a), \quad \int \log \|\alpha(a^{-1})\| q(a)\, d\mu(a)$$

soient finies. Alors, si $\Gamma_{\alpha(\mu)}$ agit de façon irréductible et proximale sur $P(V)$, on a la relation

$$\gamma_2^q - \gamma_1^q = \sup_{\eta \in M_V^{1,2}} I_\eta^{1,2} < 0.$$

Preuve. D'après le Lemme 7.1, on a

$$\gamma_2^q - \gamma_1^q = \sup_{\eta \in M_V^2} I_\eta^2 - 2 \sup_{\eta' \in M_V^1} I_{\eta'}^1. \tag{7.2}$$

Si $\eta \in M_V^2$ est donnée, le convexe compact des probabilités sur $P^{1,2}(V)$ de projection η est $\widetilde{Q}^{1,2}$-invariant. D'après le théorème de Markov–Kakutani, il existe donc $\widetilde{\eta} \in M_V^{1,2}$ de projection η sur $P^2(V)$. On a alors $I_{\widetilde{\eta}}^2 = I_\eta^2$, et la projection $\widetilde{\eta}_1$ de $\widetilde{\eta}$ sur $P^1(V)$ est aussi $\widetilde{Q}^1$-invariante : $I_{\widetilde{\eta}}^1 = I_{\widetilde{\eta}}^1 \leq \sup_{\eta' \in M_V^1} I_{\eta'}^1$.

On a donc

$$\gamma_2^q - \gamma_1^q \leq \sup_{\eta \in M_V^2} (I_\eta^2 - 2I_{\widetilde{\eta}_1}^1) = \sup_{\widetilde{\eta} \in M_V^{1,2}} (I_{\widetilde{\eta}}^2 - 2I_{\widetilde{\eta}}^1) = \sup_{\widetilde{\eta} \in M_V^{1,2}} I_{\widetilde{\eta}}^{1,2}.$$

Inversement, observons que d'après le Lemme 7.3, la proximalité de Γ_μ entraîne que pour tout $\eta \in M_V^1$, on a

$$I_\eta^1 = \sup_{\eta' \in M_V^1} I_{\eta'}^1 = \gamma_1^q.$$

Donc

$$\gamma_2^q - \gamma_1^q = \sup_{\eta \in M_V^2} I_\eta^2 - 2\gamma_1^q.$$

Si $\widetilde{\eta} \in M_V^{1,2}$ se projette sur $\eta \in M_V^2$ et sur $\widetilde{\eta}_1 \in M_V^1$, on a $\gamma_1^q = I_{\widetilde{\eta}_1}^1 = I_{\widetilde{\eta}}^1$, d'où finalement

$$\gamma_2^q - \gamma_1^q = \sup_{\widetilde{\eta} \in M_V^{1,2}} (I_{\widetilde{\eta}}^2 - 2I_{\widetilde{\eta}}^1) = \sup_{\widetilde{\eta} \in M_V^{1,2}} I_{\widetilde{\eta}}^{1,2}.$$

Pour obtenir $\gamma_2^q - \gamma_1^q < 0$, il suffit de voir que, pour chaque $\eta \in M_V^{1,2}$ on a $I_\eta^{1,2} < 0$, car $M_V^{1,2}$ est compact et $I_\eta^{1,2}$ est fonction continue de η. On considère alors la transformation $\widetilde{\theta}$ sur $X \times P^{1,2}(V) \times \Omega$, définie par $\widetilde{\theta}(x, \xi, \omega) = (a_1 x, \alpha(a_1)\xi, \theta\omega)$, la fonction $g(\xi, \omega) = \log \sigma(a_1(\omega), \xi)$, et la mesure invariante $\widetilde{\eta} = \int \delta_{(x,\xi)} \times Q_x \, d\eta(x, \xi)$.

Comme plus haut on a

$$\log \sigma(S_n(\omega), \xi) = \sum_0^{n-1} g \circ \widetilde{\theta}^k(\xi, \omega).$$

Puisque la condition (i–p) est satisfaite, le Théorème 6.4 implique que la suite $S_n^t.m$ converge Q_π-p.p. vers une mesure de Dirac.

La première partie du Lemme 6.6 entraîne alors que $\log \sigma(S_n(\omega), \xi)$ converge vers $-\infty$ si l'origine de ξ n'appartient pas à l'hyperplan $z^*(\omega)$ de $P(V)$. Supposons que l'ensemble des $(v, \omega) \in P(V) \times \Omega$ tels que $v \in z^*(\omega)$ soit de $\widetilde{\eta}$-mesure positive. Alors il existerait (v, x) tel que $z(\omega)$ appartiendrait avec Q_x-probabilité positive à l'hyperplan v^* de $P(V^*)$ définie par v. On aurait donc $\nu_x(v^*) > 0$, ce qui contredit le Théorème 6.1. On a donc $\widetilde{\eta}$-p.p.

$$\lim_n \sum_{k=0}^{n-1} g \circ \widetilde{\theta}^k = -\infty,$$

et le Lemme 7.4 donne $\widetilde{\eta}(g) = I_{\widetilde{\eta}}^{1,2} < 0$. □

Corollaire 7.7. *Avec les notations et hypothèses du Théorème* 7.6*, la suite*

$$\frac{1}{n} \sup_{x,v,v'} \int \log \frac{\delta(S_n(\omega).v, S_n(\omega).v')}{\delta(v, v')} \, dQ_x(\omega)$$

converge vers $\gamma_2^q - \gamma_1^q < 0$.

Preuve. Montrons d'abord que les valeurs d'adhérence de la suite

$$I_n = \frac{1}{n} \int \log \frac{\delta[S_n(\omega).v, S_n(\omega).v']}{\delta(v_n, v_n')} \, dQ_{x_n}(\omega)$$

sont majorées par $\gamma_2 - \gamma_1$. On écrit I_n sous la forme

$$I_n = \frac{1}{n} \int \log \wedge^2 \|S_n(\omega)(v_n \wedge v_n')\| \, dQ_{x_n}(\omega) - \frac{1}{n} \int \log \|S_n(\omega)v_n\| \, dQ_{x_n}(\omega) - \frac{1}{n} \int \log \|S_n(\omega)v_n'\| \, dQ_{x_n}(\omega),$$

où on peut supposer que chacun des trois termes converge. D'après le Lemme 7.5 on a $\lim_n I_n = I^2_{\eta_2} - 2I^1_{\eta_1}$, où η_2, η_1 sont $\widetilde{Q}$-invariantes. D'après les Lemmes 7.3 et 7.5 $I_{\eta_1} = \gamma_1^q$, $I_{\eta_2} \leq \gamma_1^q + \gamma_2^q$, d'où $\lim_n I_n \leq \gamma_2^q - \gamma_1^q$.

Inversement, soient v_n, v'_n, x_n des suites telles que, en vertu du Lemme 7.5

$$\gamma_1^q + \gamma_2^q = \lim_n \frac{1}{n} \int \log \|(\wedge^2 S_n)(\omega)(v_n \wedge v'_n)\| \, dQ_{x_n}(\omega).$$

Alors, pour ce choix de v_n, v'_n, on a

$$\lim_n I_n = (\gamma_1^q + \gamma_2^q) - 2\gamma_1^q = \gamma_2^q - \gamma_1^q.$$

Puisque

$$I_n \leq \sup_{x,v,v'} \frac{1}{n} \int \log \frac{\delta(S_n(\omega).v, S_n(\omega).v')}{\delta(v, v')} \, dQ_x(\omega),$$

la conclusion découle de l'observation initiale. □

Corollaire 7.8 (cf. Corollaire 2.9). *Soit A un ensemble fini, π une mesure de Gibbs sur $X = A^{[0,-\infty]}$ définie par un potentiel Höldérien, α une application de A dans* $\mathrm{GL}(V)$ *telle que le semi-groupe engendré par les matrices $\alpha(a)$ $(a \in A)$ possède la propriété* (i–p). *Alors les deux premiers exposants caractéristiques du produit de matrices stationnaire $S_{-n}(\omega) = \alpha(a_{-n+1}) \ldots \alpha(a_0)$ sont différents.*

Preuve. Avec les notations de la Section 2, on a $q(x, a) = e^{h(xa)}$, où $x \in X, a \in A$ et h est le potentiel associé à π. Alors $\mu(a) = 1$ si $a \in A$, et $(X, q \otimes \mu)$ est bien un système Markovien au sens de la Section 2, puisque h est Höldérienne. L'opérateur Q est ici donné par

$$Q\varphi(x) = \sum_{a \in A} e^{h(xa)} \varphi(xa),$$

et le système $(X, q \otimes \mu)$ est clairement transitif. Le semi-groupe $\Gamma_{\alpha(\mu)} \subset \mathrm{GL}(V)$ est alors le semi-groupe engendré par les matrices $\alpha(a)$ et il vérifie la condition (i–p) par hypothèse. Il suffit donc d'appliquer le Théorème 7.6. □

Corollaire 7.9 (Théorème 2.10). *Soit $(X, q \otimes \mu)$ un système Markovien transitif, α une application Borélienne de X dans* $\mathrm{GL}(V)$ *telle que $\Gamma_{\alpha(\mu)}$ vérifie la condition* ($\mathcal{F}$–i–p). *Alors les exposants caractéristiques γ_i^q $(\gamma_1^q \geq \gamma_2^q \geq \cdots \geq \gamma_d^q)$ du produit $S_n(\omega) = \alpha(a_n) \ldots \alpha(a_1)$ satisfont l'inégalité $\gamma_1^q > \gamma_2^q > \cdots > \gamma_d^q$. De plus, si δ est la distance naturelle sur $\mathcal{F}$,*

$$\lim_n \frac{1}{n} \sup_{x,b,b'} \int \log \frac{\delta[S_n(\omega).b, S_n(\omega).b']}{\delta(b, b')} \, dQ_x(\omega) = \sup_{1 \leq i < d} (\gamma_{i+1}^q - \gamma_i^q) < 0. \quad (7.3)$$

Preuve. On considère le produit de matrices $\widehat{S}_n(\omega)$ associé à $S_n(\omega)$ dans $\mathrm{GL}(\widehat{V})$. Il est facile de voir, en utilisant les poids de la représentation de $\mathrm{GL}(V)$ dans $\widehat{V}$ [43], que la

différence des deux premiers exposants de $\widehat{S}_n(\omega)$ vaut $\sup_{1\le i\le d}(\gamma_{i+1}^q - \gamma_i^q)$. On peut alors appliquer le corollaire précédent puisque $\Gamma_{\widehat{\alpha}(\mu)}$ satisfait la condition (i–p) dans $\widehat{V}$: $\sup_{1\le i<d}(\gamma_{i+1} - \gamma_i) < 0$, c'est-à-dire $\gamma_1^q > \gamma_2^q > \cdots > \gamma_d^q$, d'où (7.3). □

8 Propriétés de contraction et d'isolation spectrale

On considère ici d'abord un espace métrique compact (X, d) et l'action du semi-groupe $C(X, X)$ sur X. On se donne un noyau Markovien continu $q \otimes \mu$ sur $X \times C(X, X)$ et l'on considère la chaîne de Markov de noyau Q sur X définie par

$$Q\varphi(x) = \int \varphi(ax) q(x, a)\, d\mu(a).$$

On note par Q_x les mesures Markoviennes sur $C(X, X)^{\mathbb{N}}$ définies par $q \otimes \mu$ comme en Section 2 et on supposera au Théorème 8.7 que Q_x est continu en variation, donc que $(X, q \otimes \mu)$ est un système Markovien. On étudie les propriétés d'isolation spectrale de l'opérateur défini par Q sur des espaces de fonctions , d'abord dans le cas général (Théorème 8.7), puis dans les cas $X = P(V)$, $X = \mathcal{F}$ (Théorèmes 8.8, 8.9 et Corollaires 8.10, 8.11), μ étant alors portée par $\mathrm{PGL}(V)$, et q étant égal à q^s ou bien $q^{(s)}$. On introduit pour cela des coefficients de contraction en moyenne $\rho(\varepsilon)$ et c (cf. ci-dessous) associés à Q. En utilisant les informations obtenues en Section 7, on précise ces résultats dans le cas $X = P(V)$ ou $X = \mathcal{F}$, $Q = Q^s$ étant alors le noyau de la chaîne de Markov associée à P^s (ou $Q^{(s)}$ associée à $P^{(s)}$). Puis on étend les résultats au cas s complexe généralisant ainsi les résultats d'isolation spectrale de [36] qui correspondent au cas s imaginaire pur. On obtient ainsi le Théorème 2.2.

Pour alléger les notations générales, on omettra l'indice q dans la plupart des écritures.

8.1 Opérateurs Markoviens contractants en moyenne

Définition 8.1. Le noyau $q \otimes \mu$ étant fixé, on note, pour $\varepsilon > 0$,

$$\rho_n(\varepsilon) = \sup_{x,y\in X} \int \frac{d^\varepsilon[s_n(\omega)x, s_n(\omega)y]}{d^\varepsilon(x, y)}\, dQ_x(\omega), \quad \rho(\varepsilon) = \inf_n (\rho_n(\varepsilon))^{1/n}.$$

On pose aussi

$$c_n = \sup_{x,y\in X} \int \log \frac{d[s_n(\omega)x, s_n(\omega)y]}{d(x, y)}\, dQ_x(\omega), \quad c = \inf_n \frac{1}{n} c_n$$

et

$$[a] = \sup_{x,y\in X} \frac{d(ax, ay)}{d(x, y)}.$$

Ces nombres sont bien définis comme éléments de $\overline{\mathbb{R}_+}$ et on étudie d'abord leurs propriétés générales.

Proposition 8.2. *Supposons qu'il existe $\beta > 0$ avec $\int [a]^\beta q(a)\, d\mu(a) < +\infty$. Alors la fonction $\log \rho_n(\varepsilon)$ est définie sur $[0, \beta]$ et convexe. Sa dérivée à droite en 0 vaut c_n. La suite $(\rho_n(\varepsilon))^{1/n}$ converge sur $[0, \beta]$ vers $\rho(\varepsilon)$, et c_n converge vers $c = \inf c_n/n$.*

La preuve découlera des trois lemmes.

Lemme 8.3. *Avec les notations de la Définition* 8.1*, on a*

$$\rho_{m+n}(\varepsilon) \le \rho_m(\varepsilon)\rho_n(\varepsilon), \quad c_{m+n} \le c_m + c_n, \quad \rho(\varepsilon) < +\infty, \quad c < +\infty.$$

Preuve. On a

$$\rho_1(\varepsilon) = \sup_{x,y} \int \frac{d^\varepsilon(ax, ay)}{d^\varepsilon(x, y)} q(x, a)\, d\mu(a) \le \sup_{x,y} \left(\int \frac{d^\beta(ax, ay)}{d^\beta(x, y)} q(x, a)\, d\mu(a) \right)^{\varepsilon/\beta}$$

d'après l'inégalité de Hölder, d'où $\rho_1(\varepsilon) \le \left(\int [a]^\beta q(a)\, d\mu(a) \right)^{\varepsilon/\beta} < +\infty$.

Notons, pour abréger,

$$\overline{\sigma}(g, \xi) = \frac{d(gx, gy)}{d(x, y)}, \quad \text{où} \quad \xi = (x, y) \in X \times X,\ x \ne y,\ g \in C(X, X),$$

et observons que

$$\overline{\sigma}(gh, \xi) = \overline{\sigma}(g, h\xi)\overline{\sigma}(h, \xi).$$

Si θ désigne le décalage sur $\Omega = S_\mu^{\mathbb{N}}$, on a donc

$$\overline{\sigma}(s_{m+n}, \xi) = \overline{\sigma}(s_m \circ \theta^n, s_n\xi)\overline{\sigma}(s_n, \xi).$$

La définition de Q_x donne alors

$$\int \overline{\sigma}^\varepsilon(s_{m+n}(\omega), \xi)\, dQ_x(\omega) = \left[\int \overline{\sigma}^\varepsilon(s_n(\omega), \xi)\, dQ_x(\omega) \right] \times \left[\int \overline{\sigma}^\varepsilon(s_m(\omega), s_n\xi)\, dQ_{s_n x}(\omega) \right],$$

d'où $\rho_{m+n}(\varepsilon) \le \rho_m(\varepsilon)\rho_n(\varepsilon)$ et $\rho(\varepsilon) \le \rho_1(\varepsilon) < +\infty$.

On a aussi

$$\int \log \overline{\sigma}(s_{m+n}(\omega), \xi)\, dQ_x(\omega) \le \int dQ_x(\omega) \int \log \overline{\sigma}(s_m(\omega), s_n\xi)\, dQ_{s_n x}(\omega) + \int \log \overline{\sigma}(s_n(\omega), \xi)\, dQ_x(\omega),$$

d'où $c_{m+n} \le c_m + c_n$.

Enfin,

$$\begin{aligned} c_1 &= \sup_{x,y\in X} \int \log \frac{d(ax,ay)}{d(x,y)} q(x,a)\, d\mu(a) \\ &\le \frac{1}{\beta} \sup_{x,y\in X} \log \int \frac{d^\beta(ax,ay)}{d^\beta(x,y)} q(x,a)\, d\mu(a) \\ &\le \frac{1}{\beta} \log \int [a]^\beta q(a)\, d\mu(a) < +\infty, \end{aligned}$$

ce qui donne $c \le c_1 < +\infty$. □

Lemme 8.4. *La fonction* $\log \rho_n(\varepsilon)$ *est convexe sur* $[0, \beta]$ *et sa dérivée à droite en* 0 *vaut* c_n.

Preuve. Elle découle des estimations élémentaires qui suivent.

Pour x et y fixés dans X, la fonction de ε égale à

$$\log \int \frac{d^\varepsilon[s_n(\omega)x, s_n(\omega)y]}{d^\varepsilon(x,y)}\, dQ_x(\omega)$$

est convexe d'après l'inégalité de Hölder. Il en est donc de même de

$$\log \rho_n(\varepsilon) = \sup_{x,y\in X} \log \int \frac{d^\varepsilon[s_n(\omega)x, s_n(\omega)y]}{d^\varepsilon(x,y)}\, dQ_x(\omega),$$

d'où la première assertion.

Pour alléger les écritures notons

$$\begin{aligned} \varphi_{x,y}(\omega) &= \frac{d[s_n(\omega)x, s_n(\omega)y]}{d(x,y)}, \\ g_{x,y}(\varepsilon) &= \int \varphi^\varepsilon_{x,y}(\omega)\, dQ_x(\omega) \\ \log \rho_n(\varepsilon) &= \sup_{x,y} \log g_{x,y}(\varepsilon). \end{aligned}$$

D'après l'inégalité de Hölder, la fonction $\log g_{x,y}(\varepsilon)$ est convexe, et

$$\frac{g'_{x,y}(\varepsilon)}{g_{x,y}(\varepsilon)} = \int \log \varphi_{x,y}(\omega) \varphi^\varepsilon_{x,y}(\omega)\, dQ_x(\omega).$$

La dérivée de $\log g_{x,y}(\varepsilon)$ étant croissante, on a, pour $\varepsilon > 0$,

$$\int \log \varphi_{x,y}(\omega)\, dQ_x(\omega) \le \frac{\log g_{x,y}(\varepsilon)}{\varepsilon} \le \int \varphi^\varepsilon_{x,y}(\omega) \log \varphi_{x,y}(\omega)\, dQ_x(\omega).$$

On a donc

$$c_n \le \frac{\log \rho_n(\varepsilon)}{\varepsilon} \le \sup_{x,y} \int \varphi^\varepsilon_{x,y}(\omega) \log \varphi_{x,y}(\omega)\, dQ_x(\omega)$$

De plus,

$$\varphi_{x,y}^{\varepsilon}(\omega) \le [s_n(\omega)]^{\varepsilon} \le [s_n(\omega)]^{\beta}, \quad \varphi_{x,y}^{\varepsilon}(\omega)\ \log \varphi_{x,y}(\omega) \le [s_n(\omega)]^{\beta},$$

car $\varepsilon < \beta$.

En raison de la forme de Q_x, la condition $\int [a]^{\beta} q(a)\, d\mu(a) < +\infty$ de l'énoncé implique alors la continuité des intégrales au second membre. Pour obtenir la dernière assertion, il suffit donc de voir que si $x_k, y_k \in X$ et $\varepsilon_k > 0$ vérifient $\lim_k \varepsilon_k = 0$, $\lim_k x_k = x$, $\lim_k y_k = y$, la suite

$$\int \varphi_{x_k,y_k}^{\varepsilon_k}(\omega) \log \varphi_{x_k,y_k}(\omega)\, dQ_{x_k}(\omega)$$

a une limite majorée par c_n. Par convergence dominée on a,

$$\lim_k \int \varphi_{x_k,y_k}^{\varepsilon_k}(\omega) \log \varphi_{x_k,y_k}(\omega)\, dQ_{x_k}(\omega) = \int \log \varphi_{x,y}(\omega)\, dQ_x(\omega) \le c_n,$$

car $\lim_k \varphi_{x_k,y_k}^{\varepsilon_k}(\omega) = 1$, d'où la dernière assertion. □

Lemme 8.5. *La suite $\rho_n(\varepsilon)^{1/n}$ converge vers la fonction convexe $\rho(\varepsilon)$. La suite $\frac{1}{n}\rho_n'(0^+)$ converge vers $\rho'(0^+) = c$.*

Preuve. La convergence de $\rho_n(\varepsilon)^{1/n}$ vers $\rho(\varepsilon)$ découle classiquement de la relation $\rho_{m+n}(\varepsilon) \le \rho_m(\varepsilon)\rho_n(\varepsilon)$ montrée au Lemme 8.3. La fonction $\rho(\varepsilon)$ est convexe comme limite de fonctions convexes. On a par convexité de $\rho_n(\varepsilon)^{1/n}$

$$\frac{\rho_n'(0^+)}{n} \le \frac{\rho_n(\varepsilon)^{1/n} - 1}{\varepsilon},$$

d'où $c \le (\rho(\varepsilon) - 1)/\varepsilon$. Par définition de $\rho(\varepsilon)$

$$\frac{\rho(\varepsilon) - 1}{\varepsilon} \le \frac{\rho_n(\varepsilon)^{1/n} - 1}{\varepsilon},$$

d'où $\limsup_{\varepsilon \to 0^+} \frac{\rho(\varepsilon)-1}{\varepsilon} \le \frac{\rho_n'(0^+)}{n}$, et à la limite

$$\limsup_{\varepsilon \to 0^+} \frac{\rho(\varepsilon) - 1}{\varepsilon} = c \le \frac{\rho(\varepsilon) - 1}{\varepsilon}.$$

Soit enfin

$$c = \lim_{\varepsilon \to 0^+} \frac{\rho(\varepsilon) - 1}{\varepsilon}.$$

□

Remarque 8.6. On peut voir que, si ρ ne s'annule pas identiquement au voisinage de 0, alors $\rho(\varepsilon)$ est continue en 0 et $(\rho_n(\varepsilon))^{1/n}$ converge uniformément vers $\rho(\varepsilon)$.

Preuve de la Proposition 8.2. Les deux premières assertions découlent du Lemme 8.4. On a, d'après le Lemme 8.5, la convergence de $(\rho_n(\varepsilon))^{1/n}$ vers $\rho(\varepsilon)$, qui est convexe.

On a, d'après ce lemme la convergence des dérivées à droite en 0: $\lim_n c_n/n = c = \inf_n c_n/n$. □

Théorème 8.7. *Soit $(X, q \otimes \mu)$ un système Markovien. Supposons qu'il existe $\beta > 0$ et $C_q \geq 0$, tels que $\int [a]^\beta q(a)\, d\mu(a) < +\infty$, et que $\|Q_x - Q_y\| \leq C_q\, d^\beta(x, y)$ et notons*

$$\rho_n(\varepsilon) = \sup_{x,y \in X} \int \frac{d^\varepsilon[s_n(\omega)x, s_n(\omega)y]}{d^\varepsilon(x, y)}\, dQ_x(\omega),$$
$$\rho(\varepsilon) = \inf_n [\rho_n(\varepsilon)]^{1/n}.$$

Alors, pour $\varepsilon \in [0, \beta]$ et pour tout $\varphi \in H_\varepsilon(X)$, $n \in \mathbb{N}$ on a

$$[Q^n\varphi]_\varepsilon \leq \rho_n(\varepsilon)[\varphi]_\varepsilon + C_q|\varphi|_\infty.$$

Supposons vérifiée la condition de contraction en moyenne suivante:

$$c = \inf_n \frac{1}{n} \sup_{x,y} \int \log \frac{\delta[s_n(\omega)x, s_n(\omega)y]}{\delta(x, y)}\, dQ_x(\omega) < 0.$$

Alors il existe $\beta_0 > 0$ tel que pour tout $\varepsilon \in]0, \beta_0[$ et tout n assez grand on ait

$$\rho(\varepsilon) \leq (\rho_n(\varepsilon))^{1/n} < 1.$$

L'opérateur Q sur $H_\varepsilon(X)$ est $\rho(\varepsilon)$-quasi-compact, et 1 en est valeur spectrale isolée. De plus, le sous-espace spectral correspondant est de dimension 1, dès que les solutions continues de l'équation $Qh = h$ sont réduites aux constantes. Enfin, sous cette condition il y a une unique probabilité Q-invariante sur X.

Preuve. Puisque

$$Q^n\varphi(x) - Q^n\varphi(y) = \int (\varphi(s_n(\omega)x) - \varphi(s_n(\omega)y))\, dQ_x(\omega) + \int \varphi(s_n(\omega)y)\, d(Q_x - Q_y)(\omega),$$

on a

$$|Q^n\varphi(x) - Q^n\varphi(y)| \leq [\varphi]_\varepsilon \int \delta^\varepsilon(s_n(\omega)x, s_n(\omega)y)\, dQ_x(\omega) + |\varphi|_\infty \|Q_x - Q_y\|$$

et

$$[Q^n\varphi]_\varepsilon \leq \rho_n(\varepsilon)[\varphi]_\varepsilon + C_q|\varphi|_\infty$$

par définition de $\rho_n(\varepsilon)$ et C_q, d'où la première assertion.

La Proposition 8.2 montre que $(\rho_n(\varepsilon))^{1/n}$ converge vers $\rho(\varepsilon)$, fonction convexe sur $[0, \beta]$, dont la dérivée à droite en 0 vaut $c < 0$. Puisque $\rho(0) = 1$, il existe $\beta_0 > 0$ telle que $\rho(\varepsilon) < 1$ sur $]0, \beta_0[$. La convergence précédente montre alors que, pour tout ε fixé dans $]0, \beta_0[$ et n assez grand, on a $\rho(\varepsilon) \leq (\rho_n(\varepsilon))^{1/n} < 1$. Puisque $Q1 = 1$, le théorème de Ionescu–Tulcea et Marinescu [31], précisé en [29] montre que 1 est

valeur spectrale isolée et que, de plus, Q est $\rho(\varepsilon)$-quasi-compact. Les deux dernières assertions découlent alors de ce théorème. □

8.2 Opérateurs de transfert sur l'espace projectif

On développe ici pour les opérateurs P^s de Section 4 les conséquences des résultats de Section 7.

Théorème 8.8 (cf. Théorème 2.2 et Corollaire 2.4). *Soit μ une probabilité sur $\mathrm{GL}(V)$ telle que Γ_μ agisse de façon irréductible et proximale sur $P(V)$. On suppose qu'il existe $\sigma > 0$ tel que*

$$\int \|g\|^\sigma \, d\mu(g) < +\infty, \quad \int \|g\|^\sigma \log|\det g| \, d\mu(g) > -\infty,$$

et on considère, pour $s \in [0, \sigma[$, l'opérateur P^s sur $P(V)$ défini par

$$P^s\varphi(v) = \int \|gv\|^s \varphi(g.v) \, d\mu(g).$$

On pose

$$k(s) = \lim_n \left(\int \|g\|^s \, d\mu^n(g) \right)^{1/n}.$$

Alors, l'opérateur P^s de $H_\varepsilon(P(V))$ admet pour rayon spectral $k(s)$. Il possède une unique fonction propre e_s telle que $e_s > 0$ sur $P(V)$, et une unique mesure propre ν^s normalisées ($|e_s|_\infty = 1$, $\nu^s(e_s) = 1$), correspondant à $k(s)$. On a, pour $\varepsilon > 0$ assez petit (dépendant de s), la décomposition en somme directe

$$P^s = k(s)(\nu^s \otimes e_s + U^s),$$

où U^s commute avec le projecteur $p^s = \nu^s \otimes e_s$ ($U^s p^s = p^s U^s = 0$), et est de rayon spectral strictement inférieur à 1. La fonction $k(s)$ est analytique sur $[0, \sigma[$ et $k'(0) = \gamma_1$. Si $\sigma = +\infty$, on a

$$\lim_{s \to +\infty} \frac{\log k(s)}{s} = \gamma^\infty(S_\mu).$$

Soit Q^s l'opérateur Markovien sur $P(V)$ défini par

$$Q^s\varphi = \frac{1}{k(s)e_s} P^s(e_s\varphi),$$

et notons

$$\rho_n^s(\varepsilon) = \sup_{v,v'} \int \frac{\delta^\varepsilon(S_n(\omega).v, S_n(\omega).v')}{\delta^\varepsilon(v, v')} \, dQ_v^s(\omega),$$

$$\rho^s(\varepsilon) = \lim_n (\rho_n^s(\varepsilon))^{1/n}.$$

Alors pour $s \in [0, \sigma[$ *fixé et* ε *assez petit, on a* $\rho^s(\varepsilon) < 1$. *L'opérateur* Q^s *de* $H_\varepsilon[P(V)]$ *est* $\rho^s(\varepsilon)$*-quasi-compact, sa résolvante admet un pôle simple pour* $\lambda = 1$, *et elle est holomorphe dans le domaine* $\{\lambda \in \mathbb{C} : |\lambda| > 1 - r,\ \lambda \neq 1\}$ *pour* $r > 0$ *assez petit*.

Preuve. Les assertions concernant $e_s, k(s)$ découlent du Théorème 4.1. Pour obtenir la décomposition spectrale de P^s, on note, comme en Section 4,

$$q^s(v, g) = \frac{1}{k(s)} \frac{e_s(g.v)}{e_s(v)} \|gv\|^s,$$

de sorte que

$$Q^s\varphi(v) = \int q^s(v, g)\varphi(g.v)\, d\mu(g).$$

On montre que $q^s(v, g) = q(v, g)$ satisfait les hypothèses du Théorème 8.7.

D'après le Théorème 4.1, les seules solutions continues de l'équation $Q^s h = h$ sont les constantes. De plus, le Théorème 4.17 fournit l'inégalité :

$$\|Q^s_v - Q^s_{v'}\| \leq c_s\, \delta^{\bar{s}}(v, v').$$

Ici on a, avec $c, c' > 0$,

$$c'\|g\|^s \leq q^s(g) = \sup_v q^s(v, g) \leq c\|g\|^s.$$

Pour tout $s \in [0, \sigma[$, on a donc

$$\int \log\|g\|\, q^s(g)\, d\mu(g) < +\infty, \quad \int \log\|g^{-1}\|\, q^s(g)\, d\mu(g) < +\infty,$$

en utilisant la relation $g^{-1} = (\det g)^{-1} \left(\wedge^{d-1} g\right)^t$ et la condition de l'énoncé. Le noyau Q^s_v satisfait alors les propriétés requises en Sections 6 et 7.

Puisque Γ_μ agit de façon irréductible et proximale sur $P(V)$, on peut appliquer le Corollaire 7.7 au noyau Markovien Q^s:

$$c^s = \lim_n \sup_{v, v'} \frac{1}{n} \int \log \frac{\delta[S_n(\omega).v,\, S_n(\omega).v']}{\delta(v, v')}\, d\, Q^s_v(\omega) = \gamma^s_2 - \gamma^s_1 < 0.$$

Les conditions du Théorème 8.7 ci-dessus sont alors remplies et celui-ci donne $\rho^s(\varepsilon) < 1$, pour $\varepsilon < 0$ assez petit. D'après ce théorème, la résolvante $(\lambda I - Q^s)^{-1}$ de Q^s est méromorphe pour $|\lambda| > \rho^s(\varepsilon)$, et 1 en est un pôle simple. En particulier $(\lambda I - Q^s)^{-1}$ admet un pôle sur le cercle $|\lambda| = 1$, si et seulement si l'équation $Q^s\varphi = e^{i\alpha}\varphi$ a une solution dans $H_\varepsilon(P(V))$, avec $\alpha \in \mathbb{R}$. Cette condition équivaut à $P^s(\varphi e_s) = e^{i\alpha}k(s)(\varphi e_s)$, et implique $e^{i\alpha} = 1$, $\varphi = 1$ d'après le Théorème 4.1, d'où les propriétés annoncées de la résolvante.

Enfin si π^s désigne l'unique probabilité Q^s-invariante sur $P(V)$ (voir Corollaire 4.23), on a la décomposition $Q^s = \pi^s \otimes 1 + U^s$ avec $U^s(\pi^s \otimes 1) = (\pi^s \otimes 1)U^s =$

0 et $\lim_n \|(U^s)^n\|^{1/n} < 1$. Revenant à P^s, on obtient dans $H_\varepsilon(P(V))$

$$P^s = k(s)(\nu^s \otimes e_s + U_1^s),$$

où $U_1^s(\nu^s \otimes e_s) = (\nu^s \otimes e_s)U_1^s = 0$ et $\lim_n \|(U_1^s)^n\|^{1/n} < 1$.

Pour montrer l'analyticité de $k(s)$ pour $s \in [0, \sigma[$ on considère l'opérateur P^z avec $z \in \mathbb{C}$, $z = s + it$, et t petit, et on montre d'abord l'analyticité de P^z, pour $\Re z < \sigma$. Si γ est un lacet de $\mathbb{C}$ contenu dans le demi-plan $\Re z < \sigma$ et $\varphi \in H_\varepsilon(P(V))$ on a

$$\int_\gamma P^z\varphi(v)\,dz = \iint_{G\times\gamma} \varphi(g.v)\|gv\|^z\,d\mu(g)dz = \int_G \varphi(g.v)\,d\mu(g) \int_\gamma \|gv\|^z\,dz = 0$$

par holomorphie de $z \mapsto \|gv\|^z$. D'autre part, si γ est un petit cercle de centre $k(s) \in \mathbb{C}$, on a, puisque $k(s)$ est un pôle simple de la fonction $\lambda \mapsto (\lambda I - P^s)^{-1}$ (cf. [13, p. 566])

$$\nu^s \otimes e_s = \frac{1}{2i\pi} \int_\gamma (\lambda I - P^s)^{-1}\,d\lambda.$$

Puisque P^z dépend analytiquement de z, la fonction $(\lambda I - P^z)^{-1}$ possède aussi un unique pôle à l'intérieur du disque défini par γ, si z est proche de s. L'opérateur P^z admet alors $k(z)$ pour valeur spectrale isolée, et le projecteur correspondant est, avec des notations analogues

$$\nu^z \otimes e_z = \frac{1}{2i\pi} \int_\gamma (\lambda I - P^z)^{-1}\,d\lambda.$$

La formule ci-dessus montre que ce projecteur dépend analytiquement de z. La relation $P^z(\nu^z \otimes e_z) = k(z)(\nu^z \otimes e_z)$ montre donc que $k(z)$ est analytique au voisinage de s, d'où l'analyticité de $k(s)$ pour $s \in [0, \sigma[$.

La preuve de la Proposition 8.2 montre que la fonction convexe

$$\log k(s) = \lim_n \frac{1}{n} \log \int \|g\|^s\,d\mu^n(g)$$

admet une dérivée à droite en 0 donnée par

$$\gamma_1 = \lim_n \frac{1}{n} \int \log\|g\|\,d\mu^n(g).$$

Enfin, la dernière assertion du Théorème 4.17 entraîne

$$\lim_{s\to+\infty} \frac{\log k(s)}{s} = \gamma^\infty(S_\mu)$$

si $\sigma = +\infty$. □

Théorème 8.9 (cf. Corollaire 2.4). *Avec les notations du Théorème* 8.8*, considérons pour $z = s + it$ l'opérateur Q^z sur $P(V)$ défini par*

$$Q^z\varphi = \frac{1}{k(s)e_s} P^{s+it}(e_s\varphi).$$

Alors, pour ε assez petit, il existe une constante $C_{z,n}(\varepsilon)$ telle que l'on ait, pour tout n et tout $z \in \mathbb{C}$ avec $0 \leq \Re z < \sigma$

$$[(Q^z)^n \varphi]_\varepsilon \leq \rho_n^s(\varepsilon)[\varphi]_\varepsilon + C_{z,n}(\varepsilon)|\varphi|_\infty.$$

L'opérateur Q^z est $\rho^s(\varepsilon)$-quasi-compact, et sa résolvante admet un pôle simple pour $\lambda = 1$, si $t = 0$. Si $t \neq 0$, le rayon spectral de Q^z sur $H_\varepsilon(P(V))$ est strictement inférieur à 1.

Preuve. On a

$$\begin{aligned} Q^{s+it}\varphi(v) &= \frac{1}{k(s)} \int \varphi(g.v)\|gv\|^{s+it} \frac{e_s(g.v)}{e_s(v)}\, d\mu(g) \\ &= \int \|gv\|^{it} q^s(v,g)\varphi(g.v)\, d\mu(g) \end{aligned}$$

et

$$\begin{aligned} |Q^{s+it}\varphi(v) - Q^{s+it}\varphi(v')| \leq \int \big|\|gv^{it} - \|gv'\|^{it}\big| q^s(v,g)|\varphi(g.v)|\, d\mu(g) \\ + \int |q^s(v,g)\varphi(g.v) - q^s(v',g)\varphi(g.v')|\, d\mu(g). \end{aligned}$$

La deuxième intégrale est majorée, comme plus haut, par

$$\delta^\varepsilon(v,v')(\rho_1^s(\varepsilon)[\varphi]_\varepsilon + C_s(\varepsilon)|\varphi|_\infty).$$

Pour majorer la première intégrale on observe que $|e^{iu} - 1| \leq 2|u|^\varepsilon$, ce qui donne

$$\begin{aligned} \big|\|gv\|^{it} - \|gv'\|^{it}\big| &\leq 2|t|^\varepsilon \big|\log\|gv\| - \log\|gv'\|\big|^\varepsilon, \\ &\leq 2|t|^\varepsilon \left(\sup_{x\in P(V)} \frac{1}{\|gx\|}\right)^\varepsilon \|g(v-v')\|^\varepsilon \\ &\leq |t|^\varepsilon \sup_{x\in P(V)} \left|\frac{\|g\|}{\|gx\|}\right|^\varepsilon \delta^\varepsilon(v,v'). \end{aligned}$$

Prenant ε assez petit, on obtient en intégrant

$$\begin{aligned} |Q^{s+it}\varphi(v) - Q^{s+it}\varphi(v')| &\leq (\rho_1^s(\varepsilon)\,[\varphi]_\varepsilon + C_{s,1}(\varepsilon)|\varphi|_\infty + C_t(\varepsilon)|\varphi|_\infty)\delta^\varepsilon(v,v'), \\ [Q^{s+it}\varphi]_\varepsilon &\leq \rho_1^s(\varepsilon)[\varphi]_\varepsilon + C_{z,1}(\varepsilon)|\varphi|_\infty. \end{aligned}$$

Ce calcul s'étend à $(Q^{s+it})^n\varphi$ et donne

$$[(Q^{s+it})^n\varphi]_\varepsilon \leq \rho_n^s(\varepsilon)[\varphi]_\varepsilon + C_{z,n}(\varepsilon)|\varphi|_\infty,$$

d'où, comme dans la preuve du Théorème 8.7, pour tout z fixé avec $\Re z \in [0,\sigma[$, la méromorphie de la résolvante $(\lambda I - Q^{s+it})^{-1}$ pour $|\lambda| > \rho^s(\varepsilon)$. Si $t = 0$, alors d'après le Théorème 8.8 la résolvante $(\lambda I - Q^s)^{-1}$ a un pôle simple en $\lambda = 1$. Si le rayon spectral de Q^{s+it} vaut 1, la méromorphie de la résolvante sur le cercle $|\lambda| = 1$ et la finitude de la dimension des sous-espaces spectraux correspondants impliquent que

l'équation $Q^{s+it}\varphi = e^{i\alpha}\varphi$, avec $\varphi \in H_\varepsilon(P(V))$ et $\alpha \in \mathbb{R}$, a une solution. Comme en Section 4, cette équation implique

$$e^{i\alpha}\varphi(v) = \|gv\|^{it}\varphi(g.v) \quad \forall\, g \in \Gamma_\mu,\ v \in L_\Gamma.$$

D'après le Théorème 3.19, elle ne peut être satisfaite pour $t \neq 0$, car Γ_μ vérifie la condition (i–p). Donc, pour $t \neq 0$, le rayon spectral de Q^{s+it} est strictement inférieur à 1. □

Dans le cas $(s) \in \mathbb{C}^d$, soit $(\sigma) \in \mathbb{R}_+^d$ et notons $[0, (\sigma)[$ le produit des intervalles $[0, \sigma_i[$ $(0 \leq i \leq d)$. Alors les preuves des Théorèmes 8.8 et 8.9 donnent les analogues suivants de ces Corollaires, en remplaçant $P(V)$ par l'unique orbite compacte de $\mathrm{GL}(V)$ dans $P(\widehat{V})$. Dans les preuves des dernières assertions des Corollaires 8.10 et 8.11, on utilise le Théorème 3.26 au lieu du Théorème 3.19.

Corollaire 8.10. *Soit μ une probabilité sur $\mathrm{GL}(V)$ telle que Γ_μ vérifie la condition ($\mathcal{F}$–i–p), que $\int \| \wedge g\|^{(\sigma)}\, d\mu(g) < +\infty$, et considérons, pour $(s) \in [0, (\sigma)[$, l'opérateur $P^{(s)}$ défini par*

$$P^{(s)}\varphi(b) = \int \sigma^{(s)}(g, b)\varphi(g.b)\, d\mu(g).$$

On pose

$$k[(s)] = \lim_n \left[\int \| \wedge g\|^{(s)}\, d\mu^n(g)\right]^{1/n}.$$

Alors, l'opérateur $P^{(s)}$ de $H_\varepsilon(\mathcal{F})$ admet pour rayon spectral $k[(s)]$ et possède une unique fonction propre $e_{(s)}$, une unique mesure propre $\nu^{(s)}$ correspondant à $k[(s)]$ et normalisées. On a $e_{(s)} > 0$ sur $\mathcal{F}$ et, pour $\varepsilon > 0$ assez petit, la décomposition en somme directe

$$P^{(s)} = k[(s)]\big[\nu^{(s)} \otimes e_{(s)} + U^{(s)}\big],$$

où $U^{(s)}$ commute avec le projecteur $p^{(s)} = \nu^{(s)} \otimes e_{(s)}$ $(U^{(s)}p^{(s)} = p^{(s)}U^{(s)} = 0)$ et est de rayon spectral strictement inférieur à 1.

La fonction $\log k[(s)]$ est analytique et strictement convexe sur $[0, (\sigma)[$, et sa k-ième dérivée partielle à l'origine vaut $\gamma_1 + \gamma_2 + \cdots + \gamma_k$.

Si $k[(s)]$ est définie sur $\mathbb{R}_+^d$, on a

$$\lim_{t \to +\infty} \frac{1}{t} \log k[(ts)] = \gamma_{(s)}^{\infty}(S_\mu).$$

Soit $Q^{(s)}$ l'opérateur Markovien sur $\mathcal{F}$ associé à $P^{(s)}$, et notons

$$\rho_n^{(s)}(\varepsilon) = \sup_{b, b' \in \mathcal{F}} \int \frac{\delta^\varepsilon[S_n(\omega).b,\, S_n(\omega).b']}{\delta^\varepsilon(b, b')}\, dQ_b(\omega),$$

$$\rho^{(s)}(\varepsilon) = \lim_n [\rho_n^{(s)}(\varepsilon)]^{1/n}.$$

Alors, pour $(s) \in [0, (\sigma)[$ *et* ε *assez petit, on a* $\rho^{(s)}(\varepsilon) < 1$. *De plus, l'opérateur* $Q^{(s)}$ *est* $\rho^{(s)}(\varepsilon)$*-quasi-compact. Si* $\mathrm{Zc}(\Gamma_\mu) \supset \mathrm{SL}(V)$, *il existe* $r > 0$ *tel que sa résolvante soit holomorphe dans* $\{\lambda \in \mathbb{C} : |\lambda| > 1 - r, \lambda \neq 1\}$.

Corollaire 8.11. *Avec les notations et hypothèses du Corollaire* 8.10, *considérons, pour* $(z) = (s + it)$, *l'opérateur* $Q^{(z)}$ *sur* $\mathcal{F}$ *défini par*

$$Q^{(z)}\varphi = \frac{1}{k[(s)]e_{(s)}}\, P^{(s+it)}[\varphi e_{(s)}].$$

Supposons que $\int \| \wedge g\|^{(\sigma)}\, d\mu(g) < +\infty$ *et fixons* $(s) \in [0, (\sigma)[$ *et* $\varepsilon > 0$ *assez petit pour que* $\rho^{(s)}(\varepsilon) < 1$. *Alors, pour tout* $n \in \mathbb{N}$, *il existe une constante* $C_{(z),n}(\varepsilon)$ *telle que, pour toute* $\varphi \in H_\varepsilon[P(\mathcal{F})]$ *on ait*

$$[[Q^{(z)}]^n \varphi]_\varepsilon \leq \rho_n^{(s)}(\varepsilon)[\varphi]_\varepsilon + C_{(z),n}(\varepsilon)|\varphi|_\infty.$$

Pour n assez grand, on a $\rho_n^{(s)}(\varepsilon) < 1$.

L'opérateur $Q^{(z)}$ *est* $\rho^{(s)}(\varepsilon)$*-quasi-compact. Si* $t = 0$, *alors sa résolvante admet un pôle simple pour* $\lambda = 1$. *Si* $t \neq 0$ *et si l'adhérence de Zariski de* Γ_μ *contient* $\mathrm{SL}(V)$, *elle est holomorphe pour* $|\lambda| > 1 - r$ *avec* $r > 0$.

Remerciements. Les auteurs remercient vivement les organisateurs des journées de probabilités et mécanique statistique de Bologne (Février 2000) et ceux du semestre sur les marches aléatoires à l'Institut Schrödinger de Vienne (Juin 2001), où les principaux résultats de ce travail ont été exposés, ainsi que M. Babillot pour d'importantes remarques.

References

[1] M. Babillot, Théorie du renouvellement pour des chaînes semi-markoviennes transientes, Ann. Inst. H. Poincaré Probab. Statist. 24 (1988), 507–569.

[2] W. Ballmann, Lectures on Spaces of Non Positive Curvature. With an appendix by Misha Brin, DMV Sem. 25, Birkhäuser, Basel 1995.

[3] Y. Benoist, Propriétés asymptotiques des groupes linéaires, Geom. Funct. Anal. 7 (1997), 1–47.

[4] Y. Benoist, Propriétés asymptotiques des groupes linéaires. II, in: Analysis on Homogeneous Spaces and Representation Theory of Lie Groups, Okayama–Kyoto (1997), Adv. Stud. Pure Math. 26, Math. Soc. Japan, Tokyo 2000, 33–48.

[5] C. Bonatti and M. Viana, Lyapunov exponents with multiplicity 1 for deterministic products of matrices, preprint (2001).

[6] P. Bougerol and J. Lacroix, Products of Random Matrices with Applications to Schrödinger Operators, Progress in Probability and Statistics 8, Birkhäuser, Boston, MA, 1985.

[7] L. Breiman, Probability, Addison–Wesley, Reading, MA, 1968.

[8] A. Broise, F. Dal'bo and M. Peigné, Études Spectrales d'Opérateurs de Transfert et Applications, Astérisque 238, Société Mathématique de France, Paris 1996.

[9] A. Broise-Alamichel and Y. Guivarc'h, Exposants caractéristiques de l'algorithme de Jacobi–Perron et de la transformation associée, Ann. Inst. Fourier (Grenoble) 51 (2001), 565–686.

[10] R. Carmona and J. Lacroix, Spectral Theory of Random Schrödinger Operators, Probab. Appl., Birkhäuser, Boston, MA, 1990.

[11] J.-P. Conze and Y. Guivarc'h, Densité d'orbites d'actions de groupes linéaires et propriétés d'équidistribution de marches aléatoires, in: Rigidity in Dynamics and Geometry (Cambridge, 2000), Springer-Verlag, Berlin 2002, 39–76.

[12] W. Doeblin and R. Fortet, Sur les chaînes à liaisons complètes, Bull. Soc. Math. France 65 (1937), 132–148.

[13] N. Dunford and J.T. Schwarz, Linear Operators. I. General Theory. With the assistance of W. G. Bade and R. G. Bartle, Pure Appl. Math. 7, Interscience, New York 1958.

[14] W. Feller, An Introduction to Probability Theory and its Applications, vol. II, second edition, Wiley, New York 1971.

[15] A. Furman, Random walks on groups and random transformations, in: Handbook of Dynamical Systems, Vol. 1A, North-Holland, Amsterdam 2002, 931–1014.

[16] H. Furstenberg, Noncommuting random products, Trans. Amer. Math. Soc 108 (1963), 337–428.

[17] H. Furstenberg, Translation-invariant cones of functions on semi-simple Lie groups, Bull. Amer. Math. Soc. 71 (1965), 271–326.

[18] H. Furstenberg, Boundary theory and stochastic processes on homogeneous spaces, in: Harmonic Analysis on Homogeneous Spaces (Proc. Sympos. Pure Math., XXVI, Williams Coll., Williamstown, Mass., 1972), Amer. Math. Soc., Providence, RI, 1973, 193–229.

[19] I. Goldsheid and Y. Guivarc'h, Zariski closure and the dimension of the Gaussian law of the product of random matrices. I, Probab. Theory Related Fields 105 (1996), 109–142.

[20] I. Ya. Goldsheid and G. A. Margulis, Lyapunov exponents of a product of random matrices, Russian Math. Surveys 44 (5) (1989), 11–71.

[21] I. Ya. Goldsheid, S. A. Molchanov and L. A. Pastur, A pure point spectrum of the stochastic one-dimensional Schrödinger operator, Funct. Anal. Appl. 11 (1977), 1–8.

[22] Y. Guivarc'h, Produits de matrices aléatoires et applications aux propriétés géométriques des sous-groupes du groupe linéaire, Ergodic Theory Dynam. Systems 10 (1990), 483–512.

[23] Y. Guivarc'h, A spectral gap property for transfer operators, in: Harmonic Functions on Trees and Buildings (New York, 1995), 129–131, Contemp. Math. 206, Amer. Math. Soc., Providence, RI, 1997, 129–131.

[24] Y. Guivarc'h and A. Raugi, Products of random matrices: convergence theorems, in: Random Matrices and their Applications (Brunswick, Maine, 1984), Contemp. Math. 50, Amer. Math. Soc., Providence, RI, 1986, 31–54.

[25] Y. Guivarc'h and A. Raugi, Frontière de Furstenberg, propriétés de contraction et théorèmes de convergence, Z. Wahrsch. Verw. Gebiete 69 (1985), 187–242.

[26] Y. Guivarc'h and A. Raugi, Propriétés de contraction d'un semi-groupe de matrices inversibles. Coefficients de Liapunoff d'un produit de matrices aléatoires indépendantes, Israel J. Math. 65 (1989), 165–196.

[27] Y. Guivarc'h and R. Shah, Asymptotic properties of convolution operators and limits of triangular arrays on locally compact groups, Trans. Amer. Math. Soc. 184 (2004).

[28] Y. Guivarc'h and A. Starkov, Orbits of linear group actions, random walks on homogeneous spaces and toral automorphisms, Ergodic Theory Dynam. Systems 24 (2004), 1–36.

[29] H. Hennion, Sur un théorème spectral et son application aux noyaux lipschitziens, Proc. Amer. Math. Soc. 118 (1993), 627–634.

[30] H. Hennion and L. Hervé, Limit Theorems for Markov Chains and Stochastic Properties of Dynamical Systems, Lecture Notes in Math. 1766, Springer-Verlag, Berlin 2001.

[31] C.T. Ionescu-Tulcea and G. Marinescu, Théorie ergodique pour des classes d'opérations non complètement continues, Ann. of Math. (2) 52 (1950), 140–147.

[32] V. A. Kaimanovich, The Poisson boundary of groups with hyperbolic properties, Ann. of Math. (2) 152 (2000), 659–692.

[33] H. Kesten, Random difference equations and renewal theory for products of random matrices, Acta Math. 131 (1973), 207–248.

[34] H. Kesten, Sums of stationary sequences cannot grow slower than linearly, Proc. Amer. Math. Soc. 49 (1975), 205–211.

[35] H. Kesten, M. V. Kozlov and F. Spitzer, A limit law for random walks in a random environment, Compositio Math. 30 (1975), 145–168.

[36] E. Le Page, Théorèmes limites pour les produits de matrices aléatoires, in: Probability Measures on Groups (Oberwolfach, 1981), Lecture Notes in Math. 928, Springer-Verlag, Berlin 1982, 258–303.

[37] E. Le Page, Théorie du renouvellement pour les produits de matrices aléatoires. Équations aux coefficients aléatoires, in: Séminaires de Probabilités Rennes 1983, Publ. Sém. Math., Univ. Rennes I, Rennes 1983, 1–116.

[38] A. V. Letchikov, Products of unimodular independant random matrices, Russian Math. Surveys 51 (1) (1996), 49–96.

[39] G. A. Margulis, Discrete Subgroups of Semi-simple Lie Groups, Ergeb. Math. Grenzgeb. (3) 17, Springer-Verlag, Berlin 1991.

[40] A. L. Onishchik and E. B. Vinberg, Lie Groups and Algebraic Groups, Springer Series in Soviet Mathematics, Springer-Verlag, Berlin 1990.

[41] V. I. Oseledets, A multiplicative ergodic theorem: Lyapunov characteristic exponents for dynamical systems, Trans. Moscow Math. Soc. 19 (1968), 197–231.

[42] W. Parry and M. Pollicott, Zeta Functions and the Periodic Orbit Structure of Hyperbolic Dynamics, Astérisque 187–188, Société Mathématique de France, Paris 2000.

[43] G. R. Prasad, *R*-regular elements in Zariski-dense subgroups, Quart. J. Math. Oxford Ser. (2) 45 (1994), 541–545.

[44] Ya. G. Sinai, Gibbs measures in ergodic theory, Russian Math. Surveys 27 (4) (1972), 21–69.

[45] F. Spitzer, Principles of Random Walk, The University Series in Higher Mathematics, Van Nostrand, Princeton 1964.

[46] B. Weiss, Subshifts of finite type and sofic systems, Monatsh. Math. 77 (1973), 462–474.

Yves Guivarc'h, IRMAR, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex, France

E-mail: yves.guivarch@univ-rennes1.fr

Émile Le Page, LMAM, Université de Bretagne–Sud, Campus de Tohannic, BP 573, 56017 Vannes, France

E-mail: Emile.Le-Page@univ-ubs.fr

An introduction to the Stochastic Loewner Evolution

Gregory F. Lawler

Abstract. The stochastic Loewner evolution (SLE), introduced by Oded Schramm, is a one parameter family of conformally invariant measures on curves in the plane. It gives the continuum limit of a number of lattice models in statistical physics at criticality. This expository paper gives an introduction to SLE for probabilists. After discussing the necessary results from complex variables and the (deterministic) Loewner equation, I define SLE and discuss some of its properties as well as some recent results.

Contents

1 Introduction 262

2 Loewner differential equations 263
- 2.1 Some complex analysis . . . 263
- 2.2 Chordal Loewner equation . . . 264
- 2.3 Radial Loewner equation . . . 269
- 2.4 Whole plane Loewner equation . . . 271

3 Stochastic Loewner Evolution 272
- 3.1 Chordal version . . . 272
- 3.2 Radial and whole plane versions . . . 274
- 3.3 Locality . . . 275
- 3.4 Restriction property . . . 278
  - 3.4.1 Chordal. . . . 278
  - 3.4.2 Radial. . . . 281
- 3.5 Some discrete thoughts . . . 282

4 Calculating exponents 284
- 4.1 Crossing exponent for chordal SLE . . . 284
- 4.2 Crossing exponent for radial SLE . . . 287
- 4.3 Brownian excursions and exponents . . . 288
- 4.4 Computing Brownian exponents . . . 290

1 Introduction

There are a number of lattice models in statistical physics, e.g., random walks, self-avoiding random walks, percolation, loop-erased walks (uniform spanning trees), Potts models, that are expected to behave conformally in two dimensions. More specifically, when the value of a parameter reaches a critical value (at which a phase transition occurs in the system), the configurations are expected to have a continuum limit that is conformally invariant. This idea was used by a number of theoretical physicists to produce exact values for critical exponents (see, for example, [4, Chapter 9]). From a mathematical perspective, these arguments present a challenge for they assume that in the continuum limit there will be a "conformally invariant field". Even if one can prove the existence of the limit, it is not clear how to use this to prove results about random curves.

There has been a lot of work recently by mathematicians in a number of areas (combinatorics, complex variables, mathematical physics, probability), to try to prove the physicists' predictions on these exponents. The basic approach is to try to understand the continuum limits of these processes as conformally invariant measures on paths (or clusters) in the complex plane. One example is already well known to probabilists: the limit of simple random walk is Brownian motion which is conformally invariant in two dimensions.

Oded Schramm [21] combined an old idea of Loewner's in complex variables with some randomness to create a process he called the *stochastic Loewner evolution* (SLE). This one parameter family of clusters growing in $\mathbb{C}$ is conjectured to be the limit of many of these models; a number of these conjectures have been proved and more will probably be proved in the next few years. Also, analysis of the SLE with the aid of stochastic calculus allows one to determine many critical exponents for the process as well as for other processes. For example, Lawler, Schramm, and Werner [9, 10, 11, 12] used SLE to calculate the "intersection exponents" for Brownian motion. This last result has consequences for the geometry of Brownian motion; for example, it proved a conjecture of Mandelbrot that the outer boundary of a planar Brownian path has Hausdorff dimension $4/3$.

The purpose of this paper is to give an introduction to SLE for probabilists. The paper has three parts: Section 2 reviews some facts from complex variables and then discusses the (deterministic) Loewner equation. There are three closely related Loewner equations: chordal (for clusters growing from a boundary towards a boundary point), radial (for clusters growing from a boundary to an interior point), and whole plane (for clusters growing from one point to infinity). The next section defines the stochastic Loewner evolution SLE_κ which is the Loewner equation "driven" by a Brownian motion with variance parameter κ. Some of the properties are also discussed, in particular, the notion of "locality" for $\kappa = 6$ and "restriction" for $\kappa = 8/3$. We end this section with a short discussion of locality and restriction in terms of discrete models. The final section shows how one can compute exponents for SLE and how

this can be used to find exponents for Brownian motion. Here we use the locality property of SLE_6, but we do not use the results about restriction for $SLE_{8/3}$.

As mentioned before, the definition of SLE is due to Schramm. The results in this paper from Section 3.3 on, unless otherwise attributed, come from a collaboration of Schramm, Wendelin Werner, and myself. For more details on this work, see [9, 10, 11, 12, 15].

2 Loewner differential equations

2.1 Some complex analysis

If D, D' are two simply connected domains in $\mathbb{C}$ other than the whole plane and $w \in D, z \in D'$, the Riemann mapping theorem [1, Theorem 6.1] states that there is a unique conformal transformation of D onto D' with $f(w) = z$, $f'(w) > 0$. For this reason, much of the study of simply connected domains in $\mathbb{C}$ reduces to the study of univalent (conformal, one-to-one) mappings f on the unit disk $\mathbb{D}$. By translation and scaling, one can assume that $f(0) = 0$, $f'(0) = 1$. The set $\mathcal{S}$ of such scaled univalent functions on the unit disk has been studied extensively in complex analysis (see [3]).

Call a compact $K \subset \mathbb{C}$ containing the origin a *hull* if K is connected and larger than a single point. In this case, the complement of K, considered as a domain in the Riemann sphere $\mathbb{C} \cup \{\infty\}$, is simply connected. For any such hull, there is a unique conformal transformation

$$g_K : \mathbb{C} \setminus K \to \mathbb{C} \setminus \bar{\mathbb{D}}$$

with $g_K(\infty) = \infty$ and $g_K'(\infty) > 0$, which we can expand at infinity

$$g_K(z) = e^{-a(K)} z + O(1), \quad z \to \infty.$$

This defines the quantity $a(K)$, which is monotone in K, called the *logarithmic capacity* of K. Note that the closed disk of radius R about the origin has logarithmic capacity $\log R$. To every hull K we can associate the function $f_K \in \mathcal{S}$,

$$f_K(z) = \frac{e^{a(K)}}{g_K^{-1}(1/z)}.$$

This gives a one-to-one correspondence between $\mathcal{S}$ and the set of hulls of logarithmic capacity zero. The Koebe One-Quarter Theorem [3, Theorem 2.3] states that the image of any function in $\mathcal{S}$ includes the open disk of radius $1/4$ about the origin. Translated to capacities, this implies

$$\log[\mathrm{rad}(K)/4] \le a(K) \le \log \mathrm{rad}(K), \tag{2.1}$$

where $\mathrm{rad}(K) = \sup\{|z| : z \in K\}$. By considering the function $f(z) = [z + (1/z) - 2]/4$, we can see that $a([-R, 0]) = \log(R/4)$. The Bieberbach conjecture, proved

by de Branges, says that coefficients of any $f \in \mathcal{S}$,

$$f(z) = z + a_2 z^2 + a_3 z^3 + \cdots ,$$

satisfy $|a_n| \leq n$. A weaker estimate, $|a_n| \leq en$, goes back to Littlewood (see [3, Theorem 2.8]). The first proof of $|a_3| \leq 3$ is due to Loewner, using the Loewner differential equation. The proof uses the correspondence between hulls and functions in $\mathcal{S}$. To prove something about all functions in $\mathcal{S}$, one can analyze what happens as one moves from one function to another, i.e., as the hull grows. The Loewner differential equation describes the time evolution of a growing hull.

There is a constant C such that if $f \in \mathcal{S}$,

$$|f(z) - z| \leq C\ |z|\ |f(z)|, \quad |z| < 1.$$

For $|z| \geq 1/2$ this follows from the Koebe One-Quarter Theorem. For $|z| \leq 1/2$, Littlewood's estimate implies $|f(z) - z| \leq C' |z|^2$, and then the Koebe estimate gives $|z| \leq 4|f(z)|$. Translating to g_K, we get

$$|e^{a(K)}\ w - g_K^{-1}(w)| \leq C\ e^{a(K)}, \quad |w| > 1. \tag{2.2}$$

If D is a simply connected domain and z_1, z_2, z_3, z_4 are four distinct points ordered counterclockwise on the boundary of D , then there is a unique L such that there is a conformal transformation f of D onto the rectangle

$$R_L = \{x + iy : 0 < x < L; 0 < y < \pi\}$$

with $f(z_1) = \pi i$, $f(z_2) = 0$, $f(z_3) = L$, $f(z_4) = L + \pi i$. We call this L the π-*extremal distance* between arcs $[z_1, z_2]$ and $[z_3, z_4]$ in D. This is clearly a conformal invariant. In the complex variables literature, the quantity L/π is called the extremal distance or extremal length and π/L is called the module or conformal modulus. (We choose to use π-extremal distance rather than extremal distance in order to make formulas nicer; see, e.g., (4.6)).

2.2 Chordal Loewner equation

Let $\mathcal{H} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$ denote the upper half plane. We call a compact set $K \subset \overline{\mathcal{H}}$ such that $K = \overline{K \cap \mathcal{H}}$ and $\mathcal{H} \setminus K$ is simply connected a *hull* in $\mathcal{H}$ By the Riemann mapping theorem, there exist conformal transformations $g_K : \mathcal{H} \setminus K \to \mathcal{H}$ that map infinity to infinity. Since g_K maps $\mathbb{R} \setminus K$ into $\mathbb{R}$, the Schwarz reflection principle [1, Theorem 6.24] can be used to extend g_K to the complement of

$$K^* = \{z : z \in K \text{ or } \bar{z} \in K\}.$$

At infinity, g_K has an expansion

$$g_K(z) = bz + a_0 + a_1 z^{-1} + \cdots , \quad b > 0,\ a_j \in \mathbb{R}.$$

(One can see this by considering the expansion of $\psi(z) = 1/g_K(1/z)$ about the origin. Since ψ locally maps reals to reals, the coefficients in its expansion are real. It is easy to check that $b > 0$.) For convenience we choose the unique g_K that satisfies the *hydrodynamic normalization*,

$$\lim_{z\to\infty} g_K(z) - z = 0,$$

i.e., we choose $b = 1$, $a_0 = 0$ so that

$$g_K(z) = z + \frac{a_1}{z} + O\left(\frac{1}{|z|^2}\right), \quad z \to \infty.$$

We call the coefficient $a_1 = a_1(K)$ the *capacity* of K (in $\mathcal{H}$ from infinity). If $r > 0$, it is easy to check that $g_{rK}(z) = rg_K(z/r)$; this implies the scaling relation

$$a_1(rK) = r^2 a_1(K). \tag{2.3}$$

We can give a probabilistic interpretation of a_1. Let B_t denote a complex Brownian motion starting at $z \in \mathcal{H} \setminus K$, and let

$$\tau = \tau_K = \inf\{t : B_t \in \mathbb{R} \cup K\}.$$

Since $\mathrm{Im}(g_K(z) - z)$ is a bounded harmonic function on $\mathcal{H} \setminus K$, and $\mathrm{Im}(g_K)$ vanishes on K, the optional sampling theorem implies

$$\mathrm{Im}(g_K(z) - z) = \mathbf{E}^z[\mathrm{Im}(g(B_0) - B_0)] = -\mathbf{E}^z[\mathrm{Im}(B_\tau)].$$

Hence,

$$a_1 = \lim_{y\to\infty} y\, \mathbf{E}^{iy}[\mathrm{Im}(B_\tau)].$$

We note that $a_1(K)$ is not the same as $a(K^*)$; for example, if K is a line segment at the origin with small angle θ, then $a_1(K)$ is much smaller than $a(K^*)$.

Suppose $J \subset K$ are two hulls with corresponding transformations g_J, g_K, and let L be the hull $L = \overline{g_J(K \setminus J)}$. Then it is easy to check that $g_K = g_L \circ g_J$. By considering the expansions at infinity, we get the relation

$$a_1(K) = a_1(J) + a_1(L).$$

Suppose that $\gamma : [0, \infty) \to \bar{\mathcal{H}}$ is a simple (no self-intersections) continuous curve with $\gamma(0) = x \in \mathbb{R}$, $\gamma(0, \infty) \subset \mathcal{H}$ and $a_1(\gamma[0, t]) \to \infty$ as $t \to \infty$. For each t, let $g_t = g_{\gamma[0,t]}$ be the corresponding transformation. We assume that the parameterization of γ has been chosen so that

$$a_1(\gamma[0, t]) = 2t. \tag{2.4}$$

(It is not difficult to show that the capacity of $\gamma[0, t]$ is continuous and increasing in t, so we can always choose a parameterization so that (2.4) holds.) We state the next proposition for simple curves, but as we will see, the result also holds for many curves with self-intersections. However, self-crossings are not allowed.

Proposition 2.1. *Let $U_t = g_t(\gamma(t))$. Then for each $z \in \mathcal{H}$, $g_t(z)$ satisfies the Loewner differential equation,*

$$\partial_t g_t(z) = \frac{2}{g_t(z) - U_t}, \quad g_0(z) = z.$$

The equation is valid up to a time $T_z \in (0, \infty]$ that can be characterized as the first time t such that $g_t(z) \in \mathbb{R}$.

Proof. We will show that

$$\partial_t \mathrm{Im}[g_t(z) - z] = \partial_t \mathrm{Im}[g_t(z)] = \mathrm{Im}\left[\frac{2}{g_t(z) - U_t}\right]. \tag{2.5}$$

The Cauchy–Riemann equations then imply that

$$\mathrm{Re}[g_t(z) - z] = \int_0^t \mathrm{Re}\left[\frac{2}{g_s(z) - U_s}\right] ds + c_t,$$

where c_t is independent of z. Since $g_t(z) - z \to 0$ as $z \to \infty$, $c_t = 0$.

It suffices to prove (2.5) when $t = 0$ and $U_0 = 0$, i.e., if we write $z = x + iy$ we need to show that

$$\partial_t \mathrm{Im}[g_t(z) - z]|_{t=0} = -\frac{2y}{x^2 + y^2}. \tag{2.6}$$

Let B_t be a Brownian motion, $\epsilon_t = \sup\{|\gamma(s)|, 0 \le s \le t\}$, σ_t the first time B_t reaches $\partial[\epsilon_t \mathbb{D}]$, and τ_t the first time it reaches $\mathbb{R} \cup \gamma[0, t]$. Then

$$\mathrm{Im}[z - g_t(z)] = \mathbf{E}^z[\mathrm{Im}(B_{\tau_t})] = \mathbf{P}^z\{\sigma_t < \tau_t\}\, \mathbf{E}^z[\mathrm{Im}(B_{\tau_t}) \mid \sigma_t < \tau_t].$$

If we fix t, let $z = iy$, and let $y \to \infty$, the left hand side is asymptotic to $2t/y$; the first term on the right is asymptotic to $(4/\pi)\epsilon_t/y$ (one can check this, for example, by using separation of variables to give a series form of the exact probability). Hence the second term approaches $2t\pi/4\epsilon_t$. Now if we fix z, w, and let $t \to 0+$, we can see from a coupling argument that

$$\mathbf{E}^z[\mathrm{Im}(B_{\tau_t}) \mid \sigma_t < \tau_t] \sim \mathbf{E}^w[\mathrm{Im}(B_{\tau_t}) \mid \sigma_t < \tau_t] = \frac{(2t)\pi}{4\epsilon_t}[1 + o(1)].$$

But for fixed z, as $t \to 0+$,

$$\mathbf{P}^z\{\sigma_t < \tau_t\} \sim \frac{(4/\pi)\epsilon_t y}{x^2 + y^2}.$$

Hence,

$$\mathrm{Im}[g_t(z) - z] = -\frac{2ty}{x^2 + y^2} + o(t),$$

which establishes (2.6). □

In this proposition, we started with the curve $\gamma(t)$ (or, equivalently, the conformal transformations g_t) and derived a function U_t. We can also go the other way. Suppose that for each time t we have a nonzero Borel measure μ_t on the real line. Assume that $\mu_t(\mathbb{R})$ is uniformly bounded and for every $t_0 < \infty$ there is a compact interval that contains the supports of μ_t, $0 \le t \le t_0$. Suppose also that $t \mapsto \mu_t$ is piecewise continuous. Continuity is in the sense of the weak topology on measures; in particular, if μ_t is continuous at t, then for every $z \in \mathcal{H}$ the function

$$t \longmapsto \int \frac{\mu_t(dx)}{z-x}$$

is continuous in t. The main example we will consider is $\mu_t = 2\delta_{U_t}$ where $t \mapsto U_t$ is a continuous real valued function. For each $z \in \mathcal{H}$, consider the Loewner differential equation

$$\partial_t g_t(z) = \int \frac{\mu_t(dx)}{g_t(z)-x}, \quad g_0(z) = z. \tag{2.7}$$

Note that $|\partial_t g_t(z)| \le \mu_t(\mathbb{R})/\mathrm{Im}[g_t(z)|$ and $\partial_t \mathrm{Im}[g_t(z)] < 0$. For each z the solution is well defined up to a time $T_z \in (0,\infty]$. If $T_z < \infty$, then $\lim_{t\to T_z-} \mathrm{Im}(g_t(z)) = 0$. Let K_t be the closure of $\{z \in \mathcal{H} : T_z \le t\}$.

Proposition 2.2. *For every $t > 0$, g_t is a conformal transformation of $\mathcal{H} \setminus K_t$ onto $\mathcal{H}$ satisfying*

$$g_t(z) = z + \frac{a(t)}{z} + O(\frac{1}{|z|^2}), \quad z \to \infty, \tag{2.8}$$

where

$$a(t) = \int_0^t \mu_s(\mathbb{R})\, ds.$$

Proof. Without loss of generality we assume $t \mapsto \mu_t$ is continuous; if this is not true we can apply the argument to the intervals of continuity. As long as $g_t(z)$ stays in $\mathcal{H}$, the differential equation (2.7) is nice, and one can interchange derivatives easily. For example, differentiation of (2.7) with respect to z gives the equation

$$\partial_t g_t'(z) = -g_t'(z) \int \frac{\mu_t(dx)}{(g_t(z)-x)^2}\ .$$

(Here, and throughout this paper, $'$ refers to z-derivatives.) From this we see that $g_t'(z)$ is well defined and nonzero for $t < T_z$. If $t < \min\{T_z, T_w\}$, then

$$\partial_t[g_t(z) - g_t(w)] = -(g_t(z) - g_t(w)) \int \frac{\mu_t(dx)}{(g_t(z)-x)(g_t(w)-x)},$$

and hence $g_t(z) \neq g_t(w)$. This shows that g_t is a conformal transformation of $\mathcal{H} \setminus K_t$. For z large (how large depending on t), the equation (2.7) becomes

$$\partial_t g_t(z) = \frac{\mu_t(\mathbb{R})}{z} + O\left(\frac{1}{|z|^2}\right),$$

from which (2.8) follows. The only thing remaining to show is that $g_t(\mathcal{H} \setminus K_t) = \mathcal{H}$. Fix a t_0 and $w \in \mathcal{H}$ and let $h_t(w), 0 \leq t \leq t_0$ be the solution of the initial value problem

$$\partial_t h_t(w) = -\int \frac{\mu_{t_0-t}(dx)}{h_t(w) - x}, \quad h_0(w) = w.$$

Since $|\partial_t h_t(w)| \leq \mu_{t_0-t}(\mathbb{R})/\mathrm{Im}[h_t(w)]$ and $\partial_t \mathrm{Im}[h_t(w)] > 0$, it is easy to see that $h_t(w)$ is defined for all $0 \leq t \leq t_0$. Moreover, if $g_t(z) = h_{t_0-t}(z)$, then g_t satisfies the Loewner equation (2.7) for $0 \leq t \leq t_0$ and $z = h_{t_0}(w)$. In other words, $g_{t_0}(z) = w$ and $w \in g_{t_0}(\mathcal{H} \setminus K_t)$. □

The Loewner equation is often written in terms of the inverse transformation $f_t(z) = g_t^{-1}(z)$. Differentiating the equation $f_t(g_t(z)) = z$ with respect to t gives the following alternative form of the Loewner differential equation:

$$\partial_t f_t(z) = -f_t'(z) \int \frac{\mu_t(dx)}{z - x}, \quad f_0(z) = z. \tag{2.9}$$

The capacity of K_t is $a(t)$. Consider the case $\mu_t = 2\delta_{U_t}$. If U_t is nice enough, then we can define $\gamma_t = g_t^{-1}(U_t)$ and get a simple curve γ_t. One sufficient condition for "nice enough" is for U to be Hölder continuous of order $1/2$ with sufficiently small Hölder-$(1/2)$ norm [19]. There is a simple heuristic that shows why Hölder $1/2$ should be a critical case. Suppose U_t does not move very fast. Then in time t the curve $\gamma[0, t]$ moves away from the imaginary axis, and it gets distance of order $\sqrt{t}$ away in time t (since the capacity $a(t)$ tends to look like the square of the radius). What can prevent this nice heuristic from working is if U moves sufficiently fast that by the time the path wants to reach distance $\sqrt{t}$, U_t has moved a distance of order $\sqrt{t}$ along the real axis.

For not so nice functions, it is possible that γ is well defined but has double points. In this case K_t includes points not on $\gamma[0, t]$ for whenever γ makes a loop, all the points "inside" the loop are swallowed into K_t. One can also create examples of curves for which γ is not well defined; in these cases, the hull K_t is defined but there is no γ such $\mathcal{H} \setminus K_t$ is the unbounded component of $\mathcal{H} \setminus \gamma[0, t]$.

Example. Suppose $U_t = 0$ for all t. The Loewner equation becomes

$$\partial_t g_t(z) = \frac{2}{g_t(z)}, \quad g_0(z) = z,$$

which has the solution

$$g_t(z) = \sqrt{z^2 + 4t}.$$

Then, $\gamma(t) = 2\sqrt{t}i$ and $K_t = [0, 2\sqrt{t}i]$.

Suppose g_t is a collection of conformal maps obtained from measures μ_t. Let K_t be the corresponding hulls with capacities $a(t)$ satisfying $\partial_t a(t) = \mu_t(\mathbb{R})$. For any $s < t$, define the map $h_{s,t}$ by $g_t = h_{s,t} \circ g_s$. Let $K_{s,t}$ be the closure of $g_s(K_t) \cap \mathcal{H}$. Then $h_{s,t}$ is the unique conformal transformation of $\mathcal{H} \setminus K_{s,t}$ onto $\mathcal{H}$ such that as $z \to \infty$,

$$h_{s,t}(z) = z + \frac{a(t) - a(s)}{z} + O\left(\frac{1}{|z|^2}\right).$$

In particular, the capacity of $K_{s,t}$ is $a(t) - a(s)$.

We sometimes call the equation (2.7) (or the equation for the inverse (2.9)) the *chordal* Loewner equation because it describes the evolution of a hull starting on the boundary of a domain going to a boundary point (in this case the point at infinity in the upper half plane).

2.3 Radial Loewner equation

There is a similar equation that describes the evolution of a hull from the boundary of a domain to an interior point. This equation (or the equation for the inverse) is the one more commonly referred to as the Loewner differential equation in the complex variables literature. For ease we will choose the unit disk $\mathbb{D}$, and use the origin as the interior point. We will call a compact set $K \subset \mathbb{D} \setminus \{0\}$ a hull if $K = \overline{K \cap \mathbb{D}}$, and $\mathbb{D} \setminus K$ is simply connected. Given a hull K, the Riemann mapping theorem tells us that there is a unique conformal transformation $g_K : \mathbb{D} \setminus K \to \mathbb{D}$ such that $g_K(0) = 0$, $g_K'(0) > 0$. We call $g_K'(0)$ the capacity of K (in $\mathbb{D}$ from the origin); sometimes it is more natural to consider the logarithmic capacity $\log g_K'(0)$.

Let $t \mapsto \mu_t$ be a piecewise continuous function from $[0, \infty)$ to the set of positive Borel measures on $\partial\mathbb{D}$ (we will use μ_t for both this measure and the corresponding measure on $[0, 2\pi)$). Assume that $\mu_t(\partial\mathbb{D})$ is uniformly bounded. An important example is $\mu_t = \delta_{e^{iU_t}}$ where $t \mapsto U_t$ is a continuous real-valued function. For each $z \in \mathbb{D}$, consider the initial value problem

$$\partial_t g_t(z) = g_t(z) \int_0^{2\pi} \frac{e^{i\theta} + g_t(z)}{e^{i\theta} - g_t(z)} \, \mu_t(d\theta), \quad g_0(z) = z. \tag{2.10}$$

Note that $g_t(0) = 0$ for all t, and

$$|\partial_t g_t(z)| \le 2\,|g_t(z)|\,\mu_t(\mathbb{D})\,(1 - |g_t(z)|)^{-1} \le 2\,\mu_t(\mathbb{D})\,(1 - |g_t(z)|)^{-1}.$$

If $g_t(z) \neq 0$, we can write the differential equation as

$$\partial_t[\log g_t(z)] = \int_0^{2\pi} \frac{e^{i\theta} + g_t(z)}{e^{i\theta} - g_t(z)}\, \mu_t(d\theta),$$

where log denotes any branch of the logarithm defined locally around $g_t(z)$. Since

$$\operatorname{Re}\left[\frac{e^{i\theta} + w}{e^{i\theta} - w}\right] > 0, \quad w \in \mathbb{D},$$

we see that $\log |g_t(z)|$ and hence $|g_t(z)|$ increases with t. For any $z \in \mathbb{D}$, the solution exists up to a time $T_z \in (0, \infty]$; if $T_z < \infty$, then $\lim_{t \to T_z -} |g_t(z)| = 1$. Let K_t be the closure of $\{z \in \mathbb{D} : T_z \leq t\}$. If $z \neq 0$ and $t < T_z$, in a neighborhood of z we can write $g_t(z) = \exp\{i h_t(z)\}$. The equation (2.10) can be written as

$$\partial_t h_t(z) = -i \int_0^{2\pi} \frac{e^{i\theta} + e^{i h_t(z)}}{e^{i\theta} - e^{i h_t(z)}}\, \mu_t(d\theta) = \int_0^{2\pi} \cot\left[\frac{h_t(z) - \theta}{2}\right] \mu_t(d\theta). \quad (2.11)$$

Proposition 2.3. *For all t, g_t is the unique conformal transformation from $\mathbb{D} \setminus K_t$ to $\mathbb{D}$ with $g_t(0) = 0$ and $g_t'(0) > 0$. In fact,*

$$\log g_t'(0) = \int_0^t \mu_s(\partial\mathbb{D})\, ds. \quad (2.12)$$

Proof. By differentiating (2.10) with respect to z, we see that $g_t' = g_t'(z)$ exists and satisfies

$$\partial_t g_t' = g_t' \left[\int_0^{2\pi} \frac{e^{i\theta} + g_t}{e^{i\theta} - g_t}\, \mu_t(d\theta) + g_t \int_0^{2\pi} \frac{2e^{2i\theta}}{(e^{i\theta} - g_t)^2}\, \mu_t(d\theta) \right]. \quad (2.13)$$

From this we can see that $g_t'(z) \neq 0$ if $t < T_z$. Plugging in $z = 0$ gives

$$\partial_t g_t'(0) = g_t'(0)\, \mu_t(\partial\mathbb{D}),$$

which implies (2.12). If $z \neq w$ and $t < \min\{T_z, T_w\}$,

$$\frac{\partial_t[g_t(z) - g_t(w)]}{g_t(z) - g_t(w)} = \int \frac{e^{2i\theta} + e^{i\theta}[g_t(z) + g_t(w)] - g_t(z) g_t(w)}{(e^{i\theta} - g_t(z))(e^{i\theta} - g_t(w))}\, \mu_t(d\theta),$$

and hence $g_t(z) - g_t(w) \neq 0$. It remains to show that the image of g_t is $\mathbb{D}$. Fix $t_0 > 0$ and $w \in \mathbb{D}$, and consider the initial value problem

$$\partial_t h_t(w) = -h_t(w) \int_0^{2\pi} \frac{e^{i\theta} + h_t(z)}{e^{i\theta} - h_t(z)}\, \mu_{t_0 - t}(d\theta), \quad h_0(w) = w.$$

Note that $|h_t(w)|$ decreases as t increases, and

$$|\partial_t h_t(w)| \leq 2\, |h_t(w)|\, \mu_{t_0 - t}(\mathbb{D})\, (1 - |h_t(w)|)^{-1} \leq 2\, |h_t(w)|\, \mu_{t_0 - t}(\mathbb{D})\, (1 - |w|)^{-1}.$$

Hence $h_t(w)$ is well defined for $0 \leq t \leq t_0$. Also, $g_t(z) = h_{t_0 - t}(z)$ satisfies (2.10) with $g_0(z) = h_{t_0}(w)$. In other words, $g_{t_0}(h_{t_0}(w)) = w$. □

As in the chordal case, the Loewner differential equation is often written in terms of the inverse transformation $f_t = g_t^{-1}$. By differentiating the relation $f_t(g_t(z)) = z$ with respect to t we get the equation

$$\partial_t f_t(z) = -z\, f_t'(z) \int \frac{e^{i\theta} + z}{e^{i\theta} - z}\, \mu_t(d\theta), \quad f_0(z) = z.$$

By conformal transformation, one can have a radial process on any simply connected domain starting at a boundary point going to an interior point. One interesting case is the inversion $z \mapsto 1/z$. Suppose that g_t satisfies the radial Loewner equation (2.10) for a given μ_t, and let $\tilde{g}_t(z) = 1/g_t(1/z)$. Then $\tilde{g}_t$ satisfies the Loewner equation

$$\partial_t \tilde{g}_t = \tilde{g}_t \int_0^{2\pi} \frac{e^{i\theta} + \tilde{g}_t(z)}{e^{i\theta} - \tilde{g}_t(z)}\, \tilde{\mu}_t(d\theta), \quad g_t(z) = z, \tag{2.14}$$

where $\tilde{\mu}_t$ is the induced from μ_t from conjugation. This can be considered as the radial Loewner process growing towards infinity.

2.4 Whole plane Loewner equation

The whole plane Loewner equation describes the evolution of hulls starting with a single point, which for ease we choose to be the origin, going to infinity. It is just the radial equation started at $t = -\infty$. The hull starts with only the origin, immediately grows a small amount so that it has positive capacity, and then evolves as a radial Loewner process going to infinity.

To be more precise, assume we have a piecewise continuous function $t \mapsto \mu_t$ with $t \in (-\infty, \infty)$. Assume that $\mu_t(\mathbb{R})$ is uniformly bounded in t, and let

$$a(t) = \int_0^t \mu_s(\mathbb{R})\, ds$$

(here t can be positive or negative). We assume that

$$\lim_{t\to\infty} a(t) = \infty, \qquad \lim_{t\to-\infty} a(t) = -\infty.$$

A solution to the whole plane Loewner equation with driving measures μ_t is a growing set of compact hulls K_t, $-\infty < t < \infty$ and corresponding conformal transformations

$$g_t : \mathbb{C} \setminus K_t \to \mathbb{C} \setminus \bar{\mathbb{D}},$$

with $g_t(\infty) = \infty$, $g_t'(\infty) > 0$. The logarithmic capacity (see Section 2.1) of K_t is $a(t)$, i.e., $g_t'(\infty) = e^{-a(t)}$. In particular, $K_{-\infty} := \bigcap_t K_t = \{0\}$. The transformations

g_t satisfy the differential equation in (2.14),

$$\partial_t g_t = g_t \int_0^{2\pi} \frac{e^{i\theta} + g_t(z)}{e^{i\theta} - g_t(z)}\, \mu_t(d\theta). \tag{2.15}$$

It is not immediately obvious that a solution exists for all times $t \in (-\infty, \infty)$. If we are given K_s, g_s, then we can use the radial Loewner equation to determine K_t, g_t for all $t \geq s$. In fact, $g_t = h_{s,t} \circ g_s$ where $h_{s,t}$ satisfies

$$\partial_t h_{s,t} = h_{s,t} \int_0^{2\pi} \frac{e^{i\theta} + h_{s,t}(z)}{e^{i\theta} - h_{s,t}(z)}\, \mu_t(d\theta), \qquad t \geq s,$$

with $h_{s,s}(z) = z$. Let $\tilde{g}_s(z) = e^{-a(s)} z$ and $g_{s,t}(z) = h_{s,t} \circ \tilde{g}_s$. If $|w| > 1$, then by (2.2),

$$|g_t^{-1}(w) - g_{s,t}^{-1}(w)| = |g_s^{-1}(h_{s,t}^{-1}(w)) - \tilde{g}_s^{-1}(h_{s,t}^{-1}(w))| \leq C\, e^{a(s)}.$$

In particular, as $s \to -\infty$, $g_{s,t}^{-1}$ converges uniformly to g_t^{-1}. This gives us a way to construct g_t^{-1}: take the limit of the functions $g_{s,t}^{-1}$ as $s \to -\infty$.

3 Stochastic Loewner Evolution

3.1 Chordal version

The stochastic Loewner evolution (SLE) can be thought of as Brownian motion on the set of conformal maps g_t as described in Section 2.2. Suppose a collection of random maps g_t are given that satisfy the Loewner equation

$$\partial_t g_t(z) = \frac{2}{g_t(z) - U_t}, \qquad g_0(z) = z,$$

where U_t is a random continuous function from $[0, \infty)$ to $\mathbb{R}$ with $U_0 = 0$. As before, define maps $h_{s,t}$ by $g_t = h_{s,t} \circ g_s$, and let $\tilde{h}_{s,t}(z) = h_{s,t}(z + U_s) - U_s$. Suppose the random maps $\tilde{h}s, t$ satisfy the following assumptions that are very reminiscent of assumptions for (driftless) Brownian motion.

- **Independent increments:** For every $s < t$, the random function $\tilde{h}_{s,t}$ is independent of g_r, $0 \leq r \leq s$;
- **Identically distributed increments:** For every $s < t$, the distribution of $\tilde{h}_{s,t}$ is the same as the distribution of g_{t-s};
- **Reflection Symmetry:** The distribution of g_t is invariant under the map $x + iy \mapsto -x + iy$.

These assumptions can be translated into assumptions about the process U_t: U_t is a random continuous process satisfying

- **Independent increments:** For every $s < t$, the random variable $U_t - U_s$ is independent of U_r, $0 \le r \le s$;
- **Identically distributed increments:** For every $s < t$, the distribution of $U_{s+r} - U_s$, $0 \le r \le t - s$ is the same as the distribution of U_r, $0 \le r \le t - s$;
- **Reflection Symmetry:** the distribution of U_t is symmetric about the origin.

It is well known that the only continuous process satisfying these assumptions is driftless Brownian motion. There is one parameter we can choose, the variance parameter for the Brownian motion.

Definition. The (*chordal*) *stochastic Loewner evolution* with parameter $\kappa \ge 0$ (denoted SLE_κ) starting at $x \in \mathbb{R}$ is the random family of maps g_t obtained from the chordal Loewner equation with $U_t = \sqrt{\kappa} B_t$, and B_t is a standard one-dimensional Brownian motion with $\sqrt{\kappa} B_0 = x$.

We have defined SLE_κ in terms of the random maps g_t. In fact, we would like to define the random curve $\gamma(t) := g_t^{-1}(U_t)$. The sufficient condition we mentioned in Section 2.2 that would guarantee the existence of a simple curve γ is that U_t have sufficiently small Hölder-$(1/2)$ norm. However, Brownian paths are not Hölder continuous of order $(1/2)$ (they are close in that they are Hölder continuous of order α for every $\alpha < 1/2$). Rohde and Schramm [20] have proved that the SLE_κ curve γ is well defined at least for $\kappa \ne 8$ (the result is expected to be true for $\kappa = 8$ but a different proof will be needed).

A surprising aspect of SLE_κ is that the behavior of g_t varies greatly as the parameter κ varies, even though the paths U_t are qualitatively the same for all $\kappa > 0$. Let us first ask whether or not SLE_κ curves γ are simple (nonintersecting) curves. In order to answer this question, consider what would happen if the curve had double points. Since the "past" of the curve is conformally equivalent to a segment of the real line, it would mean that the curve had a chance to hit the real line. So we can ask a similar question: consider the hull K_t produced by SLE_κ starting at the origin. What is $K_t \cap \mathbb{R}$? For topological reasons we know that $K_t \cap \mathbb{R}$ is a compact interval containing 0, perhaps the trivial one-point interval.

Let $x > 0$, and consider the probability that $x \in K_t$. Note that if $x \notin K_t$, then the boundary of $\mathcal{H} \setminus K_t$ contains an open interval in $\mathbb{R}$ about x. In this case, $g_t(x)$ is well defined and satisfies the Loewner equation,

$$\partial_t g_t(x) = \frac{2}{g_t(x) - U_t}, \quad g_0(x) = x. \tag{3.1}$$

(Note that $\partial(\mathcal{H} \setminus K_t)$ is locally a straight line at x, and hence we can extend Loewner's equation to x.) The first time t at which $x \in K_t$ can be characterized as the first time that $g_t(x) = U_t$. Let $Y_t = (g_t(x) - U_t)/\sqrt{\kappa}$. Then Y_t satisfies the stochastic differential equation

$$dY_t = \frac{a}{Y_t}\, dt + d\tilde{B}_t,$$

where $a = 2/\kappa$ and $\tilde{B}_t = -B_t$. The solution to this equation is a Bessel process; if $a = (d-1)/2$, then the process has the same distribution as the absolute value of a d-dimensional Brownian motion. Bessel processes hit the origin if and only if $a < 1/2$ ($a = 1/2$ corresponds to $d = 2$ which is the critical dimension for Brownian paths to hit the origin). This argument (with the result of Rohde and Schramm) proves the following proposition.

Proposition 3.1. *For $\kappa \le 4$, the paths γ are simple curves. For $\kappa > 4$, with probability one,*

$$\bigcup_{t>0} K_t \cap \mathbb{R} = \mathbb{R}.$$

Chordal SLE_κ has been defined so that $a(K_t) = 2t$. This is a somewhat arbitrary choice. In fact, most of the properties we are interested in are independent of the choice of parameterization so it is important to know how time changes affect the equation. Suppose $a : [0,\infty) \to [0,\infty)$ is an increasing C^1 homeomorphism and g_t are the transformations given by SLE_κ. Then $\hat{g}_t := g_{a(t)/2}$ parameterizes the transformations so that $a(\tilde{K}_t) := a(K_{a(t)/2}) = a(t)$. Note that $\hat{g}_t$ satisfies

$$\partial_t \hat{g}_t = \frac{\partial_t a}{\hat{g}_t - U_{a(t)/2}}, \quad \hat{g}_0(z) = z,$$

If $a(\tilde{K}_t) = 2\alpha t$ for a constant $\alpha > 0$, then

$$\partial_t[\alpha^{-1/2}\hat{g}_t] = \frac{2}{\alpha^{-1/2}\hat{g}_t - \alpha^{-1/2}U_{\alpha t}}.$$

But Brownian scaling implies that $\alpha^{-1/2}B_{\alpha t}$ has the same distribution as B_t. This gives a scaling relation for SLE_κ.

Proposition 3.2 (Scaling). *If g_t are the transformations of SLE_κ and $\alpha > 0$, then $\hat{g}_t(z) := \alpha^{-1/2} g_{\alpha t}(\alpha^{1/2} z)$ also has the distribution of SLE_κ.*

It follows from this proposition that the image of a chordal SLE_κ path under the map $z \mapsto rz$ has the same distribution as (a time change of) SLE_κ. We can define chordal SLE_κ connecting any two distinct points on the boundary of a simply connected domain by conformal transformation. The conformal transformation is not uniquely determined; however, the scaling rule shows that the distribution of the image will be the same for all transformations, up to a change of time. Hence the distribution of paths given is well defined if we consider two paths equivalent if one is just an increasing reparameterization of the other.

3.2 Radial and whole plane versions

The radial version of SLE_κ is defined similarly to the chordal version. Suppose B_t is a standard Brownian motion on the line, $U_t = \sqrt{\kappa} B_t$, and $\mu_t = \delta_{e^{iU_t}}$. Then the

solution g_t to the radial Loewner equation (2.10) is called radial SLE_κ. The result of Rohde and Schramm allows us to define the random path γ from the boundary to the origin (at least if $\kappa \neq 8$).

Qualitatively, radial SLE_κ paths and chordal SLE_κ paths for the same value of κ are similar. For example, let us consider the case of double points. Suppose $U_0 = 0$, $0 < x < 2\pi$ and consider whether the hull K_t every contains e^{ix}. Again, if $e^{ix} \notin K_t$, then the map g_t extends smoothly to e^{ix}, so we can use the Loewner equation. Let us write $g_t(e^{ix}) = e^{iX_t}$; then the first time t at which $e^{ix} \in K_t$ is the same as the first time t that $Z_t := X_t - U_t$ equals 0 or 2π. Using (2.11), we see that Z_t satisfies the stochastic differential equation

$$dZ_t = \cot[Z_t/2]\, dt + \sqrt{\kappa}\, d\tilde{B}_t.$$

Note that near 0 this acts like a Bessel process with the same critical parameter $\kappa = 4$; similarly, near 2π it acts like a Bessel process repelling in the opposite direction. From this we see that radial SLE_κ paths are simple for $\kappa \leq 4$, and otherwise have double points.

We have defined radial SLE_κ as starting on the unit circle going towards the origin; we could just as easily have defined it as going from the unit circle to infinity. (In fact, radial SLE_κ can be defined in any simply connected domain connecting a boundary point to an interior point.) To obtain the whole plane SLE_κ, we let $U_t = \sqrt{\kappa} B_t$, $-\infty < t < \infty$ be a two-sided Brownian motion such that U_0 has a uniform distribution of $[0, 2\pi]$, i.e., e^{iU_0} is uniformly distributed on $\partial\mathbb{D}$. The maps g_t are then defined as in Section 2.4. Given $\gamma(t)$, $-\infty < t < t_0$, the curve evolves like radial SLE_κ going towards infinity.

3.3 Locality

Suppose A is a compact hull bounded away from 0 in $\mathcal{H}$, and let $\Phi = \Phi_A : \mathcal{H}\backslash A \to \mathcal{H}$ be the conformal transformation with $\Phi(\infty) = \infty$, $\Phi(0) = 0$, $\Phi'(\infty) = 1$ (This is a translation of the transformation described in Section 2.2.) Suppose γ is a curve obtained from chordal SLE_κ starting at the origin with corresponding hulls K_t, and let

$$T = T_A = \inf\{t : K_t \cap A \neq \emptyset\}.$$

If $\kappa \leq 4$, then $\mathbf{P}\{T = \infty\} > 0$ while $\mathbf{P}\{T < \infty\} = 1$ if $\kappa > 4$. On the event $V_t = \{T > t\}$, consider the path

$$\gamma^*(s) = \Phi \circ \gamma(s), \quad 0 \leq s \leq t.$$

The capacity $a(\Phi(K_s))$ is less than $2s$; the exact value depends on both Φ and on γ. We say that chordal SLE_κ satisfies the *locality property* if the distribution of $\gamma^*[0, t]$ on the event V_t is the same as a time change of SLE_κ; the time change is given in terms of the capacity $a(\Phi(K_s))$. Heuristically, the locality property says that an

SLE_κ path in $\mathcal{H} \setminus A$ does not "feel the boundary" until it actually hits the boundary. This is a property satisfied by planar Brownian motion paths in bounded domains — they evolve like Brownian motions in the entire plane, except for when they hit the boundary at which point they are stopped or reflected (depending on the particular boundary conditions).

Similarly, if $A \subset \mathbb{D}$ is a hull such that $\mathbb{D} \setminus A$ is a simply connected domain containing the origin, let $\Psi = \Psi_A : \mathbb{D} \setminus A \to \mathbb{D}$ be the conformal transformation with $\Psi(0) = 0$, $\Psi'(0) > 0$. If γ is the curve and K_t the hulls obtained from radial SLE_κ starting at $z \in \partial\mathbb{D} \setminus A$, and

$$T = T_A = \inf\{t : K_t \cap A \neq \emptyset\},$$

then on the event $V_t = \{T > t\}$ we can consider the curve $\gamma^*(s) = \Psi \circ \gamma(s)$. We say that radial SLE_κ satisfies the locality property if the distribution of γ^* on the event V_t is a time change of SLE_κ. In this section we will sketch the proof of the following theorem.

Theorem 3.3. *Chordal and radial SLE_κ satisfy the locality property for $\kappa = 6$ and no other values of κ.*

We note that we would not expect the locality property to hold for $\kappa \leq 4$. For example, in the chordal case with $\kappa \leq 4$, the curve γ goes to infinity without hitting the real line and hence stays a (random) positive distance away from every compact interval not containing the origin. However, γ has a positive probability of hitting A, and hence the paths of γ^* have the property that they approach the real line with positive probability.

We start with the chordal case. The basic strategy of the proof is as follows. We have a Brownian motion B_t starting at the origin, and hence $U_t = \sqrt{\kappa} B_t$, and the corresponding $\gamma(t)$, g_t, K_t given by solving the Loewner equation. Let N be a simply connected open subset of $\mathcal{H}$, disjoint from A, whose closure contains an open interval about 0. Then Φ is a conformal transformation of N onto $\Phi(N)$. Let

$$T = \inf\{t : \gamma(t) \notin \bar{N}\}.$$

On the event $V_t = \{T > t\}$, let $\gamma^*(s) = \Phi \circ \gamma(s)$, $0 \leq s \leq t$. Since γ^* is a curve, we can describe the evolution of γ^* (and the corresponding hulls K_t^* and transformations g_t^*) by a Loewner equation

$$\partial_t g_t^* = \frac{\partial_t a(K_t^*)}{g_t^* - U_t^*}.$$

This equation defines the function U_t^*. We use Ito's formula to write U_t^* as a continuous local semimartingale (i.e., a continuous local martingale plus a "drift" term of bounded variation). When we do this we will discover that the drift term is zero for all Φ if and only if $\kappa = 6$.

On the event V_t, let $h_t = g_t^* \circ \Phi \circ g_t^{-1}$. Note that if $\Phi = \Phi_A$, then h_t is the conformal transformation of $\mathcal{H} \setminus g_t(A)$ onto $\mathcal{H}$ with $h_t(z) \sim z$ as $z \to \infty$ and

$h_t(U_t) = U_t^*$. Also, h_t is smooth up to the boundary in a neighborhood of U_t. The chain rule gives

$$\partial_t h_t(z) = \frac{\partial_t a(K_t^*)}{h_t(z) - U_t^*} - \frac{2h_t'(z)}{z - U_t}. \tag{3.2}$$

Here we have used (2.9). By considering the infinitesimal effect of the scaling relation (2.3), we can see that $\partial_t a(K_t^*) = 2h_t'(U_t)^2$. Since $U_t^* = h_t(U_t)$, we get

$$\lim_{z \to U_t} \partial_t h_t(z) = \lim_{z \to U_t} \frac{2h_t'(U_t)^2}{h_t(z) - h_t(U_t)} - \frac{2h_t'(z)}{z - U_t} = -3h_t''(U_t).$$

(In taking the limit, we need to expand $h_t(z)$ up to terms of $(z - U_t)^2$, and $h_t'(z)$ up to terms of $(z - U_t)$.) We can now use Ito's formula to obtain

$$\begin{aligned} dU_t^* &= \partial_t h_t(U_t)\, dt + h_t'(U_t)\, dU_t + \frac{1}{2}\, h_t''(U_t)\, d\langle U \rangle_t \\ &= \frac{\kappa - 6}{2}\, h_t''(U_t)\, dt + \sqrt{\kappa}\, h_t'(U_t)\, dB_t. \end{aligned}$$

From this we see that this is a continuous local martingale for all Φ if and only if $\kappa = 6$. Note that the quadratic variation (for all κ) satisfies $d\langle U_t^* \rangle_t = \kappa h_t'(U_t)^2\, dt = (\partial_t a/2)\kappa\, dt$. Since every continuous local martingale is a time change of standard Brownian motion [5, Theorem 3.4.6] we get the result. Note that for $\kappa \neq 6$, the distribution of $\Phi \circ \gamma$ is absolutely continuous with respect to a time change of SLE_κ (since Brownian motion with drift is absolutely continuous with respect to Brownian motion without drift, assuming the same variance, see, e.g., [5, Section 3.3.5]).

Let Φ be a conformal transformation of $\mathcal{H}$ onto $\mathcal{H}$ with $\Phi(0) = 0$, $\Phi(\infty) = x > 0$. Then, by definition, $\gamma^*(t) = \Phi \circ \gamma(t)$ has the distribution of SLE_κ from 0 to x (this is only defined modulo time change). From the above calculation, we can see that if $\kappa = 6$, γ^* has the same distribution as (a time change of) γ up to the first time the path hits $[x, \infty)$. At this point the two distributions differ in that γ heads towards infinity while γ^* heads towards x.

The radial case is done similarly. We define $h_t = g_t^* \circ \Psi \circ g_t^{-1}$; in this case

$$\partial_t g_t^*(z) = \partial_t [\log h_t'(0)]\, g_t^*(z)\, \frac{V_t^* + g_t^*(z)}{V_t^* - g_t^*(z)},$$

where $V_t = e^{iU_t}$ and $V_t^* = e^{iU_t^*} = h_t(V_t)$. The infinitesimal change in capacity gives $\partial_t \log h_t'(0) = |h'(V_t)|^2$. Let $\phi_t(z) = -i \log h_t(e^{iz})$, so that $\phi_t(z)$ locally takes reals to reals. The chain rule gives

$$\partial_t \phi_t(z) = \phi_t'(U_t)^2 \cot\left(\frac{\phi_t(z) - \phi_t(U_t)}{2}\right) - \phi_t'(z) \cot\left(\frac{z - U_t}{2}\right). \tag{3.3}$$

Direct calculation shows that

$$\lim_{z \to U_t} \partial_t \phi_t(z) = -3\phi_t''(U_t),$$

and hence again by Ito's formula, $U_t^* = \phi_t(U_t)$ is a continuous local martingale if and only if $\kappa = 6$.

Corollary 3.4. *Suppose $D' \subset D$ are simply connected domains with $z, w \in \partial D' \cap \partial D$. Let γ', γ be chordal SLE$_6$ paths going from z to w in D', D respectively. Let T (resp., T') be the first time that γ (resp., γ') hits $D \setminus D'$. Then the distributions of $\gamma[0, T]$ and $\gamma'[0, T']$ are the same.*

Corollary 3.5. *Let $D \subset \mathbb{D}$ be simply connected with $0 \in \partial D$ such that $\partial D \cap \mathbb{D}$ is a nontrivial closed arc. Suppose $z_0 \in \partial D \cap \mathbb{D}$. Let γ be a chordal SLE$_6$ path going from z_0 to 0 in D, and let γ' be a radial SLE$_6$ path going from z_0 to 0 in $\mathbb{D}$. Let T (resp., T') be the first time that γ (resp., γ') hits $\mathbb{D} \setminus D$. Then the distributions of $\gamma[0, T]$ and $\gamma'[0, T']$ are the same.*

This last corollary relates chordal and radial SLE$_6$. Let Φ be a version of $-i \log z$ defined on D. By definition, $\Phi \circ \gamma$ is a chordal SLE$_6$ path from $\log z_0$ to infinity in $\Phi(D)$, and by the chordal locality property this agrees (up to time change) with chordal SLE$_6$ in $\mathcal{H}$ started at $\log z_0$, up to the first time the path leaves $\Phi(D)$. Hence to prove the corollary, it suffices to show that $\Phi \circ \gamma'$ is (a time change of) chordal SLE$_6$ in $\mathcal{H}$ up to the time that γ' leaves D. This can be done using the method described in this section. Here, $h_t = g_t^* \circ \Phi \circ g_t^{-1}$, where g_t^* comes from a chordal Loewner equation while g_t comes from a radial Loewner equation.

3.4 Restriction property

3.4.1 Chordal. Consider chordal SLE$_\kappa$ for $\kappa \leq 4$ which produces a simple curve $\gamma : [0, \infty) \to \bar{\mathcal{H}}$ with $\gamma(0) = 0$, $\gamma(0, \infty) \subset \mathcal{H}$, and $\gamma(t) \to \infty$ as $t \to \infty$. As in the previous section let A be a compact hull bounded away from 0 and infinity in $\mathcal{H}$ and let $\Phi_A : \mathcal{H} \setminus A \to \mathcal{H}$ be the conformal transformation fixing 0 and infinity with $\Phi_A'(\infty) = 1$. It is not difficult to see that there is a positive probability that $\gamma[0, \infty) \cap A = \emptyset$. On this event, we can consider the path $\gamma^*(t) = \Phi_A \circ \gamma(t)$. We say that SLE$_\kappa$ satisfies the (*chordal*) *restriction property* if for all such A the conditional distribution of γ^* given $\gamma[0, \infty) \cap A = \emptyset$ is the same as (a time change of) SLE$_\kappa$. The difference between the restriction property and the locality property is that the restriction property discusses the conditional distribution of the whole path $\gamma[0, \infty)$ only while the locality property describes $\gamma[0, t]$ for $t < \infty$. We will sketch a proof of the following theorem.

Theorem 3.6. *Chordal SLE$_\kappa$ satisfies the restriction property for $\kappa = 8/3$ and for no other $\kappa \leq 4$. If $\kappa = 8/3$ and A, Φ_A are as in the previous paragraph,*

$$\mathbf{P}\{\gamma[0, \infty) \cap A = \emptyset\} = \Phi_A'(0)^{5/8}.$$

The term $\Phi_A'(0)$ can be interpreted in terms of Brownian motion. Let B_t be a Brownian excursion in the upper half plane started at the origin, i.e., a Brownian motion

conditioned to stay in the upper half plane for all times. One can define such an excursion by letting the imaginary part of the Brownian motion move like a Bessel-3 process (which has the distribution of the absolute value of a three dimensional Brownian motion) and letting the real part move by an independent Brownian motion. Then $\Phi_A'(0)$ is the probability that this excursion never hits A. To see this, let B be a Brownian motion in the upper half plane and let τ_R be the first time that the imaginary part of B is at least R. Then the probability that a Brownian motion starting at $z = x + iy \in \mathcal{H}$, conditioned never to hit the real axis, never hits A is given by

$$\lim_{R\to\infty} \frac{\mathbf{P}^z\{B[0,\tau_R] \cap (\mathbb{R} \cup A) = \emptyset\}}{\mathbf{P}^z\{B[0,\tau_R] \cap \mathbb{R} = \emptyset\}}.$$

A standard calculation for one dimensional Brownian motion shows that the denominator is y/R. By conformal invariance the numerator is the probability that a Brownian motion starting at $\Phi_A(z)$ reaches $\Phi_A(\{\mathrm{Im}(z) = R\})$ before hitting the real axis. As $R \to \infty$, this is asymptotic to $\mathrm{Im}[\Phi_A(z)]/R$, and hence the probability that the conditioned Brownian motion starting at z never hits A is $\mathrm{Im}[\Phi_A(z)]/y$. The probability that an *excursion* starting at 0 never hits A is the limit of this probability at $z \to 0$ which is $\Phi_A'(0)$.

Similarly, if k is a positive integer, $\Phi_A'(0)^k$ is the probability that k independent excursions all avoid A. Heuristically, we can interpret Theorem 3.6 as "one $\mathrm{SLE}_{8/3}$ path is the same as 5/8 of a Brownian excursion." More precisely, it can be proved that the distribution of the hull generated by 8 independent $\mathrm{SLE}_{8/3}$ paths is the same as the hull generated by 5 Brownian excursions. (The hull is the entire region between the leftmost boundary and rightmost boundary of the union.)

We state Theorem 3.6 only for $\kappa \le 4$. For $\kappa > 4$, the hull created by $\gamma[0,\infty)$ is the entire upper half plane; however, Corollary 3.4 can be considered a restriction result for $\kappa = 6$.

The set of maps $\{\Phi_A\}$ forms a semigroup under composition. If the restriction property holds, then the map

$$\Phi_A \longmapsto \mathbf{P}\{\gamma[0,\infty) \cap A = \emptyset\}$$

is a homomorphism into the multiplicative semigroup [0, 1]. From this, one can show that

$$\mathbf{P}\{\gamma[0,\infty) \cap A = \emptyset\} = \Phi_A'(0)^a, \tag{3.4}$$

for some $a > 0$. Conversely, if we can show that (3.4) holds for some $a > 0$, it follows that the restriction property holds. To see this, note that the distribution of the image $\gamma[0,\infty)$ is determined by the probabilities of avoiding hulls A' (here we use the fact that γ is simple and goes to infinity). Suppose A, A' are hulls and let $A_1 = \Phi_A^{-1}(A')$.

Then $\Phi_{A_1} = \Phi_{A'} \circ \Phi_A$, and (3.4) gives

$$\begin{aligned}\mathbf{P}\{(\Phi_A \circ \gamma)[0,\infty) \cap A' = \emptyset \mid \gamma[0,\infty) \cap A = \emptyset\} \\ &= \mathbf{P}\{\gamma[0,\infty) \cap A_1 = \emptyset \mid \gamma[0,\infty) \cap A = \emptyset\} \\ &= \Phi'_{A_1}(0)^a / \Phi'_A(0)^a \\ &= \Phi'_{A'}(0)^a = \mathbf{P}\{\gamma[0,\infty) \cap A' = \emptyset\}.\end{aligned}$$

The idea of the proof of Theorem 3.6 is to find for which $\kappa \leq 4$ (3.4) holds (and for which value of a). Fix a hull A and let $\Phi = \Phi_A$. For any $t > 0$, let V_t be the event $\{\gamma[0,t] \cap A = \emptyset\}$. On V_t define g_t^* and h_t as in the chordal part of Section 3.3. If (3.4) holds, then

$$\mathbf{P}\{\gamma[0,\infty) \cap A = \emptyset \mid \gamma[0,t]\} = 1_{V_t}\, h_t'(U_t)^a,$$

and $M_t := h_t'(U_t)^a = \exp\{a \log h_t'(U_t)\}$ must be a martingale up to the hitting time of A. Differentiating (3.2) gives

$$\partial_t h_t'(z) = \frac{2h_t'(z)}{(z-U_t)^2} - \frac{2h_t''(z)}{z-U_t} - \frac{2h_t'(U_t)^2 h_t'(z)}{(h_t(z)-h_t(U_t))^2},$$

and hence

$$\lim_{z \to U_t} \partial_t h_t'(z) = \frac{h_t''(U_t)^2}{2h_t'(U_t)} - \frac{4h_t'''(U_t)}{3}.$$

(Taking this limit requires expanding $h_t(z)$ into terms up to $(z-U_t)^3$ and $h_t'(z)$ into terms up to $(z-U_t)^2$.) Ito's formula gives

$$\begin{aligned}dM_t = M_t\, [\, &\frac{a[(\kappa/2) - (4/3)]\, h_t'''(U_t)}{h_t'(U_t)} \\ &+ \frac{a\,[(\kappa/2)(a-1) + (1/2)]\, h_t''(U_t)^2}{h_t'(U_t)^2}\,]\, dt \;+\; (\cdots)\, dB_t.\end{aligned}$$

The drift term is zero for all h if and only if $\kappa = 8/3$, $a = 5/8$.

Corollary 3.7. *Suppose $D' \subset D$ are simply connected domains with $z, w \in \partial D' \cap \partial D$. Let γ', γ be chordal $SLE_{8/3}$ paths going from z to w in D', D respectively. Then the conditional distribution of $\gamma[0,\infty)$ given $\gamma(0,\infty) \subset D'$ is the same as the distribution of $\gamma'[0,\infty)$. Moreover, if $\partial D'$, ∂D are smooth near z, w, then*

$$\mathbf{P}\{\gamma(0,\infty) \subset D'\} = |\Phi'(z)|^{5/8}\, |\Phi'(w)|^{5/8},$$

where Φ is any conformal transformation of D' onto D with $\Phi(z) = z$, $\Phi(w) = w$.

Remark. Although the transformation Φ is not uniquely determined, the quantity $\Phi'(z)\Phi'(w)$ is unique. This follows since every conformal transformation $f : \mathbb{D} \to \mathbb{D}$ with $f(1) = 1$, $f(-1) = -1$ satisfies $f'(1)f'(-1) = 1$.

3.4.2 Radial. Suppose γ is a radial SLE_κ curve ($\kappa \le 4$) going from $z \in \partial\mathbb{D}$ to the origin. Let $A \subset \mathbb{D}$ be a hull not containing z such that $\mathbb{D} \setminus A$ is simply connected and contains the origin. Let $\Psi_A : \mathbb{D} \setminus A \to \mathbb{D}$ be the conformal transformation with $\Psi_A(0) = 0, \Psi_A'(0) > 0$. We say that SLE_κ satisfies the *(radial) restriction property* if for all such A the conditional distribution of $\gamma^*(t) := \Psi_A \circ \gamma(t)$ given $\{\gamma[0,\infty) \cap A = \emptyset\}$ is a time change of SLE_κ.

Theorem 3.8. *Radial SLE_κ satisfies the restriction property for $\kappa = 8/3$ and for no other $\kappa \le 4$. If $\kappa = 8/3$ and z, A, Ψ_A are as in the previous paragraph,*

$$\mathbf{P}\{\gamma[0,\infty) \cap A = \emptyset\} = \Psi_A'(z)^{5/8}\, \Psi_A'(0)^{5/48}.$$

Again, it is not hard to show that if the restriction property holds, then it must be the case that

$$\mathbf{P}\{\gamma[0,\infty) \cap A = \emptyset\} = \Psi_A'(z)^{a}\, \Psi_A'(0)^{b}, \tag{3.5}$$

for some a, b. Fix a hull A and let $\Psi = \Psi_A$. For any $t > 0$, let $V_t = \{\gamma[0,t] \cap A = \emptyset\}$. On the event V_t define g_t^* and h_t, ϕ_t as in the radial part of Section 3.3. We note that $\partial_t \log h_t'(0) = |h_t'(V_t)|^2 - 1 = \phi_t'(U_t)^2 - 1$. Let

$$M_t = h_t'(U_t)^a = \exp\{a \log \phi_t'(U_t) + b \log h_t'(0)\}.$$

If (3.5) is to hold, then M_t must be a martingale on V_t. Ito's formula shows that this requires

$$\partial_t \phi_t'(U_t) + \frac{\kappa}{2}\left[\phi_t'''(U_t) - \frac{\phi_t''(U_t)^2}{\phi_t'(U_t)}\right] + \frac{a\kappa}{2}\frac{\phi_t''(U_t)^2}{\phi_t'(U_t)} + \frac{b}{a}\phi_t'(U_t)^3 = 0. \tag{3.6}$$

Differentiation of (3.3) gives

$$\begin{aligned}\partial_t \phi_t'(z) = &-\frac{1}{2}\phi_t'(0)^2 \phi_t'(z) \csc^2\left(\frac{\phi_t(z) - \phi_t(U_t)}{2}\right) \\ &- \phi_t''(z) \cot\left(\frac{z - U_t}{2}\right) + \frac{1}{2}\phi_t'(z)\csc^2\left(\frac{z-U_t}{2}\right).\end{aligned}$$

Hence,

$$\lim_{z \to U_t} \partial_t \phi_t'(z) = -\frac{4\phi_t'''(U_t)}{3} + \frac{\phi_t''(U_t)^2}{2\phi_t'(U_t)} + \frac{\phi_t'(U_t)}{6} - \frac{\phi_t'(U_t)^3}{6}.$$

In particular, (3.6) holds for all ϕ_t if and only if $a = 5/8, b = 5/48$.

Corollary 3.9. *Suppose $D' \subset D$ are simply connected domains containing the origin and $z \in \partial D' \cap \partial D$. Let γ', γ be radial $\mathrm{SLE}_{8/3}$ paths going from z to the origin in D', D respectively. Then the conditional distribution of $\gamma[0,\infty)$ given $\gamma(0,\infty) \subset D'$ is the same as the distribution of $\gamma'[0,\infty)$. Moreover, if $\partial D'$, ∂D are smooth near z, then*

$$\mathbf{P}\{\gamma(0,\infty) \subset D'\} = |\Psi'(z)|^{5/8}\, |\Psi'(0)|^{5/48},$$

where Ψ *is the unique conformal transformation of* D' *onto* D *with* $\Psi(z) = z$, $\Psi(0) = 0$.

3.5 Some discrete thoughts

The stochastic Loewner evolution is a "continuous" object and can be studied as such. However, one of the primary reasons that mathematicians are interested in SLE is that it is the continuum limit of discrete models in statistical physics at criticality. Theoretical physicists had already predicted that these limits would be conformally invariant. The ideas of locality and restriction described in the previous two sections originally came from trying to find which value of κ gives the limit of a particular discrete system. If the discrete model satisfies "locality" or "restriction", then the limit should as well.

For ease, we will consider discrete processes in the upper half plane. (In fact, we should be able to take continuum limits of lattice models on any finitely connected domain. There are still many open questions to establish that these continuum limits exist and produce conformally invariant measures.) Let $\mathcal{H}$ be the upper half plane and for every $\delta > 0$, let $\mathcal{H}_\delta = \delta\mathbb{Z}^2 \cap \mathcal{H}$ with (discrete) boundary $\overline{\mathcal{H}}_\delta = \{(\delta k, 0) : k \in \mathbb{Z}\}$. A (simple random walk) excursion starting at the origin is a simple random walk conditioned not to return to $\overline{\mathcal{H}}_\delta$. This is easy to define as an "h-process". Let $\omega = [\omega^0, \ldots, \omega^k]$ denote nearest neighbor paths in $\mathbb{Z}^2$. For any $z \in \mathcal{H}_\delta$, let $Q_n(z)$ be the number of simple random walk paths of n steps starting at z that stay in $\mathcal{H}_\delta$. Then the excursion starting at the origin takes its first step to $(0, \delta)$ and then uses the transition probability

$$\mathbf{P}\{\omega^{k+1} = z \mid [\omega^0, \ldots, \omega^k]\} = \lim_{n\to\infty} \frac{Q_n(z)}{\sum_{|w-\omega^k|=1} Q_n(w)}. \tag{3.7}$$

(It is an easy exercise to find the limit and show that the excursion at $(j_1\delta, j_2\delta)$ goes up with probability $(j_2+1)/4j_2$, down with probability $(j_2-1)/4j_2$ and left or right with probability $1/4$ each.)

In the limit, this process approaches a Brownian excursion in $\mathcal{H}$ starting at the origin (see Section 4.3). Brownian motion or Brownian excursion, considered as a path, is *not* an SLE path; we can see this because SLE and Brownian motion act differently when they have double points: SLE paths reflect off while Brownian paths "go through". If one observes the "hull" generated by a Brownian excursion (the set of points in $\mathcal{H}$ separated from infinity) growing in time, then the place where the hull grows moves discontinuously in time.

Suppose that $A \subset \mathcal{H}$ is a hull as in Section 3.3. Consider excursions conditioned not to visit A. The conditional distribution is given by a transition probability as in (3.7) where $Q_n(z)$ is replaced with $Q_n(z; A)$, the number of walks of length n that stay in $\mathcal{H}_\delta \setminus A$. This process does not satisfy locality — we can see the effect of the conditioning locally at a point away from A. However, conformal invariance

properties of Brownian motion (Section 4.3) can be used to show that this does satisfy the restriction property. (Here we are using locality and restriction in a loose sense; since this process is not an SLE we cannot use the precise definition from the previous sections.)

Now let us consider the *infinite self-avoiding walk*. This is a process in $\mathcal{H}_\delta$ given by a transition probability as in (3.7) where $Q_n(z)$ is replaced with $Q_n^*(z;\omega)$, the set of self-avoiding (no self-intersections) walks of length n starting at z in $\mathcal{H}$ that do not hit the "past" ω. This limit is expected to exist, although it has not been proved (see [18, Section 8.3] for a proof of existence of a similar limit). It is conjectured that the limit will be conformally invariant, i.e., if we define a similar limit in another simply connected domain it is the same as the half plane limit transformed by the conformal map. It is also conjectured that the limit will give a measure on simple curves. If all these assumptions are true, then the limit should be an SLE_κ for some κ. The nature of the limit shows that it should satisfy the restriction property. This leads to the following conjecture. From this conjecture one can derive heuristically the values of a number of exponents for the walk [14].

Conjecture 3.10. *The continuum limit of the infinite self-avoiding walk is* $SLE_{8/3}$.

Another way to obtain a simple (self-avoiding) path in the limit is to replace $Q_n(z)$ in (3.7) with $Q_n(z;\omega)$, the number of nearest neighbor walks staying in $\mathcal{H}_\delta \setminus \omega$. This is the same measure as one obtains by erasing loops from Brownian excursions (see, e.g., [6]), and is called the loop-erased or Laplacian random walk. This is also conjectured to have a conformally invariant limit; under this assumption, Schramm [21] showed that the limit must be SLE_2; in [13], it was proved that loop-erased walks approach SLE_2. Note that the definition of this process in terms of the loop-erasing procedure shows that it does *not* satisfy locality or restriction.

To get a process satisfying locality, one can start with simple random walk, and instead of conditioning that it avoids A, we instead reflect the walk when it reaches A. Then the distribution of the processes reflected or not reflected at A are exactly the same up to the time the first reflection occurs. Another process that is similar is the boundary of a macroscopic percolation cluster in the plane. This was first noticed on the discrete scale on the triangular or hexagonal lattice by Schramm [21] who showed that under the assumption of conformal invariance, this interface should be an SLE_6 curve. More recently, Smirnov [22] has shown that the limit of critical percolation on the triangular lattice is conformally invariant, and hence that the limit is SLE_6. The distribution of the hull generated by SLE_6 is the same as a hull generated by Brownian motion with reflection by an angle of $\pi/3$ (see [15, 23]), although the two processes create this hull in different orders. (This reflection can be defined easily for random walks on the triangular lattice. What recent results have shown is that the triangular lattice is a natural lattice for percolation and related processes.)

4 Calculating exponents

4.1 Crossing exponent for chordal SLE

Let $R_L = (0, L) \times (0, \pi)$ be the rectangle as in Section 2.1, and suppose γ is the path of SLE_κ in R_L starting at $i\pi$ and ending at $L + i\pi$. This is the image of an SLE_κ in $\mathcal{H}$ under conformal map from $\mathcal{H}$ onto R_L sending 0 to $i\pi$ and ∞ to $L + i\pi$; the choice of time parameterization is not important. Let

$$\hat{T} = \inf\{t : \gamma(t) \in [L, L + i\pi]\}.$$

If $\kappa \le 4$, $\gamma(0, \infty) \subset R_L$ and hence $\hat{T} = \infty$. If $\kappa > 4$, $\hat{T} < \infty$. Let $V = V_L$ be the event that $\gamma[0, \hat{T}] \cap [0, L] = \emptyset$. This has probability one if and only if $\kappa \le 4$. On the event V, let $\mathcal{D}$ be the connected component of $R_L \setminus \gamma[0, \hat{T}]$ whose boundary includes $[0, L]$ and let $\hat{\mathcal{L}}$ be the π-extremal distance (as defined in Section 2.1) between $[0, \pi i] \cap \partial\mathcal{D}$ and $[L, L + \pi i] \cap \partial\mathcal{D}$ in $\mathcal{D}$. On the event V^c, let $\hat{\mathcal{L}} = \infty$. For $\beta \ge 0$, we define the *chordal crossing exponent* $\nu = \nu(\beta, \kappa)$ by the relation

$$\mathbf{E}[\exp\{-\beta\hat{L}\}] \approx e^{-\nu L}, \qquad L \to \infty.$$

(Here we use the convention $e^{-\beta\cdot\infty} = 0$ even for $\beta = 0$.) This is an SLE analogue of the intersection exponents for Brownian motion. Here we have used $\approx$ to indicate that the logarithms are asymptotic. (In fact, one can show that $\mathbf{E}[\exp\{-\beta\hat{L}\}] \asymp e^{-\nu L}$ where $\asymp$ means that both sides are bounded by constants times the other side.)

The first step is to reduce this problem to an easier computation for SLE. Here we give the main idea of the reduction but leave out a number of details that need to be verified to give a complete proof. Our first step is to move the problem to the upper half plane. Let Φ be the conformal transformation of R_L onto $\mathcal{H}$ with $\Phi(\pi i) = 0$, $\Phi(0) = 1$, $\Phi(L + \pi i) = \infty$. Then $\Phi(L) \asymp e^L$. Let γ be an SLE_κ path in $\mathcal{H}$ starting at the origin going to infinity, let

$$T = \inf\{t : \gamma(t) \in (e^L, \infty)\},$$

and let $\tilde{\mathcal{L}}$ be the π-extremal distance between $[0, 1]$ and $[e^L, \gamma(T)]$ in $\mathcal{H} \setminus \gamma[0, T]$. (If the boundary of $\mathcal{H} \setminus \gamma[0, T]$ does not intersect $[0, 1]$, then $\mathcal{L} = \infty$; otherwise, the intersection is actually $[r, 1]$ for some $r > 0$, so we could say equally well the π-extremal distance between $[r, 1]$ and $[e^L, \gamma(T)]$.) Then,

$$\mathbf{E}[\exp\{-\beta\tilde{\mathcal{L}}\}] \approx e^{-\nu L} = (e^L)^{-\nu}.$$

The scaling relation (Proposition 3.2) can be used to show that the left hand side is comparable to the quantity with $\tilde{\mathcal{L}}$ replaced by $\mathcal{L}$, the π-extremal distance between $[0, 1]$ and $[t, \infty]$ in $\mathcal{H} \setminus \gamma[0, t^2]$, where $t = e^L$. By conformal invariance, $\mathcal{L}$ is the π-extremal distance between $[g_{t^2}(0), g_{t^2}(1)]$ and $[g_{t^2}(t), \infty)$ in $\mathcal{H}$. In the terms that dominate, $g_{t^2}(t) - g_{t^2}(1)$ is of order t. Hence, we can see that

$$e^{\mathcal{L}} \approx t\,[g_{t^2}(1) - g_{t^2}(0)]^{-1}.$$

Our final approximation is $[g_{t^2}(1) - g_{t^2}(0)] \approx g'_{t^2}(1)$. If we accept all these approximations (they can be verified, but it takes some work), we get the relation

$$\mathbf{E}[g'_{t^2}(1)^\beta] = \mathbf{E}[\exp\{\beta \log g'_{t^2}(1)\}] \approx t^{-(\nu-\beta)}, \qquad t \to \infty.$$

For $t \geq 0$, $x > 0$, let

$$v(t, x) = v_{\kappa,\beta}(t, x) = \mathbf{E}[\exp\{\beta \log g'_t(x)\}].$$

We need to find the asymptotics of $v(t^2, 1)$ as $t \to \infty$. The scaling law for SLE implies that $v(t, x) = v(1, x/\sqrt{t}) = \phi(x/\sqrt{t})$ for some function ϕ of one variable, and (if ν exists), $\phi(y) \approx y^{\nu-\beta}$ as $y \to 0+$. Differentiating the Loewner equation (3.1) gives

$$\partial_t \log[[g_t(x) - U_t]'] = \partial_t \log g'_t(x) = -\frac{2}{(g_t(x) - U_t)^2},$$

or, if $Y_t = g_t(x) - U_t$,

$$\exp\{\beta \log g'_t(x)\} = \exp\left\{-\beta \int_0^t \frac{2}{Y_s^2}\, ds\right\}.$$

Since Y_t satisfies the stochastic differential equation

$$dY_t = \frac{2}{Y_t}\, dt - \sqrt{\kappa}\, dB_t,$$

the Feynman–Kac formula [5, Theorem 5.7.6] tells us that $v(t, x)$ satisfies the PDE

$$\partial_t v = \frac{\kappa}{2} v'' + \frac{2}{x} v' - \frac{2\beta}{x^2} v,$$

and hence ϕ satisfies the second order ODE

$$\phi''(y) + \left(\frac{4}{\kappa y} + \frac{4y}{\kappa}\right) \phi'(y) - \frac{4\beta}{\kappa y^2}\, \phi(y) = 0. \tag{4.1}$$

The two linear independent solutions can be given in terms of hypergeometric functions. However, since we are only interested in the behavior as $y \to 0+$, let us assume that we have a solution of the form $\phi(y) = y^{\nu-\beta} h(y)$ where h is smooth and nonzero at the origin and $\nu > \beta$. Then by plugging in and considering the lowest order term, we see that

$$(\nu - \beta)(\nu - \beta - 1) + \frac{4}{\kappa}(\nu - \beta) - \frac{4\beta}{\kappa} = 0,$$

i.e.,

$$\nu(\beta, \kappa) = \beta + \frac{\kappa - 4 + \sqrt{(\kappa - 4)^2 + 16\beta\kappa}}{2\kappa}. \tag{4.2}$$

If V is the event defined earlier in this subsection, then $\mathbf{P}(V)$ decays like $e^{-\nu(0,\kappa)L}$ as $L \to \infty$. From the formula we see that $\nu(0,\kappa) = 0$ for $\kappa \le 4$ (which, of course, is obvious since $P(V) = 1$), but $\nu(0,\kappa) > 0$ for $\kappa > 4$.

When locality holds, we can derive some stronger results. Note that

$$\nu(\beta, 6) = \frac{6\beta + 1 + \sqrt{1 + 24\beta}}{6}. \tag{4.3}$$

In particular, $\nu(0, 6) = 1/3$.

Corollary 4.1. *Suppose γ is a chordal SLE$_6$ path from πi to $L + \pi i$ in R_L ($L \ge 1$). Let $D \subset R_L$ be a simply connected subdomain with $[\pi i, L + \pi i] \subset \partial D$. Then*

$$\mathbf{P}\{\gamma[0,\infty) \cap R_L \subset D\} \asymp e^{-\mathcal{L}/3},$$

where $\mathcal{L}$ is the π-extremal distance between $\partial D \cap [0, \pi i]$ and $\partial D \cap [L, L + \pi i]$ in D.

Proof. The probability on the left is the same as the probability that γ reaches $[L, L + i\pi]$ before hitting $R_L \setminus D$. By Corollary 3.4, this probability is the same as the corresponding probability for chordal SLE$_6$ in D, and by conformal invariance it is the same as the probability that chordal SLE$_6$ in $R_{\mathcal{L}}$ from πi to $\mathcal{L} + \pi i$ reaches $[\mathcal{L}, \mathcal{L} + \pi i]$ before hitting $[0, \mathcal{L}]$. □

The exponent $\nu = \nu(\beta, \kappa)$ can be considered a *one-sided* crossing exponent as opposed to the *two-sided* exponent $\nu^* = \nu^*(\beta_1, \beta_2, \kappa)$ that we describe now. Suppose γ is a chordal SLE$_\kappa$ path connecting $(\pi/2)i$ and $L + (\pi/2)i$ in R_L and let T be the first time that it reaches $[L, L + \pi i]$. . Let $\mathcal{D}_+$ (resp. $\mathcal{D}_-$) be the connected component of $R_L \setminus \gamma[0, T]$ whose boundary contains $[\pi i, L + \pi i]$ (resp. $[0, L]$), and let $\mathcal{L}_+$ ($\mathcal{L}_-$) denote the π-extremal distance between $[0, \pi i]$ and $[L, L + \pi i]$ in $\mathcal{D}_+$ ($\mathcal{D}_-$). (If $\gamma[0, T]$ hits one of the horizontal boundaries, then one of these domains is not well defined; in this case, we say that the π-extremal distance is infinite.) The exponent ν^* is defined by the relation

$$\mathbf{E}[\exp\{-\beta_1 \mathcal{L}_+ - \beta_2 \mathcal{L}_-\}] \asymp e^{-\nu^* L}, \quad L \to \infty.$$

This exponent can be calculated in a similar way [11],

$$\nu^* = \nu^*(\beta_1, \beta_2, \kappa) = \frac{\left(\sqrt{(\kappa-4)^2 + 16\kappa\beta_1} + \sqrt{(\kappa-4)^2 + 16\kappa\beta_2} + \kappa\right)^2 - (8-\kappa)^2}{16\kappa}.$$

In particular,

$$\nu^*(0, 0, 6) = 1, \tag{4.4}$$

i.e., the probability that an SLE$_6$ path crosses the rectangle R_L without hitting either horizontal boundary decays like e^{-L}.

4.2 Crossing exponent for radial SLE

For any L, let $\mathcal{A}_L$ be the annulus $\{e^{-L} < |z| < 1\}$, with boundaries $C := \partial\mathbb{D}$ and $C_L := \partial(e^{-L}\mathbb{D})$. Consider radial SLE_κ started at 1 going towards the origin, and let

$$T = T_L = \inf\{t : \gamma(t) \in C_L\}.$$

Let $\mathcal{D}$ be the connected component of $\mathcal{A}_L \setminus \gamma[0, T]$ whose boundary contains C_L, and let $V = V_L$ be the event that $\partial\mathcal{D}$ also contains a nontrivial subarc of C. As in the previous section, $\mathbf{P}(V) = 1$ if and only if $\kappa \le 4$. Let $\mathcal{L}$ denote the π-extremal distance between C_L and $C \cap \partial\mathbb{D}$ in $\mathcal{A}$, where $\mathcal{L} = \infty$ on V^c. Define the *radial crossing exponent* $\zeta = \zeta(\beta, \kappa)$ by the relation

$$\mathbf{E}[\exp\{-\beta\mathcal{L}\}] \approx e^{-\zeta L}, \quad L \to \infty.$$

This can be reduced in a similar way to an expectation involving a derivative. In particular, we replace $e^{-\mathcal{L}}$ with $g_T'(e^{ix})$ for some $x \in (0, 2\pi)$. Since $g_T'(0) = e^T$, it follows from (2.1) that $|T - L| \le \log 4$. Hence, it will suffice to consider

$$u(t, x) = u_{\kappa,\beta}(t, x) = \mathbf{E}\,[\, g_t'(e^{ix})^\beta\,] = \mathbf{E}\,[\,\exp\{\beta \log g_t'(e^{ix})\}\,],$$

and to find the ζ such that $u(t, \pi) \asymp e^{-\zeta t}$ as $t \to \infty$. As in Section 3.2, we let $g_t(e^{ix}) = e^{iX_t}$ and $Z_t = X_t - U_t$ so that Z_t satisfies

$$dZ_t = \cot\left(\frac{Z_t}{2}\right) dt - \sqrt{\kappa}\, dB_t.$$

By differentiating the Loewner equation (2.11) with respect to x we get

$$\partial_t \log g_t'(e^{ix}) = -\frac{1}{2}\csc^2\left(\frac{Z_t}{2}\right),$$

i.e.,

$$\log g_t'(e^{ix}) = -\frac{1}{2}\int_0^t \csc^2\left(\frac{Z_s}{2}\right) ds,$$

Using the Feynman–Kac formula, we get that $u(t, x)$ satisfies

$$\partial_t u = \frac{\kappa}{2}u'' + \cot(x/2)u' - \frac{\beta}{2}\csc^2(x/2)u = 0.$$

For large times t, the solution of this equation looks like $e^{-\zeta t}\phi(x)$ where ϕ is a positive solution to

$$\frac{\kappa}{2}\phi'' + \cot(x/2)\phi' - \frac{\beta}{2}\csc^2(x/2)\phi + \zeta\phi = 0 \tag{4.5}$$

with $\phi(0) = \phi(2\pi) = 0$. Making a natural guess $\phi(x) = \sin^b(x/2)$ and plugging in we see that the only positive value for b is the same one as the value of $\nu - \beta$ in (4.2).

For this b, we solve for ζ,

$$\zeta(\beta,\kappa) \;=\; \frac{\beta}{2}+\frac{\kappa b}{8} \;=\; \frac{\beta}{2}+\frac{\kappa-4+\sqrt{(\kappa-4)^2+16\beta\kappa}}{16}.$$

One case of the last formula is well known. If $\kappa = 0$, the path γ is a radius pointing towards the origin from 1 to e^{-L}. In this case the domain $\mathcal{D} = \mathcal{A}_L\backslash[e^{-L}, 1]$ is conformally equivalent to $R_{L/2}$, and hence $e^{-\beta\mathcal{L}} = e^{-(\beta/2)L}$. The Beurling projection theorem (see, e.g.. [2, Theorem V.4.1]) implies that for any curve γ connecting C to C_L in $\mathcal{A}_L$, $e^{-\mathcal{L}} \geq e^{-L/2}$. From this we could have deduced initially that $\zeta(\beta,\kappa) \geq \beta/2$. If V is the event described earlier in this section, then $\mathbf{P}(V) \asymp e^{-\zeta(0,\kappa)L}$, with

$$\zeta(0,\kappa) = \begin{cases} (\kappa-4)/8, & \kappa > 4,\\ 0, & \kappa \leq 4.\end{cases}$$

In the case $\kappa = 6$, we have

$$\zeta(\beta,6) = \frac{\beta}{2}+\frac{1}{8}+\frac{1}{8}\sqrt{1+24\beta}.$$

4.3 Brownian excursions and exponents

If D is a simply connected domain other than the entire plane and $z \in D$, let $\mu_{D,z}$ denote the probability measure on paths obtained by starting a complex valued Brownian motion at z and stopping the path when it reaches the boundary. Throughout this section, two paths will be considered equivalent if one is a reparameterization of the other. Under this equivalence, the measure $\mu_{D,z}$ is conformally invariant, i.e., if $f : D \to D'$ is a conformal transformation, then $f \circ \mu_{D,z} = \mu_{f(D),f(z)}$ (see [2, Chapter V], e.g., for a discussion of conformal invariance properties of Brownian motion). If $z \in \partial D$ and ∂D is smooth near z, we define the *excursion measure* $\mu_{D,z}$ starting at z and ending at ∂D by

$$\mu_{D,z} = \lim_{\epsilon\to 0+} \epsilon^{-1}\mu_{D,z+\epsilon n}.$$

Here n is the unit normal pointing into D at z. Note that $\mu_{D,z}$ is an infinite measure, but if A is a nontrivial closed subarc of ∂D not containing z, then the $\mu_{D,z}$ measure of paths that leave D at A is positive and finite. This measure transforms under conformal transformations f by $f\circ\mu_{D,z} = |f'(z)|\,\mu_{f(D),f(z)}$. (The correction term comes from the fact that $|f(z+\epsilon n) - f(z)| \sim |f'(z)|\epsilon$.) The excursion measure on D is defined (for piecewise smooth ∂D) by

$$\mu_D = \int_{\partial D} \mu_{D,z}\,|dz|.$$

From the transformation rule for $\mu_{D,z}$ we can see that this measure is conformally invariant, $f \circ \mu_D = \mu_{f(D)}$. If A_1, A_2 are disjoint subarcs of ∂D, let $\mu_{D;A_1,A_2}$ be μ_D restricted to paths that start on A_1 and end at A_2; the total mass $|\mu_{D;A_1,A_2}|$ is a

conformal invariant. If D is the rectangle R_L and A_1, A_2 are the vertical boundaries, it is straightforward to check that $|\mu_{D;A_1,A_2}| \asymp e^{-L}$. Since the excursion measure is conformally invariant, we can conclude that

$$|\mu_{D;A_1,A_2}| \asymp \exp\{-\mathcal{L}(D; A_1, A_2)\}, \quad \mathcal{L}(D; A_1, A_2) \to \infty, \tag{4.6}$$

where $\mathcal{L}(D; A_1, A_2)$ denotes the π-extremal distance between A_1 and A_2 in D.

It is immediate from the definition that the excursion measure satisfies the following restriction principle. Suppose $D' \subset D$ are two simply connected domains and A_1, A_2 are disjoint subarcs of both $\partial D'$ and ∂D. Then $\mu_{D';A_1,A_2}$ is the same as $\mu_{D;A_1,A_2}$ restricted to paths that stay in D'.

If $\omega[0,t]$ is a path in R_L connecting the two vertical boundaries A_1, A_2, let D_- be the connected component of $R_L \setminus \omega[0,t]$ that contains $[0, L]$ on its boundary. Let $\mathcal{L}_-$ denote the π-extremal distance between $A_1 \cap \partial D_-$ and $A_2 \cap \partial D_-$ in D_-. The *half-space* or *rectangle Brownian intersection exponent* $\tilde{\xi}(1,\lambda)$ is defined by the relation

$$\mu_{R_L;A_1,A_2}\,[\,\exp\{-\lambda\mathcal{L}_-\}\,] \asymp e^{-\tilde{\xi}(1,\lambda)L}, \quad L \to \infty. \tag{4.7}$$

(We are using the shorthand notation $\mu(f) = \int f\, d\mu$.) It is not hard to prove that a $\tilde{\xi}(1,\lambda)$ satisfying (4.7) exists, but it is a challenge to find the value. Similarly, if we let D_+, $\mathcal{L}_+$ be the corresponding quantities for the component of R_L whose boundary includes $[\pi i, L + \pi i]$, we can define $\tilde{\xi}(\lambda_1, 1, \lambda_2)$ by

$$\mu_{R_L;A_1,A_2}\,[\,\exp\{-\lambda_1\mathcal{L}_+ - \lambda_2\mathcal{L}_-\}\,] \asymp e^{-\tilde{\xi}(\lambda_1,1,\lambda_2)L}, \quad L \to \infty.$$

In [16] it is shown how to define $\tilde{\xi}(a_1, \ldots, a_n)$ for any nonnegative real $a_1, \ldots, a_n$, so that $\tilde{\xi}$ is symmetric and satisfies the "cascade relation"

$$\tilde{\xi}(a_1, \ldots, a_n) = \tilde{\xi}(a_1, \ldots, a_j, \tilde{\xi}(a_{j+1}, \ldots, a_n)).$$

The Brownian excursion measure $\mu_{\mathcal{A}_L;C,C_L}$ can also be defined on the annulus $\mathcal{A}_L$ for curves starting on C and ending at C_L. It is easy to check that $|\mu_{\mathcal{A}_L;C;C_L}| \asymp 1/L$. Suppose $\gamma[0,t]$ is a curve in $\mathcal{A}_L$ connecting C and C_L, and let D be the connected component of $\mathcal{A}_L \setminus \gamma[0,t]$ whose boundary contains C_L. We say that $\gamma[0,t]$ is *nondisconnecting* if ∂D also contains C. If $\gamma[0,t]$ is nondisconnecting, we let $\mathcal{L}$ be the π-extremal distance between C and C_L in D; if γ is disconnecting, then $\mathcal{L} = \infty$. The (*whole space*) *Brownian intersection exponent* $\xi(1,\lambda)$ is defined by

$$\mu_{\mathcal{A}_L;C,C_L}\,[\,\exp\{-\lambda L\}\,] \asymp e^{-\xi(1,\lambda)L}, \quad L \to \infty,$$

where again we use the convention $e^{-0\cdot\infty} = 0$. More generally, if k is a positive integer, and $\gamma^1[0,t_1], \ldots, \gamma^k[0,t_k]$ are curves in $\mathcal{A}_L$ connecting C and C_L, we call the k-tuple nondisconnecting if there exists a component D of $\mathcal{A}_L \setminus (\gamma^1[0,t_1] \cup \cdots \cup \gamma^k[0,t_k])$ whose boundary includes nontrivial arcs in both C and C_L. If the k-tuple is nondisconnecting we let $\mathcal{L}$ be the π-extremal distance between the two circles in this domain D (if, by chance, there is more than one D, choose the D with smallest

π-extremal distance). Then the intersection exponent $\xi(k, \lambda)$ is defined by

$$\mu_{\mathcal{A}_L;C,C_L} \times \cdots \times \mu_{\mathcal{A}_L;C,C_L} [\, \exp\{-\lambda L\}\,] \asymp e^{-\xi(k,\lambda)L}, \quad L \to \infty.$$

The exponent $\xi(k, 0) = \xi(k, 0+)$ is sometimes called the *k-disconnection exponent*. Again, one can prove directly that such exponents exist. Also, these exponents are directly related to fractal properties of a Brownian path. For example, if B_t, $0 \le t \le 1$, is a complex Brownian motion then with probability one the Hausdorff dimension of the set of cut points is $2 - \xi(1, 1)$ and the Hausdorff dimension of the frontier (outer boundary) is $2 - \xi(2, 0)$ [7, 8]. These facts were proved without knowing the values of $\xi(1, 1), \xi(2, 0)$.

Analogously, we can define $\xi(j_1, \lambda_1, \ldots, j_n, \lambda_n)$ for any positive integers $j_1, \ldots, j_n$ and nonnegative $\lambda_1, \ldots, \lambda_n$. In [16], the definition of $\xi(a_1, \ldots, a_n)$ is extended to nonnegative reals (at least two of which are at least 1) as a symmetric function satisfying the "cascade relation"

$$\xi(a_1, \ldots, a_n) = \xi(a_1, \ldots, a_j, \tilde{\xi}(a_{j+1}, \ldots, a_n)).$$

4.4 Computing Brownian exponents

In this section we describe how to compute the values of the Brownian intersection exponent from the values of SLE_6 crossing exponents. Let γ be the path of an SLE_6 curve from πi to $L + \pi i$ in the rectangle R_L, and let $\hat{T}$, V, $\mathcal{D}$, $\hat{\mathcal{L}}$ be as in Section 4.1. Let μ_L denote the Brownian excursion measure on paths in R_L connecting $[0, \pi i]$ to $[L, L + \pi i]$. We will use ω to denote the image of such an excursion in R_L. We can also consider the Brownian excursion measure of paths from $A_1 := [0, \pi i] \cap \partial\mathcal{D}$ to $A_2 := [L, L + \pi i] \cap \partial\mathcal{D}$ in $\mathcal{D}$. By the restriction property this is the same as $1_{\omega \subset \mathcal{D}}\, \mu_D$. Let U be the set of paths ω contained in D; note that $\mu_D(U) \asymp e^{-\tilde{\mathcal{L}}}$. Let $\mathcal{D}_+$ be the connected component of $\mathcal{D} \setminus \omega$ whose boundary includes part of $\gamma(0, T)$ and let $\mathcal{D}_-$ be the component of $\mathcal{D} \setminus \omega$ whose boundary includes $[0, L]$. Let $\mathcal{L}_+$, $\mathcal{L}_-$ denote the respective π-extremal distances between $[0, \pi i]$ and $[L, L + \pi i]$ in these domains. By conformal invariance and symmetry, the conditional distributions of $\mathcal{L}_+$ and $\mathcal{L}_-$ given $\mathcal{D}$ are the same. Consider the exponent $\alpha = \alpha(\lambda)$, defined by

$$(\mathbf{P} \times \mu_L)[\exp\{-\lambda\mathcal{L}_+\}] = (\mathbf{P} \times \mu_L)[\exp\{-\lambda\mathcal{L}_-\}] \asymp e^{-\alpha L}, \quad L \to \infty.$$

Here $\mathbf{P}$ denotes the probability measure for the SLE_6 path γ. We will find two different expressions for α: one by choosing ω first and the other by choosing γ first.

Given ω, let $\tilde{\mathcal{D}}_+$ be the connected component of $R_L \setminus \omega$ whose boundary includes $[\pi i, L + \pi i]$ and let $\tilde{\mathcal{L}}_+$ denote the corresponding π-extremal distance between the vertical boundaries. By Corollary 4.1, the probability that $\gamma(0, \hat{T})$ lies in $\tilde{\mathcal{D}}_+$ is comparable to $e^{-\tilde{\mathcal{L}}_+/3}$. Hence,

$$(\mathbf{P} \times \mu_L)[\exp\{-\lambda\mathcal{L}_-\}] \asymp \mu[\exp\{-\tilde{\mathcal{L}}_+/3\} \exp\{-\lambda\mathcal{L}_-\}] \asymp e^{-\tilde{\xi}(1/3,1,\lambda)L}.$$

Therefore $\alpha = \tilde{\xi}(1/3, 1, \lambda)$. The cascade relation for the Brownian intersection exponent shows that $\tilde{\xi}(1/3, 1, \lambda) = \tilde{\xi}(1/3, \tilde{\xi}(1, \lambda))$.

However, given γ, and hence $\mathcal{D}$ and $\mathcal{L}$, the conditional expectation

$$\mu_{\mathcal{D}}[\exp\{-\lambda\mathcal{L}_-\} \mid \gamma] \asymp \exp\{-\tilde{\xi}(1, \lambda)\mathcal{L}\}.$$

Therefore,

$$(\mathbf{P} \times \mu_L)[\exp\{-\lambda\mathcal{L}_-\}] \asymp \mathbf{E}[\exp\{-\tilde{\xi}(1, \lambda)\mathcal{L}\}],$$

where $\mathbf{E}$ denotes expectation over the SLE_6 path γ. The right hand side is comparable to $e^{-\nu(\tilde{\xi}(1,\lambda),6)L}$ where $\nu(\beta, \kappa)$ is as defined in Section 4.1. Hence $\alpha = \nu(\tilde{\xi}(1, \lambda), 6)$. By letting λ vary from 0 to infinity, we can conclude (see (4.3))

$$\tilde{\xi}(1/3, s) = \nu(s, 6) = \frac{6s + 1 + \sqrt{1 + 24s}}{6}, \quad s \geq 1.$$

Note that $\tilde{\xi}(1/3, 1/3) = 1$. Using the cascade relation, we can conclude that

$$\tilde{\xi}(1, s) = \tilde{\xi}(\tilde{\xi}(1/3, 1/3), s) = \tilde{\xi}(1/3, \tilde{\xi}(1/3, s)).$$

We will do a similar argument in the radial case. Let us write $\hat{\mu}_L$ for $\mu_{\mathcal{A}_L;C,C_L}$. Let γ be a radial SLE_6 path from 1 to the origin and let ω^1, ω^2 be the paths of two independent excursions in $\mathcal{A}_L$ from C to C_L. Consider the event $V = V_L$ that $\gamma, \omega^1, \omega^2$ are disjoint. On the event V, let $\mathcal{D}^1, \mathcal{D}^2$ be the two connected components of $\mathcal{A}_L \setminus (\omega^1 \cup \omega^2)$ whose boundary intersects both C and C_L. Choose $\mathcal{D}^1$ so that $1 \in \partial\mathcal{D}^1$ (and hence $\gamma \cap \mathcal{A}_L \subset \mathcal{D}^1$). Let $\mathcal{L}^1, \mathcal{L}^2$ denote the π-extremal distances between C and C_L in $\mathcal{D}^1, \mathcal{D}^2$, respectively. We will be interested in the $\alpha = \alpha(\lambda)$ such that

$$\mathbf{P} \times \hat{\mu}_L \times \hat{\mu}_L \, [\, 1_V \, \exp\{-\lambda\mathcal{L}^2\} \,] \approx e^{-\alpha L}.$$

Here $\mathbf{P}$ denotes the probability for the path γ.

If we fix γ first we see that

$$\alpha = \zeta(\tilde{\xi}(1, \lambda, 1), 6) = \frac{\tilde{\xi}(1, \lambda, 1)}{2} + \frac{1}{8} + \frac{1}{8}\sqrt{1 + 24\tilde{\xi}(1, \lambda, 1)}.$$

However, if we fix ω^1, ω^2 first we get

$$\alpha = \zeta(b, 1, \lambda, 1),$$

where b is the exponent for SLE_6 defined by saying the probability that an SLE_6 path in R_L from $(\pi/2)i$ to $L + (\pi/2)i$ does not intersect $[0, L] \cup [\pi i, L + \pi i]$ decays like e^{-bL}. This exponent can be computed (see (4.4), $b = 1$. Hence $\alpha = \zeta(1, 1, \lambda, 1) = \xi(1, \tilde{\xi}(1, \lambda, 1))$. Letting $s = \tilde{\xi}(1, \lambda, 1)$, we see that

$$\xi(1, s) = \frac{s}{2} + \frac{1}{8} + \frac{1}{8}\sqrt{1 + 24s}.$$

Actually, this argument only establishes the above relation for all $s \geq \tilde{\xi}(1, 0, 1) = 2$. However, it can be shown [12] that $\xi(1, s)$ is real analytic for $s > 0$ and continuous at 0, so this holds for all $s \geq 0$. Similarly,

$$\xi(2, s) = \xi(\tilde{\xi}(1, 1/3), s) = \xi(1, \tilde{\xi}(1/3, 2)) = \frac{s}{2} + \frac{11}{24} + \frac{5}{24}\sqrt{1 + 24s}.$$

In particular, the dimension of the set of cut points of a Brownian path is $2 - \xi(1, 1) = 3/4$, and the dimension of the frontier is $2 - \xi(2, 0) = 4/3$.

Acknowledgment. Most of what I know about SLE has been learned from my collaboration with Oded Schramm and Wendelin Werner. Needless to say, without them this paper would not exist. I also thank them for their useful comments on this paper. I also thank all those who made comments and pointed out errors in an earlier draft of this paper; special thanks go to Jean Bricmont, Svante Jansson, Michael Kozdron, Antti Kupiainen, Carl Mueller, Jeff Steif, and the referee. Much of this paper was written while the author was visiting the Erwin Schrödinger Institute for Mathematical Physics. The author was partially supported by the National Science Foundation.

References

[1] L. Ahlfors, Complex Analysis. An Introduction to the Theory of Analytic Functions of One Complex Variable. Third edition, International Series in Pure and Applied Mathematics, McGraw-Hill, New York 1978.

[2] R. F. Bass, Probabilistic Techniques in Analysis, Probab. Appl. (New York), Springer-Verlag, New York 1995.

[3] P. L. Duren, Univalent Functions, Grundlehren Math. Wiss. 259, Springer-Verlag, New York 1983.

[4] C. Itzykson and J.-M. Drouffe, Statistical Field Theory. Vol. 2. Strong Coupling, Monte Carlo Methods, Conformal Field Theory, and Random Systems, Cambridge Monogr. Math. Phys., Cambridge University Press, Cambridge 1989.

[5] I. Karatzas and S. Shreve, Brownian Motion and Stochastic Calculus. Second edition, Grad. Texts in Math. 113, Springer-Verlag, New York 1991.

[6] G. F. Lawler, Intersections of Random Walks, Probability and its Applications, Birkhäuser, Boston, MA, 1991.

[7] G. F. Lawler, Hausdorff dimension of cut points for Brownian motion, Electron. J. Probab. 1 (1996) (electronic).

[8] G. F. Lawler (1996), The dimension of the frontier of planar Brownian motion, Electron. Comm. Probab. 1 (1996), 29–47 (electronic).

[9] G.F. Lawler, O. Schramm, W. Werner, Values of Brownian intersection exponents. I. Half-plane exponents, Acta. Math. 187 (2001), 237–273.

[10] G. F. Lawler, O. Schramm, W. Werner, Values of Brownian intersection exponents. II. Plane exponents, Acta Math. 187 (2001), 275–308.

[11] G. F. Lawler, O. Schramm, W. Werner, Values of Brownian intersection exponents. III. Two sided exponents, Ann. Inst. H. Poincaré Probab. Statist. 38 (2002), 109–123.

[12] G. F. Lawler, O. Schramm, W. Werner, Analyticity of intersection exponents for planar Brownian motion, Acta Math. 189 (2002), 179–201.

[13] G. F. Lawler, O. Schramm, W. Werner, Conformal invariance of planar loop-erased random walk and uniform spanning trees, preprint (2002).

[14] G. F. Lawler, O. Schramm, W. Werner, On the scaling limit of planar self-avoiding walk, preprint (2002).

[15] G. F. Lawler, O. Schramm, W. Werner, Conformal restriction properties, in preparation.

[16] G. F. Lawler, W. Werner, Intersection exponents for planar Brownian motion, Ann. Probab. 27 (1999), 1601–1642.

[17] G. F. Lawler, W. Werner, Universality for conformally invariant intersection exponents, J. Eur. Math. Soc. (JEMS) 2 (2000), 291–328.

[18] N. Madras and G. Slade, The Self-Avoiding Walk, Probab. Appl., Birkhäuser, Boston, MA, 1993.

[19] D. Marshall and S. Rohde, The Loewner differential equation and slit mappings, preprint (2001).

[20] S. Rohde and O. Schramm, Basic properties of SLE, preprint (2001).

[21] O. Schramm, Scaling limits of loop-erased random walks and uniform spanning trees, Israel J. Math. 118 (2000), 221–288.

[22] S. Smirnov, Critical percolation in the plane: conformal invariance, Cardy's formula, scaling limits, C. R. Acad. Sci. Paris Sér. I Math. 333 (2001), 239–244.

[23] W. Werner, Critical exponents, conformal invariance and planar Brownian motion, in: European Congress of Mathematics, Vol. II (Barcelona, 2000), Progr. Math. 202, Birkhäuser, Basel 2001, 87–103.

Gregory F. Lawler, Department of Mathematics, Duke University, Durham, NC 27708-0320 and Department of Mathematics, Cornell University, Ithaca, NY 14853-4201

E-mail: jose@math.duke.edu, lawler@math.cornell.edu

A canonical form for automorphisms of totally disconnected locally compact groups

George A. Willis

Abstract. Automorphisms of totally disconnected locally compact groups have a canonical form described in terms of the action of the automorphism on certain compact open subgroups known as *tidy* subgroups. Identifying tidy subgroups is analogous to finding a basis which triangularises a linear transformation and the canonical form for automorphisms of general totally disconnected groups provides analogues of the eigenspaces and eigenvalues of automorphisms of the Lie algebra of a Lie group. These ideas have been used to answer some long-standing questions about locally compact groups and potential further applications are discussed at the end of the paper.

Contents

1 Introduction 295

2 Examples of totally disconnected groups 296

3 Tidy subgroups and the scale of an automorphism 299
3.1 The tidying procedure 303
3.2 Dynamical description of tidy subgroups 304

4 The scale function on G 305

5 Tidy subgroups for groups of automorphisms 307

6 Applications and further developments 309

1 Introduction

Each locally compact group, G, is a canonical extension of its connected component of the identity, G_0, by the totally disconnected group G/G_0, [25, Theorem II.7.3]. Questions concerning the structure of locally compact groups may therefore be dealt with by treating separately the cases where G is connected and where G is totally disconnected and then combining the two answers.

Connected groups can be approximated by Lie groups in the sense that, if G is connected, then each neighbourhood of the identity in G contains a compact normal subgroup, K, such that G/K is a Lie group, [13, 35, 40]. Lie group techniques may then be used to analyse the structure of connected groups and their automorphisms, see for example [8, 21, 28, 33]. The power of the Lie techniques derives from the fact that they bring tools from linear algebra, such as the Jordan canonical form and eigenspaces and eigenvectors, to the analysis of group automorphisms.

Totally disconnected groups, on the other hand, are not so well understood. However, a canonical form for automorphisms of totally disconnected locally compact groups has been developed in [53, 56, 58]. It provides analogues of eigenspaces and eigenvalues of automorphisms and thus promises to play a similar role in the study of totally disconnected groups as approximation by Lie groups does in the case of connected groups.

The main part of the paper explains this canonical form for automorphisms of totally disconnected groups. The last section describes some applications of the theory as developed so far, including one to random walks, and indicates directions for further research. The class of groups to which these new techniques apply is very large and we begin, in the next section, by describing some examples which will illustrate the ideas.

2 Examples of totally disconnected groups

The inductive dimension of a totally disconnected locally compact Hausdorff space is 0, see [25, Theorem 3.5]. Totally disconnected groups are consequently also known as 0-dimensional groups. This class includes, of course, all discrete groups. The results described below do not directly say anything new about discrete groups and so, unless stated otherwise, G will denote throughout a *non-discrete* totally disconnected locally compact group. Here are some examples of such groups.

Example 2.1. (a) Let $\{F_\alpha\}_{\alpha\in A}$ be a family of finite groups. Then $\prod_{\alpha\in A} F_\alpha$ becomes a compact totally disconnected group when equipped with the product topology and group operations. Any closed subgroup of such a product is also a compact totally disconnected group.

It will be seen later, in Corollary 3.2, that every compact totally disconnected group is profinite and is therefore isomorphic to a closed subgroup of such a product of finite groups.

(b) Any Lie group over a non-discrete totally disconnected locally compact field is a totally disconnected locally compact group. These fields have been completely classified up to isomorphism by van Dantzig, see for example [9, 29, 30, 52], and fall into two families, namely:

(i) the fields of p-adic numbers, $\mathbb{Q}_p$ and their finite extensions; and

(ii) the fields of formal Laurent series, $\boldsymbol{k}((X))$, over a finite field $\boldsymbol{k}$.

Non-discrete characteristic 0 totally disconnected fields are in the first family and non-discrete finite characteristic fields are in the second family. Section 12.3.4 of [42] gives a good summary of the classification along with further references.

In particular, for each $n \geq 2$ the matrix group $GL(n, \mathbb{K})$, where $\mathbb{K}$ is either $\mathbb{Q}_p$ or $\boldsymbol{k}((X))$, with the topology inherited as a subset of $\mathbb{K}^{n^2}$ is a totally disconnected locally compact group.

Note that $\mathbb{Q}_p$ is locally compact because it has $\mathbb{Z}_p$, the ring of p-adic integers, as a compact open subring and $\boldsymbol{k}((X))$ is locally compact because it has $\boldsymbol{k}(X)$, the ring of formal Taylor series over $\boldsymbol{k}$, as a compact open subring. The Lie groups over the field are locally compact because they have neighbourhoods of the identity homeomorphic to $\mathbb{Z}_p^n$ or $\boldsymbol{k}(X)^n$ where n is the dimension of the Lie group. Unlike connected groups, this is not the same as the topological dimension of the group.

(c) Automorphism groups of connected locally finite graphs are totally disconnected when equipped with the compact-open topology, that is, the topology of uniform convergence on compact sets. Let $\mathcal{G}$ be a locally finite graph. Then a base of neighbourhoods for the compact-open topology on $\mathrm{Aut}(\mathcal{G})$ is given by sets of the form

$$\mathcal{N}(x; F) := \left\{ y \in \mathrm{Aut}(\mathcal{G}) \mid y(v) = x(v),\ v \in F \right\},$$

where $x \in \mathrm{Aut}(\mathcal{G})$ and F is a finite subset of $V(\mathcal{G})$.

Note that any automorphism of $\mathcal{G}$ which stabilises a vertex v will permute $B(v, n)$ for each n, where $B(v, n)$ denotes the (finite) set of vertices of distance n from v. Hence $\mathcal{N}(e; \{v\})$ is isomorphic to a closed subgroup of the product of finite permutation groups $\prod_n S_{B(v,n)}$. Therefore $\mathcal{N}(e; \{v\})$ is a compact open neighbourhood of e and $\mathcal{G}$ is locally compact.

(d) A subgroup H of a group G is said to be *almost normal* if the H-orbits in G/H are finite. Almost normal subgroups have been studied recently in connection their associated *Hecke algebra* and its enveloping C*-algebra, [4, 7, 22, 38], and in this context (G, H) is said to be a *Hecke pair*. Hecke algebras consist of functions on the double coset space $H\backslash G/H$ and the product is defined by means of counting H-orbits in G/H.

(i) If U is a compact open subgroup of the totally disconnected group G, then G/U is discrete because U is open. Since U is also compact, it then follows that U-orbits in G/U are finite. Hence U is almost normal in G.

(ii) Conversely, if H is almost normal in G, then there is a totally disconnected locally compact group, $\widetilde{G}$, and a homomorphism $\varphi : G \to \widetilde{G}$, having dense range, such that $\overline{\varphi(H)}$ is compact and open and the G-space $\widetilde{G}/\overline{\varphi(H)}$ is equivalent to G/H. (In the construction of $\widetilde{G}$, $\overline{\varphi(H)}$ is seen to be compact because it is isomorphic to a closed subgroup of the product of the finite permutation groups of the H-orbits in G/H.)

In view of (i) and (ii), the study of almost normal subgroups thus reduces to that of compact open subgroups of totally disconnected locally compact groups. It is shown by Tzanev, [51], who builds on the proof of (ii) in [47], that the Hecke algebra associated with the Hecke pair (G, H) can be realised as the subalgebra $m * C_{00}(\widetilde{G}) * m$ of the group convolution algebra $L^1(\widetilde{G})$, where m denotes the normalised Haar measure on the compact open subgroup $\overline{\varphi(H)}$. This can be used to give better proofs of some results about Hecke algebras, [17].

Of the above classes of examples, the compact groups and Lie groups are the best understood. The canonical form for automorphisms to be described does not add to our knowledge of these groups but, rather, extends Lie group techniques to general non-compact and non-discrete totally disconnected locally compact groups.

Note first of all that, if $G = \mathbb{Q}_p^n$ for some n and $\alpha \in \mathrm{Aut}(G)$, then α is a linear transformation of the vector space $\mathbb{Q}_p^n$. Hence, passing to a finite extension, L, of $\mathbb{Q}_p$ if necessary, α has eigenvalues and a Jordan canonical form. For each eigenvalue, λ, denote the generalised eigenspace in $\mathbb{Q}_p^n \otimes_{\mathbb{Q}_p} L$ corresponding to λ by $\mathfrak{V}_\lambda$ and define

$$\mathfrak{V}_+ := \mathbb{Q}_p^n \cap \bigoplus_{|\lambda|_p>1} \mathfrak{V}_\lambda,\ \mathfrak{V}_0 := \mathbb{Q}_p^n \cap \bigoplus_{|\lambda|_p=1} \mathfrak{V}_\lambda \text{ and } \mathfrak{V}_- := \mathbb{Q}_p^n \cap \bigoplus_{|\lambda|_p<1} \mathfrak{V}_\lambda. \quad (2.1)$$

Then α is expanding on $\mathfrak{V}_+$, contracting on $\mathfrak{V}_-$ and neutral on V_0, and $\mathbb{Q}_p^n$ may be decomposed as the direct sum

$$\mathbb{Q}_p^n = \mathfrak{V}_+ \oplus \mathfrak{V}_0 \oplus \mathfrak{V}_-. \quad (2.2)$$

In the next section it will be seen that automorphisms of general totally disconnected locally compact groups give rise to an analogous decomposition of the group into expanding and contracting subgroups. For single automorphisms there is no analogue of the further decomposition of these subgroups into eigenspaces but it will be seen in Section 5 that such an analogue can be produced by considering several commuting automorphisms at once.

In the case of non-abelian Lie groups the decomposition follows from the above example via the Lie algebra. Let G be a p-adic Lie group and $\mathfrak{g}$ be its Lie algebra. Then each continuous automorphism, $\alpha \in \mathrm{Aut}(G)$, induces a linear transformation, $\mathrm{ad}\,\alpha \in GL(\mathfrak{g})$. Applying the decomposition in (2.2) to $\mathrm{ad}\,\alpha$, $\mathfrak{g}$ may be decomposed into subalgebras

$$\mathfrak{g} = \mathfrak{g}_+ \oplus \mathfrak{g}_0 \oplus \mathfrak{g}_-, \quad (2.3)$$

where $\mathrm{ad}\,\alpha$ is respectively expanding, neutral and contracting on $\mathfrak{g}_+$, $\mathfrak{g}_0$ and $\mathfrak{g}_-$. This decomposition does not exponentiate to give a corresponding decomposition of the whole of G. However the exponential map is defined on some neighbourhood of 0 and the image of the neighbourhood is a compact open subgroup of G which decomposes into subgroups on which α is expanding, neutral and contracting.

In the general case the decomposition must be proved directly. It holds for certain compact open subgroups of G called the tidy subgroups.

3 Tidy subgroups and the scale of an automorphism

All the groups introduced in the previous section have compact open subgroups. One of the earliest results about topological groups shows that all totally disconnected locally compact groups possess many such subgroups.

Theorem 3.1 (van Dantzig, 1931). *Let G be a totally disconnected locally compact group and $\mathcal{O}$ be a neighbourhood of e. Then there is a compact open subgroup, U, contained in $\mathcal{O}$.*

The compact open subgroups guaranteed by the theorem need not be normal. Examples of groups having no normal compact open subgroups were given by van Dantzig at the same time as his theorem, but it turns out that many Lie groups and automorphism groups of locally finite graphs do not have compact open normal subgroups. Furthermore, the concept of almost normal subgroup is only of interest when the subgroup (H or $\overline{\phi(H)}$) is not normal.

When G is a compact group, each open subgroup U has finite index, and the kernel of the action of G on G/U is a compact open *normal* subgroup. The theorem thus has the following immediate consequence, [10].

Corollary 3.2 (van Dantzig, 1936). *Let G be compact and totally disconnected, and let $\mathcal{O}$ be a neighbourhood of e. Then there is a compact open normal subgroup U contained in $\mathcal{O}$. It follows that G is profinite and is isomorphic to a closed subgroup of a product of finite groups.*

The corollary describes compact totally disconnected groups completely. Non-compact totally disconnected groups cannot be described in a similar way, that is, as being prodiscrete, because they do not in general have normal compact open subgroups. However the almost normality of compact open subgroups yields information about inner automorphisms and the structure of G. Most definitions and arguments apply equally well to general automorphisms and will, Section 4 excepted, be stated in these terms.

Let α be in $\mathrm{Aut}(G)$ and U be a compact open subgroup of G. Then

$$[\alpha(U) : U \cap \alpha(U)] < \infty \tag{3.1}$$

because $U \cap \alpha(U)$ is an open subgroup of the compact group $\alpha(U)$. (Finiteness of these indices for all inner automorphisms is equivalent to almost normality of U.)

Definition 3.3. Define the *scale of* α to be

$$s(\alpha) = \min\big\{[\alpha(U) : U \cap \alpha(U)] \mid U \text{ is a compact open subgroup of } G\big\}. \tag{3.2}$$

The subgroup U is *tidy for* α if this minimum is attained at U.

Note that $s(\alpha)$ is the minimum of a set of positive integers and so it is always attained. Hence subgroups tidy for α always exist. They are those which are closest to being invariant under α. Tidy subgroups have the following structural characterisation.

Theorem 3.4 (Structure of Tidy Subgroups). *Let W be a compact open subgroup of G and let $\alpha \in Aut(G)$. Define $W_\pm := \bigcap_{k\geq 0} \alpha^{\pm k}(W)$. Then W is tidy for α if and only if either*

(i) **T1:** $W = W_+W_-$, *and*

T2: $W_{++} := \bigcup_{k\geq 0} \alpha^k(W_+)$ *is closed;*

or

(ii) $s(\alpha) = [\alpha(W_+) : W_+]$.

The theorem combines results from [53] and [56]. It was shown in [53] that compact open subgroups satisfying **T1** and **T2** exist. The proof is an algorithm for producing such a subgroup from a given compact open subgroup. That these conditions are equivalent to the minimising condition in Definition 3.3 is shown in [56] by checking that a modified version of the algorithm reduces the index (3.1) at each step.

It follows from the definitions that $\alpha(W_+) \geq W_+$ and hence that W_{++} is a group. Similarly, $\alpha(W_-) \leq W_-$ and so **T1** expresses W as the product of a subgroup where α expands and a subgroup where α shrinks. In the case when G is a p-adic Lie group this factoring of W corresponds to the decomposition of the Lie algebra given in (2.3).

Part (ii) of the theorem shows that $s(\alpha)$ is the factor by which α "scales up" or expands W_+. For p-adic Lie groups and linear groups over skew fields, H. Glöckner has computed $s(\alpha)$ in terms of the valuations of the eigenvalues of $\operatorname{ad}\alpha$, see [14, 15]. The following calculation for a p-adic linear group is closer to the approach taken in [14]. This and Example 3.6 will illustrate Theorem 3.4. The algorithm entailed in the proof of the theorem will be stated and explained with the aid of these examples in Subsection 3.1.

Example 3.5. Let $G = \mathrm{SL}(2, \mathbb{Q}_p)$, and let α be the automorphism of G given by

$$\alpha : y \mapsto xyx^{-1}, \text{ where } x = \begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix}.$$

A natural choice of compact open subgroup of G is $U = \mathrm{SL}(2, \mathbb{Z}_p)$. We then have

$$\alpha(U) = \left\{ \begin{pmatrix} a & pb \\ p^{-1}c & d \end{pmatrix} \mid a, b, c, d \in \mathbb{Z}_p,\ ad - bc = 1 \right\}$$

and

$$U \cap \alpha(U) = \left\{ \begin{pmatrix} a & pb \\ c & d \end{pmatrix} \mid a, b, c, d \in \mathbb{Z}_p,\ ad - pbc = 1 \right\},$$

and it may be verified that coset representatives of $\alpha(U)$ over $U \cap \alpha(U)$ are

$$\begin{pmatrix} 1 & 0 \\ p^{-1}c_0 & 1 \end{pmatrix}, \text{ where } c_0 \in \{0, 1, \ldots, p-1\}, \text{ and } \begin{pmatrix} 0 & -p \\ p^{-1} & 0 \end{pmatrix}.$$

Hence $[\alpha(U) : U \cap \alpha(U)] = p + 1$. However,

$$U_+ = \left\{ \begin{pmatrix} a & 0 \\ c & d \end{pmatrix} \mid a, c, d \in \mathbb{Z}_p,\ ad = 1 \right\}$$
$$\text{and} \quad U_- = \left\{ \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \mid a, b, d \in \mathbb{Z}_p,\ ad = 1 \right\}.$$

It follows that U does not satisfy **T1** because $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ belongs to U but is not in U_+U_-.

Next, let

$$V = U \cap \alpha(U) = \left\{ \begin{pmatrix} a & pb \\ c & d \end{pmatrix} \mid a, b, c, d \in \mathbb{Z}_p,\ ad - pbc = 1 \right\}.$$

Then

$$V_+ = \left\{ \begin{pmatrix} a & 0 \\ c & d \end{pmatrix} \mid a, c, d \in \mathbb{Z}_p,\ ad = 1 \right\}$$
$$\text{and} \quad V_- = \left\{ \begin{pmatrix} a & pb \\ 0 & d \end{pmatrix} \mid a, b, d \in \mathbb{Z}_p,\ ad = 1 \right\}.$$

Any element of V must have $|a|_p = 1$, and so we may use Gaussian elimination and pivot on the top left entry without leaving $\mathrm{SL}(2, \mathbb{Z}_p)$. Doing so expresses each element of V as the product of a matrix in V_+ and one in V_-. Hence V satisfies **T1**. Since we also have

$$V_{++} = \left\{ \begin{pmatrix} a & 0 \\ c & d \end{pmatrix} \mid a, d \in \mathbb{Z}_p,\ ad = 1 \text{ and } c \in \mathbb{Q}_p \right\}$$

which is closed, V is tidy for α. It may be calculated that $[\alpha(V) : V \cap \alpha(V)] = [\alpha(V_+) : V_+] = p$ and this is the value of $s(\alpha)$.

The value of the index $[\alpha(V) : V \cap \alpha(V)]$ is reduced because the coset corresponding to $\begin{pmatrix} 0 & -p \\ p^{-1} & 0 \end{pmatrix} = \alpha \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ does not occur.

The calculation of $s(\alpha)$ in Example 3.5 is straightforward because the element x used to define α is diagonal. The occurrence in the example that the tidy subgroup V has the form $U \cap \alpha(U)$ where U is a given compact open subgroup is no accident. It is shown in [53] that, given a compact open subgroup U, there is an integer k such that the intersection $\bigcap_{n=0}^{k} \alpha^n(U)$ satisfies condition **T1**. Replacing U by this intersection is the first step of the algorithm for finding tidy subgroups.

Another example which may be calculated directly is when $G = \mathbb{Q}^n$, in which case automorphisms of G are linear transformations. If α is diagonal with respect to the natural basis, then $\mathbb{Z}_p^n$ is tidy for α, and

$$s(\alpha) = \prod \left\{ |\lambda|_p \mid \lambda \text{ is an eigenvalue and } |\lambda|_p > 1 \right\}.$$

Otherwise, α must be put into Jordan canonical form to identify tidy subgroups, see [12, 58].

More generally, the calculation in [15] is made by putting $\operatorname{ad}\alpha$ in Jordan canonical form and identifying the subspaces $\mathfrak{g}_+$, $\mathfrak{g}_0$ and $\mathfrak{g}_-$ defined in Equation (2.1). In [14] condition **T1** reduces to factorisation of elements of the tidy subgroup into upper and lower triangular matrices, just as in Example 3.5. Condition **T2** is automatic for Lie groups.

Condition **T2** is not always automatic, as the next example shows.

Example 3.6. Let $\mathcal{T}_q$ denote the homogeneous tree in which every vertex has degree $q+1$ and let $G = \operatorname{Aut}(\mathcal{T}_q)$. Choose an infinite path ℓ through the tree, and let x be a translation by one unit of the tree along ℓ, as illustrated with $q = 2$ in Figure 1. Let α be the inner automorphism of G given by conjugation by x.

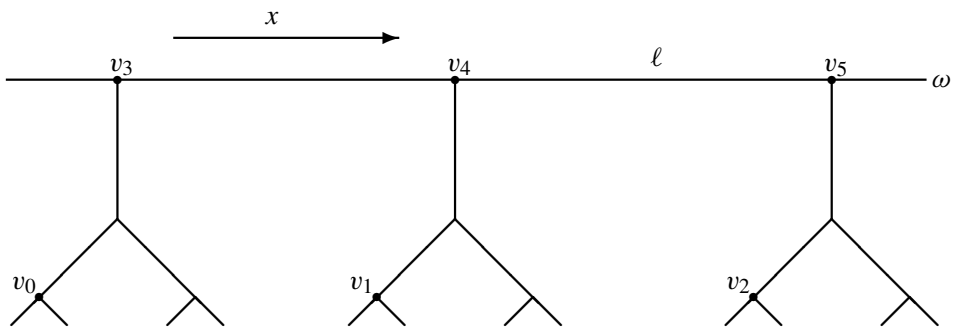


Figure 1. A translation automorphism of the tree $\mathcal{T}_2$

First, let $U = \mathcal{N}(e, v_0)$, the stabiliser of a vertex at distance 2 from ℓ. Then $\alpha(U)$ is the stabiliser of $v_1 := \alpha(v_0)$, and $U \cap \alpha(U)$ is the subgroup of G consisting of automorphisms of $\mathcal{T}_q$ which fix every vertex on the unique path from v_0 to v_1. Hence $[\alpha(U) : U \cap \alpha(U)] = (q+1)q^4$. Similar considerations show that U_+ consists of those automorphisms which fix all vertices on the subtree spanned by $\alpha^n(v_0)$, $n \geq 0$, and U_- consists of those fixing all vertices on the subtree spanned by $\alpha^n(v_0)$, $n \leq 0$. In particular, automorphisms in either U_+ or U_- fix all vertices on the path joining v_0 to ℓ, and so any automorphism in U_+U_- does the same. Since automorphisms in U may move vertices on this path, it follows that U does not satisfy condition **T1**.

Next, let $V = U \cap \alpha(U)$. Then $\alpha(V)$ fixes all vertices on the path from v_1 to $v_2 := \alpha^2(v_0)$, and $V \cap \alpha(V)$ fixes all vertices in the subtree spanned by v_0, v_1 and v_2. Hence $[\alpha(V) : V \cap \alpha(V)] = (q-1)q^2$. The index is reduced therefore if U is replaced by V. Also, V_+ consists of those automorphisms fixing all vertices in the subtree spanned by $\alpha^n(v_0)$, $n \geq 0$, and V_- consists of those fixing all vertices on the subtree spanned by $\alpha^n(v_1)$, $n \leq 0$. Let $\mathcal{T}_+$ be that part of $\mathcal{T}_q$ to the right of the edge (v_3, v_4), and $\mathcal{T}_-$ be that part to the left. Then each $y \in V$ leaves $\mathcal{T}_+$ and $\mathcal{T}_-$ invariant

and so is the product $y = y_+ y_-$, where y_+ agrees with y on $\mathcal{T}_+$ and fixes all vertices in $\mathcal{T}_-$, and y_- agrees with y on $\mathcal{T}_-$ and fixes all vertices in $\mathcal{T}_+$. Since $y_\pm \in V_\pm$, we have verified that V satisfies **T1**.

It may be seen that V_{++} is not closed however. Each $y \in \alpha^k(V_+)$ fixes all the vertices $\alpha^n(v_0)$, $n \geq k$. Hence each y in $V_{++} = \bigcup_{n\geq 0} \alpha^n(V_+)$ fixes the right hand end, ω, of ℓ. It is easily seen that V_{++} is in fact dense in S_ω, the stabiliser group of ω. Since S_ω contains automorphisms which move all vertices with distance 2 from ℓ and no element of V_{++} has this property, it follows that V_{++} is not closed and V does not satisfy condition **T2**.

Finally, let $W = \mathcal{N}(e, \{v_3, v_4\})$, the subgroup which fixes two adjacent vertices on ℓ. Then $\alpha(W) = \mathcal{N}(e, \{v_4, v_5\})$, and it follows that

$$W \cap \alpha(W) = \mathcal{N}(e, \{v_3, v_4, v_5\}) \quad \text{and} \quad [\alpha(W) : W \cap \alpha(W)] = q.$$

Similar reasoning to that used in the previous paragraph shows that in this case $W = W_+ W_-$ and $W_{++} = S_\omega$, which is closed. Therefore W is tidy for α and $s(\alpha) = q$.

It may be shown that the subgroups tidy for α all lie along ℓ in the sense that any tidy subgroup must fix at least two vertices on ℓ and not fix any vertex off ℓ (or at distance at least 2 from ℓ when $q = 2$).

3.1 The tidying procedure

Here is the algorithm for finding tidy subgroups. This procedure takes a given compact open subgroup U and "tidies" it to produce a subgroup W satisfying **T1** and **T2**. The index $[\alpha(U) : U \cap \alpha(U)]$ is reduced at each step.

Step 1: *There is an integer N such that $\bigcap_{n=0}^{N} \alpha^n(U)$ satisfies* **T1**. *Choose N to be the least such integer and put $V = \bigcap_{n=0}^{N} \alpha^n(U)$. Then*

$$[\alpha(V) : V \cap \alpha(V)] \leq [\alpha(U) : U \cap \alpha(U)]$$

with equality if and only if $N = 0$.

It may be seen that $N = 1$ in both the above examples. The number N may be bigger than 1 when G is a higher rank Lie group or the automorphism group of a higher rank building.

Step 2: *Form the group*

$$\mathcal{L} = \left\{ v \in G \mid \alpha^n(v) \in V \text{ for all but finitely many } n \in \mathbb{Z} \right\}$$

and put $L = \overline{\mathcal{L}}$. Then L is a compact group.

In Example 3.5 $\mathcal{L}$ is the group of diagonal matrices in $\mathrm{SL}(2, \mathbb{Z}_p)$ and is already closed. In Example 3.6

$$\mathcal{L} = \left\{ y \in \mathrm{Aut}(\mathcal{T}_q) \mid y \text{ fixes all but finitely many of the vertices } \alpha^n(v_0),\ n \in \mathbb{Z} \right\},$$

and L is the subgroup of $\mathrm{Aut}(\mathcal{T}_q)$ of automorphisms fixing all the vertices on ℓ.

Step 3: *Define* $V' = \left\{v \in V \mid lvl^{-1} \in VL,\ \forall l \in L\right\}$ *and put* $W = V'L$. *Then* W *is a compact open subgroup of* G *and is tidy for* α. *Furthermore*

$$[\alpha(W) : W \cap \alpha(W)] \leq [\alpha(V) : V \cap \alpha(V)]$$

with equality if and only if $\mathcal{L}$ *is closed. In the case when* $\mathcal{L}$ *is closed,* $L \leq V$ *and so* $W = V$.

In Example 3.5, and in any p-adic Lie group, $\mathcal{L}$ is closed and the procedure is complete after Step 1. The subgroup W in Example 3.6 is that produced by Step 3 from the given V.

A more natural choice of compact open subgroup to begin with in Example 3.6 might be thought to be $U'' = \mathcal{N}(e, \{v_3\})$, that is, the stabiliser of a vertex on ℓ. This group is not tidy for α because it does not satisfy **T1**. However $U'' \cap \alpha(U'') = W$ and is tidy. The group $\mathrm{SL}(2, \mathbb{Q}_p)$ has a natural action on $\mathcal{T}_p$, [48], and under this action U'' corresponds to the group $U = \mathrm{SL}(n, \mathbb{Z}_p)$ chosen in Example 3.5.

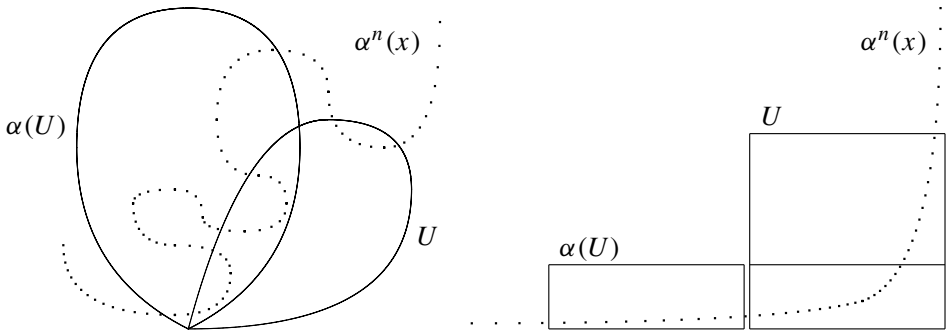
Note that Step 1 reduces U to obtain a subgroup satisfying **T1** while Step 3 increases it again to obtain a subgroup satisfying **T2** as well. At each step the index $[\alpha(U) : U \cap \alpha(U)]$ is not increased and is strictly decreased if and only if the desired property is not already satisfied. It is shown in [53] that $[\alpha(W) : W \cap \alpha(W)]$ is the same for all groups satisfying **T1** and **T2**. It follows that the tidying procedure finds a subgroup where this index is minimised.

3.2 Dynamical description of tidy subgroups

The properties **T1** and **T2** characterising tidy subgroups can be formulated in terms of the action of α.

General subgroups and the relations between them are often depicted as in Figure 3.2(i). If U is tidy however condition **T1** shows that the relation between U and $\alpha(U)$ is better depicted as in Figure 3.2(ii). In this picture U is the product of U_+ (on the bottom edge) and U_- (on the right side). When α is applied, U_+ expands and U_- contracts, and $\alpha(U)$ is the product of $\alpha(U_+)$ and $\alpha(U_-)$. Figure 3.2(ii) shows an automorphism α where $s(\alpha) = 2$ and $s(\alpha^{-1}) = 3$.

Condition **T2** is equivalent to a statement about α-orbits. For a tidy subgroup, U, each α-orbit can enter and leave U at most once, as shown in Figure 3.2(ii). The figure does not represent the orbit accurately because there will be at most one point in U but not in $\alpha(U)$. However, the orbits may be thought of as hyperbolic with asymptotes U_{--} and U_{++}. Of course an orbit might not enter U at all. It might also enter U at U_- and never leave or start in U_+ and then depart and never return. If U does not satisfy **T2**, then α-orbits may enter and leave any number of times.



(i) U is a general subgroup (ii) U is tidy

Figure 2. The action of α on a compact open subgroup U

4 The scale function on G

The scale is a function $s : \mathrm{Aut}(G) \to \mathbb{Z}^+$. Composing with the map $x \mapsto \alpha_x : G \to \mathrm{Aut}(G)$, where α_x is the inner automorphism $\alpha_x : y \mapsto xyx^{-1}$, yields a function on G which will also be denoted by s. It is shown in [53] that this function is continuous from the given locally compact topology on G to the discrete topology on $\mathbb{Z}^+$. This fact and other properties of the scale easily derived from Theorem 3.4 are collected in the next result.

Proposition 4.1. *Let $s : G \to \mathbb{Z}^+$ be the scale function. Then s is continuous and satisfies:*

S1: $s(x) = 1 = s(x^{-1})$ *if and only if there is a compact open subgroup U with $xUx^{-1} = U$;*

S2: $s(x^n) = s(x)^n$, *for every $x \in G$ and $n \geq 0$;*

S3: $\Delta(x) = s(x)/s(x^{-1})$, *where $\Delta : G \to \mathbb{Q}^+$ is the modular function on G;*

S4: $s(\alpha(x)) = s(x)$ *for every $x \in G$ and $\alpha \in \mathrm{Aut}(G)$.*

Note that property **S3** implies that the modular function on a totally disconnected group takes only rational values. This fact seems to have been noticed first by G. Schlichting, [46, Lemma 1]. Since the modular function is a homomorphism, $\Delta(G)$ is a finitely generated subgroup of $(\mathbb{Q}^+, \times)$ when G is compactly generated. In view of this, property **S3** suggests that $s(G)$ should have a finite number of prime divisors when G is compactly generated. The multiplicative property **S2** is not sufficiently strong to make that deduction in the same way as for the modular function but it can be shown, nevertheless, by direct calculation, [57].

In analogy with the term *unimodular* for a group where the modular function is identically 1, a group where the scale function is identically 1 is called *uniscalar*, [42, p. 1360]. The groups $\mathrm{SL}(n, \mathbb{Q}_p)$ and $\mathrm{Aut}(\mathcal{T}_q)$ are unimodular but it was seen in Examples 3.5 and 3.6 that they are not uniscalar. Any group which has a compact open normal subgroup is uniscalar. This class includes all totally disconnected groups which are abelian, compact, discrete or compactly generated and nilpotent, see [27] for the latter.

Property **S1** says that, if G is uniscalar, then every element of G normalises some compact open subgroup. It does not follow however that every uniscalar group has a compact open normal subgroup. Examples which are not compactly generated are easily found, see [27, 55], and compactly generated examples can also be found with more effort – it was shown in [37] that such groups exist if there exist permutation groups having certain properties, and such a permutation group was constructed in [3, Theorem 1.2]. However p-adic Lie groups are uniscalar only when they have compact open normal subgroups, see [18, 43].

The scale function encodes a significant amount of information about the group it is defined on, and the existence of a function having the properties described in Proposition 4.1 has structural consequences for totally disconnected groups. The scale function was used in [54] to answer a question of K. H. Hofmann, [26], as follows.

An element x of a locally compact group is *periodic* if $\langle x\rangle^-$ is compact. The set of periodic elements in G is denoted by $P(G)$. This set need not be closed when G is connected – for example, rotations are periodic elements in the group of affine motions of the plane and translations are not, but translations may be approximated uniformly on compact sets by rotations. On the other hand, $P(G)$ is obviously closed when G is discrete. K. H. Hofmann asked whether $P(G)$ must always be closed when G is totally disconnected.

Corollary 4.2. *Let G be a totally disconnected locally compact group. Then $P(G)$ is closed.*

Proof. Let $y \in P(G)$. Then $s(\langle y\rangle^-)$ is a bounded subset of $\mathbb{Z}^+$ because s is continuous and $\langle y\rangle^-$ is compact. On the other hand, $s(y^n) = s(y)^n$ by property **S2**. Hence $s(y) = 1$ and $s(y^{-1}) = 1$ also.

Next, let $x \in P(G)^-$. Then continuity of s implies that $s(x) = 1 = s(x^{-1})$ by the first paragraph. Condition **S1** now shows that there is a compact open subgroup U such that $xUx^{-1} = U$. Since $x \in P(G)^-$, there is $y \in P(G) \cap xU$. Hence $x \in yU$, and so $\langle x\rangle^- \subset \langle yU\rangle^-$. Since $yUy^{-1} = U$, the last set equals $\langle y\rangle^- U$, which is compact. Therefore x belongs to $P(G)$. □

5 Tidy subgroups for groups of automorphisms

Theorem 3.4, Examples 3.5 and 3.6, and the work of H. Glöckner support the idea that tidy subgroups for automorphisms of totally disconnected locally compact groups are analogues of the Jordan canonical form of linear transformations. They do so to the extent that they generalise the decomposition, in (2.1), of the Lie algebra of G into expansive and contractive subspaces for a single automorphism. In further development of this analogy, it will now be seen that the notions of eigenspace and eigenvalue can be generalised by considering the intersections of expansive and contractive subspaces for several automorphisms.

Since commuting linear transformations can be simultaneously triangularised, the analogy suggests that commuting automorphisms of G should have a common tidy subgroup. That is indeed the case and leads to a refinement of the factoring of tidy subgroups given in Theorem 3.4.

Theorem 5.1. *Let $\mathfrak{H}$ be a finitely generated abelian subgroup of* $\mathrm{Aut}(G)$. *Then there is a compact open U tidy for every $\alpha \in \mathfrak{H}$. Furthermore:*

(i) *there are closed subgroups $U_0, U_1, \ldots, U_q$ of U such that*

$$U = U_0 U_1 \ldots U_q,$$

and for each $\alpha \in \mathfrak{H}$ we have $\alpha(U_0) = U_0$ and either $\alpha(U_j) \geq U_j$ or $\alpha(U_j) \leq U_j$ for $j \in \{1, \ldots, q\}$;

(ii) *the subgroups $\widehat{U}_j := \bigcup_{\alpha \in \mathfrak{H}} \alpha(U_j)$ are closed;*

(iii) *for each $j \in \{1, \ldots q\}$ there are homomorphisms $\rho_j : \mathfrak{H} \to \mathbb{Z}$ and positive integers $t_j > 1$ such that*

$$t_j^{\rho_j(\alpha)} = \begin{cases} [\alpha(U_j) : U_j], & \text{if } \alpha(U_j) \geq U_j, \\ [U_j : \alpha(U_j)]^{-1}, & \text{if } \alpha(U_j) \leq U_j; \end{cases}$$

(iv) *for all $\alpha \in \mathfrak{H}$,*

$$s(\alpha) = \prod_{\rho_j(\alpha) > 0} t_j^{\rho_j(\alpha)}.$$

This theorem combines statements from Theorems 3.4, 5.5, 6.8 and 6.14 of [58]. The subgroups $\widehat{U}_j$, $j \in \{1, \ldots, q\}$, are analogues of common eigenspaces for commuting linear transformations, and the homomorphisms

$$\alpha \mapsto t_j^{\rho_j(\alpha)} : \mathfrak{H} \to (\mathbb{R}^+, \times)$$

are analogues of the (absolute values of) the corresponding eigenvalues. These homomorphisms are independent of the choice of tidy subgroup U, but the subgroups $\widehat{U}_j$ are only independent of this choice modulo U_0.

The number q of factors U_j is at least equal to the free rank of $\mathfrak{H}$, but there is no upper bound for q in terms of this rank, see Example 6.10 in [58]. If $\mathfrak{H}$ is the group of inner automorphisms corresponding to a Cartan subgroup of $\mathrm{SL}(n, \mathbb{Q}_p)$, then the factors U_j are the root subgroups of $\mathrm{SL}(n, \mathbb{Q}_p)$, see Example 6.11 in [58].

In the other direction, invertible linear transformations which can be simultaneously triangularised commute modulo unipotent transformations. Here is the corresponding statement for automorphisms of a totally disconnected locally compact group having a common tidy subgroup.

Theorem 5.2. *Let $\mathfrak{H} \leq \mathrm{Aut}(G)$ and suppose that there is a compact open subgroup of G that is tidy for every $\alpha \in \mathfrak{H}$. Let*

$$\mathfrak{H}(1) = \left\{\alpha \in \mathfrak{H} \mid s(\alpha) = 1 = s(\alpha^{-1})\right\} \tag{5.1}$$

$$= \{\alpha \in \mathfrak{H} \mid \alpha(U) = U\}, \tag{5.2}$$

be the subgroup of uniscalar elements *of $\mathfrak{H}$. Then $\mathfrak{H}(1) \lhd \mathfrak{H}$, and $\mathfrak{H}/\mathfrak{H}(1)$ is a free abelian group.*

Furthermore, if $\mathfrak{H}/\mathfrak{H}(1)$ is finitely generated, then (i)–(iv) *from Theorem* 5.1 *hold.*

See Corollary 6.15 of [58]. The hypothesis that $\mathfrak{H}$ is a *group* of automorphisms is essential for the theorem. It is not true in general that, if automorphisms α and β have a common tidy subgroup, then there is a subgroup tidy for $\langle\alpha, \beta\rangle$. For instance (cf. Example 3.6), let ℓ and m be two paths through $\mathcal{T}_q$ which have more than two, but only finitely many, vertices in common, and let x and y be translations along ℓ and m, respectively. Let v_1 and v_2 be adjacent vertices common to ℓ and m, and let U be the compact open subgroup which fixes v_1 and v_2. Then U is tidy for the inner automorphisms, α and β say, of $\mathrm{Aut}(\mathcal{T}_q)$ induced by x and y. However the element $z = x^r y x^{-r}$ is a translation along the path $x^r(m)$, and that path has no vertices in common with both ℓ and m when r is sufficiently large. It follows that there is no subgroup tidy for $\langle\alpha, \beta\rangle$.

Versions of Theorems 5.1 and 5.2 for infinitely generated groups are proved in [58]. They are not included here because they are more complicated to state.

One contrast between the notion of tidy subgroup and the Jordan canonical form is that, while the full set of eigenvalues and corresponding eigenspaces may be computed for a single linear transformation, the factoring of a tidy subgroup into more than two factors requires at least two automorphisms. The reason for this is that eigenvalues of a linear transformation are computed inside the algebra it generates and the linear combinations of powers provide extra degrees of freedom. On the other hand, when identifying tidy subgroups we have only powers of the automorphism to work with. To refine the factoring of W given in Theorem 3.4 further automorphisms are needed but we cannot produce them as linear combinations of powers of α.

When G is abelian more can be said about an individual automorphism α. In that case, there is an open α-invariant subgroup H and a compact α-invariant subgroup K such that H/K is the sum of finitely many Sylow p-subgroups. Each of these

subgroups is invariant under α and the restriction of α to each p-subgroup is close to being a sum of linear transformations. The tidy subgroups for α then correspond to a canonical form, and the scale of α can be computed in terms of "eigenvalues", see [12].

6 Applications and further developments

This final section describes some problems which tidy subgroups have been used to solve, some open problems where they might be used, and outlines possible further developments of the structure theory of totally disconnected locally compact groups. Some of these problems relate to random walks and geometry.

Random walks on groups. The stimulus for the introduction of the scale function and the notion of tidy subgroup was a problem concerning the so-called concentration functions of a random walk.

Let μ be a probability measure on the group G and define, for each $n > 0$ and compact subset K of G,

$$f_n(K) = \sup\left\{\mu^{*n}(Kx) : x \in G\right\}.$$

Then the functions f_n defined on the set $\mathbb{K}$ of all compact subsets of G are called the *concentration functions*. The name derives from the fact that, if there is a K such that $f_n(K) \not\to 0$, then the random walk with law μ remains concentrated on translates of K. This can occur if μ is supported on a coset zH where H is a compact subgroup of G and $zH = Hz$. K. H. Hofmann and A. Mukherjea addressed in [28] the circumstances under which random walks remain concentrated.

The probability measure is said to be *irreducible* if the closed semigroup generated by the support of μ is equal to G. Hofmann and Mukherjea set out to show that, if G supports an irreducible measure μ such that $f_n(K) \not\to 0$ for some K, then G is compact. They reduced the problem to a structural question by showing that, if G supports such a probability measure, then G contains:

1. a cocompact, closed normal subgroup N;
2. a compact subgroup H of N; and
3. an element z such that for every neighbourhood V of H we have

$$N \subset \overline{\Big(\bigcup_{n=1}^{\infty} z^n V z^{-n}\Big)}.$$

They called a locally compact group having these properties *strange* and conjectured that there were no strange groups.

Hofmann and Mukherjea proved their conjecture in the case of connected groups by means of an approximation by Lie groups argument. A key step in their argument was to show that, when G is a Lie group, the eigenvalues of $\operatorname{ad} z$ all have modulus one.

Their argument reduced the conjecture to the case of totally disconnected groups and that case was proved in [31]. A key step in the argument was to show that $s(z) = 1$. That eigenvalues of $\operatorname{ad} z$ having modulus one in the connected case corresponds in the totally disconnected case to the scale of z being equal to one was the first indication that the scale provides an analogue of eigenvalues.

Ergodic automorphisms. The notion of tidy subgroup can also be used to prove the totally disconnected case of a conjecture of P. R. Halmos. This conjecture is obviously related to the question on concentration functions, and it is surprising that the two questions do not seem to have been considered together before.

Halmos conjectured in [23] that any locally compact group which has an ergodic automorphism must be compact. This conjecture was proved in the case when G is connected by R. Kaufman and M. Rajagopalan in [36] and also announced by T. S. Wu in [60]. It was shown moreover by S. G. Dani, in [8], that the same holds for affine automorphisms (i.e., maps ϕ of the form $\phi(y) = z\alpha(y)$, where z is an element of G and α is an automorphism).

For totally disconnected G, N. Aoki showed by means of a topological dynamics argument that, if G has an ergodic automorphism, then G is compact, see [1]. M. Dateyama and T. Kasuga used the same techniques to prove a version of this result for affine maps, see [11]. However a much shorter proof that only compact groups can have ergodic automorphisms is given in [44], and that proof uses properties of tidy subgroups.

Derivations on $L^1(G)$. For each bounded measure μ on G the map D_μ on $L^1(G)$ defined by

$$D_\mu(f) := f * \mu - \mu * f \quad (f \in L^1(G)),$$

is a derivation. (Note that the range of D_μ is contained in $L^1(G)$ because $L^1(G)$ is an ideal in $M(G)$.) The question of whether every derivation on $L^1(G)$ arises in this way (i.e., of whether $H^1(L^1(G), M(G)) = 0$) has been a very fruitful one.

B. E. Johnson showed in [32] that $H^1(L^1(G), M(G)) = 0$ if G is amenable or an [SIN]-group. It was in the course of the proof that he introduced the concept of an amenable Banach algebra. More recently, he has shown that $H^1(L^1(G), M(G)) = 0$ if G is connected [33], and this proof uses an approximation by Lie groups argument.

Although Johnson's result for [SIN]-groups covers the discrete group case, the question remains open for non-discrete totally disconnected groups. A totally disconnected group is a [SIN]-group if and only if it has a base of neighbourhoods of the identity consisting of compact open normal subgroups. General tools seem to be needed, playing the role filled by approximation by Lie groups in the connected

case, for dealing with groups which are not [SIN]. The techniques described above are likely to fill this role but further development may be required. This question will not be completely answered even if the connected and totally disconnected cases are solved: there are extensions of connected groups by discrete for which it is not known whether $H^1(L^1(G), M(G)) = 0$. Some of the difficulties with the extension problem also arise in the case of totally disconnected groups because [IN]-groups are not covered by Johnson's results but they are uniscalar.

The centre of $M(G)$ and $VN(G)$. The centre of the group convolution algebra $L^1(G)$ is non-zero if and only if G is an [IN]-group. The central functions are then supported on the union of all the compact invariant neighbourhoods of the identity, see [41]. However, the centre of the algebra of bounded measures $M(G)$ of the locally compact group G has not been described except in the case when G is connected, see [19, 20, 45]. The method of approximation by Lie groups is used in the description of the centre in this case. Tidy subgroups and associated ideas might be useful for the description of the centre of $M(G)$ when G is totally disconnected.

One of the basic examples of a von Neumann algebra which is a factor (i.e., has trivial centre) is the algebra $VN(G)$ when G is a discrete group having no finite conjugacy classes other than that of the identity, see for example [34, §6.7]. There seems to be little known about the centre of $VN(G)$ for non-discrete G however. This appears to be a more difficult problem than describing the centre of $M(G)$.

Simple totally disconnected groups. A simple connected group is in fact a Lie group, and the simple Lie groups have been classified, [49]. The simple totally disconnected locally compact groups have not even begun to be classified but Theorems 5.1 and 5.2 provide a starting point. The rank of the free abelian group $\mathfrak{H}/\mathfrak{H}(1)$ identified in Theorem 5.2 leads to a notion of rank of a totally disconnected locally compact group, [58], and this number will be an important invariant in any classification.

A possible new feature of the potential classification is that there could be rank-0 simple groups. These would be uniscalar simple groups in which every element normalises some compact open subgroup. Simplicity of the group would imply that there is no compact open normal subgroup. There are compactly generated uniscalar groups, see [3, 37], but the construction does not yield simple groups.

Another important invariant in a classification will be the set of values of the scale function on the simple groups. This set extends to general totally disconnected groups the number p which is characteristic of p-adic Lie groups. If G is compactly generated, the range of the scale function has only finitely many prime divisors, see [57], and so this set can be expected to distinguish between such simple groups quite well.

In fact, a classification of *compactly generated* simple totally disconnected groups is probably the best which can be hoped for. It may be shown that, if G is compactly generated and simple, then no compact open subgroup of G is solvable, and relating the simplicity of G to its local structure in this way will obviously play a role in any

classification. However, there are (non-compactly generated and uniscalar) simple groups which have abelian compact open subgroups.

Combinatorial geometry. As already seen in Example 2.1(c), automorphism groups of locally finite graphs are totally disconnected locally compact groups. These automorphism groups have been studied in [50] and [59] for example. The relationship between the work of V. I. Trofimov and the ideas described here has yet to be thoroughly explored. In [39], R. Möller has interpreted and given new proofs for many of the results in [53] in the language of groups operating on graphs. He has also found new characterisations of the scale function in this context.

In the other direction, an important technique in the study of semisimple p-adic Lie groups is to represent them as acting on trees and buildings, [5, 48]. Theorems 5.1 and 5.2 provide the starting point for an extension of this representation to a more general class of totally disconnected locally compact groups. The author has begun developing such a representation in collaboration with Udo Baumgartner. Another important ingredient is the notion of a *contraction group* for commuting automorphisms,[2].

Finitely generated groups. Tidy subgroups and the scale function yield no direct information about discrete groups but they might be used as tools for the investigation of finitely generated groups nevertheless.

The Cayley graph of a group G with a given finite generating set $\mathcal{F}$ is locally finite, and so the automorphism group, $\widetilde{G}_{\mathcal{F}}$ say, of the Cayley graph is totally disconnected, compactly generated and locally compact. The rank of $\widetilde{G}_{\mathcal{F}}$, the set of values of the scale function on $\widetilde{G}_{\mathcal{F}}$ and other invariants for totally disconnected groups calculated on $\widetilde{G}_{\mathcal{F}}$ will then be distinguishing features of G and its generating set $\mathcal{F}$. Since G itself acts on the Cayley graph by translation and is thus identified with a cocompact subgroup of $\widetilde{G}_{\mathcal{F}}$, there is likely to be a close relationship between G and $\widetilde{G}_{\mathcal{F}}$.

There are many questions to be answered.

It is not always the case that $\widetilde{G}_{\mathcal{F}}$ is non-discrete: if G is abelian, for example, then $\widetilde{G}_{\mathcal{F}}$ is discrete for every generating set $\mathcal{F}$. How large is the class of groups G such that $\widetilde{G}_{\mathcal{F}}$ is discrete for every generating set $\mathcal{F}$? Does it include all groups with less than exponential growth? How large is the class of groups such that $\widetilde{G}_{\mathcal{F}}$ always has a compact open normal subgroup? is uniscalar? How do invariants such as the range of the scale function on $\widetilde{G}_{\mathcal{F}}$ depend on the generating set $\mathcal{F}$? Is there, for each finitely generated group G, a finite set of prime numbers such that scale function values on $\widetilde{G}_{\mathcal{F}}$ are a product of these primes for all choices of $\mathcal{F}$? (Recall that the number of prime factors of the scale function on a compactly generated group is always finite, [57].)

The scale function is potentially applicable only to those finitely generated groups G which possess a generating set $\mathcal{F}$ such that $\widetilde{G}_{\mathcal{F}}$ is not uniscalar. There are many such groups however because many vertex transitive, locally finite graphs are the Cayley graphs of some finitely generated group. For example, trees and some higher

rank buildings can be realised in this way, [6]. Which vertex transitive, locally finite graphs are Cayley graphs?

What are the computability properties of the scale function on $\widetilde{G}_{\mathcal{F}}$? Under what circumstances can the scale on the elements of $\mathcal{F}$ be computed? Is it possible when G with generating set $\mathcal{F}$ is finitely presented, for example? Recall that tidy subgroups and the scale function are computed for automorphism groups of trees, which are Cayley graphs of free groups, in the final Section of [53].

Lie theory. A structure theory for totally disconnected groups completely parallel to that for connected groups would require functors to Lie algebras. That is not possible in general, but the extent to which it is possible has been thoroughly explored by Helge Glöckner in [16]. Work on classifying simple totally disconnected groups and on representing totally disconnected groups on discrete geometries may lead to an analogue of Lie algebras arising from Theorems 5.1 and 5.2.

Acknowledgement. I have had the benefit of discussions with many people about this work, particularly Udo Baumgartner, Jacqui Ramagge and Helge Glöckner. The research has been supported by the Australian Research Council Grant A69700321 and I am also grateful to the Erwin Schrödinger Institute for its support during the writing of this paper.

References

[1] I. N. Aoki, Dense orbits of automorphisms and compactness of groups, Topology Appl. 20 (1985), 1–15.

[2] U. Baumgartner and G. A. Willis, Contraction groups and scales of automorphisms of totally disconnected locally compact groups, Israel J. Math., to appear.

[3] M. Bhattacharjee and D. MacPherson, Strange permutation representations of free groups, J. Austral. Math. Soc. 74 (2003), 267–285.

[4] J.-B. Bost and A. Connes, Hecke algebras, type III factors and phase transitions with spontaneous symmetry breaking in number theory, Selecta Math. (N.S.) 1 (1995), 411–457.

[5] K. S. Brown, Buildings, Springer-Verlag, New York 1989.

[6] D. I. Cartwright, A. M. Mantero, T. Steger, A. Zappa, Groups acting simply transitively on the vertices of a building of type $\tilde{A}_2$. I, Geom. Dedicata 47 (1993), 143–166.

[7] A. Connes, Noncommutative Geometry, Academic Press, San Diego, CA, 1994.

[8] S. G. Dani, Dense orbits of affine automorphisms and compactness of groups, J. London Math. Soc. (2) 25 (1982), 241–245.

[9] D. van Dantzig, Studien über topologische Algebra, Dissertation, Amsterdam 1931.

[10] D. van Dantzig, Zur topologisches Algebra. III. Brouwersche und Cantorsche Gruppen, Compositio Math. 3 (1936), 408–426.

[11] M. Dateyama and T. Kasuga, Ergodic affine maps of locally compact groups, J. Math. Soc. Japan 37 (1985), 363–372.

[12] S. Evans-Riley and G. A. Willis, Automorphisms of abelian totally disconnected groups, in preparation.

[13] A. M. Gleason, Groups without small subgroups, Ann. of Math. (2) 56 (1952), 193–212.

[14] H. Glöckner, Scale functions on linear groups over local skew fields, J. Algebra 205 (1998), 525–541.

[15] H. Glöckner, Scale functions on p-adic Lie groups, Manuscripta Math. 97 (1998), 205–215.

[16] H. Glöckner, Real and p-adic Lie algebra functors on the category of topological groups, Pacific J. Math. 203 (2002), 321–368.

[17] H. Glöckner, Details of Hall's category equivalence, preprint (2000).

[18] H. Glöckner and G. A. Willis, Uniscalar p-adic Lie groups, Forum Math. 13 (2001), 413–421.

[19] F. P. Greenleaf, M. Moskowitz and L. P. Rothschild, Unbounded conjugacy classes in Lie groups and location of central measures, Acta Math. 132 (1974), 225–243.

[20] F. P. Greenleaf, M. Moskowitz and L. P. Rothschild, Central idempotent measures on connected locally compact groups, J. Funct. Anal. 15 (1974), 22–32.

[21] S. Grosser and M. A. Moskowitz, Compactness conditions in topological groups, J. Reine Angew. Math. 246 (1971), 1–40.

[22] R. Hall, Hecke C^*-algebras, Dissertation, The Pennsylvania State University, 1999.

[23] P. R. Halmos, Lectures on Ergodic Theory, Publ. Math. Soc. Japan, Tokyo 1956.

[24] E. P. Herman, Totally disconnected locally compact topological groups, Dissertation, University of Oregon, 1997.

[25] E. Hewitt and K. A. Ross, Abstract Harmonic Analysis I, Grundlehren Math. Wiss. 115, Springer-Verlag, Berlin 1963.

[26] K. H. Hofmann, Characteristic subgroups in locally compact totally disconnected groups and their applications to a problem on random walks on locally compact groups, preprint 606, Technische Hochschule Darmstadt (1981).

[27] K. H. Hofmann, J. R. Liukkonen and M. W. Mislove, Compact extensions of compactly generated nilpotent groups are pro-Lie, Proc. Amer. Math. Soc. 84 (1982), 443–448.

[28] K. H. Hofmann and A. Mukherjea, Concentration functions and a class of non-compact groups, Math. Ann. 256 (1981), 535–548.

[29] N. Jacobson, Totally disconnected locally compact rings, Amer. J. Math. 58 (1936), 433–449.

[30] N. Jacobson, Basic Algebra II, Freeman, San Francisco 1980.

[31] W. Jaworski, J. M. Rosenblatt and G. A. Willis, Concentration functions in locally compact groups, Math. Ann. 305 (1996), 673–691.

[32] B. E. Johnson, Cohomology in Banach Algebras, Mem. Amer. Math. Soc. 127 (1972).

[33] B. E. Johnson, The derivation problem for group algebras of connected locally compact groups, J. London Math. Soc. (2) 63 (2001), 441–452.

[34] R. V. Kadison and J. R. Ringrose, Fundamentals of the Theory of Operator Algebras. II, Academic Press, New York 1986.

[35] I. Kaplansky, Lie Algebras and Locally Compact Groups, The University of Chicago Press, Chicago 1971.

[36] R. Kaufman and M. Rajagopolan, On automorphisms of a locally compact group, Michigan Math. J. 13 (1966), 373–374.

[37] A. Kepert and G. Willis, Scale functions and tree ends, J. Austral. Math. Soc. Ser. 70 (2001), 273–292.

[38] M. Laca and I. Raeburn, The ideal structure of the Hecke C^*-algebra of Bost and Connes, Math. Ann. 318 (2000), 433–451.

[39] R. G. Möller, Structure theory of totally disconnected locally compact groups via graphs and permutations, Canad. J. Math. 54 (2002), 795–827.

[40] D. Montgomery and L. Zippin, Topological Transformation Groups, Interscience, New York–London 1955.

[41] R. D.Mosak, Central functions in group algebras, Proc. Amer. Math. Soc. 29 (1971) 613–616.

[42] T. W. Palmer, Banach Algebras and the General Theory of $*$-algebras. Volume II. $*$-algebras, Encyclopedia of Mathematics 79, Cambridge University Press, Cambridge 2001.

[43] A. Parreau, Sous-groupes elliptiques de groupes linéares sur un corps valué, J. Lie Theory 13 (2003), 271–278.

[44] W. H. Previts and T. S. Wu, Dense orbits and compactness of groups, Bull. Austral. Math. Soc. 68 (2003), 155–159.

[45] D. L. Ragozin and L. P. Rothschild, Central measures on semisimple Lie groups have essentially compact support, Proc. Amer. Math. Soc. 32 (1972), 585–589.

[46] G. Schlichting, Polynomidentitäten und Permutationsdarstellungen lokalkompakter Gruppen, Invent. Math. 55 (1979), 97–106.

[47] G. Schlichting, On the periodicity of group operations, in: Group Theory (Singapore 1987), Walter de Gruyter, Berlin 1989, 507–517.

[48] J.-P. Serre, Arbres, Amalgames, SL_2, Astérisque 46, Société Mathématique de France, Paris 1977.

[49] J. Tits, Tabellen zu den einfachen Lie Gruppen und ihren Darstellungen, Springer-Verlag, Berlin 1967.

[50] V. I. Trofimov, On the action of a group on a graph, Acta Appl. Math. 29 (1992), 161–170.

[51] K. Tzanev, C^*-algèbres de Hecke et K-théorie, Doctoral Thesis, Université Paris 7, 2000.

[52] A. Weil, Basic Number Theory, 2nd ed., Grundlehren Math. Wiss. 144, Springer-Verlag, Berlin 1973.

[53] G. A. Willis, The structure of totally disconnected, locally compact groups, Math. Ann. 300 (1994), 341–363.

[54] G. A. Willis, Totally disconnected groups and proofs of conjectures of Hofmann and Mukherjea, Bull. Austral. Math. Soc. 51 (1995), 489–494.

[55] G. A. Willis, Totally disconnected, nilpotent, locally compact groups, Bull. Austral. Math. Soc. 55 (1997), 143–146.

[56] G. A. Willis, Further properties of the scale function on a totally disconnected group, J. Algebra 237 (2001), 142–164.

[57] G. A. Willis, The number of prime factors of the scale function on a compactly generated group is finite, Bull. London Math. Soc. 33 (2001), 168–174.

[58] G. A. Willis, Tidy subgroups for commuting automorphisms of totally disconnected groups: an analogue of simultaneous triangularisation of matrices, New York J. Math. 10 (2004), 1–35.

[59] W. Woess, Amenable group actions on infinite graphs, Math. Ann. 284 (1989), 251–265.

[60] T. S. Wu, Continuous automorphisms on locally compact groups, Abstract 66T-104, Notices Amer. Math. Soc. 13 (1966), 238.

George A. Willis, Department of Mathematics, University of Newcastle, Newcastle (Callaghan), NSW 2308, Australia

E-mail: george@frey.newcastle.edu.au

Research communications

On the classification of invariant measures for horosphere foliations on nilpotent covers of negatively curved manifolds

Martine Babillot

Abstract. We classify those invariant measures of the horocycle flow on nilpotent covers of compact hyperbolic surfaces which have the additional property of being quasi-invariant for the geodesic flow. For this, we use the potential theory at infinity, thus giving a purely geometrical proof of a result of Aaronson et al. [1] which was obtained using symbolic dynamics. We also extend this classification to the horosphere foliation on covers of negatively curved manifolds.

1 Introduction

Let $M = \Gamma\backslash\widetilde{M}$ be a negatively curved Riemannian manifold, with universal cover $\widetilde{M}$ and fundamental group Γ. For a compact manifold M, a result of Bowen and Marcus asserts that the horosphere foliation of the unit tangent bundle T^1M is uniquely ergodic, or, equivalently, that there exists a unique measure on the space of horospheres of $\widetilde{M}$ which is invariant under the action of the group Γ [6]. This extends Furstenberg's celebrated result [10] on the unique ergodicity of the horocycle flow on a compact hyperbolic surface $S = \Gamma\backslash\boldsymbol{D}$, which was proved by showing that the Lebesgue measure on $\mathbb{R}^2 \setminus \{0\}$ (which is essentially the space of horospheres of the hyperbolic disc $\boldsymbol{D}$) is the unique invariant measure under the linear action of Γ. Since then similar results were obtained in the algebraic context for unipotent flows acting on finite volume quotients of Lie groups [20], or in the more geometrical setup for the horocycle flow and the horosphere foliation of negatively curved geometrically finite manifolds [7, 21]. In all these cases, either the dynamical system is uniquely ergodic, or there exists a complete classification of the invariant measures.

However, for manifolds which are not geometrically finite the situation is far from being understood, and, for instance, it may exist an infinite family of ergodic invariant Radon measures. This was observed in [3] for the horocycle flow on an infinite abelian cover $S = \Gamma\backslash\boldsymbol{D}$ of a hyperbolic compact surface $S_0 = \Gamma_0\backslash\boldsymbol{D}$. It is still an open question whether the invariant measures constructed in [3] are the only ones which are ergodic and conservative.

Partial answers to this question are known under additional assumptions. For instance, these measures are the only ergodic ones which are also quasi-invariant under the action of the deck group $A = \Gamma_0/\Gamma$ [3]. A recent result of Aaronson, Sarig and Solomyak shows that these are also the only ones which are quasi-invariant under the geodesic flow [1]. Their proof uses symbolic dynamics.

The purpose of this note is to suggest a purely geometrical approach to the latter result. Following Furstenberg and Sullivan, we use the potential theory on the hyperbolic disc in order to explain the connection between invariant measures of the horocycle flow, conformal measures on the boundary S^1 and harmonic functions. By applying the notion of harmonic measures on foliations as developed in [11, 12], we then show how this approach extends to general negatively curved manifolds. More precisely, we prove

Main Theorem. *Let $M_0 = \Gamma_0\backslash\widetilde{M}$ be a compact negatively curved Riemannian manifold, with universal cover $\widetilde{M}$ and fundamental group Γ_0. Let $M = \Gamma\backslash\widetilde{M}$ be a regular nilpotent cover of M_0. Then the ergodic invariant measures of the horosphere foliation of M which are quasi-invariant under the geodesic flow are in a one-to-one correspondence with the cohomology classes of M_0 which vanish on Γ.*

We refer to Sections 3 and 4 for more details on this correspondence. We emphasize that the ideas developed here can be seen as an extension to negatively curved manifolds of the fundamental work by Sullivan [24, 25, 26, 27] on hyperbolic manifolds, and that our main purpose here is to present these ideas, together with related recent results, in a unified way.

The text is organized in the following way. In Section 2, we set up the notations and explain the correspondence between conformal measures and invariant measures of the horosphere foliation which are quasi-invariant for the geodesic flow (Proposition 2.1). In Section 3 we study conformal measures in the constant curvature case. Section 4 is devoted to the variable negative curvature case.

2 Preliminaries

We shall consider a complete simply connected Riemannian manifold $\widetilde{M}$ with pinched negative sectional curvatures $-b^2 \leq K \leq -a^2$. For background information on non-positively curved manifolds, we refer to [4]. The main example to have in mind is the *hyperbolic disc* $\boldsymbol{D}$, i.e., the unit disc of the complex plane endowed with the Riemannian metric $ds^2 = |dz|^2/\left[4(1-|z|^2)^2\right]$. This is a Riemannian manifold of constant negative curvature -1. We recall that the *boundary at infinity* $\partial\widetilde{M}$ is the set of equivalence classes of geodesic rays of $\widetilde{M}$, where two rays $r_1, r_2 : [0, +\infty) \to \widetilde{M}$ are equivalent if and only if the distance between $r_1(t)$ and $r_2(t)$ remains bounded. This boundary can be identified with the unit circle S^1 when $\widetilde{M} = \boldsymbol{D}$, and in general comes with a topology such that $\overline{M} = M \cup \partial\widetilde{M}$ is a compactification of M.

With each point ξ at infinity is associated the *Busemann cocycle* $b_\xi : \widetilde{M}^2 \to \mathbb{R}$ defined as

$$b_\xi(p,q) := \lim_{t\to+\infty}\big[d(p,r(t)) - d(q,r(t))\big] = \text{“}d(p,\xi) - d(q,\xi)\text{”},$$

where $r(t)$ is any ray pointing towards ξ. In the case of the disc, with origin o at the center,

$$b_\xi(z,o) = -\log\frac{1-|z|^2}{|z-\xi|^2} \tag{2.1}$$

coincides with minus the logarithm of the classical Poisson kernel $P(z,\xi)$.

The *horosphere* centered at $\xi \in \partial\widetilde{M}$ and passing through a point p of $\widetilde{M}$ is the level set

$$H(\xi,p) = \{q \in \widetilde{M};\ b_\xi(p,q) = 0\}.$$

Each horosphere lifts to the unit tangent bundle $T^1\widetilde{M}$ by considering the set $H^+(\xi,p)$ of all unit vectors normal to $H(\xi,p)$ and pointing inwards. This leads to the (strong stable) *horosphere foliation* $\widetilde{\mathcal{W}}^{ss}$ of $T^1\widetilde{M}$, where, for a vector u pointing to the point at infinity ξ and based at p, one defines the leaf $\widetilde{W}^{ss}(u)$ of u to be $H^+(\xi,p)$. The space of leaves of this foliation is therefore the space of horospheres $\mathcal{H}$ of $\widetilde{M}$, which can be parameterized by $\partial\widetilde{M}\times\mathbb{R}$ using the map

$$H(\xi,p) \to \big(\xi, r = b_\xi(p,o)\big) \in \partial\widetilde{M}\times\mathbb{R},$$

where o is any fixed point in $\widetilde{M}$. The *geodesic flow* (g_t) sends a horosphere centered at ξ to the horosphere centered at ξ at distance $|t|$ from the first one (it is inside the original horosphere if $t>0$ and outside if $t<0$) and acts in the above coordinates as

$$g_t(\xi,r) = (\xi, r-t).$$

The action of the isometry group of $\widetilde{M}$ extends continuously to the boundary at infinity $\partial\widetilde{M}$, and thus to the space of horospheres. Since an isometry g sends the horosphere $H(\xi,p)$ to $H(g.\xi, g.p)$, the relations

$$b_{g.\xi}(g.p,o) = b_{g.\xi}(g.p,g.o) + b_{g.\xi}(g.o,o) = b_\xi(p,o) + b_\xi(o,g^{-1}.o)$$

show that the action of $\mathrm{Iso}(\widetilde{M})$ on $\mathcal{H}$ parameterized by $\partial\widetilde{M}\times\mathbb{R}$ is given by the formula

$$g.(\xi,r) = \big(g.\xi, r + b_\xi(o,g^{-1}.o)\big).$$

We note that the quantity

$$b(g,\xi) := b_\xi(o,g^{-1}.o),$$

which appears in the right-hand side, satisfies the identity

$$b(g_1g_2,\xi) = b(g_1, g_2.\xi) + b(g_2,\xi),$$

and, therefore, it is a cocycle for the action of $\mathrm{Iso}(\widetilde{M})$ on the boundary $\partial\widetilde{M}$. Its cohomology class does not depend on the choice of the origin.

For the hyperbolic disc with the isometry group $G = \mathrm{PSL}(2,\mathbb{R})$, the space $\mathcal{H}$ can be identified with $\mathbb{R}^2 \setminus \{0\}/\pm 1 \simeq P^1(\mathbb{R}) \times \mathbb{R}$, and the boundary $\partial \boldsymbol{D}$ with the projective line $P^1(\mathbb{R})$. The action of G on $\mathcal{H}$ amounts to the standard linear action of $\mathrm{PSL}(2,\mathbb{R})$ on $\mathbb{R}^2 \setminus \{0\}/\pm 1$, whereas the geodesic flow is conjugate with the action of $\mathbb{R}$ by dilatations. In this model, one has

$$b(g,\xi) = 2\log\|g\dot{\xi}\|,$$

where $\|\cdot\|$ is the usual Euclidean norm on $\mathbb{R}^2$, and $\dot{\xi}$ is any unit vector representing $\xi \in P^1(\mathbb{R})$.

Yet another description of this cocycle can be given in terms of the pseudo-metric on the boundary $|\cdot|$ defined by the formula

$$|\xi - \xi'| = e^{-(\xi|\xi')_o},$$

where

$$(\xi|\xi')_o = \lim_{\substack{p\to\xi \\ p'\to\xi'}} \frac{1}{2}\big[d(o,p) + d(o,p') - d(p,p')\big].$$

is the Gromov product of ξ and ξ' with respect to the origin o.

For the n-dimensional hyperbolic space $\boldsymbol{H}^n$ modelled by the unit ball B^n of $\mathbb{R}^n$, this metric coincides with the chordal Euclidean metric on S^{n-1}, whereas Poincaré's theorem identifies $\mathrm{Iso}(\boldsymbol{H}^n)$ with the group of conformal transformations of S^{n-1}. In the general case, this formula defines a metric only if the sectional curvature is less than -1, but a power of it is indeed a metric [5]. An isometry g of $\widetilde{M}$ acts quasi-conformally on $(\partial\widetilde{M}, |\cdot|)$ with infinitesimal dilatation coefficient $|g'(\xi)|$ being precisely

$$|g'(\xi)| = e^{b(g,\xi)}.$$

In particular, the obvious relation

$$2(g.\xi, g.\xi')_o = 2(\xi,\xi')_o - b(g,\xi) - b(g,\xi')$$

leads to the identity

$$|g.\xi - g.\xi'|^2 = |g'(\xi)| \cdot |g'(\xi')| \cdot |\xi - \xi'|^2. \tag{2.2}$$

Let Γ be a discrete subgroup of $\mathrm{Iso}(\widetilde{M})$. Since Γ preserves the horosphere foliation $\widetilde{\mathcal{W}}^{\mathrm{ss}}$ of $T^1\widetilde{M}$, one gets a foliation of $\Gamma\backslash T^1\widetilde{M}$ by passing to the quotient: this is the horosphere foliation of the unit tangent bundle of the manifold $M = \Gamma\backslash\widetilde{M}$ when Γ is torsion-free. Looking for (transverse) invariant measures of this foliation thus amounts to looking for invariant measures of the foliation $\widetilde{\mathcal{W}}^{\mathrm{ss}}$ which are Γ-invariant and therefore to looking for Radon invariant measures of the action of Γ on the space of horospheres. In the rest of this paper, we shall always adopt this point of view, and thus leave aside the foliation.

We shall use the following notations: for a measure m on a measured space $(X, \mathcal{B})$, and a transformation g of X, we denote by gm the image of m by g, i.e. the measure defined by $gm(A) = m(g^{-1}A)$ for any $A \in \mathcal{B}$.

Let us now explain the relation between Γ-invariant measures on $\mathcal{H}$ and conformal measures. Recall that a Borel measure ν on $\partial\widetilde{M}$ is said to be Γ-*conformal* if it is quasi-invariant for the action of Γ with Radon–Nikodym cocycle given by

$$\frac{d\gamma^{-1}\nu}{d\nu}(\xi) = e^{\delta\, b(\gamma,\xi)} = |\gamma'(\xi)|^{\delta}$$

for some $\delta \geq 0$, which is called the *dimension* of the measure ν. For instance, in the case of the hyperbolic n-space $\boldsymbol{H}^n$, the Lebesgue measure on the sphere S^{n-1} is a conformal measure of dimension $n-1$ for any Kleinian group $\Gamma \subset \mathrm{Iso}(\boldsymbol{H}^n)$.

By a result of Patterson for hyperbolic manifolds, which easily extends to the case of variable curvature, there always exists a conformal measure with dimension equal to the critical exponent δ_Γ of the group Γ

$$\delta_\Gamma = \limsup_{R\to+\infty} \mathrm{card}\,\{\gamma \in \Gamma;\ d(o, \gamma.o) \leq R\}.$$

Moreover, all conformal measures have dimension $\delta \geq \delta_\Gamma$, and in particular have strictly positive dimension when Γ is non-elementary [24].

Proposition 2.1. *Let Γ be a discrete group of isometries of a simply connected negatively curved manifold $\widetilde{M}$. If ν is a Γ-conformal measure with dimension δ, then the measure m^ν on the space of horospheres $\mathcal{H} \simeq \partial\widetilde{M} \times \mathbb{R}$ defined as*

$$dm^\nu(\xi, r) = d\nu(\xi)e^{-\delta r}dr$$

is Γ-invariant and quasi-invariant under the geodesic flow. Conversely, if m is a Radon measure on $\mathcal{H}$ which is Γ-invariant and ergodic, and quasi-invariant under the geodesic flow, then there exists a Γ-conformal measure ν such that $m = m^\nu$.

Proof. The measure m^ν obviously satisfies the relation $g_t m^\nu = e^{-\delta t}m^\nu$ and is therefore quasi-invariant under the geodesic flow. Moreover, for any isometry $\gamma \in \Gamma$ and any compactly supported continuous function φ on $\mathcal{H}$, one has

$$\begin{aligned}
\gamma^{-1}m^\nu(\varphi) &= \int_{\partial\widetilde{M}\times\mathbb{R}} \varphi(\gamma^{-1}(\xi, r))\ d\nu(\xi)e^{-\delta r}dr \\
&= \int_{\partial\widetilde{M}\times\mathbb{R}} \varphi((\gamma^{-1}.\xi, r + b(\gamma^{-1}, \xi))\ d\nu(\xi)e^{-\delta r}dr \\
&= \int_{\partial\widetilde{M}\times\mathbb{R}} \varphi((\xi, r + b(\gamma^{-1}, \gamma.\xi))\ e^{\delta b(\gamma,\xi)}\ d\nu(\xi)e^{-\delta r}dr \\
&= \int_{\partial\widetilde{M}\times\mathbb{R}} \varphi((\xi, r - b(\gamma, \xi))\ e^{-\delta(r-b(\gamma,\xi))}\ d\nu(\xi)dr = m^\nu(\varphi),
\end{aligned}$$

so that m^ν is Γ-invariant . Conversely, let m be a Γ-invariant and ergodic measure on $\mathcal{H}$ which is quasi-invariant under the geodesic flow. Then, since the actions of (g_t)

and Γ commute, the measures m and $g_t m$ are both ergodic for the action of Γ, and since they are equivalent, they must be proportional. This being true for all $t \in \mathbb{R}$, there exists $\delta \in \mathbb{R}$ such that $g_t m^\nu = e^{-\delta t} m^\nu$, and therefore such that the measure $e^{\delta r} dm(\xi, r)$ is invariant by translations on the second coordinate. Such a measure is necessarily of the type $\nu \otimes dr$, with ν a finite measure on $\partial \widetilde{M}$. The above computation shows that ν has to be conformal, from which it follows that the dimension δ is ≥ 0, since a finite measure cannot be conformal with negative dimension. □

Remark 2.2. If m^ν is ergodic, then obviously the conformal measure ν has to be ergodic. The converse assertion has been shown to be true in the case of a general Fuchsian group when ν is the Lebesgue measure on S^1 and for some other cases [14, 15], but, to the best of our knowledge, it is still unknown in general. However, this is true for the groups that will be considered here.

Using this correspondence, one sees that the classification of Γ-invariant ergodic measures on $\mathcal{H}$ which are quasi-invariant under the geodesic flow reduces to that of conformal measures. In the next two sections, we shall see how these conformal measures can be classified using potential theory.

3 Conformal measures for Fuchsian groups

In this section, we concentrate on the case where $\widetilde{M} = \boldsymbol{D}$ is the hyperbolic disc, and relate conformal measures with eigenfunctions of the hyperbolic Laplacian $L = (1 - z^2)^2 \Delta$. Then, we prove the theorem in the case where $M_0 = S_0$ is a compact hyperbolic surface.

For a given $\lambda \geq -1/4$, denote by $H^+(\Gamma, \lambda)$ the convex cone of positive Γ-invariant λ-eigenfunctions. Note that when Γ is torsion-free, these eigenfunctions can be seen as eigenfunctions for the Laplacian on the hyperbolic surface $S = \Gamma \backslash \boldsymbol{D}$.

The fact that the Busemann cocycle can be expressed in terms of the Poisson kernel (see (2.1)) implies the following

Proposition 3.1. *Let Γ be a Fuchsian group. If ν is a Γ-conformal measure of dimension δ, the function $h^\nu : \boldsymbol{D} \to [0, +\infty)$ defined as*

$$h^\nu(z) = \int_{S^1} P(z, \xi)^\delta \, d\nu(\xi)$$

is a Γ-invariant positive $\delta(\delta - 1)$-eigenfunction of the hyperbolic Laplacian. Conversely, any positive Γ-invariant eigenfunction of the Laplacian is of the form h^ν for some Γ-conformal measure ν with dimension $\delta \geq 1/2$.

Proof. It is a classical fact that the Poisson kernel raised to the power δ produces $\delta(\delta - 1)$-eigenfunctions of the hyperbolic Laplacian, so that h^ν, being a convex combination of eigenfunctions, is still an eigenfunction. The fact that h^ν is Γ-invariant

follows from the conformality of ν. Indeed, we may write

$$h^\nu(z) = \int_{S^1} e^{-\delta\, b_\xi(z,o)} d\nu(\xi),$$

so that for any isometry $\gamma \in \Gamma$,

$$\begin{aligned} h^\nu(\gamma.z) &= \int_{S^1} e^{-\delta\, b_\xi(\gamma.z,o)} d\nu(\xi) = \int_{S^1} e^{-\delta\, b_{\gamma^{-1}.\xi}(z,\gamma^{-1}.o)} d\nu(\xi) \\ &= \int_{S^1} e^{-\delta\, b_\xi(z,\gamma^{-1}.o)} e^{\delta b_\xi(o,\gamma^{-1}.o)} d\nu(\xi) = \int_{S^1} e^{-\delta\, b_\xi(z,o)} d\nu(\xi) = h^\nu(z). \end{aligned}$$

Conversely, if h is a positive λ-eigenfunction of the hyperbolic Laplacian, then λ belongs to the positive spectrum of L, and therefore $\lambda \geq -1/4$, e.g., see [27]. Let then δ be the solution $\geq 1/2$ of the equation $\delta(\delta - 1) = \lambda$. Then the family $\{P(\cdot, \xi)^\delta, \xi \in S^1\}$ is a complete set of extremal positive λ-eigenfunctions, i.e., any positive λ-eigenfunction h can be presented as

$$h = \int_{S^1} P(\cdot, \xi)^\delta \, d\nu(\xi)$$

for a unique measure ν on S^1. If h is Γ-invariant, then $h \circ \gamma = h$ for any $\gamma \in \Gamma$, and the above computation shows that ν has to be conformal, by the uniqueness of the representing measure ν of h. □

Recall that a point h in a convex cone C of a vector space is said to be *extremal* if $h - g \in C$ for some $g \in C$ implies that $g = th$ for some $t \in [0, 1]$. In the case of the cone $H^+(\Gamma, \lambda)$, this translates into the condition that $g \leq h$ for some $g \in H^+(\Gamma, \lambda)$ implies $g = th$. Ergodicity of a Γ-conformal measure means that it is extremal in the convex cone of Γ-conformal measures of a given dimension, so that we have

Corollary 3.2. *A Γ-conformal measure ν of dimension $\delta \geq 1/2$ is ergodic for the action of Γ on S^1 if and only if the function h^ν is extremal in the cone $H^+(\Gamma, \lambda)$, where $\lambda = \delta(\delta - 1)$.*

Remark 3.3. If ν is conformal, one may consider the family $\{\nu_z, z \in \boldsymbol{D}\}$ of measures on S^1 defined as

$$d\nu_z(\xi) = e^{-\delta\, b_\xi(z,o)} d\nu(\xi),$$

so that the function h^ν is given by $h^\nu(z) = \nu_z(S^1)$. The conformality of ν translates into the invariance relation

$$\gamma \nu_z = \nu_{\gamma.z},$$

and $\{\nu_z, z \in \boldsymbol{D}\}$ is called a Γ-*conformal density* in [24].

Remark 3.4. The fact that the Poisson kernel raised to a power which is less than 1/2 is not extremal is a consequence of the following formula: for $\delta > 1/2$,

$$P(z,\xi)^{1-\delta} = \frac{1}{c(\delta)} \int_{S^1} \frac{P(z,\xi')^{\delta}}{|\xi - \xi'|^{2(1-\delta)}} d\xi' \tag{3.1}$$

for some constant $c(\delta) > 0$, the condition $\delta > 1/2$ ensuring that the kernel

$$\xi' \to 1/|\xi - \xi'|^{2(1-\delta)}$$

belongs to $L^1(S^1, d\xi')$. This relation shows that $P^{1-\delta}(\cdot,\xi)$ is the barycenter of $\{P^{\delta}(\cdot,\xi'), \xi' \in S^1\}$ with respect to the measure $d\xi'/|\xi - \xi'|^{2(1-\delta)}$, and it can be proved using formula (2.2) in the following way. For $z = g.o$, we may write

$$\begin{aligned}
\int_{S^1} \frac{P(z,\xi')^{\delta}}{|\xi - \xi'|^{2(1-\delta)}} d\xi' &= \int_{S^1} \frac{|(g^{-1})'\xi'|^{\delta}}{|\xi - \xi'|^{2(1-\delta)}} d\xi' \\
&= \int_{S^1} \frac{|(g^{-1})'(g.\xi')|^{\delta} |g'(\xi')|}{|\xi - g.\xi'|^{2(1-\delta)}} d\xi' \\
&= \int_{S^1} \frac{|g'(\xi')|^{1-\delta}}{|\xi - g.\xi'|^{2(1-\delta)}} d\xi' \\
&= |(g^{-1})'(\xi)|^{1-\delta} \int_{S^1} \frac{1}{|g^{-1}\xi - \xi'|^{2(1-\delta)}} d\xi'.
\end{aligned}$$

It suffices now to notice that, by rotation invariance, the quantity

$$\int_{S^1} \frac{d\xi'}{|\theta - \xi'|^{2(1-\delta)}}$$

does not depend on $\theta \in S^1$ and is, therefore, a constant $c(\delta)$, in order to get (3.1). There is an analogous formula for $\delta = 1/2$ which can be derived by differentiating (3.1) with respect to δ.

Remark 3.5. A similar computation shows that if ν' is a Γ-conformal measure with dimension $\delta' < 1/2$, then the probability measure ν defined by

$$d\nu(\xi) = \frac{1}{c(\delta)} \int_{S^1} \frac{d\nu(\xi')}{|\xi - \xi'|^{2\delta'}} d\xi$$

is a Γ- conformal measure with dimension $\delta = 1 - \delta' > 1/2$, which, by (3.1), leads to the same eigenfunction $h^{\nu} = h^{\nu'}$.

Let Γ_0 be a co-compact torsion free Fuchsian group, and denote by S_0 the hyperbolic surface $S_0 = \Gamma_0 \backslash \boldsymbol{D}$. Since S_0 is compact, it follows from the maximum principle that harmonic functions ($\lambda = 0$) for the Laplacian on S_0 reduce to the constant functions. Since eigenfunctions are square-integrable, those functions with non-zero eigenvalue which are orthogonal to the function $\mathbf{1}$ must have vanishing integral. This implies that the positive eigenfunctions for the Laplacian on S_0 reduce to

the constant functions. Since, by the classical Poisson representation formula,

$$1 \equiv \int_{S^1} P(z, \xi)\, d\xi,$$

one sees that the Lebesgue measure is the unique (up to a multiplicative constant) conformal measure for the co-compact Fuchsian group Γ_0. In particular, we recover from the discussion in Section 2 the fact that the measure $d\xi \otimes e^{-r} dr$ is the unique measure on the space of horocycles which is Γ_0-invariant and quasi-invariant under the geodesic flow, a very weak form of Furstenberg's theorem.

We now consider a normal subgroup Γ of Γ_0 such that the group $N = \Gamma_0/\Gamma$ is nilpotent. The hyperbolic surface $S = \Gamma\backslash \boldsymbol{D}$ is a nilpotent regular cover of S_0. Let us recall some general results concerning the classification of eigenfunctions on a nilpotent Riemannian cover M of a compact manifold M_0 with the deck group N. The *positive spectrum* of M, i.e., the set of $\lambda \in \mathbb{R}$ such that there exists a positive λ-eigenfunction, equals $[0, +\infty)$ [27], and for $\lambda = 0$, positive harmonic functions on M are constant [19]. (In our setup, this translates into the fact that any conformal measure for Γ has dimension $\delta \geq 1$ and that the Lebesgue measure is the unique Γ-conformal measure of dimension $\delta = 1$).

For $\lambda > 0$, positive λ-eigenfunctions have an integral representation in terms of extremal λ-eigenfunctions. The work of Conze and Guivarc'h [8] shows that extremal eigenfunctions are in a one-to-one correspondence with *exponentials* of N, i.e., group homomorphisms from N to $\mathbb{R}_+^*$. More precisely, to each exponential $\chi : N \to \mathbb{R}_+^*$ corresponds a unique (up to a constant multiplier) eigenfunction h^χ with the property that for any point $x \in M$ and any $n \in N$,

$$h^\chi(n.x) = \chi(n) h^\chi(x) \tag{3.2}$$

(here $x \mapsto n.x$ denotes the action of n on M by deck transformations), and h^χ is extremal. Conversely, any extremal positive eigenfunction is of the form h^χ for some exponential χ.

We are now in a position to state a more precise version of Main Theorem in the particular case of hyperbolic surfaces.

Theorem A. *Let $S = \Gamma\backslash \boldsymbol{D}$ be a nilpotent cover of a compact hyperbolic surface $S_0 = \Gamma_0\backslash \boldsymbol{D}$, with the deck group $N = \Gamma_0/\Gamma$. We fix a point $o \in S$, and denote by A^{tf} the torsion free component of $N/[N, N]$. Then, the following sets are in a one-to-one correspondence:*

(1) *exponentials of N;*

(1′) *exponentials of A^{tf};*

(1″) *cohomology classes of S_0 which vanish on Γ;*

(2) (*rays of*) *extremal positive eigenfunctions for the Laplacian on S* (*with non-negative eigenvalue*)*;*

(3) (*rays of*) *ergodic* Γ*-conformal measures on* S^1 (*of dimension* ≥ 1)*;*

(4) (*rays of*) *Radon measures on the space of horocycles* $\mathcal{H} \simeq S^1 \times \mathbb{R}$ *which are invariant and ergodic for the action of* Γ *and quasi-invariant for the geodesic flow;*

(5) (*rays of*) *ergodic invariant measures for the horocycle flow on* T^1S *which are quasi-invariant under the geodesic flow.*

Proof. To see the correspondence between (1) and (1′), note that an exponential of N is trivial on the commutator group of N and on finite subgroups, so that it can be identified with an exponential of A^{tf}, the torsion free component of the abelian group $A = N/[N, N]$. Conversely, any exponential of A^{tf} can be lifted to an exponential of N. In a similar way, an exponential χ of $N = \Gamma_0/\Gamma$ can be lifted to an exponential of Γ_0 which is trivial on Γ. Such an exponential, being also trivial on the commutator group of Γ_0, is in fact the exponential of an additive homomorphism of $H_1^{\mathrm{tf}}(S_0, \mathbb{Z}) = (\Gamma_0/[\Gamma_0, \Gamma_0])^{\mathrm{tf}}$, and therefore is of the form

$$\chi(\gamma) = e^{\langle \alpha | [\gamma] \rangle}$$

for some linear form $\alpha : H_1^{\mathrm{tf}}(S_0, \mathbb{Z}) \to \mathbb{R}$. By the classical de Rham duality, such a linear form can be identified with a cohomology class $\alpha \in H^1(S_0, \mathbb{R})$. Since χ is trivial on Γ, this class vanishes on (loops of) Γ. Going the other way round gives the correspondence between (1) and (1″).

The correspondence between (1) and (2) is the result of Conze & Guivarc'h discussed above.

The correspondence between (2) and (3) follows from Proposition 3.1 and Corollary 3.2.

In order to prove the correspondence between (3) and (4), we recall from Proposition 2.1 that a measure on the space of horocycles which is $invariant$ and ergodic under the action of Γ and quasi-invariant under the geodesic flow is of the form $m^\nu = \nu \otimes e^{-\delta r} dr$ for some $\delta \geq 0$ and some Γ-conformal measure ν on S^1. Since m^ν is assumed to be ergodic, then trivially the conformal measure ν has to be Γ-ergodic. The opposite direction is not so obvious, because the ergodicity of ν on S^1 does not *a priori* guarantee the ergodicity of the extension m^ν on $S^1 \times \mathbb{R}$. However, it has been observed by Kaimanovich that the action of Γ on $S^1 \times \mathbb{R}$ is the so-called *Radon–Nikodym extension* of the action of Γ on (S^1, ν), since the Busemann cocycle is proportional to the logarithm of the Radon–Nikodym cocycle of ν. In this setup, m^ν is ergodic if and only if ν is ergodic and the ratio set of ν coincides with $\mathbb{R}$ ([23], see also [14, 15, 16]). The latter condition is satisfied here, because the surface S_0 is compact (see [14] for more details). This leads to the correspondence between (3) and (4).

The correspondence between (4) and (5) has been discussed in Section 2

This ends the proof of Theorem A. □

Remark 3.6. Given a cohomology class $\alpha \in H^1(S_0, \mathbb{R})$ which vanishes on Γ, Theorem A ensures the existence of conformal measure ν_α associated with α. We shall briefly remind its explicit construction from [2] which is based on using a twisted Poincaré series. For $s \in \mathbb{R}$ put

$$P_\alpha(s) = \sum_{\gamma \in \Gamma_0} e^{-sd(o,\gamma.o)+\alpha([\gamma])}$$

where $[\gamma]$ denotes the homology class of γ. Denote by δ_α the exponent of convergence of this series. It can be shown, using general results on Gibbs measures for the compact manifold M_0 (e.g., see [18]), that $\delta_\alpha \geq \delta_0 = 1$ and that $P_\alpha(\delta_\alpha) = +\infty$. Moreover, the probability measures ν_α^s (viewed as probability measures on the compact space $\overline{\boldsymbol{D}} = \boldsymbol{D} \cup S^1$) given by the formula

$$\nu_\alpha^s = \frac{1}{P_\alpha(s)} \sum_{\gamma \in \Gamma_0} e^{-sd(o,\gamma.o)+\alpha([\gamma])} \delta_{\gamma.o}$$

converge, as s decreases to δ_α, to a probability measure ν_α on S^1 with the property

$$\frac{d\gamma^{-1}\nu}{d\nu}(\xi) = e^{\alpha([\gamma])}|\gamma'(\xi)|^{\delta_\alpha} \quad \forall\, \gamma \in \Gamma_0$$

Since α vanishes on Γ, the probability measure ν_α is indeed a conformal measure for Γ with dimension $\delta_\alpha \geq 1$. It is also easy to see that the corresponding Radon measures m^{ν_α} on the space of horocycles satisfy the relation

$$\gamma^{-1} m^{\nu_\alpha} = e^{\alpha([\gamma])} m^{\nu_\alpha} \quad \forall\, \gamma \in \Gamma_0,$$

and that the corresponding eigenfunction h^{ν_α} on $S = \Gamma \backslash \boldsymbol{D}$ transforms according to the rule (3.2), where $\chi(\gamma) = e^{\alpha([\gamma])}$.

Remark 3.7. Recall the identification of the space of horocycles $\mathcal{H}$ with the space $\mathbb{R}^2 \setminus \{0\}/\{\pm\ \mathrm{Id}\} \simeq P^1(\mathbb{R}) \times (0, +\infty)$ (see Section 2). In this model, the action of the group of isometries on $\mathcal{H}$ corresponds to the linear action of $\mathrm{SL}(2, \mathbb{R})$ on $\mathbb{R}^2$. The Lebesgue measure on $\mathbb{R}^2$ being $\mathrm{SL}(2, \mathbb{R})$-invariant, it corresponds to a measure on $\mathbb{R}^2 \setminus \{0\}/\{\pm\ \mathrm{Id}\}$ which is invariant with respect to any Fuchsian group: this is the measure $\nu_0 \otimes \rho d\rho$, where ν_0 is the Lebesgue measure on $P^1(\mathbb{R})$. The above construction shows that there exist other invariant measures on $\mathbb{R}^2 \setminus \{0\}/\{\pm\ \mathrm{Id}\}$ for the action of $\Gamma \lhd \Gamma_0$ as above, which are of the form $\nu^\alpha \otimes \rho^{2\delta_\alpha - 1} d\rho$.

Remark 3.8. The results of this section extend almost *verbatim* to Kleinian groups in the higher dimensional hyperbolic space $\boldsymbol{H}^n$, the only difference being that conformal measures of dimension δ correspond to positive λ-eigenfunctions, where λ is now given by the formula $\lambda = \delta(\delta - (n-1))$. Therefore, only conformal measures of dimension $\geq (n-1)/2$ need to be considered for an arbitrary Kleinian group Γ. In the case $\Gamma \lhd \Gamma_0$ with co-compact Γ_0 and nilpotent quotient, only conformal measures of dimension $\geq n-1$ occur.

Remark 3.9. Similar results also hold for rank 1 Riemannian symmetric spaces, since they still have the key property that a power of the Poisson kernel is an eigenfunction.

4 General negatively curved manifolds

For arbitrary negatively curved manifolds, there is no obvious relationship between the Poisson kernel $P(g^{-1}.o, \xi)$ and the conformal coefficient $|g'(\xi)|$ as was the case above. However, the work of Hamenstädt [11, 12] still allows one to use potential theory for studying conformal measures on $\partial\widetilde{M}$.

We recall from Section 2 that $\widetilde{M}$ denotes a simply connected Riemannian manifold with negatively pinched sectional curvature, $\partial\widetilde{M}$ – its boundary at infinity, and $b_\xi(\cdot,\cdot)$ – the Busemann cocycle determined by a point $\xi \in \partial\widetilde{M}$. In the following, o denotes a fixed point of $\widetilde{M}$, the origin. If Γ is a discrete subgroup of $\mathrm{Iso}(\widetilde{M})$, then a Γ-*conformal density* $\{\nu_p\}_{p\in\widetilde{M}}$ of dimension δ is a family of measures on $\partial\widetilde{M}$ in the same measure class such that, for any $\gamma \in \Gamma$ and $(p,q) \in \widetilde{M}^2$,

$$\gamma\nu_p = \nu_{\gamma.p} \quad \text{and} \quad \frac{d\nu_p}{d\nu_q}(\xi) = e^{-\delta b_\xi(p,q)}.$$

This density is said to be *normalized* if $\nu_o(\partial\widetilde{M}) = 1$. As we have seen in Remark 3.3, one may always construct a conformal density from a conformal measure.

We now introduce the potential theory that will be used in order to replace harmonic functions. Recall that $T^1\widetilde{M}$ is foliated by the stable foliation $\widetilde{\mathcal{W}}^{\mathrm{s}}$ of the geodesic flow $\widetilde{g}_t$ of $\widetilde{M}$, where the leaf of a unit vector v is

$$W^{\mathrm{s}}(v) = \Big\{w \in T^1\widetilde{M} : \sup_{t\geq 0} D(\widetilde{g}_t(v), \widetilde{g}_t(w)) < +\infty\Big\} = \bigcup_{t\in\mathbb{R}} \widetilde{g}_t W^{\mathrm{ss}}(v).$$

(here we use the Sasaki metric D on $T^1\widetilde{M}$). In particular, the map $T^1\widetilde{M} \to \widetilde{M}$ which associates to a vector v its base point allows one to identify each leaf with $\widetilde{M}$, so that the Laplace–Beltrami operator $\widetilde{\Delta}$ of $\widetilde{M}$ induces on each leaf $W^{\mathrm{s}}(v)$ a differential operator Δ^{s}_v. The operator Δ^{s} is then the foliated differential operator given leafwise by the family $(\Delta^{\mathrm{s}}_v)_{v\in T^1\widetilde{M}}$

We also denote by X the geodesic spray, i.e. the vector field on $T^1\widetilde{M}$ defined as $X\phi(v) = \lim_{t\to 0}[\phi(\widetilde{g}_t v) - \phi(v)]/t$. We note that $X\phi(v)$ only depends on the values of ϕ on $W^{\mathrm{s}}(v)$. Geometrically, X_v can be identified with the set of vectors pointing to the point at infinity of v.

A Radon measure η on $T^1\widetilde{M}$ is called *harmonic for* $\Delta^{\mathrm{s}} + \delta X$, or δ-*harmonic* if for any $\phi \in C^2_c(T^1\widetilde{M})$, one has

$$\int_{T^1\widetilde{M}} (\Delta^{\mathrm{s}} + \delta X)\phi \, d\eta = 0.$$

It is *δ-self-adjoint* if for any $\phi, \psi \in C_c^2(T^1\widetilde{M})$,

$$\int_{T^1\widetilde{M}} \phi\ (\Delta^s + \delta X)\psi\ d\eta = \int_{T^1\widetilde{M}} \psi\ (\Delta^s + \delta X)\phi\ d\eta. \tag{4.1}$$

A δ-self-adjoint measure is δ-harmonic.

We remark that the Laplace operator commutes with any isometry of $\widetilde{M}$, so that if Γ is any discrete subgroup of $\mathrm{Iso}(\widetilde{M})$ and $M = \Gamma\backslash\widetilde{M}$, one may define accordingly the foliated operator Δ_Γ^s on $\Gamma\backslash T^1\widetilde{M} = T^1M$. A δ-harmonic (resp. δ-self-adjoint) measure for Δ_Γ^s is a Radon measure on T^1M which satisfies on T^1M a relation similar to (4.1):

$$\int_{T^1M} (\Delta_\Gamma^s + \delta X)\phi\, d\eta = 0 \quad \text{resp.} \quad \int_{T^1M} \psi(\Delta_\Gamma^s + \delta X)\phi\, d\eta = \int_{T^1M} \phi(\Delta_\Gamma^s + \delta X)\psi\, d\eta$$

for any $\phi, \psi \in C_c^2(T^1M)$. Obviously, δ-self-adjoint measures on T^1M are in a one-to-one correspondence with δ-self-adjoint measures on $T^1\widetilde{M}$ which are Γ-invariant.

Proposition 4.1 ([11]). *Let Γ be a torsion free discrete subgroup of $\mathrm{Iso}(\widetilde{M})$, and $M = \Gamma\backslash\widetilde{M}$. We fix a fundamental domain M_0 for the action of Γ on $\widetilde{M}$, and parameterize T^1M by $M_0 \times \partial\widetilde{M}$. If $\nu = (\nu_p)_{p\in\widetilde{M}}$ is a Γ-conformal density of dimension δ, then the measure η^ν on T^1M defined as*

$$\eta^\nu(\phi) = \int_{M_0\times\partial\widetilde{M}} \widetilde{\phi}(p,\xi) dp d\nu_p(\xi)$$

is a self-adjoint measure for $\Delta_\Gamma^s + \delta X$. Conversely, any self-adjoint measure for Δ_Γ^s is of the form η^ν for a unique conformal density ν.

This result has been first proved in [18] for compact negatively curved manifolds. It was extended to a much more general setting in [11, Corollary 2.6, pp. 48, 49]. We note that the proof of Corollary 2.6 only depends on local computations and does not require the manifold to be compact, as opposed to [11, Corollary 2.4], where the uniqueness of a self-adjoint harmonic measure relies heavily on the compactness of M.

The space $H^s(\Gamma, \delta)$ of self-adjoint measures for $\Delta_\Gamma^s + \delta X$ is a convex cone, which, endowed with the weak*-topology, is homeomorphic by the previous proposition to the space of Γ-conformal measures of dimension δ. Take the *base* of this cone consisting of normalized δ-self-adjoint measures, where a δ- harmonic measure η^ν is called *normalized* if the corresponding conformal density ν is normalized. Since the set of δ-conformal probability measures is closed inside the compact set of probability measures on $\partial\widetilde{M}$, this base is compact.

As in Corollary 3.2, one has

Proposition 4.2 ([12]). *The measure class of a Γ-conformal density ν of dimension δ is ergodic if and only if the harmonic measure η^ν on T^1M is extremal in the cone $H^s(\Gamma, \delta)$ of self-adjoint measures for $\Delta_\Gamma^s + \delta X$.*

Let us first consider the case when $\Gamma = \Gamma_0$ is co-compact. Then, by [13], see also [18], there exists a unique Γ-conformal density, of dimension $\delta = h$, where h denotes

the volume growth of $\widetilde{M}$:

$$h = \lim_{R\to+\infty} \frac{1}{R} \log \text{vol } B(p, R)$$

(which does not depends on the chosen point $p \in \widetilde{M}$). It implies that there are no δ-conformal measures for Γ_0 other than the one obtained from Patterson's construction, for $\delta = h$.

We now turn to the case when $\Gamma \lhd \Gamma_0$ is a normal subgroup of Γ_0 with nilpotent quotient $N = \Gamma_0/\Gamma$.

We now quote the following result of Hamenstädt, which is analogous to the result by Conze et Guivarc'h described in Section 3:

Proposition 4.3 ([12]). *For any extremal point η in $H^s(\Gamma, \delta)$, there exists an exponential χ of N such that, for any deck transformation $n \in N$, acting on T^1M*

$$n\eta = \chi(n)\eta \quad \forall n \in N.$$

These Propositions put together lead to the following

Corollary 4.4. *Let Γ be a normal subgroup of a co-compact subgroup Γ_0 of $\text{Iso}(\widetilde{M})$ such that $N = \Gamma_0/\Gamma$ is nilpotent. Then any Γ-conformal ergodic measure on $\partial\widetilde{M}$ is quasi-invariant under the action of Γ_0: there exist $\delta \geq h$ and an exponential χ of Γ_0 which is trivial on Γ such that for any γ_0 in Γ_0,*

$$\frac{d\gamma_0\nu}{d\nu}(\xi) = \chi(\gamma_0)|\gamma_0'(\xi)|^\delta. \tag{4.2}$$

For γ_0 in Γ_0, let $[\gamma]$ be the torsion free homology class of γ. An exponential of Γ_0 can be written as

$$\chi(\gamma_0) = e^{\alpha([\gamma_0])}$$

for some cohomology class $\alpha \in H_1(M_0, \mathbb{R})$.

As outlined in Remark 3.6, one may conversely associate to a class α vanishing on Γ a Γ-conformal measure of dimension $\delta^\alpha \geq h$, such that ν^α is quasi-invariant under the action of Γ_0 and transforms according to (4.2). That ν^α is the unique conformal measure satisfying (4.2) follows from the fact that ν^α can also be constructed as the measure at infinity induced by the equilibrium state of the function ϕ_α on T^1M induced by α; the exponent δ_α is then the pressure of ϕ_α [3].

To summarize, we get:

Theorem B. *Let $M = \Gamma\backslash\widetilde{M}$ be a nilpotent cover of a compact negatively curved manifold $M_0 = \Gamma_0\backslash\widetilde{M}$ with the deck group $N = \Gamma_0/\Gamma$. We let h be the volume entropy of $\widetilde{M}$. We fix a point $o \in M$, and denote by A^{tf} the torsion free component of $N/[N, N]$. Then, the following sets are in a one-to-one correspondence:*

(1) *exponentials of N;*

(1′) *exponentials of A^{tf};*

(1″) *cohomology classes of M_0 which vanish on Γ;*

(2) *extremal normalized self-adjoint δ-harmonic measures (for $\delta \geq h$);*

(3) *ergodic Γ-conformal probability measures (of dimension $\geq h$) on $\partial\widetilde{M}$;*

(3′) *Γ-conformal probability measures on $\partial\widetilde{M}$ which are quasi-invariant and ergodic under the action of Γ_0;*

(4) *(rays of) Radon measures on the space of horospheres $\mathcal{H} \simeq \partial\widetilde{M} \times \mathbb{R}$ which are invariant and ergodic for the action of Γ and quasi-invariant for the geodesic flow;*

(5) *(rays of) ergodic invariant measures for the horosphere foliation on T^1M which are quasi-invariant under the geodesic flow.*

Proof. The only point which has not been discussed is the fact that if ν is conformal ergodic for Γ, then the measure m^ν induced on the space of horospheres is ergodic. For this we refer to [14]. One may also refer to [16] or [9], since an ergodic conformal measure ν is of the measure class at infinity induced by a Gibbs state. □

Acknowledgement. I wish to thank the organizers of the conference in Vienna, for giving me the opportunity to discuss these problems, in particular with R. Solomyak.

Added in proof. Since this paper had been written a more direct approach to the problem of classification of conformal measures was proposed by Thomas Roblin [22].

Editor's comment (March 2004). The recent preprint "Invariant Radon measures for horocycle flows on abelian covers" by Omri Sarig contains a complete classification of ergodic invariant Radon measures for the horocycle flow on abelian covers of compact Riemann surfaces of negative curvature (more generally, it is also applicable to the horosphere foliation in the higher dimensional case) proving the conjecture of Babillot and Ledrappier from [3] discussed in the Introduction to this article.

References

[1] J.Aaronson, O. Sarig and R. Solomyak, Tail-invariant measures for some suspension semiflows, Discrete Contin. Dynam. Systems 8 (2002), 725–735.

[2] M. Babillot, Géodésiques et horocycles sur le revêtement d'homologie d'une surface hyperbolique, Sémin. Théor. Spectr. Géom. 14 (1995–1996), 89–104.

[3] M. Babillot and F. Ledrappier, Geodesic paths and horocycle flow on abelian covers, in: Lie Groups and Ergodic Theory (Mumbai, 1996), Tata Inst. Fund. Res. Stud. Math. 14, Tata Inst. Fund. Res., Bombay 1998, 1–32.

[4] W. Ballmann, Lectures on Spaces of Nonpositive Curvature. With an appendix by Misha Brin, DMV Sem. 25, Birkhäuser, Basel 1995.

[5] M. Bourdon, Structure conforme au bord et flot géodésique d'un CAT(-1)-espace, Enseign. Math. (2) 41 (1995), 63–102.

[6] R. Bowen and B. Marcus, Unique ergodicity for horocycle foliations, Israel J. Math. 26 (1977), 43–67.

[7] M. Burger, Horocycle flow on geometrically finite surfaces, Duke Math. J. 61 (1990), 779–803.

[8] J.-P. Conze and Y. Guivarc'h, Propriété de droite fixe et fonctions propres des opérateurs de convolution, in: Théorie du Potentiel et Analyse Harmonique (Exposés des Journées de la Soc. Math. France, Inst. Recherche Math. Avancée, Strasbourg, 1973), Lecture Notes in Math. 404, Springer-Verlag, Berlin 1974, 126–132.

[9] Y. Coudène, Cocycles and stable foliations of Axiom A flows, Ergodic Theory Dynam. Systems 21 (2001), 767–775.

[10] H. Furstenberg, The unique ergodicity of the horocycle flow, in: Recent Advances in Topological Dynamics (Proc. Conf., Yale Univ., New Haven, Conn., 1972; in honor of Gustav Arnold Hedlund), Lecture Notes in Math. 318, Springer-Verlag, Berlin 1973, 95–115.

[11] U. Hamenstädt, Harmonic measures for compact negatively curved manifolds, Acta Math. 178 (1997), 39–107.

[12] U. Hamenstädt, Ergodic properties of Gibbs measures on nilpotent covers, Ergodic Theory Dynam. Systems 22 (2002), 1169–1179.

[13] V. A. Kaimanovich, Invariant measures of the geodesic flow and measures at infinity on negatively curved manifolds, Ann. Inst. H. Poincaré Phys. Théor. 53 (1990), 361–393.

[14] V. A. Kaimanovich, Ergodic properties of the horocycle flow and classification of Fuchsian groups, J. Dynam. Control Systems 6 (2000), 21–56.

[15] V. A. Kaimanovich, SAT actions and ergodic properties of the horosphere foliation, in: Rigidity in Dynamics and Geometry (Cambridge, 2000), 261–282, Springer-Verlag, Berlin 2002, 261–282.

[16] V. A. Kaimanovich and K. Schmidt, Ergodicity of cocycles. I: General theory, preprint (2000).

[17] F. Ledrappier, Ergodic properties of the stable foliations, in: Ergodic Theory and Related Topics, III (Gustrow, 1990), Lecture Notes in Math. 1514, Springer-Verlag, Berlin 1992, 131–145.

[18] F. Ledrappier, Structure au bord des variétés a courbure négative, Sémin. Théor. Spectr. Géom. 13 (1994–1995), 97–122.

[19] G. A. Margulis, Positive harmonic functions on nilpotent groups, Soviet Math. Dokl. 7 (1966), 241–244.

[20] M. Ratner, Interactions between ergodic theory, Lie groups and number theory, in: Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994), Birkhäuser, Basel 1995, 157–182.

[21] T. Roblin, Ergodicité et unique ergodicité du feuilletage horosphérique, mélange du flot géodésique et équidistributions diverses dans les groupes discrets en courbure négative, preprint (2001).

[22] T. Roblin, Un théorème de Fatou pour les densités conformes avec applications aux revêtements galoisiens en courbure négative, preprint (2003).

[23] K. Schmidt, Cocycles of Ergodic Transformation Groups, McMillan (India), New Delhi 1977.

[24] D. Sullivan, The density at infinity of a discrete group of hyperbolic motions, Inst. Hautes Études Sci. Publ. Math. 50 (1979), 171–202.

[25] D. Sullivan, On the ergodic theory at infinity of an arbitrary discrete group of hyperbolic motions, in: Riemann Surfaces and Related Topics: Proceedings of the 1978 Stony Brook Conference (State Univ. New York, Stony Brook, NY, 1978), Ann. of Math. Stud. 97, Princeton University Press, Princeton, NJ, 1981, 465–496.

[26] D. Sullivan, Entropy, Hausdorff measures old and new and limit sets of geometrically finite Kleinian groups, Acta Math. 153 (1984), 259–277.

[27] D. Sullivan, Related aspects of positivity in Riemannian geometry, J. Differential Geom. 25 (1987), 327–351.

Markov processes on vermiculated spaces

Martin T. Barlow and Steven N. Evans

Abstract. A general technique is given for constructing new Markov processes from existing ones. The new process and its state space are both projective limits of sequences built by an iterative scheme. The space at each stage in the scheme is obtained by taking disjoint copies of the space at the previous stage and quotienting to identify certain distinguished points. Away from the distinguished points, the process at each stage evolves like the one constructed at the previous stage on some copy of the previous state space, but when the process hits a distinguished point it enters at random another of the copies "pinned" at that point. Special cases of this construction produce diffusions on fractal-like objects that have been studied recently.

1 Introduction

In this paper we present a procedure for constructing new Markov processes from existing ones. This procedure can, for example, produce processes on rather exotic fractal-like spaces starting with processes on more familiar spaces such as Euclidean space.

The state space of the new process is built as a projective limit of an iterative scheme. The space at each stage of the iterative construction is produced by taking a disjoint collection of copies of the space coming from the previous stage and performing a quotient operation that identifies certain distinguished points, which we call "wormholes".

A particular case of the state space construction, beginning with $[0, 1]$, was given in [13] and used to give examples of analytically regular spaces L_Q with Hausdorff dimension Q for any $Q \in [1, \infty)$. (This answered a question in [10]). A similar construction in a discrete setting was used in [1] to show that the only constraints on the volume growth and the anomalous diffusion exponent are those already known in the literature.

These spaces are interesting in their own right, and also useful building blocks for counterexamples. For example if $Q > 1$ and two copies of L_Q are joined at a single point the resulting space $L_Q \cup L'_Q$ satisfies a $(2, 2)$ Poincaré inequality. Hence, by [9, 14], the heat kernel has the usual Gaussian upper and lower bounds. On the other hand the stronger $(1, 1)$ Poincaré inequality fails for this space, which shows that isoperimetric regularity is a strictly stronger condition than heat kernel

regularity. Similarly one can use the spaces constructed in [1] to show that the property of satisfying an elliptic Harnack inequality is not stable under products.

While the construction in [13] is analytic, it gives rise to a "Laplacian" type operator on the spaces L_Q, so that it is natural to ask about the properties of the corresponding Brownian motion. In this paper we present a probabilistic construction of this process, which works for a wide class of base spaces and Markov processes.

To give the flavour of the construction, we begin with a very simple example described in somewhat fanciful terms. Consider a "universe" that is a subinterval F of $\mathbb{R}$ and another "parallel" universe that is just a copy of F. At the same locations in each universe there is a set of "anchor points" B_1. Figure 1 shows a universe with 3 such anchor points.

Figure 1. The space F with B_1 marked as *'s

In the language of innumerable science fiction stories, imagine that the two universes are connected by "wormholes" at the anchor points. This produces a composite structure F_1 that, mathematically, is the product space $F \times \{0, 1\}$ with points of the form $(b, 0)$ and $(b, 1)$, $b \in B_1$, identified. The space F_1 with the usual quotient topology looks like the subset of the plane drawn in Figure 2. For future purposes, it will

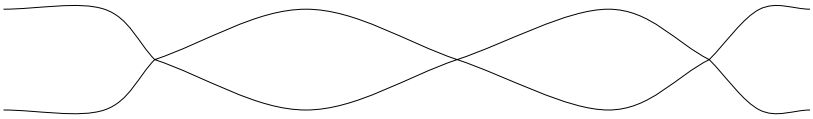


Figure 2. The space F_1 embedded in the plane

be more convenient to represent F_1 schematically as in Figure 3 as two copies of F with the points at either end of an arrow identified.

Figure 3. The space F_1 as two copies of F with identified points joined by $\updownarrow$

At an intuitive level, there is an obvious way to "lift" a base Markov process on F to a Markov process on F_1. Namely, the lifted process evolves as the base process away from the anchor points, but when it hits an anchor point it chooses at random to either keep evolving in its current universe or to jump through the attached wormhole

into the alternate universe. Of course, if the base process is something like Brownian motion, then there is some work that needs to be done to make this idea precise because of the fact that a Brownian motion returns to its starting point infinitely often in any neighbourhood of the origin. For Brownian motion, the technicalities involved in making sense of idea are of the same sort as those encountered in the construction of Walsh's spider and Brownian motion on more general graphs (see [18, 17, 3, 4, 7, 8, 12, 16, 2, 5]). A substantial generalisation of the spider construction that applies to base processes on state spaces more general than the real line is given in [6]. That generalisation is basic to this paper and is reviewed in Section 2.

The construction that produced F_1 from F can be iterated. Suppose that F now has "first order" anchor points B_1 and "second order" anchor points B_2 as in Figure 4. An

Figure 4. The space F with B_1 as $*$'s and B_2 as $+$'s

obvious quotient construction on $F_1 \times \{0, 1\}$ produces a space F_2 shown schematically in Figure 5 as four copies of F with the points at either end of an arrow identified. The

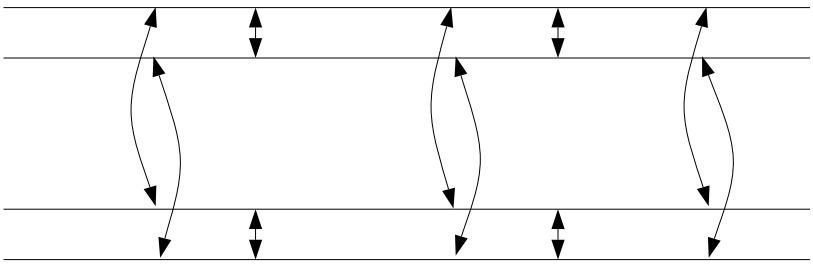


Figure 5. The space F_2 as four copies of F with identified points joined by $\updownarrow$

Markov process that was constructed on F_1 can be lifted to F_2 by once again making suitable random choices whenever a second order anchor point is encountered.

Continuing in this manner produces a sequence of spaces $F_1, F_2, \ldots$ These spaces form a projective system and hence converge to a projective limit space F_∞. We will refer to the limit spaces produced by a generalisation of this construction (see Section 3) as *vermiculated* (that is, riddled with worm holes). Furthermore, the associated Markov processes have a natural projective structure, and consequently they give rise to a limit process on F_∞. The details are carried out in great generality in Section 4.

Properties of the limit Markov process for the case of a Brownian or Lévy base process will be investigated in a subsequent paper. The potential theory of such processes is particularly interesting. For example, it is possible that the base process

hits points whereas the limit process does not if the anchor points and the number of copies at each stage in the iterative scheme are chosen correctly.

2 The pinching and twisting construction

In this section we review quickly a construction from [6] that produces one Markov process from another by means of a partial collapse of the state space and the introduction of appropriate extra randomisation. As we noted in the introduction, this construction can be seen as a generalisation of the construction that produces Walsh's spider from Brownian motion on the line.

We begin with some topological ingredients. Let E and $\hat{E}$ be two Hausdorff, locally compact, second countable topological spaces. Thus E and $\hat{E}$ are, in particular, Polish (that is, metrisable as complete, separable metric spaces). Fix a continuous surjection $\psi : E \to \hat{E}$ such that $\psi^{-1}(K)$ is compact for any compact subset K of $\hat{E}$ and a closed set $A \subseteq E$ such that $\psi^{-1}(\psi(A)) = A$.

Set

$$\tilde{E} := (E \setminus A) \cup \psi(A),$$

this being a disjoint union. Define the map $\pi : E \to \tilde{E}$ by

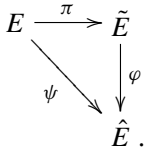
$$\pi(x) := \begin{cases} x, & \text{if } x \in E \setminus A, \\ \psi(x), & \text{if } x \in A, \end{cases}$$

and give $\tilde{E}$ the topology induced by π. That is, $U \subset \tilde{E}$ is open in the topology of $\tilde{E}$ if and only if $\pi^{-1}(U)$ is open in the topology of E. (Equivalently, we can think of $\tilde{E}$ as the quotient topological space of the topological space E under the equivalence relation that declares two points x' and x'' equivalent if and only if $\pi(x') = \pi(x'')$.) Assume that $\tilde{E}$ with this topology is Hausdorff, locally compact, and second countable (and hence Polish).

Define a continuous map $\varphi : \tilde{E} \to \hat{E}$ by

$$\varphi(x) := \begin{cases} \psi(x), & \text{if } x \in E \setminus A, \\ x, & \text{if } x \in \psi(A). \end{cases}$$

Then, $\psi = \varphi \circ \pi$, or, equivalently, we have the commutative diagram



Notation 2.1. Given a Hausdorff, locally compact, second countable topological space S, write $B(S)$ for the space of bounded real-valued Borel functions on S and

let $B^+(S)$ be the collection of nonnegative elements of $B(S)$. Let $C_0(S)$ be the Banach space of real-valued continuous functions on S that vanish at infinity (if S is compact, then of course $C_0(S) = C(S)$, the Banach space of continuous functions on S). For any subset R of S, define $B(S; R) := \{f \in B(S) : f\big|_R \equiv 0\}$ and $C_0(S; R) := C_0(S) \cap B(S; R)$. If S' is a second locally compact space and ξ is a Borel map from S to S', we define $\xi^* : B(S') \to B(S)$ as $\xi^* f := f \circ \xi$. If ξ is continuous and $\xi^{-1}(K)$ is a compact subset of S for all compact subsets K of S', then $\xi^* : C_0(S') \to C_0(S)$.

Thus,

$$\begin{array}{ccc} B(E) & \xleftarrow{\pi^*} & B(\tilde{E}) \\ & \nwarrow^{\psi^*} & \uparrow{\scriptstyle \varphi^*} \\ & & B(\hat{E})\ . \end{array}$$

Moreover, $\varphi^* : C_0(\hat{E}) \to C_0(\tilde{E})$, $\pi^* : C_0(\tilde{E}) \to C_0(E)$, and $\psi^* : C_0(\hat{E}) \to C_0(E)$. Define $\check{\pi}_* : B(E; A) \to B(\tilde{E}; \psi(A))$ by

$$(\check{\pi}_* f)(x) := \begin{cases} f(x), & \text{if } x \in E \setminus A, \\ 0, & \text{if } x \in \psi(A). \end{cases}$$

Note that $\check{\pi}_* : C_0(E; A) \to C_0(\tilde{E}; \psi(A))$.

We now introduce the probabilistic ingredients of the construction.

Assumption 2.2. Let $X = (\Omega, \mathcal{F}, \mathcal{F}_t, X_t, \theta_t, \mathbf{P}^x)$ (resp., $\hat{X} = (\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathcal{F}}_t, \hat{X}_t, \hat{\theta}_t, \hat{\mathbf{P}}^x)$) be a quasi-left-continuous Borel right process with state space E (resp., $\hat{E}$) and transition semigroup (P_t) (resp., $(\hat{P}_t)$). Let $k : \hat{E} \times \mathcal{B}(E) \to \mathbb{R}$ be a probability kernel. Define a linear operator $K : B(E) \to B(\hat{E})$ by $Kf(x) := \int_{y \in E} f(y) k(x, dy)$. We assume that

$$k(x, \psi^{-1}\{x\}) = 1$$

for all $x \in \hat{E}$. Assume that (P_t), $(\hat{P}_t)$, ψ and K satisfy the *Dynkin intertwining relation*

$$P_t \psi^* = \psi^* \hat{P}_t \tag{2.1}$$

and the *Carmona–Petit–Yor intertwining relation*

$$K P_t = \hat{P}_t K. \tag{2.2}$$

Notation 2.3. Define the stopping times

$$T(\omega) := \inf\{t \geq 0 : X_t(\omega) \in A\}, \qquad \omega \in \Omega,$$
$$\hat{T}(\hat{\omega}) := \inf\{t \geq 0 : \hat{X}_t(\hat{\omega}) \in \psi(A)\}, \quad \hat{\omega} \in \hat{\Omega}.$$

Since $\psi^{-1}(\psi(A)) = A$, the $\mathbf{P}^x$-law of T is the same as the $\hat{\mathbf{P}}^{\psi(x)}$ law of $\hat{T}$ for any $x \in E$. Let (Q_t) and $(\hat{Q}_t)$ be, respectively, the semigroups for X stopped at T and $\hat{X}$ stopped at $\hat{T}$; that is,

$$(Q_t f)(x) := \mathbf{P}^x [f(X_{t\wedge T})], \quad f \in B(E),\ x \in E,$$
$$(\hat{Q}_t f)(x) := \hat{\mathbf{P}}^x \left[f(\hat{X}_{t\wedge \hat{T}}) \right], \quad f \in B(\hat{E}),\ x \in \hat{E}.$$

For $\alpha > 0$, define the operators

$$\begin{aligned} U^\alpha f(x) &:= \int_0^\infty e^{-\alpha t} P_t f(x)\, dt = \mathbf{P}^x \left[\int_0^\infty e^{-\alpha t} f(X_t)\, dt \right], \\ P_T^\alpha f(x) &:= \mathbf{P}^x \left[e^{-\alpha T} f(X_T) \right], \\ V^\alpha f(x) &:= \int_0^\infty e^{-\alpha t} Q_t f(x)\, dt = \mathbf{P}^x \left[\int_0^\infty e^{-\alpha t} f(X_{t\wedge T})\, dt \right] \\ &= U^\alpha f(x) - P_T^\alpha U^\alpha f(x) + \alpha^{-1} P_T^\alpha f(x) \end{aligned} \tag{2.3}$$

for all $f \in B(E)$ and $x \in E$; U^α is the α–resolvent of the semigroup (P_t), and V^α is the α–resolvent of (Q_t). Similarly define

$$\begin{aligned} \hat{U}^\alpha f(x) &:= \int_0^\infty e^{-\alpha t} \hat{P}_t f(x)\, dt = \hat{\mathbf{P}}^x \left[\int_0^\infty e^{-\alpha t} f(\hat{X}_t)\, dt \right], \\ \hat{P}_{\hat{T}}^\alpha f(x) &:= \hat{\mathbf{P}}^x \left[e^{-\alpha \hat{T}} f(\hat{X}_{\hat{T}}) \right], \\ \hat{V}^\alpha f(x) &:= \int_0^\infty e^{-\alpha t} \hat{Q}_t f(x)\, dt = \hat{\mathbf{P}}^x \left[\int_0^\infty e^{-\alpha t} f(\hat{X}_{t\wedge \hat{T}})\, dt \right] \\ &= \hat{U}^\alpha f(x) - \hat{P}_{\hat{T}}^\alpha \hat{U}^\alpha f(x) + \alpha^{-1} \hat{P}_{\hat{T}}^\alpha f(x) \end{aligned} \tag{2.4}$$

for all $f \in B(\hat{E})$ and $x \in \hat{E}$.

The following result is proved in [6]. Intuitively, it describes the construction of a process on $\tilde{E}$ that evolves as X on $E \setminus A$ and as $\hat{X} \equiv \psi \circ X$ on $\psi(A)$. When this process passes from $\psi(A)$ into $E \setminus A$ it undergoes a random twist according to the kernel k. We refer the reader to [6] for more details and several examples.

Theorem 2.4. *Suppose that $U^\alpha C_0(E) \subseteq C_0(E)$ and $P_T^\alpha C_0(E) \subseteq C_0(E)$ for each $\alpha > 0$, and that $K C_0(E) \subseteq C_0(\hat{E})$. Then the following hold.*

a) *There is a quasi–left–continuous Borel right process*

$$\tilde{X} = (\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathcal{F}}_t, \tilde{X}_t, \tilde{\theta}_t, \tilde{\mathbf{P}}^x)$$

with state space $\tilde{E}$, with transition semigroup $(\tilde{P}_t)$ given by

$$\tilde{P}_t = \check{\pi}_* Q_t \pi^* \left(I_{\tilde{E}} - \phi^* K \pi^* \right) + \phi^* \hat{P}_t K \pi^*, \tag{2.5}$$

this expression being well-defined, and with resolvent $(\tilde{U}^\alpha)$ *given by*

$$\tilde{U}^\alpha = \check{\pi}_* V^\alpha \pi^* \left(I_{\tilde{E}} - \phi^* K \pi^*\right) + \phi^* \hat{U}^\alpha K \pi^*, \tag{2.6}$$

this expression being well-defined.

b) *For each* $x \in \tilde{E}$*, the law of* $\phi \circ X$ *under* $\tilde{\mathbf{P}}^x$ *coincides with that of* $\hat{X}$ *under* $\hat{\mathbf{P}}^{\phi(x)}$*; in particular,* $\tilde{P}_t \phi^* = \phi^* \hat{P}_t$*.*

c) *Define a stopping time by*

$$\tilde{T}(\tilde{\omega}) := \inf\left\{t \geq 0 :\ \tilde{X}_t(\tilde{\omega}) \in \psi(A)\right\}, \quad \tilde{\omega} \in \tilde{\Omega},$$

and define a semigroup

$$\tilde{Q}_t f(x) := \tilde{\mathbf{P}}^x \left[f\left(\tilde{X}_{t \wedge \tilde{T}}\right)\right], \quad t \geq 0,\ f \in B(\tilde{E}),\ x \in \tilde{E}.$$

For each $x \in E$*, the* $\tilde{\mathbf{P}}^{\pi(x)}$*-law of* $\{\tilde{X}_t;\ 0 \leq t < \tilde{T}\}$ *is equal to the* $\mathbf{P}^x$*-law of* $\{\pi(X_t);\ 0 \leq t < T\}$*; in particular,* $\pi^* \tilde{Q}_t = Q_t \pi^*$*.*

d) *The semigroup* $(\tilde{P}_t)$ *is Feller (that is,* $\tilde{P}_t C_0(\tilde{E}) \subseteq C_0(\tilde{E})$ *for each* $t \geq 0$*, and* $\lim_{t \downarrow 0} \sup_{x \in \tilde{E}} |\tilde{P}_t f(x) - f(x)| = 0$ *for all* $f \in C_0(\tilde{E})$*).*

The following stopped version of Theorem 2.4 is not proved in [6], but follows using the same ideas.

Notation 2.5. Let $B \subseteq E$ be a closed set such that $\psi^{-1}(\psi(B)) = B$. Write $(\underline{P}_t)$ and $(\underline{V}^\alpha)_{\alpha>0}$ for the transition semigroup and resolvent of X stopped at the first hitting time $\underline{T}$ of $A \cup B$. Denote by $(\underline{\tilde{P}}_t)$ and $(\underline{\tilde{U}}^\alpha)_{\alpha>0}$ (resp., $(\underline{\hat{P}}_t)$ and $(\underline{\hat{U}}^\alpha)_{\alpha>0}$) the transition semigroup and resolvent of $\tilde{X}$ (resp., $\hat{X}$) stopped on hitting $\pi(B)$ (resp., $\psi(B)$).

Corollary 2.6. *Suppose that the assumptions of Theorem* 2.4 *hold. Suppose further that* $\underline{U}^\alpha C_0(E) \subseteq C_0(E)$ *and* $P^\alpha_{\underline{T}} C_0(E) \subseteq C_0(E)$ *for each* $\alpha > 0$*. Then*

$$\underline{\tilde{P}}_t = \check{\pi}_* \underline{P}_t \pi^* \left(I_{\tilde{E}} - \varphi^* K \pi^*\right) + \varphi^* \underline{\hat{P}}_t \pi^*,$$

and

$$\underline{\tilde{U}}^\alpha = \check{\pi}_* \underline{V}^\alpha \pi^* \left(I_{\tilde{E}} - \varphi^* K \pi^*\right) + \varphi^* \underline{\hat{U}}^\alpha K \pi^*.$$

Moreover, the semigroup $(\underline{\tilde{P}}_t)$ *is Feller.*

3 Vermiculated spaces

We begin with a generalisation of the iterative state space construction outlined in the introduction.

Let F be a Hausdorff, locally compact, second countable topological space. Suppose that $B_1, B_2, \ldots$ are closed subsets of F. Let $G_1, G_2, \ldots$ be Hausdorff, compact, second countable topological spaces. The space G_n will index the collection of "alternate universes" at stage n of the construction. In the example described in the introduction, $G_1 = G_2 = \{0, 1\}$.

Put $F_0 := F$ and

$$E_1 := F_0 \times G_1,$$
$$\hat{E}_1 := F_0,$$

and

$$A_1 := B_1 \times G_1 \subseteq E_1.$$

Define $\psi_1 : E_1 \to \hat{E}_1$ by

$$\psi_1(y, z) = y.$$

Now apply the general state space construction of Section 2 with the ingredients $E = E_1$, $\hat{E} = \hat{E}_1$, $A = A_1$, and $\psi = \psi_1$. It is clear that the conditions of the construction hold.

Write $\tilde{E}_1$ for the resulting space denoted by $\tilde{E}$ in the general construction, and π_1, φ_1 for the maps denoted by π, φ in the general construction. Thus

$$\tilde{E}_1 = (E_1 \setminus A_1) \cup \psi_1(A_1) = ((F_0 \setminus B_1) \times G_1) \cup B_1,$$

and $\pi_1 : E_1 \to \tilde{E}_1$ is given by

$$\pi_1(y, z) = \begin{cases} (y, z), & \text{if } y \in F_0 \setminus B_1 \ , \\ y, & \text{if } y \in B_1 \ ; \end{cases}$$

and $\tilde{E}_1$ is equipped with the topology induced by π_1. Set $F_1 := \tilde{E}_1$ and write $\varphi_1 : F_1 \to F_0$ for the map denoted by φ in the general construction. Thus

$$\varphi_1(y, z) = y, \quad (y, z) \in E_1 \setminus A_1 = (F_0 \setminus B_1) \times G_1,$$
$$\varphi_1(y) = y, \quad y \in \psi_1(A_1) = B_1.$$

Suppose now that Hausdorff, locally compact, second countable topological spaces F_m, $0 \le m \le n$, and continuous surjections $\varphi_m : F_m \to F_{m-1}$, $1 \le m \le n$, have already been constructed. Define F_{n+1} and a continuous surjection $\varphi_{n+1} : F_{n+1} \to F_n$ as follows.

Put

$$E_{n+1} := F_n \times G_{n+1},$$
$$\hat{E}_{n+1} := F_n,$$

and

$$A_{n+1} := \varphi_{n,0}^{-1}(B_{n+1}) \times G_{n+1} \subseteq E_{n+1},$$

where

$$\varphi_{j,i} := \varphi_{i+1} \circ \varphi_{i+2} \circ \cdots \circ \varphi_j, \quad 0 \le i < j \le n.$$

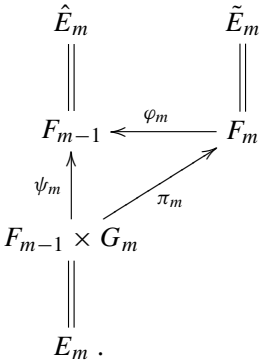
Define $\psi_{n+1} : E_{n+1} \to \hat{E}_{n+1}$ by

$$\psi_{n+1}(y, z) = y.$$

Now apply the general state space construction of Section 2 with the ingredients $E = E_{n+1}$, $\hat{E} = \hat{E}_{n+1}$, $A = A_{n+1}$, and $\psi = \psi_{n+1}$. Once again, it is clear that the conditions of the construction hold. Set $F_{n+1} := \tilde{E}_{n+1}$ and write φ_{n+1} for the map denoted by φ in the general construction; that is,

$$\varphi_{n+1}(y, z) = y, \quad (y, z) \in E_{n+1} \setminus A_{n+1} = (F_n \setminus \varphi_{n,0}^{-1}(B_{n+1})) \times G_{n+1},$$
$$\varphi_{n+1}(y) = y, \quad y \in \psi_{n+1}(A_{n+1}) = \varphi_{n,0}^{-1}(B_{n+1}).$$

The salient points of the construction are summarised as follows:



The sequence of spaces $(F_n)_{n=0}^{\infty}$ equipped with the maps $(\varphi_{j,i})_{0 \le i < j < \infty}$ is a *projective system* of topological spaces (sometimes also called an *inverse system*). Therefore this system has a *projective limit* topological space $F_\infty := \lim_{\leftarrow} F_n$ (also called an *inverse limit*) equipped with a family of continuous surjections $\Phi_n : F_\infty \to F_n$, $0 \le n \le \infty$, satisfying $\varphi_{j,i} \circ \Phi_j = \Phi_i$, $0 \le i < j < \infty$. By general facts about projective limits, the space F_∞ is Hausdorff, locally compact and second countable (see [11, Section 2-14]).

4 Construction of a projective limit process

We continue with the development begun in Section 3.

Let ξ be a quasi-left-continuous Borel right process with state space F. The process ξ is the base process that we will successively lift up to $F_1, F_2, \ldots$ in the manner outlined in the introduction. Write Π^x for the probability measure governing

ξ started at $x \in F$. Let μ_n be a Borel probability measure on G_n, $n \geq 1$. Recall that G_n indexes the various alternate universes at the n^{th} stage of the iterative part of the state space construction. The probability measure μ_n describes how an alternate universe is chosen when the n^{th} stage process hits an anchor point. In the example described in the introduction, μ_1 and μ_2 are both the uniform measure on $\{0, 1\}$.

Assumption 4.1. Write $\mathcal{C}$ for the collection consisting of the empty set and finite unions of sets drawn from $B_1, B_2, \ldots$. Given $C \in \mathcal{C}$, set $\tau := \inf\{t \geq 0 : \xi_t \in C\}$. Assume for each $C \in \mathcal{C}$ and $\alpha > 0$ that $x \mapsto \Pi^x[\int_0^\infty e^{-\alpha t} f(\xi_{t\wedge\tau})\, dt]$ and $x \mapsto \Pi^x[e^{-\alpha\tau} f(\xi_\tau)]$ are in $C_0(F)$ when f is in $C_0(F)$.

Example 4.2. Assumption 4.1 holds for $F = \mathbb{R}$, ξ a standard Brownian motion, and any closed sets $B_1, B_2, \ldots \subseteq \mathbb{R}$.

Set X^0 to be the Markov process ξ on $F_0 = F$. Suppose that quasi-left-continuous Borel right processes $X^m = (\Omega^m, \mathcal{F}^m, \mathcal{F}_t^m, X_t^m, \theta_t^m, \mathbf{P}_m^x)$, $0 \leq m \leq n$, have been defined. For $C \in \mathcal{C}$, write $(U_{C,m}^\alpha)_{\alpha>0}$ for the resolvent of X^m stopped on hitting $\varphi_{m,0}^{-1}(C)$ and $P_{C,m}^\alpha$ for the corresponding α-hitting operator. Suppose further that $U_{C,m}^\alpha C_0(F_m) \subseteq C_0(F_m)$ and $P_{C,m}^\alpha C_0(F_m) \subseteq C_0(F_m)$ for all $C \in \mathcal{C}, \alpha > 0, 0 \leq m \leq n$.

Consider the construction of Section 2 with the ingredients:

- $E = F_n \times G_{n+1}$,
- $\hat{E} = F_n$,
- $A = \varphi_{n,0}^{-1}(B_{n+1}) \times G_{n+1} \subseteq E$,
- $\psi(y, z) = y$, $(y, z) \in F_n \times G_{n+1}$,
- X under $\mathbf{P}^{(y,z)}$ has the law of $\{(X^n, z) : t \geq 0\}$ when X_n is under $\mathbf{P}_n^y$,
- $\hat{X} = X^n$,
- $k(y, \cdot) = \delta_y \otimes \mu_{n+1}$.

It is clear that the conditions of Theorem 2.4 hold. Moreover, the conditions of Corollary 2.6 hold with the set B given by $\varphi_{n,0}^{-1}(C) \times G_{n+1}$ for any $C \in \mathcal{C}$.

Let

$$X_{n+1} = (\Omega^{n+1}, \mathcal{F}^{n+1}, \mathcal{F}_t^{n+1}, X_t^{n+1}, \theta_t^{n+1}, \mathbf{P}_{n+1}^x)$$

be the quasi-left-continuous Borel right processes on $\tilde{E} = F_{n+1}$ produced by Theorem 2.4. Write $(U_{C,n+1}^\alpha)_{\alpha>0}$ for the resolvent of X^{n+1} stopped on hitting $\varphi_{n+1,0}^{-1}(C)$, $C \in \mathcal{C}$, and $P_{C,n+1}^\alpha$ for the corresponding α-hitting operator. Corollary 2.6 guarantees that $U_{C,n+1}^\alpha C_0(F_{n+1}) \subseteq C_0(F_{n+1})$ and $P_{C,n+1}^\alpha C_0(F_{n+1}) \subseteq C_0(F_{n+1})$ for all $\alpha > 0$.

Denote by $(U_n^\alpha)_{\alpha>0}$ the resolvent of X^n. It follows from Theorem 2.4 that

$$U_j^\alpha \varphi_{j,i}^* = \varphi_{j,i}^* U_i^\alpha, \quad 0 \leq i < j < \infty. \tag{4.1}$$

The subspace $\bigcup_n \Phi_n^* C_0(F_n)$ is dense in $C_0(F_\infty)$ by construction. By (4.1), there are well-defined linear operators U_∞^α, $\alpha > 0$, defined on $\bigcup_n \Phi_n^* C_0(F_n)$ such that

$$U_\infty^\alpha \Phi_n^* = \Phi_n^* U_n^\alpha, \quad \text{for all } n. \tag{4.2}$$

It is clear that each U_∞^α extends by continuity to a Markov operator on $C_0(F_\infty)$, and this extension still satisfies the Dynkin intertwining relation (4.2). By (4.2) and the resolvent equation for $(U_n^\alpha)_{\alpha>0}$,

$$U_\infty^\alpha \Phi_n^* = \Phi_n^* U_n^\alpha = \Phi_n^* \left(U_n^\beta + (\beta-\alpha) U_n^\alpha U_n^\beta\right) = \left(U_\infty^\beta + (\beta-\alpha) U_\infty^\alpha U_\infty^\beta\right) \Phi_n^*,$$

and hence, by continuity, $(U_\infty^\alpha)_{\alpha>0}$ obeys the resolvent equation on $C_0(F_\infty)$. Again by continuity,

$$\lim_{\alpha\to\infty} U_\infty^\alpha f = f \text{ pointwise}, \ f \in C_0(F_\infty). \tag{4.3}$$

Standard arguments (see [15, Theorem 9.26]) and (4.2) now give the following.

Theorem 4.3. *Under Assumption* 4.1 *there is a quasi-left-continuous Borel right process*

$$X^\infty = (\Omega^\infty, \mathcal{F}^\infty, \mathcal{F}_t^\infty, X_t^\infty, \theta_t^\infty, \mathbf{P}_\infty^x)$$

on F_∞ with resolvent $(U_\infty^\alpha)_{\alpha>0}$. The law of $\Phi_n \circ X^\infty$ under $\mathbf{P}_\infty^x$ is that of X^n under $\mathbf{P}_n^{\Phi_n(x)}$.

Remark 4.4. It is worth pointing out what can go wrong if Assumption 4.1 does not hold. For example, consider the first stage of the inductive part of the construction with $F = \mathbb{R}^2$, $B_1 = \{0\}$, $G_1 = \{0, 1\}$, $\mu(\{0\}) = \mu(\{1\}) = \frac{1}{2}$, and ξ a standard planar Brownian motion. If our construction "worked" in this instance it would produce a process X^1 that, when started at the anchor point 0, picks at random between the two copies of $\mathbb{R}^2$ pinned at 0 and then never leaves this copy. The identity of the chosen copy of $\mathbb{R}^2$ is thus a non-trivial random variable measurable with respect to the germ σ-field of X^1 at time 0, and so even the Blumenthal zero–one law fails for X^1.

Acknowledgement. This research was initiated while the authors were visiting the Erwin Schrödinger International Institute for Mathematical Physics in 2001 as part of the program on Random Walks. Martin Barlow was supported by NSERC (Canada) and CNRS (France). Steven Evans was supported by NSF grant DMS-0071468 (U.S.A.) and a Miller Institute for Basic Research in Science Research Professorship.

References

[1] M. T. Barlow, Which values of the volume growth and escape time exponent are possible for a graph?, preprint (2001).

[2] M. T. Barlow, M. Émery, F. B. Knight, S. Song and M. Yor, Autour d'un théorème de Tsirelson sur les filtrations Browniennes et non Browniennes, in: Séminaire de Probabilites, XXXII, Lecture Notes in Math. 1686, Springer-Verlag, Berlin 1998, 264–305.

[3] M. T. Barlow, J. Pitman, and M. Yor, On Walsh's Brownian motions, in: Séminaire de Probabilites, XXIII, Lecture Notes in Math. 1372, Springer-Verlag, Berlin 1989, 275–293.

[4] D. S. Dean and K. M. Jansons, Brownian excursions on combs, J. Statist. Phys. 70 (1993), 1313–1332.

[5] S. N. Evans, Snakes and spiders: Brownian motion on $\mathbb{R}$-trees, Probab. Theory Related Fields 117 (2000), 361–386.

[6] S. N. Evans and R. B. Sowers, Pinching and twisting Markov processes, Ann. Probab. 31 (2003), 486–527.

[7] M. I. Freidlin and A. D. Wentzell, Diffusion processes on graphs and the averaging principle, Ann. Probab. 21 (1993), 2215–2245.

[8] M. I. Freidlin and A. D. Wentzell, Random Perturbations of Hamiltonian Systems, Mem. Amer. Math. Soc. 109 (1994), no. 523.

[9] A. A. Grigor′yan, The heat equation on noncompact Riemannian manifolds, Math. USSR-Sb. 72 (1992), 47–77.

[10] J. Heinonen and S. Semmes, Thirty-three yes or no questions about mappings, measures, and metrics, Conform. Geom. Dyn. 1 (1997), 1–12 (electronic).

[11] J. G. Hocking and G. S. Young, Topology, Addison–Wesley, Reading, MA, 1961.

[12] W. B. Krebs, Brownian motion on the continuum tree, Probab. Theory Related Fields 101 (1995), 421–433.

[13] T. J. Laakso, Ahlfors Q-regular spaces with arbitrary $Q > 1$ admitting weak Poincaré inequality, Geom. Funct. Anal. 10 (2000), 111–123.

[14] L. Saloff-Coste, A note on Poincaré, Sobolev, and Harnack inequalities, Internat. Math. Res. Notices (2) (1992), 27–38.

[15] M. Sharpe, General Theory of Markov Processes, Academic Press, San Diego 1988.

[16] B. Tsirelson, Triple points: from non-Brownian filtrations to harmonic measures, Geom. Funct. Anal. 7 (1997), 1096–1142.

[17] N. Th. Varopoulos, Long range estimates for Markov chains, Bull. Sci. Math. (2) 109 (1985), 225–252.

[18] J. B. Walsh, A diffusion with discontinuous local time, in: Temps Locaux, Astérisque 52–53, Société Mathématique de France, Paris 1978, 37–45.

Martin T. Barlow, Department of Mathematics, University of British Columbia, 1984 Mathematics Rd, Vancouver BC V6T 1Z2, Canada
E-mail: barlow@math.ubc.ca

Steven N. Evans, Department of Statistics #3860, University of California at Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, USA
E-mail: evans@stat.berkeley.edu

Cactus trees and lower bounds on the spectral radius of vertex-transitive graphs

Laurent Bartholdi[*]

Abstract. This paper gives lower bounds on the spectral radius of vertex-transitive graphs, based on the number of "prime cycles" at a vertex. The bounds are obtained by constructing circuits in the graph that resemble "cactus trees", and enumerating them. Counting these circuits gives a coefficient-wise underestimation of the Green function of the graph, and hence an underestimation of its spectral radius.

The bounds obtained are very good for the Cayley graph of surface groups of genus $g \geq 2$ with standard generators (these graphs are the 1-skeletons of tessellations of hyperbolic plane by $4g$-gons, $4g$ per vertex). We have for example for $g = 2$

$$0.662418 \leq \|M\| \leq 0.662816,$$

and for $g = 3$

$$0.552773 \leq \|M\| \leq 0.552792.$$

1 Introduction: Groups

Throughout this paper, Γ will be a group generated by a finite, symmetric set S, of cardinality $\#S = d$. "Symmetric" means that $S = S^{-1}$. Many of the objects we define will depend heavily on the choice of S, even though we will not make it explicit in the notation.

The *Cayley graph* $\mathfrak{G}$ of Γ is the graph with vertex set Γ, and vertices connected under the right action of S; i.e., γ and γs are joined for all $\gamma \in \Gamma$ and $s \in S$. The *Markov operator* $M : \ell^2(\Gamma) \to \ell^2(\Gamma)$ is defined by

$$(Mf)(\gamma) = \frac{1}{d} \sum_{s \in S} f(\gamma s).$$

It is used to study the simple random walk on Γ; for instance, the probability of return in n steps is $p_n = \langle M^n \delta_1 | \delta_1 \rangle$, where δ_1 is the Dirac function at $1 \in \Gamma$.

The first step in understanding M is the computation of its (operator) norm $\|M\|$, also called the *spectral radius* of $\mathfrak{G}$; indeed, the probabilities p_n satisfy the relation

[*]The author acknowledges support from *Fonds national de la recherche scientifique*.

$\limsup_{n\to\infty}\sqrt[n]{p_n} = \|M\|$. Harry Kesten showed in [9, 10] the group-theoretical importance of $\|M\|$: we always have

$$\frac{2\sqrt{d-1}}{d} \le \|M\| \le 1, \tag{1.1}$$

with equality on the left if and only if $\mathfrak{G}$ is a tree, i.e., Γ is a free product of $\mathbb{Z}/2$'s and $\mathbb{Z}$'s, with S consisting of the standard generators and their inverses; and equality holds on the right if and only if Γ is amenable.

Assume now that Γ is not free; say it has a relation of length $k \ge 3$. William Paschke obtained in [15] the estimate

$$\|M_\Gamma\| \ge \min_{s>0}\left\{2\cosh(s) + (d-2)Q\left(\frac{\cosh(ks)+1}{\sinh(s)\sinh(ks)}\right)\right\}, \tag{1.2}$$

where $Q(t) = (\sqrt{t^2+1}-1)/t$.

The purpose of this paper is to show that the lower bound (1.2) on $\|M\|$ can be improved if additional hypotheses are made on the number of relations of Γ, and on the number of distinct cyclic permutations of these relations.

Definition 1.1. The *Green function* of Γ is the formal power series

$$G(t) = \sum_{w\in S^*:\, w\equiv_\Gamma 1} t^{|w|} = \sum_{n\ge 0} p_n d^n t^n,$$

i.e., the growth series of the words representing 1 in Γ.

For two power series $G(t) = \sum_{n\ge0} g_n t^n$ and $H(t) = \sum_{n\ge0} h_n t^n$, define $G \precsim H$ to mean $g_n \le h_n$ for all $n \ge 0$.

A *prime relator* is a word $w \in S^*$ such that $w \equiv_\Gamma 1$, and such that $v \not\equiv_\Gamma 1$ for all proper subwords v of w.

A set $\mathcal{R}$ of prime relators satisfies the *small cancellation condition* $\mathcal{O}(\eta)$ if for any $w, w' \in \mathcal{R}$, and any factorization $w = uv$ and $w' = vu'$ we have either $u = u'$ or $|v| \le \eta\cdot\min\{|w|, |w'|\}$.

Note that a prime relator is necessarily a cyclically freely reduced word. The symbol "$\mathcal{O}$" stands for "overlap". It is a notion close, but strictly weaker than the $C(\eta)$ in small cancellation theory.

The main result of this paper, stated for finitely generated groups, is the following. See in Subsection 2.1 the more general form stated for vertex-transitive graphs:

Corollary 1.2. *Let Γ be a group generated by a finite symmetric set S of cardinality d, and let M be its Markov operator. Assume that Γ has a set $\mathcal{R}$ of prime relators satisfying a small cancellation condition $\mathcal{O}(\eta)$.*

Let $f(t) = \sum_{w\in\mathcal{R}} t^{|w|}$ be the growth series of $\mathcal{R}$, and let ζ satisfy

$$\zeta - 1 + (1-1/d)\zeta f(\zeta^{\eta-1}/(d-1)) = 0.$$

Construct the power series

$$h(t) = \frac{2(d-1)}{d-2+d\sqrt{1-4(d-1)t^2}}, \qquad (*)$$

$$H(t,u) = \frac{1-(1-u)^2t^2}{1+(1-u)(d-1+u)t^2}\, h\left(\frac{t}{1+(1-u)(d-1+u)t^2}\right),$$

$$g_1(t,u) = H(\zeta t, u),$$

$$g_2(t) = g_1\left(t\frac{d-f(t)}{d-(d-1)f(t)}, \frac{(d-2)f(t)}{d-f(t)}\right),$$

$$g_3(t) = h(t)g_2\left(\frac{1-\sqrt{1-4(d-1)t^2}}{2(d-1)t}\right).$$

Then the Green function of $\mathfrak{G}$ *satisfies* $G \gtrsim g_3$.

Let ρ *be the radius of convergence of* g_3. *Then* $\|M\| \geq 1/(d\rho)$.

Note that Corollary 1.2 does not supersede Paschke's result (1.2), in that the bound it gives for the group $(\mathbb{Z}/k) * (\mathbb{Z}/2) * \cdots * (\mathbb{Z}/2)$ is inferior to Paschke's. It does, however, give a superior bound for many groups, and in particular surface groups.

1.1 Surface groups

Consider the fundamental group of a surface of genus $g \geq 2$

$$\Gamma_g = \langle a_1, b_1, \ldots, a_g, b_g \,|\, [a_1,b_1]\ldots[a_g,b_g] = 1\rangle.$$

This group is non-elementary hyperbolic, hence non-amenable; its spectral radius is $O(g^{-1/2})$ for large g. Simple lower and upper bounds come respectively from Γ_g being a quotient of a free group of rank $2g$, and containing by Magnus' Freiheitssatz a $(2g-1)$-generated free subgroup $\langle a_1, b_1, \ldots, a_{g-1}, b_{g-1}, a_g\rangle$:

$$\frac{\sqrt{4g-1}}{2g} \leq \|M_g\| \leq \frac{\sqrt{4g-3}+2}{2g};$$

For more details, see [2]. Many improvements on these bounds were obtained, see [15, 5, 2, 17, 13, 3]. Currently, the best known bounds are

Theorem 1.3. *The spectral radius of the surface groups of genus* 2 *and* 3 *satisfy*

$$0.662418 \leq \|M_2\| \leq 0.662816,$$
$$0.552773 \leq \|M_3\| \leq 0.552792.$$

These upper bounds are due to Tatiana Nagnibeda [13, 14].

1.2 Reduction to graphs

Equations (1.1) and (1.2) hold more generally for any vertex-transitive graph, i.e., for any graph whose group of automorphisms acts transitively on vertices. Remember from [15] that there are vertex-transitive graphs with no simply vertex-transitive automorphism group – for instance, the 1-skeleton of a dodecahedron. Such a graph cannot be the Cayley graph of a group.

In the context of d-regular vertex-transitive graphs, $p_n d^n$ is the number of closed paths of length n at any fixed vertex in $\mathfrak{G}$, and $d\|M\|$ is the asymptotic exponential growth rate of these numbers of closed paths.

In just the same way as any group is a quotient of a free group, any d-regular graph is covered by the d-regular tree. Since closed paths in the tree remain closed in the quotient graph, the spectral radius of any d-regular graph is bounded from below by the spectral radius of the d-regular tree; this proves the left inequality of (1.1).

Similarly, any vertex-transitive, d-regular graph with a loop of length k at each vertex is covered by the graph $P_{k,d}$ obtained from a $k(d-2)$-regular tree by replacing each vertex by a k-gon and equidistributing the edges on the k-gon's vertices [15, Proposition 2.4]. The spectral radius of $P_{k,d}$ can be computed using

Theorem 1.4 ([16]; [4]; [1, Theorem 9.2]). *Let $\mathcal{X}_1$, $\mathcal{X}_2$ be vertex-transitive graphs, and let $\mathcal{X}$ be their free product. Let $G_1(t)$, $G_2(t)$ and $G(t)$ be their corresponding Green functions. Then*

$$\frac{1}{(tG(t))^{-1}} = \frac{1}{(tG_1(t))^{-1}} + \frac{1}{(tG_2(t))^{-1}} - \frac{1}{t},$$

where $F^{-1}(t)$ denotes the formal inverse, i.e., the series $E(t)$ such that $E(F(t)) = F(E(t)) = t$.

Indeed, if we take $\mathcal{X}_1$ a $(d-2)$-regular tree and $\mathcal{X}_d$ a k-cycle, then $\mathcal{X}$ is $P_{k,d}$.

Clearly, if $F(t)$ is an algebraic series, then $F^{-1}(t)$ is algebraic of same degree, because $P(t, F(t)) = 0$ implies $P(F^{-1}(t), t) = 0$; also, $\deg(F_1(t) + F_2(t)) \leq \deg F_1(t) \deg F_2(t)$. Simple computations show that for $\mathcal{X}_1$ and $\mathcal{X}_2$ as above G_1 is algebraic of degree 2 and G_2 is algebraic of degree $(k+1)/2$. It follows that $\|M_P\|$ is an algebraic number of degree at most $k+1$.

Another computation shows that $\|M_P\|$ is the right-hand side of (1.2); the inequality then follows.

Note that Paschke's estimate is valid for all vertex-transitive graphs; the inequality is obtained by constructing a cover for all graphs containing a k-cycle. We note *a posteriori* that $P_{k,d}$ is the Cayley graph of $(\mathbb{Z}/k) * (\mathbb{Z}/2) * \cdots * (\mathbb{Z}/2)$, with $d-2$ copies of $\mathbb{Z}/2$, but there is no *a priori* reason for the graph of smallest norm, $P_{k,d}$, to be the Cayley graph of a group. Our method can also be understood as constructing a transitive graph of minimal norm satisfying some conditions on its cycles; however, this graph of minimal norm will not be the Cayley graph of a group.

2 Preliminaries: graphs

An (oriented) *graph* $\mathfrak{G}$ is a pair (V, E) of sets called respectively *vertices* and *edges*, with maps $\alpha, \omega : E \to V$ called respectively *start* and *end*, and an involution $e \mapsto \overline{e}$ of E such that $\alpha(\overline{e}) = \omega(e)$. By

$$\deg(v) = \#\{e \in E : \alpha(e) = v\}$$

we denote the *degree* of a vertex v. The graph is *d-regular* if each vertex has degree d.

A *path* is a sequence $p = (p_1, \ldots, p_n)$ of edges, with $\omega(p_i) = \alpha(p_{i+1})$ for all $i \in \{1, \ldots, n-1\}$. Its *length* is $|p| = n$, and its *start* and *end* are $\alpha(p) = \alpha(p_1)$ and $\omega(p) = \omega(p_n)$. It is a *circuit* if $\alpha(p) = \omega(p)$. Paths are multiplied by concatenation; therefore in this definition a graph is nothing but a small $*$-category with object set V and arrow set E.

A *spike* in a path p is an index $i \in \{1, \ldots, |p| - 1\}$ such that $p_i = \overline{p_{i+1}}$. A path is *reduced* if it has no spikes. The *spike count* of a path p is the number $\mathrm{sc}(p)$ of spikes in p. A circuit is *prime* if it is reduced, non-trivial, and $\alpha(p_i) \neq \alpha(p_j)$ for all $i \neq j$.

The *Markov operator* on $\mathfrak{G}$ is the operator $M : \ell^2(V, \deg) \to \ell^2(V, \deg)$ given by

$$(Mf)(v) = \frac{1}{\deg(v)} \sum_{\{e \in E : \alpha(e) = v\}} f(\omega(e)).$$

It governs the behaviour of the simple random walk on $\mathfrak{G}$: given two vertices $v, w \in V$ the probability of going from v to w in n steps is $\langle M^n \delta_w | \delta_v \rangle$, where δ_x is the Dirac function at $x \in V$.

The graph $\mathfrak{G}$ is *transitive* if for any two vertices $v, w \in V$ there is a graph automorphism mapping v to w. Transitivity implies that the probability of return $p_n = \langle M^n \delta_v | \delta_v \rangle$ does not depend on v; its exponential rate of decay is $\|M\|$, the spectral radius of the random walk.

If Γ is a group with symmetric generating set S, its *Cayley graph* $\mathfrak{G}$ has vertex set Γ and edge set $\Gamma \times S$, with $\alpha(\gamma, s) = \gamma$ and $\omega(\gamma, s) = \gamma s$ and $\overline{(\gamma, s)} = (\gamma s, s^{-1})$. Clearly, $\mathfrak{G}$ is a transitive graph under the left action of Γ.

Let $\mathfrak{G}$ be a transitive graph, and let $\mathcal{P}$ be a set of prime circuits in $\mathfrak{G}$ at a given vertex v. By transitivity, there exists a translate $\mathcal{P}_w$ of $\mathcal{P}$ at any vertex w. A *spread* of $\mathcal{P}$ is the union of the $\mathcal{P}_w$ for all $w \in V$.

The set $\mathcal{P}$ satisfies the *small cancellation condition* $\mathcal{O}(\eta)$ if it has a spread $\widehat{\mathcal{P}}$ such that for any $p, p' \in \widehat{\mathcal{P}}$, and any factorization $p = qr$ and $p' = rq'$ we have either $q = q'$ or $|r| \leq \eta \cdot \min\{|q|, |q'|\}$.

The *growth series* of a set $\mathcal{P}$ of paths is the formal power series $f(t) = \sum_{p \in \mathcal{P}} t^{|p|}$. The *Green function* of $\mathfrak{G}$ is the growth series of the set of circuits at an irrelevant but fixed vertex.

2.1 Main result

The definitions given above were tailor-cut to make Corollary 1.2 a direct consequence of the following result on graphs:

Theorem 2.1. *Let $\mathfrak{G}$ be a vertex-transitive, d-regular graph, and let $\mathcal{P}$ be a set of prime circuits at a fixed vertex in $\mathfrak{G}$ satisfying $\mathcal{O}(\eta)$. Let $f(t)$ be the growth series of $\mathcal{P}$. Construct the power series $g_1(t), \dots, g_3(t)$ as in* (∗)*, Corollary* 1.2*. Then the Green function of $\mathfrak{G}$ satisfies $G \succsim g_3$.*

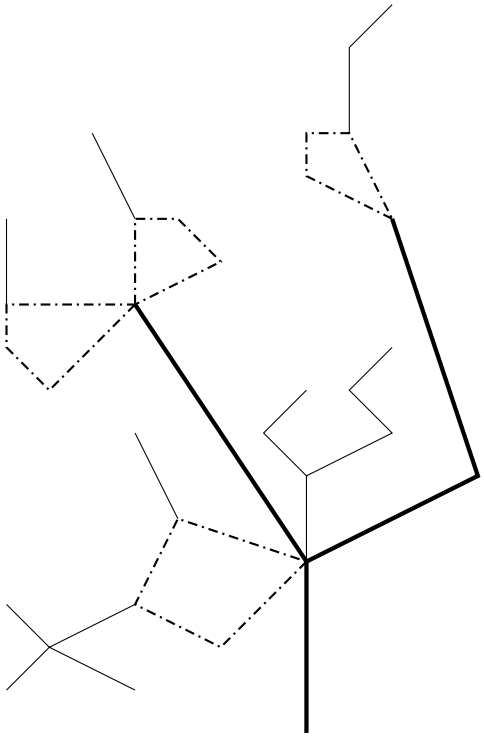
Let ρ be the radius of convergence of g_3. Then $\|M\| \geq 1/(d\rho)$.

The main idea behind Theorem 2.1 is the construction of *cactus trees* in $\mathfrak{G}$. A *cactus tree* is a circuit in $\mathfrak{G}$ that is built in three stages: at the first stage, a "trunk" is constructed, i.e. a circuit that freely reduces to the trivial circuit. This trunk may not contain any subword of length $(1-\eta)|p|$ of any prime circuit p.

At the second stage, "fruits", i.e. prime circuits, are inserted at all vertices of the circuit, in such a way that the resulting circuit is reduced.

At the third stage, "spikes", i.e. circuits that freely reduce to the trivial circuit, are inserted at all vertices of the cactus tree.

Here is an example of cactus, with the three stages of construction indicated in solid bold, dashed medium and solid thin lines:



Consider the graph $X_{d,m}$ that is the 1-skeleton of the tessellation of the hyperbolic plane by m-gons, d per vertex. Numerically, the growth of cactus trees is a very good underestimation of the growth of circuits in the graph $X_{d,m}$. The Cayley graph of the surface group Γ_g is $X_{4g,4g}$. This explains the bounds given in Subsection 1.1.

3 A formula: cogrowth

This section recalls the main result of [1]. Fix a vertex $v \in V$, and consider a set $\mathcal{C}$ of reduced circuits at v. The *saturation* $\langle\mathcal{C}\rangle$ of $\mathcal{C}$ is the closure of $\mathcal{C}$ under the iterated operation of inserting spikes in paths; i.e., $\langle\mathcal{C}\rangle$ is the smallest set of circuits containing $\mathcal{C}$ and such that for any product of paths $pq \in \langle\mathcal{C}\rangle$ and for any $e \in E$ with $\alpha(e) = \omega(p) = \alpha(q)$ we have $pe\bar{e}q \in \langle\mathcal{C}\rangle$. (We allow p, q to be empty, in which case by convention $\omega(p) = \alpha(q) = v$).

Let $\mathcal{P}$ be a set of paths. Its *spiky growth series* is the formal 2-variable power series

$$G_{\mathcal{P}}(t,u) = \sum_{p\in\mathcal{P}} t^{|p|}u^{\mathrm{sc}(p)}.$$

The specialization $G_{\mathcal{P}}(t,1)$ "forgets" the number of spikes in the paths while "remembering" only their lengths, and in the case when $\mathcal{P}$ consists of all the circuits at v is often called the *Green function* of $\mathfrak{G}$ at v. The specialization $G_{\mathcal{P}}(t,0)$ counts only reduced paths in $\mathcal{P}$.

Theorem 3.1 ([1], Corollary 2.6). *Let $\mathcal{C}$ be an arbitrary set of reduced paths in a d-regular graph $\mathfrak{G}$, and let $\langle\mathcal{C}\rangle$ be its saturation. Then*

$$\frac{G_{\langle\mathcal{C}\rangle}(t,u)}{1-(1-u)^2t^2} = \frac{G_{\langle\mathcal{C}\rangle}\left(\frac{t}{1+(1-u)(d-1+u)t^2},1\right)}{1+(1-u)(d-1+u)t^2}.$$

In particular, we have

$$\frac{G_{\mathcal{C}}(t,0)}{1-t^2} = \frac{G_{\langle\mathcal{C}\rangle}\left(\frac{t}{1+(d-1)t^2},1\right)}{1+(d-1)t^2}.$$

Assume that $\mathfrak{G}$ is transitive and d-regular, and recall that p_n denotes the probability of return in n steps of the simple random walk on $\mathfrak{G}$. Then at each vertex there are $d^n p_n$ circuits of length n, and we have the following connection between spectral radius and counting of paths:

Proposition 3.2. *Let $\mathcal{C}$ be a set of circuits at $v \in V$ in a d-regular graph $\mathfrak{G}$, and let ρ be the convergence radius of $G_{\langle\mathcal{C}\rangle}(t,1)$. Then $\|M\| \geq 1/(d\rho)$.*

3.1 Forbidden words

Let F be a set of words over an alphabet A, and consider the problem of estimating the number of words over A not containing any element of F as a subword. We have the following lemma, which is in essence the Lovasz local lemma:

Lemma 3.3. *Let W be either A^* or the set of reduced words $A^* \setminus \{A^* a a^{-1} A^*\}_{a\in A}$, if A has an involution $a \leftrightarrow a^{-1}$.*

Let $F \subset W$ be a set of "forbidden" words, and set $L = W \setminus WFW$ be the language of words in W not containing an element of F as a subword. Let Φ denote the growth series of F, and let λ, ρ denote the growth rate of L, W respectively; ρ is either $\#A$ or $\#A - 1$.

Then we have $\lambda \geq \rho\zeta$, where ζ satisfies the equation $\zeta - 1 + \frac{\rho\zeta}{\#A}\Phi(\rho^{-1}\zeta^{-1}) = 0$.

Proof. Take the uniform distribution on the set $S_n = W \cap A^n$, and define events Q_j, R_j on S_n as follows: Q_j is the set of words $w \in S_n$ for which there are no factorizations $w = efg$ with $|ef| = j$ and $f \in F$; in other words, Q_j is the event of not containing a forbidden word ending at index j. Set $R_j = Q_1 \cap Q_2 \cap \cdots \cap Q_{j-1}$. Note that $\mathbb{P}(R_{n+1}) = \mathbb{P}(Q_n|R_n)\mathbb{P}(R_n)$.

We claim that $\mathbb{P}(Q_j|R_j) \geq \zeta$ for all $j \in \{1, \dots, n\}$, and proceed by induction, the basis being $\mathbb{P}(Q_0|R_0) = 1 \geq \zeta$ by definition of ζ:

$$\begin{aligned}
\mathbb{P}(Q_j|R_j) &\geq 1 - \sum_{f\in F} \frac{\mathbb{P}(\text{“occurrence of } f \text{ ending at index } j\text{”} \cap R_j)}{\mathbb{P}(R_j)} \\
&\geq 1 - \sum_{f\in F} \frac{\mathbb{P}(\text{“occurence of } f \text{ ending at index } j\text{”} \cap R_{j-|f|+1})}{\mathbb{P}(Q_{j-1}|R_{j-1}) \dots \mathbb{P}(Q_{j-|f|+1}|R_{j-|f|+1})\mathbb{P}(R_{j-|f|+1})} \\
&\geq 1 - \sum_{f\in F} \frac{\#A^{-1}\rho^{1-|f|} \quad \mathbb{P}(R_{j-|f|+1})}{\zeta^{|f|-1}\mathbb{P}(R_{j-|f|+1})} \\
&= 1 - \rho\zeta\Phi(\rho^{-1}\zeta^{-1})/\#A = \zeta.
\end{aligned}$$

(In the third equality, we use $\mathbb{P}(\text{“occurence of } f \text{ ending at index } j\text{”}) = \#A^{-1}\rho^{1-|f|}$ and independence with $R_{j-|f|+1}$.)

Therefore, $\#(L \cap S_n) = \rho^n \mathbb{P}(R_{n+1}) \geq (\rho\zeta)^n$. □

As an application, consider a d-regular transitive graph $\mathfrak{G}$, and a spread $\mathcal{P}$ of "forbidden paths" in $\mathfrak{G}$. Fix a vertex $*$ in $\mathfrak{G}$, and let T denote the set of circuits at $*$. By Theorem 3.1, the spiky growth series of T is

$$\Theta(t, 1-u) = \frac{2(d-1)(1-u^2t^2)}{(d-2)(1+u(d-u)t^2) + d\sqrt{(1+u(d-u)t^2)^2 - 4(d-1)t^2}}.$$

Lemma 3.4. *If F is a set of freely reduced words, then with the notation of Lemma 3.3 the spiky growth series of $T \cap L$ is coefficient-wise at least $\Theta(\zeta t, u)$.*

Proof. For $n \in \mathbb{N}$ consider $T \cap A^n$; then ζ^n is a lower bound of the probability, for a word of length n chosen with uniform probability, to belong to L. For any $i < j \in \{1, \dots, n\}$, if there exists $w \in T \cap A^n$ whose subword $w_i \dots w_j$ is reduced, then the restriction $T \cap A^n \to A^{j-i+1}$ to indices $i \dots j$ yields a uniform reduced word of length $j - i + 1$. If follows that

$$\zeta^n \leq \mathbb{P}(w \in F | w \in A^n) \leq \mathbb{P}(w \in F | w \in T \cap A^n). \qquad \square$$

Note that in the case when a more detailed description of F is available, strengthenings of the above result are possible. With surface groups in mind, we consider the concrete case $A = \{a_1^{\pm 1}, b_1^{\pm 1}, \dots, a_g^{\pm 1}, b_g^{\pm 1}\}$; W is either A^* or the set F_A of reduced words over A; and $r = [a_1, b_1] \dots [a_g, b_g]$. We take for F the set of subwords of the cyclic permutations of $r^{\pm 1}$ of length $4g - 1$. As above we write ρ for the growth rate of W. Let $\Lambda(t)$ denote the growth series of L. Then

$$\Lambda \succsim \frac{1}{1 - \rho t + 8gt^{4g-1} \frac{1-t}{1-t^{4g-1}}}. \tag{3.3}$$

Indeed, Lemma 3.3 applies, and can be slightly improved by letting F_i denote the set of subwords of length i of $rrr \dots$ and $r^{-1}r^{-1}r^{-1} \dots$, for all $r \in F$, and writing Φ_i for the generating series of F_i. The equation defining Λ is then

$$\Lambda = \frac{1}{1 - \rho t}\big(1 - \big(\Phi_{4g-1} - \Phi_{4g} + \Phi_{8g-2} - \Phi_{8g-1} + - \cdots\big)\Lambda\big);$$

indeed, a word in W is either in L, or is of the form fgh, with $h \in L$ and $g \in F_i$ for i maximal. The result follows by inclusion-exclusion.

4 Main result: proof of Theorem 2.1

Assume throughout this section that a vertex $v \in V$ has been fixed in the transitive, d-regular graph $\mathfrak{G}$. By Proposition 3.2, a lower bound on $\|M\|$ can be obtained by evaluating $G_{\mathcal{C}}(t, 1)$ for some set $\mathcal{C}$ of circuits at v.

Let $\mathcal{P}$ be a set of prime circuits at v satisfying the small cancellation condition $\mathcal{O}(\eta)$, and let $f = G_{\mathcal{P}}(t, 1)$ be its growth series. Let $\widehat{\mathcal{P}}$ be an arbitrary but fixed spread of $\mathcal{P}$, i.e., a choice of a translate of every circuit in $\mathcal{P}$ at every vertex of $\mathfrak{G}$.

Start with the set $\mathcal{C}_0 = \{\cdot\}$ consisting only of the empty circuit; its growth series is $G_{\mathcal{C}_0} = 1$. Let $\widetilde{\mathcal{C}_1}$ be the saturation $\langle \mathcal{C}_0 \rangle$ of $\mathcal{C}_0$, and let $H(t, u)$ be its spiky growth series, obtained via Theorem 3.1.

Let $\mathcal{C}_1$ be the set of circuits $c \in \widetilde{\mathcal{C}_1}$ such that, for all $p \in \widehat{\mathcal{P}}$, the circuit c does not contain any prefix p' of p with $|p'| \geq (1 - \eta)|p|$.

Since $f(t)$ counts the prime circuits p, we get $f(t^{1-\eta})$ as the growth of "forbidden" prefixes p'. Solve $\zeta - 1 + (1 - 1/d)\zeta f(\zeta^{\eta-1}/(d-1)) = 0$ for ζ; then by Lemma 3.4

the spiky growth series of $\mathcal{C}_1$ is bounded from below as

$$G_{\mathcal{C}_1}(t,u) \succsim H(\zeta t, u).$$

Consider next $\mathcal{C}_2$, the circuits obtained from $\mathcal{C}_1$ by inserting at each vertex a non-negative number of prime circuits such that the resulting circuit is reduced. Insertion of 0 prime circuits can be done in 0 or 1 ways, depending on whether the vertex is a spike or not, and insertion of $i \geq 1$ circuits can be done in at least $(1-\frac{2}{d})(1-\frac{1}{d})^{i-1} f(t)^i$ ways, counting the insert's length. Indeed, to guarantee that the resulting circuit is reduced, it suffices to forbid one out of d starting edges for all prime circuits inserted, except for the last one, for which a starting and an ending edge must be forbidden.

Summing over i gives the generating functions

$$\frac{(d-2)f(t)}{d-(d-1)f(t)}, \qquad \frac{d-f(t)}{d-(d-1)f(t)}$$

counting possible insertions at a spike and non-spike vertex respectively. The growth series of $\mathcal{C}_2$ is therefore minorized as

$$G_{\mathcal{C}_2}(t,1) \succsim G_{\mathcal{C}_1}\left(t\frac{d-f(t)}{d-(d-1)f(t)}, \frac{(d-2)f(t)}{d-f(t)}\right).$$

Finally let $\mathcal{C}_3 = \langle \mathcal{C}_2 \rangle$; a last application of Theorem 3.1 gives a lower bound for $G_{\mathcal{C}_3}(t,1)$, which in turn, using Proposition 3.2, gives a lower bound on $\|M\|$.

Proof of Theorem 2.1. The proof relies on the construction of cactus trees described in this section. In a cactus tree, i.e., a path constructed as above, mark the edges with the alphabet $\{1,2,3\}$ according to the stage at which that edge appeared in the cactus tree.

The arguments given above show that the growth series $G_{\mathcal{C}_3}(t,1)$ undercounts marked cactus trees; indeed, different choices of initial tree (in $\mathcal{C}_1$), prime circuits (in $\mathcal{C}_2$) or final spikes (in $\mathcal{C}_3$) yield different marked cactus trees. To show that $G_{\mathcal{C}_3}(t,1)$ undercounts circuits, it therefore suffices to show that the markings on cactus trees are uniquely determined; i.e., that two distinct marked cactus trees remain distinct after the marks are erased.

Consider a cactus tree. After removal of all spikes, it gives rise to a unique reduced path; in other words, the order in which the spikes are removed does not change the resulting reduced path.

Now, in this reduced circuit, locate the first subword in $\widehat{\mathcal{P}}$, and pluck it; and repeat till no such subword can be removed. These subwords are necessarily the prime circuits that were inserted in constructing $\mathcal{C}_2$; the only other possibility would be that some circuit $p = p'p'' \in \mathcal{C}_1$ is such that $p'qp''$ contains a prime circuit r at a position before q. This r would then be either a subword of p', which is forbidden by our construction of $\mathcal{C}_1$, or a subword of $p'q$ containing a part of q; by the small cancellation condition, a large part of r would subsist in p', and this is also forbidden by our construction. □

5 Computations: surface groups

Here we make explicit the arguments in the previous section. Even though our main motivation is to obtain lower bounds for the spectral radius of surface groups, all computations are performed on $X_{d,m}$, the 1-skeleton of a tessellation of the hyperbolic plane by m-gons, d per vertex, introduced above.

Fix a base vertex v of $X_{d,m}$. Consider as prime circuits the $2d$ distinct m-gons touching v, in both orientations. Then $\widehat{\mathcal{P}}$ is the set of m-gons of $X_{d,m}$, taken with all possible base points and orientations.

This choice amounts to taking $f(t) = 2dt^m$. Since two m-gons touch in at most one edge, we have a $\mathcal{O}(1/m)$ small cancellation condition, i.e., we take $\eta = 1/m$.

By Equation (3.3), we find ζ by solving $\zeta - 1 + 2(d\zeta - 1)/(d^m\zeta^m - 1) = 0$ for ζ's largest positive root. Numerically, this root is close to 1, and slightly smaller; i.e., $\zeta = 1 - \mathcal{O}(e^{-d})$. For $d = 8$ and $d = 12$, we obtain respectively

$$\zeta(8) = 1 - 0.63 \cdot 10^{-5}, \quad \zeta(12) = 1 - 0.29 \cdot 10^{-10}.$$

We now follow the steps of Theorem 2.1 for $d = m = 8$, corresponding to the surface group of genus 2. We obtain

$$\begin{aligned}
h(t) &= \frac{2 \cdot 7}{6 + 8\sqrt{1 - 4 \cdot 7t^2}}, \\
H(t,u) &= \frac{1 - (1-u)^2t^2}{1 + (1-u)(7+u)t^2} h\left(\frac{t}{1 + (1-u)(7+u)t^2}\right), \\
g_1(t,u) &= H(t\zeta(8), u), \\
g_2(t) &= g_1\left(t\frac{d - f(t)}{d - (d-1)f(t)}, \frac{(d-2)f(t)}{d - f(t)}\right) \\
&\cong \frac{14(1-t^2)(1-14t^8)}{(1 + 7t^2 - 14t^8 - 2t^{10})\left(6 + 8\sqrt{1 - 28\left(\frac{\zeta(8)t(1-2t^8)}{1+7t^2-14t^8-2t^{10}}\right)^2}\right)}, \\
g_3(t) &= h(t) g_2\left(\frac{1 - \sqrt{1 - 28t^2}}{14t}\right).
\end{aligned}$$

In particular, g_1 and g_2 are algebraic functions of degree 2, and g_3 is algebraic of degree 4. Clearly, g_3's radius of convergence is given by the $g_2(\cdot)$ term, since g_2 counts more circuits than h; the radius of convergence of g_2 is given by the vanishing of the surd expression

$$1 - 28\left(\frac{\zeta(8)t(1-2t^8)}{1 + 7t^2 - 14t^8 - 2t^{10}}\right)^2,$$

which is a polynomial equation of degree 20, with least solution $\alpha \approx 0.357936$. Now the radius of convergence of g_4 is the minimal t such that $(1 - \sqrt{1 - 28t^2})/14t = \alpha$,

namely $\rho = \alpha/(1+7\alpha^2) \approx 0.188702$. We then have $\|M_2\| \geq 1/(8\rho)$, so

$$\|M_2\| \geq 0.662418.$$

Similar computations give

$$\|M_3\| \geq 0.552773.$$

The best upper bounds were obtained by Tatiana Nagnibeda [13], by applying Gabber's lemma [6] to a function on the edges of $X_{4g,4g}$ defined by the edge-origin's cone type. She obtained

$$\|M_2\| \leq 0.662816, \quad \|M_3\| \leq 0.552792.$$

5.1 Isoperimetric constants

Let $\mathfrak{G}$ be a connected graph. For a subset $K \subset V$ of vertices denote by $\partial K = \{e \in E : \alpha(e) \in K \text{ and } \omega(e) \notin K\}$ the *boundary* of K. The number

$$\iota(\mathfrak{G}) = \inf_{\text{finite, non-empty } K \subset V} \frac{\#\partial K}{\#K} \tag{5.4}$$

is called the (*edge-*)*isoperimetric constant* of $\mathfrak{G}$. It is connected to the spectral radius by the following

Theorem 5.1 ([11, 12]). *Let $\mathfrak{G}$ be an infinite d-regular graph. Then one has*

$$\frac{d^2-(d-2)\iota(\mathfrak{G})}{d^2+\iota(\mathfrak{G})} \leq \|M_{\mathfrak{G}}\| \leq \sqrt{1-\iota(\mathfrak{G})^2/d^2}. \tag{5.5}$$

Note that, by (1.1) and (5.5), $\mathfrak{G}$ is amenable if and only if $\iota(\mathfrak{G}) = 0$.

The isoperimetric constant $\iota(X_{d,m})$ has recently been computed independently by Yusuke Higuchi and Tomoyuki Shirai [8] and by Olle Häggström, Johan Jonasson and Russell Lyons [7]. They obtained the values

$$\iota(X_{d,m}) = (d-2)\sqrt{1-\frac{4}{(d-2)(m-2)}}. \tag{5.6}$$

In particular, they obtained $\iota(X_{8,8}) = 4\sqrt{2}$, which, together with (5.5) gives

$$0.431 \approx \frac{16-6\sqrt{2}}{16+\sqrt{2}} \leq \|M_2\| \leq \frac{1}{\sqrt{2}} \approx 0.707.$$

Evidently, these bounds are much weaker than the ones given in this paper. It may be worthwhile to improve the connection between the isoperimetric constant and the spectral radius – probably Theorem 5.1 is not the last word in this topic.

Acknowledgment. The author thanks Tullio Ceccherini-Silberstein, Pierre de la Harpe, Russ Lyons, Tatiana Nagnibeda, and Yuval Peres, who provided valuable comments and encouraged him to write this note.

References

[1] L. Bartholdi, Counting paths in graphs, Enseign. Math. (2) 45 (1999), 83–131.

[2] L. Bartholdi, S. Cantat, T. G. Ceccherini-Silberstein and P. de la Harpe, Estimates for simple random walks on fundamental groups of surfaces, Colloq. Math. 72 (1997), 173–193.

[3] L. Bartholdi and T. G. Ceccherini-Silberstein, Growth functions and random walks on surface graphs, Monatsh. Math. 136 (2002), 181–202.

[4] D. I. Cartwright and P. M. Soardi, Random walks on free products, quotients and amalgams, Nagoya Math. J. 102 (1986), 163–180.

[5] P.-A. Cherix and A. Valette, On spectra of simple random walks on one-relator groups. With an appendix by Paul Jolissaint, Pacific J. Math. 175 (1996), 417–438.

[6] O. Gabber and Z. Galil, Explicit constructions of linear-sized superconcentrators, J. Comput. System Sci. 22 (1981), 407–420.

[7] O. Häggström, J. Jonasson, and R. Lyons, Explicit isoperimetric constants and phase transitions in the random-cluster model, Ann. Probab. 30 (2002), 443–473.

[8] Y. Higuchi and T. Shirai, Isoperimetric constants of (d, f)-regular planar graphs, preprint (2000).

[9] H. Kesten, Full Banach mean values on countable groups, Math. Scand. 7 (1959), 146–156.

[10] H. Kesten, Symmetric random walks on groups, Trans. Amer. Math. Soc. 92 (1959), 336–354.

[11] B. Mohar, Isoperimetric inequalities, growth, and the spectrum of graphs, Linear Algebra Appl. 103 (1988), 119–131.

[12] B. Mohar and W. Woess, A survey on spectra of infinite graphs, Bull. London Math. Soc. 21 (1989), 209–234.

[13] T. Nagnibeda, An upper bound for the spectral radius of a random walk on surface groups, J. Math. Sci. (New York) 96 (1999), 3542–3549.

[14] T. Nagnibeda, Random walks, spectral radii and Ramanujan graphs, in: Random Walks and Geometry, Proceedings of a Workshop at the Erwin Schrödinger Institute, Walter de Gruyter, Berlin 2004, 487–500.

[15] W. L. Paschke, Lower bound for the norm of a vertex-transitive graph, Math. Z. 213 (1993), 225–239.

[16] W. Woess, Nearest neighbour random walks on free products of discrete groups, Boll. Un. Mat. Ital. B (6) 5 (1986), 961–982.

[17] A. Żuk, A remark on the norm of a random walk on surface groups, Colloq. Math. 72 (1997), 195–206.

Laurent Bartholdi, EPFL, IGAT, Bâtiment BCH, 1015 Lausanne, Switzerland
E-mail: laurent.bartholdi@epfl.ch

Equilibrium measure, Poisson kernel and effective resistance on networks

Enrique Bendito, Ángeles Carmona and Andrés M. Encinas

Abstract. We consider the Laplacian of a finite network as a kernel on the vertex set. The properties of this kernel allow us to assign to every proper set an equilibrium measure and a capacity. So, we can build a discrete Potential Theory with respect to the Laplacian kernel on networks. We aim here at showing how equilibrium measures can be used to obtain simple expressions for both Poisson and Green kernels and hence to deduce nice expressions for the effective resistance and the hitting time.

1 Introduction

A systematic treatment of Potential Theory with respect to a kernel can be found in the work of B. Fuglede [7]. Although in that work the particular case in which the underlying space is finite was not considered explicitly, the main results in this context can be deduced without difficulty from the general case. The finite case was explicitly considered in classical works of G. Choquet and J. Deny, for instance, see [4]. However, the kernels considered in all these papers are always non negative, which corresponds to developing Potential Theory with respect to the Green kernel.

In the present work we deal with the Potential Theory of a signed kernel on a finite network. Since any linear operator on a finite space can be considered as an integral operator, the Laplace operator of a network can be interpreted as a kernel on the vertex set. This approach (see [2, 3]) differs from both the classical Potential Theory with respect to the Green kernel and from the Dirichlet Forms Theory.

In Section 2 we present, for the sake of completeness, some of the results on the Discrete Potential Theory with respect to the Laplacian kernel obtained in [2, 3]. Essentially, these results follow from the fact that this kernel satisfies two fundamental principles, namely the energy and Frostman's maximum principles, which allow us to solve equilibrium problems whose solutions will be the basic tool in the rest of the work. Moreover, the Wiener capacity with respect to the Laplacian kernel has some remarkable properties. In particular, it gives information about the connection

between a subset and its complement and allows us to characterize independent vertex sets.

We further study some relevant concepts both in the context of electrical network and in the context of reversible Markov Chains. Specifically, we study the effective resistance and the hitting time. Since we want to characterize these concepts for all the subsets of a network, and such concepts are obtained from the solution of suitable boundary value problems, we are focused on the analysis of the kernels that solve the corresponding boundary value problems. Hence in Section 3 we deal with the study of the Poisson kernel.

The authors gave in [3] a relation between the Poisson kernel and the normal derivative of the Green kernel, analogous to the continuous case. Here we obtain new and simple expressions for the Poisson and Green kernels in terms of equilibrium measures which, in particular, allow us to prove that the Laplacian verifies the condenser principle. In addition, we get some of the more relevant properties of these kernels. This link between kernels and equilibrium measures enables us to get explicit expressions for the effective resistance and for the hitting time in terms of equilibrium measures (Section 4). The expression of the Green function in terms of effective resistances (inverse resistive) has been proved by different authors by using distinct techniques, see for instance Coppersmith et al. [6], Metz [8] and Ponzio [9]. The results presented here allow us to obtain a straightforward proof of that relation.

Throughout the paper $\Gamma = (V, E, c)$ denotes an *electrical network*, that is, a simple and finite connected graph, with vertex set V and edge set E, in which each edge (x, y) has been assigned a *conductance* $c(x, y) > 0$. The *order of* Γ is $n = |V|$. Given a subset $F \subset V$, we denote by F^c its complement in V, and by $\delta(F) = \{x \in F^c : c(x, y) > 0 \text{ for some } y \in F\}$ and $\partial F = \{(x, y) \in E : x \in F, y \in F^c\}$ its *vertex boundary* and its *edge boundary*, respectively. We also use the notation $\overline{F} = F \cup \delta(F)$. The *Laplacian* of Γ is the matrix of order n whose entries are $\mathcal{L}(x, y) = -c(x, y)$ for all $x \neq y$ and $\mathcal{L}(x, x) = c(x)$, where $c(x) = \sum_{y \in V} c(x, y)$.

2 Potential Theory in finite spaces

Let V be a finite space with n points endowed with the discrete topology, and F be a non-empty subset of V. Then, the set of functions on V, denoted by $\mathcal{C}(V)$, and the set of non-negative functions on V, $\mathcal{C}^+(V)$, are naturally identified with $\mathbb{R}^n$ and the positive cone of $\mathbb{R}^n$, respectively. If $u \in \mathcal{C}(V)$, its *support* is the subset $\text{supp}(u) = \{x \in V : u(x) \neq 0\}$. Moreover, we consider the sets $\mathcal{C}(F) = \{u \in \mathcal{C}(V) : \text{supp}(u) \subset F\}$ and $\mathcal{C}^+(F) = \mathcal{C}(F) \cap \mathcal{C}^+(V)$.

A symmetric function $\mathcal{K} : V \times V \longrightarrow \mathbb{R}$ will be called a *kernel* on V. Clearly, a kernel on V is identified with a real symmetric matrix of order n.

On the other hand, the set of Radon measures on V, denoted by $\mathcal{M}(V)$, is identified with $\mathcal{C}(V)$ and hence, if $\mu \in \mathcal{M}(V)$, its *support* is defined as above. Therefore, the

sets of Radon measures supported by F, $\mathcal{M}(F)$, and of positive Radon measures supported by F, $\mathcal{M}^+(F)$, are identified with $\mathcal{C}(F)$ and $\mathcal{C}^+(F)$, respectively. In addition, if $\mu \in \mathcal{M}(V)$ its *mass* is given by $||\mu|| = \sum_{x\in V} \mu(x)$ and we denote by $\mathcal{M}^1(F)$, the set of positive Radon measures supported by F with unit mass. Finally, for each $x \in V$, ε_x stands for the Dirac measure on x, whereas the measure $\sum_{x\in F} \varepsilon_x$ will be denoted by $\mathbf{1}_F$. When $F = V$, the subscript in the above expression will be omitted.

If $\Gamma = (V, E, c)$ is a network then its Laplacian can be considered as a kernel on V. So, given $\mu \in \mathcal{M}(V)$ we will call *potential* and *energy of μ with respect to $\mathcal{L}$*, the function and the real number given respectively by

$$\mathcal{L}\mu(x) = \sum_{y\in V} c(x, y)\big(\mu(x) - \mu(y)\big) \quad x \in V \quad \text{and} \quad \mathcal{I}(\mu) = \langle \mathcal{L}\mu, \mu\rangle.$$

The application $\mathcal{I} : \mathcal{M}(V) \longrightarrow \mathbb{R}$ that assigns to each measure its energy, is a quadratic functional, whose associated bilinear form will be denoted by $\mathcal{I}(\cdot, \cdot)$ and is given by

$$\begin{aligned}\mathcal{I}(\mu_1, \mu_2) = \langle \mathcal{L}\mu_1, \mu_2\rangle &= \frac{1}{2}\sum_{x,y\in V} c(x, y)(\mu_1(x) - \mu_1(y))(\mu_2(x) - \mu_2(y)) \\ &= \langle \mu_1, \mathcal{L}\mu_2\rangle.\end{aligned}$$

Proposition 2.1. *The Laplacian kernel $\mathcal{L}$ satisfies the energy principle, i.e. $\mathcal{L}$ is strictly positive definite on $\{\mu \in \mathcal{M}(V) : ||\mu|| = 0\}$.*

Proof. Note that $\mathcal{I}(\mu) = \sum_{x,y\in V} c(x, y)(\mu(x) - \mu(y))^2 \geq 0$. Moreover, $\mathcal{I}(\mu) = 0$ iff $\mu = a\mathbf{1}$, $a \in \mathbb{R}$, since Γ is connected. □

Proposition 2.2. *The Laplacian kernel $\mathcal{L}$ satisfies Frostman's maximum principle, i.e.* $\max_{x\in V}\{\mathcal{L}\mu(x)\} = \max_{x\in \mathrm{supp}(\mu)}\{\mathcal{L}\mu(x)\}$, *for all $\mu \in \mathcal{M}^+(V)$.*

Proof. Let $\mu \in \mathcal{M}^+(V)$ and $F = \mathrm{supp}(\mu)$. If we consider $x \in F$ such that $\mu(x) = \max_{y\in F} \mu(y)$, then $\mathcal{L}\mu(x) \geq 0$. Moreover, for any $x \in F^c$, $\mathcal{L}\mu(x) \leq 0$. □

Proposition 2.3. *The Laplacian kernel satisfies the equilibrium principle, i.e., for every proper set $F \subset V$ there exists a unique $\nu^F \in \mathcal{M}^+(F)$ such that*

$$\mathcal{L}\nu^F = 1 \text{ on } F.$$

Moreover, $\mathrm{supp}(\nu^F) = F$.

Proof. It is known that if a kernel satisfies the energy and the maximum principles then

$$\min_{\mu\in\mathcal{M}^1(F)} \{\mathcal{I}(\mu)\} = \min_{\mu\in\mathcal{M}^1(F)} \max_{x\in V}\{\mathcal{L}\mu(x)\},$$

see, for instance, [7] for the general case and [3] for the discrete setting.

Moreover, the unique solution of the above problem, $\sigma \in \mathcal{M}^1(F)$, is such that $\mathcal{L}\sigma = \mathcal{I}(\sigma)$ in F and hence $\nu^F = \mathcal{I}(\sigma)^{-1}\,\sigma$.

To prove the last claim, suppose that there exists $x \in F$ such that $\nu^F(x) = 0$. Then $\mathcal{L}\nu^F(x) = -\sum_{y\in V} c(x, y)\nu^F(y) \leq 0$, which contradicts $\mathcal{L}\nu^F = \mathbf{1}$ in F. □

For each proper set F the measure ν^F is called *the equilibrium measure for* F.

Next result is the well-known minimum principle for superharmonic functions. Here we give the proof for completeness and because monotonicity of equilibrium measures can be deduced straightforwardly from it.

Proposition 2.4. *Let F be a proper subset of V. The Laplacian $\mathcal{L}$, as an operator, verifies the minimum principle, i.e., if $u \in \mathcal{C}(\overline{F})$ is such that $\mathcal{L}(u) \geq 0$ on F, then*

$$\min_{x\in\delta(F)} \{u(x)\} \leq \min_{x\in F}\{u(x)\}.$$

Proof. Let $m = \min_{x\in\delta(F)}\{u(x)\}$, and consider $w = u - m\mathbf{1}_{\overline{F}}$. Then, $\mathcal{L}w = \mathcal{L}u \geq 0$ on F and $w \geq 0$ on F^c. Let $x \in F$ be such that $w(x) = \min_{z\in F}\{w(z)\}$. To conclude, it is enough to show that $w(x) \leq 0$ implies $w(x) = 0$.

Suppose that $w(x) \leq 0$. Then, $w(x) \leq w(z)$ for all $z \in V$ and therefore

$$0 \leq \mathcal{L}w(x) = \sum_{z\in V} c(x, z)(w(x) - w(z)) \leq 0,$$

which implies that $w(x) = w(z)$ for each $z \in V$ such that $c(x, z) \neq 0$. Of course, if $c(x, z) > 0$ for some $z \in F^c$, as $w(z) \geq 0$, necessarily $w(x) = 0$. Otherwise as Γ is connected and F is a proper set there exists $y \in F$ such that $w(y) = w(x)$ and $c(y, z) > 0$ for some $z \in F^c$, and hence $w(x) = 0$. □

Corollary 2.5. *If F and H are proper subsets such that $F \subset H$, then $\nu^F \leq \nu^H$.*

The consideration of the Laplacian as a kernel in the context of Potential Theory allows us to introduce the Wiener capacity of a subset. We show that this concept is useful in order to obtain information about the connection between a subset of vertices and its complement. In [2] the authors developed an analogous approach in the case of graphs.

For each $F \subset V$, the value $I(F) = \inf_{\mu\in\mathcal{M}^1(F)} \mathcal{I}(\mu)$ is called the *energy of* F. Moreover, the value $\mathrm{cap}(F) = \frac{1}{I(F)}$ is known as the *Wiener capacity of* F. The unique measure $\sigma \in \mathcal{M}^1(F)$ such that $I(F) = \mathcal{I}(\sigma)$ is called *the capacitary measure for* F. It is easy to check that the Wiener capacity is a monotone set function, i.e., $\mathrm{cap}(F) \leq \mathrm{cap}(H)$ when $F \subset H$. On the other hand, if F is a proper set and ν^F is its equilibrium measure, then $\mathcal{I}(\nu^F)^{-1} = ||\nu^F|| = \mathrm{cap}(F)$.

Proposition 2.6. *Let $F \subset V$ be such that $F = \bigcup_{i=1}^s F_i$, where F_i, $i = 1, \ldots, s$, are the vertex sets of the connected components of the subnetwork induced by F. Then*

$$\mathrm{cap}(F) = \sum_{i=1}^{s} \mathrm{cap}(F_i).$$

Proof. If $F = V$, then $s = 1$, because Γ is connected. Hence, the result holds.

Suppose that F is a proper subset of V. For each $i = 1, \ldots, s$, let μ^i be the capacitary measure for F_i. If $\beta = \sum_{i=1}^{s} \operatorname{cap}(F_i)$ and we consider $\mu \in \mathcal{M}^1(F)$ defined as

$$\mu = \frac{1}{\beta} \sum_{i=1}^{s} \operatorname{cap}(F_i)\mu^i,$$

then $\mathcal{L}\mu = \frac{1}{\beta} \sum_{i=1}^{s} \operatorname{cap}(F_i)\mathcal{L}\mu^i$. If $x \in F$, there exists k such that $x \in F_k$. Moreover, as $x \notin \overline{F_j}$ for all $j \neq k$, then $\mathcal{L}\mu^i(x) = I(F_k)$ if $i = k$ and $\mathcal{L}\mu^i(x) = 0$ otherwise. Hence, $\mathcal{L}\mu(x) = \frac{1}{\beta}$ for all $x \in F$, which implies that $I(F) = \frac{1}{\beta}$ and the result follows. □

The above result states that the Wiener capacity is additive with respect to the connected components of an induced subnetwork. However, it is not true for an arbitrary union of subsets of V. In fact, the following corollary shows that the Wiener capacity is not subadditive.

Corollary 2.7. *If $F \subset V$, then $\operatorname{cap}(F) \geq \sum_{x \in F} \operatorname{cap}(\{x\})$. Moreover, the equality holds iff F is an independent vertex set.*

Proof. Note that $I(\{x\}) = c(x)$, since $\mathcal{L}\varepsilon_x(x) = c(x)$. Consider the equilibrium measure ν^F for F, and let $\sigma(x) = \nu^F(x) - \frac{1}{c(x)}\mathbf{1}_F(x)$, $x \in V$. Then, for each $x \in F$

$$\mathcal{L}\sigma(x) = \sum_{y \in F} \frac{c(x, y)}{c(y)} \geq 0.$$

Hence, $\nu^F(x) \geq \frac{1}{c(x)}$ since $\mathcal{L}$ satisfies the minimum principle, which implies that $\operatorname{cap}(F) \geq \sum_{x \in F} \frac{1}{c(x)}$.

On the other hand, if F is an independent vertex set the equality follows from the above proposition. Conversely, suppose that F is a set of non independent vertices. Then there exist $x_0, y_0 \in F$ such that $c(x_0, y_0) > 0$ which implies that $\mathcal{L}\sigma(x_0) > 0$ and hence $\nu^F(x_0) > \frac{1}{c(x_0)}$. □

The Wiener capacity for the Laplacian kernel is not subadditive due to the fact that the Laplacian is not a positive kernel. If $q = \max_{x,y \in V}\{c(x, y)\}$ and $F \subset V$, the value $(I(F) + q)^{-1}$ can be seen as the Wiener capacity of F with respect to the positive kernel $\mathcal{L} + qJ$, where J denotes the kernel whose values are equal to one. In fact, the Wiener capacity is subadditive for a positive kernel, see [7, p. 157].

Proposition 2.8. *Let $F_1, \ldots, F_s \subset V$ and $F = \bigcup_{i=1}^{s} F_i$. Then*

$$(I(F) + q)^{-1} \leq \sum_{i=1}^{s} (I(F_i) + q)^{-1}.$$

In particular, we have the following result.

Corollary 2.9. *Let $F \subset V$ be a proper subset. Then* $\operatorname{cap}(F)\operatorname{cap}(F^c) \geq \frac{1}{q^2}$.

Before ending this section let us determine the Wiener capacities and the capacitary measures for connected proper subsets of a weighted path which will help us to study the sharpness of the lower bound in the above corollary.

Given a path P_n with $n \geq 2$ vertices and conductances $c_{ii+1}, i = 1, \ldots, n-1$, consider a proper subset $F = \{x_1, \ldots, x_s\}$ of P_n. Then $\sigma(x_j) = \sum_{i=j}^{s} \frac{i}{c_{ii+1}}$ for any $x_j \in F$, and $\operatorname{cap}(F) = \sum_{i=1}^{s} \frac{i^2}{c_{ii+1}}$. By considering $p = \min_{(x,y)\in E}\{c(x, y)\}$, we get that

$$\frac{1}{6q^2}(n-1)n(2n-1) \leq \operatorname{cap}(F)\operatorname{cap}(F^c) \leq \frac{1}{24^2 p^2} n^2(n+1)^2(n+2)^2.$$

On the other hand, in a complete network in which each edge has conductance q, the product of capacities attains its minimum value for any subset. The differences in the behaviour of the capacity products are due to the different degrees of connection between the vertices of F and F^c. In particular, the following result characterizes when equality holds in Corollary 2.9.

Proposition 2.10. *Let $F \subset V$ be a proper subset. Then,*

$$\operatorname{cap}(F)\operatorname{cap}(F^c) = \frac{1}{q^2} \iff |\partial F| = |F||F^c| \text{ and } c(x, y) = q,\ \forall (x, y) \in \partial F.$$

Moreover, the capacitary measures for F and F^c are the uniform measures on F and F^c respectively.

Proof. Note that $|\partial F| = |F||F^c|$ and $c(x, y) = q,\ \forall (x, y) \in \partial F$ iff $\sum_{y\in F^c} c(x, y) = q|F^c|$ for all $x \in F$ and $\sum_{x\in F} c(x, y) = q|F|$ for all $y \in F^c$. In addition, the uniform measures on F and F^c, $\mu_1 = \frac{1}{|F|}\mathbf{1}_F$ and $\mu_2 = \frac{1}{|F^c|}\mathbf{1}_{F^c}$, satisfy $\mathcal{L}\mu_1 = \frac{q|F^c|}{|F|}$ on F and $\mathcal{L}\mu_2 = \frac{q|F|}{|F^c|}$ on F^c, respectively. Therefore, they are the capacitary measures for F and F^c, and $\operatorname{cap}(F)\operatorname{cap}(F^c) = \frac{1}{q^2}$.

Conversely, if $\mathcal{K} = \mathcal{L} + qJ$ and we consider $\mathbf{1} = \mathbf{1}_F + \mathbf{1}_{F^c}$, then

$$\begin{aligned} qn^2 &= \langle \mathcal{K}\mathbf{1}, \mathbf{1}\rangle = \langle \mathcal{K}\mathbf{1}_F, \mathbf{1}_F\rangle + \langle \mathcal{K}\mathbf{1}_{F^c}, \mathbf{1}_{F^c}\rangle + 2\langle \mathcal{K}\mathbf{1}_F, \mathbf{1}_{F^c}\rangle \\ &\geq (\mathcal{I}(\mathbf{1}_F) + q|F|^2) + (\mathcal{I}(\mathbf{1}_{F^c}) + q|F^c|^2) \\ &\geq |F|^2(I(F) + q) + |F^c|^2(\mathcal{I}(F^c) + q) \\ &\geq \frac{(|F|+|F^c|)^2}{\frac{1}{I(F)+q} + \frac{1}{I(F^c)+q}} = \frac{n^2}{\frac{1}{I(F)+q} + \frac{1}{I(F^c)+q}}\ . \end{aligned}$$

On the other hand, $\operatorname{cap}(F)\operatorname{cap}(F^c) = \frac{1}{q^2}$ iff $\frac{1}{(I(F)+q)} + \frac{1}{(I(F^c)+q)} = \frac{1}{q}$. Therefore, by using the above inequalities, we conclude that

$$\operatorname{cap}(F)\operatorname{cap}(F^c) = \frac{1}{q^2} \implies \langle \mathcal{K}\mathbf{1}_F, \mathbf{1}_{F^c}\rangle = 0.$$

Finally, it is enough to observe that

$$|\partial F| = |F||F^c| \text{ and } c(x,y) = q,\ \forall (x,y) \in \partial F \text{ iff } \langle \mathcal{K}\mathbf{1}_F, \mathbf{1}_{F^c}\rangle = 0,$$

because $\langle \mathcal{K}\mathbf{1}_F, \mathbf{1}_{F^c}\rangle = q|F||F^c| - \sum_{(x,y)\in\partial F} c(x,y)$. □

3 The Poisson Kernel

In this section we consider a semihomogeneous Dirichlet Problem, and we solve it by considering the associated Poisson Kernel. Our purpose is to use such a kernel for finding an expression of the effective resistance between subsets, in the next section.

Let $F \subset V$ be a proper subset. A *Semihomogeneous Dirichlet Problem* on F consists in finding, given $g \in \mathcal{C}(\delta(F))$, a function $u \in \mathcal{C}(\overline{F})$ such that

$$\left.\begin{aligned} \mathcal{L}u(x) &= 0, & x &\in F,\\ u(x) &= g(x), & x &\in \delta(F). \end{aligned}\right\} \tag{3.1}$$

As shown in [3], this problem has a unique solution which can be obtained by means of the Poisson Kernel.

A function $P : \overline{F} \times \delta(F) \longrightarrow \mathbb{R}$ is called the *Poisson kernel* for Problem (3.1) if for all $y \in \delta(F)$, $P_y = P(\cdot, y)$ is the solution of the following problem:

$$\left.\begin{aligned} \mathcal{L}P_y &= 0, & &\text{in } F,\\ P_y &= \varepsilon_y, & &\text{in } \delta(F). \end{aligned}\right\} \tag{3.2}$$

From the definition, it is easy to check that the function $u(x) = \sum_{y\in\delta(F)} P(x,y)g(y)$ is the solution of Problem (3.1). Now we show how the use of equilibrium measures allow us to obtain an expression of the Poisson Kernel for every proper set F such that $|\delta(F)| \geq 2$.

Proposition 3.1. *Let $F \subset V$ be a proper set. If $F = V \setminus \{y\}$, then $\delta(F) = \{y\}$ and $P_y = \mathbf{1}$. Otherwise, the Poisson Kernel for F is given by:*

$$P(x,y) = \left(\nu^{F\cup\{y\}}(y)\right)^{-1}\left(\nu^{F\cup\{y\}}(x) - \nu^F(x)\right).$$

As the Poisson kernel for Problem (3.1) only depends on F, we will call it the Poisson kernel for F and it will be denoted by P^F.

Corollary 3.2 (The condenser principle). *Let F be a proper subset of V and $\{A, B\}$ a partition of $\delta(F)$. Then, $u \in \mathcal{C}(\overline{F})$, the unique solution of the boundary value problem*

$$\left.\begin{array}{lll} \mathcal{L}u(x) & = 0 & \text{if } x \in F, \\ u(x) & = 1 & \text{if } x \in A, \\ u(x) & = 0 & \text{if } x \in B, \end{array}\right\}$$

is such that $0 \leq u \leq 1$ on V, $\mathcal{L}u \geq 0$ on A, and $\mathcal{L}u \leq 0$ on B.

Proof. From Proposition 3.1, the solution of the boundary value problem is

$$u = \sum_{y \in A} \frac{\nu^{F \cup \{y\}} - \nu^F}{\nu^{F \cup \{y\}}(y)}.$$

Moreover, from Corollary 2.5, $u \geq 0$ on V, which implies that if $x \in B$ then $\mathcal{L}u(x) = -\sum_{z \in V} c(x, z)u(z) \leq 0$. Consider now $v = \mathbf{1} - u$, then v is the solution of

$$\left.\begin{array}{lll} \mathcal{L}v(x) & = 0 & \text{if } x \in F, \\ v(x) & = 0 & \text{if } x \in A, \\ v(x) & = 1 & \text{if } x \in B. \end{array}\right\}$$

Therefore, reasoning as above, $v \geq 0$ on V, $\mathcal{L}v \leq 0$ on A and *a fortiori* $u \leq \mathbf{1}$ and $\mathcal{L}u \geq 0$ on A. □

Proposition 3.3. *If F is a proper subset of V, then $0 \leq P_y^F \leq 1$ for each $y \in \delta(F)$. Moreover, if $F \neq V - \{y\}$ then $\mathcal{L}P_y^F(y) > 0$, and when F is connected, $0 < P_y^F < 1$ on F for each $y \in \delta(F)$.*

Proof. If $y \in \delta(F)$ and we consider $u = P_y^F$, then u satisfies $\mathcal{L}u = 0$ on F, $u(y) = 1$ and $u = 0$ on $\delta(F) - \{y\}$. Then, applying the condenser principle with $A = \{y\}$ and $B = \delta(F) - \{y\}$, we get that $0 \leq u \leq 1$ and moreover $\mathcal{L}u(y) \geq 0$. As $\langle \mathcal{L}(u), u \rangle = \mathcal{L}u(y)$, it follows that $\mathcal{L}u(y) = 0$ iff $u = a\mathbf{1}$, $a \in \mathbb{R}$. Moreover, if $F \neq V \setminus \{y\}$, $a = 0$ since $u = 0$ on $\delta(F) \setminus \{y\}$. So, in this case $\mathcal{L}u(y) > 0$. On the other hand, if $x \in F$, necessarily $u(x) > 0$, since otherwise

$$0 = \mathcal{L}(\nu^{F \cup \{y\}} - \nu^F)(x) = -\sum_{z \in V} c(x, z)\big(\nu^{F \cup \{y\}}(z) - \nu^F(z)\big) \leq 0,$$

because of Corollary 2.5, and therefore $\nu^{F \cup \{y\}}(z) = \nu^F(z)$ for all z such that $c(x, z) \neq 0$. Since Γ is connected, we get that $\nu^{F \cup \{y\}}(y) = \nu^F(y) = 0$, which is a contradiction. Reasoning analogously for $v = 1 - u$, we get that $u < 1$ for all $z \in F$. □

It is well known that a Dirichlet Problem can be solved by using the associated Green kernel, so that there should be a relation between the Poisson and Green kernels.

This relation is well known in the continuous case and we investigate it next for the discrete case. Let us start by defining the Green kernel for a subset.

A function $G : \overline{F} \times F \longrightarrow \mathbb{R}$ is called the *Green kernel* for Problem (3.1) if for all $y \in F$, $G_y = G(\cdot, y)$ is the solution of the following problem:

$$\left.\begin{aligned} \mathcal{L}G_y &= \varepsilon_y \quad \text{in } F, \\ G_y &= 0 \quad \text{in } \delta(F). \end{aligned}\right\} \tag{3.3}$$

Clearly the solution of (3.1) is given by $u(x) = g(x) - \sum_{y\in F} G(x, y)\mathcal{L}g(y)$. As before, the Green kernel depends only of F so that it will be denoted by G^F.

Proposition 3.4. *Let F be a proper subset of V. Then*

(i) *G^F is a symmetric kernel, and $G^F_y(y) > 0$, $P_y^{F-\{y\}} = \frac{G^F_y}{G^F_y(y)}$ and $0 \le G^F_y \le G^F_y(y)$ for each $y \in F$.*

(ii) *$P^F(x, y) = \varepsilon_x(y) - \frac{\partial}{\partial\eta_y} G^F(x, y)$, where $\frac{\partial}{\partial\eta_y} G^F(x, y) = -\sum_{z\in V} c(x, y)G^F(x, y)$ denotes the normal derivative of $G^F(x, \cdot)$.*

(iii) *$G^F(x, y) = \frac{\nu^F(y)}{||\nu^F||-||\nu^F_y||}\big(\nu^F(x) - \nu^F_y(x)\big)$, where ν^F_y denotes the equilibrium measure for the set $F - \{y\}$, $y \in F$.*

Proof. (i) Clearly, G^F is symmetric and non-negative from Proposition 2.4. On the other hand, if $y \in F$, then $G^F_y(y) > 0$ since otherwise we would have $\mathcal{L}G^F_y(y) = -\sum_{z\in V} c(y, z)G^F_y(z) \le 0$, in contradiction with $\mathcal{L}G^F_y = \varepsilon_y$. Moreover, if we consider $u = \frac{1}{G^F_y(y)} G^F_y$, it follows that $u(y) = 1$, $u = 0$ on $\delta(F)$ and in addition $\mathcal{L}u = 0$ on $F\setminus\{y\}$. Therefore, $u = P_y^{F\setminus\{y\}}$. The last part follows from Proposition 3.3.

(ii) For each $y \in \delta(F)$ and each $x \in \overline{F}$, let $u(x) = \varepsilon_x(y) - \frac{\partial}{\partial\eta_y} G^F(x, y)$. Then $\mathcal{L}u = 0$ on F, $u(y) = 1$ and $u = 0$ on $\delta(F) \setminus \{y\}$. Hence $u = P^F_y$.

(iii) From (i), $G^F_y = G^F_y(y)P_y^{F\setminus\{y\}}$. It is enough to find the value $G^F_y(y)$ by imposing the conditions satisfied by the Green kernel. So,

$$\begin{aligned} 1 = \mathcal{L}G^F_y(y) &= \frac{G^F_y(y)}{\nu^F(y)}\big(\mathcal{L}\nu^F(y) - \mathcal{L}\nu^F_y(y)\big) \\ &= \frac{G^F_y(y)}{\nu^F(y)}\big(1 - \mathcal{L}\nu^F_y(y)\big) = \frac{G^F_y(y)}{\nu^F(y)^2}\big(||\nu^F|| - ||\nu^F_y||\big), \end{aligned}$$

since $||\nu^F_y|| = \langle\nu^F_y, \mathbf{1}\rangle = \langle\nu^F_y, \mathcal{L}\nu^F\rangle = \langle\mathcal{L}\nu^F_y, \nu^F\rangle = ||\nu^F|| - \nu^F(y)\big(1 - \mathcal{L}\nu^F_y(y)\big)$. □

4 The effective resistance

The effective resistance between two vertices or two subsets of a network is defined as the inverse of the current arising from applying to them the unit potential. It is well known that the effective resistance is the inverse of the energy of the solution of the Dirichlet problem. It is also known that we can restrict ourselves to the case in which the source and the sink are single vertices joined with an edge. Here we look (similarly to the continuous case [1]) at the situation when the vertex boundary of a set F is partitioned into three subsets: two of them are the source (one vertex) and the sink (one vertex), and the remaining part of the boundary is insulated. The general effective resistance can be defined in terms of the energy of the solution of the corresponding mixed boundary value problem which is solved by using the Poisson kernel. Then, using equilibrium measures, we deduce an expression for the effective resistance between two vertices of an electrical network when the remaining part of the boundary is insulated. If the insulated subset is empty, then we obtain a formula for the standard effective resistance. We also examine the probabilistic interpretation of these results.

Let $F \subset V$ and $\delta(F) = \{y\} \cup \{z\} \cup D$, a partition of the vertex boundary. If the unit potential is applied across vertices y and z, then, according to Kirchhoff's Laws, the potential u at vertices of the network has to be the solution of the following *mixed boundary value problem*:

$$\left.\begin{aligned} \mathcal{L}u(x) &= 0 \ \text{ if } \ x \in F, \\ u(y) &= 1, \\ u(z) &= 0, \\ \tfrac{\partial u}{\partial \eta}(x) &= 0 \ \text{ if } \ x \in D, \end{aligned}\right\} \tag{4.4}$$

where $\frac{\partial u}{\partial \eta}(x) = \sum_{t \in F} c(x,t)\big(u(x) - u(t)\big)$ is the *normal derivative* of u at a point $x \in D$. When $D = \emptyset$, Problem (4.4) gives the standard notion of effective resistance between vertices y and z. From now on we suppose that $F \cup \{y\} \cup \{z\} \cup D = V$, because the potentials at vertices of the set $V \setminus (F \cup \{y\} \cup \{z\} \cup D)$ are zero and hence these vertices can be identified with z.

On the other hand, as shown in [3], this problem has a unique solution and can be transformed into a Dirichlet problem in the following way: consider a new network built from the subnetwork induced by F adding its edge and vertex boundaries. Specifically, given F, we define the network $\overline{\Gamma}(F) = (\overline{F}, \overline{E}, \overline{c})$, where $\overline{E} = \{(x,t) \in E : x \in F\}$, and the conductance function $\overline{c}$ is the restriction of c to $\overline{E}$. We denote the Laplacian of this network by $\overline{\mathcal{L}} = \mathcal{L}(\overline{\Gamma})$. Note that $\overline{\mathcal{L}}u(x) = \mathcal{L}u(x)$ if $x \in F$, and $\overline{\mathcal{L}}u(x) = \frac{\partial u}{\partial \eta}(x)$ if $x \in \delta(F)$. So, $u \in \mathcal{C}(\overline{F})$ is the solution of Problem (4.4) iff u is

the solution of the following Dirichlet problem:

$$\left.\begin{aligned} \overline{\mathcal{L}}u(x) &= 0, \ x \in F \cup D, \\ u(y) &= 1, \\ u(z) &= 0. \end{aligned}\right\} \tag{4.5}$$

We define the *effective resistance between y and z when D is insulated* as $R^D_{yz} = \overline{\mathcal{I}}(u)^{-1}$, where $\overline{\mathcal{I}}$ is the energy with respect to the kernel $\overline{\mathcal{L}}$. As $u(x) = P^{F\cup D}(x, y)$, to find an expression for the effective resistance we must know the value of $P^{F\cup D}_y$. Let us see that in this case we can get a simpler expression for that kernel. To simplify the notation, we will write P^{yz} instead of $P^{V\setminus\{y,z\}}$.

Proposition 4.1. *Let $F \cup D = V \setminus \{y, z\}$. Then the Poisson kernel for $F \cup D$ is given by:*

$$P^{yz}(x, y) = \frac{\nu_z(x) - \nu_y(x) + \nu_y(z)}{\nu_z(y) + \nu_y(z)} \quad \textit{and} \quad P^{yz}(x, z) = \frac{\nu_y(x) - \nu_z(x) + \nu_z(y)}{\nu_z(y) + \nu_y(z)},$$

where ν_t denotes the equilibrium measure for the set $V\setminus\{t\}$ with respect to the kernel $\overline{\mathcal{L}}$.

Proof. Consider the unique solutions $u, v \in \mathcal{C}(V)$ of the problems

$$\left.\begin{aligned} \overline{\mathcal{L}}u(x) &= 0, \ x \in F \cup D \\ u(y) &= 1 \\ u(z) &= 0 \end{aligned}\right\} \quad \text{and} \quad \left.\begin{aligned} \overline{\mathcal{L}}v(x) &= 0, \ x \in F \cup D \\ v(y) &= 0 \\ v(z) &= 1 \end{aligned}\right\}$$

respectively. Then, both solutions are determined by the identities

$$u(x) = P^{yz}(x, y) \quad \text{and} \quad v(x) = P^{yz}(x, z).$$

Keeping in mind the expression for P^{yz} given in Proposition 3.1, we obtain that $u = \frac{\nu_z - \nu_{yz}}{\nu_z(y)}$ and $v = \frac{\nu_y - \nu_{yz}}{\nu_y(z)}$, respectively, where ν_{yz} denotes the equilibrium measure for the set $V \setminus \{y, z\}$. On the other hand, as $u + v = \mathbf{1}$, adding the above expressions gives

$$\nu_{yz} = \frac{\nu_z\nu_y(z) + \nu_y\nu_z(y) - \nu_y(z)\nu_z(y)}{\nu_y(z) + \nu_z(y)}.$$

The result then follows from substituting the above expression in the formulas for $P^{yz}(x, y)$ and $P^{yz}(x, z)$. □

Corollary 4.2. *Let F be a subset of V such that $\delta(F) = \{y\} \cup \{z\} \cup D$. Then*

$$R^D_{yz} = \frac{\nu_y(z) + \nu_z(y)}{n}.$$

Proof. Let u be the solution of problem (4.5). Then, keeping in mind the expression for the Poisson kernel of the set $V \setminus \{y, z\}$ obtained in Proposition 4.1, we get that

$$\overline{\mathit{I}}(u) = \overline{\mathcal{L}}u(y) = \frac{1}{\nu_y(z) + \nu_z(y)} \overline{\mathcal{L}}(\nu_z - \nu_y)(y) = \frac{n}{\nu_y(z) + \nu_z(y)},$$

because $0 = \langle \nu_y, \overline{\mathcal{L}}\mathbf{1} \rangle = \langle \overline{\mathcal{L}}\nu_y, \mathbf{1} \rangle = n - 1 + \overline{\mathcal{L}}\nu_y(y)$. □

Corollary 4.3. *Let $z \in V$. If G^z denotes the Green kernel for the set $V \setminus \{z\}$, then $G^z(y, y) = R_{yz}$. Moreover,*

$$G^z(x, y) = \frac{1}{n}\big(\nu_z(x) - \nu_y(x) + \nu_y(z)\big).$$

Proof. From the proof of Proposition 3.4 (iii), we get that

$$G^z(y, y) = \frac{\nu_z(y)}{1 - \mathcal{L}\nu_{yz}(y)}.$$

On the other hand, keeping in mind the expression for ν_{yz} obtained in the proof of Proposition 4.1, we get $\mathcal{L}\nu_{yz}(y) = \frac{1}{\nu_y(z)+\nu_z(y)}(\nu_y(z) + \nu_z(y)(1 - n)) = 1 - \frac{\nu_z(y)}{R_{yz}}$. Hence, $G^z(y, y) = R_{yz}$. Moreover from Proposition 3.4 (i),

$$G^z(x, y) = \frac{\nu_z(x) - \nu_y(x) + \nu_y(z)}{\nu_z(y) + \nu_y(z)} R_{yz} = \frac{1}{n}\big(\nu_z(x) - \nu_y(x) + \nu_y(z)\big). \qquad \square$$

We can obtain direct formulas for the Poisson kernel of the set $V \setminus \{y, z\}$ and for the Green kernel for the set $V \setminus \{z\}$ in terms of the effective resistance between two vertices.

Proposition 4.4. *Let $y, z \in V$, then*

(i) *The Green kernel for the set $V \setminus \{z\}$ is given by*

$$G^z(x, y) = \frac{1}{2}\big(R_{xz} + R_{yz} - R_{xy}\big).$$

(i) *The Poisson kernel for the set $V \setminus \{y, z\}$ is given by*

$$P^{yz}(x, y) = \frac{1}{2R_{yz}}\big(R_{xz} + R_{yz} - R_{xy}\big), \quad P^{yz}(x, z) = \frac{1}{2R_{yz}}\big(R_{xy} + R_{yz} - R_{xz}\big).$$

Proof. (i) From Corollary 4.3 we know that

$$\begin{aligned} G^z(x, y) &= \tfrac{1}{n}\big(\nu_z(x) - \nu_y(x) + \nu_y(z)\big) \\ &= \tfrac{1}{2n}\big(\nu_z(x) - \nu_y(x) + \nu_y(z) + \nu_z(y) - \nu_x(y) + \nu_x(z)\big) \\ &= \tfrac{1}{2}\big(R_{xz} + R_{yz} - R_{xy}\big), \end{aligned}$$

where the second identity follows from the symmetry of the Green kernel.

(ii) It follows from the previous point and from part (i) of Proposition 3.4. □

The expression obtained in the above proposition for the Green kernel is well known in the context of the so-called *resistive inverse*. Specifically, given the matrix $(R_{xy})_{x,y\in V}$ of effective resistances of a network, we are interested in finding the matrix of conductances of the network. Coppersmith et al. [6] gave a simple but obscure four-step algorithm for computing the resistive inverse. Later, Ponzio [9] gave a self-contained combinatorial explanation of this algorithm. This relation was also given by Metz [8] using Dirichlet forms. Here we have obtained a new and simple proof of that algorithm in term of equilibrium measures.

Some of the concepts considered here have a well-known probabilistic interpretation. For instance, the effective resistance is related with the escape probability for a reversible Markov chain. Also, Problem 4.4 can be described in terms of the Neumann random walk, see [5]. Hence, the general concept of effective resistance corresponds to a generalization of the escape probability.

Finally, let us consider another Dirichlet problem whose solution has an important probabilistic meaning. For that, let $\Gamma = (V, E, c)$ be the network that has as vertices the states of a reversible Markov chain and as conductances $c(x, y) = \pi(x)p(x, y)$, where $p(x, y)$ is the transition probability from state x to state y and $\pi(x)$ is the value of the stationary distribution at state x. Then, the *hitting time* $H(x, y)$ *from* x *to* y, defined as the expected number of steps in order to reach the state y from the state x, satisfies the following relations:

$$\begin{aligned} \mathcal{L}H_y(x) &= c(x), \quad x \in V \setminus \{y\}, \\ H_y(y) &= 0. \end{aligned}$$

Therefore, by using the expression for the Green kernel G^y we obtain

$$H(x, y) = \frac{1}{n} \sum_{z\in V} c(z)\big(v_z(y) + v_y(x) - v_z(x)\big),$$

and also the well known relation between the hitting time and the effective resistance, see [10]:

$$H(x, y) = \frac{1}{2} \sum_{z\in V} c(z)\big(R_{xy} + R_{yz} - R_{xz}\big).$$

Acknowledgement. This work has been supported by the Comisión Interministerial de Ciencia y Tecnología (Spanish Research Council) under project BFM2000-1063 and the ETSECCPB.

References

[1] L. V. Ahlfors and L. Sario, Riemann Surfaces, Princeton Univ. Press, Princeton, NJ, 1960.

[2] E. Bendito, A. Carmona and A. M. Encinas, Shortest paths in distance-regular graphs, Europ. J. Combin. 21 (2000), 153–166.

[3] E. Bendito, A. Carmona and A. M. Encinas, Solving boundary value problems on networks using equilibrium measures, J. Funct. Anal. 171 (2000), 155–176.

[4] G. Choquet, J. Deny, Modèles finis en théorie du potentiel, J. Analyse Math. 5 (1956/57), 77–135.

[5] F. R. K. Chung, Spectral Graph Theory, CBMS Regional Conference Series in Mathematics, 92, Amer. Math. Soc., Providence, RI, 1997.

[6] D. Coppersmith, P. Doyle, P. Raghavan and M. Snir, Random walks on weighted graphs and applications to on-line algorithms, J. Assoc. Comput. Mach. 40 (1993), 421–453.

[7] B. Fuglede, On the theory of potentials in locally compact spaces, Acta Math. 103 (1960), 139–215.

[8] V. Metz, Shorted operators: an application in potential Theory, Linear Algebra Appl. 264 (1997), 439–455.

[9] S. Ponzio, The combinatorics of effective resistances and resistive inverses, Inform. and Comput. 147 (1998), 209–223.

[10] P. Tetali, Random walks and the effective resistance of networks, J. Theoret. Probab. 4 (1991), 101–109.

Enrique Bendito, Angeles Carmona, Andrés M. Encinas, Departament de Matemàtica Aplicada III, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain

E-mail: enrique.bendito@upc.es
angeles.carmona@upc.es
andres.marcos.encinas@upc.es

Internal diffusion limited aggregation on discrete groups of polynomial growth

Sébastien Blachère

Abstract. The Internal Diffusion Limited Aggregation has been introduced by Diaconis and Fulton in 1991. It is a growth model defined on an infinite set and associated to a Markov chain on this set. We present first an overview of its behavior on $\mathbb{Z}^d$. In that case, under moment conditions, when the Markov chain is a centered random walk, the model has a limiting shape which coincides with the trace of certain Euclidean balls of $\mathbb{R}^d$. Simulations show that the non-centered case leads to a different behavior for which only very partial results are available. In the second part, we study this model for symmetric random walk on discrete groups with polynomial growth. We obtain bounds for the shape of the model, less precise than those on $\mathbb{Z}^d$.

1 Introduction

Let $\mathcal{G}$ be an infinite discrete set. Let $S(k)$ be an irreducible Markov chain on the elements of $\mathcal{G}$, where $S(0) = O$ a fixed element. The Internal Diffusion Limited Aggregation (Internal DLA) is a Markov chain $A(n)$ ($n \in \mathbb{N}$) of increasing subsets of $\mathcal{G}$ defined as follow:

- $A(1) = \{O\}$,
- $\mathbf{P}\{A(n+1) = A(n) \cup \{y\} \mid A(n)\} = \mathbf{P}\{S(\tau(A^c(n))) = y\}$,

where $\tau(A)$ is the hitting time of the set A.

Thus, the set $A(n)$ grows as follows. At each discrete time n, we start a particle at O, wait until the particle leaves the previous set $A(n-1)$, and add the first element visited outside $A(n-1)$, to obtain $A(n)$. Note that this Markov chain is well-defined because, for all n, $\tau(A^c(n)) < \infty$ almost surely, as $A(n)$ is finite and $S(k)$ is irreducible.

The name Internal DLA comes from the similarity with the (external) DLA. In the latter model, the particles start "at infinity", conditioned to hit the previous cluster, and are stuck just before hitting it. See for example [7] for a precise definition and some properties.

The Internal DLA was introduced by Diaconis and Fulton in 1991 [2]. They gave the first shape theorem in the case when $\mathcal{G}$ is the natural graph of $\mathbb{Z}$, and $S(k)$

is the Simple Random Walk (nearest neighbor random walk with uniform transition probabilities). Later, other shape theorems were proved on $\mathbb{Z}^d$ ([9, 8, 1]). They will be reviewed in Section 2.

In Section 3 we study this model when $\mathcal{G}$ is the Cayley graph of a polynomial growth group. Here, weaker shape theorem are obtained, based on the comparison between the shape of the model and level lines of the Green function.

2 Internal DLA on $\mathbb{Z}^d$

In [2], Diaconis and Fulton obtained the first shape theorem for the Internal DLA on $\mathbb{Z}$ associated with the Simple Random Walk. Their result uses the fact that in this special case the model can be considered as a generalized urn model.

Theorem 2.1 ([2, Proposition 3.2]). *Denote by $A^+(n)$ the number of positive sites visited at time n. Then as n goes to infinity,*

$$\mathbf{P}\left\{\frac{\mathrm{Vol}(A^+(n)) - n/2}{\sqrt{n/12}} \leq t\right\} \longrightarrow \frac{1}{\sqrt{2\pi}}\int_0^t \exp(-x^2/2)dx.$$

Lawler, Bramson, and Griffeath [9] studied the case of $\mathbb{Z}^d$ for $d \geq 2$ (for simplicity, we write $\mathbb{Z}^d$ for both the graph and the set of vertices) with the Simple Random Walk. They proved that the limiting shape of the cluster exists and is the trace of the Euclidean balls in $\mathbb{Z}^d$. In [8], sharper estimates on the fluctuations of the cluster around these balls were obtained.

More precisely, let $|\cdot|$ be the Euclidean norm on $\mathbb{R}^d$, and ω_d be the volume of the unit ball in $\mathbb{R}^d$. Let $B(n)$ be the trace of the Euclidean balls on $\mathbb{Z}^d$, $B(n) = \{z \in \mathbb{Z}^d : |z| < n\}$, so that $B(n)$ contains approximately $[\omega_d n^d]$ points. Let $\delta_{\mathrm{I}}(n)$ and $\delta_{\mathrm{O}}(n)$ be two random variables describing the inner and outer error:

$$n - \delta_{\mathrm{I}}(n) = \inf\left\{|z| : z \notin A([\omega_d n^d])\right\},$$
$$n + \delta_{\mathrm{O}}(n) = \sup\left\{|z| : z \in A([\omega_d n^d])\right\}.$$

Note that $\delta_{\mathrm{I}}(n)$ and $\delta_{\mathrm{O}}(n)$ are functions of the random set $A([\omega_d n^d])$. The following result is due to Lawler. It should be compared to [6] where the average fluctuations in a similar model (the Diffusion Limited Erosion) are studied in the context of statistical physics.

Theorem 2.2 ([8, Theorem 1]). *If $d \geq 2$, then with probability* 1,

$$\delta_{\mathrm{I}}(n) = o((\ln n)^2 n^{1/3}),$$

and

$$\delta_{\mathrm{O}}(n) = o((\ln n)^4 n^{1/3}).$$

This shape theorem, in the case $d = 2$, can be illustrated by the simulation presented in Figure 1 (taken from [10]).

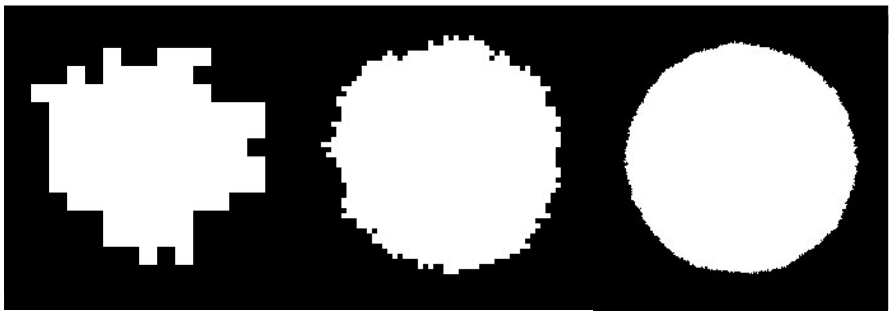


Figure 1. 2-dimensional (normalized) Internal DLA at times 100, 1600, and 25600

Still on $\mathbb{Z}^d$, we can look at more general random walks, typically irreducible and centered ones. Then the expected limiting shape corresponds to balls of $\mathbb{R}^d$ with a specific norm: let $p(x) = \mathbf{P}\{X = x\}$ (where X is the increment of the random walk) and let Q be the covariance matrix of X. The matrix Q is symmetric and positive definite, with determinant $|Q|$. The norm we consider is then

$$\|x\| = (d^{-1} x^t Q^{-1} x)^{1/2}.$$

Note that this is the norm which appears in central limit theorems.

Remark 2.3. When $S(k)$ is the Simple Random Walk, this norm coincides with the Euclidean one. And for $d = 1$, Q is simply the variance of X.

The volume of balls in $\mathbb{R}^d$ with this norm becomes

$$\mathrm{Vol}(\mathcal{B}(0, n)) = d^{d/2} |Q|^{1/2} \omega_d n^d = \omega_d^Q n^d \ .$$

The new definitions of the inner error and outer error are

$$n - \delta_{\mathrm{I}}^Q(n) = \inf \left\{ \|z\| : z \notin A([\omega_d^Q n^d]) \right\},$$
$$n + \delta_{\mathrm{O}}^Q(n) = \sup \left\{ \|z\| : z \in A([\omega_d^Q n^d]) \right\}.$$

The existence of a limiting shape and the precision about fluctuations will depend on moment conditions for the random walk:

$$\mathcal{A}: \begin{cases} \mathbf{E}\{\|X\|^4\} < \infty & \text{if } d \le 4, \\ \mathbf{E}\{\|X\|^{d-1}\} < \infty & \text{if } d \ge 5. \end{cases}$$

$$\mathcal{B}: \begin{cases} \mathbf{E}\{\|X\|^4\} < \infty & \text{if } d \le 3, \\ \mathbf{E}\{\|X\|^4 \ln \|X\|\} < \infty & \text{if } d = 4, \\ \mathbf{E}\{\|X\|^d\} < \infty\} & \text{if } d \ge 5. \end{cases}$$

$$\mathcal{C}(\psi): \begin{cases} \mathbf{E}\{\|X\|^{3+(2/\psi)} (\ln \|X\|)^6\} < \infty & \text{if } d = 1, \\ \mathbf{E}\{\|X\|^{3+(d+2)/\psi} (\ln \|X\|)^4\} < \infty & \text{if } d \ge 2. \end{cases}$$

$$\mathcal{D}(\psi): \begin{cases} \mathbf{E}\{\|X\|^{3+(d+2/3)/\psi} (\ln \|X\|)^6\} < \infty & \text{if } d \ge 2. \end{cases}$$

The shape theorem is then

Theorem 2.4 ([1, Theorem 2.3]). *Consider a centered irreducible random walk on $\mathbb{Z}^d$. Then the inner error with probability 1 satisfies the following estimates:*

- $\delta_{\mathrm{I}}^{Q}(n) = o(n^{1/2} \ln n)$ *for* $d = 1$, *under* $\mathcal{A}$;
- $\delta_{\mathrm{I}}^{Q}(n) = O(n^{1/2})$ *for* $d \ge 2$, *under* $\mathcal{A}$;
- $\delta_{\mathrm{I}}^{Q}(n) = o(n^{1/3} \ln n)$ *for* $d \ge 2$, *under* $\mathcal{B}$ *and assuming that the random walk is symmetric*.

The outer error with probability 1 *satisfies the following estimates* (*where one puts* $\psi = 0$ *for a finite range random walk, and replaces* n^{ψ} *with* $\ln n$ *for a random walk with exponential moments*):

- $\delta_{\mathrm{O}}^{Q}(n) = o(n^{\psi+1/2}(\ln n)^3)$ *for* $d = 1$, *under* $\mathcal{C}(\psi)$ *for* ψ *such that* $0 < \psi < 1/2$;
- $\delta_{\mathrm{O}}^{Q}(n) = o(n^{2\psi+1/2}(\ln n)^2)$ *for* $d \ge 2$, *under* $\mathcal{C}(\psi)$ *for* ψ *such that* $0 < \psi < 1/4$;
- $\delta_{\mathrm{O}}^{Q}(n) = o(n^{2\psi+1/3}(\ln n)^3)$ *for* $d \ge 2$, *under* $\mathcal{D}(\psi)$ *for* ψ *such that* $0 < \psi < 1/3$ *and assuming that the random walk is symmetric*.

Note that Theorem 2.4 gives polynomial bounds for the fluctuations of the model (between $o(n^{1/3})$ and $o(n)$), whereas Theorem 2.2 leads to logarithmic fluctuations in the special case of the Simple Random Walk.

In [1], a necessary moment condition is obtained:

Proposition 2.5 ([1, Proposition 2.20]). *If* $\delta_{\mathrm{I}}^{Q}(n) = o(n)$ *and* $\delta_{\mathrm{O}}^{Q}(n) = o(n)$ *almost surely, then the random walk must have at least* $d + 1$ *moments*.

Finally, the non-centered case leads to a behavior only partially understood when $d \ge 2$, see [1, Theorem 3.5]. In dimension 1, under moment conditions, the limiting

shape of $A(n)$ is the interval $[0, n]$ or $[-n, 0]$ (depending on the direction of the drift) with precise estimates on the fluctuations, see [1, Lemma 3.1 and Theorem 3.4].

As an example of a 2-dimensional non-centered random walk, take the one, starting at $(0, 0)$, with transition probabilities defined as

$$p((x, y), (x + 1, y + 1)) = p((x, y), (x + 1, y - 1)) = 1/2.$$

Then the Internal DLA can be simulated and leads to an interesting and hypnotical shape (Figure 2, the two lines correspond to $y = \pm x$). Nevertheless, no shape theorem is known for this example.

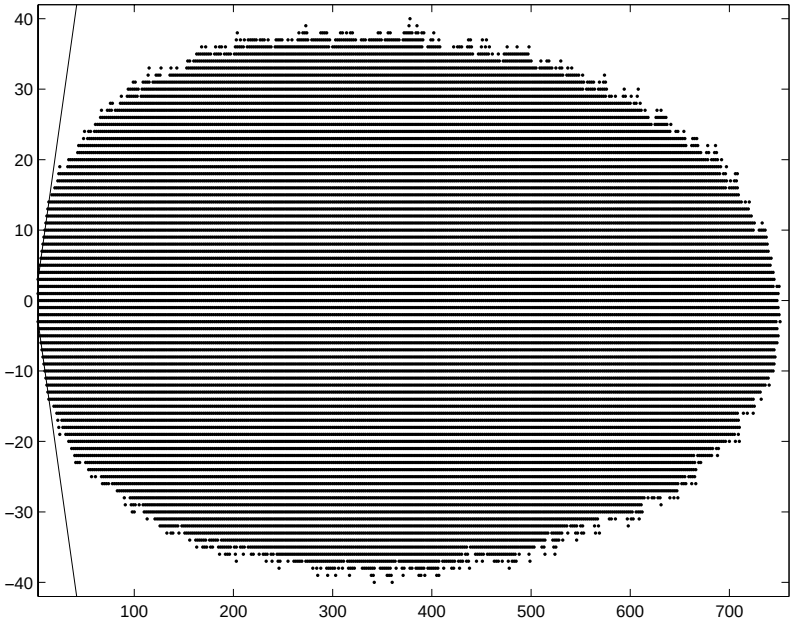


Figure 2. Example at time $n = 23000$

3 Internal DLA on groups of polynomial growth

Let Γ be a finitely generated group of polynomial growth. Let $\mathcal{S}$ be a symmetric set of generators. We write each element x of Γ as a word in the alphabet $\mathcal{S}$. The minimal length of words representing x is called the word length $d(x)$. It determines the word distance on Γ. According to this distance, we define the balls $B_w(n) = B_w(e, n) = \{x \in \Gamma : d(x) \leq n\}$ and their volume $V_w(n)$. The polynomial growth property

corresponds to

$$c^{-1}n^D \leq V_w(n) \leq cn^D, \tag{3.1}$$

for some constant $c > 1$ and some integer D, called the degree of Γ. Note that, by a theorem of Gromov [3], an upper bound of the form $V_w(n) \leq an^A$ for all n, implies that (3.1) holds for some integer D.

Let $S(k)$ be an irreducible symmetric random walk on Γ, with finite support. Hence, the support of its increment can be taken as a set of generators $\mathcal{S}$ of Γ. Then, $S(k)$ is a nearest neighbor symmetric random walk on the Cayley graph $\mathcal{G} = (\Gamma, \mathcal{S})$, on which we study the Internal DLA.

We consider only the transient case, that is when $D \geq 3$ since we strongly use the Green function whose definition makes sense only in this case. We will prove a much weaker result than the one on $\mathbb{Z}^D$ in the sense that we do not know whether or not a limiting shape exists. Note that, even for $\mathbb{Z}^D$, the word distance is not adapted to describe the limiting shape. Thus, we define new spheres as the level lines of the Green function. Precisely, let $G(x) = G(e, x)$ be the Green function associated to $S(k)$, that is the expected number of visits at x, starting at e. Hebisch and Saloff-Coste [4, Theorem 5.1] proved the following estimates: there exists a constant C such that (for $x \neq e$)

$$(Cd(x))^{2-D} \leq G(x) \leq C(d(x))^{2-D}, \tag{3.2}$$

$$\nabla G(x) \leq C(d(x))^{1-D}, \tag{3.3}$$

where $\nabla G(x) = \max_{y \in x\mathcal{S}} |G(y) - G(x)|$ is a discrete version of the gradient of G. We define, for $n \geq 1$, our new balls and spheres,

$$B(n) = \{x \in \Gamma : \ G(x) \geq (\delta/n)^{D-2}\},$$
$$\partial B(n) = \{x \in B(n) : \ x\mathcal{S} \cap B(n)^c \neq \varnothing\}.$$

For some suitable constant δ, James and Peres [5, Lemma 2.3] proved that the spheres $\partial B(n)$ are separated ($\partial B(n) \cap B(n-1) = \emptyset$). By (3.1) and (3.2) the volume $V(n)$ of $B(n)$ is of order n^D, namely, there exist two constants $\underline{v}$ and $\overline{v}$ such that for every n

$$\underline{v}n^D \leq V(n) \leq \overline{v}n^D. \tag{3.4}$$

The result is the following:

Theorem 3.1. *There exist two constants c_1 and c_2, such that, with probability* 1*, for all n large enough*

$$B(c_1 n) \subset A(n^D) \subset B(c_2 n).$$

First, remark that this statement holds also for B_w. But we will need these new balls B in the proof. For instance, on $\mathbb{Z}^D$ these balls coincide (for large n) with the trace of the Euclidean ones.

We split the proof into two parts. The lower bound proof is adapted from [8], whereas the upper bound proof uses simpler techniques from [9].

The superscript in the notations $\mathbf{P}^z$ and $\mathbf{E}^z$ refers to random walks starting at z. We omit it when $z = e$. By c, c' we denote constants whose value may change from one line to another. We denote by $\xi_n = \inf\{k \geq 1 :\ S(k) \in \partial B(n)\}$ the hitting time of the sphere of radius n.

3.1 Proof of the lower bound

The idea is to bound $\mathbf{P}\{z \notin A(n^D)\}$ for $z \in B(c_1 n)$ by a quantity sufficiently small to get the result by the Borel–Cantelli Lemma. Estimates of this probability will come from comparison of two random variables related to the behavior, with respect to z, of the n^D random walks. Then, using estimates on the expectation of these random variables, the event $\{z \notin A(n^D)\}$ will be seen in terms of large deviation from the average for those random variables.

Let $G_n(x, y)$ be the stopped Green function, that is, the expected number of visits to y, starting at x, and before ξ_n,

$$G_n(x, y) = \mathbf{E}^x\Big\{ \sum_{k=0}^{\xi_n - 1} \mathbf{1}^x\{S(k) = y\}\Big\}.$$

Lemma 3.2. *There exists a non-negative constant c_a such that, for every z within $B(n)\backslash\partial B(n)$*

$$c_a G(z) \leq G_n(e, z) \leq G(z). \tag{3.5}$$

Proof. The upper bound is obvious. For the lower bound, write

$$\begin{aligned} G_n(e, z) &= \mathbf{E}\Big\{ \sum_{k=0}^{\xi_n - 1} \mathbf{1}\{S(k) = z\}\Big\} \\ &= \mathbf{E}\Big\{ \sum_{k=0}^{\infty} \mathbf{1}\{S(k) = z, k < \xi_n\}\Big\} \\ &= \sum_{k=0}^{\infty} \mathbf{P}\{S(k) = z\}\mathbf{P}\{k < \xi_n | S(k) = z\}. \end{aligned}$$

We denote $\tau_z = \inf\{j \geq 1 :\ S(j) = z\}$ and $\tau_z^n = \inf\{j \geq \xi_n :\ S(j) = z\}$, assuming that these times are infinite when the infimum is not defined. Then

$$\mathbf{P}\{k < \xi_n | S(k) = z\} \geq \mathbf{P}\{\tau_z^n = \infty\} = \sum_{x \in \partial B(n)} \mathbf{P}^x\{\tau_z = \infty\}\mathbf{P}\{S(\xi_n) = x\}.$$

But $\mathbf{P}^x\{\tau_z = \infty\} \geq \min_{y \in z\mathcal{S}} \mathbf{P}^y\{\tau_z = \infty\} \geq c_a > 0$ by transience of the random walk. So,

$$G_n(e, z) \geq c_a G(z). \qquad \square$$

Lemma 3.3. *There exist two non-negative constants c_b and c'_b such that, for every z in $B(n/2)$*

$$c'_b n^2 \leq \mathbf{E}^z\{\xi_n\} \leq c_b n^2. \tag{3.6}$$

Proof. We first prove the upper bound (true for every z in $B(n)$).

$$\begin{aligned}
\mathbf{E}^z\{\xi_n\} &= \sum_{k=1}^{n^2-1} k\mathbf{P}^z\{\xi_n = k\} + \sum_{k=n^2}^{\infty} k\mathbf{P}^z\{\xi_n = k\} \\
&\leq n^2\mathbf{P}^z\{\xi_n \leq n^2 - 1\} + \sum_{k=n^2}^{\infty} k[\mathbf{P}^z\{\xi_n > k-1\} - \mathbf{P}^z\{\xi_n > k\}] \\
&\leq n^2[\mathbf{P}^z\{\xi_n \leq n^2 - 1\} + \mathbf{P}^z\{\xi_n > n^2 - 1\}] + \sum_{k=n^2}^{\infty} \mathbf{P}^z\{\xi_n > k\} \\
&\leq n^2 + \sum_{k=n^2}^{\infty} \mathbf{P}^z\{\xi_n > k\}.
\end{aligned}$$

So, we are left with the last term in the right hand side. Take the Dirichlet kernel $\mathbf{p}^k_{B(n)}(z, y) = \mathbf{P}^z\{S(k) = y,\ k \leq \xi_n\}$ with the Dirichlet (killing) condition on the boundary $\partial B(n)$, defined as the transition kernel of the random walk $S(k)$ restricted to $\{k \leq \xi_n\}$. Then

$$\mathbf{P}^z\{\xi_n > k\} = \sum_{y \in B(n)\setminus\partial B(n)} \mathbf{p}^k_{B(n)}(z, y).$$

By [4, Theorem 5.1],

$$\mathbf{p}^k_{B(n)}(z, y) \leq \mathbf{P}^z\{S(k) = y\} \leq \frac{J}{V(\sqrt{k})},$$

for some constant J which can be chosen strictly greater than $[\overline{v}\underline{v}^{-1}]^{1/(-1+D/2)}$. Then, by (3.4) and as $D \geq 3$, there exists a constant $J' < 1$ such that

$$J'V(\sqrt{Jk}) \geq JV(\sqrt{k}).$$

Therefore,

$$\mathbf{p}^{Jk}_{B(n)}(z, y) \leq \frac{J}{V(\sqrt{Jk})} \leq \frac{J'}{V(\sqrt{k})}. \tag{3.7}$$

The semi-group property of the kernel $\mathbf{p}^k_{B(n)}(\cdot, \cdot)$ implies

$$\begin{aligned}
\mathbf{P}^z\{\xi_n > k\} = \sum_{y_1,\ldots,y_a, y \in B(n)\setminus\partial B(n)} &\mathbf{p}^{Jn^2}_{B(n)}(z, y_1) \times \mathbf{p}^{Jn^2}_{B(n)}(y_1, y_2) \times \cdots \\
&\cdots \times \mathbf{p}^{Jn^2}_{B(n)}(y_{a-1}, y_a) \times \mathbf{p}^{b}_{B(n)}(y_a, y),
\end{aligned}$$

with $a = [k/(Jn^2)]$ and $b = k - aJn^2$ if $k \geq Jn^2$. If $k < Jn^2$, then the sum is only over y.

When $k \geq Jn^2$, as $\sum_{y \in B(n) \setminus \partial B(n)} \mathbf{p}^b_{B(n)}(y_a, y) \leq 1$, by (3.7),

$$\begin{aligned} \mathbf{P}^z\{\xi_n > k\} &\leq \sum_{y_1, \ldots, y_a \in B(n) \setminus \partial B(n)} \left[\frac{J'}{V(n)}\right]^a \leq (J')^a \\ &\leq \exp\left(\frac{k}{Jn^2} \ln J'\right) = \exp\left(-\frac{k}{cn^2}\right), \end{aligned} \tag{3.8}$$

with $c > 0$ because $J' < 1$. If $k < Jn^2$,

$$\mathbf{P}^z\{\xi_n > k\} \leq \mathrm{e} \cdot \exp\left(-\frac{k}{Jn^2}\right).$$

So,

$$\sum_{k=n^2}^{\infty} \mathbf{P}^z\{\xi_n > k\} \leq c \sum_{k=n^2}^{\infty} \exp\left(-\frac{k}{cn^2}\right) \leq cn^2.$$

Thus, we obtain the upper bound of Lemma 3.3. For the lower bound, we use [11, Lemma 3] as in [4, Theorem 9.1], namely

$$\mathbf{P}\{\xi_n < k\} \leq c \exp\left(-\frac{n^2}{ck}\right), \tag{3.9}$$

for some constant $c > \mathrm{e}$. For any $\alpha > 0$,

$$\begin{aligned} \mathbf{E}^z\{\xi_n\} &= \sum_{k \geq \alpha n^2} k\mathbf{P}^z\{\xi_n = k\} + \sum_{k < \alpha n^2} k\mathbf{P}^z\{\xi_n = k\} \\ &\geq \alpha n^2 \mathbf{P}^z\{\xi_n \geq \alpha n^2\} + (\alpha n^2 - 1)\mathbf{P}^z\{\xi_n < \alpha n^2 + 1\} \\ &\quad - \sum_{k < \alpha n^2} \mathbf{P}^z\{\xi_n < k\} \\ &\geq \alpha n^2 - 1 - \sum_{k < \alpha n^2} \mathbf{P}\{\xi_{n/2} < k\} \\ &\geq \alpha n^2 - 1 - c \int_0^{\alpha n^2} \exp\left(-\frac{n^2}{4cx}\right) dx \\ &\geq \alpha n^2 - 1 - 4c^2\alpha^2 n^2 \exp\left(-\frac{1}{4\alpha c}\right). \end{aligned}$$

Hence, if we choose $\alpha < 1/(4c \ln c)$, then $\mathbf{E}^z\{\xi_n\} \geq c'_b n^2$. □

Let $\beta \geq 2$ be such that

$$\underline{v}(\delta\beta)^{D-2} c_a > c_b.$$

Now we prove that, with probability 1, for all n large enough

$$B(n/\beta) \subset A(n^D).$$

Among the n^D random walks that define $A(m)$, the j-th one is associated to the random walk $S^j(k)$ starting at e, and we define:

$$\begin{aligned}
\xi_n^j &= \inf\{k \geq 1 :\ S^j(k) \in \partial B(n)\} \quad \text{hitting time of } \partial B(n),\\
\rho^j &= \inf\{k \geq 0 :\ S^j(k) \notin A(j-1)\} \quad \text{adding time to the cluster,}\\
\tau_z^j &= \inf\{k \geq 1 :\ S^j(k) = z\} \quad \text{hitting time of } z.
\end{aligned}$$

Without the j, these stopping times refer to a general random walk starting at e, and ξ_n^y, ρ^y, τ_z^y refer to a general random walk starting at y. We define two random variables associated to $z \in B(n/\beta)$.

$$\begin{aligned}
M = M(n,z) &= \sum_{j=1}^{n^D} \mathbf{1}\{\tau_z^j < \xi_n^j\}\\
&= \#\text{ of } n^D \text{ walks that visit } z \text{ before hitting } \partial B(n),
\end{aligned}$$

and

$$\begin{aligned}
L = L(n,z) &= \sum_{j=1}^{n^D} \mathbf{1}\{\rho^j \leq \tau_z^j < \xi_n^j\}\\
&= \#\text{ of } n^D \text{ walks that visit } z \text{ after adding to the cluster and before hitting } \partial B(n).
\end{aligned}$$

Remark that, by irreducibility, for any fixed n_0,

$$\mathbf{P}\{B(n_0/\beta) \not\subset A(n^D) \text{ for infinitely many } n\} = 0.$$

Then, if we write

$$B(n/\beta) = \Big[\bigcup_{k=n_0}^{n} B(k/\beta)\backslash B(k/(2\beta))\Big] \cup B(n_0/\beta),$$

we obtain

$$\{B(n/\beta) \not\subset A(n^D)\} \subset \{B(n_0/\beta) \not\subset A(n^D)\} \bigcup_{k=n_o}^{n} \{B(k/\beta)\backslash B(k/(2\beta)) \not\subset A(k^D)\},$$

and so, using the Borel–Cantelli Lemma, the lower bound in Theorem 3.1 follows from

$$\sum_{n=1}^{\infty} \sum_{z \in B\left(\frac{n}{\beta}\right)\backslash B\left(\frac{n}{2\beta}\right)} \mathbf{P}\{z \notin A(n^D)\} < \infty.$$

From now, let n and $z \in B(n/\beta)\backslash B(n/(2\beta))$ be fixed. Note that for any $a > 0$

$$\mathbf{P}\{z \notin A(n^D)\} \leq \mathbf{P}\{M = L\} \leq \mathbf{P}\{M \leq a\} + \mathbf{P}\{L \geq a\}.$$

To estimate the right hand side we will need to compute the expectations of M and L. The summands of M are i.i.d., then we can remove the index j, and obtain

$$\mathbf{E}\{M\} = n^D \mathbf{P}\{\tau_z < \xi_n\}.$$

The summands of L are not i.i.d., but only the particles such that $\rho^j < \xi_n^j$ contribute to the sum. Moreover, for each $y \in B(n)$ there is at most one j such that $S^j(\rho^j) = y$. So,

$$L \leq \sum_{y \in B(n)\backslash \partial B(n)} \mathbf{1}^y\{\tau_z^y < \xi_n^y\} = \widetilde{L}.$$

Now, the summands of $\widetilde{L}$ are i.i.d., so that

$$\mathbf{E}\{\widetilde{L}\} = \sum_{y \in B(n)\backslash \partial B(n)} \mathbf{P}^y\{\tau_z < \xi_n\},$$

and

$$\mathbf{P}\{z \notin A(n^D)\} \leq \mathbf{P}\{M \leq a\} + \mathbf{P}\{\widetilde{L} \geq a\}. \tag{3.10}$$

Remark that

$$G_n(z, z)\mathbf{E}\{M\} = n^D G_n(e, z),$$

and

$$G_n(z, z)\mathbf{E}\{\widetilde{L}\} = \sum_{y \in B(n)\backslash \partial B(n)} G_n(y, z) = \sum_{y \in B(n)\backslash \partial B(n)} G_n(z, y) = \mathbf{E}^z\{\xi_n\}.$$

So, by Lemma 3.2 and the fact that $z \in B(n/\beta)\backslash B(n/(2\beta))$,

$$\begin{aligned} &\mathbf{E}\{M\} \leq \overline{v} n^D [G_n(z, z)]^{-1} G(z) \leq (2\delta\beta)^{D-2} n^2 \overline{v} [G_n(z, z)]^{-1} \\ &\mathbf{E}\{M\} \geq \underline{v} c_a n^D [G_n(z, z)]^{-1} G(z) \geq (\delta\beta)^{D-2} n^2 \underline{v} c_a [G_n(z, z)]^{-1}. \end{aligned} \tag{3.11}$$

By Lemma 3.3,

$$\begin{aligned} &\mathbf{E}\{\widetilde{L}\} \leq c_b n^2 [G_n(z, z)]^{-1} \\ &\mathbf{E}\{\widetilde{L}\} \geq c_b' n^2 [G_n(z, z)]^{-1}. \end{aligned} \tag{3.12}$$

Then

$$\mathbf{E}\{M\}^{2/3} + \mathbf{E}\{\widetilde{L}\}^{2/3} \leq c n^{4/3} [G_n(z, z)]^{-2/3},$$

and, by the choice of β

$$\begin{aligned}\mathbf{E}\{M\} - \mathbf{E}\{\widetilde{L}\} &\geq cn^2[G_n(z,z)]^{-1} \\ &\geq c[\mathbf{E}\{M\}^{2/3} + \mathbf{E}\{\widetilde{L}\}^{2/3}][G_n(z,z)]^{-1/3} \\ &\geq c[\mathbf{E}\{M\}^{2/3} + \mathbf{E}\{\widetilde{L}\}^{2/3}].\end{aligned}$$

The last inequality comes from $G_n(z,z) \leq G(z,z) = G(e)$.

To conclude, we need a large deviation theorem.

Lemma 3.4 ([9, Lemma 4]). *Let $\mathcal{T}$ be a sum of k independent indicator functions, and $\mu = \mathbf{E}\{\mathcal{T}\}$. Then for all sufficiently large k and all $\lambda \in (0, 1/4)$ and $c > 0$,*

$$\mathbf{P}\{|\mathcal{T} - \mu| \geq (c/2)\mu^{1/2+\lambda}\} \leq 2\exp(-c\mu^{2\lambda}/8).$$

Taking $a = \mathbf{E}\{M\} - (c/2)[\mathbf{E}\{M\}^{2/3} + \mathbf{E}\{\widetilde{L}\}^{2/3}]$ leads to

$$\mathbf{P}\{M \leq a\} \leq \mathbf{P}\{|M - \mathbf{E}\{M\}| \geq (c/2)\mathbf{E}\{M\}^{2/3}\},$$

and

$$\mathbf{P}\{\widetilde{L} \geq a\} \leq \mathbf{P}\{|\widetilde{L} - \mathbf{E}\{\widetilde{L}\}| \geq (c/2)\mathbf{E}\{\widetilde{L}\}^{2/3}\}.$$

The result follows then from Lemma 3.4 with $\lambda = 1/6$, since both $\exp(-\mathbf{E}\{M\}^{1/3})$ and $\exp(-\mathbf{E}\{\widetilde{L}\}^{1/3})$ tend exponentially fast to 0 by (3.11) and (3.12) together with $G_n(z,z) \leq G(e)$.

3.2 Proof of the upper bound

Let $\partial' B(k) = \{z \in \partial B(k) : z\mathcal{S} \cap B(k)\backslash\partial B(k) \neq \emptyset\}$, this is the part of $\partial B(k)$ where lies $S(\xi_k)$. To prove the upper bound in Theorem 3.1, we will estimate the decay of the expected number of points in $A(n^D)$ that lie within the shells $\partial' B(k)$ for $k > n$. Checking that this decay is exponential in k will lead to the desired result.

Lemma 3.5. *There exists a constant C_h such that for every $\alpha < 1$, k large enough, and $z \in \partial' B(k)$*

$$\mathbf{P}^z\{\xi_{[\alpha k]} < \xi_k\} \leq C_h(\alpha^{-1/2} - 1)k^{-1}.$$

Proof. We denote $\xi'_k = \inf\{t \geq 0 : S(t) \in \partial B(k)\}$ and $\tau = \xi_{[\alpha k]} \wedge \xi'_k$. Then

$$\mathbf{P}^z\{\xi_{[\alpha k]} < \xi_k\} = \sum_{y \in z\mathcal{S}} p(z^{-1}y)\mathbf{P}^y\{\xi_{[\alpha k]} < \xi'_k\} \leq \max_{y \in z\mathcal{S}} \mathbf{P}^y\{\xi_{[\alpha k]} < \xi'_k\}. \tag{3.13}$$

We denote by y the point in the neighborhood of z where this maximum is attained. Remark that $y \in B(k)\backslash\partial B(k)$, otherwise this probability is 0. As the random walk is transient, the Green function G is bounded and harmonic on $\Gamma\backslash\{e\}$. So, the sequence $M_l = G(S^y(l \wedge \tau))$ is a bounded martingale [7, Proposition 1.4.1]. Therefore, by the

optional sampling theorem $\mathbf{E}^y\{M_0\} = \mathbf{E}^y\{M_\tau\}$. So,

$$\begin{aligned} G(y) &= \mathbf{E}^y\{G(S(\tau))\} \\ &= \mathbf{P}^y\{\tau = \xi_{[\alpha k]}\}\mathbf{E}^y\{G(S(\tau))|\tau = \xi_{[\alpha k]}\} \\ &\quad + (1 - \mathbf{P}^y\{\tau = \xi_{[\alpha k]}\})\mathbf{E}^y\{G(S(\tau))|\tau = \xi'_k\}. \end{aligned}$$

Then

$$\begin{aligned} \mathbf{P}^y\{\xi_{[\alpha k]} < \xi'_k\} &= \mathbf{P}^y\{\tau = \xi_{[\alpha k]}\} \\ &= \frac{G(y) - \mathbf{E}^y\{G(S(\tau))|\tau = \xi'_k\}}{\mathbf{E}^y\{G(S(\tau))|\tau = \xi_{[\alpha k]}\} - \mathbf{E}^y\{G(S(\tau))|\tau = \xi'_k\}}. \end{aligned}$$

By definition of $\partial B([\alpha k])$,

$$\mathbf{E}^y\{G(S(\tau))|\tau = \xi_{[\alpha k]}\} \geq \delta^{D-2}(k\alpha)^{2-D},$$

and by [5, Lemma 2.3], with the constant C taken from (3.2) and (3.3),

$$\mathbf{E}^y\{G(S(\tau))|\tau = \xi'_k\} \leq (1 + \delta C^D/k)\delta^{D-2}k^{2-D} \leq \alpha^{-D+5/2}\delta^{D-2}k^{2-D},$$

for large k, since $\alpha < 1$. So,

$$\mathbf{E}^y\{G(S(\tau))|\tau = \xi_{[\alpha k]}\} - \mathbf{E}^y\{G(S(\tau))|\tau = \xi'_k\} \geq (\alpha^{-1/2} - 1)\delta^{D-2}k^{2-D}.$$

Again by [5, Lemma 2.3] and [4, Theorem 5.1], as y lies in the neighborhood of $z \in \partial B(k)$,

$$G(y) \leq G(z) + \nabla G(z) \leq (1 + \delta C^D/k)\delta^{D-2}k^{2-D} + ck^{1-D},$$

and

$$\mathbf{E}^y\{G(S(\tau))|\tau = \xi'_k\} \geq \delta^{D-2}k^{2-D}.$$

so

$$G(y) - \mathbf{E}^y\{G(S(\tau)) \mid \tau = \xi'_k\} \leq C^D\delta^{D-1}k^{1-D} + ck^{1-D} \leq c'k^{1-D}.$$

Finally,

$$\mathbf{P}^z\{\xi_{[\alpha k]} < \xi_k\} \leq C_h(\alpha^{-1/2} - 1)k^{-1},$$

for some constant C_h. □

Lemma 3.6. *There exists a constant C'_h such that for every $k \in \mathbb{N}$ and $U \in \partial' B(k)$*

$$h(\partial B(k), U) \leq C'_h(\#U)k^{1-D}.$$

where $h(\partial B(k), U)$ is the probability to hit $\partial B(k)$ in U.

Proof. It suffices to prove the statement for $U = \{z\}$ and conclude for a general U by summing over all its elements. We denote $\mu_k = \xi_k \wedge \tau_e$. Then, taking $\alpha = 1/4$, a last

decomposition argument (see [7, Lemma 2.1.1]) gives

$$h(\partial B(k), z) = \mathbf{P}\{S(\xi_k) = z\} = G_k(e, e)\mathbf{P}\{S(\mu_k) = z\}.$$

So, by symmetry,

$$\begin{aligned} h(\partial B(k), z) &\leq G_k(e, e)\mathbf{P}^z\{\xi_{[\alpha k]} < \xi_k\} \\ &\quad \times \sum_{x \in \partial B([\alpha k])} \mathbf{P}^z\{S(\mu_k) = e, S(\xi_{[\alpha k]}) = x \mid \xi_{[\alpha k]} < \xi_k\} \\ &\leq C_h k^{-1} G_k(e, e) \sup_{x \in \partial B([\alpha k])} \mathbf{P}^x\{S(\mu_k) = e\} \\ &\leq C_h k^{-1} \sup_{x \in \partial B([\alpha k])} G(x) \\ &\leq C'_h k^{1-D}. \end{aligned}$$

The second inequality uses the Markov property and Lemma 3.5. The third one uses the symmetry of the random walk. □

We define $Z_k(j) = \mathrm{Vol}(A(j) \cap \partial' B(k))$ $(k \geq 1)$, and $\nu_k(j) = \mathbf{E}\{Z_k(j)\}$, then

$$\nu_{k+1}(l+1) - \nu_{k+1}(l) = \mathbf{E}\{h(A(l)^c, \partial B(k+1))\}.$$

But, if a particle hits $A(l)^c$ in $\partial B(k+1)$, that implies it has stayed within the cluster before. So, by Lemma 3.6,

$$\nu_{k+1}(l+1) - \nu_{k+1}(l) \leq \mathbf{E}\{h(\partial B(k), A(l))\} \leq \mathbf{E}\{C'_h Z_k(l) k^{1-D}\} = C'_h \nu_k(l) k^{1-D}.$$

But, for any $k \geq 1$, $\nu_k(0) = 0$, then by summing these telescoping inequalities from $l = 0$ to $j - 1$, we obtain

$$\nu_{k+1}(j) \leq C'_h k^{1-D} \sum_{l=1}^{j-1} \nu_k(l).$$

We prove, by induction on $k \geq n$, that

$$\forall j\,, \quad \nu_k(j) \leq [C'_h n^{1-D}]^{k-n} \, \frac{j^{k-n+1}}{(k-n+1)!}. \tag{3.14}$$

Indeed, for $k = n$, (3.14) becomes $\nu_n(j) \leq j$ which is always true. On the other hand, if the assumption is true until $k \geq n$,

$$\nu_{k+1}(j) \leq C'_h k^{1-D} [C'_h n^{1-D}]^{k-n} \sum_{l=1}^{j-1} \frac{l^{k-n+1}}{(k-n+1)!} \leq [C'_h n^{1-D}]^{k+1-n} \frac{j^{k-n+2}}{(k-n+2)!}.$$

As $k! \geq k^k \exp(-k)$, we obtain

$$\nu_k(j) \leq \left[C'_h n^{1-D} \frac{\mathrm{e}\, j}{k-n+1} \right]^{k-n} \frac{\mathrm{e}\, j}{k-n+1}.$$

Then for any $c > 1$

$$\mathbf{P}\{A(n^D) \not\subset B(cn)\} \le \mathbf{P}\{Z_{cn}(n^D) \ge 1\} \le \nu_{cn}(n^D) \le \left[\frac{C'_h \mathrm{e}}{c-1}\right]^{(c-1)n} \frac{\mathrm{e} n^{D-1}}{c-1}.$$

Taking $c = c_2 > C'_h e + 1$, the right-hand side tends exponentially fast to zero and so, by the Borel–Cantelli Lemma we obtain the upper bound of Theorem 3.1.

Acknowledgement. I would like to thank Laurent Saloff-Coste for his great help.

References

[1] S. Blachère, Agrégation limitée par diffusion interne sur $\mathbb{Z}^d$, Ann. Inst. H. Poincaré Probab. Statist. 38 (2002), 613–648.

[2] P. Diaconis and W. Fulton, A growth model, a game, an algebra, Lagrange inversion, and characteristic classes, Rend. Sem. Mat. Univ. Pol. Torino 49 (1991), 95–119.

[3] M. Gromov, Groups of polynomial growth and expanding maps, Inst. Hautes Études Sci. Publ. Math. 53 (1981), 53–73.

[4] W. Hebisch and L. Saloff-Coste, Gaussian estimates for Markov chains and random walks on groups, Ann. Probab. 21 (1993), 673–709.

[5] N. James and Y. Peres, Cutpoints and exchangeable events for random walks, Theory Probab. Appl. 41 (1996), 666-677.

[6] J. Krug and P. Meakin, Kinetic roughening of Laplacian fronts, Phys. Rev. Lett. 66 (1991), 703–706.

[7] G. Lawler, Intersections of Random Walks, Birkhäuser, Boston 1991.

[8] G. Lawler, Subdiffusive fluctuations for internal diffusion limited aggregation, Ann. Probab. 23 (1995), 71–86.

[9] G. Lawler, M. Bramson, and D. Griffeath, Internal diffusion limited aggregation, Ann. Probab. 20 (1992), 2117–2140.

[10] C. Moore and J. Machta, Internal diffusion-limited aggregation: parallel algorithms and complexity, J. Statist. Phys. 99 (2000), 661–690.

[11] L. Saloff-Coste and D. W. Stroock, Opérateurs uniformément sous-elliptiques sur les groupes de Lie, J. Funct. Anal. 98 (1991), 97–121.

Sébastien Blachère, Laboratoire de Statistique et Probabilités, Université Toulouse III, and École Polytechnique Fédérale de Lausanne

E-mail: sebastien.blachere@epfl.ch

URL: http://dmawww.epfl.ch/~blachere/

On the physical relevance of random walks: an example of random walks on a randomly oriented lattice

Massimo Campanino and Dimitri Petritis

Abstract. Random walks on general graphs play an important role in the understanding of the general theory of stochastic processes. Beyond their fundamental interest in probability theory, they arise also as simple models of physical systems. A brief survey of the physical relevance of the notion of random walk on both undirected and directed graphs is given followed by the exposition of some recent results on random walks on randomly oriented lattices. It is worth noticing that general undirected graphs are associated with (not necessarily Abelian) groups while directed graphs are associated with (not necessarily Abelian) C^*-algebras. Since quantum mechanics is naturally formulated in terms of C^*-algebras, the study of random walks on directed lattices has been motivated lately by the development of the new field of quantum information and communication.

1 Introduction

Random walks are mathematical objects thoroughly studied nowadays by probabilists for their own interest but also for shedding new light into a variety of mathematical problems like diffusions on manifolds, harmonic analysis, infinite graph theory, group theory, etc. The other contributions in this volume give an overview of the most recent developments of random walks in connection with most of these mathematical disciplines.

An interesting class of problems concerns random walks in random environments. The present contribution intends to present some recent results on a particular kind of random environment defined by the orientation of some edges of the graph on which the random walk evolves. However, we felt that instead of merely announcing these results and reproducing the proofs – that can anyway be found elsewhere [9] – it should be useful for the mathematical community to enlarge this report by giving a short overview on the physical relevance of random walks both on undirected lattices and directed lattices (based on some recent connections between them and quantum evolution, established recently in [29] and in [35]). Therefore, this contribution is organised as follows. In Section 2 we give the notations, definitions, and our main

results on the probabilistic problem we have studied [9]. In Section 3 we give a short overview of the usefulness of random walks on undirected lattices as representations of physical quantities like Green's functions. In section 4 we recall an algebraic construction of (partially) oriented lattices in terms of C^*-algebras and demonstrate the usefulness of random walks in a particular example stemming from quantum communication. Finally in Section 5 we sketch the proof of our results; for detailed proofs the reader can consult [9].

2 Notation, definitions, and main results

We use the notations $\mathbb{N} = \{0, 1, 2, \dots\}$ and $\mathbb{N}_+ = \{1, 2, 3, \dots\}$. An *oriented* (or equivalently *directed*) graph $\mathbb{G}$ is the quadruple $\mathbb{G} = (G^0, G^1, r, s)$, where G^0 and G^1 are denumerable sets called respectively *vertex* and *edge* sets, and r, s are mappings $G^1 \to G^0$ called respectively *range* and *source* functions. For every $v \in G^0$, the set $N_v^- = s^{-1}(\{v\}) \subseteq G^1$ represents the set of outgoing edges from vertex v, and its cardinality $d_v^- = \mathrm{card} N_v^-$ is called the *outwards degree* of v. We assume that our graphs are *row-finite*, i.e., $d_v^- < \infty, \forall v \in G^0$. Similarly, we denote by $N_v^+ = r^{-1}(\{v\}) \subseteq G^1$ the set of incoming edges at vertex v and by $d_v^+ = \mathrm{card} N_v^+$ the *inwards degree* of v. We assume that our graphs are *locally finite*, i.e., $d_v^+ < \infty, \forall v \in G^0$. Finally, we denote $N_v = N_v^+ \cup N_v^-$ and $d_v = \mathrm{card} N_v$. The sets G^0 and G^1 are primitive objects for the graph $\mathbb{G}$. For every $n \in \mathbb{N}_+$, we define a sequence of derived objects $G^n = \{\alpha = (a_1, \dots, a_n) : a_i \in G^1,\ i = 1, \dots, n \text{ and } r(a_i) = s(a_{i+1}),\ i = 1, \dots, n-1\}$ called *oriented paths* of length n, i.e., composed of n edges of G^1. For $\alpha \in G^n$, we denote by $|\alpha| = n$ its length. The mappings r, s, initially defined on G^1 are naturally extended to G^n: for $\alpha = a_1 \dots a_n$, $r(\alpha) \equiv r(a_n)$ and $s(\alpha) \equiv s(a_1)$. The vertices $v \in G^0$ are considered as paths of length 0, and if $\alpha \in G^0$ then $r(\alpha) = s(\alpha) = \alpha$.

The set $G^* = \bigcup_{n \in \mathbb{N}} G^n$ represents the set of oriented paths of arbitrary (finite) length, and

$$\partial G^* \equiv G^\infty = \{\alpha = (a_i)_{i=0}^\infty : s(a_{i+1}) = r(a_i), i \in \mathbb{N}\}$$

represents the set of infinite paths. Thus, the set of oriented paths of $\mathbb{G}$ acquires a natural tree structure.

All graphs considered here are assumed *transitive*, i.e., for every $u, v \in G^0$, there is an $\alpha \in G^*$ such that $s(\alpha) = u$ and $r(\alpha) = v$. Transitivity implies in particular the *no-sink* condition $d_v^- \geq 1$, for all $v \in G^0$.

Remark 2.1. Notice that the above definition of $\mathbb{G}$ is quite general, namely:

- It does not exclude multiple edges since no conditions are imposed on the mappings r, s. If for any two edges $e, f \in G^1$ we cannot have simultaneously

$r(e) = r(f)$ and $s(e) = s(f)$, then all edges are simple and moreover each edge can be thought of as a pair of vertices, i.e., in that case $G^1 \subseteq G^0 \times G^0$.

- It does not exclude loops (of length 1) since it is not excluded that $r(\alpha) = s(\alpha)$ for some $\alpha \in G^1$. We say that the graph has no loops if $r(\alpha) \neq s(\alpha)$ for all $\alpha \in G^1$.
- Finally, the existence of unoriented edges is not excluded since an unoriented edge can be thought of as a pair of oriented edges $\alpha, \beta \in G^1$ with $r(\alpha) = s(\beta)$ and $r(\beta) = s(\alpha)$.

Definition 2.2. Let $\mathbb{G} = (G^0, G^1, r, s)$ be an oriented, transitive, row-finite graph. A *simple random walk* on $\mathbb{G}$ is a G^0-valued Markov chain $(\mathbf{M}_n)_{n\in\mathbb{N}}$ with stochastic matrix defined by

$$P_{u,v} = \mathbb{P}(\mathbf{M}_{n+1} = v | \mathbf{M}_n = u) = \begin{cases} 1/d_u^- & \text{if } \exists \alpha \in G^1 : r(\alpha) = v, s(\alpha) = u \\ 0 & \text{otherwise.} \end{cases}$$

In the sequel, the graphs we shall consider are always transitive, row-finite, without multiple edges and without loops (of length 1). To simplify notation, we denote $\mathbb{V} \equiv G^0$ and $\mathbb{A} \equiv G^1 \subseteq \mathbb{V} \times \mathbb{V} \setminus \mathrm{diag}(\mathbb{V} \times \mathbb{V})$. The range and source functions for the edges are then naturally defined, and we denote the graph simply by $\mathbb{G} = (\mathbb{V}, \mathbb{A})$.

Definition 2.3. Let $\mathbb{V} = \mathbb{V}_1 \times \mathbb{V}_2 = \mathbb{Z}^2$, with $\mathbb{V}_1$ and $\mathbb{V}_2$ isomorphic to $\mathbb{Z}$ and $\boldsymbol{\epsilon} = (\epsilon_y)_{y\in\mathbb{V}_2}$ be a sequence of $\{-1, 1\}$-valued variables assigned to each ordinate. The *$\boldsymbol{\epsilon}$-horizontally oriented lattice* $\mathbb{G} = \mathbb{G}(\mathbb{V}, \boldsymbol{\epsilon})$ is the directed graph with the vertex set $\mathbb{V} = \mathbb{Z}^2$ and the edge set $\mathbb{A}$ defined by the condition that $(\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{A}$ if and only if $\boldsymbol{u}$ and $\boldsymbol{v}$ are distinct vertices such that

(i) either $v_1 = u_1$ and $v_2 = u_2 \pm 1$,

(ii) or $v_2 = u_2$ and $v_1 = u_1 + \epsilon_{u_2}$.

Example 2.4 (The alternate lattice $\mathbb{L}$). In that case, $\boldsymbol{\epsilon}$ is the deterministic sequence $\epsilon_y = (-1)^y$ for $y \in \mathbb{V}_2$. A part of this graph is presented in Figure 1.

Example 2.5 (The half-plane one-way lattice $\mathbb{H}$). Here $\boldsymbol{\epsilon}$ is the deterministic sequence

$$\epsilon_y = \begin{cases} 1 & \text{if } \ y \geq 0 \\ -1 & \text{if } \ y < 0. \end{cases}$$

A part of this graph is presented in Figure 2.

Example 2.6 (The lattice with random horizontal orientations $\mathbb{O}_{\boldsymbol{\epsilon}}$). Here $\boldsymbol{\epsilon}$ is a Rademacher sequence, i.e., it consists of $\{-1, 1\}$-valued symmetric Bernoulli random variables which are independent for different values of y. Figure 3 depicts a part of a realisation of this graph. The random sequence $\boldsymbol{\epsilon}$ is also termed the *environment of random horizontal directions*.

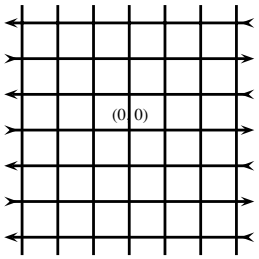


Figure 1. The alternately directed lattice $\mathbb{L}$ corresponding to the choice $\epsilon_y = (-1)^y$.

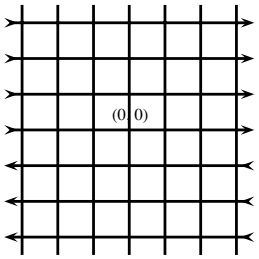


Figure 2. The half-plane one-way lattice $\mathbb{H}$ with $\epsilon_y = -1$, if $y < 0$ and $\epsilon_y = 1$, if $y \geq 0$.

Now we can state our main results.

Theorem 2.7. *The simple random walk on the alternate lattice $\mathbb{L}$ is recurrent.*

Remark 2.8. This result can be easily generalised to any lattice with periodically alternating horizontal directions (for every finite period).

Theorem 2.9. *The simple random walk on the half-plane one-way lattice $\mathbb{H}$ is transient.*

Remark 2.10. The result concerning transience in Theorem 2.9 is robust. In particular, perturbing the orientation of any finite set of horizontal lines either by reversing the orientation of these lines or by transforming them into two-ways does not change the transient behaviour of the simple random walk. Therefore, the half-plane one-way lattice is so deeply in the transience region that the asymptotic behaviour of the simple random walk cannot be changed by simply modifying the transition probabilities along a lower dimensional manifold as was the case in [32] where the bulk behaviour is on the critical point and it can be changed by lower-dimensional perturbations.

Theorem 2.11. *For almost all realisations of the environment $\boldsymbol{\epsilon}$, the simple random walk on the randomly horizontally oriented lattice $\mathbb{O}_{\boldsymbol{\epsilon}}$ is transient, and its speed is* 0.

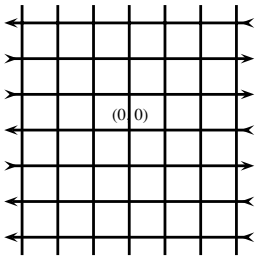


Figure 3. The randomly horizontally directed lattice $\mathbb{O}_{\boldsymbol{\epsilon}}$ with $(\epsilon_y)_{y\in\mathbb{Z}}$ an independent and identically distributed sequence of Rademacher random variables.

3 The physical relevance of random walks on undirected lattices

3.1 A brief history

The original impetus for the study of the continuous time analogue of a random walk, the Brownian motion, was given by the seminal work of Einstein [14] on diffusions[1].

The discrete time process, we nowadays call a simple random walk, was first studied by Pólya [36]. It is remarkable however that, contrary to Brownian motion whose physical motivation lies in the very definition of the model, the intrinsic physical relevance of random walks was discovered much later, with the development of polymer physics [16]. Polymers are long, *topologically* one dimensional molecules, composed by repeating several times (typically 100–10000) the same structural unit. The structural unit can be viewed as a small straight and rigid segment that can be glued with subsequent segments by loose bonds in such a way that if the first segment is held fixed, the second segment forms with the previous one a given angle θ (depending only on the chemical nature of the molecule), but otherwise its position is arbitrary. Assuming that the structural units have unit length, they merely represent directions $\hat{x} \in \mathbb{S}^2$. Hence, denoting $C(\hat{x}, e)$ the cone of opening 2θ, having its apex at the endpoint e of a segment, and its axis collinear with $\hat{x}$, the second segment will lie on an arbitrary separatrix of $C(\hat{x}, e)$. One thus immediately recognises a $\mathbb{R}^3$-valued discrete time random process $(S_n)_{n\in\mathbb{N}}$ with $S_0 = 0$, $S_1 = \xi_1$, and for $n \geq 2$, $S_n = S_{n-1} + \xi_n$ with ξ_1 uniformly distributed in $\mathbb{S}^2(0, 1)$, and ξ_n uniformly distributed in $\mathbb{S}^2(S_{n-1}, 1) \cap C(\xi_{n-1}, S_{n-1})$, where $\mathbb{S}^2(x, r)$ is the sphere of center x and radius r. Here the "time" n indexing the process corresponds to the order of appearance of a given monomer inside the macromolecular chain. For any bounded measurable

[1]The commented scientific biography of Einstein [33], which exposes the main ideas in an informal but fascinating style, is more accessible than the original paper [14]

function $f : \mathbb{R}^3 \to \mathbb{R}$, we have then

$$\mathbb{E}(f(S_n)|\mathcal{F}_{n-1}) = \int_0^{2\pi} f(S_{n-1} + \sin\theta\cos\phi e_1 + \sin\theta\sin\phi e_2 + \cos\theta e_3))\frac{d\phi}{2\pi},$$

where $\mathcal{F}_n = \sigma(S_k, k \leq n)$ and e_1, e_2, e_3 is the canonical basis of $\mathbb{R}^3$. A natural simplification of the model consists in considering a sequence (ξ_n) of independent and identically distributed random variables, getting thus a simple random walk on $\mathbb{R}^3$, an object that has been extensively studied and generalised in various respects and especially on non-commutative groups. It is not the intention of the authors to report further in this direction since it is perfectly well known by the community of probabilists and excellent monographs have been devoted to the subject (for instance, see [40], [39], [45]), but to report on some aspects developed mainly by physicists and less known to probabilists.

The model of simple random walk is too naïve to realistically model physical polymers: two different atoms cannot occupy the same position. Hence a realistic model must be self-avoiding, spoiling thus the Markovian character of the process. Consider the simplest random walk on $\mathbb{Z}^d$, the nearest neighbour random walk, i.e., let $E_d = \{\pm e_1, \ldots, \pm e_d\}$, where $(e_i)_{i=1,\ldots,d}$ denote the standard basis of $\mathbb{Z}^d$, and let $(\xi_n)_{n\in\mathbb{N}}$ be an independent and identically uniformly distributed sequence of E_d-valued random variables. Then the process defined by $S_0 = 0$ and $S_n = S_{n-1} + \xi_n$ for $n \geq 1$, provides the Markovian description of the ordinary random walk. If we are interested only in a finite sequence $(S_n)_{n=0}^N$, an equivalent description is provided by the trajectory space

$$\Omega_N = \{\omega : \{0, \ldots, N\} \to \mathbb{Z}^d \mid \omega(0) = 0, \omega(i) - \omega(i-1) \in E_d, i = 1, \ldots, N\},$$

equipped with the uniform probability measure $\mu_N(\omega) = 1/c_N$ for all $\omega \in \Omega_N$, where $c_n = \mathrm{card}\Omega_N = (2d)^N$. For $k : 0 \leq k \leq N$, the canonical projection $S_k(\omega) = \omega_k$ has the same law as the random walk defined by the sum $\sum_{i=1}^k \xi_i$, showing thus the equivalence of the Markovian and trajectorial descriptions for simple random walks. Adding the self-avoiding condition can be performed on the trajectorial description but not on the Markovian one. More precisely, let

$$\Omega_N^{\mathrm{saw}} = \{\omega \in \Omega_N : \omega(i) \neq \omega(j), \text{ for } 0 \leq i < j \leq N\},$$

and $\mu_N^{\mathrm{saw}}(\omega) = 1/c_N^{\mathrm{saw}}$ for all $\omega \in \Omega_N^{\mathrm{saw}}$ with $c_N^{\mathrm{saw}} = \mathrm{card}\Omega_N^{\mathrm{saw}}$. Notice however that the numerical sequence $(c_N^{\mathrm{saw}})_{N\in\mathbb{N}}$ is not explicitly known for $d \geq 2$ hence the model of self-avoiding random walk has been so far intractable. Nevertheless, the sequence of probability measures $(\mu_N^{\mathrm{saw}})_{N\in\mathbb{N}}$ – the probability μ_N^{saw} being defined on Ω_N^{saw} for every $N \in \mathbb{N}$ – is perfectly well-defined.

Instead of defining μ_N^{saw} on Ω_N^{saw} we can also define it on Ω_N by

$$\mu_N^{\mathrm{saw}}(d\omega) = \frac{1}{Z_N^{\mathrm{saw}}}\,\mathrm{Id}_{\Omega_N^{\mathrm{saw}}}(\omega)\mu_N(d\omega), \tag{3.1}$$

where Z_N^{saw} is a normalising factor, in fact $Z_N^{\text{saw}} = c_N^{\text{saw}}/(2d)^N$. Physicists have introduced various approximations to deal with the untractable measure μ_N^{saw}. One of them consists in approximating the indicator appearing in the previous formula and defining

$$\mu_{N,\beta}(d\omega) = \frac{1}{Z_n(\beta)} \exp(-\beta H_N(\omega))\mu_N(d\omega), \tag{3.2}$$

where $H_N(\omega) = \text{card}\{k : 2 \le k \le N | \exists j < k \text{ with } \omega(j) = \omega(k)\}$ (or some variant of this number) counts the self-intersections of the trajectory $\omega \in \Omega_N$, the real parameter β is non-negative, and $Z_N(\beta)$ is a normalising factor, in fact $Z_N(\beta) = \int_{\Omega_N} \exp(-\beta H_N(\omega))\mu_N(d\omega)$. The continuous version of this model has been introduced by Edwards in [13]. The discrete version, defined in (3.2), is known as discrete Edwards random walk or weakly self-avoiding random walk. The reason for this terminology is the following. For fixed N, we have $\lim_{\beta\to 0} \mu_{N,\beta} = \mu_N$ and $\lim_{\beta\to\infty} \mu_{N,\beta} = \mu_N^{\text{saw}}$, so that the weakly self-avoiding walk interpolates between ordinary and self-avoiding random walks. A much more difficult limit to study is $N \to \infty$ for fixed $\beta \in (0, \infty)$; this limit is dimension dependent and several authors have contributed to its study [41, 7, 43, 43, 28, 27, 19, 30, 18, 42, 8, 20, 24, 12, 22], either rigorously or numerically[2].

Formula (3.2) is interesting far beyond its application to the study of self-avoiding random walks since it is reminiscent of ideas at the basis of the Gibbs formulation of statistical mechanics and quantum field theory. The Radon–Nikodým derivative $d\mu_{N,\beta}/d\mu_N = \exp(-\beta H_N(\omega))/Z_N(\beta)$ is interpreted as a Boltzmann factor, making more rare trajectories with many self-intersections in the statistical sample described by the measure μ_N. The limit $N \to \infty$ is also a standard procedure in statistical mechanics and quantum field theory known as thermodynamic limit or infrared limit, respectively. Thus, this formula is naturally connecting random walks and various other physical theories defined on an infinite graph for which the fundamental quantities can be written as (formal) random walk expansions. In some particular situations, these formal expansions converge; they provide therefore a valuable probabilistic tool for the study of the asymptotic behaviour of physical models.

3.2 An example of random walk expansion

It is immediate to see that the Markov operator for a simple random walk on an undirected graph is essentially the discrete Laplacian on the graph. Random walk expansions can take a very sophisticated formulation; all of them can be seen however as (non-trivial) generalisations of a very simple formula of inversion of the Markov operator that connects the Green function of the regularised Laplacian with a power series on random walks of arbitrary length given in the following

[2]The day of submission of the present contribution, we learnt about a new result on weakly self-avoiding walks [3].

Lemma 3.1. *Let Δ be the difference Laplacian on $\mathbb{Z}^d$ and $m \neq 0$ a fixed parameter (the free mass of the quantum field theory). Then for any $x, y \in \mathbb{Z}^d$,*

$$(m^2 \cdot \mathrm{Id} - \Delta)^{-1}_{xy} = \sum_{\omega \in G^*(x,y)} (2d + m^2)^{-|\omega|},$$

where $G^*(x, y) = \bigcup_{n=0}^{\infty} G^n(x, y)$, and the set $G^n(x, y)$ is the set of paths of length n on the non-oriented graph $\mathbb{G}$ having $\mathbb{Z}^d$ as vertex set and with nearest neighbours vertices as edge set, that start at point x and end at point y.

Proof. Write $\Delta = J - 2d \cdot \mathrm{Id}$, where

$$J_{xy} = \begin{cases} 1, & \text{if } y - x \in E_d, \\ 0, & \text{otherwise,} \end{cases}$$

and develop $((m^2 + 2d)\,\mathrm{Id} - J)^{-1}$ as formal series in powers of J. For $m \neq 0$ the series converges defining thus the left-hand side of the formula. □

Consider now a general graph $\mathbb{G} = (G^0, G^1, r, s)$ where G^1 is a subset of $G^0 \times G^0 \setminus \mathrm{diag}(G^0 \times G^0)$ (i.e., the graph is simple and without loops); we assume moreover that the graph is undirected, i.e., if $(u, v) \in G^1$ then $(v, u) \in G^1$. Let J and L be $G^0 \times G^0$ matrices such that $J_{uv} > 0$ if $v \in N_u$ and $J_{uv} = 0$ otherwise, and $L_{uv} = \lambda_u \delta_{uv}$, with $\lambda_u > 0$ for all $u \in G^0$.

Lemma 3.2. *Suppose that the parameters $(\lambda_u)_{u \in G^0}$ are large enough. Then*

(i) $$(L - J)^{-1}_{uv} = \sum_{\omega \in G^*(u,v)} \prod_{a \in \omega} J_a \prod_{v \in G^0} \lambda_v^{-\eta_{|\omega|}(v,\omega)},$$
where $\eta_{|\omega|}(v, \omega) = \sum_{k=0}^{|\omega|} \mathrm{Id}_{\{v\}}(\omega_k)$ is the occupation time of the vertex v by the trajectory ω.

(ii) $$\det(L - J)^{-1} = \Big(\prod_{v \in G^0} \lambda_v \Big)^{-1} \exp\Big(\sum_{\omega \in \mathcal{L}^*} J_\omega \big(\prod_{u \in G^0} \lambda_u^{-\eta_{|\omega|}(u,\omega)} \big) \Big),$$
where $J_\omega = \prod_{a \in \omega} J_a$, and $\mathcal{L}^*$ is the set of loops of arbitrary length (i.e., equivalence classes of random walks of arbitrary length starting and ending on the same vertex).

Proof. The matrix L is diagonal and invertible. The formulae are again obtained by standard formal power series expansions that converge when λ_u are chosen sufficiently large (see [6] for details.) □

We shall always consider a simple undirected graph without loops $\mathbb{G} = (G^0, G^1, r, s)$ and let $f_v : \mathbb{R} \to \mathbb{R}$ be a family of maps indexed by $v \in G^0$ such that $\lim_{t \to \infty} f_v(t) \exp(ct) = 0$ for some $c > 0$. Denote by

$$\mathbb{X} = \{x : G^0 \to \mathbb{R}^\nu\} \simeq (\mathbb{R}^\nu)^{G^0}$$

the configuration space, so that each configuration $x \in \mathbb{X}$ is the collection $(x_v)_{v\in G^0}$ with $x_v \in \mathbb{R}^\nu$. The space $\mathbb{R}^\nu$ is equipped with its Borel σ-algebra $\mathcal{B}(\mathbb{R}^\nu)$. Let

$$\kappa_u(dx_u) = f_u(x^2)dx_u = f_u(x_{u,1}^2 + \cdots + x_{u,\nu}^2)dx_{u,1}\ldots dx_{u,\nu}$$

be a family of measures on $(\mathbb{R}^\nu, \mathcal{B}(\mathbb{R}^\nu))$ absolutely continuous with respect to the ν-dimensional Lebesgue measure. The configuration space can be naturally equipped with a product measure structure $\prod_{u\in G^0}\kappa_u$. However, a product measure structure is not very interesting for physical purposes since it corresponds to an infinite system of non-interacting components. To introduce some interaction, let J be an infinite $G^0 \times G^0$ matrix with $J_{uv} = J_{vu} > 0$ when $v \in N_u$ and $J_{uv} = 0$ otherwise. Since the matrix is symmetric, it defines a quadratic form on $\mathbb{X}$, known as the (formal) Hamiltonian

$$H(x) = -\frac{1}{2}\sum_{u,v\in G^0}(x_u, J_{uv}x_v) \equiv -\frac{1}{2}\sum_{u,v\in G^0}\sum_{\alpha=1}^{\nu} x_{u,\alpha}J_{uv}x_{v,\alpha}.$$

We can then define (formally) a probability μ on $(\mathbb{X}, \mathcal{F})$, where $\mathcal{F}$ is the natural σ-algebra, by

$$\mu(A) = \frac{1}{Z}\int_A \exp(-H(x))\prod_{v\in G^0}(f_v(x_v^2)dx_v)),$$

where Z is a normalising factor. There are various standard procedures to give a mathematical meaning to the above expressions. One of them is the following: suppose that the graph $\mathbb{G}$ is isometrically embedded in $\mathbb{R}^d$ for some d. Let $\Lambda_n = [-n, n]^d \cap G^0$ be the set of vertices of the graph inside the hypercube with edge length $2n + 1$. Then we define the finite volume Hamiltonian

$$H_n(x) = -\frac{1}{2}\sum_{u,v\in\Lambda_n}(x_u, J_{uv}x_v)$$

and the finite volume probability

$$\mu_n(A) = \frac{1}{Z_n}\int_A \exp(-H_n(x))\prod_{v\in\Lambda_n}(f_v(x_v^2)dx_v)),$$

where Z_n is the corresponding normalising factor. The sequence (μ_n) is a perfectly well defined sequence of probability measures. When this sequence converges weakly, we call the weak limit infinite volume Gibbs measure associated with the Hamiltonian H. Notice however that the existence of the weak limit is highly non trivial and it is granted only for some cases (for instance, see [41, 6, 17, 31]).

Theorem 3.3 (Symanzik). *Assume that the Hamiltonian is such that the weak limit μ exists. For $p = 1, 2, \ldots$ let $\{v_1, \ldots, v_{2p}\}$ be a given set of vertices; partition this set into p disjoint pairs. For each such pair of vertices let $\omega^{(l)}$ be a random walk of*

arbitrary length starting at one vertex and ending at the other vertex of the pair. Then

$$\mu(x_{v_1,\alpha_1}\dots x_{v_{2p},\alpha_{2p}}) \equiv \int x_{v_1,\alpha_1}\dots x_{v_{2p},\alpha_{2p}}\mu(dx) = \sum_{\omega^{(1)},\dots,\omega^{(p)}} \frac{W(\omega^{(1)},\dots,\omega^{(p)})}{Z},$$

where the sum extends over all partitions of the set of vertices and all random walk defined on the p pairs, and

$$W(\omega^{(1)},\dots,\omega^{(p)}) = \sum_{n=0}^{\infty}\frac{1}{n!}\left(\frac{\nu}{2}\right)^n \sum_{\gamma^{(1)},\dots,\gamma^{(n)}\in\mathcal{L}^*}\prod_{m=1}^{n} J_{\gamma^{(m)}}\prod_{l=1}^{p} J_{\omega^{(l)}} \exp(-U(\gamma^{(1)},\dots,\gamma^{(n)},\omega^{(1)},\dots,\omega^{(p)})),$$

where $\mathcal{L}^*$ denotes the set of loops of arbitrary length. Moreover, the normalising factor is

$$Z = \sum_{n=0}^{\infty}\frac{1}{n!}\left(\frac{\nu}{2}\right)^n \sum_{\gamma^{(1)},\dots,\gamma^{(n)}\in\mathcal{L}^*}\prod_{m=1}^{n} J_{\gamma^{(m)}} \exp(-U(\gamma^{(1)},\dots,\gamma^{(n)})),$$

and the mapping U defined on an arbitrary Cartesian product of random walks is given by

$$\exp(-U(\omega^{(1)},\dots,\omega^{(k)})) = \prod_v \int_\Gamma \hat{f}_v(z_v)(2iz_v)^{-h_v(\omega^{(1)},\dots,\omega^{(k)})}dz_v,$$

$\hat{f}$ being the Fourier transform of f_v,

$$h_v(\omega^{(1)},\dots,\omega^{(k)}) = \sum_{l=1}^{k}\eta_{|\omega^{(l)}|}(v,\omega^{(l)}) + \frac{\nu}{2},$$

and Γ an appropriately chosen integration contour.

Remark 3.4. The previous theorem looks formidable. It is worth noticing however that it is nothing more than a clever combinatorial recombination of terms appearing in the power series expansion of the exponential and that far-reaching results in quantum field theory [1, 15] – impossible or much more difficult to obtain otherwise – are obtained by the random walk representation it provides. Variants of this representation – known under the generic name of cluster expansions or abstract polymer models [31, 17] – are used in many different contexts, like statistical mechanics, disordered systems, etc., and in the cases it can be rigorously applied it provides a very powerful probabilistic tool for the study of covariance properties of limit Gibbs measures.

4 The physical relevance of random walks on oriented lattices

For any simple directed graph without loops $\mathbb{G} = (G^0, G^1, r, s)$ we define two operators [34], $D : l^2(G^0) \to l^2(G^1)$ by

$$(Df)(a) = f(s(a)) - f(r(a)) \quad \forall a \in G^1,$$

and its adjoint $D^* : l^2(G^1) \to l^2(G^0)$ by

$$(D^*\phi)(v) = \sum_{a \in s^{-1}(v)} \phi(a) - \sum_{a \in r^{-1}(v)} \phi(a) \quad \forall v \in G^0.$$

Then

$$(-D^*Df)(v) = -d_v f(v) + \sum_{u \in N_v} f(u) \equiv (\Delta f)(v).$$

For the simple random walk on $\mathbb{G}$, defined by its stochastic matrix $(P(u, v))_{u,v \in G^0}$, the operator M defined on bounded functions f by

$$Mf(u) = \sum_{u \in G^0} P(u, v) f(v) - f(u)$$

can not be expressed in terms of the Laplacian (contrary to the case of unoriented graphs where it is expressed in terms of the Laplacian), because

$$Mf(u) = \frac{1}{d_u^-} \sum_{a \in s^{-1}(v)} Df(a) \neq \frac{1}{d_u^-} (\Delta f)(u).$$

As a matter of fact, M is (roughly) reminiscent of the Dirac operator on the graph, providing thus the first hint that random walks on oriented lattices are relevant for non-commutative geometry.

4.1 A C^*-algebraic description of oriented lattices

With every oriented graph we can associate a C^*-algebra of operators, known as the Cuntz-Krieger algebra [11] of the graph [25, 26].

Let $(V_i)_{i \in I}$ be a finite or denumerable family of non-zero partial isometries and A an $I \times I$ $\{0, 1\}$-valued matrix whose rows contain a finite number of 1's. The Cuntz-Krieger algebra $\mathcal{O}_A$ associated with the matrix A is the C^*-algebra defined by the relations

$$V_i^* V_i = \sum_{j \in I} A_{ij} V_j V_j^*, \qquad i \in I.$$

The algebra $\mathcal{O}_A$ can be connected with oriented graphs in the following way. Let $\mathbb{G} = (G^0, G^1, r, s)$ be a row-finite, locally finite graph, and consider the corresponding

path space G^* defined in Section 2. Let $(P_v)_{v\in G^0}$ be a set of mutually orthogonal projections and $(V_a)_{a\in G^1}$ a set of non-zero partial isometries satisfying the relations

$$V_a^* V_a = P_{r(a)} \quad \forall\, a \in G^1$$

and

$$P_v = \sum_{a\in s^{-1}(v)} V_a V_a^* \quad \forall\, v \in G^0.$$

Define the edge-matrix $(A_{\mathbb{G}}(a,b))_{a,b\in G^1}$ of the graph $\mathbb{G}$ as

$$A_{\mathbb{G}}(a,b) = \begin{cases} 1, & \text{if } r(a) = s(b), \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$P_{r(a)} = V_a^* V_a = \sum_{b\in G^1 : r(a)=s(b)} V_b V_b^* = \sum_{b\in G^1} A_{\mathbb{G}}(a,b) V_b V_b^*.$$

The G^1-indexed Cuntz-Krieger C^*-algebra $\mathcal{O}_{A_{\mathbb{G}}}$ is called the C^*-algebra of the graph $\mathbb{G}$ and is denoted $C^*(\mathbb{G})$. The partial isometries V_a defined on G^1 are naturally extended to the path space G^* for every $\alpha \in G^*$ by putting

$$V_\alpha = V_{a_1} \cdots V_{a_{|\alpha|}}.$$

Then we have

Theorem 4.1 (Kumjian, Pask, and Raeburn [25]). *Let $\{P_v, V_a;\, v \in G^0, a \in G^1\}$ be the Cuntz-Krieger algebra associated with $\mathbb{G}$, and let β and γ be arbitrary paths of G^*. Then*

$$V_\beta^* V_\gamma = \begin{cases} V_{\gamma'}, & \text{if } \gamma = \beta\gamma',\ \gamma' \ni G^0, \\ P_{r(\gamma)}, & \text{if } \gamma = \beta, \\ V_{\beta'}^*, & \text{if } \beta = \gamma\beta',\ \beta' \neq G^0, \\ 0, & \text{otherwise.} \end{cases}$$

This construction associates with every path in G^* an operator of $C^*(\mathbb{G})$. Therefore the C^*-algebra $C^*(\mathbb{G})$ can be thought of as the non-commutative analogue of the subshift space [5], corresponding to the matrix $A_{\mathbb{G}}$. It is worth noticing that the Cuntz-Krieger algebras also arise in various other situations, like wavelets [4], tilings [23, 2], generalised sub-shifts [5], non-commutative geometry [10], etc.

4.2 Non-reversible evolution of quantum states

It has been convincingly argued lately that the progress in semiconductor technology will soon reach the limits of applicability of classical reasoning used in information

theory and computer science because quantum effects will start to be determining [44, 38, 21, 35].

In the classical physics a microscopic state of a multi-component system is described as an element of the Cartesian product of state spaces of individual components; for instance, in order to determine the microstate of a litre of gas, it is necessary to know the precise positions and momenta of all its molecules. The set of all microstates is called *configuration space*. Macroscopic states are classically probability measures on the configuration space and physical observables are bounded measurable real-valued functions on the configuration space. Time evolution is implemented as a Markov semi-group acting on macrostates; it can be reversible when the stochastic kernel of the semigroup is deterministic, or irreversible in general.

In the quantum physics the configuration space of microstates is a complex separable Hilbert space $\mathcal{H}$ (in general infinite dimensional but finite dimensional spaces are also of interest.) The macrostates are self-adjoint, positive operators with unit trace that are projections. Such operators are usually called *density matrices*. Observables are self-adjoint bounded operators acting on $\mathcal{H}$. Time evolution is implemented by a completely positive transformation on the set of density matrices $\rho \to \sum_{i\in I} T_i \rho T_i^*$. Such evolutions can be reversible – when $I = \{0\}$ is a singleton and T_0 is unitary – or irreversible – when we only require that $\sum_{i\in I} T_i T_i^* \leq \mathrm{id}_{\mathcal{H}}$. One immediately remarks that C^*-algebras provide a unified approach to both classical (Abelian ones) and quantum (non Abelian ones) systems.

In the context of applications in quantum information and communication, unitary transformations correspond to quantum logical gates, while irreversible ones to noisy transmission through quantum channels or to measurements. When a quantum macrostate ρ is transmitted through a noisy channel, different operators from the family $(T_i)_{i\in I}$ will sequentially act on ρ to get $\rho \to \hat{\rho}_n = \sum_{i_1,\ldots,i_n} T_{i_n} \ldots T_{i_1} \rho T_{i_1}^* \ldots T_{i_n}^*$. It is thus clear that products of operators will appear in the form of products of partial isometries along paths of oriented graphs. The main difference is that we have not required here the evolution to be implemented by partial isometries but by general non-commuting operators satisfying only $\sum_{i\in I} T_i T_i^* \leq \mathrm{id}_{\mathcal{H}}$. Nevertheless the analogy can be made complete by virtue of the following result:

Theorem 4.2 (Popescu [37]). *For every sequence (finite or denumerable) $(T_i)_{i\in I}$ of non-commuting operators acting on a Hilbert space $\mathcal{H}$, such that $\sum_{i\in I} T_i T_i^* \leq \mathrm{id}_{\mathcal{H}}$, there exists a minimal isometric dilation into operators $(V_i)_{i\in I}$ acting on a Hilbert space $\mathcal{K} \supset \mathcal{H}$, uniquely determined up to isomorphisms, such that the family $(V_i)_{i\in I}$ is composed by non-zero partial isometries on $\mathcal{K}$.*

Now it is evident that random walks on a (partially) directed lattice induce a random walk on the space of density matrices; recovering information from the perturbed macrostate $\hat{\rho}_n$ will be better when $\hat{\rho}_n$ is close (with respect to an appropriate topology) to ρ. This remark was made for the first time in [29] and gives a practical physical and technological relevance to questions of recurrence of random walks on randomly oriented lattices.

5 Sketch of the proofs

The idea of the proof of Theorems 2.7, 2.9, and 2.11 is to decompose the Markov chain $(\mathbf{M}_n)$ into a vertical skeleton (Y_n) and a horizontal embedded random walk (X_n) that – when sampled on a particular sequence of random times defined in terms of the vertical skeleton – has the same recurrence/transience properties as the original random walk $(\mathbf{M}_n)$.

Let $(\psi_n)_{n\in\mathbb{N}_+}$ be a sequence of independent, identically distributed, $\{-1, 1\}$-valued symmetric Bernoulli variables. Let

$$\begin{cases} Y_0 = 0, \\ Y_n = \sum_{k=1}^n \psi_k, \quad n = 1, 2, \dots, \end{cases}$$

be the simple $\mathbb{V}_2$-valued symmetric one-dimensional random walk. We call the process $(Y_n)_{n\in\mathbb{N}}$ the *vertical skeleton*. We denote by

$$\eta_n(y) = \sum_{k=0}^{n} \mathrm{Id}_{\{Y_k=y\}}, \quad n \in \mathbb{N}, y \in \mathbb{V}_2$$

its *occupation time* at level y. If $\mathcal{F}_n = \sigma(\psi_i, i \le n)$ then $\eta_n(y)$ is obviously $\mathcal{F}_n$-measurable.

Define $\sigma_0 = 0$ and recursively, for $n = 1, 2, \dots, \sigma_n = \inf\{k \ge \sigma_{n-1} : Y_k = 0\} > \sigma_{n-1}$, the n^{th} *return to the origin* for the vertical skeleton. It is known by the standard theory of simple symmetric one-dimensional random walk that almost surely $\sigma_n < \infty$ for all n.

To define the horizontal embedded random walk, let $(\xi_n^{(y)})_{n\in\mathbb{N}_+, y\in\mathbb{V}_2}$ be a doubly infinite sequence of independent identically distributed $\mathbb{N}$-valued geometric random variables of parameters $p = 2/3$ and $q = 1 - p = 1/3$, i.e., $\mathbb{P}(\xi_1^y = \ell) = pq^\ell, \quad \ell = 0, 1, 2, \dots$. Let moreover

$$T_n = n + \sum_{y\in\mathbb{V}_2} \sum_{i=1}^{\eta_{n-1}(y)} \xi_i^{(y)}$$

be the instant just after the random walk $(\mathbf{M}_k)$ has performed its n^{th} vertical move (with the convention that the sum $\sum_{i=1}^{\eta_{n-1}(y)}$ vanishes whenever $\eta_{n-1}(y) = 0$.) Then

$$\mathbf{M}_{T_n} = (X_n, Y_n),$$

where (Y_n) is the vertical skeleton and

$$X_n = \sum_{y\in\mathbb{V}_2} \epsilon_y \sum_{i=1}^{\eta_{n-1}(y)} \xi_i^{(y)}$$

represents the total horizontal displacement when the random walk $(\mathbf{M}_k)$ has completed exactly n vertical moves. Notice also that the horizontal embedded random walk

(X_n) can be viewed as a random walk with unbounded jumps in a random scenery, and that $\mathbf{M}_{T_{\sigma_n}} = (X_{\sigma_n}, 0)$. Now $(\mathbf{M}_n)$ can return to the origin if and only if both vertical and horizontal components vanish simultaneously. Since the vertical component can vanish only at instants $\sigma_n, n \in \mathbb{N}$, we prove in [9] the following

Lemma 5.1. (i) $\sum_{n=0}^{\infty} \mathbb{P}(X_{\sigma_n} = 0) = \infty$ *if and only if the random walk* $(\mathbf{M}_l)$ *is recurrent.*

(ii) $\sum_{n=0}^{\infty} \mathbb{P}(X_{\sigma_n} = 0) < \infty$ *if and only if the random walk* $(\mathbf{M}_l)$ *is transient.*

Introduce now the characteristic function

$$\chi(\theta) = \mathbb{E}\exp(i\theta\xi_1^{(0)}) = \frac{p}{1 - q\exp(i\theta)} = r(\theta)\exp(i\alpha(\theta))$$

and observe that r is an even function of θ while α is odd. Hence, denoting $\mathcal{F} = \mathcal{F}_\infty$ and $\mathcal{G} = \sigma(\epsilon_y, y \in \mathbb{V}_2)$, we have

$$\begin{aligned}
\mathbb{E}\exp(i\theta X_n) &= \mathbb{E}\Big(\mathbb{E}(\exp(i\theta \sum_{y\in\mathbb{V}_2} \epsilon_y \sum_{i=1}^{\eta_{n-1}(y)} \xi_i^{(y)} | \mathcal{F} \vee \mathcal{G})\Big) \\
&= \mathbb{E}\Big(\prod_{y\in\mathbb{V}_2} \chi(\theta\epsilon_y)^{\eta_{n-1}(y)}\Big) \\
&= \mathbb{E}(r(\theta)^{\sum_{y\in\mathbb{V}_2}\eta_{n-1}(y)}\exp(i\alpha(\theta)\Delta_n)) \\
&= r(\theta)^n \mathbb{E}(\exp(i\alpha(\theta)\Delta_n)),
\end{aligned}$$

where $\Delta_n = \sum_{y\in\mathbb{V}_2} \epsilon_y \eta_{n-1}(y)$.

Lemma 5.2. *For the* $\mathbb{L}$ *lattice,* $\Delta_{\sigma_n} = 0$.

Proof. This is an elementary combinatorial result a complete proof of which is given in [9]. □

The proof of Theorem 2.7 follows now immediately, since

$$\mathbb{E}\exp(i\theta X_{\sigma_n}) = \mathbb{E}(r(\theta)^{\sigma_n}) = (\sqrt{1 - r(\theta)^2})^n.$$

Hence

$$\mathbb{P}(X_{\sigma_n} = 0) = \lim_{\epsilon\to 0} 2\int_\epsilon^\pi \frac{1}{\sqrt{1 - r(\theta)^2}} d\theta = \infty.$$

Remark 5.3. Notice that for the random walk on the $\mathbb{L}$ lattice, various other more elegant proofs can be given; we presented the most elementary one.

For the corresponding result on the $\mathbb{H}$ lattice, the proof is based on the following lemma shown in [9]

Lemma 5.4. *Denote by $(\rho_k)_{k\in\mathbb{N}}$ a sequence of independent identically distributed Rademacher variables and by $(\tau_k)_{k\in\mathbb{N}}$ a sequence of independent, identically distributed random variables, independent of the sequence $(\rho_k)_{k\in\mathbb{N}}$, such that $\tau_1 \stackrel{d}{=} \sigma_1$, i.e., the random variables τ_k have the same law as the time of the first return to the origin for the skeleton random walk. Then*

$$\Delta_{\sigma_n} \stackrel{d}{=} \sum_{k=1}^{n} \rho_k(\tau_k - 1) + n.$$

The proof of Theorem 2.9 follows then from the equality $\mathbb{E}\left(\exp(i\theta X_{\sigma_n})\right) = g(\theta)^n$, where

$$g(\theta) = \frac{1}{2}\chi(\theta)\left[\left(1 - \sqrt{1-\chi(\theta)^2}\right)\exp(-i\alpha(\theta)) + \left(1 - \sqrt{1-\overline{\chi}(\theta)^2}\right)\exp(i\alpha(\theta))\right].$$

This expression allows an explicit estimate for $\sum_{n=1}^{\infty}\mathbb{P}(X_{\sigma_n}=0)$ that converges in the present case, establishing transience of the random walk.

Finally for the $\mathbb{O}_\epsilon$ lattice, no simple decomposition can be made and we need joint estimates. The idea of the proof is to decompose the probability $p_n = \mathbb{P}(X_{2n} = 0, Y_{2n} = 0)$ into $p_n = p_{n,1} + p_{n,2} + p_{n,3}$, where

$$\begin{aligned}
p_{n,1} &= \mathbb{P}(I(X_{2n}, -\epsilon_0 Z) \ni 0; Y_{2n} = 0; B_n),\\
p_{n,2} &= \mathbb{P}(I(X_{2n}, -\epsilon_0 Z) \ni 0; Y_{2n} = 0; A_n \setminus B_n),\\
p_{n,3} &= \mathbb{P}(I(X_{2n}, -\epsilon_0 Z) \ni 0; Y_{2n} = 0; A_n^c),
\end{aligned}$$

and $A_n = A_{n,1} \cap A_{n,2}$ with

$$\begin{aligned}
A_{n,1} &= \left\{\omega\in\Omega : \max_{0\le k\le 2n}|Y_k| < n^{\frac{1}{2}+\delta_1}\right\} \text{ for some } \delta_1 > 0,\\
A_{n,2} &= \left\{\omega\in\Omega : \max_{y\in\mathbb{V}_2}\eta_{2n-1}(y) < n^{\frac{1}{2}+\delta_2}\right\} \text{ for some } \delta_2 > 0,\\
B_n &= \left\{\omega\in A_n : \left|\sum_{y\in\mathbb{V}_2}\epsilon_y\eta_{2n-1}(y)\right| > n^{\frac{1}{2}+\delta_3}\right\} \text{ for some } \delta_3 > 0.
\end{aligned}$$

The technical part of the proof consists in showing that $p_{n,1}$ and $p_{n,3}$ are of order $\mathcal{O}(\exp(-n^\delta))$ with some $\delta > 0$ for large n while the main part of the mass charging the event $\{X_{2n} = 0, Y_{2n} = 0\}$ is supported by the set $A_n \setminus B_n$. More precisely, it is shown in [9] that on the set $A_n \setminus B_n$ we have $\mathbb{P}(X_{2n} = 0|\mathcal{F}\vee\mathcal{G}) = \mathcal{O}\left(\sqrt{\frac{\ln n}{n}}\right)$, while $\mathbb{P}(A_n \setminus B_n|\mathcal{F}\vee\mathcal{G}) = \mathcal{O}\left(n^{-1/4+\delta_4}\right)$, so that together with the standard estimate $\mathbb{P}(Y_{2n} = 0) = \mathcal{O}\left(1/\sqrt{n}\right)$ the series $\sum_n p_n$ is shown to converge.

Notice also that some additional results concerning the mean speed and law of large numbers are presented in [9].

References

[1] M. Aizenman and J. Fröhlich, Topological anomalies in the n dependence of the n-states Potts lattice gauge theory, Nuclear Phys. B 235 (1984), FS11, 1–18.

[2] J. Bellissard, R. Benedetti and J.-M. Gambaudo, Spaces of tilings, finite telescopic approximations and gap-labelling, preprint, arXiv:math.DS/0109062 (2001).

[3] E. Bolthausen and C. Ritzmann, A central limit theorem for convolution equations and weakly self-avoiding walks, preprint, arXiv:math.PR/0103218 (2001).

[4] O. Bratteli and P. E. T. Jorgensen, A connection between multiresolution wavelet theory of scale N and representations of the Cuntz algebra $\mathcal{O}_N$, in: Operator Algebras and Quantum Field Theory (Rome, 1996), International Press, Cambridge, MA, 1997, 151–163.

[5] O. Bratteli and P. E. T. Jorgensen, Iterated Function Systems and Permutation Representations of the Cuntz Algebra, Mem. Amer. Math. Soc. 139 (1999), no. 663.

[6] D. Brydges, J. Fröhlich and T. Spencer, The random walk representation of classical spin systems and correlation inequalities, Comm. Math. Phys. 83 (1982), 123–150.

[7] D. Brydges and T. Spencer, Self-avoiding walk in 5 or more dimensions, Comm. Math. Phys. 97 (1985), 125–148.

[8] B. Cadre, Une preuve standard du principe d'invariance de Stoll, in: Séminaire de Probabilités, XXXI, Lecture Notes in Math. 1655, Springer-Verlag, Berlin 1997, 85–102.

[9] M. Campanino and D. Petritis, Random walks on randomly oriented lattices, Markov Process. Related Fields 9 (2003), 391–412.

[10] A. Connes, Noncommutative Geometry, Academic Press, San Diego, CA, 1994.

[11] J. Cuntz and W. Krieger, A class of C^*-algebras and topological Markov chains, Invent. Math. 56 (1980), 251–268.

[12] Ph. de Forcrand, J. Pasche and D. Petritis, Critical behaviour of Edwards random walk in two dimensions: a case where the fractal and Hausdorff dimensions are not equal, J. Phys. A 21 (1988), 3771–3782.

[13] S. F. Edwards, The statistical mechanics of polymers with excluded volume. Proc. Phys. Soc. 85 (1965), 613–624.

[14] A. Einstein, Über die Theorie der Brownschewegung, Ann. der Physik, 19 (1906), 371–381.

[15] R. Fernández, J. Fröhlich and A. Sokal, Random Walks, Critical Phenomena and Triviality in Quantum Field Theory, Texts Monogr. Phys., Springer-Verlag, Berlin 1992.

[16] P. Flory, The configuration of real polymer chain, J. Chem. Phys. 17 (1949) 303–310.

[17] J. Glimm and A. Jaffe, Quantum Physics. A Functional Integral Point of View. Second edition, Springer-Verlag, New York 1987.

[18] N. Guillotin, Marches aléatoires à une dimension avec auto-intéraction, d'après H. Zoladek, in: Seminaires de Probabilites de Rennes (1995), Publ. Inst. Rech. Math. Rennes, 1995, Univ. Rennes I, Rennes 1995.

[19] T. Hara and G. Slade, Self-avoiding walk in five or more dimensions. I. The critical behaviour, Comm. Math. Phys. 147 (1992), 101–136.

[20] R. van der Hofstad, F. den Hollander and G. Slade, A new inductive approach to the lace expansion for self-avoiding walks, Probab. Theory Related Fields 111 (1998), 253–286.

[21] A. Holevo, Statistical Structure of Quantum Theory, Lecture Notes in Physics. Monographs 67, Springer-Verlag, Berlin 2001.

[22] I. Hueter, Proof of the conjecture that planar self-avoiding walk has root mean square displacement exponent 3/4, preprint (2001).

[23] J. Kellendonk, Noncommutative geometry of tilings and gap labelling, Rev. Math. Phys, 7 (1995), 1133–1180.

[24] F. Koukiou, J. Pasche and D. Petritis, The Hausdorff dimension of the two-dimensional Edwards' random walk, J. Phys. A 22 (1989), 1385–1391.

[25] A. Kumjian, D. Pask and I. Raeburn, Cuntz–Krieger algebras of directed graphs, Pacific J. Math. 184 (1998), 161–174.

[26] A. Kumjian, D. Pask, I. Raeburn and J. Renault, Graphs, groupoids, and Cuntz–Krieger algebras, J. Funct. Anal. 144 (1997), 505–541.

[27] G. F. Lawler, The infinite self-avoiding walk in high dimensions, Ann. Probab. 17 (1989), 1367–1376.

[28] G. F. Lawler, Intersections of Random Walks, Probab. Appl., Birkhäuser, Boston, MA, 1991.

[29] Ph. Leroux, Description algébrique des graphes orientés pondérés et apllications, Ph.D. Thesis, Université de Rennes I, 2002.

[30] N. Madras and G. Slade, The Self-Avoiding Walk, Probability and its Applications, Birkhäuser, Boston, MA, 1993.

[31] V. A. Malyshev, Cluster expansions in lattice models of statistical physics and quantum theory of fields, Russian Math. Surveys 35 (2) (1980), 1–62.

[32] M. Menshikov and D. Petritis, Markov chains in a wedge with excitable boundaries, preprint, Université Rennes-1 (2000).

[33] A. Pais, "Subtle is the Lord...". The science and life of Albert Einstein, The Clarendon Press, Oxford University Press, New York 1982.

[34] W. L. Paschke, The flow space of a directed G-graph, Pacific J. Math. 159 (1993), 127–138.

[35] D. Petritis, Introduction to quantum information and communication, in preparation (2002).

[36] G. Pólya, Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt in Straßennetz, Math. Ann. 84 (1921), 149–160.

[37] G. Popescu, Isometric dilations for infinite sequences of noncommuting operators, Trans. Amer. Math. Soc. 316(1989), 523–536.

[38] J. Preskill, Quantum computing: pro and con, in: Quantum Coherence and Decoherence (Santa Barbara, CA, 1996), R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci. 454 (1998), 469–486.

[39] D. Revuz, Markov Chains. Second edition, North-Holland Math. Library 11, North-Holland, Amsterdam 1984.

[40] F. Spitzer, Principles of Random Walks. Second edition, Grad. Texts in Math. 34, Springer-Verlag, New York 1976.

[41] K. Symanzik, Euclidean quantum field theory, in: Local Quantum Field Theory (Varenna 1968), Academic Press, New York 1969, 152–226.

[42] F. Vermet, Phase transition and law of large numbers for a non-symmetric one-dimensional random walk with self-interactions, J. Appl. Probab. 35 (1998), 55–63.

[43] M. J. Westwater, On Edwards' model for long polymer chains, Comm. Math. Phys. 72 (1980), 131–174.

[44] C. P. Williams and S. H Clearwater, Explorations in Quantum Computing, Springer-Verlag, New York 1998.

[45] W. Woess, Random Walks on Infinite Graphs and Groups, Cambridge Tracts in Math. 138, Cambridge University Press, Cambridge 2000.

Massimo Campanino, Dipartimento di Matematica, Università degli Studi di Bologna, piazza di porta San Donato 5, I-40126 Bologna, Italy
E-mail: campanin@dm.unibo.it

Dimitri Petritis, Institut de Recherche Mathématique de Rennes, Université de Rennes I and CNRS (UMR 6625), Campus de Beaulieu, F-35042 Rennes Cedex, France
E-mail: Dimitri.Petritis@univ-rennes1.fr

Random walks, entropy and hopfianity of free groups

*Tullio Ceccherini-Silberstein*and Fabio Scarabotti*

Abstract. In this note we present two proofs of Nielsen's theorem on the hopfianity of free groups. The first one is based on Kesten's characterization of free groups via random walks; the second one uses an entropy inequality proved with a method of Gromov.

1 Introduction: hopfianity

A group G is termed *hopfian* if it is not isomorphic to a proper quotient of itself, equivalently G is hopfian if any epimorphism $\phi : G \to G$ is an isomorphism.

In 1921 Nielsen [21, 19] showed that if $\mathbb{F}_n = F(x_1, x_2, \ldots, x_n)$ is the free group on generators $x_1, x_2, \ldots, x_n$, N is a normal subgroup, and the quotient $\mathbb{F}_n/N$ is a free group, then there exist (possibly new) free generators $y_1, y_2, \ldots, y_n$ for $\mathbb{F}_n$ such that N is the normal subgroup generated by $y_1, y_2, \ldots, y_k$ ($k \leq n$). From this he deduced that a proper quotient of $\mathbb{F}_n$ can be free of rank at most $n - 1$, thus showing

Theorem (Nielsen). *A free group of finite rank is hopfian.*

Different proofs of this theorem were given, later, by Magnus, Fuchs and Rabinovitch, Federer and Jónsson, Lyndon, and M. Hall (see [18, 19]).

There are large families of groups which are hopfian, for instance finitely generated, residually nilpotent groups, residually free groups, residually finite groups (a celebrated theorem of Mal'cev), fundamental groups of surfaces (G. Baumslag and, independently, Frederick), see [18, 19] or, more generally torsion-free *hyperbolic groups*, a remarkable result of Z. Sela [25].

In 1931 H. Hopf [12], motivated by investigations in a topological context, asked whether a finitely generated group could be isomorphic to a proper quotient of itself. This question was answered in the affirmative by B. H. Neumann in 1950 who presented a two-generator non-hopfian group with infinitely many relations (non finitely presentable); one year later Higman produced a non-hopfian three-generator group

*Partially supported by the Swiss National Science Foundation.

with two defining relations, and finally G. Baumslag and Solitar in 1962 presented the simplest example of a non–hopfian group: the two-generated one-relator group $G =< a, b : ab^2a^{-1} = b^3 >$ (see [18, 19]). We also remind the interesting examples of Meier [20] including that of a group isomorphic with its own square: $G \cong G \times G$.

We finally mention that in [8] non-hopfian groups are used in some constructions yielding new examples of groups with intermediate growth and potential examples of groups of exponential but not *uniformly-exponential* growth (see [7]).

Grigorchuk, in the middle of the 70s, motivated by his investigations on *cogrowth* [6], conjectured that a proof of Nielsen's theorem could be derived from Kesten's characterization of finitely generated free groups, in terms of the spectral radius of the simple random walk on their Cayley graphs. In §2, in accordance with Grigorchuk's intuition, we present our first proof of the hopfianity of free groups.

In §3 we determine a relation between the entropy of finitely-generated free groups and the entropy of their proper quotients: this is based on some techniques of Gromov [10] and avoids the Perron–Frobenius theory. Combining this relation with the *minimality* of the *uniform exponential growth* of free groups we obtain our second proof of Nielsen's Theorem.

2 Random walks

Given a discrete group G, denote by $\ell^2(G) = \{f : G \longrightarrow \mathbf{C} : \sum_{g\in G} |f(g)|^2 < \infty\}$ the Hilbert space of square summable complex functions on G. Suppose that G is finitely generated and let A be a finite and *symmetric* ($A = A^{-1}$) system of generators for G. The *Markov operator* $M = M(G, A) : \ell^2(G) \to \ell^2(G)$, defined as $M[f](g) = \frac{1}{|A|} \sum_{a\in A} f(ag),\ f \in \ell^2(G),\ g \in G$, is bounded and self-adjoint. Thus, its spectral radius $\sup\{|z| : M - z\mathbf{I}$ has no bounded inverse$\}$ coincides with its norm $\|M\|$: this common value, denoted by $\rho(G, A)$, is the *spectral radius of the simple random walk on G with respect to A*. This number also has the following probabilistic interpretation. Let $X(G, A)$ be the Cayley graph of G with respect to A, and denote by

$$p_n = \frac{\#\{\text{closed paths of base 1 of length } n\}}{\#\{\text{paths of length } n \text{ starting at } 1\}}$$

the probability that a random walker starting from 1 arrives back to 1 in n steps; then $\rho(G, A) = \limsup_{n\to\infty} \sqrt[n]{p_n}$.

We state together the fundamental results of Kesten from [13] and [14].

Theorem 2.1 (Kesten). *Let G be a finitely generated group and A a finite symmetric generating system.*

1. *Suppose that $N \trianglelefteq G$ is a normal subgroup and denote by $\overline{g} = g \bmod N \in \overline{G} = G/N$ the image of $g \in G$ under the canonical quotient homomorphism. Then $\overline{A}$ is a*

finite symmetric system of generators for $\overline{G}$*, and one has* $\rho(G, A) \leq \rho(\overline{G}, \overline{A})$ *with equality if and only if* N *is amenable.*

2. *Suppose that in* A *there are no involutions, so that* $A = A_+ \coprod (A_+)^{-1}$*, and set* $n = |A_+| = |A|/2$*. Then, the estimates* $\frac{\sqrt{2n-1}}{n} \equiv \frac{2\sqrt{|A|-1}}{|A|} \leq \rho(G, A) \leq 1$ *hold, with equality on the RHS if and only if* G *is amenable, and with equality on the LHS if and only if* G *is the free group* $\mathbb{F}_n$ *of rank* n *freely generated by* A_+*.*

Proof of Nielsen's Theorem via random walks. Let $\mathbb{F}_n$ be a free group and A_+ a free basis. Let $N \trianglelefteq F$ be a non–trivial normal subgroup: then N is free non-abelian and thus non-amenable. Suppose, by contradiction, that $\overline{\mathbb{F}_n} = \mathbb{F}_n/N$ is free of rank n. Then, with the above notation, we have that $\overline{A_+}$ is a free basis for the quotient group $\overline{\mathbb{F}_n}$ so that, setting $A = A_+ \coprod (A_+)^{-1}$, one has

$$\frac{\sqrt{2n-1}}{n} = \frac{2\sqrt{|A|-1}}{|A|} = \rho(\mathbb{F}_n, A) < \rho(\overline{\mathbb{F}_n}, \overline{A}) = \frac{2\sqrt{|\overline{A}|-1}}{|\overline{A}|} = \frac{\sqrt{2n-1}}{n},$$

where the inequality comes from Kesten's theorem, a contradiction. □

3 Entropy

Let G be a finitely generated group, and A be a finite and symmetric system of generators. The *A-length* of an element $g \in G$ is by definition $|g|_A = 0$ if $g = 1$, and the minimal length n of a word in the generators in A expressing g, namely $|g|_A = \min\{n \in \mathbb{N} : g = a_1 a_2 \dots a_n,\ \ a_i \in A\}$, otherwise.

The function $\sigma_A^G(n) = |\{g \in G : |g|_A = n\}|$ that counts the elements in G of A-length equal to n, is called the *spherical growth function* of G with respect to A. The *entropy* $\mathrm{ent}(G, A)$ of the couple (G, A) is the limit

$$\mathrm{ent}(G, A) = \lim_{n\to\infty} \frac{\log \sigma_A^G(n)}{n}$$

which always exists (by the Fekete–Polya lemma on subadditive sequences, see e.g. [11, 17]). The quantity $\omega_A(G) = \exp(\mathrm{ent}(G, A)) = \lim_{n\to\infty} \sqrt[n]{\sigma_A^G(n)}$ is usually called *the growth rate* of G with respect to A. In particular, G is said to have *exponential growth* if $\omega_A(G) > 1$ (equivalently $\mathrm{ent}(G, A) > 0$) for some (and therefore for any other) finite system of generators A; also

$$\mathrm{ent}(G) = \inf_A \mathrm{ent}(G, A), \tag{3.1}$$

where the infimum is taken over all finite generating systems, is the *entropy* of G. The group G has *uniform exponential growth* if $\omega(G) = \exp(\mathrm{ent}(G)) > 1$. This last concept is due to Avez [2], and it is discussed in [7, 9].

Remark 3.1. The function $\gamma_A^G(n) = |\{g \in G : |g|_A \leq n\}|$ that counts the elements in G of A-length less or equal to n, is called *the growth function* of (G, A). This is more used than the spherical growth function, but anyway they give the same growth rate, i.e. $\lim_{n\to\infty} \sqrt[n]{\gamma_A^G(n)} = \lim_{n\to\infty} \sqrt[n]{\sigma_A^G(n)}$; see [11].

The simplest example of a group with uniformly exponential growth is provided by the free group $\mathbb{F}_r$ of finite rank $r \geq 2$; indeed for these groups the infimum in (3.1) is even attained:

Proposition 3.2. *Let $\mathbb{F}_r$ be the free group of rank $r \geq 2$ and A_+ a free basis. Then* $\mathrm{ent}(\mathbb{F}_r) = \mathrm{ent}(\mathbb{F}_r, A_+ \coprod (A_+)^{-1}) = \log(2r-1)$*, and, in particular, $\mathbb{F}_r$ is of uniformly exponential growth.*

Proof. Set $A = A_+ \coprod (A_+)^{-1}$, where $A_+ = \{a_1, a_2, \ldots, a_k\}$ is a finite system of generators. Clearly $k \geq r = \mathrm{rank}\, \mathbb{F}_r$.

If $k = r$, then A_+ is a free basis and $\mathrm{ent}(\mathbb{F}_r, A) = \log(2r-1)$.

Suppose $k > r$ and denote by

$$\begin{array}{rccc} \pi: & \mathbb{F}_r & \longrightarrow & \mathbb{F}_r/[\mathbb{F}_r, \mathbb{F}_r] \cong \mathbb{Z}^r \\ & x & \longmapsto & \pi(x) = \overline{x} \end{array}$$

the canonical quotient homomorphism. Then $\overline{A_+} = \{\overline{a}_1, \overline{a}_2, \ldots, \overline{a}_k\}$, the image of A_+ via π, is a generating system for $\mathbb{Z}^r$ so that there exists $\{i_1, , i_2, \ldots, i_r\} \subset \{1, 2, \ldots, k\}$ such that $\{\overline{a}_{i_1}, \ldots, \overline{a}_{i_r}\}$ generates a subgroup $\overline{Q} \cong \mathbb{Z}^r$. Then, as subgroups of free groups are always free, $A'_+ := \{a_{i_1}, \ldots, a_{i_r}\}$ generates a subgroup $Q \cong \mathbb{F}_\ell$ of rank at most r (i.e., $\ell \leq r$). As

$$\mathbb{Z}^r \cong \overline{Q} = \pi(Q) \cong Q/([\mathbb{F}_r, \mathbb{F}_r] \cap Q) \text{ is a quotient of } Q/[Q, Q] \cong \mathbb{Z}^\ell,$$

we have $r = \ell$ and therefore, setting $A' = A'_+ \coprod (A'_+)^{-1}$, one has

$$\mathrm{ent}(\mathbb{F}_r, A) \geq \mathrm{ent}(Q, A') = 2r - 1$$

yielding $\mathrm{ent}(\mathbb{F}_r) = 2r - 1$. □

The following theorem is essentially a particular case of an entropic inequality in Symbolic Dynamics and in the Theory of Formal Languages. Indeed, the free group is the language of a symbolic system of finite type [17]. The proof presented here is a variation of a Lemma of Gromov [10]; see also [1, 3, 4, 5, 22, 23, 24]. Note that this approach also provides an explicit estimate of the decrease of entropy in terms of the rank of the ambient free group and of the length of a word in the normal subgroup.

Theorem 3.3. *Let $F = \mathbb{F}_r$ denote the free group of rank $r \geq 2$ freely generated by a set A_+. Set $A = A_+ \coprod (A_+)^{-1}$. Then for any nontrivial normal subgroup $N \trianglelefteq F$ one has* $\mathrm{ent}(\overline{F}, \overline{A}) < \mathrm{ent}(F, A)$*, where $F \ni g \longmapsto \overline{g} = g \bmod N \in \overline{F} = F/N$ denotes the canonical quotient homomorphism.*

Proof. Denote by $B_k(F) = \{g \in G : |g|_A = k\}$ the sphere of radius k (with respect to A) in F; similarly set $B_k(\overline{F})$ (with respect to $\overline{A}$). Let w be a nontrivial word in N, and let W_k denote the set of all words in $B_k(F)$ not containing w. Clearly, $|B_k(\overline{F})| \leq |W_k|$. We want to show that if $|w| = \ell$, setting $n = \ell + 2$, one has

$$|W_{kn}| \leq \left(1 - \frac{4(r-1)^2}{(2r-1)^n}\right)^k |B_{kn}(F)| \tag{3.2}$$

for any $k \in \mathbb{N}$. Denote by π_h the set of all words $a_1 a_2 \dots a_{kn} \in B_{kn}(F)$ such that $a_{(h-1)n+2} a_{(h-1)n+3} \dots a_{hn-1} = w$. Each word $a_{n+1} a_{n+2} \dots a_{kn} \in B_{(k-1)n}(F)$ may be extended to a word $a_1 w a_n a_{n+1} a_{n+2} \dots a_{kn} \in \pi_1$ in at least $4(r-1)^2$ ways (the possible choices of a_1 and a_n). Therefore, $|\pi_1| \geq 4(r-1)^2 |B_{(k-1)n}(F)|$, and

$$|B_{kn}(F) \setminus \pi_1| \leq |B_{kn}(F)| - 4(r-1)^2 |B_{(k-1)n}(F)| = \left(1 - \frac{4(r-1)^2}{(2r-1)^n}\right) |B_{kn}(F)|.$$

Set

$$B^h = B_{kn}(F) \setminus \bigcup_{t=1}^{h} \pi_t,$$

and

$$C^h = \{a_1 a_2 \dots a_{kn} \in B^h : a_{hn+2} a_{hn+3} \dots a_{(h+1)n-1} = w\} \equiv \pi_{h+1} \cap B^h.$$

Also define D^h as the set of all couples (v_1, v_2) such that $v_1 = a_1 a_2 \dots a_{hn} \in B_{hn}(F)$, with $a_{(t-1)n+2} a_{(t-1)n+3} \dots a_{tn-1} \neq w$ for $t = 1, 2, \dots, h$ and $v_2 \in B_{(k-h-1)n}(F)$. For any $(v_1, v_2) \in D^h$ there exist $a, b \in A$ such that $v_1 a w b v_2 \in C^h$. Therefore, we have again: $|C^h| \geq 4(r-1)^2 |D^h| \geq \frac{4(r-1)^2}{(2r-1)^n} |B^h|$, and so:

$$\begin{aligned} |B_{kn}(F) \setminus \bigcup_{t=1}^{h} \pi_t| &= |(B_{kn}(F) \setminus \bigcup_{t=1}^{h-1} \pi_t) \setminus \pi_h| = |B^{h-1} \setminus \pi_h| = |B^{h-1} \setminus C^{h-1}| \\ &\leq \left(1 - \frac{4(r-1)^2}{(2r-1)^n}\right) |B^{h-1}| \leq \left(1 - \frac{4(r-1)^2}{(2r-1)^n}\right)^h |B_{kn}(F)|, \end{aligned}$$

where the last inequality follows by an obvious inductive argument on h. Then for $h = k$ we obtain (3.2). Taking the logarithms, it follows that

$$\frac{\log |B_{kn}(\overline{F})|}{kn} \leq \frac{\log |W_{kn}|}{kn} \leq \frac{1}{n} \cdot \log \left(1 - \frac{4(r-1)^2}{(2r-1)^n}\right) + \frac{\log |B_{kn}(F)|}{kn},$$

and therefore letting $k \to \infty$ we obtain

$$\mathrm{ent}(\overline{F}, \overline{A}) \leq \frac{1}{n} \cdot \log \left(1 - \frac{4(r-1)^2}{(2r-1)^n}\right) + \mathrm{ent}(F, A) < \mathrm{ent}(F, A) \qquad \square.$$

Proof of Nielsen's Theorem via the entropy inequality. Let F be a free group of rank, say n, A_+ a free set of generators, N a normal subgroup and $\overline{F} = F/N$ the quotient group. Then with $A = A_+ \coprod (A_+)^{-1}$ we have

$$\mathrm{ent}(\overline{F}) \leq \mathrm{ent}(\overline{F}, \overline{A}) < \mathrm{ent}(F, A) = \mathrm{ent}(F) = 2n - 1,$$

the strict inequality comes from the above theorem. Therefore, by Proposition 3.2, $\overline{F}$ cannot be a free group of rank n. □

Remark 3.4. As we mentioned in the introduction, Nielsen's theorem has been generalized to torsion-free hyperbolic groups by Z. Sela [25]. One could then ask whether the entropic method presented in this section could be generalized to this class of groups yielding a new proof of Sela's theorem. We remark that Koubi [16] has shown that the entropy of non-elementary hyperbolic groups is positive, that is these groups are of uniform exponential growth. The problem is that for hyperbolic groups (indeed even for surface groups) it is not known whether the (minimal) entropy is attained or not, i.e. an analogue of Proposition 3.2 holds.

On the other hand, Arzhantseva and Lysenok [1], generalizing Theorem 3.3, have shown that non-elementary hyperbolic groups are *growth-tight* (see [7, 23]).

Acknowledgments. We express our gratitude to Slava Grigorchuk for several advices and interesting discussions. T.C.-S. acknowledges partial support from the Swiss National Science Foundation.

References

[1] G. N. Arzhantseva and I. G. Lysenok, Growth tightness for word hyperbolic groups, Math. Z. 241 (2002), 597–611.

[2] A. Avez, Entropie des groupes de type fini, C. R. Acad. Sci. Paris Sér. A-B 275 (1972), A1363–A1366.

[3] T. Ceccherini-Silberstein, F. Fiorenzi and F. Scarabotti, Garden of Eden theorems for cellular automata and for symbolic dynamical systems, in: Random Walks and Geometry, Proceedings of a Workshop at the Erwin Schrödinger Institute, Walter de Gruyter, Berlin 2004, 73–108.

[4] T. Ceccherini-Silberstein, A. Machì and F. Scarabotti, On the entropy of regular languages, Theoret. Comput. Sci. (to appear).

[5] T. Ceccherini-Silberstein and W. Woess, Growth and ergodicity of context-free languages, Trans. Amer. Math. Soc. 354 (2002), 4597–4625.

[6] R. I. Grigorchuk, Symmetric random walks on discrete groups, in: Multicomponent Random Systems, Adv. Probab. Related Topics 6, Dekker, New York 1980, 285–325.

[7] R. I. Grigorchuk and P. de la Harpe, On problems related to growth, entropy and spectrum in group theory, J. Dynam. Control Systems 3 (1997), 51–89.

[8] R. I. Grigorchuk and M. J. Mamaghani, On use of iterates of endomorphisms for constructing groups with specific properties, Mat. Stud. 8 (1997), 198–206, 238.

[9] M. Gromov, Structures métriques pour les variétés riemanniennes, rédigé par J, Lafontaine et P. Pansu, Textes mathématiques 1, CEDIC/Fernand Nathan, Paris 1981.

[10] M. Gromov, Endomorphisms of symbolic algebraic varieties, J. Eur. Math. Soc. (JEMS) 1 (1999), 109–197.

[11] P. de la Harpe, Topics in Geometric Group Theory, University of Chicago Press, Chicago 2000.

[12] H. Hopf, Beiträge zur Klassifizierung der Flächenabbildungen, J. Reine Angew. Math. 165 (1931), 225–236.

[13] H. Kesten, Symmetric random walks on groups, Trans. Amer. Math. Soc. 92 (1959), 336–354.

[14] H. Kesten, Full Banach mean values on countable groups, Math. Scand. 7 (1959), 146–156.

[15] M. Koubi, On exponential growth rates for free groups, Publ. Mat. 42 (1998), 499–507.

[16] M. Koubi, Croissance uniforme dans les groupes hyperboliques, Ann. Inst. Fourier (Grenoble) 48 (1998), 1441–1453.

[17] D. Lind and B. Marcus, An Introduction to Symbolic Dynamics and Coding, Cambridge University Press, Cambridge 1995.

[18] R. Lyndon and P. E. Schupp, Combinatorial Group Theory, Ergeb. Math. Grenzgeb. 89, Springer-Verlag, Berlin 1977.

[19] W. Magnus, A. Karrass and D. Solitar, Combinatorial Group Theory; Presentations of Groups in Terms of Generators and Relations, Interscience, New York 1966.

[20] D. Meier, Non-Hopfian groups, J. London Math. Soc. (2) 26 (1982), 265–270.

[21] J. Nielsen, Om regning med ikke kommutative Faktoren og dens Anvedelse i Gruppenteorien, Matematisk Tidsskrift, B (1921) 77–94.

[22] A. Sambusetti, Minimal growth of non-Hopfian free prducts, C. R. Acad. Sci. Paris, Sér. I Math. 329 (1999), 943–946.

[23] A. Sambusetti, Growth tightness of free and amalgamated products, Ann. Sci. École Norm. Sup. (to appear).

[24] F. Scarabotti, On a lemma of Gromov and the entropy of a graph, European J. Combin. 23 (2002), 631–633

[25] Z. Sela, Endomorphisms of hyperbolic groups. I. The Hopf property, Topology 38 (1999), 301–321.

Tullio Ceccherini-Silberstein, Dipartimento di Ingegneria, Università del Sannio. C.so Garibaldi 107, 82100 Benevento, Italy

E-mail: tceccher@mat.uniroma1.it

Fabio Scarabotti, Dipartimento MeMoMat, Università degli Studi di Roma “La Sapienza”. Via A. Scarpa 16, 00161 Roma, Italy

E-mail: scarabot@dmmm.uniroma1.it

Growth rates of small cancellation groups

Anna Erschler[*]

Abstract. We estimate growth rate for small cancellation groups. In particular we show that there is a continuum of possible values for exponential growth rates.

1 Introduction

Let G be a group with some fixed set of generators S, and let d denote the word metric corresponding to this set of generators.

Let $B_{G,S}(n)$ denote the ball $\{g \in G : d(g,e) \leq n\}$ and $v_S(G)$ denote the (exponential) growth rate

$$v_S(G) = \lim \sqrt[n]{\#B_{G,S}(n)}.$$

It is well known and easy to check that this limit always exists and it is equal to

$$\inf_n \sqrt[n]{\#B_{G,S}(n)},$$

e.g., see [8]. The group G has exponential growth, if for some (and hence for all) set of generators of G one has $v_S(G) > 1$. There are many examples of groups of exponential growth (e.g., free groups and more generally infinite hyperbolic groups, non-amenable groups, solvable non virtually nilpotent groups [14], [18]). On the other hand, all known examples of finitely presented groups of non-exponential growth have polynomial growth (and hence are virtually nilpotent by a theorem of Gromov [9]). We remind that there exist groups of intermediate growth, but all known examples are not finitely presented. The existence of such groups was first discovered by R. Grigorchuk [6].

Yet not much is known about the values of exponential growth rates. Set

$$\Omega^G = \{v \in [1, \infty) \mid v = v_S(G) \text{ for some set of generators } S \text{ of } G\}.$$

A well known question [10] asked whether there exists a group G of exponential growth such that

$$\inf \Omega^G = 1$$

[*]Supported in part by the Swiss National Science Foundation.

(i.e., whether there exists a group of exponential, but not uniformly exponential growth). Some partial results for this question were obtained (see the references in [12]). Namely, it was proved for some classes of groups that exponential growth implies uniformly exponential growth. Recently I learned from P. de la Harpe that the question was settled by J. Wilson[1], who constructed a group of exponential, but not uniformly exponential growth.

Another problem that arises in this context is to describe the set of growth rates of a single group (or of the generic group in a given class).

If we fix the number n of generators of G then there are obvious estimates

$$1 \leq v_S(G) \leq 2n - 1,$$

where the equality in the last inequality is achieved if and only if the group is free on the $2n - 1$ given generators. There are no other known general restrictions for $v_S(G)$. But for some classes of groups there exist such restrictions. For example, any growth rate of any hyperbolic group is an integer algebraic number [11], [5].

The groups we consider in this paper are close to hyperbolic ones. We consider non finitely presented groups with small cancellation condition. We remind that having a finite presentation implies hyperbolicity. Still the situation for these groups differs from the hyperbolic case.

In this paper we study the set

$$\Omega = \{v \in [1, \infty) \mid v = v_S(G) \text{ for some } G \text{ and generating set } S\}.$$

It is known ([8], [12]) that Ω is everywhere dense in $[1, \infty]$. It was asked in [12] whether Ω has the cardinality of continuum.

The following theorem answers this question.

Theorem 1.1. *Let*

$$\Omega_2 = \{v \in [1, \infty) \mid v = v_S(G) \textit{ for some two generated group } G\}.$$

Then $\#\Omega_2 = 2^{\aleph_0}$.

This theorem follows from the proposition below.

Proposition 1.2. *There exists a function $E : \mathbb{N} \to \mathbb{N}$ with the following property. Let $r_i = (a^{E(i)}b^{E(i)})^{100}$. For any $J \subset \mathbb{N}$ define G_J as*

$$G_J = \langle a, b \mid r_j = e, j \in J \rangle.$$

Then $I \neq J$ implies that $v_{\{a,b\}}(G_I) \neq v_{\{a,b\}}(G_J)$.

Remark 1.3. Similar groups were first introduced by B. H. Bowditch. He used them to construct continuously many quasi-isometry classes of groups [2]. Similar groups also provide examples of a group with non-homeomorphic asymptotic cones [17].

The results of this paper were announced in [3].

[1] Added in proof: J. S. Wilson, On exponential growth and uniformly exponential growth, Invent. Math. 155 (2) (2004), 287–303.

2 Definitions and preliminary observations

Let $|w|$ denote the length of a word w in the corresponding free group.

For a non-negative matrix A it is known [4] that there exists a non-negative eigenvalue v of A such that for any other eigenvalue v' it holds $|v'| \leq v$. The non-negative eigenvalue v is called the Perron-Frobenius eigenvalue, and the corresponding eigenvector is called the Perron-Frobenius eigenvector. Let $v(A)$ denote this maximal eigenvalue and $\sum(A)$ denote the sum of elements of A.

Let G be a group and let $\langle S, R\rangle$ be a presentation of G. *Symmetrization* R_* consists of all distinct cyclic permutations of the defining relators $r \in R$ and their inverses. A word u is a *piece* relative to R, if R_* contains two distinct elements of the form uv' and uv''.

Definition 2.1 (e.g., see [16]). Let $\lambda > 0$. We say that $\langle S, R\rangle$ satisfies *a small cancellation condition* $C'(\lambda)$ if for any $r \in R_*$ and any piece u of it

$$|u| < \lambda |r|.$$

If some finite presentation of a group satisfies $C'(1/6)$, then this group is hyperbolic [16]. Hence there is a Markov grammar corresponding to G (and preserving the word length) [11], [5].

Lemma 2.2. *Let*

$$H = \langle S \mid w_i = e, i \in \mathbb{N}\rangle \quad \textit{and} \quad H_k = \langle S \mid w_i = e, i \leq k\rangle.$$

Then $\lim_{k\to\infty} v_S(H_k) = v_S(H)$.

Proof. It is clear that $v_S(H_k) \geq v_S(H_j)$ for $k < j$. Hence, there exists a limit $v = \lim_{k\to\infty} v_S(H_k)$. Then $v_S(H) \leq v$, because $v_S(H) \leq v_S(H_k)$ for any k.

On the other hand,

$$\#B_{H_k}(r) \geq v_S(H_k)^r \geq v^r,$$

and for each r there exists k such that

$$\#B_H(r) = \#B_{H_k}(r).$$

Hence, for any r we have $\#B_H(r) \geq v^r$ and $v_S(H) \geq v$. □

Definition 2.3. Consider a Markov grammar with states $1, 2, \ldots, N$. Adjacency matrix of this grammar M is the matrix such that $M(i, j)$ is the number of oriented edges from state i to state j.

If we say that a Markov grammar corresponds to some group we assume that the correspondence preserves word length.

The growth function of a language over some finite alphabet $B(n)$ is the number of words of length at most n in this language. The exponential growth rate v of a

language is defined as

$$v = \limsup \sqrt[n]{\#B(n)},$$

where $B(n)$ is the set of words of length at most n. Is is easy to check that for a language given by a Markov grammar

$$v = \lim \sqrt[n]{\#B(n)}.$$

Lemma 2.4. *Consider a language given by some Markov grammar and let M be the adjacency matrix of this grammar. Then the growth rate of the language is equal to $v(M)$.*

Proof. It is clear that the growth rate is not greater than $v(M)$. On the other hand, we know that there is a non-negative eigenvector corresponding to $v(M)$ [4]. This implies that $\sum(M^r) \geq Cv(M)^r$ for some $C > 0$, so that the growth rate is not less than $v(M)$. □

3 Estimate from below

Lemma 3.1. *Consider a Markov grammar with s states and k labels. Let v be the growth rate of the corresponding language. Suppose that $s \leq N$ and fix some p words of length $4N$. We call them long forbidden words. Consider the following new language. It consists of the words of the old language that do not contain any of the long forbidden words. Let v_{new} be the growth rate of the new language. Then*

$$v_{\text{new}}^N \geq (v^N/s^2 - 4Np)/k^s.$$

Proof. Let M be the adjacency matrix of the Markov grammar. Then $p_{i,j}^L = M^L(i, j)$ is the number of paths of length L starting from the state i and ending in the state j.

Note that there exists $\tilde{s} \leq s$ such that for some i

$$p_{i,i}^{N+\tilde{s}} \geq v^N/s^2.$$

In fact, let $A_1, A_2, \ldots, A_f$ be irreducible blocks of M (see [4] for the definition of irreducible blocks). Then there exists g such that $v(A_g) = v(M) = v$. Since $\sum(A_g^N) \geq v(A_g)^N = v^N$, we see that there exist i and j in the block A_g such that

$$p_{i,j}^N = M^N(i, j) = A_g^N(i, j) \geq v^N/s^2.$$

There exists $\tilde{s} \leq s$ such that we can go from the state i to the state j in $\tilde{s}$ steps. This implies that

$$p_{i,i}^{N+\tilde{s}} \geq v^N/s^2.$$

Let $N' = N + \tilde{s}$. Note that $N' \leq 2N$. Consider all subwords of length N' of long forbidden words. We will call them short forbidden words. Note that there are at most

$4Np$ short forbidden words. Take some word w. Suppose that

$$w = a_1 a_2 \dots a_l$$

and that $|a_m| = N'$ for any $m < l$, $|a_l| \leq N'$. Note that if none of a_m is a short forbidden word, then w contains no long forbidden words. This implies that

$$v_{\text{new}}^{N'} \geq v^N / s^2 - 4Np.$$

Since $v_{\text{new}} \leq v \leq k$, we deduce that

$$v_{\text{new}}^{N} = v_{\text{new}}^{N'} / v_{\text{new}}^{s} \geq v_{\text{new}}^{N'} / k^s \geq\geq (v^N / s^2 - 4Np)/k^s. \qquad \square$$

Lemma 3.2. *Let*

$$G = \langle S \mid w_1 = w_2 = \dots = w_{n+1} = e \rangle$$

and

$$H = \langle S \mid w_1 = w_2 = \dots = w_n = e \rangle.$$

Suppose that these presentations satisfy $C'(1/6)$ and that $|w_i| > 6$ for any i. We call a word w $1/12$-reduced with respect to w_{n+1} if it contains no subword of w_{n+1} of length $\geq 1/12|w_{n+1}|$ and if it is freely reduced. Then

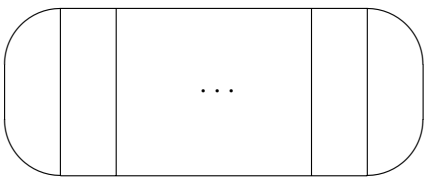
1. *If w is $1/12$-reduced with respect to w_{n+1}, and w is geodesic in H, then it is geodesic in G.*

2. *Let γ_1 and γ_2 be $1/12$-reduced with respect to w_{n+1}. If γ_1 and γ_2 are geodesic in H, and $\gamma_1 \neq \gamma_2$ in H, then $\gamma_1 \neq \gamma_2$ in G.*

Proof. 1. Let us use induction on $|w|$.

Base. Note that if $|w| = 1$, then the $C'(1/6)$ condition implies that $w \neq e$ in G and hence it is geodesic in G.

Induction step. Suppose that $w = w_1 w_2$ and that $|w_2| = 1$. Since $|w_1| = |w| - 1$ and w_1 is geodesic in H, we know already that w_1 is geodesic in G. Suppose that $w = w_3$ in G and that $|w_3| < |w|$. Consider a triangle in G with sides w_1, w_2, w_3. Without loss of generality we may assume that this is a simple reduced triangle, that is that its sides intersect only in its vertices. In fact, otherwise we could replace w_1 and w_3 with some of its end subwords. Consider a reduced diagram of this triangle and its tiling. Note that the tiling contains at most 2 distinguished tiles. (A tile is said to be distinguished, if it contains a vertex of a triangle inside some of its open exterior edges, see [16]). Then [16, Part 3 of Theorem 35] implies that the tiling looks like one on the



picture.

Note that if some tile corresponds to w_{n+1}, then it has at least 3 interior edges. So no tile can be labelled with w_{n+1}. This means that we have this triangle not only in G but also in H. So $w = w_3$ in H, but this is impossible.

2. By the first part of the lemma we know that γ_1, γ_2 are geodesic in G. Suppose that $\gamma_1 = \gamma_2$ in G. Consider a digon with sides γ_1 and γ_2. Again without loss of generality we may assume that the digon is reduced and apply [16, Part 3 of Theorem 35]. No tile can be labelled with w_{n+1} and this implies that $\gamma_1 = \gamma_2$ in H. □

The two previous lemmas imply the following corollary.

Corollary 3.3. *Let*

$$G = \langle S \mid w_1 = w_2 = \cdots = w_{n+1} = e \rangle$$

and

$$H = \langle S \mid w_1 = w_2 = \cdots = w_n = e \rangle.$$

Suppose that these presentations satisfy $C'(1/6)$ and that there is a Markov grammar for H with s states. There exists $\gamma(s,k)$ such that if

$$|w_{n+1}| > \gamma(s,k),$$

then

$$v_S(G) \geq v_S(H) - 200/\sqrt{|w_{n+1}|}.$$

Proof. Let N be the integer part of $|w_{n+1}|/52$. We may assume that

$$N > 1000 s^2 (2k)^{s+6}.$$

Consider all subwords of length $4N$ of w_{n+1}. There are at most $52(N+1)$ such subwords. By the previous lemma we can estimate the growth rate of G by the growth rate of the grammar corresponding to H where these words of length $4N$ are forbidden. Then we apply Lemma 3.1 and get that

$$v_S(G)^N \geq (v_S(H)^N/s^2 - 4N52(N+1))/(2k)^s \geq (v_S(H)^N - N^4)N.$$

Note that without loss of generality we may assume that $v_S(H) > N^4 + 1$, since otherwise

$$v_S(H) < \sqrt[N]{1+N^4} \leq 1 + 1/\sqrt{N}$$

(note that $N > 1000$).

But if $v_S(H) > N^4 + 1$, then

$$\sqrt[N]{v_S(H)^N - N} \geq v_S(H) - 1/\sqrt{N},$$

and

$$\sqrt[N]{N} \leq 1 + 1/N^{\frac{2}{3}}.$$

Hence,

$$v_S(G) \geq \frac{v_S(H) - 1/\sqrt{N}}{1 + 1/N^{\frac{2}{3}}} \geq (v_S(H) - 1/\sqrt{N})(v_S(H) - 1/N^{\frac{2}{3}}).$$

But since $\sqrt[6]{N} \geq 2k \geq v_S(H)$,

$$(v_S(H) - 1/\sqrt{N})(v_S(H) - 1/N^{\frac{2}{3}}) \geq v_S(H) - 2/\sqrt{N} \geq v_S(H) - 200/\sqrt{|w_{n+1}|}.$$

□

The previous corollary implies another corollary:

Corollary 3.4. *Consider a group*

$$H = \langle S \mid w_1 = w_2 = \cdots = w_n = e \rangle$$

and a sequence of groups

$$H_k = \langle S \mid w_1 = w_2 = \cdots = w_n = \tilde{w}_k = e \rangle.$$

Suppose that these presentations satisfy $C'(1/6)$ and that $|\tilde{w}_k| \to \infty$ when $k \to \infty$. Then $\lim_{k\to\infty} v_S(H_k) = v_S(H)$.

This generalizes a corollary from [15] (where H was a free group).

4 Proof of the main result

Notations.

- Choose $\alpha : \mathbb{N} \to \mathbb{N}$ such that if

  $$H_1 = \langle a, b \mid r_i^1 = e, i \in I \rangle, \quad H_2 = \langle a, b \mid r_j^2 = e, j \in J \rangle,$$

  and $|r_i^1|, |r_j^2| \leq N$, then either

  $$v_{\{a,b\}}(H_1) = v_{\{a,b\}}(H_2),$$

  or

  $$|v_{\{a,b\}}(H_1) - v_{\{a,b\}}(H_2)| \geq 1/\alpha(N).$$

- Choose $\beta : \mathbb{N} \to \mathbb{N}$ such that if

  $$H = \langle a, b \mid r_i = e, i \in I \rangle,$$

  $|r_i| \leq N$ for any i, and there is a Markov grammar for H (and for these generators a, b), then one of these Markov grammars contains at most $\beta(N)$ states (note that for example we can take any $C'(1/6)$ group for H).

Note that α and β are well defined, since there are finitely many groups H such that

$$H = \langle a, b \mid r_i = e, i \in I \rangle,$$

and $|r_i| \leq N$ for any i.

Observation 4.1. *Suppose that*

$$H_1 = \langle a, b \mid r_1 = \dots r_n = e \rangle, \quad H_2 = \langle a, b \mid r_1 = \dots = r_{n+1} = e \rangle,$$

both these presentation satisfy $C'(1/6)$ small cancellation condition, and $|r_i| \leq N$ for any $1 \leq i \leq n+1$. Then

$$v_{\{a,b\}}(H_2) < v_{\{a,b\}}(H_1) - \alpha(N).$$

Proof. First note that $v_{\{a,b\}}(H_2) < v_{\{a,b\}}(H_1)$, because H_2 is an infinite quotient of a word-hyperbolic group H_1 (see [1]). Hence the observation follows from the definition of α. □

Proof of Proposition 1.2. Consider $E : \mathbb{N} \to \mathbb{N}$ such that

$$E(i+1) > 400E(i),$$
$$400/\sqrt{E(i+1)} < \alpha(200E(i)),$$
$$E(i+1) > \gamma(\beta(200E(i)), 2),$$

where γ is as in (the statement of) Corollary 3.3.

Suppose that $J \neq M$ are subsets of $\mathbb{N}$. Let

$$J = \{j_1, j_2, \dots, j_k, \dots\}, \quad M = \{m_1, m_2, \dots, m_k, \dots\}$$

with $j_1 < j_2 < \dots < j_k < \dots$ and $m_1 < m_2 < \dots < m_k < \dots$.

Suppose that $j_i = m_i$ for $i \leq k$ and that $j_{k+1} \neq m_{k+1}$. Without loss of generality we may assume that $j_{k+1} > m_{k+1}$. Let $L = \{j_1, j_2, \dots, j_k\}$, $v = v_{\{a,b\}}(G_L)$, $v_1 = v_{\{a,b\}}(G_J)$ and $v_2 = v_{\{a,b\}}(G_M)$. From the estimate from below and from Lemma 2.2 we conclude that

$$\begin{aligned} v_1 &\geq v - 200/\sqrt{E(j_{k+1})} - 200/\sqrt{E(j_{k+2})} - 200/\sqrt{E(j_{k+3})} - \dots \\ &\geq v - 400/\sqrt{E(j_{k+1})} > v - \alpha(200E(j_{k+1} - 1)) \geq v - \alpha(200E(m_{k+1})). \end{aligned}$$

Let

$$L' = \{m_1, m_2, \dots, m_k, m_{k+1}\},$$

and let $v' = v_{\{a,b\}}(G_{L'})$. From Observation 4.1 we get

$$v' \leq v - \alpha(200E(m_{k+1})),$$

and hence

$$v_2 \leq v' < v_1.$$

So we have proved that $v_1 \neq v_2$. □

Remark 4.2. In fact one can construct $E(i)$ described in Proposition 1.2 explicitly. To do that one should estimate $\alpha(x)$ and $\beta(x)$ for the case of small cancellation groups. This can be done (for the example from the proof of the main theorem of [5] one can estimate the number of states of Markov grammar in terms of the hyperbolicity constant). Finally one can see that one can take

$$E(i) = \underbrace{1000^{1000^{\cdots^{1000}}}}_{10i \quad \text{times}}.$$

We omit the details.

Acknowledgement. I would like to thank Pierre de la Harpe for the encouragement and many helpful comments on this paper. A part of this paper was written during the author's stay at the University of Geneva. The author gratefully acknowledges the support of the Swiss National Science Foundation.

References

[1] G. N. Arzhantseva and I. G. Lysenok, Growth tightness for word hyperbolic groups, Math. Z. 241 (2002), 597–611.

[2] B. H. Bowditch, Continuously many quasi-isometry classes of 2-generator groups, Comment. Math. Helv. 73 (1998), 232–236.

[3] A. Erschler (Dyubina), On the values of exponential growth rates for groups with a small cancellation condition, Funct. Anal. Appl. 36 (1) (2002), 79–81.

[4] F. R. Gantmacher, The Theory of Matrices, Vols. 1, 2, Chelsea, New York 1959.

[5] E. Ghys and P. de la Harpe, La propriete de Markov pour les groupes hyperboliques, in: Sur les Groupes Hyperboliques d'après Mikhael Gromov (Bern, 1988), Progr. Math. 83, Birkhäuser Boston, Boston, MA, 1990, 165–187.

[6] R. I. Grigorchuk, Degrees of growth of finitely generated groups and the theory of invariant means, Izv. Akad. Nauk SSSR Ser. Mat. 48 (1984), 939–985.

[7] R. I. Grigorchuk and P. de la Harpe, On problems related to growth, entropy and spectrum in group theory, J. Dynam. Control Systems 3 (1997), 51–89.

[8] R. Grigorchuk and P. de la Harpe, Limit behavior of exponential growth rates for finitely generated groups, in: Essays on Geometry and Related Topics, Vols. 1, 2, Monogr. Enseign. Math. 38, Enseignement Math., Geneva 2001, 351–370.

[9] M. Gromov, Groups of polynomial growth and expanding maps, Inst. Hautes Études Sci. Publ. Math. 53 (1981), 53–73.

[10] M. Gromov, Structures métriques pour les variétés riemanniennes, rédigé par J. Lafontaine et P. Pansu, Textes mathématiques 1, CEDIC/Fernand Nathan, Paris 1981.

[11] M. Gromov, Hyperbolic groups, in: Essays in Group Theory, Math. Sci. Res. Inst. Publ. 8, Springer-Verlag, New York 1987, 75–263.

[12] P. de la Harpe, Topics in Geometric Group Theory, University of Chicago Press, Chicago 2000.

[13] R. Lyndon and P. E. Schupp, Combinatorial Group Theory, Ergeb. Math. Grenzgeb. 89, Springer-Verlag, Berlin 1977.

[14] J. Milnor, Growth of finitely generated solvable groups, J. Differential Geom. 2 (1968) 447–449.

[15] A. G. Shukhov, On the dependence of the growth exponent on the length of the defining relation, Math. Notes 65 (1999), 510–515.

[16] R. Strebel, Appendix. Small cancellation groups, in: Sur les groupes hyperboliques d'après Mikhael Gromov (Bern, 1988), Progr. Math. 83, Birkhäuser, Boston, MA, 1990, 227–273.

[17] S. Thomas, B. Velickovic, Asymptotic cones of finitely generated groups, Bull. London Math. Soc. 32 (2000), 203–208.

[18] J. A. Wolf, Growth of finitely generated solvable groups and curvature of Riemannian manifolds, J. Differential Geom. 2 (1968), 421–446.

Anna Erschler, CNRS, University Lille 1, UFR de Mathematiques, 59655 Villeneuve d'Ascq Cedex, France

E-mail: erschler@math.univ-lille1.fr, erschler@pdmi.ras.ru

Recurrence properties of random walks on finite volume homogeneous manifolds

*Alex Eskin*and Gregory Margulis***

Abstract. Let G be a semisimple Lie group, and Γ a nonuniform irreducible lattice in G. Recurrence properties of the action of a unipotent one-parameter subgroup of G on the quotient space G/Γ were studied by Dani and Margulis. The aim of this paper is to show that similar results hold under some conditions for random walks on G/Γ.

1 Introduction

Let G be a semisimple Lie group, and Γ a nonuniform irreducible lattice in G. Let π denote the natural projection from G to G/Γ. We recall some results from [2, 3, 4, 5, 9]. Let $U = \{u_t\}$ be a unipotent one-parameter subgroup of G.

Theorem 1.1. *For* every *point $x \in G/\Gamma$ and every $\epsilon > 0$ there exists a compact set $K \subset G/\Gamma$ such that for all $T > 0$*

$$|\{t \in [0, T] : u_t x \in K\}| > (1 - \epsilon)T. \tag{1.1}$$

More generally, for every compact set $C \in G/\Gamma$ and every $\epsilon > 0$ there exists a compact set K such that (1.1) *holds for all $x \in C$, and all $T > 0$.*

A parabolic subgroup P of G is called Γ-*rational* if the unipotent radical of P intersects Γ in a lattice. Note that if Γ is arithmetic then a parabolic subgroup of G is Γ-rational if and only if it is defined over $\mathbb{Q}$.

Theorem 1.2. *There exists a compact set K with the following property: Let $g \in G$ be any element such that gUg^{-1} is not contained in a Γ-rational parabolic subgroup of G. Then $U\pi(g)$ intersects K.*

Remark 1.3. Essentially the assertion of Theorem 1.2 is that every orbit of U returns to a fixed compact set K. An example that the algebraic condition on gUg^{-1} in

*Research partially supported by NSF grant DMS-9704845, the Sloan Foundation and the Packard Foundation

**Research partially supported by NSF grant DMS-9800607

Theorem 1.2 is needed can be given as follows: Let G be $\mathrm{SL}(n,\mathbb{R})$ and $\Gamma = \mathrm{SL}(n,\mathbb{Z})$. Suppose gUg^{-1} is contained in a $\mathbb{Q}$-parabolic P. Let V be the subspace stabilized by P. Since V is defined over $\mathbb{Q}$, its intersection with $\mathbb{Z}^n$ is a lattice L. Note that the volume of the torus $V/V \cap L$ is an invariant of the orbit; if it is sufficiently small, the orbit will not intersect the fixed compact set K.

We emphasize that the key point of Theorem 1.1 and Theorem 1.2 is that Theorem 1.1 holds for every point x, and Theorem 1.2 holds for every element g satisfying an explicit algebraic condition (the fact that the theorems hold almost everywhere being an easy consequence of ergodicity and the Birkhoff ergodic theorem). The proofs rely heavily on the polynomial nature of the unipotent flow, and are based mostly on the techniques of [9].

The aim of this paper is to show that similar results hold under some conditions when considering random walks on G/Γ. Let μ be a measure on G satisfying the condition

$$\int_G \|g\|^\delta \, d\mu(g) < \infty \tag{1.2}$$

for sufficiently small $\delta > 0$, where $\|\cdot\|$ denotes some norm in some faithful finite dimensional representation of G. To state the results we formulate some properties analogous to the conclusions of Theorem 1.1 and Theorem 1.2 which may be possessed by the random walk defined by μ.

Definition 1.4. Let μ be a probability measure on G satisfying (1.2). Let $\mu^{(m)}$ denote the convolution of μ with itself m times, and for $x \in G$ let δ_x denote the probability measure supported on the point x. The following may hold:

(R1) For every compact set $C \subset G/\Gamma$ and every $\epsilon > 0$ there exists a compact set $K \supset C$ such that for every $x \in C$ and every $m > 0$,

$$(\mu^{(m)} * \delta_x)(K) > (1 - \epsilon).$$

(R2) For every $\epsilon > 0$ there exists a compact set K such that for every $x \in G/\Gamma$, there exists $M = M(x) > 0$ such that for $m > M(x)$,

$$(\mu^{(m)} * \delta_x)(K) > (1 - \epsilon).$$

The constant $M(x)$ can be chosen so that for every compact subset C of G/Γ, $\sup_{x \in C} M(x) < \infty$.

(S) For every $\epsilon > 0$ there exists a compact set K such that for every $g \in G$ either $g(\mathrm{supp}\mu)g^{-1}$ is contained in a Γ-rational parabolic P, or there exists $M > 0$ such that for $m > M$, $(\mu^{(m)} * \delta_{\pi(g)})(K) > (1 - \epsilon)$.

The property (R1) is a version of the conclusion of Theorem 1.1 which makes sense in the current context. The property (S) has the same relation to the conclusion of Theorem 1.2.

2 Recurrence properties of random walks

To simplify the terminology, we assume that G is a connected algebraic group. All the claims can be easily reduced to this case.

Notation. Let H_μ denote noncompact part of the Zariski closure of the subgroup generated by the support of μ. (By the noncompact part of an algebraic group H we mean the subgroup of H generated by unipotent elements, together with the split part of H).

Our result is the following:

Theorem 2.1. *Suppose H_μ is semisimple, and for all $g \in G$, $gH_\mu g^{-1}$ is not contained in any proper Γ-rational parabolic subgroup of G. Then μ has properties* (R1) *and* (R2). *In particular, if $H_\mu = G$, then μ has properties* (R1) *and* (R2).

Recall that a measure ν on a G-space X is μ-stationary if and only if $\mu * \nu = \nu$. We note the following general lemma (proved in §5):

Lemma 2.2. *Let μ be any measure on G satisfying* (R2). *Then any μ-stationary locally finite measure on G/Γ is finite.*

As a corollary of Theorem 2.1 and Lemma 2.2 we have the following result conjectured by N. Shah in [10]:

Theorem 2.3. *Let G be a semisimple Lie group, and Γ a nonuniform irreducible lattice in G. Let $\Lambda \subset G$ be a countable Zariski dense subgroup (or, more generally, a countable subgroup with semisimple Zariski closure which is not contained in any conjugate of a proper Γ-rational parabolic subgroup of G). Let $x \in G/\Gamma$ be a point such that the orbit $\Lambda x \subset G/\Gamma$ is discrete. Then Λx is finite.*

Proof of Theorem 2.3. Since Λ is countable, we can find a measure μ supported on $\Lambda \subset G$, such that $\mathrm{supp}\mu$ generates Λ, and (1.2) holds. Let σ denote counting measure on the discrete set Λx (i.e., $\sigma(C)$ is the cardinality of $C \cap \Lambda x$). Then σ is invariant, hence stationary. Then, in view of Theorem 2.1 and Lemma 2.2, σ is finite. Hence Λx is finite. □

As another corollary, we obtain another proof, in a special case, of the following well-known theorem of Borel and Harish-Chandra:

Theorem 2.4. *Let $\boldsymbol{H}$ be a semisimple algebraic group defined over $\mathbb{Q}$. Then $\boldsymbol{H}(\mathbb{Z})$ is a lattice in $\boldsymbol{H}(\mathbb{R})$ (i.e., $\boldsymbol{H}(\mathbb{R})/\boldsymbol{H}(\mathbb{Z})$ has finite $\mathbf{H}(\mathbb{R})$-invariant measure).*

We give a proof of this theorem under the assumption that there exists a faithful representation of $\boldsymbol{H}$ which is defined over $\mathbb{Q}$ and is irreducible over $\mathbb{R}$.

Proof of Theorem 2.4 (under the assumption). Let $H = \mathbf{H}(\mathbb{R})$, $\Lambda = \mathbf{H}(\mathbb{Z})$. We are assuming that there is a faithful representation $\rho : H \to \mathrm{SL}(N, \mathbb{R})$ which is irreducible

over $\mathbb{R}$ and defined over $\mathbb{Q}$. Since ρ is defined over $\mathbb{Q}$, we have, after possibly replacing Λ by a subgroup of finite index, $\rho(\Lambda) \subset \mathrm{SL}(N,\mathbb{Z})$. Hence, we have $\rho(H)/\rho(\Lambda) \subset \mathrm{SL}(N,\mathbb{R})/\mathrm{SL}(N,\mathbb{Z})$. Let σ denote the $\rho(H)$-invariant measure on $\rho(H)/\rho(\Lambda)$. Then σ is locally finite. Let μ be any compactly supported absolutely continuous measure on H. Then, by Theorem 2.1, and since ρ is irreducible over $\mathbb{R}$, $\rho(\mu)$ has property (R2). Also since σ is $\rho(H)$-invariant, it is also $\rho(H)$-stationary. Thus, by Lemma 2.2, σ is finite. Hence Λ is a lattice in H. □

We also make the following conjecture:

Conjecture 2.5. Suppose H_μ is semisimple (or more generally generated by unipotent elements). Then μ has properties (R1) and (S).

Theorem 2.1 can be somewhat generalized: see Proposition 2.6 below.

The set Δ. We now define a certain finite collection Δ of maximal parabolic subgroups of G. If G has real rank 1, then, since Γ is non-uniform, there exists a Γ-rational minimal parabolic subgroup P_0 of G, and we define $\Delta = \{P_0\}$. If the real rank of G is at least 2, then Γ is arithmetic, hence there exists a $\mathbb{Q}$-structure on G such that Γ is commensurable with the set $G(\mathbb{Z})$ of integer points (with respect to this $\mathbb{Q}$-structure). We fix a maximal $\mathbb{Q}$-split torus A_0 for G, and we let $\Delta = \{P_1, \dots, P_r\}$, where the P_k are the standard parabolic subgroups with respect to A_0. For every $P_i \in \Delta$, there exists a representation $\rho_i : G \to \mathrm{GL}(V_i)$ and vectors $w_i \in V_i$ such that the stabilizer of $\mathbb{R}w_i$ is P_i.

Condition A. We say that μ satisfies *Condition A* if it satisfies (1.2), and for all sufficiently small $\delta > 0$ there exist $c < 1$ and $n > 0$ such that for all i, $1 \le i \le r$ and all $v \in Gw_i$,

$$\int_G \frac{1}{\|\rho_i(g)v\|^\delta}\, d\mu^{(n)}(g) \le \frac{c}{\|v\|^\delta}. \tag{2.1}$$

Theorem 2.1 follows immediately from the following two propositions:

Proposition 2.6. *Suppose μ satisfies Condition A. Then μ has properties* (R1) *and* (R2).

Proposition 2.7. *Suppose H_μ is semisimple and for any $g \in G$, $gH_\mu g^{-1}$ is not contained in any Γ-rational parabolic subgroup of G. Then μ satisfies Condition A.*

Proposition 2.6 will be proved in §3, and Proposition 2.7 will be proved in §4. Lemma 2.2 will be proved in §5.

3 Systems of inequalities

In this section we prove Proposition 2.6. The proof is based on the following:

Lemma 3.1. *Suppose that there exists a positive function $u : G/\Gamma \to \mathbb{R}$ with the following properties:*

(i) $u(x) \to \infty$ *as* $x \to \infty$ *in* G/Γ.

(ii) *There exists constants* $c_1 < 1$ *and* $b > 0$ *and* $n > 0$ *such that for any* $x \in G/\Gamma$,

$$\int_G u(gx)\, d\mu^{(n)}(g) \le c_1 u(x) + b. \tag{3.1}$$

Then μ has properties (R1) *and* (R2).

Proof. We note that in view of (1.2), it is enough to prove (R1) and (R2) for m in an arithmetic progression. After iterating (3.1) and summing the geometric series, we obtain for any multiple m of n,

$$\int_G u(gx)\, d\mu^{(m)}(g) \le c_1^{m/n} u(x) + b_1, \tag{3.2}$$

where b_1 is independent of m and x. Since for any $R > 0$ the set $\{y \in G/\Gamma : u(y) < R\}$ is compact, this immediately implies (R1), since we may choose the compact set $K = \{y \in G/\Gamma : u(y) < (u(x) + b_1)/\epsilon\}$; then

$$\frac{(u(x) + b_1)}{\epsilon}(\mu^{(m)} * \delta_x)(K^c) \le \int_G u(hx)\, d\mu^{(m)}(h) \le c_1^{m/n} u(x) + b_1,$$

hence $(\mu^{(m)} * \delta_x)(K^c) < \epsilon$ as required. To get (R2), we choose $K = \{y \in G/\Gamma : u(y) < 2b_1/\epsilon\}$; then, for m large enough so that $c_1^{m/n} u(x) < b_1$, arguing as above we see that $(\mu^{(m)} * \delta_x)(K^c) < \epsilon$ in view of (3.2). □

3.1 Construction of the function u in the $\mathrm{SL}(d, \mathbb{R})/\mathrm{SL}(d, \mathbb{Z})$ case

In this case the construction is more transparent, and follows closely that of [6]. The representation ρ_i of Condition A is the representation of $G = \mathrm{SL}(d, \mathbb{R})$ on the i'th exterior power of $\mathbb{R}^d$, and we can take $w_i = e_1 \wedge \cdots \wedge e_i$, where $\{e_1, \ldots, e_d\}$ is the standard basis for $\mathbb{R}^n$. We now recall some notation and results from [6]:

Let Δ be a lattice in $\mathbb{R}^d$. We say that a subspace L of $\mathbb{R}^d$ is Δ-*rational* if $L \cap \Delta$ is a lattice in L. For any Δ-rational subspace L, we denote by $d_\Delta(L)$ or simply by $d(L)$ the volume of $L/(L \cap \Delta)$. Let us note that $d(L)$ is equal to the norm of $u_1 \wedge \cdots \wedge u_\ell$ in the exterior power $\bigwedge^\ell(\mathbb{R}^d)$, where $\ell = \dim L$, $(u_1, \ldots, u_\ell)$ is a basis over $\mathbb{Z}$ of $L \cap \Delta$, and the norm on $\bigwedge(\mathbb{R}^d)$ is induced from the Euclidean norm on $\mathbb{R}^d$. If $L = \{0\}$ we write $d(L) = 1$. A lattice Δ is *unimodular* if $d_\Delta(\mathbb{R}^d) = 1$. The space of unimodular lattices is canonically identified with $\mathrm{SL}(d, \mathbb{R})/\mathrm{SL}(d, \mathbb{Z})$.

We introduce the following notation:

$$\alpha_i(\Delta) = \sup\left\{\frac{1}{d(L)} \;\middle|\; L \text{ is a } \Delta\text{-rational subspace of dimension } i\right\}, \quad 0 \le i \le d,$$
$$\alpha(\Delta) = \max_{0\le i\le d} \alpha_i(\Delta). \tag{3.3}$$

Lemma 3.2 ([6], Lemma 5.6). *For any two Δ-rational subspaces L and M*

$$d(L)d(M) \ge d(L\cap M)d(L+M). \tag{3.4}$$

In view of (1.2) we may write $\mu^{(n)} = \mu_1 + \mu_2$, where μ_1 has compact support, and for any $1 \le i \le d-1$,

$$\int_G \|\rho_i(g)\|^\delta \, d\mu_2(g) \le \frac{1-c}{3}. \tag{3.5}$$

Lemma 3.3. *Suppose μ satisfies condition A, and let n, δ and c be as in* (2.1). *Let $\mu_1 < \mu$ be any compactly supported measure. Then there exists a constant $\omega > 1$ such that for any lattice Λ in $\mathbb{R}^d$, and any $1 \le i \le d-1$,*

$$\int_G \alpha_i(g\Lambda)^\delta \, d\mu_1(g) < c\,\alpha_i(\Lambda)^\delta + \omega^{2\delta} \max_{0<j\le \min\{d-i,i\}} \left(\sqrt{\alpha_{i+j}(\Lambda)\alpha_{i-j}(\Lambda)}\,\right)^\delta. \tag{3.6}$$

Proof. Let Λ be a lattice in $\mathbb{R}^d$, and let M be a Λ-rational subspace of $\mathbb{R}^d$. Then, for any $g \in G$, gM is also a $g\Lambda$-rational subspace. By (2.1) applied to $v_1 \wedge \cdots \wedge v_l$, where $(v_1, \ldots, v_\ell)$ is a basis over $\mathbb{Z}$ of $M \cap \Lambda$, we have

$$\int_G \frac{1}{d_{g\Lambda}(gM)^\delta} \, d\mu_1(g) \le c\frac{1}{d(M)^\delta}. \tag{3.7}$$

There exists a Λ-rational subspace L_i of dimension i such that

$$\frac{1}{d_\Lambda(L_i)} = \alpha_i(\Lambda). \tag{3.8}$$

Inequality (3.7) implies

$$\int_G \frac{d\mu_1(g)}{d_{g\Lambda}(gL_i)^\delta} < c\frac{1}{d_\Lambda(L_i)^\delta}. \tag{3.9}$$

Let

$$\omega = \max_{\substack{g\in \mathrm{supp}\mu_1 \\ 0<j<d}} \|\rho_j(g)^{\pm 1}\|.$$

We have that for any Λ-rational subspace L, and any $g \in \mathrm{supp}\mu_1$,

$$\omega^{-1} \le \frac{d_{g\Lambda}(gL)}{d_\Lambda(L)} \le \omega. \tag{3.10}$$

Let us denote the set of Λ-rational subspaces L of dimension i with $d_\Lambda(L) < \omega^2 d_\Lambda(L_i)$ by Ψ_i. We get from (3.10) that for a Λ-rational i-dimensional subspace $L \notin \Psi_i$

$$d_{g\Lambda}(gL) \geq d_{g\Lambda}(gL_i), \quad g \in \mathrm{supp}\mu_1. \tag{3.11}$$

It follows from (3.9), (3.11) and the definition of α_i that

$$\int_G \alpha_i(g\Lambda)^\delta \, d\mu_1(g) < c\alpha_i(\Lambda)^\delta \quad \text{if } \Psi_i = \{L_i\}. \tag{3.12}$$

Assume now that $\Psi_i \neq \{L_i\}$. Let $M \in \Psi_i$, $M \neq L_i$. Then $\dim(M + L_i) = i + j$, $j > 0$. Now using (3.8), (3.10) and Lemma 3.2 we get that for any $g \in \mathrm{supp}\mu_1$,

$$\begin{aligned} \alpha_i(g\Lambda) < \omega\alpha_i(\Lambda) = \frac{\omega}{d_\Lambda(L_i)} &< \frac{\omega^2}{\sqrt{d_\Lambda(L_i)d_\Lambda(M)}} \\ &\leq \frac{\omega^2}{\sqrt{d_\Lambda(L_i \cap M)d_\Lambda(L_i + M)}} \\ &\leq \omega^2\sqrt{\alpha_{i+j}(\Lambda)\alpha_{i-j}(\Lambda)}. \end{aligned} \tag{3.13}$$

Hence, if $\Psi_i \neq \{L_i\}$ then

$$\int_G \alpha_i(g\Lambda)^\delta \, d\mu_1(g) \leq \omega^{2\delta} \max_{0<j\leq \min\{d-i,i\}} \left(\sqrt{\alpha_{i+j}(\Lambda)\alpha_{i-j}(\Lambda)}\right)^\delta. \tag{3.14}$$

Combining (3.12) and (3.14) we get that for any lattice $\Lambda \subset \mathbb{R}^n$, (3.6) holds. □

Lemma 3.4. *For sufficiently small $\epsilon > 0$ and $\delta > 0$ the function*

$$u(\Lambda) = \sum_{i=0}^{d} \epsilon^{i(d-i)}\alpha_i(\Lambda)^\delta \tag{3.15}$$

satisfies the conditions of Lemma 3.1.

Proof. Note that $u(\Lambda)$ satisfies (i) of Lemma 3.1. Let $\delta_0 > 0$ be some choice of δ satisfying (1.2) and Condition A, and let $\delta = \delta_0/d$. Since for any $1 \leq j \leq d-1$ and any $g \in G$, $\|\rho_j(g)\| \leq \|g\|^d$, we have for some constant C,

$$u(g\Lambda) \leq C\|g\|^{d\delta}u(\Lambda).$$

Hence, in view of (1.2), we may decompose $\mu^{(n)} = \mu_1 + \mu_2$ such that μ_1 is compactly supported and

$$\int_G u(g\Lambda) \, d\mu_2(g) \leq \frac{1-c}{3}u(\Lambda), \tag{3.16}$$

where $c < 1$ and $n > 0$ are as in (2.1). Let A_1 denote the averaging operator on G/Γ given by

$$(A_1 f)(\Lambda) = \int_G f(g\Lambda) \, d\mu_1(g). \tag{3.17}$$

Let $q(i) = i(d-i)$. Then by direct computations $2q(i) - q(i+j) - q(i-j) = 2j^2$. Therefore, we get from Lemma 3.3 that for any i, $0 < i < d$, and any positive $\epsilon < 1$

$$\begin{aligned} &A_1(\epsilon^{q(i)}\alpha_i^\delta) \\ &\quad < c\,\epsilon^{q(i)}\alpha_i^\delta + \omega^2 \max_{0<j\le\min\{d-i,i\}} \epsilon^{q(i)-\frac{q(i+j)+q(i-j)}{2}} \sqrt{\epsilon^{q(i+j)}\alpha_{i+j}^\delta \epsilon^{q(i-j)}\alpha_{i-j}^\delta} \\ &\quad \le c\,\epsilon^{q(i)}\alpha_i^\delta + \epsilon\,\omega^2 \max_{0<j\le\min\{d-i,i\}} \sqrt{\epsilon^{q(i+j)}\alpha_{i+j}^\delta \epsilon^{q(i-j)}\alpha_{i-j}^\delta}. \end{aligned} \tag{3.18}$$

Since $\epsilon^{q(i)}\alpha_i^\delta < u$, $\alpha_0 = 1$ and $\alpha_d = 1/d(\Delta) = 1$, the inequalities (3.18) imply the following inequality:

$$(A_1 u)(\Lambda) < 1 + 1 + c\,u(\Lambda) + d\epsilon\omega^2 u(\Lambda).$$

Taking $\epsilon = \frac{1-c}{3d\omega^2}$ we see that

$$(A_1 u)(\Lambda) = \int_G u(gx)\,d\mu_1(\Lambda) < \frac{1+2c}{3}u(\Lambda) + 2. \tag{3.19}$$

Now, in view of (3.19) and (3.16), (ii) of Lemma 3.1 also holds. This completes the proof of Lemma 3.4, and hence the proof of Proposition 2.6 in the $\mathrm{SL}(d,\mathbb{R})/\mathrm{SL}(d,\mathbb{Z})$ case. □

3.2 Construction of the function u in the general case

Let P_0 denote the minimal Γ-rational parabolic subgroup of G. Then we have the Langlands decomposition $P_0 = M_0 A_0 N_0$, where M_0 is semisimple, N_0 is the unipotent radical of P_0, and A_0 is as in the definition of Δ. From the general theory, $\Gamma \cap M_0$ is a cocompact lattice in M_0, and $\Gamma \cap N_0$ is a cocompact lattice in N_0. Let $\mathfrak{a}$ denote the Lie algebra of A_0. We identify $\mathfrak{a}$ with its dual using the Killing form.

Let $\alpha_1, \dots, \alpha_r$ denote the roots, which we view as elements of the dual of $\mathfrak{a}$. A *Siegel set* is a set $\mathfrak{S} = K\mathcal{M}\mathcal{A}\mathcal{N}$, where K is a maximal compact subgroup of G, $\mathcal{M} \in M_0$ and $\mathcal{N} \in N_0$ are compact, and $\mathcal{A} = \{a \in A : \alpha_k(\log a) < C,\ 1 \le k \le r\}$, where $C > 0$ is some positive constant. It follows from reduction theory that for appropriate choices of $\mathcal{M}$, $\mathcal{N}$ and C there exists a finite set $J \in G$ such that for every $g \in G$, the intersection $\mathfrak{S} \cap g\Gamma J$ is not empty.

Since $G = KP_0 = KM_0A_0N_0$ we may decompose

$$g_1 = k(g_1)m_0(g_1)\exp H(g_1)n_0(g_1), \tag{3.20}$$

where $k(g_1) \in K$, $m_0(g_1) \in M_0$, $\exp H(g_1) \in A_0$, and $n_0(g_1) \in N_0$. Then, for $g_1 \in \mathfrak{S}$, we have for any root α_j,

$$\alpha_j(-H(g_1)) > C, \tag{3.21}$$

where $C < 0$ is an absolute constant. (Hence $-H(g_1)$ is a finite distance from the positive Weyl chamber).

Let ρ_k, w_k be as in Condition A. Let $P_k = M_k A_k N_k$ be the Langlands decomposition of P_k, where as above $N_k \subset N_0$ is the unipotent radical of P_k, $M_k \supset M_0$ is semisimple and $A_k \subset A_0$ is one dimensional. Let $d_k(g) = \|\rho_k(gw_k)\|$. Then, by construction, if $g = kman$, where k is in the maximal compact K, $m \in M_k$, $a \in a_k$ and $n \in N_k$, then $d_k(g) = d_k(a)$. It follows from structure theory that $|\log d_k(a) - c_k\omega_k(\log a)| < C_1$, where $C_1 < 0$ and $c_k > 0$ are absolute constants, and ω_k is the coroot corresponding to the root α_k; i.e., $\omega_k(\alpha_k) = 1$, and $\omega_k(\alpha_j) = 0$ if $j \neq k$. Hence, if $g = km_0 \exp(H(g))n_0$, where $k \in K$, $m_0 \in M_0$, $a_0 = \exp(H(g)) \in A_0$, $n_0 \in N_0$, then $\log d_k(g) = \omega_k(\log a_0) = \omega_k(H(g))$.

Let

$$\beta_k(g) = \max_{\gamma \in \Gamma} \frac{1}{d_k(g\gamma)^{1/c_k}}. \tag{3.22}$$

It follows from reduction theory that up to an absolute constant, $\beta_k(g) = \beta_k(g_1)$, where g_1 is any element of $g\Gamma J \cap \mathfrak{S}$. Hence,

$$|\log \beta_k(g) - \omega_k(-H(g_1))| < C, \tag{3.23}$$

where $H(g_1)$ is as in (3.20), and C is an absolute constant. In particular, it follows that the function $\beta_k : G \to \mathbb{R}$ is bounded from below away from 0.

We may apply the identity

$$\alpha_k = \sum_j \langle \alpha_k, \alpha_j \rangle \omega_j \tag{3.24}$$

to $H(g_1)$ and combine with (3.21) and (3.23) in order to obtain that for all $g \in G$ and all k,

$$\sum_j \langle \alpha_k, \alpha_j \rangle \log \beta_j(g) > C', \tag{3.25}$$

where C' is an absolute constant.

Furthermore, for every constant C there exists a constant C' such that if for some k and some $g \in G$ and $g_1 \in g\Gamma J \cap \mathfrak{S}$ there exists $g_2 \in g\Gamma J$ such that $\frac{1}{d_k(g_2)} > \frac{1}{Cd_k(g_1)}$, $d_k(g_2) \neq d_k(g_1)$, then it follows from reduction theory that

$$\alpha_k(-H(g_1))) < C' \tag{3.26}$$

(i.e., $-H(g_1)$ is "near the k'th wall" of the positive Weyl chamber). Hence, in this case

$$\sum_j \langle \alpha_k, \alpha_j \rangle \log \beta_j(g) < C''. \tag{3.27}$$

Choose $q_j > 0$ such that $\sum_j q_j \omega_j$ belongs to the positive Weyl chamber. Then for any k,

$$\sum_j q_j \langle \alpha_j, \alpha_k \rangle > 0. \tag{3.28}$$

Let $u_j(g) = \beta_j(g)^{1/q_j}$, so that $\log \beta_j(g) = q_j \log u_j(g)$. Hence, if (3.27) holds,

$$\sum_j \langle \alpha_k, \alpha_j \rangle q_j \log u_j(g) < C''. \tag{3.29}$$

Since $\langle \alpha_k, \alpha_k \rangle > 0$ and $\langle \alpha_k, \alpha_j \rangle \leq 0$ for $j \neq k$, (3.29) may be rewritten as

$$u_k \leq C \prod_{j \neq k} u_j^{\lambda_{jk}}, \tag{3.30}$$

where

$$\lambda_{jk} = \frac{q_j |\langle \alpha_j, \alpha_k \rangle|}{q_k \langle \alpha_k, \alpha_k \rangle}. \tag{3.31}$$

Note that $\lambda_{jk} \geq 0$, and in view of (3.28),

$$\sum_{j \neq k} \lambda_{jk} < 1. \tag{3.32}$$

We note that the functions u_k satisfy the estimate

$$u_k(g'g) \leq \|g'\|_k^{1/q_k} u_k(g), \tag{3.33}$$

where $\| \cdot \|_k$ is the operator norm in the representation ρ_k.

Now choose $\delta > 0$ such that (1.2) and Condition A hold for δ/q_k instead of δ, $1 \leq k \leq r$. Let c and n be as in Condition A (with δ/q_k instead of δ). In view of (1.2) we may write $\mu^{(n)} = \mu_1 + \mu_2$, where μ_1 has compact support, and for any $1 \leq k \leq r$,

$$\int_G \|g\|_k^\delta \, d\mu_2(g) < \frac{1-c}{3}. \tag{3.34}$$

Let A_1 denote the averaging operator

$$(A_1 f)(g) = \int_G f(g'g) \, d\mu_1(g').$$

Let $n > 0$ be as in Condition A, and for sufficiently small $\delta > 0$ and $g \in G$ consider the average $(A_1 u_k^\delta)(g)$. If for every h in the support of μ_1 the maximum in the definition of β_k (hence of u_k) is achieved by the same $\gamma \in \Gamma$, then by Condition A we have $(A_1 u_k)(g) \leq c u_k(g)$, with $c < 1$. If the maximum is achieved by different γ depending on the choice of g' in the (compact) support of μ_1, then (3.26) holds. Hence, in that case (3.30) also holds. Thus, in all cases we obtain the system of

inequalities:

$$A_1 u_k^\delta \le c u_k^\delta + C' \prod_{j \neq k} u_j^{\delta \lambda_{jk}} + B', \tag{3.35}$$

where $c < 1$, $C' < \infty$, $B' < \infty$, and we have used again the compactness of the support of μ_1. The additive constant B' arises because we have assumed that the u_i are bounded from below.

For $\epsilon > 0$, (3.35) may be rewritten as

$$A_1(\epsilon u_k)^\delta \le c(\epsilon u_k)^\delta + \epsilon_1 \prod_{j \neq k} (\epsilon u_j)^{\delta \lambda_{jk}} + B'',$$

and in view of (3.32), $\epsilon_1 > 0$ can be made arbitrarily small by choosing $\epsilon > 0$ small enough. Note that by (3.32), Jensen's inequality and the fact that the u_j are bounded from below we have

$$\prod_{j \neq k} (\epsilon u_j)^{\delta \lambda_{jk}} \le C_1 \sum_{j \neq k} (\epsilon u_j)^\delta.$$

Hence $u = \sum_k (\epsilon u_k)^\delta$ satisfies the inequality

$$A_1 u \le c' u + b,$$

where $c' < 1 - \frac{(1-c)}{3}$ and $b > 0$. Now, in view of (3.34) we have

$$Au \le c_1 u + b, \tag{3.36}$$

where $c_1 < 1$ and $(Af)(g) = \int_G f(g'g)\, d\mu^{(n)}(g')$.

4 Averaging operators

In this section we prove Proposition 2.7.

Lemma 4.1. *Suppose H is a semisimple algebraic subgroup of* $\mathrm{GL}(V)$ *without compact factors, such that V does not have any H-invariant vectors. Suppose μ a Zariski dense measure on H satisfying* (1.2). *Then there exists $N > 0$ such that for all $n > N$ and all nonzero $v \in V$,*

$$\frac{1}{n} \int_H \log \frac{\|hv\|}{\|v\|}\, d\mu^{(n)}(h) > c > 0.$$

Proof. See [1, Chapter III, Corollary 3.4 p. 53-54]. Also see [7] and the original paper [8] for closely related statements. □

Lemma 4.2. *Suppose H is a semisimple algebraic subgroup of* $\mathrm{GL}(V)$ *without compact factors such that V does not contain any H-invariant vectors. Then for all*

sufficiently small $\delta > 0$ there exist c, $0 < c < 1$, and $N > 0$ such that for all $n > N$ and all $v \in V$,

$$\int_H \|hv\|^{-\delta}\, d\mu^{(n)}(h) < c\|v\|^{-\delta}. \tag{4.1}$$

Proof. Without loss of generality, we assume that $\|v\| = 1$. We first show that there exists $n > 0$ for which (4.1) holds. Let n, c be as in Lemma 4.1. Note that if $\|hv\| \geq 1$, then $\|hv\|^{-\delta} \leq 1 - \delta \log \|hv\|$. If $\|hv\| \leq 1$, then using Taylor's theorem, for $\delta \in [0, \delta_0]$,

$$\|hv\|^{-\delta} \leq 1 - \delta \log \|hv\| + \frac{1}{2}\delta^2 (\log \|hv\|)^2 \|hv\|^{-\delta_0}.$$

Hence, using Lemma 4.1,

$$\begin{aligned}
\int_H \|hv\|^{-\delta}\, d\mu^{(n)}(h) &\leq 1 - \delta c + \frac{1}{2}\delta^2 \int_H (\log \|hv\|)^2 \|hv\|^{-\delta_0}\, d\mu^{(n)}(h) \\
&\leq 1 - \delta c + \frac{1}{2}\delta^2 \int_H (\log \|h^{-1}\|)^2 \|h^{-1}\|^{\delta_0}\, d\mu^{(n)}(h) \\
&\leq 1 - \delta c + \delta^2 C_n(\delta_0),
\end{aligned}$$

where in the last line we have used the condition (1.2). Now choose $\delta < C_n(\delta_0)/c$. Then (4.1) holds for n. To see that it holds for $m > n$, note that $\mu^{(m)} = \mu^{(n)} * \mu^{(m-n)}$, and $\mu^{(m-n)}$ is a probability measure. □

Proof of Proposition 2.7. Let P_k, ρ_k, V_k, w_k be as in condition A. Let $V_k^{H_\mu}$ denote the H_μ-invariant subspace of V_k on which the action of H_μ is trivial, and let V_k' denote the complementary H_μ-invariant subspace of V_k. The assumption of Proposition 2.7 implies that Gw_k does not intersect $V_k^{H_\mu}$. Let π_k' denote the projection onto V_k'. Let K denote the maximal compact subgroup of G. Since $G = KP_k$, and P_k stabilizes the span of w_k, Gw_k is projectively compact. Hence there exists a constant $c_0 > 0$ such that for all $v \in Gw_k$, $\|\pi_k'(v)\| > c_0\|v\|$. Now the proposition follows from Lemma 4.2 applied to $\pi_k'(v)$. □

5 Proof of Lemma 2.2

Let $X = G/\Gamma$, and let σ be a locally finite μ-stationary measure on X. Consider the space $G \times X$ with the measure $\mu \times \sigma$. Then, since σ is μ-stationary, for any (compact) $K \subset X$,

$$(\mu \times \sigma)\{(g, x) \,:\, gx \in K\} = \sigma(K).$$

Hence,

$$\int_X \mu\{\, g \,:\, gx \in K\}\, d\sigma(x) = \sigma(K).$$

Hence, for any compact subset C of X,

$$\int_C \mu\{\, g \,:\, gx \in K\}\, d\sigma(x) \leq \sigma(K).$$

Now, after convolving we see that for any $n > 0$,

$$\int_C \mu^{(n)}\{\, g \,:\, gx \in K\}\, d\sigma(x) \leq \sigma(K).$$

But by (R2), for n sufficiently large, for any $x \in C$, $\mu^{(n)}\{\, g \,:\, gx \in K\} \geq (1 - \epsilon)$. Then,

$$\sigma(C) \leq \frac{\sigma(K)}{1-\epsilon}.$$

Since C is arbitrary, we see that $\sigma(X) \leq \frac{\sigma(K)}{1-\epsilon}$.

References

[1] P. Bougerol and J. Lacroix, Products of Random Matrices with Applications to Schrödinger Operators, Progr. Probab. Statist. 8, Birkhäuser, Boston MA 1985.

[2] S. G. Dani, On invariant measures, minimal sets and a lemma of Margulis, Invent. Math. 51 (1979), 239–260.

[3] S. G. Dani, Invariant measures and minimal sets of horospherical flows, Invent. Math. 64 (1981), 357–385.

[4] S. G. Dani, On orbits of unipotent flows on homogeneous spaces, Ergodic Theory Dynam. Systems 4 (1984), 25–34.

[5] S. G. Dani, On orbits of unipotent flows on homogeneous spaces. II, Ergodic Theory Dynam. Systems 6 (1986), 167–182.

[6] A. Eskin, G. Margulis and S. Mozes, Upper bounds and asymptotics in a quantitative version of the Oppenheim conjecture, Ann. of Math. (2) 147 (1998), 93–141.

[7] A. Furman, Random walks on groups and random transformations, in: Handbook of Dynamical Systems, Vol. 1A, North-Holland, Amsterdam 2002, 931–1014.

[8] H. Furstenberg, Noncommuting random products, Trans. Amer. Math. Soc. 108 (1963), 377–428.

[9] G. A. Margulis, On the action of unipotent groups in the space of lattices, in: Lie Groups and their Representations (Proc. Summer School, Bolyai, Janos Math. Soc., Budapest, 1971), Halsted, New York 1975, 365–370.

[10] N. A. Shah, Invariant measures and orbit closures on homogeneous spaces for actions of subgroups generated by unipotent elements, in: Lie Groups and Ergodic Theory (Mumbai, 1996), Tata Inst. Fund. Res. Stud. Math. 14, Tata Inst. Fund. Res., Bombay 1998, 229–271.

Alex Eskin, Department of Mathematics, University of Chicago, Chicago, IL 60637, USA
E-mail: eskin@math.uchicago.edu

Gregory Margulis, Department of Mathematics, Yale University, New Haven, CT 06520, USA
E-mail: margulis@math.yale.edu

On the cohomology of foliations with amenable groupoid

Alessandra Iozzi

Abstract. We illustrate the proof of a vanishing theorem for the tangential de Rham cohomology of a compact foliated space with amenable fundamental groupoid, by using the existence of bounded primitives of closed bounded differential forms in degree above the rank (for an appropriate notion). In the case of foliated bundles we give a proof of a related theorem asserting the vanishing of the tangential singular cohomology, by using methods in homological algebra.

1 A discussion of the main result

Given a differentiable manifold M, it is a classical problem to study the relation between the topology and the geometry of M, in particular, which restriction the fundamental group of M imposes on the possible Riemannian geometries of M. A fundamental result in this direction is the following:

Theorem 1.1 ([15, 8]). *Let M be a compact Riemannian manifold with non-positive sectional curvature $\kappa \leq 0$ and solvable fundamental group $\pi_1(M)$. Then $\kappa = 0$ (and, in fact, $\pi_1(M)$ is virtually abelian).*

More generally,

Theorem 1.2 ([16]). *Let M be a compact Riemannian manifold such that $\kappa \leq 0$ and $\pi_1(M)$ is amenable. Then $\kappa = 0$.*

We refer the reader to § 2.3 for a discussion of amenability and related topics, and we limit ourselves to point out here that solvable groups are amenable.

The purpose of this paper is to illustrate some results whose motivation stems from the proof of Theorem 1.2 for negatively curved manifolds which is due to Gromov and Thurston and can be summarized in two steps:

1. The bounded cohomology $\mathrm{H}^\bullet_\mathrm{b}(X)$ of any topological space X is defined as the singular cohomology of X, where we restrict our attention only to bounded cochains, that is cochains c such that

$$\|c\|_\infty = \sup\{|c(\sigma)| : \sigma \text{ is a singular simplex in } M\} < \infty.$$

Then we have the following striking result (for a complete proof see [10]):

Theorem 1.3 ([9, 3]). *For any countable* CW *complex X,*

$$\mathrm{H}_{\mathrm{b}}^{\bullet}(X) \simeq \mathrm{H}_{\mathrm{b}}^{\bullet}(\pi_1(X)).$$

Since the bounded cohomology of an amenable group vanishes (see for instance Remark 2.9 and Corollary 4.2 with $T = \{\mathrm{pt}\}$), it follows that if X has amenable fundamental group, then $\mathrm{H}_{\mathrm{b}}^{\bullet}(X) = 0$.

2. The second part of the proof follows from the following result:

Theorem 1.4 ([14]). *Let M be a compact manifold with strictly negative sectional curvature. Then there is a surjection*

$$\mathrm{H}_{\mathrm{b}}^{j}(M) \twoheadrightarrow \mathrm{H}^{j}(M)$$

in degree $j > 1$.

Hence for a compact manifold, since $\mathrm{H}^{\dim(M)}(M) \neq 0$, Theorems 1.3 and 1.4 imply the incompatibility between the amenability of the fundamental group and strictly negative sectional curvature. However, if one extends the realm of generality of the above results, one can obtain the following vanishing theorem:

Theorem 1.5 ([6]). *Let $(X, \mathcal{F})$ be a compact foliated space whose leaves are uniformly of rank at most r. If the fundamental groupoid of the foliation is amenable, then the tangential de Rham cohomology $\mathrm{H}_{\mathrm{dR}}^{j}(X, \mathcal{F})$ vanishes for all $j > r$.*

We refer the reader to § 2 for all the relevant definitions. However, we mention here that a prominent example of such situation is a compact space foliated by locally symmetric spaces of $\mathbb{R}$-rank r. Moreover, tangential de Rham cohomology has been considered by several authors with various degrees of regularity in the direction transverse to the leaves (see [13], for example, for an extensive list of references), thus obtaining different theories (see § 2.1 for an example).

The initial approach to the proof of Theorem 1.5 was along the lines of Gromov's proof of Theorem 1.2, in the special case of foliated bundles whose leaves have strictly negative curvature. The proof that eventually appeared in print in [6], and whose outline is presented in § 3, does not make any use of bounded cohomology, but uses rather a direct approach via an analogue of the Poincaré Lemma with estimates (Lemma 3.1). In that original approach Gromov's definition of bounded cohomology was used. Here we want to present instead a proof of a related vanishing theorem in the special case of foliated bundles (see Example 2.1 for the definition), in which the functorial approach to the bounded cohomology of locally compact groups developed by Burger and Monod in [5] is exploited. Although the definitions are *ad hoc*, it indicates a possible use of a systematic development of the theory of the bounded cohomology of groupoids applied to general foliations.

Theorem 1.6. *Let Y be a compact locally* CAT(-1) *space with fundamental group* Γ *and universal covering* $\widetilde{Y}$. *If* (T, μ) *is a standard measure space with a measure class preserving amenable* Γ*-action, and* $X = (Y \times T)/\Gamma$, *then the tangential singular cohomology* $\mathrm{H}^j_{\mathrm{s}}(X, \mathcal{F})$ *vanishes for all* $j > 1$.

2 Definitions and examples

We collect here the definitions needed in the sequel. We shall often prefer to give illustrative examples rather than technical definitions.

2.1 Foliations

Let $(X, \mathcal{F})$ be a topological space X with a foliation $\mathcal{F}$ whose leaves are smooth Riemannian manifolds and such that the Riemannian structure is smooth along the leaves and globally continuous. Assume that there is a measure λ which is obtained by combining a transverse measure, whose class is invariant under holonomy, with the Lebesgue measure along the leaves.

Example 2.1.

- Any locally free smooth action of a connected Lie group on a manifold determines a foliation.
- The space $X = \mathbb{R}^p \times \mathbb{R}^{n-p}$ is a foliation, and, in fact, any foliation looks locally like a product $U \times \Sigma$, where $U \subset \mathbb{R}^{n-p}$ is an open set and Σ is a topological space. More generally, if Y is a Riemannian manifold and Σ is a topological space, then $X = Y \times \Sigma$ is a topological space with a foliation whose leaves are $Y \times \{\sigma\}$, $\sigma \in \Sigma$.
- If Y and Σ are as above, if Γ acts properly discontinuously on Y and with no fixed points, and, moreover, if Γ acts on Σ, then $X = (Y \times \Sigma)/\Gamma$ is a topological space with a foliation with leaves $(Y \times \{\sigma\})/\Gamma_\sigma$, where Γ_σ is the stabilizer of $\sigma \in \Sigma$. The foliated space X is often referred to as a foliated bundle.

Since the leaves of the foliation are Riemannian manifolds, they admit tangent spaces which can then be assembled together to form the foliated tangent space $T\mathcal{F}$. Let $T^*\mathcal{F}$ be the foliated cotangent bundle and $\Lambda^j T^*\mathcal{F}$ be its j-th exterior power.

Definition 2.2 ([6]). If $(X, \mathcal{F})$ is a foliated space, its tangential de Rham cohomology $\mathrm{H}^\bullet_{\mathrm{dR}}(X, \mathcal{F})$ is the cohomology of the complex

$$\Omega^j(X, \mathcal{F}) = \{\omega : X \to \Lambda^j T^*\mathcal{F} \ : \ \omega, d\omega \in L^\infty(X, \Lambda^j T^*\mathcal{F}), \text{ and } \omega, d\omega \text{ are } C^\infty \text{ along the leaves}\},$$

where the differential is taken in the direction of the leaves, and where

$$\begin{aligned}\|\omega\|_\infty &= \operatorname*{esssup}_{x\in X} \|\omega_x\| \\ &= \operatorname*{esssup}_{x\in X} \sup\{|\omega_x(v_1\wedge\cdots\wedge v_j)| : \ v_1,\dots,v_j \in T_x\mathcal{F} \text{ are orthonormal}\}.\end{aligned}$$

As mentioned in § 1, one can choose to require various degrees of regularity in the direction transversal to the leaves. For instance, if one takes differential forms which are just measurable on the total space without any assumption of boundedness, then it was observed by Zimmer that the tangential de Rham cohomology thus defined vanishes in degree above one, provided that almost every leaf is contractible (see [6]).

2.2 Fundamental groupoid

Definition 2.3. A groupoid $\mathcal{G}$ is a small category in which each morphism is an isomorphism.

Hence the information which characterizes a groupoid is encoded by the set of units $\mathrm{Obj}(\mathcal{G})$ and the set of morphisms $\mathrm{Mor}(\mathcal{G})$. We have moreover source and target maps, $s, t : \mathrm{Mor}(\mathcal{G}) \to \mathrm{Obj}(\mathcal{G})$ which determine when two morphisms m_1, m_2 are composable, namely, if and only if $s(m_2) = t(m_1)$, in which case the multiplication is $(m_1, m_2) \mapsto m_1 \circ m_2$. A few examples will serve the purpose of clarifying this concept:

Example 2.4.

- Let G be a group acting on a space X. Then the groupoid $\mathcal{G}$ associated to the action is such that $\mathrm{Obj}(\mathcal{G}) = X$ and $\mathrm{Mor}(\mathcal{G}) = \{(x, g) \in X \times G\}$; moreover, $s : \mathrm{Mor}(\mathcal{G}) \to \mathrm{Obj}(\mathcal{G})$ and $t : \mathrm{Mor}(\mathcal{G}) \to \mathrm{Obj}(\mathcal{G})$ are respectively defined by $s(x, g) := x$ and $t(x, g) := xg$, and two morphisms (x, g) and (x', g') are composable if and only if $xg = x'$, in which case $(x, g) \circ (x', g') = (x, gg')$.

- Let $\mathcal{R} \subset X \times X$ be an equivalence relation on X. Then the groupoid $\mathcal{G}_\mathcal{R}$ associated to $\mathcal{R}$ is such that $\mathrm{Obj}(\mathcal{G}_\mathcal{R}) = X$ and $\mathrm{Mor}(\mathcal{G}_\mathcal{R}) = \{(x, y) \in \mathcal{R}\}$; Here $s(x, y) = x$ and $t(x, y) = y$, so that two morphisms $(x, y), (z, w) \in \mathcal{R}$ are composable if and only if $y = z$, in which case $(x, y) \circ (y, w) = (x, w)$.

- If X is any topological space, its fundamental groupoid $\mathcal{G}_X$ is such that $\mathrm{Obj}(\mathcal{G}_X) = X$ and $\mathrm{Mor}(\mathcal{G}_X)$ is the set of homotopy classes (with fixed endpoints) of paths. Evidently, two morphisms are composable if and only if the endpoint of a path (or, more precisely, of an equivalence class of paths) coincides with the beginning point of the other path.

- As a generalization of the previous example, we finally have the definition of the fundamental groupoid of a foliated topological space:

Definition 2.5. If $(X, \mathcal{F})$ is a foliated topological space, the fundamental groupoid of the foliation $\mathcal{G}_{(X,\mathcal{F})}$ is the groupoid whose set of units $\mathrm{Obj}(\mathcal{G}_{(X,\mathcal{F})})$ is X and whose set of morphisms $\mathrm{Mor}(\mathcal{G}_{(X,\mathcal{F})})$ is the set of homotopy classes (endpoints fixing) of paths contained in a leaf.

2.3 Amenability

One of the many classical equivalent definitions of amenability of a topological group G requires that for every compact metric space X on which G acts continuously, there exists on X a G-invariant Borel probability measure μ. Note that the space $C(X)$ of continuous functions on X with the supremum norm is a separable Banach space with an isometric G-action, and the space of Borel probability measures $\mathcal{M}(X)$ is a compact convex G-invariant subset of the unit ball of the dual $C^*(X)_1$ (in the weak* topology). Then an invariant measure $\mu \in \mathcal{M}(X)$ is nothing but a fixed point for the G-action on $\mathcal{M}(X)$, and one is hence lead to the following definition:

Definition 2.6. A group G is amenable if and only if there exists a fixed point in any affine G-space, that is in any compact convex G-invariant subset $A \subset E_1^*$ in the unit ball of the dual of a separable Banach space, on which G acts isometrically and continuously.

We mentioned already that cyclic groups and, more generally, solvable groups are amenable (see for instance [17, Ch. 4, § 1]). Moreover, we shall use in what follows that, among the parabolic subgroups of Lie groups, the only ones which are amenable are the minimal parabolics.

In order to extend the definition of amenability of a group to a groupoid, we first need to define the notion of action of a groupoid. Let E be a separable Banach space, $V \to X$ an isometric Banach bundle with fiber E (that is, a fiber bundle with fiber E such that there is a covering of X and a corresponding trivialization of V with transition functions in $\mathrm{Iso}(E)$), and let $V^* \to X$ be its dual Banach bundle. If V_x is the fiber of $V \to X$ above the point $x \in X$, let $\mathrm{Iso}(V)$ be the groupoid with $\mathrm{Obj}(\mathrm{Iso}(V)) = X$ and morphisms $\mathrm{Mor}(\mathrm{Iso}(V)) = \left\{ \mathrm{Iso}(V_x, V_y) : x, y \in X \right\}$, that is the linear isomorphisms between fibers.

Definition 2.7. An action of a groupoid $\mathcal{G}$ on V is a functor from $\mathcal{G}$ to $\mathrm{Iso}(V)$ which is the identity on objects, that is a map

$$\begin{array}{rccc} \rho : & \mathrm{Mor}(\mathcal{G}) & \to & \mathrm{Mor}(\mathrm{Iso}(V)) \\ & (g : x \to y) & \mapsto & (\rho(g) : V_x \to V_y) \end{array}$$

such that $\rho(gh) = \rho(g)\rho(h)$ whenever g and h are composable.

Once we have an action of $\mathcal{G}$ on V, a field of compact convex subsets of V^* parameterized by X is a subset $\mathcal{A} \subset V^*$ such that each subset $A_x \subset V_x^*$ is a compact

convex subset of $(V_x)_1^*$. We say that $\mathcal{A}$ is ρ-invariant if for any morphism $g : x \to y$ in $\mathrm{Mor}(\mathcal{G})$ and almost every $x \in X$, we have that $\rho(g^{-1})^* A_x \subset A_y$, where $\rho(g^{-1})^* : V_x^* \to V_y^*$. We finally have:

Definition 2.8. A groupoid $\mathcal{G}$ is amenable is for every Borel representation of $\mathcal{G}$ on an isometric Banach bundle $V \to X$ with separable fiber and any ρ-invariant Borel field $\mathcal{A}$ of compact convex subsets of V^*, there exists a ρ-invariant section of $\mathcal{A}$, that is a Borel map $s : X \to V^*$ with $s(x) \in A_x$ and such that $\rho(g^{-1})^*(s(x)) = s(y)$ for almost every $x \in X$ and all morphisms $g : x \to y$.

Remark 2.9.

- If $\mathcal{G}$ is the groupoid of an action, then $\mathcal{G}$ is amenable if and only if the action is amenable [17, Definition 4.3.1].

- Recall that a transitive action is amenable if and only if the stabilizer of a point is amenable. More generally, an action is amenable if and only if the equivalence relation of the action is amenable and the stabilizers are amenable ([1] or [2]). Analogously, the fundamental groupoid of a foliation is amenable if and only if the foliation is amenable (that is, the equivalence relation induced on any transversal is amenable) and the fundamental groups of the leaves are amenable (for example, see [2]).

We can now give examples of foliations with amenable fundamental groupoid.

Example 2.10.

- Let M be a compact Riemannian manifold with negative sectional curvature and universal cover $\widetilde{M}$, and let $\Sigma = \widetilde{M}(\infty)$ be the set of equivalence classes of asymptotic geodesic rays. If $\Gamma = \pi_1(M)$, then $X = (\widetilde{M} \times \Sigma)/\Gamma$ is a foliated space with amenable fundamental groupoid, since the equivalence relation of the transversal Σ is amenable, and the fundamental group Γ_σ of the leaf $\mathcal{L}_\sigma = (\widetilde{M} \times \{\sigma\})/\Gamma_\sigma$ is amenable because cyclic.

- Let Y be a symmetric space of noncompact type, $G = \mathrm{Iso}(Y)$ be its isometry group (hence a semisimple group), $\Gamma < G$ a cocompact torsionfree lattice, and Q a parabolic subgroup. Then $X = (Y \times (G/Q))/\Gamma$ is a space foliated by leaves $(Y \times [x])/\Gamma_{[x]}$, and the fundamental groupoid of the foliation is amenable if and only if Q is minimal parabolic. Note that in this case the nonamenability of Q is reflected in the nonamenability of the foliation, although the fundamental groups of the leaves might still be amenable. This is the case, for instance, if $G = \mathrm{SL}((p-1)/2, \mathbb{C})$ for p a prime congruent to 3 modulo 4 and Q is the parabolic subgroup which stabilizes the vector $(1, 0, \ldots, 0) \in \mathbb{C}^{(p-1)/2}$, in which case one might choose Γ so that for each $[x] = gQ$ the leaf $\mathcal{L}_{[x]}$ has abelian fundamental group $\Gamma_{[x]} = g^{-1}\Gamma g \cap Q$, [6, Corollary 4.5].

2.4 Rank of a manifold

The notion of rank that is needed in Theorem 1.5 is somewhat different from any of the standard definitions. We say that a manifold M of nonpositive curvature has rank r at a point m and with respect to a tangent vector $v \in TM_m$ if r is the largest dimension of a subspace $W \subset TM_m$ containing v such that every plane in W containing v has sectional curvature zero. The uniform notion of rank that is needed is then the following:

Definition 2.11. Let M be a complete simply connected Riemannian manifold with nonpositive sectional curvature. We say that M is uniformly of rank at most r if there is a positive constant C such that, for every subspace of dimension $r+1$ of every tangent space to M and every nonzero vector v in the subspace, there is a plane with sectional curvature at most $-C$ containing v.

Notice that if M is a symmetric space this notion of rank coincides with the usual one in terms of maximal dimension of flats.

2.5 Remarks

We give here some indication of examples which show that the hypotheses of Theorem 1.5 are sharp. For instance, one cannot expect to have vanishing of the tangential cohomology in degree smaller than or equal to the rank of the manifold, since, already for the one leaf foliation consisting of a flat torus, the de Rham cohomology does not vanish in top degree.

Moreover, also the full strength of the amenability of the fundamental groupoid is necessary. In fact, on the one hand one can consider once again the foliation consisting of just one leaf which is a compact quotient of a symmetric space of noncompact type. In this case the equivalence relation on a transversal is amenable (being the trivial one), but the fundamental group of the leaf is typically not amenable. In many of such examples one has nonvanishing of the de Rham cohomology in degree above the rank, as one can see for instance by taking any compact quotient of any symmetric space of noncompact type, in which case the volume form gives a nonvanishing class in top degree.

On the other hand, one can construct examples of foliated bundles with nonamenable equivalence relation, but such that the leaves have abelian fundamental groups, and for which the tangential de Rham cohomology groups do not vanish in some degree above the rank. In fact:

Proposition 2.12 ([6]). *For $n \neq 3$, let $G = \mathrm{SL}(n, \mathbb{C})$, $\Gamma < G$ a cocompact lattice, $Q < G$ the parabolic subgroup which stabilizes the vector $(1, 0, \ldots, 0) \in \mathbb{C}^n$, and $Y = G/\mathrm{SU}(n)$. Then for all j odd, with $3 \leq j \leq 2n-3$, $\mathrm{H}^j_{\mathrm{dR}}((Y \times (G/Q))/\Gamma) \neq 0$.*

Collecting the information from the above proposition and from Example 2.10, that if $p \geq 7$ is a prime such that $p \equiv 3 \pmod 4$, for all $(p-1)/2 \leq j \leq p-4$ and j odd, we have that $\mathrm{H}^j_{\mathrm{dR}}((Y \times (G/Q))/\Gamma \neq 0$ despite the fact that $\mathbb{R}$-rank$(\mathrm{SL}((p-1)/2, \mathbb{C})) = (p-3)/2$.

We want to conclude this section by mentioning a possible relation between our theorem and the main theorem in [16]. There Zimmer considered the case of a measure space X with a Riemannian measurable foliation $\mathcal{F}$ of finite total volume, such that almost every leaf is a complete simply connected manifold of nonpositive sectional curvature. He proved that if the foliation is amenable and if there exists a transversally invariant measure, then almost every leaf is flat. Although this theorem is much more general in that, for example, there is no rank assumption on the leaves, Theorem 1.5 should imply this result in the case in which both can be applied. In fact, in view of the simple connectivity of the leaves, amenability of the foliation coincides with amenability of the fundamental groupoid. Now suppose that the leaves satisfy the uniform rank condition in Definition 2.11, for instance are locally symmetric spaces of dimension n. Then, if one were to prove an analogue of a theorem of Ruelle and Sullivan (see [13, Corollary 4.25], for example), the existence of an absolutely continuous transversally invariant measure would imply the existence of a nonzero class in $\mathrm{H}^n_{\mathrm{dR}}(X, \mathcal{F})$. Hence, by Theorem 1.5, we must have that $n \leq \mathbb{R}$-rank(G), that is the leaves are flat.

3 A sketch of the proof of Theorem 1.5

The idea is simple. For each leaf $\mathcal{L}$ of the foliation and each leafwise closed differential form α of degree at least equal to the rank, there exists a canonical convex set of bounded primitives of α, once α is restricted to the leaf $\mathcal{L}$ and lifted to its universal cover $\widetilde{\alpha}$. Then, by using the amenability of the fundamental groupoid, it is possible to choose primitives from these convex sets coherently for all leaves. More specifically:

Lemma 3.1. *Let M be a complete simply connected Riemannian manifold with nonpositive sectional curvature which is uniformly of rank at most r, and let $\alpha \in \Omega^j(M)$ be a bounded smooth closed differential j-form, $r < j \leq \dim M$. If $M(\infty)$ is the boundary consisting of equivalence classes of asymptotic geodesic rays, then there exists a Borel map $\beta : M(\infty) \to \Omega^{j-1}(M)$, $\beta(\xi) := \beta_\xi$, such that $d\beta_\xi = \alpha$ and $\|\beta\| = \sup_\xi \|\beta_\xi\| < \infty$. Moreover β is equivariant with respect to isometries.*

The proof of the lemma is basically the same as the proof of the Poincaré Lemma with estimates. Let $\varphi_\xi(t)$ be the gradient flow associated to the gradient vector field of the Busemann function $b_\xi : M \to \mathbb{R}$. Define a map $\Phi_\xi : M \times [0,1] \to M$, by $\Phi_\xi(m,t) = \varphi_\xi(t)(m)$ to use as a homotopy in the classical Poincaré Lemma. Namely, if $\Phi_\xi^*(\alpha) = \omega_0(t) + \omega_1(t) \wedge dt$, define $\beta_\xi = \int_0^\infty \omega_1(t)dt$. Note that the existence of

the map β uses the fact that $\varphi_\xi(t)$ is a contraction on j tangent vector, $j \geq r+1$, i.e., that $\|\varphi_\xi(t)(X_1 \wedge \cdots \wedge X_k)\|$ decays exponentially.

We observe now the first consequence of the amenability of the fundamental groupoid $\mathcal{G}_{(X,\mathcal{F})}$, for which we need to define an appropriate action. Let $\mathcal{L}_{x_0}$ be the leaf through $x_0 \in X$, and if $x \in \mathcal{L}_{x_0}$, let

$$\widetilde{\mathcal{L}_x} = \{([d], z) : z \in \mathcal{L}_{x_0}, [d] \text{ is a homotopy class of paths from } x \text{ to } z\}$$

be its universal covering based at x. If y is another point in $\mathcal{L}_{x_0}$, any homotopy class $[c]$ from x to y defines an isometry $\rho([c]) : \widetilde{\mathcal{L}_x} \to \widetilde{\mathcal{L}_y}$, by $\rho([c])([d], z) = ([c^{-1} \circ d], z)$, which extends to a homeomorphism of the associated ideal boundaries $\rho([c]) : \widetilde{\mathcal{L}_x}(\infty) \to \widetilde{\mathcal{L}_y}(\infty)$. Since $\widetilde{\mathcal{L}_x}(\infty)$ is compact and metrizable, for every $x \in X$ the space of continuous functions $C(\widetilde{\mathcal{L}_x}(\infty))$ is a separable Banach space, so that we can consider the isometric Banach bundle $V \to X$ with fiber $C(\widetilde{\mathcal{L}_x}(\infty))$, on which $\mathcal{G}_{(X,\mathcal{F})}$ acts via $\rho : \mathrm{Mor}(\mathcal{G}_{(X,\mathcal{F})}) \to \mathrm{Mor}(\mathrm{Iso}(V))$.

Hence we have a field of compact convex subsets of V^* parameterized by X, $x \to \mathcal{M}(\widetilde{\mathcal{L}_x}) \subset C(\widetilde{\mathcal{L}_x})^*_1$ consisting of probability measures on $\widetilde{\mathcal{L}_x}$, which can be easily seen to be $\mathcal{G}_{(X,\mathcal{F})}$-invariant. The amenability of $\mathcal{G}_{(X,\mathcal{F})}$ implies the existence of a $\mathcal{G}_{(X,\mathcal{F})}$-invariant Borel section $s : X \to \mathcal{M}(\widetilde{\mathcal{L}_x})$.

To conclude, let now $\alpha \in \Omega^j(X, \mathcal{F})$ be a closed form. If $p_x : \widetilde{\mathcal{L}_x} \to \mathcal{L}_x$ is the projection, let us consider $p_x^*(\alpha|_{\mathcal{L}_x}) \in \Omega^j(\widetilde{\mathcal{L}_x})$, where $\alpha|_{\mathcal{L}_x}$ is the restriction of α to $\mathcal{L}_x$. By Lemma 3.1, there exists a Borel map $\beta : \widetilde{\mathcal{L}_x}(\infty) \to \Omega^{j-1}(\widetilde{\mathcal{L}_x})$ such that $d\beta_\xi = p_x^*(\alpha|_{\mathcal{L}_x})$ for every $\xi \in \widetilde{\mathcal{L}_x}$, and such that β is bounded uniformly in ξ. Define now

$$\beta_x = \int_{\widetilde{\mathcal{L}_x}(\infty)} \beta_\xi \, ds_x(\xi) \in \Omega^{j-1}(\widetilde{\mathcal{L}_x}),$$

which has still the property that $d\beta_x = p_x^*(\alpha|_{\mathcal{L}_x})$. Now we use twice the invariance of the section s. Firstly, since s is invariant for morphisms $x \mapsto x$ (that is, for homotopy paths in $\pi_1(\mathcal{L}_x)$), we obtain that there exists $\omega_x \in \Omega^j(\mathcal{L}_x)$ such that $\beta_x = p_x^*(\omega_x)$; secondly, since s is $\mathcal{G}_{(X,\mathcal{F})}$-invariant (that is, invariant with respect to all morphisms $x \mapsto y$), we deduce that the differential form ω_x is independent of the choice of the basepoint, namely that $\omega_x = \omega_y$ if $\mathcal{L}_x = \mathcal{L}_y$. We have hence defined a tangential form $\omega \in \Omega^{j-1}(X, \mathcal{F})$ which inherits its Borel measurability from s. □

4 Proof of Theorem 1.6

Given a discrete group Γ, if $\mathrm{C_b}(\Gamma^j)$ denotes the space of bounded functions on the j-fold cartesian product Γ^j, the bounded cohomology of Γ can be defined as the cohomology of the complex

$$0 \longrightarrow \mathrm{C_b}(\Gamma)^\Gamma \longrightarrow \mathrm{C_b}(\Gamma^2)^\Gamma \longrightarrow \mathrm{C_b}(\Gamma^3)^\Gamma \longrightarrow \cdots$$

with the usual homogeneous coboundary operator. However, just like in the case of ordinary group cohomology, one can use instead the homological algebra approach which has the advantage of being more flexible in that one can use resolutions which are more appropriate to specific situations, as long as they satisfy certain properties. In other words, it can be proven that the cohomology of any admissible resolution by relatively injective Γ-modules is isomorphic to the bounded cohomology of Γ. As in the case of ordinary group cohomology, admissibility of a resolution involves the existence of homotopy operators, which in this case should be bounded in norm. Moreover, amenability of a Γ-action is intimately related to certain functions spaces being relatively injective Γ-modules, which makes this theory particularly fitting in this case and, more generally, whenever there is a suitable boundary. All of this is very vague and it is just to give some of the flavor of what follows: we refer the reader to [5], [12] and [4], where this theory was developed (in much greater generality) for the background and the precise definitions.

Let Y be a countable cellular space, $\pi_1(Y) = \Gamma$ its fundamental group, $\widetilde{Y}$ its universal covering, and (T, μ) a standard measure space with a measure class preserving Γ-action. If $S_j(\widetilde{Y})$ is the space of singular simplices in $\widetilde{Y}$, let $L^\infty_{w*}(T, \ell^\infty(S_j(\widetilde{Y})))$ denote the space of (equivalence classes of) maps $\alpha : T \to \ell^\infty(S_j(\widetilde{Y}))$ which are measurable when $\ell^\infty(S_j(\widetilde{Y}))$ is endowed with the weak* topology as the dual of $\ell^1(S_j(\widetilde{Y}))$, and which are essentially bounded. We then define the singular tangential bounded cohomology $\mathrm{H}^\bullet_{s,b}(X, \mathcal{F})$ of the foliated bundle $X = (\widetilde{Y} \times T)/\Gamma$ as the cohomology of the complex

$$0 \longrightarrow L^\infty_{w*}(T, \ell^\infty(S_0(\widetilde{Y})))^\Gamma \xrightarrow{d} L^\infty_{w*}(T, \ell^\infty(S_1(\widetilde{Y})))^\Gamma \xrightarrow{d} \cdots \tag{4.1}$$

with boundary operator

$$d\alpha(t)(s) := \alpha(t)(ds),$$

where $\alpha \in L^\infty_{w*}(T, \ell^\infty(S_j(\widetilde{Y})))$ and $s \in S_{j+1}(\widetilde{Y})$.

The first application of the homological algebra approach to bounded cohomology is the following:

Proposition 4.1. $\mathrm{H}^\bullet_{s,b}(X, \mathcal{F}) \simeq \mathrm{H}^\bullet_b(\Gamma, L^\infty(T))$.

Proof. First of all observe that we have the identification

$$L^\infty_{w*}(T, \ell^\infty(S_j(\widetilde{Y}))) \simeq L^\infty(T \times S_j(\widetilde{Y})) \simeq \ell^\infty(S_j(\widetilde{Y}), L^\infty(T)),$$

so that the complex in (4.1) can be rewritten as the complex

$$0 \longrightarrow \ell^\infty(S_0(\widetilde{Y}), L^\infty(T))^\Gamma \longrightarrow \ell^\infty(S_1(\widetilde{Y}), L^\infty(T))^\Gamma \longrightarrow \cdots.$$

Since this is the non-augmented subcomplex of invariant vectors of the complex

$$0 \longrightarrow L^\infty(T) \longrightarrow \ell^\infty(S_0(\widetilde{Y}), L^\infty(T)) \longrightarrow \ell^\infty(S_1(\widetilde{Y}), L^\infty(T)) \longrightarrow \cdots, \tag{4.2}$$

to prove the proposition it will be enough to show that (4.2) is an admissible resolution by relatively injective Γ-modules (see [5] or [12]).

We start by observing that the properness of the action of Γ on $\widetilde{Y}$ implies that, for all $j \geq 0$, the spaces $\ell^\infty(S_j(\widetilde{Y}), L^\infty(T))$ are relatively injective objects in the category of isometric Γ-Banach spaces, [12, Definition 4.1.2 and Theorem 4.5.2].

We need to define now appropriate homotopy operators. By using the usual coning procedure (since $\widetilde{Y}$ is contractible) there are homotopy operators h_j's

$$\ell^\infty(S_{j-1}(\widetilde{Y})) \underset{h_j}{\overset{d}{\rightleftarrows}} \ell^\infty(S_j(\widetilde{Y}))$$

which are norm continuous, and such that $\|h_j\| \leq 1$ [9, 10]. We can now define contracting homotopy operators [12, § 7.1]

$$\ell^\infty(S_{j-1}(\widetilde{Y}), L^\infty(T)) \xleftarrow{H_j} \ell^\infty(S_j(\widetilde{Y}), L^\infty(T))$$

as follows: let $\alpha : S_j(\widetilde{Y}) \to L^\infty(T)$ be a cochain, and for $f \in L^1(T)$ define $\alpha_f : S_j(\widetilde{Y}) \to \mathbb{R}$ by $\alpha_f(s_{(j)}) := \langle \alpha(s_{(j)}), f \rangle$ for $s_{(j)} \in S_j(\widetilde{Y})$. Then $f \mapsto h_j(\alpha_f)(s_{(j-1)})$ is a continuous linear form on $L^1(T)$, giving thus an element in $L^\infty(T)$ denoted $H_j(\alpha)(s_{(j-1)})$. This defines a norm continuous H_j, and hence the cohomology of the complex

$$0 \longrightarrow \ell^\infty(S_0(\widetilde{Y}), L^\infty(T))^\Gamma \longrightarrow \ell^\infty(S_1(\widetilde{Y}), L^\infty(T))^\Gamma \longrightarrow \cdots$$

is isomorphic to $\mathrm{H}_{\mathrm{b}}^\bullet(\Gamma, L^\infty(T))$ [12, Proposition 8.1.1]. □

This is the point where the amenability of the Γ-action on T plays an essential role.

Corollary 4.2. *If Γ acts amenably on T, then $\mathrm{H}_{\mathrm{s,b}}^\bullet(X, \mathcal{F}) = 0$.*

Proof. The amenability of the Γ-action implies that $L^\infty(T)$ is a relatively injective Γ-module, which in turns implies easily that $\mathrm{H}_{\mathrm{b}}^\bullet(\Gamma, L^\infty(T)) = 0$ [12, Proposition 7.4.1]. □

Now we need to relate the ordinary group cohomology of Γ to the singular cohomology of the foliated bundle. The idea is to use spaces very similar to those used in the case of singular bounded cohomology, but with no requirement on the boundedness in the direction of the leaves. To this purpose, if Y is a compact locally CAT(-1) space (that is a generalization, in the singular context, of a $\mathbb{R}$-rank one symmetric space), let $\sigma_j(\widetilde{Y})$ denote the set of j-simplices lifted to $\widetilde{Y}$ of any finite simplicial decomposition of Y. Observe that $\sigma_j(\widetilde{Y})$ is countable. Let $L_\infty(T, \mathrm{Maps}(\sigma_j(\widetilde{Y}), \mathbb{R}))$ be the space of all maps $\alpha : T \to \mathrm{Maps}(\sigma_j(\widetilde{Y}), \mathbb{R})$ such that for every $s_{(j)} \in \sigma_j(\widetilde{Y})$ the function $t \mapsto \alpha(t)(s_{(j)})$ is in $L^\infty(T)$, and define the singular tangential cohomology $\mathrm{H}_{\mathrm{s}}^\bullet(X, \mathcal{F})$ of $\mathcal{F}$ as the cohomology of the complex

$$0 \longrightarrow L_\infty(T, \mathrm{Maps}(\sigma_0(\widetilde{Y}), \mathbb{R}))^\Gamma \longrightarrow L_\infty(T, \mathrm{Maps}(\sigma_1(\widetilde{Y}), \mathbb{R}))^\Gamma \longrightarrow \cdots.$$

Observe that $L_\infty(T, \mathrm{Maps}(\sigma_j(\widetilde{Y}), \mathbb{R})) \simeq \mathrm{Maps}(\sigma_j(\widetilde{Y}), L^\infty(T))$; then a classical argument in ordinary group cohomology analogous to the one in the proof of Proposition 4.1 shows that the resolution

$$0 \longrightarrow \mathrm{Maps}(\sigma_0(\widetilde{Y}), L^\infty(T))^\Gamma \longrightarrow \mathrm{Maps}(\sigma_1(\widetilde{Y}), L^\infty(T))^\Gamma \longrightarrow \cdots$$

is an admissible resolution by relatively injective modules (where all the concepts have to be interpreted now in ordinary group cohomology) and hence its cohomology computes $\mathrm{H}^\bullet(\Gamma, L^\infty(T))$.

Now that all cohomology spaces have been defined, finally the punchline. Since Y is a compact locally CAT(-1) space, its fundamental group Γ is a Gromov hyperbolic group [7]. The essential step now is a result of Mineyev [11], which states that the map

$$\mathrm{H}^j_{\mathrm{b}}(\Gamma, V) \longrightarrow\!\!\!\!\!\rightarrow \mathrm{H}^j(\Gamma, V)$$

is surjective for all $j \geq 2$ and all isometric Banach Γ-modules V. In particular, the map

$$\mathrm{H}^j_{\mathrm{b}}(\Gamma, L^\infty(T)) \longrightarrow\!\!\!\!\!\rightarrow \mathrm{H}^j(\Gamma, L^\infty(T)) \tag{4.3}$$

is surjective for $j \geq 2$.

Collecting the isomorphisms $\mathrm{H}^\bullet_{\mathrm{s,b}}(X, \mathcal{F}) \simeq \mathrm{H}^\bullet_{\mathrm{b}}(\Gamma, L^\infty(T))$ and $\mathrm{H}^\bullet_{\mathrm{s}}(X, \mathcal{F}) \simeq \mathrm{H}^\bullet(\Gamma, L^\infty(T))$, and using (4.3), we have:

Corollary 4.3. *The map*

$$\mathrm{H}^j_{\mathrm{s,b}}(X, \mathcal{F}) \longrightarrow\!\!\!\!\!\rightarrow \mathrm{H}^j_{\mathrm{s}}(X, \mathcal{F})$$

is surjective for every $j \geq 2$.

Then Corollaries 4.2 and 4.3 immediately imply Theorem 1.6 if the Γ-action on T is amenable.

Acknowledgement. The proof in §4 is a part of an ongoing project with M. Burger. I want to thank: the Erwin Schrödinger International Institute in Vienna for their hospitality and support; V. Kaimanovich, K. Schmidt, and W. Woess for having given me the opportunity to participate to the workshop; and V. Kaimanovich for having undertaken the task of collecting this volume of Proceedings.

References

[1] S. Adams, Generalities on amenable actions, unpublished notes.

[2] C. Anantharaman-Delaroche and J. Renault, Amenable Groupoids. With a foreword by Georges Skandalis and Appendix B by E. Germain, Monographies de L'Enseignement Mathématique 36, L'Enseignement Mathématique, Geneva 2000.

[3] R. Brooks, Some remarks on bounded cohomology, in: Riemann Surfaces and Related Topics, Proceedings of the 1978 Stony Brook Conference (State Univ. New York, Stony Brook, N.Y., 1978), Ann. of Math. Stud. 97, Princeton University Press, Princeton, NJ, 1981, 53–63.

[4] M. Burger and A. Iozzi, Boundary maps in bounded cohomology, Appendix to Continuous bounded cohomology and applications to rigidity theory, by M. Burger and N. Monod, Geom. Funct. Anal. 12 (2002), 281–292.

[5] M. Burger and N. Monod, Continuous bounded cohomology and applications to rigidity theory, Geom. Funct. Anal. 12 (2002), 219–280.

[6] K. Corlette, L. Hernández Lamoneda and A. Iozzi, A vanishing theorem for the tangential de Rham cohomology of a foliation with amenable fundamental groupoid, Geom. Dedicata (to appear).

[7] É. Ghys and P. de la Harpe (eds.), Sur les Groupes Hyperboliques d'après Mikhael Gromov. Papers from the Swiss Seminar on Hyperbolic Groups held in Bern, 1988, Progr. Math. 83, Birkhäuser, Boston, MA, 1990.

[8] D. Gromoll and J. Wolff, Some relations between the metric structure and the algebraic structure of the fundamental group in manifolds of non-positive curvature, Bull. Amer. Math. Soc. 77 (1971), 545–552.

[9] M. Gromov, Volume and bounded cohomology, Inst. Hautes Études Sci. Publ. Math. 56 (1982), 5–99.

[10] N. V. Ivanov, Foundations of the theory of bounded cohomology, J. Soviet Math. 37 (1987), 1090–1115.

[11] I. Mineyev, Straightening and bounded cohomology of hyperbolic groups, Geom. Funct. Anal. 11 (2001), 807–839.

[12] N. Monod, Continuous Bounded Cohomology of Locally Compact Groups, Lecture Notes in Math. 1758, Springer-Verlag, Berlin 2001.

[13] C. C. Moore and C. Schochet, Global Analysis on Foliated Spaces. With appendices by S. Hurder, Moore, Schochet and Robert J. Zimmer, Math. Sci. Res. Inst. Publ. 9, Springer-Verlag, New York 1988.

[14] W. Thurston, Geometry and Topology of 3-Manifolds, Notes from Princeton University, Princeton, NJ, 1978.

[15] S. T. Yau, On the fundamental group of compact manifolds of non-positive curavture, Ann. of Math. (2) 93 (1971), 579–585.

[16] R. J. Zimmer, Curvature of leaves in amenable foliations, Amer. J. Math. 105 (1983), 1011–1022.

[17] R. J. Zimmer, Ergodic Theory and Semisimple Groups, Monogr. Math. 81, Birkhäuser, Basel 1984.

Alessandra Iozzi, FIM, ETH Zentrum, CH-8092, Zürich, Switzerland
E-mail: iozzi@math.ethz.ch

Linear rate of escape and convergence in direction

Anders Karlsson

Abstract. This paper describes some situations when random walks (or related processes) of linear rate of escape converge in direction in various senses. We discuss random walks on isometry groups of fairly general metric spaces, and more specifically, random walks on isometry groups of nonpositive curvature, isometry groups of reflexive Banach spaces, and linear groups preserving a proper cone. We give an alternative proof of the main tool from subadditive ergodic theory and we make a conjecture in this context involving Busemann functions.

1 Introduction

The well-known classical phenomenon of the nonexistence versus the existence of non-constant bounded harmonic functions in the plane and the unit disk, respectively, may be understood from observing that standard random walks in the Euclidean and the hyperbolic geometry behave quite differently. Brownian motion (or simple symmetric random walk on a lattice) in the Euclidean space does not converge in direction as time goes to infinity, while this is the case in the hyperbolic space, e.g., see [12] and [14]. Many contributions have extended this by showing that in many "hyperbolic" geometric situations convergence in direction (almost surely) occurs (e.g., [30, 39, 2, 40, 15, 29, 18, 20, 4, 8, 19, 1, 9, 26]). The present article points out some recent results illustrating that in several situation convergence in direction is a consequence of linear rate of escape of trajectories rather than of hyperbolicity (e.g., the main theorem in [42], as well as Theorems 4.1 and 5.2 below) extending the law of large numbers. We also explain two situations where convergence to points on some hyperbolic-type boundary takes place (Sections 6 and 7).

Our contributions are mostly relevant for spaces with large isometry groups, while many important works, some of which are listed above, deal with general, not necessarily homogeneous, situations. We apologize for omitted references.

2 Cocycles of semicontractions

Let S be a semigroup of semicontractions $D \to D$, where D is a nonempty subset of a metric space (Y, d), and fix a point $y \in D$.

Furthermore, let (X, μ) be a measure space with $\mu(X) = 1$, and let $L : X \to X$ be an ergodic and measure preserving transformation. Given a measurable map $w : X \to S$, put

$$u(n, x) = w(x)w(Lx) \dots w(L^{n-1}x), \tag{2.1}$$

and denote $u(n, x)y$ by $y_n(x)$. Note that by multiplying the transformation in this order makes the orbit $\{y_n(x)\}_{n=0}^{\infty}$ look like a trajectory of some kind of random walk. Assume that

$$\int_X d(y, w(x)y)d\mu(x) < \infty. \tag{2.2}$$

Let $a(n, x) = d(y, y_n(x))$. By the triangle inequality, the equality (2.1) and the semi-contraction property,

$$\begin{aligned} a(m + n, x) &\leq a(m, x) + d(u(m, x)y, u(m, x)u(n, L^m x)y) \\ &\leq a(m, x) + a(n, L^m x), \end{aligned}$$

hence a is a subadditive cocycle (see below). Furthermore, by the assumption (2.2),

$$\int_X a^+(1, x)d\mu(x) = \int_X d(y, w(x)y)d\mu(x) < \infty,$$

which means that the cocycle a satisfies the basic integrability condition. The subadditive ergodic theorem (see the next section) then implies that

$$\lim_{n\to\infty} \frac{1}{n} d(y, y_n(x)) = A \geq 0 \tag{2.3}$$

for almost every $x \in X$. This number A is called the *rate of escape*, and if $A > 0$ this is referred to as almost every trajectory $y_n(x)$ is of linear rate of escape.

3 Subadditive ergodic theory

Let (X, μ) be a measure space with $\mu(X) = 1$ and L a measure preserving transformation. A *subadditive cocycle* a is a measurable map $a : \mathbb{N}\times X \to \mathbb{R}$ such that

$$a(n + m, x) \leq a(n, x) + a(m, L^n x)$$

for $n, m \geq 1$ and μ-almost every x. Assume that a is integrable, that is,

$$\int_X a^+(1, x)d\mu(x) < \infty,$$

where $f^+(x) := \max\{f(x), 0\}$.

Kingman's subadditive ergodic theorem [32] asserts that for almost every x, the limit

$$\lim_{n\to\infty} \frac{1}{n} a(n, x)$$

exists. The following lemma will be the basic tool from ergodic theory that we use in most results discussed in this paper. It was proved and used by Margulis and the present author in [27].

Lemma 3.1 ([27]). *For each $\varepsilon > 0$, let E_ε be the set of x in X for which there exist an integer $K = K(x)$ and infinitely many n such that*

$$a(n, x) - a(n - k, L^k x) \geq (A - \varepsilon)k$$

for all k, $K \leq k \leq n$. Then $\mu\big(\bigcap_{\varepsilon>0} E_\varepsilon\big) = 1$.

Lemma 3.1 was proved in [27] using the so-called lemma about leaders. Here we describe an alternative proof and raise the question whether a stronger statement is true. Now follows an outline of the alternative proof of Lemma 3.1:

Define $v(n, x)$ by the formula

$$a(n, x) = v(n, x) + \sum_{k=0}^{n-1} a(1, L^k x).$$

It is immediate that $v(n, x)$ is a subadditive cocycle, and in addition $v(n, x) \leq 0$. The additive part of a (the above sum) is taken care of with Birkhoff's pointwise ergodic theorem, and the subadditive nonpositive part $v(n, x)$ is dealt with using the following lemma. Assume that

$$\gamma(v) := \lim_{n\to\infty} \frac{1}{n} \int_X v(n, x) d\mu(x) > -\infty.$$

Lemma 3.2. *Let $\lambda < 0$ and*

$$B = \{x \mid \exists K : \forall n > K, \min_{1\leq k\leq n} \frac{1}{k}(v(n, x) - v(n - k, L^k x) < \lambda\}.$$

Then

$$\mu(B) \leq \frac{\gamma(v)}{\lambda}.$$

This lemma can be proved in exactly the same way as [34, Lemma 5.10], where it was proved that for any integer K

$$\mu(B_K) \leq \frac{\gamma(v)}{\lambda},$$

where

$$B_K = \big\{x \mid \forall n > K, \min_{1\leq k\leq n} \frac{1}{k}(v(n, x) - v(n - k, L^k x) < \lambda\big\}.$$

Combining Lemma 3.2 and Birkhoff's ergodic theorem, we get that $\mu(E_\varepsilon) > 0$ for every ε. It is easy to see that $L^l E_\varepsilon \subset E_{2\varepsilon}$ for all $l \geq 0$, and assuming ergodicity it then follows that $\mu(E_{2\varepsilon}) = 1$. Since this holds for every $\varepsilon > 0$ and $E_\varepsilon \subset E_{\varepsilon'}$, whenever $\varepsilon < \varepsilon'$, Lemma 3.1 is proved.

In view of Sections 2, Section 5, and also [25], the following question arises. Fix $\varepsilon_i \to 0$ and consider the set F of x for which there are $n_i = n_i(x) \to \infty$ such that

$$a(n_i, x) - a(n_i - k, L^k x) \geq (A - \varepsilon_j)k$$

for all $j \leq i$ and $n_j \leq k \leq n_i$. This set is L-invariant, and for any additive cocycle a, $\mu(F) = 1$ by Birkhoff's theorem. Furthermore, for a subadditive sequence $a(n, x) = a_n$, it holds that $\mu(F) = 1$. For a general subadditive cocycle a, can it happen that $\mu(F) = 0$?

4 Nonpositive curvature

A *Hadamard space* is a complete metric space (Y, d) satisfying the following semi-parallelogram law: for any $x, y \in Y$ there exists a point z such that

$$d(x, y)^2 + 4d(z, w)^2 \leq 2d(x, w)^2 + 2d(y, w)^2$$

for any $w \in Y$. For basic facts about these spaces see [7]. A *geodesic ray* is a map $\gamma : [0, \infty) \to Y$ such that

$$d(\gamma(t), \gamma(s)) = |t - s|$$

for every s, t.

The following multiplicative ergodic theorem was proved by Margulis and the author using Lemma 3.1 and some geometric arguments:

Theorem 4.1 ([27]). *Assume that (Y, d) is a Hadamard space. Then for almost every x there exist $A \geq 0$ and a geodesic ray $\gamma(\cdot, x)$ starting at y such that*

$$\lim_{n\to\infty} \frac{1}{n} d(\gamma(An, x), u(n, x)y) = 0.$$

If $A > 0$, then the rays $\gamma(\cdot, x)$ are unique, and the orbit $u(n, x)y$ converges to this point on the boundary at infinity. As explained in [27] and [25], this theorem contains as special cases (the convergence statement of) the ergodic theorems of von Neumann, Birkhoff, and Oseledec. Note that the theorem is proved in [27] under the more general condition of a uniformly convex, nonpositively curved in the sense of Busemann, complete metric space Y.

The following remark is taken from [27]: assume that $S = \Gamma$ is a discrete cocompact group of isometries of a Cartan–Hadamard manifold Y. Consider a Markov process on Y/Γ with absolutely continuous transition probabilities, for example, the Brownian motion. Let X be the space of all bi-infinite trajectories on Y with the

measure μ coming from the process and a chosen stationary initial measure on a fundamental domain of Y/Γ. Let $w : X \to \Gamma$ be the map coming from the time 1 map and the chosen fundamental domain. For L we take the time 1 shift operator which is measure preserving. The theorem can then be applied to yield the result that for almost every sample path there is a geodesic ray such that the distance from the sample path to this geodesic grows sublinearly in n. In this context, we refer to Ballmann's paper [4] for comparison. In this paper Ballmann deals with the special case of independent, identically distributed increments of isometries of a space belonging to a certain rank 1 class of locally compact Hadamard spaces). He therefore needs a more sophisticated approximation scheme (following the method of Furstenberg and Lyons–Sullivan [35]) to transfer the Markov process to a random walk on a group of isometries. Then a result of Guivarc'h [17] can be used to guarantee that $A > 0$ whenever the group in question is nonamenable (which most of the time is the case here).

We now establish the link between Theorem 4.1 and the conjecture in Section 8. The Busemann function b_γ corresponding to γ is (see also Section 8):

$$b_\gamma(z) = \lim_{n\to\infty} d(\gamma(t), z) - d(\gamma(t), y).$$

(The triangle inequality implies that the limit exists.)

Proposition 4.2. *For Y a Hadamard space the conclusion in Theorem* 4.1 *is equivalent to the conclusion in Conjecture* 8.1 *below.*

Proof. For Hadamard spaces it is known that every horofunction is a Busemann function corresponding to a geodesic ray as above. Let y_n be an arbitrary sequence of points such that $d(y, y_n)/n \to A > 0$. Assume that $-b_\gamma(y_n) \sim An$, and denote by $\bar{y}_n$ the point on γ closest to y_n. By the cosine law, a property of projections, and the fact that horoballs are geodesically convex:

$$d(y, y_n)^2 \geq d(y, \bar{y}_n)^2 + d(\bar{y}_n, y_n)^2 \geq b_\gamma(y_n)^2 + d(\bar{y}_n, y_n)^2.$$

This implies that $d(\bar{y}_n, y_n) = o(n)$, and by the triangle inequality that $d(\gamma(An), y_n) = o(n)$ as desired. The converse holds for any metric space: assume $d(\gamma(An), y_n) = o(n)$. It is a general fact that

$$b_\gamma(y_n) \leq d(\gamma(An), y_n) - d(\gamma(An), y),$$

which in our case implies that $-b_\gamma(y_n) \sim An$. □

5 Continuous linear functionals

In this section we assume that Y is a normed real vector space and S is a semigroup of semicontractions $D \to D$, where the subset D for convenience is assumed to contain $y = 0$.

Proposition 5.1. *For almost every x and for any $\varepsilon > 0$ there exists an element f_x^ε in the topological dual of Y with norm 1 such that*

$$\liminf_{n\to\infty} \frac{1}{n} f_x^\varepsilon(y(n,x)) \geq A - \varepsilon.$$

Proof. If $A = 0$, then any f would do. If $A > 0$, then consider $x \in E_\varepsilon$ for some $\varepsilon > 0$ (Lemma 3.1). It follows from the Hahn-Banach theorem (see [11, p. 65]) that we can find elements f_n of norm 1 in the dual space such that $f_n(y(n,x)) = a(n,x)$. Take a sequence of n_i and a $k \geq K$ such that the inequality in the lemma holds. By picking subsequences and applying the diagonal process we may assume that $f_{n_i}(y(k,x)$ converges for every $k \geq K$. This defines a linear functional of norm at most 1 on the linear span of the orbit $y(k,x)$, $k \geq K$, which we may extend to a linear functional with the same norm on the whole space again by the Hahn-Banach theorem. We have

$$\begin{aligned} f_{n_i}(y(k,x)) &= a(n_i,x) - f_{n_i}(y(n_i,x) - y(k,x)) \\ &\geq a(n_i,x) - ||y(n_i,x) - y(k,x)|| \\ &\geq a(n_i,x) - a(n_i - k, L^k x) \\ &\geq (A-\varepsilon)k. \end{aligned}$$

Therefore,

$$\frac{1}{k} f(y(k,x)) \geq A - \varepsilon$$

for all $k \geq K$. □

Whenever $x \in F$ (see Section 3), we can remove ε and replace lim inf by lim in the above proposition. Since it is not clear to the author when this is the case, we can only prove the following by adding assumptions on Y.

Theorem 5.2. *Assume that Y is a reflexive Banach space. For almost every x there exists an element f_x in the dual of Y with norm 1 such that*

$$\lim_{n\to\infty} \frac{1}{n} f_x(y(n,x)) = A.$$

Proof. We may assume that Y is separable as we can, if necessary, replace it with the closed linear span of the orbit. Therefore, and due to reflexivity, the closed unit balls in Y and in Y^* are sequentially compact in the respective weak topology, see [11, p. 68]. Suppress x and pick $\varepsilon_i \to 0$ such that f_{ε_i} converges to some f in the weak*-topology. Given any infinite subsequence n_j, pick a weak limit point $\bar{y}$ of $y(n_j,x)/n$, so that

$$f_{\varepsilon_i}(y(n_j,x)/n_j) \to f_{\varepsilon_i}(\bar{y})$$

along the subsequence of n_j for which the points converge to $\bar{y}$. Therefore, $f_{\varepsilon_i}(\bar{y}) \geq A - \varepsilon_i$, but since

$$f_{\varepsilon_i}(y) \to f(y)$$

for any y, we must have that $f(\bar{y}) \geq A$. Finally note that as f has norm 1, it trivially holds that

$$\lim \sup_{n\to\infty} \frac{1}{n} f(y(n,x)) \leq A,$$

and the theorem is proved. (Instead of arguing with limit points $\bar{y}$ we could have applied S. Mazur's theorem on closures of convex sets.) □

Corollary 5.3 (Cf. [27]). *Assume that Y is a reflexive Banach space whose dual has Fréchet differentiable norm. Then for almost every x*

$$\frac{1}{n} y(n,x)$$

converges in norm.

Proof. It is known (due to Šmulian) and not difficult to show that the dual has Fréchet differentiable norm if and only if every sequence y_n in Y satisfying $||y_n|| = 1$ and $f(y_n) \to 1$ for some $f \in Y^*$ with norm 1, must converge, see [10] for a proof. □

Uniform convexity implies that the dual has Fréchet differentiable norm. The above corollary improves on Theorem 4.1 for Banach spaces. The author believes that the assumption that Y is a reflexive Banach space in the above results may be relaxed, which would have implications for random products of continuous linear operators, see the last section of [25]. One idea of relaxing the conditions on the Banach spaces could be to use the known fact that any separable space can be renormed to have a locally uniformly convex norm. Note however that, except possibly for the reflexivity, the above assumption (the differentiability of the norm in the dual) is best possible in Corollary 5.3 in view of a counterexample constructed in [33]. There are several other papers studying the iteration of a single non-expansive map (e.g., [37, 38]).

The random mean ergodic theorem of Beck–Schwartz [6] can be deduced from Corollary 5.3 (although with a less general Y), compare with [25].

6 Conformal or Floyd-type boundaries

The construction here of a hyperbolic type boundary is a restrictive version of the one given by Gromov [16, Section 7.2.K "A conformal view on the boundary"] , which extends Floyd [13], which in turn is "based on an idea of Thurston's and inspired by a construction of Sullivan's".

Assume Y is a complete, geodesic metric space. The *length* of a continuous curve $\alpha : [a,b] \to Y$ is defined to be

$$L(\alpha) = \sup \sum_{i=1}^{k} d(\alpha(t_{i-1}), \alpha(t_i)),$$

where the supremum is taken over all finite partitions $a = t_0 < t_1 < ... < t_k = b$. When this supremum is finite, α is said to be *rectifiable*. For such α we can define the *arc length* $s : [a, b] \to [0, \infty)$ by

$$s(t) = L(\alpha|_{[a,t]}),$$

which is a function of bounded variation.

Given a continuous, (strictly) positive function f on Y, we define the *f-length* of a rectifiable curve α to be

$$L_f(\alpha) = \int_\alpha f\, ds = \int_a^b f(\alpha(t))ds(t).$$

If $f \equiv 1$, then $L_f = L$.

A new distance d_f is defined by

$$d_f(x, y) = \inf L_f(\alpha),$$

where the infimum is taken over all rectifiable curves α with $\alpha(a) = x$ and $\alpha(b) = y$.

For simplicity we choose $f(z) = d(y, z)^{-2}$, where y is a fixed base point. Let the *f-boundary* of Y be the space $\partial_f Y := \overline{Y_f} - Y$, where $\overline{Y_f}$ denotes the metric space completion of (Y, d_f). In [26] we prove using Lemma 3.1:

Theorem 6.1. *Assume that $A > 0$. Then for almost every x the trajectory $u(n, x)y$ converges to a point $\xi = \xi(x) \in \partial_f Y$.*

Proof. Here is a sketch of a proof somewhat different from the one in [26]. Note that for appropriate k and n in the sense of the Lemma 3.1 we have:

$$\begin{aligned}(y_n(x)|y_k(x))_y &:= \frac{1}{2}(d(y_n(x), y) + d(y_k(x), y) - d(y_n(x), y_k(x)) \\ &\geq \frac{1}{2}(a(n, x) + a(k, x) - a(n - k, L^k x)) \\ &\geq (A - \varepsilon)k.\end{aligned}$$

In view of the lemma in Section 5 of [26] it follows from this estimate that for a fixed positive $\varepsilon < A$ and $n_i \to \infty$ for which the inequality in Lemma 3.1 is satisfied, the sequence $\{y_{n_i}(x)\}$ is d_f-Cauchy and hence converges to a point in $\partial_f Y$. Moreover, it then follows that the whole sequence $y_k(x)$ converges to this boundary point as well. □

An interesting special case is a random walk on Y being the Cayley graph of a finitely generated group Γ. In [26] also some visibility properties are shown, in particular, we demonstrate that Kaimanovich's conditions (CP), (CS), and (CG) in [21] hold. The arguments in [21] therefore provide an alternative approach (not using Lemma 3.1) to the convergence in direction and which moreover show that if $\partial_f \Gamma$ is non-trivial then it is indeed maximal.

For more on random walks on groups and graphs we refer to the book by Woess [43] and the references therein.

7 Hilbert's projective metric

Assume that (Y, d) is a bounded convex domain in $\mathbb{R}^N$ equipped with Hilbert's metric and let ∂Y be the natural boundary of the domain. Similar to the proof of Theorem 6.1 above, cf. also [24], and in view of the weak hyperbolicity of Hilbert's metric established by Noskov and the author in [28] (extending a result of Beardon) we have:

Theorem 7.1. *Assume that $A > 0$. Then for almost every x, there is a point $\gamma_x \in \partial Y$ such that any other limit point of $y_n(x)$ may be connected by a line segment contained in ∂Y to γ_x. In particular, if Y is strictly convex, then $y_n(x) \to \gamma_x$ for $n \to \infty$.*

In the case of a strictly convex domain and $u(n, x)$ is a random walk (the increments are i.i.d.) taking values in the isometry group, one can probably use Furstenberg's ideas of combining proximality properties with the martingale convergence theorem (without assuming $A > 0$) to show the convergence in direction. In this situation we also have Oseledec's theorem [36] at our disposal since the isometry group is the subgroup of the projective linear group preserving the convex set.

8 Busemann functions

Let (Y, d) be a metric space, and let $C(Y)$ denote the space of continuous functions on Y equipped with the topology of uniform convergence on bounded subsets. Fixing a point y, the space Y is continuously injected into $C(Y)$ by

$$\Phi : z \mapsto d(z, \cdot) - d(z, y).$$

A metric space is called *proper* if every closed ball is compact. If Y is a proper metric space, then the Arzela-Ascoli theorem asserts that the closure of the image $\overline{\Phi(Y)}$ is compact. The points on the boundary $\partial Y := \overline{\Phi(Y)} \setminus \Phi(Y)$ are called *Busemann* (or *horo) functions*, see [5] for more on this topic.

In the setup of Section 2 we formulate the following conjecture:

Conjecture 8.1. *Assume that (Y, d) is a proper metric space. For almost every x there exists a horofunction b_x such that*

$$\lim_{n\to\infty} -\frac{1}{n} b_x(u(n, x)y) = A.$$

Evidence for the truth of this statement: it holds for one transformation $u(n, x) = \phi^n$, see [24]. It holds for complete metric spaces (not necessarily locally compact!)

of nonpositive curvature, see Section 4. It would hold in general if $\mu(F) = 1$, see Section 3. Theorem 5.2 also provides some evidence. Moreover, the above type of limits with respect to Busemann functions should exist fairly generally for the following reason:

Assume that w takes its values in the isometry group of Y. Let $\widehat{X} = X \times \partial Y$ be the product measurable space, and define

$$\widehat{L} : (x, \gamma) \mapsto (Lx, w(x)^{-1}\gamma).$$

By a standard argument (using Tychonoff's fixed point theorem) due to the compactness of ∂Y, there exists an ergodic $\widehat{L}$-invariant measure $\widehat{\mu}$ such that $\widehat{\mu}(\widehat{X}) = 1$ and the projection of $\widehat{\mu}$ onto X coincides with μ.

Let $z_i \to \gamma \in \partial Y$, and denote by

$$b_\gamma^y(\cdot) = \lim_{i\to\infty} d(\cdot, z_i) - d(z_i, y)$$

the Busemann function centered at γ (and based at y).

Proposition 8.2. *For $\widehat{\mu}$-almost every (x, γ)*

$$\lim_{n\to\infty} \frac{1}{n} b_\gamma^y(u(n, x)y) = \int_{\widehat{X}} b_\xi^y(w(x')y) d\widehat{\mu}(x', \xi).$$

Proof. For any w the following is a trivial identity:

$$b_\gamma^y(\cdot) = b_\gamma^y(w) + b_\gamma^w(\cdot). \tag{8.1}$$

Let g and h be two isometries. It follows that

$$b_\gamma^{g(y)}(gh(y)) = b_{g^{-1}\gamma}^y(h(y)).$$

In view of this equality and (8.1) we have

$$\begin{aligned} b_\gamma^y(u(n+m, x)y) &= b_\gamma^y(u(m, x)y) + b_\gamma^{u(m,x)y}(u(n+m, x)y) \\ &= b_\gamma^y(u(m, x)y) + b_{u(m,x)^{-1}\gamma}^y(u(n, L^m x)y). \end{aligned}$$

Thus we have an additive cocycle on the skew product system

$$v(n, (x, \gamma)) := b_\gamma^y(u(n, x)y),$$

and it is integrable because $|b_\gamma(w(x)y)| \le d(y, w(x)y)$. The assertion is now just Birkhoff's ergodic theorem. □

9 Random randomness

Recently the subject of random walks in random environment and random walks with random transition probabilities has attracted much attention. (See the books by Kifer [31], L. Arnold [3] and Sznitman [41]). This subject was advertized in some form already by Pitt, von Neumann–Ulam, and Kakutani, see [23]. In particular, they noted that a random individual ergodic theorem follows by a simple trick from the individual ergodic theorem of Birkhoff itself. Another result from the 1950s is the random mean ergodic theorem due to Beck–Schwartz, which in fact can be deduced from Theorem 4.1 or Corollary 5.3 above (note however that their assumption on the Banach space is somewhat weaker), see [25].

The recent paper [22] studies various notions of measure theoretical boundaries and Poisson formulas associated with random walks with random transition probabilities. In the last section of their paper they give some examples of the identification of the Poisson boundary using Theorem 4.1.

We would also like to mention the law of large numbers for certain random walks in random environment obtained by Sznitman–Zerner in [42] as it exemplifies the title of the present paper. The proof of their theorem is based on a nice argument establishing, under some transience conditions, a renewal structure: there are times τ_i occuring often enough (integrability), at which the walk reaches a new peak in the transience direction and never again returns to the halfplane it just left.

Acknowledgement. I would like to thank Professors V. Kaimanovich, H. Abels, and M. Burger for inviting me to the Erwin Schrödinger Institute, the Universität Bielefeld, and the ETH-Zürich, respectively.

References

[1] A. Ancona, Convexity at infinity and Brownian motion on manifolds with unbounded negative curvature, Rev. Mat. Iberoamericana 10 (1994), 189–220.

[2] M. T. Anderson, The Dirichlet problem at infinity for manifolds of negative curvature, J. Differential Geom. 18 (1983), 701–721

[3] L. Arnold, Random Dynamical Systems, Springer Monogr. Math. Springer-Verlag, Berlin 1998.

[4] W. Ballmann, On the Dirichlet problem at infinity for manifolds of nonpositive curvature, Forum Math. 1 (1989), 201–213.

[5] W. Ballmann, M. Gromov and V. Schroeder, Manifolds of Nonpositive Curvature, Progr. Math. 61, Birkhäuser, Boston, MA, 1985.

[6] A. Beck and J. T. Schwartz, A vector-valued random mean ergodic theorem, Proc. Amer. Math. Soc. 8 (1957), 1049–1059

[7] M. Bridson and A. Haefliger, Metric Spaces of Non-positive Curvature, Grundlehren Math. Wiss. 319, Springer-Verlag, Berlin 1999.

[8] D. I. Cartwright and P. M. Soardi, Convergence to ends for random walks on the automorphism group of a tree, Proc. Amer. Math. Soc. 107 (1989), 817–823.

[9] M. Cranston, W. S. Kendall and Y. Kifer, Gromov's hyperbolicity and Picard's little theorem for harmonic maps, in: Stochastic Analysis and Applications (Powys, 1995), World Sci. Publishing, River Edge, NJ, 1996, 139–164.

[10] J. Diestel, Geometry of Banach Spaces — Selected Topics, Lecture Notes in Math. 485, Springer-Verlag, Berlin 1975.

[11] N. Dunford and J. T. Schwartz, Linear Operators. I. General Theory. With the assistance of W. G. Bade and R. G. Bartle, Pure . Appl. Math. 7, Interscience, New York 1958.

[12] E. B. Dynkin, Markov processes and problems in analysis, in: Proc. Internat. Congr. Mathematicians (Stockholm, 1962), Inst. Mittag-Leffler, Djursholm 1963, 36–58.

[13] W. J. Floyd, Group completions and limit sets of Kleinian groups, Invent. Math. 57 (1980), 205–218.

[14] H. Furstenberg, A Poisson formula for semi-simple Lie groups, Ann. of Math. (2) 77 (1963), 335–386.

[15] S. I. Goldberg, C. Mueller, Brownian motion, geometry, and generalizations of Picard's little theorem, Ann. Prob. 11 (1983), 833–846.

[16] M. Gromov, Hyperbolic groups, in: Essays in Group Theory, Math. Sci. Res. Inst. Publ. 8, Springer-Verlag, New York 1987, 75–263.

[17] Y. Guivarc'h, Sur la loi des grands nombres et le rayon spectral d'une marche aléatoire, in: Conference on Random Walks (Kleebach, 1979), Astérisque 74, Soc. Math. France, Paris 1980, 47–98.

[18] P. Hsu, P. March, The limiting angle of certain Riemannian Brownian motions, Comm. Pure Appl. Math. 38 (1985), 755–768.

[19] P. Hsu, W. S. Kendall, Limiting angle of Brownian motion in certain two-dimensional Cartan–Hadamard manifolds, Ann. Fac. Sci. Toulouse Math. (6) 1 (1992), 169–186.

[20] V. A. Kaimanovich, Lyapunov exponents, symmetric spaces and multiplicative ergodic theorem for semisimple Lie groups, J. Soviet Math. 47 (1989), 2387–2398.

[21] V. A. Kaimanovich, The Poisson formula for groups with hyperbolic properties, Ann. of Math. (2) 152 (2000), 659–692.

[22] V. A. Kaimanovich, Y. Kifer, B.-Z. Rubshtein, Boundaries and harmonic functions for random walks with random transition probabilities, ESI-preprint (2001).

[23] S. Kakutani, Ergodic theory, in: Proceedings of the International Congress of Mathematicians, Cambridge, Mass., 1950, vol. 2, Amer. Math. Soc., Providence, RI, 1952, 128–142.

[24] A. Karlsson, Non-expanding maps and Busemann functions, Ergodic Theory Dynam. Systems 21 (2001), 1447–1457.

[25] A. Karlsson, Nonexpanding maps, Busemann functions, and multiplicative ergodic theory, in: Rigidity in Dynamics and Geometry (Cambridge, 2000), Springer-Verlag, Berlin 2002, 283–294.

[26] A. Karlsson, Boundaries and random walks on finitely generated infinite group, Ark. Mat. 41 (2003), 295–306.

[27] A. Karlsson and G. A. Margulis, A multiplicative ergodic theorem and nonpositively curved spaces, Comm. Math. Phys. 208 (1999), 107–123.

[28] A. Karlsson and G. A. Noskov, The Hilbert metric and Gromov hyperbolicity, Enseign. Math. (2) 48 (2002), 73–89.

[29] W. S. Kendall, Brownian motion and a generalised little Picard's theorem, Trans. Amer. Math. Soc. 275 (1983), 751–760.

[30] Y. Kifer, Limit theorems for a conditional Brownian motion in Euclidean and Lobachevskian spaces, (Russian) Uspehi Mat. Nauk 26 (3) (1971), 203–204.

[31] Y. Kifer, Ergodic Theory of Random Transformations, Progr. Probab. Statist. 10, Birkhäuser, Boston, MA, 1986.

[32] J. F. C. Kingman, The ergodic theory of subadditive ergodic processes, J. Roy. Statist. Soc. Ser. B 30 (1968), 499–510.

[33] E. Kohlberg, A. Neyman, Asymptotic behaviour of nonexpansive mappings in normed linear spaces, Israel J. Math. 38 (1981), 269–274.

[34] U. Krengel, Ergodic Theorems. With a supplement by Antoine Brunel, de Gruyter Stud. Math. 6, Walter de Gruyter, Berlin 1985.

[35] T. Lyons, D. Sullivan, Function theory, random paths and covering spaces, J. Differential Geom. 19 (1984), 299-323

[36] V. I. Oseledets, A multiplicative ergodic theorem: Lyapunov characteristic exponents for dynamical systems, Trans. Moscow Math. Soc. 19 (1968), 197–231.

[37] A. Pazy, Asymptotic behaviuor of contractions in Hilbert space, Israel J. Math. 9 (1971), 235–240.

[38] A. Plant, S. Reich, The asymptotics of nonexpansive iterations, J. Funct. Anal. 54 (1983), 308–319.

[39] J.-J. Prat, Étude asymptotique et convergence angulaire du mouvement brownien sur une variété à courbure négative, C. R. Acad. Sci. Paris Sér. A-B 280 (1975), A1539–A1542.

[40] D. Sullivan, The Dirichlet problem at infinity for a negatively curved manifold, J. Differential Geom. 18 (1983), 723–732.

[41] A.-S. Sznitman, Brownian Motion, Obstacles and Random Media, Springer Monogr. Math., Springer-Verlag, Berlin 1998.

[42] A.-S. Sznitman, M. Zerner, A law of large numbers for random walks in random environment. Ann. Probab. 27 (1999), 1851–1869.

[43] W. Woess, Random Walks on Infinite Graphs and Groups, Cambridge Tracts in Math. 138, Cambridge University Press, Cambridge 2000.

Anders Karlsson, Department of Mathematics, Royal Institute of Technology, 100 44 Stockholm, Sweden

E-mail: akarl@math.kth.se

Remarks on harmonic functions on affine buildings

Anna Maria Mantero and Anna Zappa

Abstract. Let Δ be a thick affine building of type $\tilde{A}_2$. We prove that each positive harmonic function on Δ is the Poisson transform of a positive measure on the maximal boundary Ω. Moreover we prove that if f is weakly harmonic and bounded, then f is harmonic.

1 Introduction

This paper deals with harmonic functions on affine buildings of type $\widetilde{A}_2$. We recall that a function f on an open subset of a symmetric space $X = G/K$ is said to be harmonic if $D(f) = 0$ for any G-invariant differential operator D on X such that $D(1) = 0$ (see [6]). In the p-adic case, for a p-adic reductive group G and a compact open subgroup K, the space which plays the role of the symmetric space is an affine building Δ, and on this space the correct analogue of the G-invariant differential operators are the operators of convolution with compactly supported bi-K-invariant functions defined on the special vertices of Δ. These operators form the Hecke algebra $\mathcal{H}$ which, in the case we are considering, is generated by two operators $\mathcal{L}_1$, $\mathcal{L}_2$, called Laplacians. This suggests to call "harmonic" any function f such that $\mathcal{L}_1 f = \mathcal{L}_2 f = f$. One can give a definition of $\mathcal{H}$ which uses only the geometry of the building, and which therefore applies also to buildings not arising from linear groups (see [2] and [7]). In [8] we proved, for a building of type $\tilde{A}_2$, that every joint eigenfunction of the operators $\mathcal{L}_1$, $\mathcal{L}_2$ is the Poisson transform of a suitable finitely additive measure on the maximal boundary Ω. Using this characterization, in this paper we prove that if f is harmonic and positive, then the corresponding measure on Ω is also positive.

Besides the above definition, on a symmetric space there is a weaker definition of harmonicity (see [6]). Namely, a function f is said to be weakly harmonic if $D_0(f) = 0$, where D_0 is the Laplace–Beltrami operator on X, and more generally it is said to be θ-harmonic if $D_\theta(f) = 0$, where D_θ is the operator on X corresponding to a probability measure θ on the group G. Hence, in the buildings setup, we call a function f "weakly harmonic" if $\mathcal{L}_\theta f = f$, where $\mathcal{L}_\theta$ is the averaging operator associated with the random walk determined by transition probabilities θ. In this paper we compare these two definitions of harmonicity for buildings and show that they are not equivalent in general, but, in the same way as for symmetric spaces [3], weakly

harmonic functions which are bounded are harmonic. This result extends Theorem 15.5 of [5] (stated for linear buildings) to all buildings of type $\widetilde{A}_2$.

The results of the paper can be generalized to all rank 2 buildings using the characterization of the eigenfunctions of the Laplacians given in [10] and [11] for buildings of types $\widetilde{G}_2$ and $\widetilde{B}_2$, respectively.

We suggest [4] for results on invariant operators and [1] for a general discussion about harmonic functions on $\widetilde{A}_n$ buildings.

2 Preliminaries

We recall here only the fundamental definitions and the principal results we need in this paper. The reader is referred to [12] for background information on buildings of type $\widetilde{A}_2$ and to [8] for the proofs of all results stated in this section.

We denote by Δ a building of type $\widetilde{A}_2$ and by $\mathbb{A}$ its abstract apartment (called the fundamental apartment of Δ); moreover we denote by $\mathcal{V}$ and $\overline{\mathcal{V}}$ the sets of all vertices of Δ and $\mathbb{A}$, respectively. By $\tau(x)$ we denote the type of a vertex x.

We fix a vertex e (resp., O), say $\tau(e) = \tau(O) = 0$, in Δ (resp., in $\mathbb{A}$) and a sector Q_0 in $\mathbb{A}$. We define $\mathcal{B}_k = \{x \in \mathcal{V} : \ d(x, e) \leq k\}$, for every $k \geq 0$, where $d(x, y)$ denotes the usual graph-theoretic distance between vertices x and y.

Given a vertex x, denote by (m, n) the lengths of the sides of the convex hull $\mathcal{C}h[e, x]$ of the set $\{e, x\}$. We set $\mathcal{C}h[e, x] \sim \mathcal{C}h[e, x']$ if x and x' have the same coordinates in the sectors based at e and containing x and x', respectively. We denote by $X_{m,n}$ the vertex of $\mathbb{A}$ with coordinates (m, n) with respect to O.

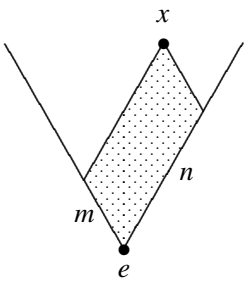


Figure 1.

We denote by Ω the maximal boundary of Δ; as it is shown in [7], Ω may be endowed with a totally disconnected compact Hausdorff topology. This topology is generated by the family

$$\mathcal{B} = \{\ \Omega(c),\ c \in \mathcal{C}\},$$

where $\mathcal{C}$ is the set of all chambers of Δ and, for every c,

$$\Omega(c) = \{\omega \in \Omega : c \subset Q_e(\omega)\}.$$

The Laplace operators of the building are the following averaging operators defined on the complex valued functions on $\mathcal{V}$,

$$\mathcal{L}_i f(x) = \frac{1}{|\mathcal{S}_i(x)|} \sum_{y \in \mathcal{S}_i(x)} f(y), \quad x \in \mathcal{V},\ i = 1, 2,$$

where

$$\mathcal{S}_i(x) = \{y \in \mathcal{V} : d(x, y) = 1, \tau(y) = \tau(x) + i\}, \quad i = 1, 2.$$

For every pair $(\gamma_1, \gamma_2) \in \mathbb{C}^2$, we denote by $\mathcal{S}(\gamma_1, \gamma_2)$ the joint eigenspace of the Laplace operators associated with the eigenvalues γ_1, γ_2 :

$$\mathcal{S}(\gamma_1, \gamma_2) = \{f : \mathcal{V} \to \mathbb{C};\ \mathcal{L}_1(f) = \gamma_1 f,\ \mathcal{L}_2(f) = \gamma_2 f\}.$$

We recall the definitions of the Poisson kernel and the Poisson transform.

Definition 2.1. For every triple $\boldsymbol{\alpha} = (a_1, a_2, a_3)$ of complex numbers, such that $a_1 a_2 a_3 = 1$, let $\phi_{\boldsymbol{\alpha}}$ be the multiplicative function on $\overline{\mathcal{V}}$ defined, with respect to the coordinate system associated to Q_0, as :

$$\phi_{\boldsymbol{\alpha}}(X_{m,n}) = a_1^m a_3^{-n}, \quad \forall (m, n) \in \mathbb{Z}^2.$$

We call the Poisson kernel (of initial point $\overline{x}$ and of parameter $\boldsymbol{\alpha}$) the function

$$P_{\boldsymbol{\alpha}}^{\overline{x}}(x, \omega) = \phi_{\boldsymbol{\alpha}}(r_\omega^{\overline{x}}(x)), \quad \forall x \in \mathcal{V},\ \forall \omega \in \Omega,$$

where $r_\omega^{\overline{x}}$ is the retraction of Δ onto $\mathbb{A}$ with respect to ω and the initial point $\overline{x}$.

For every $x \in \mathcal{V}$, the function $P_{\boldsymbol{\alpha}}^{\overline{x}}(x, \cdot)$ belongs to the linear space $H(\Omega)$ of all locally constant functions on Ω. The dual space $H'(\Omega)$ consists of all finitely additive measures defined on the algebra generated by the family $\mathcal{B}$.

Definition 2.2. Let $\nu \in H'(\Omega)$; the Poisson transform of ν (of initial point $\overline{x}$ and parameter $\boldsymbol{\alpha}$) is the function

$$\mathcal{P}_{\boldsymbol{\alpha}}^{\overline{x}} \nu(x) = \int_\Omega P_{\boldsymbol{\alpha}}^{\overline{x}}(x, \omega) d\nu(\omega), \quad \forall x \in \mathcal{V}.$$

In particular, for every $F \in H(\Omega)$,

$$\mathcal{P}_{\boldsymbol{\alpha}}^{\overline{x}}(F)(x) = \int_\Omega P_{\boldsymbol{\alpha}}^{\overline{x}}(x, \omega) F(\omega) d\mu(\omega),$$

where μ is the probability measure on Ω defined in [7, p. 427].

It is easy to check that $\mathcal{P}^{\overline{x}}_{\boldsymbol{\alpha}} v$ is a joint eigenfunction of the operators $\mathcal{L}_1$, $\mathcal{L}_2$ with the eigenvalues

$$\begin{aligned} \gamma_1 = \gamma_1(\boldsymbol{\alpha}) &= \frac{1}{q^2+q+1}(a_1 + qa_2 + q^2 a_3), \\ \gamma_2 = \gamma_2(\boldsymbol{\alpha}) &= \frac{1}{q^2+q+1}(a_3^{-1} + qa_2^{-1} + q^2 a_1^{-1}). \end{aligned} \tag{2.1}$$

The main result of [8] is the following theorem.

Theorem 2.3. *For every* $(\gamma_1, \gamma_2) \in \mathbb{C}^2$ *there exists a triple* $\boldsymbol{\alpha} = (a_1, a_2, a_3)$, *with* $\frac{|a_1|}{q} \geq |a_2| \geq q|a_3|$, *such that* $\gamma_1 = \gamma_1(\boldsymbol{\alpha})$, $\gamma_2 = \gamma_2(\boldsymbol{\alpha})$, *and for every* $f \in \mathcal{S}(\gamma_1, \gamma_2)$ *there exists a unique* $v \in H'(\Omega)$ *such that* $f = \mathcal{P}^{\overline{x}}_{\boldsymbol{\alpha}} v$.

3 Characterization of positive harmonic functions

Definition 3.1. A function f defined on $\mathcal{V}$ is said to be harmonic if

$$\mathcal{L}_1(f) = \mathcal{L}_2(f) = f.$$

Hence, $\mathcal{S}(1,1)$ is the space of all harmonic functions on $\mathcal{V}$, and a function f is harmonic if and only if $f = \mathcal{P}^{\overline{x}}_{\boldsymbol{\alpha}_0} v$, for some suitable $v \in H'(\Omega)$ and $\boldsymbol{\alpha}_0 = (q^2, 1, q^{-2})$. For ease of notation, we simply write $\mathcal{P}^{\overline{x}} v = \mathcal{P}^{\overline{x}}_{\boldsymbol{\alpha}_0} v$.

Let $c \in \mathcal{C}$; for $i = 0, 1, 2$ we denote by x_i its type i vertex. We consider the decomposition of the boundary Ω into a disjoint union of six subsets $\Omega_{c,j}$ canonically associated with c.

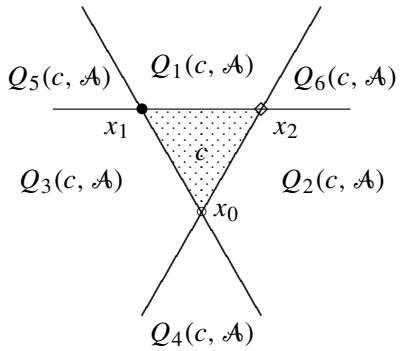


Figure 2.

Definition 3.2. For every apartment $\mathcal{A}$ containing c, we define (see Figure 2):

$$Q_j(c,\mathcal{A}) = \begin{cases} \text{the sector of base vertex } x_{j-1} \text{ containing } c, & \text{if } j = 1,2,3\,; \\ \text{the sector of base vertex } x_{j-4} \text{ opposite to } c, & \text{if } j = 4,5,6. \end{cases}$$

Then we define, for $j = 1, \ldots, 6$,

$$\Omega_{c,j} = \{\omega \in \Omega :\ Q_e(\omega) \sim Q_j(c,\mathcal{A}) \text{ for some } \mathcal{A}\}.$$

If $C = (X_0, X_1, X_2)$, then we denote by Q_j, $j = 1, \ldots, 6$, the sector $Q_j(C, \mathbb{A})$, and we denote by ω_j^0 the boundary point of $\mathbb{A}$ associated to Q_j. If we consider the retraction r_c of the building onto the fundamental apartment with respect to the chamber c, assuming $r_c(c) = C$, then

$$r_c(\omega) = \omega_j^0, \quad \forall \omega \in \Omega_{c,j},\ \forall j = 1, \ldots, 6.$$

We recall that $\Omega_{c,j_c} = \Omega(c)$, for a value j_c, depending on c.

Let f be any harmonic function, and let ν be the finitely additive measure on the boundary such that $f = \mathcal{P}^{\overline{x}}\nu$. Then, by setting $P^{\overline{x}}(x,\omega) = P^{\overline{x}}_{\boldsymbol{\alpha}_0}(x,\omega)$,

$$f(x) = \int_\Omega P^{\overline{x}}(x,\omega)d\nu(\omega) = \sum_{j=1}^{6}\int_{\Omega_{c,j}} P^{\overline{x}}(x,\omega)d\nu(\omega), \quad \forall x \in \mathcal{V}.$$

Then, if $\overline{x}$ belongs to c, the retraction of $P^{\overline{x}}(\cdot,\omega)$ with respect to c, $\widetilde{(P^{\overline{x}})_c}(\cdot,\omega)$, does not depend on the choice of ω in $\Omega_{c,j}$, for $j = 1, \ldots, 6$. Therefore the retraction of f with respect to c, $\widetilde{f_c}$, can be expressed as

$$\widetilde{f_c}(X) = \sum_{j=1}^{6} \mu_j(X)\nu(\Omega_{c,j}), \quad \forall X \in \overline{\mathcal{V}},$$

where $\mu_j(X) = \widetilde{(P^{\overline{x}})_c}(X,\omega)$, for all $\omega \in \Omega_{c,j}$. We point out that the coefficients $\mu_1(X), \ldots, \mu_6(X)$ are positive and independent of the function f.

Lemma 3.3. *If $f_0(x) = 1$ for every $x \in \mathcal{V}$, and $f_0 = \mathcal{P}^{\overline{x}}\nu_0$, then $\nu_0(\Omega(c)) > 0$ for every $c \in \mathcal{C}$.*

Proof. If c does not contain $\overline{x}$, and y is a vertex of c, then there exists a measure ν_y, absolutely continuous with respect to ν_0, such that $f_0 = \mathcal{P}^y\nu_y$; moreover (see [8, Lemma 2.15]), $d\nu_y(\omega) = P^{\overline{x}}(y,\omega)d\nu_0(\omega)$. Hence $\nu_0(\Omega(c))$ is positive if $\nu_y(\Omega(c))$ is so. Thus it suffices to prove the result for a chamber c containing the initial vertex $\overline{x}$. With this assumption, for all $X \in \overline{\mathcal{V}}$,

$$\sum_{j=1}^{6} \mu_j(X)\nu_0(\Omega_{c,j}) = 1. \tag{3.1}$$

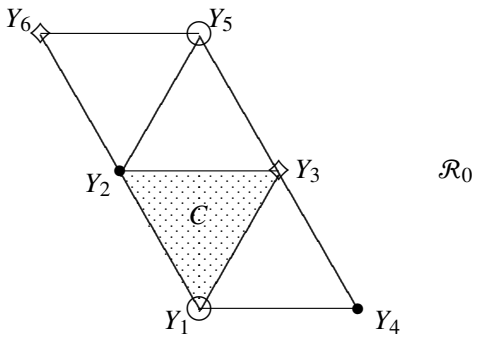


Figure 3.

In particular, if $\{Y_1, \dots, Y_6\}$ are the vertices of the region $\mathcal{R}_0$ of $\mathbb{A}$ pictured in Figure 3, then the matrix $M = (\mu_j(X_k))$ can be computed as in [8, Proposition 3.10], for $\alpha = \alpha_0$, and it is

$$M = \begin{pmatrix} 1 & q^2 & q^2 & 1 & q^3+1-q^{-1} & q^2+q-q^{-2} \\ 1 & q^{-2} & 1 & q+q^{-2}-q^{-3} & q^{-2} & q^{-4} \\ 1 & 1 & q^{-2} & q^{-2} & q^{-2} & q^{-1}+q^{-2}-q^{-3} \\ 1 & q^{-2} & q^{-2} & q^{-1}+q^{-2}-q^{-3} & q^{-4} & q^{-4} \\ 1 & q^2 & 1 & q^{-2} & q+1-q^{-1} & q^2+q-q^{-1} \\ 1 & 1 & q^2 & q+1-q^{-1} & q+1-q^{-1} & q^{-1}+q^{-2}-q^{-3} \end{pmatrix}.$$

Since M is non singular by [8, Proposition 3.10], the linear system

$$\sum_{j=1}^{6} \mu_j(Y_k)\nu_0(\Omega_{c,j}) = 1, \quad k = 1, \dots, 6, \tag{3.2}$$

has a unique solution $(\nu_0(\Omega_{c,1}), \dots, \nu_0(\Omega_{c,6}))$. A direct computation shows that $\nu_0(\Omega_{c,j})$ are strictly positive for all j, then in particular $\nu_0(\Omega(c)) > 0$. □

We denote by Q_l^0 the sector based at X_0 associated with ω_l^0, and by (m_l, n_l) the coordinates of a vertex X of $\mathbb{A}$ with respect to Q_l^0; we say that X tends to ω_l^0, and we write $X \to \omega_l^0$, if $m_l,\ n_l \to +\infty$.

Lemma 3.4. *Let $\nu \in H'(\Omega)$ supported on $\Omega_{c,j}$ for some j, and $\overline{x} \in c$. Then, for $l \neq j$,*

$$\lim_{X \to \omega_l^0} \widetilde{(\mathcal{P}^{\overline{x}}\nu)}_c(X) = 0.$$

Proof. If ν is supported on $\Omega_{c,j}$, then $\widetilde{(\mathcal{P}^{\overline{x}}\nu)}_c(X) = \mu_j(X)\nu(\Omega_{c,j})$ for every X. We shall prove that, for every $l \neq j$, $\lim_{X \to \omega_l^0} \mu_j(X) = 0$.

Let us assume, for ease of notation, $j = 1$. By definition

$$\mu_1(X) = \widetilde{(P^{\overline{x}})_c}(X, \omega_1) = \frac{1}{|r_c^{-1}(X)|} \sum_{x \in r_c^{-1}(X)} P^{\overline{x}}(x, \omega_1)$$

$$= \frac{1}{|r_c^{-1}(X)|} \sum_{x \in r_c^{-1}(X)} \phi(r_{\omega_1}^{\overline{x}}(x)),$$

where $\phi = \phi_{\boldsymbol{\alpha}_0}$, and ω_1 denotes an element of $\Omega_{c,1}$.

We denote simply by (m, n) the coordinates (with respect to Q_1^0) of the vertex X and by (m', n') the coordinates of the vertex $r_{\omega_1}^{\overline{x}}(x)$, for any $x \in r_c^{-1}(X)$. It is easy to check that $m' + n' \leq m + n$. Then

$$\sum_{x \in r_c^{-1}(X)} \phi(r_{\omega_1}^{\overline{x}}(x))$$

is a linear combination of terms $q^{2(m'+n')}$, with $m' + n' \leq m + n$.

If X belongs to the sector Q_l^0, for $l = 2, 3, 4$, then

$$\lim_{X \to \omega_l^0} q^{2(m+n)} = 0.$$

Actually, if X belongs to Q_4^0, then $m = -n_l,\ n = -m_l$; if X belongs to Q_2^0, then $m = -(m_l + n_l),\ n = m_l$; finally, if X belongs to Q_3^0, then $m = n_l,\ n = -(m_l + n_l)$. So, if $X \to \omega_l^0\ \ (l = 2, 3, 4)$, then

$$\sum_{x \in r_c^{-1}(X)} \phi(r_{\omega_1}^{\overline{x}}(x)) \to 0.$$

Assume now $X \in Q_5^0$ (resp., Q_6^0). In this case $m = (m_l + n_l),\ n = -n_l$ (resp., $m = -m_l,\ n = (m_l + n_l)$); so that $m + n = m_l \geq 0$. On the other hand,

$$|r_c^{-1}(X)| = q^d,$$

where d denotes the length of a minimal gallery connecting $C = r_c(c)$ to X. By a direct computation we get $d = 2((m_l + n_l) - 1)$; therefore, $d > 2(m + n)$, for $n_l > 1$ (resp., $m_l > 1$). This implies that, also for $l = 5, 6$, if $X \to \omega_l^0$, then

$$\frac{1}{|r_c^{-1}(X)|} \sum_{x \in r_c^{-1}(X)} \phi(r_{\omega_1}^{\overline{x}}(x)) \to 0.$$

□

Lemmas 3.3 and 3.4 allow to prove the following theorem.

Theorem 3.5. *Let f be any positive harmonic function on $\mathcal{V}$, and let ν be the finitely additive measure on Ω such that $f = \mathcal{P}^{\overline{x}}\nu$. Then ν is a positive measure.*

Proof. It suffices to prove that $\nu(\Omega(c))$ is positive for every chamber c containing the initial vertex $\overline{x}$.

We denote by ν_j, $j = 1, \dots, 6$, the finitely additive measure on Ω obtained as restriction to $\Omega_{c,j}$ of the measure ν, and by f_j the harmonic function $f_j = \mathcal{P}^{\overline{x}}\nu_j$. It is easy to check that $\nu = \nu_1 + \cdots + \nu_6$, and therefore $f = f_1 + \cdots + f_6$. Hence

$$\widetilde{f}_c(X) = (\widetilde{f}_1)_c(X) + \cdots + (\widetilde{f}_6)_c(X) = \mu_1(X)\nu_1(\Omega_{c,1}) + \cdots + \mu_6(X)\nu_6(\Omega_{c,6}).$$

Fix l; if $X \to \omega_l^0$, then Lemma 3.4 implies that $\lim_{X\to\omega_l^0}(\widetilde{f}_j)_c(X) = 0$ for $j \neq l$; thus

$$\lim_{X\to\omega_l^0}(\widetilde{f})_c(X) = \lim_{X\to\omega_l^0}(\widetilde{f}_l)_c(X) = \lim_{X\to\omega_l^0}\mu_l(X)\nu_l(\Omega_{c,l}). \tag{3.3}$$

Since $\lim_{X\to\omega_l^0}\mu_l(X)$ is independent of the choice of f, we compute this limit for the function $f_0 = 1$ and the associated measure ν_0; we obtain

$$1 = \lim_{X\to\omega_l^0}(\widetilde{f}_0)_c(X) = \lim_{X\to\omega_l^0}\mu_l(X)(\nu_0)_l(\Omega_{c,l}) = \lim_{X\to\omega_l^0}\mu_l(X)\nu_0(\Omega_{c,l}),$$

and hence

$$\lim_{X\to\omega_l^0}\mu_l(X) = 1/\nu_0(\Omega_{c,l}) > 0,$$

since $\nu_0(\Omega_{c,l}) > 0$ by Lemma 3.3. Therefore (3.3) implies that $\nu_l(\Omega_{c,l})$ is positive whenever f is so. Choosing l such that $\Omega_{c,l} = \Omega(c)$, we conclude. □

4 Weakly harmonic functions

Let $\mathcal{L}_\theta$ be the linear operator

$$\mathcal{L}_\theta f(x) = \sum_y \theta(x, y) f(y),$$

where $\sum_x \theta(x, y) = 1$ for every y, and $\theta(x, y) = \theta(y, x) \geq 0$.

We assume that $\theta(x, y)$ depends only on the shape of the convex hull $\mathcal{Ch}[x, y]$, and that, for every fixed y, $\theta(x, y) = 0$ for all but a finite number of x. Thus the transition matrix $\theta = (\theta(x, y))$ governs a random walk $(\mathbb{X}_j)$ on the set of vertices of $\mathcal{V}$, with

$$\mathcal{P}r[\mathbb{X}_j = x \mid \mathbb{X}_{j-1} = y] = \theta(x, y).$$

A function f is said to be θ-harmonic if $\mathcal{L}_\theta f = f$, and weakly harmonic if it is θ-harmonic for some θ.

If we choose

$$\theta_0(x, y) = \begin{cases} \frac{1}{2(q^2+q+1)}, & \text{if } d(x, y) = 1, \\ 0, & \text{otherwise,} \end{cases}$$

the operator $\mathcal{L}_0 = \mathcal{L}_{\theta_0}$ plays the role of (the exponent of) the Laplace–Beltrami operator on a symmetric space. From now on we focus our attention on this operator, even if all results actually hold for any $\mathcal{L}_\theta$.

Proposition 4.1. *There exist weakly harmonic positive functions which are not harmonic.*

Proof. Let $\boldsymbol{\alpha} = (a_1, a_2, a_3)$; by setting $a_1 = qb_1$, $a_2 = b_2$ and $a_3 = \frac{1}{q}b_3$, we proved in [8] that $P_{\boldsymbol{\alpha}}(\,\cdot\,, \omega)$ is harmonic if and only if

$$b_1 + b_2 + b_3 = b_1^{-1} + b_2^{-1} + b_3^{-1} = (q^2 + q + 1)/q. \tag{4.1}$$

Analogously we can prove that $P_{\boldsymbol{\alpha}}(\,\cdot\,, \omega)$ is θ_0-harmonic if and only if

$$b_1 + b_2 + b_3 + b_1^{-1} + b_2^{-1} + b_3^{-1} = 2(q^2 + q + 1)/q. \tag{4.2}$$

As only six triples $\boldsymbol{\beta} = (b_1, b_2, b_3)$ satisfy (4.1), while the possible choices for (4.2) are more, we can conclude. □

In analogy with the case of symmetric spaces of noncompact type, we shall prove that any bounded weakly harmonic function is harmonic.

Definition 4.2. For every function f on $\mathcal{V}$ and for every integer $k \geq 1$, we define:

$$\mathcal{E}_k(f)(x) = \begin{cases} f(x), & \text{if } x \in \mathcal{B}_k, \\ |\mathcal{D}_k(x)|^{-1} \sum_{x' \in \mathcal{D}_k(x)} f(x'), & \text{if } x \in \mathcal{V} \setminus \mathcal{B}_k, \end{cases}$$

where (see Figure 4)

$$\mathcal{D}_k(x) = \{x' \in \mathcal{V} : \mathcal{CH}[e, x'] \sim \mathcal{CH}[e, x],\ \mathcal{CH}[e, x'] \cap \mathcal{B}_k = \mathcal{CH}[e, x] \cap \mathcal{B}_k\}.$$

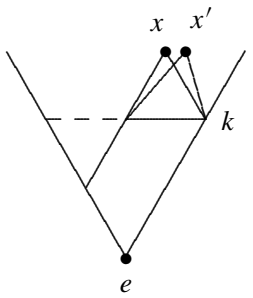


Figure 4.

Keeping in mind the properties of the function $\theta_0(x, y)$, it is easy to verify that $\mathcal{E}_k(f)$ is weakly harmonic if f is so.

Definition 4.3. Fix $\omega \in \Omega$; for every $k \geq 1$, let us define (see Figure 5)

$$V_k(\omega) = \{x \in \mathcal{V} \ : \mathcal{CH}[e, x] \cap \mathcal{B}_k = Q_e(\omega) \cap \mathcal{B}_k\}.$$

A sequence (x_j) is said to converge to ω if, for every $k \geq 1$, $x_j \in V_k(\omega)$ for j big enough, and both its coordinates m_j, n_j (with respect to e) are bigger than k.

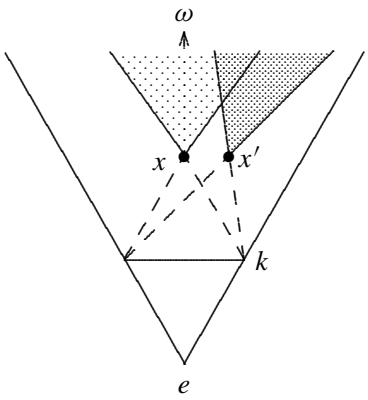


Figure 5.

Lemma 4.4. *Let f be a bounded positive weakly harmonic function. If $f = \mathcal{E}_k(f)$ for some k, then there exists a locally constant function F on Ω such that, for every $\omega \in \Omega$,*

$$f(x_j) \to F(\omega), \quad if \quad x_j \to \omega.$$

Proof. Since $f = \mathcal{E}_k(f)$, we have $f(x) = f(x')$ if $x, x' \in V_k(\omega)$ and $\mathcal{C}h[e, x] \sim \mathcal{C}h[e, x']$. Therefore the restriction of f to $V_k(\omega)$ can be identified with the bounded function $\widetilde{f}$ defined, on the special vertices of the subsector Q of Q_0 based at $X_{k,k}$, by setting $\widetilde{f}(X) = f(x)$ if $\mathcal{C}h[O, X] \sim \mathcal{C}h[e, x]$ (see Figure 6).

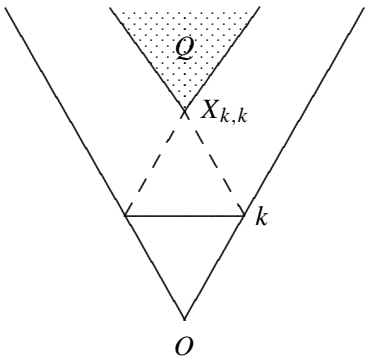


Figure 6.

The function $\widetilde{f}$ satisfies the equation

$$L_0 \widetilde{f}(X) = \sum_Y \widetilde{\theta}_0(X, Y)\, \widetilde{f}(Y) = \widetilde{f}(X), \quad \forall X \in Q,$$

where $\widetilde{\theta}_0(X, Y) = 0$, if $d(X, Y) \neq 1$, and, for $d(X, Y) = 1$,

$$\widetilde{\theta}_0(X, Y) = \begin{cases} \frac{q^2}{1+q+q^2}, & \text{if } d(Y, O) > d(X, O), \\ \frac{q}{1+q+q^2}, & \text{if } d(Y, O) = d(X, O), \\ \frac{1}{1+q+q^2}, & \text{if } d(Y, O) < d(X, O). \end{cases}$$

Therefore the random walk governed by $\widetilde{\theta}_0$ has a drift in the direction of the central axis of Q. This implies that, for every $\epsilon > 0$, the random walk has probability less than ϵ to reach a wall of Q, if the coordinates of its starting vertex X are big enough. Thus, on a suitable subsector $Q' \subset Q$, the function $\widetilde{f}$ is arbitrarily closed to a function ψ which is $\widetilde{\theta}_0$-harmonic and bounded on the whole apartment $\mathbb{A}$. Because of the natural identification of $\mathbb{A}$ with $\mathbb{Z}^2$, we can deduce that the function ψ is constant on $\mathbb{A}$.

This implies that, if (X_j) is a sequence of vertices of Q going to infinity increasing their distance from the walls of Q, then, for every $\epsilon > 0$, $\quad |\widetilde{f}(X_j) - \widetilde{f}(X_k)| < \epsilon$, for j, k big enough, and hence $\widetilde{f}(X_j)$ converges to a value l_ω.

Therefore, if x_j eventually belongs to $V_k(\omega)$ and m_j, n_j tend to infinity, then $f(x_j)$ also converges to the limit l_ω.

Finally we observe that, if $Q_e(\omega) \cap B_k = Q_e(\omega') \cap B_k$, then $V_k(\omega) = V_k(\omega')$ and hence $l_\omega = l_{\omega'}$. Thus, if we choose the locally constant function $F(\omega) = l_\omega$, the lemma is proved. □

Proposition 4.5. *Let $f = \mathcal{E}_k(f)$ for some k; if f is bounded and weakly harmonic, then $f = \mathcal{P}^e(F)$.*

Proof. Let F be the locally constant function defined in Lemma 4.4, and

$$g = f - \mathcal{P}^e(F).$$

We observe that if a sequence (x_j) tends to some $\omega \in \Omega$, then $\mathcal{P}^e(F)(x_j) \to F(\omega)$ (see [9] for more details). Therefore, by Lemma 4.4, $g(x_j) \to 0$ whenever $x_j \to \omega$. Since for every k there are finitely many sets $V_k(\omega)$, this implies that $g(x_j) \to 0$ for every sequence whose coordinates m_j, n_j tend to infinity.

Fix x_0 and consider the random walk $(\mathbb{X}_j)_{j \geq 0}$ starting from x_0 and associated with the operator $\mathcal{L}_0$. Let ρ be the retraction (with respect to the vertex e) of the building on the sector Q_0; then $(\rho(\mathbb{X}_j))$ is a random walk on Q_0, whose transition matrix Θ coincides in the subsector Q of Q_0 with the matrix $\widetilde{\theta}_0$, defined in Lemma 4.4. Since $(\rho(\mathbb{X}_j))$ has a drift in the central axis direction of Q_0, both its coordinates, and hence those of $(\mathbb{X}_j)$, tend to infinity with probability 1. This implies that $(g(\mathbb{X}_j))$ tends to zero with probability 1.

As a consequence, if $\mathbf{E}(g(\mathbb{X}_j))$ denotes the expectation of $g(\mathbb{X}_j)$, the bounded convergence theorem implies

$$\lim_{j \to \infty} \mathbf{E}(g(\mathbb{X}_j)) = 0.$$

Now we observe that $\mathbf{E}(g(\mathbb{X}_j)) = \mathbf{E}(g(\mathbb{X}_{j-1}))$ for any $j \geq 1$. In fact, for every vertex x,

$$\mathbf{E}(g(\mathbb{X}_j) \mid \mathbb{X}_{j-1} = x) = \frac{1}{2(q^2+q+1)} \sum_{d(x,y)=1} g(y) = g(x),$$

and therefore

$$\begin{aligned}\mathbf{E}(g(\mathbb{X}_j)) &= \sum_x \mathbf{E}(g(\mathbb{X}_j) \mid \mathbb{X}_{j-1} = x)\ \ \mathcal{P}r[\mathbb{X}_{j-1} = x] \\ &= \sum_x g(x)\ \mathcal{P}r[\mathbb{X}_{j-1} = x] = \mathbf{E}(g(\mathbb{X}_{j-1})).\end{aligned}$$

Since $\mathbf{E}(g(\mathbb{X}_0)) = g(x_0)$, we conclude that $g(x_0) = 0$. □

Theorem 4.6. *If f is bounded and weakly harmonic, then f is harmonic.*

Proof. We assume $f \geq 0$. For every $k \geq 1$, we consider $\mathcal{E}_k(f)$; Proposition 4.5 implies that $\mathcal{E}_k(f) = \mathcal{P}^e(F_k)$, for some locally constant F_k on Ω. Let ν_k be the positive measure such that $d\nu_k = F_k\, d\mu$. Since

$$f(0) = \mathcal{E}_k(f)(0) = \mathcal{P}^e(F_k)(0) = \int_\Omega d\nu_k(\omega) = \nu_k(\Omega),$$

the sequence (ν_k) is bounded and hence, by the Banach–Alaoglou theorem, there exists a subsequence weakly convergent to a positive measure ν. Finally, for all $x \in \mathcal{V}$, we have:

$$f(x) = \lim_{k\to\infty} \mathcal{E}_k(f)(x) = \lim_{k\to\infty} \int_\Omega P^e(x,\omega) d\nu_k(\omega) = \int_\Omega P^e(x,\omega) d\nu(\omega)\ .$$

This proves that f is harmonic. □

Acknowledgement. We express our thanks to Tim Steger for his availability to discuss this subject with us.

References

[1] D. I. Cartwright, Harmonic functions on buildings of type $\widetilde{A}_n$, in: Random Walks and Discrete Potential Theory (Cortona, 1997), Sympos. Math. XXXIX, Cambridge University Press, Cambridge 1999, 104–138.

[2] D. I. Cartwright and W. Młotkowski, Harmonic analysis for groups acting on triangle buildings, J. Austral. Math. Soc. Ser. A 56 (1994), 345–383.

[3] H. Furstenberg, A Poisson formula for semi-simple Lie groups, Ann. of Math. (2) 77 (1963), 335-386.

[4] P. Gerardin and K. F. Lai, Opérateurs invariants sur les immeubles affines de type A, C. R. Acad. Sci. Paris Sér. I Math. 329 (1999), 1–4.

[5] Y. Guivarc'h, L. Ji and J. C. Taylor, Compactifications of Symmetric Spaces, Progr. Math. 156, Birkhäuser, Boston, MA, 1998.

[6] A. Koranyi, Harmonic functions on symmetric spaces, in: Symmetric Spaces (Short Courses, Washington Univ., St. Louis, Mo., 1969–1970), Pure Appl. Math. 8, Dekker, New York 1972, 379–412.

[7] A. M. Mantero and A. Zappa, Spherical functions and spectrum of the Laplace operators on buildings of rank 2, Boll. Un. Mat. Ital. B (7) 8 (1994), 419–475.

[8] A. M. Mantero and A. Zappa, Eigenfunctions of the Laplace operators for a building of type $\tilde{A}_2$, J. Geom. Anal. 10 (2000), 339–363.

[9] A. M. Mantero and A. Zappa, Boundary behaviour of Poisson integrals on buildings of type $\tilde{A}_2$, Forum Math. 15 (2003), 23–35.

[10] A. M. Mantero and A. Zappa, Eigenfunctions of the Laplace operators for a Building of type $\tilde{G}_2$, preprint 365, Dipartimento di Matematica, Università di Genova (1998).

[11] A. M. Mantero and A. Zappa, Eigenfunctions of the Laplace operators for a Building of type $\tilde{B}_2$, Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat. (8) 5 (2002), 163–195.

[12] M. A. Ronan, Lectures on Buildings, Perspect. Math. 7, Academic Press, Boston, MA, 1989.

Anna Maria Mantero, D.S.A. Facoltá di Architettura, Universitá di Genova, Str. S. Agostino 37, 16123 Genova, Italy.
E-mail: mantero@dima.unige.it

Anna Zappa, Dipartimento di Matematica, Universitá di Genova, Via Dodecaneso 35, 16146 Genova, Italy.
E-mail: zappa@dima.unige.it

Random walks, spectral radii, and Ramanujan graphs

Tatiana Nagnibeda

Abstract. We investigate properties of random walks on trees with finitely many cone types and apply our results to get estimates on spectral radii of groups and to check whether a given finite graph is Ramanujan.

1 Introduction

The notion of a cone type in a graph was introduced by Cannon in [2] (see also [3]). Probabilities on infinite trees with finitely many cone types were first studied by Lyons in [9]. An investigation of random walks on such trees was undertaken in the author's PhD thesis [10] and was pursued in a joint paper with Woess [12]. Two main motivations for this investigation were given by two important classes of trees with finitely many cone types: trees of geodesics of hyperbolic groups (more generally, groups with finitely many cone types, e.g., also all Coxeter groups), and universal covers of finite graphs. Computation of the spectral radius of a certain random walk on the tree allows, in the first case, to get a good estimate on the spectral radius of the group and, in the second case, to decide whether a given finite graph is Ramanujan.

We shall begin by reviewing relevant results on random walks on trees with finitely many cone types and then turn our attention to the two applications.

2 Random walks on trees with finitely many cone types

Let Γ be an infinite, locally finite, connected, simple graph with the vertex set $V(\Gamma)$. For $x, y \in V(\Gamma)$ a *path* of length n from x to y is a sequence $x = x_1, x_2, \ldots, x_{n+1} = y$ of vertices, such that, for each $1 \leq i \leq n$, the vertices x_i and x_{i+1} are neighbours, $x_i \sim x_{i+1}$. The *distance* between two vertices $d(x, y)$ is the length of a shortest path from x to y. A shortest path is called *geodesic*. With a base point $x_0 \in V(\Gamma)$ fixed, the *norm* of a vertex x is $|x| = d(x_0, x)$.

A random walk on Γ is described by a stochastic transition matrix

$$P = \big(p(x, y)\big)_{x,y\in V(\Gamma)},$$

The *Green kernel* on Γ associated with the random walk P is defined by

$$G(x, y \mid z) = \sum_{n=0}^{\infty} p^{(n)}(x, y)z^n,$$

where $p^{(n)}(x, y)$ denotes the probability to go from x to y in n steps. The radius of convergence of the series $G(x, y \mid z)$ is $R_G = 1/\rho \geq 1$ with

$$\rho = \limsup_{n\to\infty} \big(p^{(n)}(x, y)\big)^{\frac{1}{n}}.$$

It is independent of x and y if P is irreducible. The number ρ is called the *spectral radius* of the random walk P. It is indeed the spectral radius of the transition operator P acting on the space $l^2_\sharp(V(\Gamma))$ with the inner product $\langle\phi, \psi\rangle_\sharp = \sum_{x\in V(\Gamma)} \phi(x)\psi(x)\deg(x)$.

Denote by $q^{(n)}(x, y)$ the probability, starting from x, to reach y for the first time after n steps; and consider the kernel

$$F(x, y \mid z) = \sum_{n=0}^{\infty} q^{(n)}(x, y)z^n.$$

In the disc of convergence of G we have

$$\begin{aligned} G(x, y \mid z) &= I(x, y) + F(x, y \mid z)G(y, y \mid z), \\ G(x, x \mid z) &= \big(1 - F(x, x \mid z)\big)^{-1}. \end{aligned} \tag{2.1}$$

The function $F(x, x \mid z)$ is analytic, thus continuous in its disc of convergence; $F(x, x \mid 0) = 0$; and it is increasing in the intersection of $\mathbb{R}_+$ with its disc of convergence. Therefore, either $F(x, x \mid R_F) \leq 1$, or, by continuity, there exists a unique real positive z_0 in the disc of convergence of $F(x, x \mid z)$, such that $F(x, x \mid z_0) = 1$. Therefore R_G either coincides with R_F, if $F(x, x \mid R_F) \leq 1$, or is equal to z_0. The power series $F(x, y \mid z)$ (resp., $G(x, y \mid z)$) has non-negative coefficients, hence by Prinsgheim's Theorem R_F (resp., R_G) is a singular point of $F(x, y \mid z)$ (resp., $G(x, y \mid z)$).

Geometry of the graph is important for the study of random walks on it. *Trees* are particularly nice in this respect as any two vertices x, y are connected by a unique path $(x_1 = x, x_2, \ldots, x_k = y)$ in the tree, and consequently $F(x, y \mid z) = \prod_{i=1}^{k-1} F(x_i, x_{i+1} \mid z)$. The book [13] contains much information and extensive bibliography on random walks on trees.

In this paper we shall turn our attention to *trees with finitely many cone types* which are defined as follows. Let T be an infinite, locally finite tree. We say that a vertex x is a *predecessor* of a vertex y (and y is a *successor* of x), if x and y are neighbours (i.e.,

connected by an edge), and $|x| < |y|$. We define the *cone* $\mathcal{C}(x)$ of a vertex x of T as the induced subgraph of T rooted at x with $V\big(\mathcal{C}(x)\big) = \{y \in V(T) \mid x \text{ belongs to a geodesic from } x_0 \text{ to } y\}$. Two vertices x and y are said to be of a same *cone type* if their cones are isomorphic as rooted graphs. Consider a function $t : V(T) \to \mathbb{Z}_+$, such that $t(x) = 0$ if and only if $x = x_0$ and $t(x_1) = t(x_2)$ if and only if $x_1, x_2 \in V(T) \setminus \{x_0\}$ and are of a same cone type. We will say that a vertex x is *of type* $t(x)$,and t is a *type function* on T. (Note that we assume the type of the base point different from the type of any other vertex, even if the cone of x_0 is isomorphic to the cone of some other vertex.) Since T is a tree, each vertex x of T except x_0 has exactly 1 predecessor $pr(x)$. The number $s_{i,j}$ of successors of type j of a vertex of type i is well defined for every $i \in \mathbb{Z}_+$ and $j \in \mathbb{N}$.

Suppose that there exists a type function t on T which takes values in a finite set $\{0, 1, \ldots, K\}$. We say then that T has finitely many cone types. There is a way to view every infinite tree with finitely many cone types as a cover of a finite digraph. Suppose G is a finite directed graph with a base vertex x_0. The *directed cover of* G *with respect to* x_0 is a rooted infinite tree T^{G,x_0} such that the set of its vertices is in one-to-one correspondence with the set of directed paths in G starting in x_0. Two vertices are connected by an edge (*a priori* undirected) in T^{G,x_0} if one of the corresponding directed paths in G is the expansion of the other one by a directed edge.

Proposition 2.1. *For every finite digraph G and every $x_0 \in V(G)$, the tree T^{G,x_0} has finitely many cone types. Conversely, for every tree T with finitely many cone types there exists a finite digraph G and $x_0 \in V(G)$, such that $T = T^{G,x_0}$.*

Proof. Suppose that two directed paths starting at x_0 in G end at a same vertex. Then the cones of the corresponding vertices in T^{G,x_0} are isomorphic. Thus the number of the cone types in T^{G,x_0} is less than or equal to $1 + |V(G)|$.

Let now T be a tree with $K+1$ cone types. Let $s_{i,j}$ denote the number of successors of type j of a vertex of type i in T. Construct a finite directed graph G as follows. There are $K+1$ vertices $\{x_0, \ldots, x_K\}$ in G. For each $i = 0, 1, \ldots, K$, $j = 1, \ldots, K$, there are $s_{i,j}$ directed edges connecting the vertex x_i with the vertex x_j. Obviously, $T = T^{G,x_0}$. □

Consider an infinite, locally finite tree T with $K+1$ cone types. On T consider a nearest neighbour random walk $P = (p(x, y))_{x,y \in V(T)}$ such that $p(x, pr(x))$ depends only on the cone type of x for any $x \in V(T) \setminus \{x_0\}$, and $p(x, y)$ depends only on the cone types of x and y for any $x \in V(T)$ and for any successor y of x. Consequently, the probability $p(x, pr(x))$ for a vertex of type i, $i \geq 1$ will be denoted by p_{-i}, and the probability $p(x, y)$ will be denoted by $p_{i,j}$ for x of type i and y a successor of x of type j. We have

$$p_{-i} + \sum_{j=1,\ldots,K} s_{i,j}\, p_{i,j} = 1, \quad j = 1, \ldots, K;$$

$$\sum_{j=1,\ldots,K} s_{0,j}\, p_{0,j} = 1.$$

Theorem 2.2. *For every $x, y \in V(T)$, the Green function $G(x, y \mid z)$ is an algebraic element over the field $\mathbb{Q}(z, \{p_{-i}z\}, \{p_{i,j}z\})$. More precisely, there exists a unique (up to a sign) non-constant irreducible polynomial $\mathcal{P}_{x,y}(w, z_1, z_2, \ldots, z_{1+K(K+2)})$ in $2 + K(K+2)$ variables, with integer coefficients, such that*

$$\mathcal{P}_{x,y}\big(G(x, y \mid z), z, p_{-1}z, \ldots, p_{-K}z, \{p_{i,j}z\}_{\substack{i=0,1,\ldots,K \\ j=1,\ldots,K}}\big) \equiv 0.$$

Moreover, if $(x_0, x_1, \ldots, x_k = x)$ denotes the geodesic from x_0 to x, and $(x_0, y_1, \ldots, y_r = y)$ denotes the geodesic from x_0 to y, then the coefficients of the polynomial $\mathcal{P}_{x,y}$ depend only on the cone types of the vertices $x_1, \ldots, x_k, y_1, \ldots, y_r$.

Lemma 2.3. *For any $x \in V(T)\backslash\{x_0\}$ and for any $n \in \mathbb{N}$, the probability $q^{(n)}(x, pr(x))$ depends only on the cone type of x.*

Proof. Obviously $q^{(n)}(x, pr(x)) \neq 0$ only if n is odd. For $n = 1$ we have

$$q^{(1)}(x, pr(x)) = p(x, pr(x)) = p_{-i},$$

where $t(x) = i$. By induction, for $n \geq 3$, n odd,

$$q^{(n)}(x, pr(x)) = \sum_{y \text{ successor of } x} p(x, y) \sum_{k=1}^{\frac{n-1}{2}} q^{(2k-1)}(y, pr(y)) q^{(n-2k)}(x, pr(x))$$

depends only on the cone type of x. □

We can now denote by $q_{-i}^{(n)}$ the probability, starting from a point of type i, to approach x_0 for the first time after n steps. We also denote by $F_{-i}(z)$ the function $F\big(x, pr(x) \mid z\big)$ for any x of type i, and we have

$$F_{-i}(z) = \sum_{n=0}^{\infty} q_{-i}^{(n)} z^n.$$

We shall see that the finite collection of functions $\{F_{-i}(z)\}_{i=1,\ldots,K}$ contains all the information about the Green kernel on T and its radius of convergence. (Note that as T is a tree, it therefore also contains all the information about the Martin boundary of (T, P).)

Proposition 2.4. *The power series $F_{-i}(z)$, $i = 1, \ldots, K$, satisfy the following system of polynomial equations:*

$$w_i = p_{-i}z + z \sum_{j=1}^{K} s_{i,j} p_{i,j} w_i w_j, \quad i = 1, \ldots, K. \tag{2.2}$$

Proof. As in Lemma 2.3, we have

$$q_{-i}^{(1)} = p_{-i},$$

and, for $n \geq 3$

$$q_{-i}^{(n)} = \sum_{j=1}^{K} s_{i,j}\, p_{i,j} \sum_{k=1}^{\frac{n-1}{2}} q_{-j}^{(2k-1)} q_{-i}^{(n-2k)}.$$

Thus,

$$\begin{aligned} F_{-i}(z) &= \sum_{n=0}^{\infty} q_{-i}^{(n)} z^n = z p_{-i} + z \sum_{j=1}^{K} s_{i,j}\, p_{i,j} \sum_{n=3}^{\infty} \sum_{k=1}^{\frac{n-1}{2}} q_{-j}^{(2k-1)} z^{2k-1} q_{-i}^{(n-2k)} z^{n-2k} \\ &= z p_{-i} + z \sum_{j=1}^{K} s_{i,j}\, p_{i,j} F_{-i}(z) F_{-j}(z). \end{aligned}$$

□

Let us now express functions $F(x, y \mid z)$, $x, y \in V(T)$, $x \sim y$, in terms of the functions $\{F_{-i}(z)\}_{i=1,\ldots,K}$. First, we have

$$F(x_0, x_0 \mid z) = \sum_{j=1}^{K} s_{0j}\, p_{0j}\, z F_{-j}(z). \tag{2.3}$$

Now let x be a vertex of T and y a successor of x. Then, for $n \geq 3$, n odd,

$$\begin{aligned} q^{(n)}(x, y) = &\sum_{y' \text{ a successor of } x} p(x, y') \sum_{r=1}^{\frac{n-1}{2}} q^{(2r-1)}(y', pr(y')) q^{(n-2r)}(x, y) \\ &+ p(x, pr(x)) \sum_{r=1}^{\frac{n-1}{2}} q^{(2r-1)}(pr(x), x) q^{(n-2r)}(x, y). \end{aligned}$$

As follows from Lemma 2.3, the probabilities $p(x, y')$ and $q^{(r)}(y', pr(y'))$ depend only on the cone type of x and its successors (which are determined by the cone type of x). On the contrary, the cone type of $pr(x)$ is not determined by the cone type of x, thus the probabilities $p(x, pr(x))$ and $q^{(r)}(pr(x), x)$ depend also on the cone type of $pr(x)$. The probabilities $q^{(r)}(pr(x), x)$ depend in turn on the cone types of x, $pr(x)$, and $pr(pr(x))$. Each x in $V(T)$ is uniquely determined by the sequence $(x, pr(x), pr(pr(x)), \ldots, pr^n(x) = x_0)$, which coincides with the unique geodesic from x to x_0 in T. Hence, the probabilities $q^{(n)}(x, y)$ depend on the cone types of y, x, and all the vertices which lie on the geodesic connecting x and x_0 in T. More precisely, the following holds.

Theorem 2.5. *Let x be a vertex of T, y be a successor of x of type j, and let $(x_0, x_1, \ldots, x_k = x)$ with $t(x_m) = i_m$ for $m = 1, \ldots, k$ be the unique geodesic*

from x_0 to x. Then

$$F(x, y \mid z) = \cfrac{zp_{i_k,j}}{H_k(z) + \cfrac{z^2 p_{i_{k-1},i_k}}{H_{k-1}(z) + \cfrac{z^2 p_{i_{k-2},i_{k-1}}}{\ddots \\ H_1(z) + \cfrac{z^2 p_{0,i_1}}{H_0(z)}}}},$$

where

$$H_k(z) = 1 - z \sum_{\substack{l=1 \\ l \neq j}}^{K} p_{i_k,l} F_{-l}(z) - z p_{i_k,j}(s_{i_k,j} - 1) F_{-j}(z);$$

for $m = k-1, \ldots, 1$:

$$H_m(z) = 1 - z \sum_{\substack{l=1 \\ l \neq i_{m+1}}}^{K} p_{i_m,l} s_{i_m,l} F_{-l}(z) - z p_{i_m,i_{m+1}}(s_{i_m,i_{m+1}} - 1) F_{-i_{m+1}}(z);$$

$$H_0(z) = 1 - z \sum_{\substack{l=1 \\ l \neq i_1}}^{K} p_{0,l} s_{0,l} F_{-l}(z) - z p_{0,i_1}(s_{0,i_1} - 1) F_{-i_1}(z).$$

Proof. As shown above, the function $F(x, y \mid z)$ depends only on $j, i_1, \ldots, i_k$, and we denote $F(x, y \mid z) = F_{0,i_1,\ldots,i_k,j}(z)$. The easiest case is that of the function $F_{0,i_1}(z)$. We have the following formulas for the probability to go from x_0 to one of its neighbours x of type i:

$$q_{0,i}^{(1)} = p_{0,i},$$

$$q_{0,i}^{(n)} = \sum_{\substack{j=0 \\ j \neq i}}^{K} p_{0,j} s_{0,j} \sum_{k=1}^{\frac{n-1}{2}} q_{-j}^{(2k-1)} q_{0,i}^{(n-2k)} + p_{0,i}(s_{0,i} - 1) \sum_{k=1}^{\frac{n-1}{2}} q_{-i}^{(2k-1)} q_{0,i}^{(n-2k)}, \quad n \geq 3.$$

Thus,

$$F_{0,i}(z) = z p_{0,i} + z \sum_{\substack{j=1 \\ j \neq i}}^{K} p_{0,j} s_{0,j} F_{-j}(z) F_{0,i}(z) + z p_{0,i}(s_{0,i} - 1) F_{-i}(z) F_{0,i}(z),$$

and

$$F_{0,i}(z) = \frac{z p_{0,i}}{1 - z \sum_{\substack{j=1 \\ j \neq i}}^{K} p_{0,j} s_{0,j} F_{-j}(z) - z p_{0,i}(s_{0,i} - 1) F_{-i}(z)}.$$

The probability of going from a point x (to which we associate the geodesic $(x_0, x_1, \dots, x_k = x)$ with $t(x_m) = i_m, m = 1, \dots, k$) to one of its successors of type j in n steps, can be computed similarly:

$$\begin{aligned} q^{(1)}_{0,i_1,\dots,i_k,j} &= p_{i_k,j}, \\ q^{(n)}_{0,i_1,\dots,i_k,j} &= \sum_{\substack{l=0 \\ l\neq j}}^{K} p_{i_k,l}s_{i_k,l} \sum_{r=1}^{\frac{n-1}{2}} q^{(2r-1)}_{-l} q^{(n-2r)}_{0,i_1,\dots,i_k,j} \\ &\quad + p_{i_k,j}(s_{i_k,j}-1) \sum_{r=1}^{\frac{n-1}{2}} q^{(2r-1)}_{-j} q^{(n-2r)}_{0,i_1,\dots,i_k,j} \\ &\quad + p_{-i_k} \sum_{r=1}^{\frac{n-1}{2}} q^{(2r-1)}_{0,i_1,\dots,i_{k-1},i_k} q^{(n-2r)}_{0,i_1,\dots,i_k,j}, \quad n \geq 3, \end{aligned}$$

which implies

$$\begin{aligned} F_{0,i_1,\dots,i_k,j}(z) &= zp_{i_k,j} + z \sum_{\substack{l=1 \\ l\neq j}}^{K} p_{i_k,l}s_{i_k,l}F_{-l}(z)F_{0,i_1,\dots,i_k,j}(z) \\ &\quad + zp_{i_k,j}(s_{i_k,j}-1)F_{-j}(z)F_{0,i_1,\dots,i_k,j}(z) \\ &\quad + zp_{-i_k}F_{0,i_1,\dots,i_{k-1},i_k}(z)F_{0,i_1,\dots,i_k,j}(z). \end{aligned}$$

Hence we have

$$F_{0,i_1,\dots,i_k,j}(z) = \frac{zp_{i_k,j}}{\Delta},$$

where

$$\Delta = 1 - z \sum_{\substack{l=1 \\ l\neq j}}^{K} p_{i_k,l}s_{i_k,l}F_{-l}(z) - zp_{i_k,j}(s_{i_k,j}-1)F_{-j}(z) - zp_{-i_k}F_{0,i_1,\dots,i_{k-1},i_k}(z).$$

In this formula for $F_{0,i_1,\dots,i_k,j}(z)$, we can replace $F_{0,i_1,\dots,i_{k-1},i_k}(z)$ by the similar expression, in which only the functions $\{F_{-i}(z)\}$ and the function $F_{0,i_1,\dots,i_{k-2},i_{k-1}}(z)$ appear. Repeating this procedure $k-1$ times, we get an expression for $F_{0,i_1,\dots,i_{k-1},i_k}$ in terms of the functions $\{F_{-i}(z)\}$ and the function $F_{0,i_1}(z)$, which in its turn can be expressed only in terms of the functions $\{F_{-i}(z)\}$. Putting all these expressions together, we get the desired formula for $F_{0,i_1,\dots,i_k,j}(z)$. □

Proof of Theorem 2.2. It follows from Proposition 2.4 that the functions $F_{-i}(z)$ are algebraic elements over the field $\mathbb{Q}(z, \{p_{-i}z\}, \{p_{i,j}z\})$. Theorem 2.5 shows that the functions $F(x, y \mid z)$ can be obtained from $\{F_{-i}(z)\}$ by a finite number of operations of addition, multiplication by a scalar from the field $\mathbb{Q}(z, \{p_{-i}z\}, \{p_{i,j}z\})$, and taking

inverse. Thus each function $F(x, y \mid z)$ is an algebraic element over the same field. Consequently, $G(x, y \mid z)$ is also algebraic over the same field by formula (2.1). □

It can be deduced from Theorem 2.2 that the spectral radius of a tree with finitely many cone types is an algebraic number. Moreover, we shall explain presently how it can be computed from the system (2.2). Computations of this type had appeared in many different places, in a similar context, e.g., in Lalley's study of finite range random walks on a free group [7], see also [13]. A similar method was used in [12] to compute the rate of escape of the random walk P on T.

Lemma 2.6. *Suppose that for every vertex $x \in V(T)$ the cone $\mathcal{C}(x)$ contains vertices of all types $1, \dots, K$. Denote by R_i the radius of convergence of $F_{-i}(z)$, $i = 1, \dots, K$. Then $R_F = R_1 = \cdots = R_K$.*

Proof. Let x be a vertex of type i. For $j \in \{1, \dots, K\}$, denote by $n_{i,j}$ the minimal distance between x and a vertex of type j in the cone $\mathcal{C}(x)$. Let $y \in \mathcal{C}(x)$ be a vertex of type j which lies at distance $n_{i,j}$ from x. Denote by $(x_0 = x, x_1, \dots, x_{n_{i,j}-1}, x_{n_{i,j}} = y)$ the geodesic between x and y, and by t_k the type of x_k, $k = 1, \dots, n_{i,j} - 1$. Put

$$C_{i,j} := p_{it_1} p_{t_1 t_2} \cdots p_{t_{n_{i,j}-1} j} p_{-j} p_{-t_{n_{i,j}-1}} \cdots p_{-t_1} p_{-i} \ .$$

Then $F_{-i}(z) \geq C_{i,j} F_{-j}(z)$, but also $F_{-j}(z) \geq C_{j,i} F_{-i}(z)$. Thus $R_i = R_j$. We further have $R_F = R_1 = \cdots = R_K$ by (2.3). □

Denote by $J(z)$ the Jacobian matrix of the system (2.2)

$$J(z) = \left(\frac{\partial \mathcal{P}_i(z, w_1, \dots, w_K)}{\partial w_j} \right)_{i,j=1,\dots,K},$$

where $\mathcal{P}_i(z, w_1, \dots w_K) = p_{-i} z + z \sum_{j=1}^{K} s_{i,j} p_{i,j} w_i w_j$, $i = 1, \dots, K$. Under the assumption of Lemma 2.6, the matrix $J(z)$ is irreducible, and the Perron–Frobenius Theorem can be applied to it to conclude that $J(z)$ has a positive simple eigenvalue $\lambda_*(z)$ which maximizes the eigenvalues of $J(z)$ in absolute value.

Theorem 2.7. *Under the assumption of Lemma* 2.6

$$R_F = \min\{z > 0 \mid \lambda_*(z) = 1\}.$$

Proof. The argument from [13, p. 211] applies without any changes. □

Remark 2.8. The inverse of the spectral radius, R_G, is equal to R_F unless the random walk is ρ-positive recurrent (see [12, 2.4]). In this latter case R_G is a pole and the unique solution of $F(x_0, x_0 \mid z) = 1$.

3 Spectral radius of a group with finitely many cone types

Let G be a finitely generated group, S a finite symmetric system of generators in G, $|S| = m$. By the simple random walk on (G, S) we mean the simple random walk on the Cayley graph of G with respect to S. The spectral radius plays an important role in the study of random walks on groups. To compute it explicitly is often a difficult task, and efforts have been made to get good estimates on the spectral radius for different groups. A celebrated theorem of Kesten [5, 6] states that $2\sqrt{m-1}/m \leq \rho_G \leq 1$ with equality on the left if and only if G is a free product of several copies of $\mathbb{Z}$ and $\mathbb{Z}_2$, and S are standard generators; and with equality on the right if and only if G is amenable.

The aim of this section is to provide a method of estimating the spectral radius ρ_G of the simple random walk on (G, S) in the situation when the corresponding Cayley graph is bipartite and has finitely many cone types, as, for example, when G is hyperbolic with all relations of even length, or when (G, S) is a Coxeter system. (Let us mention that no example is known of a group which has finitely many cone types with respect to some but not all generating sets.)

We are going to "simulate" the simple random walk on (G, S) by constructing a special random walk on its *tree of geodesics* $T_{G,S}$. The (infinite) vertex set $V(T_{G,S})$ of this tree is the set of geodesic segments of the form $\{[1_G, g]\}_{g \in G}$ in the Cayley graph $\mathrm{Cay}(G, S)$. Two vertices γ_1, γ_2 in $T_{G,S}$ are connected by an edge if the corresponding geodesic segment γ_1 in $\mathrm{Cay}(G, S)$ is a one-step extension of the segment γ_2. The tree $T_{G,S}$ has naturally a base vertex γ_0 corresponding to the null geodesic segment $[1_G, 1_G]$.

There is of course a natural projection of $T_{G,S}$ to $\mathrm{Cay}(G, S)$:

$$\begin{array}{rccc} \theta: & V(T_{G,S}) & \to & G \\ & \gamma = [1_G, g] & \mapsto & g. \end{array}$$

This map is locally injective as the induced map on the set of edges is of the form $\big([1_G, g], [1_G, gs]\big) \mapsto (g, gs)$. However, θ is not a covering map because it is not a local isomorphism.

Observe that the notion of cone type defined in Section 2 for trees, can be repeated without any changes for any bipartite graph. The number of predecessors of a vertex is determined by its cone type and will be denoted r_i for a vertex of type i. The tree of geodesics $T_{G,S}$ has the same number of cone types as $\mathrm{Cay}(G, S)$. Indeed, $\theta(\mathcal{C}(\gamma)) = \mathcal{C}(\theta(\gamma))$ for every $\gamma \in V(T_{G,S})$.

Suppose the number of cone types in $\mathrm{Cay}(G, S)$ is $K + 1$. Consider the nearest neighbour random walk P_G on $T_{G,S}$ defined by the following transition probabilities:

$$\begin{aligned} p_{i,j} &= 1/m, \quad i = 0, \ldots, K; \ j = 1, \ldots, K. \\ p_{-i} &= r_i/m \quad i = 1, \ldots, K. \end{aligned}$$

Theorem 3.1. *The spectral radius ρ_G of the simple random walk on G with respect to S is bounded from above by the spectral radius ρ_T of the random walk P_G on $T_{G,S}$.*

For $i = 0, \dots, K$, consider the functions $f_i : \mathbb{R}_+^K \to \mathbb{R}_+$ defined by

$$f_i(c_1, \dots, c_K) = r_i c_i + \sum_{j=1}^{K} \frac{s_{i,j}}{c_j}. \tag{3.1}$$

Lemma 3.2 ([11, Section 2]). *The minimum*

$$\min_{(c_1,\dots,c_K)\in\mathbb{R}_+^K} \max_{i=1,\dots,K} f_i(c_1, \dots, c_K)$$

exists and is at least $m\rho_G$.

Proof of Theorem 3.1. By Lemma 3.2 it is enough to show that

$$\rho_T \ \geq \ \frac{1}{m} \min_{(c_1,\dots,c_K)\in\mathbb{R}_+^K} \max_{i=1,\dots,K} f_i(c_1, \dots, c_K). \tag{3.2}$$

Recall that by (2.1), $1/\rho_T \leq R_F$. Formula (2.3) implies in turn that R_F is equal to the minimum (over $i = 1, \dots, K$) of the smallest real positive singularities of the functions $F_{-i}(z)$. These functions satisfy the system of equations (2.2). For a fixed z this system can be rewritten in the following form

$$\frac{m}{z} = \frac{r_i}{w_i} + \sum_{j=1}^{K} s_{i,j} w_j, \quad i = 1, \dots, K.$$

For each $\epsilon > 0$, let $z_\epsilon = R_F - \epsilon \in \mathbb{R}_+$. Then there exists a solution $\big(F_{-1}(z_\epsilon), \dots, F_{-K}(z_\epsilon)\big) \in \mathbb{R}_+^K$ of the system above. The functions $F_{-i}(z)$ are increasing in the intersection of their disc of convergence with $\mathbb{R}_+$, thus $F_{-i}(z_\epsilon) \neq 0$. We can put $c_i = 1/F_{-i}(z_\epsilon)$ in (3.1), and get

$$\begin{aligned} f_1\big(1/F_{-1}(z_\epsilon), \dots, 1/F_{-K}(z_\epsilon)\big) \ &= \cdots = f_K\big(1/F_{-1}(z_\epsilon), \dots, 1/F_{-K}(z_\epsilon)\big) \\ &= \max_{i=1,\dots,K} f_i\big(1/F_{-1}(z_\epsilon), \dots, 1/F_{-K}(z_\epsilon)\big) = \frac{m}{z_\epsilon} = \frac{m}{R_F - \epsilon}. \end{aligned}$$

Consequently,

$$\min_{(c_1,\dots,c_K)\in\mathbb{R}_+^K} \max_{i=1,\dots,K} f_i(c_1, \dots, c_K) \ \leq \ \frac{m}{R_F - \epsilon},$$

and

$$\rho_T \ \geq \ \frac{1}{m} \min_{(c_1,\dots,c_K)\in\mathbb{R}_+^K} \max_{i=1,\dots,K} f_i(c_1, \dots, c_K).$$

□

Remark 3.3. It can be shown that the inequality (3.2) is in fact an equality.

An advantage of the estimate provided by Theorem 3.1 is that it can be computed. Indeed, the tree of geodesics $T_{G,S}$ has finitely many cone types and the random walk

P_G is of the type we have considered in Section 2. So the system of equations (2.2) can be written for it, and the radius of convergence of its Green function can be found.

One particular instance in which this procedure can be carried out resulting in a good numerical estimate of the spectral radius ρ_G, is that of surface groups. Consider the fundamental group of an orientable surface of genus $g \geq 2$

$$G_g = \left\langle a_1, \ldots a_g, b_1, \ldots, b_g \mid \prod_{i=1}^{g} [a_i, b_i] \right\rangle.$$

The following estimates hold for the spectral radii ρ_g for the groups G_g of small genus:

Theorem 3.4.

$$0.662420 \leq \rho_2 \leq 0.662816,$$
$$0.552773 \leq \rho_3 \leq 0.552792.$$

The upper bounds are deduced from Theorem 3.1. They first appeared in [11] where a different method was used for computing (3.2). The lower bounds were obtained by L. Bartholdi [1].

4 How to recognize a Ramanujan graph?

Let now Γ be a finite connected graph on n vertices, and denote by $\lambda_1(\Gamma) > \lambda_2(\Gamma) \geq \lambda_3(\Gamma) \cdots \geq \lambda_n(\Gamma)$ the spectrum of its adjacency matrix. If Γ is k-regular, $\lambda_1(\Gamma) = k$. Recall that a k-regular connected finite graph Γ is called *Ramanujan* if

$$\lambda_2(\Gamma) \leq 2\sqrt{k-1}.$$

The number on the right-hand side of this inequality is nothing else than the spectral radius of the adjacency operator on the infinite k-regular tree, which is the universal covering tree of any k-regular graph.

In analogy with this classical definition, one can say that an arbitrary finite connected graph Γ is Ramanujan if

$$\lambda_2(\Gamma) \leq r(\tilde{\Gamma}),$$

where $\tilde{\Gamma}$ denotes the universal covering tree of Γ and r is the spectral radius of the adjacency operator on it. Note that if an infinite tree covers a finite graph then it covers infinitely many finite graphs (such trees are *uniform*), and so it makes sense to speak of infinite sequences of these, not necessarily regular Ramanujan graphs. Moreover, all finite graphs covered by the same tree have the same λ_1, as follows from Leighton's Theorem.

Such graphs were studied by Greenberg [4], who proved a generalization of Alon-Boppana theorem, namely that $r(T)$ is the best upper bound on $\lambda_2(\Gamma)$ which holds for infinitely many graphs Γ covered by the same uniform tree T. Later Lubotzky and the

author showed that the famous classical problem of existence, for every regular tree, of an infinite family of Ramanujan quotients, gets negative solution if the assumption of regularity is dropped. More precisely, infinitely many uniform trees such that none of their quotient is Ramanujan, were exhibited in [8].

Alex Lubotzky has asked, while discussing different aspects of the notion of Ramanujan graph, whether an algorithm exists which, given a finite connected graph, determines in finite time whether this graph is Ramanujan. According to the definition above, the question is whether one can compute algorithmically the value of the spectral radius of a uniform tree. Results of Section 2 can be applied to solve this problem.

Proposition 4.1. *A uniform tree has finitely many cone types.*

Proof. Suppose T is a uniform tree, i.e., there exists a finite connected graph Γ such that $T = \tilde{\Gamma}$. Then the number of cone types in T (counted with respect to any base point) is at most $1 + \sum_{v \in V(\Gamma)} \deg(v)$. Indeed, let π denote the covering map $\tilde{\Gamma} \to \Gamma$. Fix a base point x_0 in Γ and some $\tilde{x}_0 \in \pi^{-1}(x_0)$. Consider the set $\mathbb{E}(\Gamma)$ of oriented edges of Γ, of cardinality $2|E(\Gamma)|$. A set $E_e(\tilde{\Gamma})$ of edges in $\tilde{\Gamma}$ is associated naturally to each edge e in $\mathbb{E}(\Gamma)$, so that $\pi(\epsilon) = e$ for each $\epsilon \in E_e(\tilde{\Gamma})$. Each edge $\epsilon \in E(\tilde{\Gamma})$ can be represented as (ϵ_+, ϵ_-) with $\epsilon_+, \epsilon_- \in V(\tilde{\Gamma})$ and $d(\epsilon_+, \tilde{x}_0) = d(\epsilon_-, \tilde{x}_0) + 1$. With each such ϵ we can associate the cone of its extremity ϵ_+. The cones which correspond to the edges from some $E_e(\tilde{\Gamma})$ are isomorphic. As we assumed that the cone type of the base vertex is different from all others, it follows that the number of cone types minus 1 is bounded from above by the number of oriented edges in Γ, $\sum_{v \in V(\Gamma)} \deg(v)$. □

Note however that not every tree with finitely many cone types is the universal cover of a finite graph. For example, trees of geodesics of finitely generated groups studied in Section 3 are in general not uniform.

Proposition 4.1 above insures that results of Section 2 can be used for finding spectral radii of uniform trees. The proof shows that the assumption of irreducibility, as stated in Lemma 2.6, is satisfied. Thus Theorem 2.7 can be applied to $T = \tilde{\Gamma}$. Note that Theorem 2.7 concerns the spectral radius of the simple random walk, whereas it is the spectral radius of the adjacency operator which appears in the definition of a Ramanujan graph. Though there is no simple formula connecting these two numbers in the case when the tree is not regular, a complete analogue of Theorem 2.7 holds for the adjacency operator. Namely, the generating functions $\{F_{-i}(z)\}_{i=1,\dots,K}$, and the function $\lambda_*(z)$ (defined for the adjacency operator) satisfy the following system of polynomial equations:

$$\begin{cases} w_i = z + z \sum_{j=1}^{K} s_{i,j} w_i w_j, \quad i = 1, \dots, K \\ \det(J(z) - \lambda I) = 0. \end{cases}$$

The *elimination theory* of algebraic geometry ensures the existence of polynomials $\mathcal{Q}$, $\mathcal{Q}_i$, $i = 1, \dots, K$, in two variables, with integer coefficients, such that $\mathcal{Q}(z, \lambda_*(z)) \equiv$

0 and $\mathcal{Q}_i(z, F_{-i}(z)) \equiv 0$. Moreover, there are algorithms running in polynomial time to find coefficients of these polynomials from the coefficients of the system of equations. The radius of convergence R_F can then be found as a solution of the equation $\mathcal{Q}(z, 1) = 0$.

Being an algebraic function, each $F_{-i}(z)$ has an expansion as a Puiseux series, and the constant term in it is exactly the value $F_{-i}(R_i)$. These values can be found for all $i = 1, \ldots, K$ by plugging the Puiseux expansions in the system (2.2) (see [12, Section 2.7].) Therefore we also get the value $F(x_0, x_0 \mid R_F)$ via (2.3). If it is at most 1, the spectral radius is equal to $1/R_F$. Finally, if it appears that $F(x_0, x_0 \mid R_F) > 1$ and the radius of convergence of the Green function is a pole, polynomials $\mathcal{Q}_i$, $i = 1, \ldots, K$ can be used to find the polynomial $\mathcal{P}$ such that $\mathcal{P}(z, F(x_0, x_0 \mid z)) \equiv 0$, and the spectral radius can be found as a solution of the equation $\mathcal{P}(z, 1) = 0$.

Acknowledgement. The author would like to acknowledge the hospitality of the Erwin Schrödinger Institute and thank Vadim Kaimanovich, Klaus Schmidt and Wolfgang Woess for organizing an excellent semester on Random Walks in Vienna in 2001.

References

[1] L. Bartholdi, Cactus trees and lower bounds on the spectral radius of vertex-transitive graphs, in: Random Walks and Geometry, Proceedings of a Workshop at the Erwin Schrödinger Institute, Walter de Gruyter, Berlin 2004, 349–361.

[2] J. Cannon, The growth of the closed surface groups and the compact hyperbolic Coxeter groups, preprint (1980).

[3] J. Cannon, The combinatorial structure of cocompact discrete hyperbolic groups, Geom. Dedicata 16 (1984), 123–148.

[4] Y. Greenberg, Ph.D. thesis, Hebrew University of Jerusalem, 1995.

[5] H. Kesten, Symmetric random walks on groups, Trans. Amer. Math. Soc. 22 (1959), 336–354.

[6] H. Kesten, Full Banach mean values on countable groups, Math. Scand. 7 (1959), 146–156.

[7] S. Lalley, Finite range random walk on free groups and homogeneous trees, Ann. Probab. 21 (1993), 2087–2130.

[8] A. Lubotzky, T. Nagnibeda, Not every uniform tree covers Ramanujan graphs, J. Combin. Theory Ser. B 74 (1998), 202–212.

[9] R. Lyons, Random walks and percolation on trees, Ann. Probab. 18 (1990), 931–958.

[10] T. Nagnibeda, On random walks and growth in groups with finitely many cone types, Ph.D. thesis, University of Geneva, 1997.

[11] T. Nagnibeda, An upper bound for the spectral radius of a random walk on surface groups, J. Math. Sci. (New York) 96 (1999), 3542–3549.

[12] T. Smirnova-Nagnibeda, W. Woess, Random walks on trees with finitely many cone types, J. Theoret. Probab. 15 (2002), 383–422.

[13] W. Woess, Random Walks on Infinite Graphs and Groups, Cambridge Tracts in Math. 138, Cambridge University Press, Cambridge 2000.

Tatiana Nagnibeda, Department of Mathematics, Royal Institute of Technology, S-10044 Stockholm, Sweden

E-mail: tatiana@math.kth.se

Cogrowth of arbitrary graphs

Sam Northshield

Abstract. A "cogrowth set" of a graph G is the set of vertices in the universal cover of G which are mapped by the universal covering map onto a given vertex of G. Roughly speaking, a cogrowth set is large if and only if G is small. In particular, when G is regular, a cogrowth constant (a measure of the size of the cogrowth set) exists and has been shown to be as large as possible if and only if G is amenable.

We present two approaches to the problem of extending this to the non-regular case. First, we show that the result above extends to the case when G is not regular but is the cover of a finite graph. This proof is based on some properties of a family of Laplacians related to the zeta function of the covered graph. An example is given where this result fails when G does not cover a finite graph. Second, for any graph with transient covering tree, we define a new cogrowth constant expressed in terms of harmonic measure and show that G is amenable if and only if this constant is 1.

Finally, we show that if G covers a finite graph, then the radial limit set of a cogrowth set has largest possible Hausdorff dimension if and only if G is amenable.

1 Introduction

The concept of amenability originated with von Neumann who once conjectured, though not in these words, that every non-amenable group was an extension of a free group on two generators. This conjecture turned out to be false but it was not until 1984 that a counterexample was found. Ol'shanskii elaborately constructed a group which was neither a finite extension of F_2 nor was amenable; this last step utilized a criterion for amenability developed by Grigorchuk. Essentially, a finitely generated group is not amenable if the number of reduced words of length n grows at the same rate as the number of reduced words of length less than n. Since every finitely generated group is the quotient of a free group F, it is conceivable that a coset in the quotient of F is big in some well defined way if and only if G is amenable. Grigorchuk's criterion is this: G is amenable if and only if the number of words of length n in a coset grows as fast as the total number of words of length n in F grows.

It was later noticed that this result can be extended to regular graphs (see [8], for example). The concept of amenability was extended to graphs by Gerl: we say that a

graph is amenable if and only if

$$\inf_K \frac{|\partial K|}{|K|} = 0,$$

where the infimum is over all finite non-empty sets of vertices in G, and ∂K is the set of all edges connecting vertices of K to vertices not in K.

A d-regular graph is covered by a d-regular tree T (i.e., there exists a map θ from the vertices of T onto the vertices of G which preserves vertex degree and adjacency). Clearly, the number of vertices of distance n from a fixed vertex o in T is asymptotic to $(d-1)^n$, and we say that the "growth number" of T is $\mathrm{gr}(T) = d-1$. The "coset" $[o] = \theta^{-1}(o)$ also has a growth rate which we call the "cogrowth number" of G and define by

$$\mathrm{cogr}(G) = \limsup_{n\to\infty} |S_n(o) \cap [o]|^{\frac{1}{n}}, \tag{1.1}$$

where $S_n(x)$ is the metric sphere in T of radius n and center x. It has been shown (see [8]) that G is amenable if and only if

$$\mathrm{cogr}(G) = d-1.$$

Our aim is to extend this to the case when G has bounded vertex degree but is not necessarily regular. First, in the non-regular case, although $\mathrm{cogr}(G)$ will still be defined by (1.1), the quantity $d-1$ no longer represents the growth of T; we define the growth number of T to be, in general,

$$\mathrm{gr}(T) = \limsup_{n\to\infty} |S_n(o)|^{\frac{1}{n}}.$$

We note that our definition of $\mathrm{gr}(T)$ differs from that in Lyons [7] (he uses the lim inf), but, under the additional hypothesis that G covers a finite graph, $\lim_{n\to\infty} |S_n(o)|^{\frac{1}{n}}$ exists and thus both definitions agree. A natural conjecture is that $\mathrm{cogr}(G) = \mathrm{gr}(T)$ if and only if G is amenable. Unfortunately, this is not true in general (see example below); a main difficulty is that if G has arbitrarily long chains (i.e., sequences of adjacent vertices of degree 2) then G is amenable but $\mathrm{cogr}(G) < \mathrm{gr}(T)$. One criterion that eliminates these possibilities is that G covers a finite graph. Then indeed we get the desired result:

Theorem 1. *Let G be a simple connected graph which covers a finite graph. Then G is amenable if and only if* $\mathrm{cogr}(G) = \mathrm{gr}(T)$.

This result seems fairly "tight" in that the hypothesis that G covers a finite graph is used several times in seemingly independent places in the proof.

Our second generalization involves the topological boundary ∂T of T. The random walk in T, if transient, converges to a point in ∂T, and the distribution of that random point is "harmonic" measure (called that since every harmonic function on T has an integral representation with respect to harmonic measure). In the regular case,

harmonic measure has a particularly simple form:

$$\mu(T_\eta) = (d-1)^{-|\eta|},$$

where T_η is the set of rays starting at o which go through η. This hints at how to extend $\mathrm{gr}(T)$ and $\mathrm{cogr}(G)$: let

$$\mathrm{cogr}_\mu(G) = \limsup_{n\to\infty}[\sum_{\eta\in S_n(o)\cap[o]} \mu(T_\eta)]^{1/n} \tag{1.2}$$

and

$$\mathrm{gr}_\mu(T) = \limsup_{n\to\infty}[\sum_{\eta\in S_n(o)} \mu(T_\eta)]^{1/n}.$$

Note that $\mathrm{gr}_\mu(T) = 1$ and, in the regular case, $\mathrm{cogr}_\mu(G) = \mathrm{cogr}(G)/(d-1)$. We shall prove

Theorem 2. *Let G be a simple connected graph with bounded vertex degree for which the random walk on the covering tree T is transient. Then G is amenable if and only if* $\mathrm{cogr}_\mu(G) = 1$.

Finally, we consider measuring the size of the cogrowth set $[o]$ by how big its limit set in ∂T is. That is, let R be the set of rays in ∂T that hit $[o]$ infinitely often. As was proved for the regular case in [N3], we show:

Theorem 3. *If G is the cover of a finite graph then G is amenable if and only if* $\dim(R) = \dim(\partial T)$.

2 First situation

In this section, we shall prove Theorem 1. Given a graph G, which we assume is simple and connected, we additionally assume that it covers a finite graph G_0. That is, there exists a function $\theta_0 : G \to G_0$ such that θ_0 preserves adjacency and vertex degree (i.e., the vertex degree of x in G equals the vertex degree of $\theta_0(x)$ in G_0 and, if $x, y \in G$ are adjacent, then $\theta_0(x)$ and $\theta_0(y)$ are adjacent in G_0). Such a map is a discrete analog of a "local homeomorphism". In general, such a map is called a "cover" (of G_0 by G), and every graph is covered by a graph (if only by itself). The largest such cover of G is necessarily a tree, called the "universal covering tree", and denoted by T. Let $\theta : T \to G$ denote the covering map of T onto G. Given a vertex $x \in G$, let $[x] = \theta^{-1}(\theta(x))$; equivalently, $[x]$ is the equivalence class containing x with respect to the equivalence relation induced by θ. Since G covers a finite graph, it has a bounded vertex degree; let M denote an upper bound for the vertex degrees of G.

Let

$$K_u(x, y) = \sum_{n=0}^{\infty} |S_n(x) \cap [y]| u^n. \tag{2.1}$$

Since $K_u(x, y)$ can also be written as $\sum_{z\in[y]} u^{d(x,z)}$, it is clear that the convergence of K_u is independent of the choice of x and y. By (1.1), it is clear that K_u exists if $u < 1/\operatorname{cogr}(G)$, but diverges if $u > 1/\operatorname{cogr}(G)$. Even if K_u exists, the result when applied to a function need not. Consider K_u applied to a constant function:

$$K_u 1(x) = \sum_y K_u(x, y) = \sum_{n=0}^{\infty} |S_n(x)| u^n = \sum_z u^{d(x,z)}.$$

By the last equality, it is clear that the convergence of $K_u 1(x)$ is independent of x and, by the definition of $\operatorname{gr}(T)$, $K_u 1$ exists if $u < 1/\operatorname{gr}(T)$ and diverges if $u > 1/\operatorname{gr}(T)$. For convenience, let

$$u_0 = 1/\operatorname{gr}(T).$$

Then there is a gap between $\operatorname{cogr}(G)$ and $\operatorname{gr}(T)$ if and only if $K_{u_0+\epsilon}$ exists for some $\epsilon > 0$.

As a first step in studying the kernels K_u, we first find their inverses. A useful tool for this is the study of the "covering operators".

We say that a function $\hat{f} : T \to R$ covers $f : G \to R$ if

$$\hat{f} = f \circ \theta,$$

and we say that a kernel (i.e., generalized matrix) $\widehat{M} : T \to R$ covers the kernel $M : G \to R$ if

$$M(\theta(\xi), \theta(\eta)) = \sum_{\rho\in[\eta]} \widehat{M}(\xi, \rho).$$

An example of this last case is afforded by the "adjacency matrices" of T and G: for $x, y \in G$, let $A(x, y)$ be 1 or 0 according to whether x and y are adjacent or not. Similarly, let $\widehat{A}$ denote the adjacency matrix of T. Since θ preserves vertex degree, $\widehat{A}$ covers A.

Another such matrix is Q on G and $\widehat{Q}$ on T defined by

$$Qf(x) = (d(x) - 1) f(x).$$

Clearly, $\widehat{Q}$ covers Q.

It is easy to verify that the covering relation is preserved by matrix multiplication (i.e., $\widehat{MN} = \widehat{M}\widehat{N}$ by which we mean: if $\widehat{M}$ covers M and $\widehat{N}$ covers N, then $\widehat{M}\widehat{N}$ covers MN). Also, if $\widehat{f}$ covers f, then $\widehat{Mf} = \widehat{M}\widehat{f}$. If we define $\widehat{K}_u(x, y) = u^{d(x,y)}$, then $\widehat{K}_u$ covers K_u as defined by (2.1).

Lemma 1. $(I - uA + u^2 Q)K_u = K_u(I - uA + u^2 Q) = (1 - u^2)I$.

Proof. Note that

$$\begin{aligned}\widehat{A}\widehat{K}_u(\xi,\eta) &= \sum_{\rho\sim\xi} u^{d(\rho,\eta)} \\ &= u^{d(\xi,\eta)}[(d(\xi)-1)u + 1/u - (1/u-u)\hat{I}(\xi,\eta)] \\ &= \widehat{K}_u(\xi,\eta)[d(\xi)u + (1/u-u)(1-\hat{I}(\xi,\eta)],\end{aligned}$$

and so

$$\widehat{A}\widehat{K}_u = u\widehat{D}\widehat{K}_u + (1/u-u)(\widehat{K}_u - \hat{I}).$$

Hence

$$AK_u = uDK_u + (1/u-u)(K_u - I),$$

and so

$$(I - uA + u^2Q)K_u = (1-u^2)I.$$

The equality

$$K_u(I - uA + u^2Q) = (1-u^2)I$$

can be treated similarly or by using the facts that $\widehat{K}_u\widehat{A}$ and $\widehat{K}_u\widehat{Q}$ are the transposes of $\widehat{A}\widehat{K}_u$ and $\widehat{Q}\widehat{K}_u$ respectively. □

We define a generalized *Laplacian* by $\Delta_u \equiv I - uA + u^2Q$. This terminology is motivated by the fact that $\Delta_1 = D - A$ is equivalent (i.e., equal up to multiplication by a bounded function which is also bounded away from 0) to the usual Laplacian on graphs ($\Delta = D^{-1}A - I$). In general, Δ_u is equivalent to the Schrödinger operator $\Delta + q$, where $q(x) = u - u^2 - \frac{1-u^2}{d(x)}$ (which is constant when G is regular). The operator Δ_u has long appeared (though not with this notation) in the literature on zeta functions for graphs. For example, Bass [1] was the first to prove:

$$Z(u) \equiv \prod_C (1-u^{|C|})^{-1} = \frac{1}{(1-u^2)^r \det(\Delta_u)},$$

where Z, the zeta function of a finite graph, is the product over "prime" cycles C, and r is the Betti number of the graph. See also papers [9, 11, 6] for other proofs of this generalization of Ihara's theorem.

Lemma 1 then states that Δ_u is essentially the inverse of K_u. We say that a function is u-superharmonic if $\Delta_u f \geq 0$. As in the usual case, Harnack's inequality holds.

Lemma 2. *If f is non-negative and u-superharmonic for some $u > 0$, then there exists $C > 0$ such that $f(y) \leq Cf(x)$ for all pairs of adjacent vertices x, y.*

Proof.

$$f(y) \le \sum_{z\sim x} f(z) = Af(x)$$
$$\le (1+u^2q(x))f(x)/u \le f(x)(1+u^2(M-1))/u. \qquad \square$$

Lemma 3. *If $f > 0$ and $\Delta_u f \ge \lambda f$, then $\forall \sigma < \lambda : \exists \epsilon > 0 : \Delta_{u+\epsilon} f \ge (\lambda - \sigma) f$.*

Proof. By Harnack's inequality and bounded vertex degree, choose $\epsilon > 0$ such that $Af(x) \le \frac{\sigma}{\epsilon} f$. Then

$$-\epsilon Af + 2\epsilon u Qf + \epsilon^2 Qf \ge -\sigma f,$$

and so by hypothesis,

$$\Delta_{u+\epsilon} f = f - uAf + u^2 Qf - \epsilon Af + 2\epsilon u Qf + \epsilon^2 Qf \ge (\lambda - \sigma) f. \qquad \square$$

A necessary and sufficient condition for $K_{u_0+\epsilon}$ to exist (equivalently, for $\operatorname{cogr}(G) < \operatorname{gr}(T)$) follows.

Proposition 1. *Suppose G is a simple connected graph which covers a finite graph. Then $\Delta_{u_0} f \ge \lambda f$ for some positive function f and some $\lambda > 0$ if and only if $K_{u_0+\epsilon}$ exists for some $\epsilon > 0$.*

Proof. The idea of the proof here is that $K_{u_0+\epsilon}$ is an analogue of the resolvent kernel G^ϵ for the usual Laplacian and we merely follow the proof of the analogous theorem in the usual case. One difficulty, that arises here is that there is no "resolvent equation". However, it turns out that equation (2.2) below is sufficient for our purposes.

Suppose $f > 0$ and $\Delta_{u_0} f \ge \lambda f$ for some $\lambda > 0$. By Lemma 3, for sufficiently small ϵ, there exists $\sigma < \lambda$ such that $\Delta_{u_0+\epsilon} f \ge (\lambda - \sigma) f$. Then $\widehat{\Delta}_{u_0+\epsilon} \hat{f} \ge (\lambda - \sigma) \hat{f}$. On T, $\widehat{K}_{u_0+\epsilon}$ exists, and

$$[1 - (u_0+\epsilon)^2]\hat{f} = \widehat{K}_{u_0+\epsilon} \widehat{\Delta}_{u_0+\epsilon} \hat{f} \ge (\lambda - \sigma) \widehat{K}_{u_0+\epsilon} \hat{f},$$

from which it follows that

$$K_{u_0+\epsilon} f(x) \le \frac{1 - (u_0+\epsilon)^2}{\lambda - \sigma} f(x),$$

and therefore $K_{u_0+\epsilon}$ exists.

To prove the other way, we note that $K_u 1$ takes on only finitely many values (since K_u covers a corresponding kernel on a finite graph), and thus $K_u 1$ is bounded if $u < u_0$.

Suppose that, for some C,

$$\widehat{K}_{u_0} \widehat{K}_{u_0+\epsilon} \le C \widehat{K}_{u_0+\epsilon}. \tag{2.2}$$

Then $g(x) \equiv K_{u_0+\epsilon}(x, x_0)$ satisfies $K_{u_0} g \leq Cg$ and $g \geq 0$. Choose λ such that $(1 - u_0^2)g \geq \lambda K_{u_0} g$. By Lemma 1, $\Delta_{u_0} K_{u_0} g \geq \lambda K_{u_0} g$. Letting $f = K_{u_0} g$, we find $f > 0$ and $\Delta_{u_0} f \geq \lambda f$.

It remains to prove (2.2) under the hypothesis that $K_u 1(x)$ is bounded for $u < u_0$. Fix $\xi, \eta \in T$ and suppose $\gamma = (\xi = \gamma_0, \gamma_1, \ldots, \gamma_n = \eta)$ is the path connecting ξ to η in T. Define $T(i) = \{\rho : d(\rho, \gamma_i) = d(\rho, \gamma)\}$. For convenience, let $s = u_0$ and $t = u_0 + \epsilon$. Then

$$\begin{aligned}
\widehat{K}_{u_0} \widehat{K}_{u_0+\epsilon}(\xi, \eta) &= \sum_{\rho} s^{d(\xi,\rho)} t^{d(\rho,\eta)} \\
&= \sum_{i=0}^{n} \sum_{\rho \in T(i)} s^{d(\rho,\gamma_i)+i} t^{d(\rho,\gamma_i)+n-i} \\
&= t^n \sum_{i=0}^{n} (s/t)^i \sum_{\rho \in T(i)} (st)^{d(\rho,\gamma_i)} \\
&= t^n \sum_{i=0}^{n} (s/t)^i \sum_{k=0}^{\infty} (st)^k |S_k(\gamma_i) \cap T(i)| \\
&\leq t^n \sum_{i=0}^{n} (s/t)^i \sum_{k=0}^{\infty} (st)^k |S_k(\gamma_i)| \\
&\leq t^n \frac{1}{1 - \frac{s}{t}} \sup_{\xi} \widehat{K}_{st} 1(\xi).
\end{aligned}$$

The result follows, since $t^n = \widehat{K}_t(\xi, \eta)$ and $st < s = u_0$. □

Essential to the proof of Theorem 1 will be the fact that there exists a positive u_0-harmonic function on G. The proof of this, basically an application of the Perron–Frobenius theorem, appears in a paper on zeta functions on graphs by Kotani and Sunada [6].

Lemma 4. *There exists $h > 0$ such that $\Delta_{u_0} h = 0$ on G.*

Proof. Let G_0 be a finite graph covered by G, and let $\theta_0 : G \to G_0$ be the covering map. By theorem 1.6 of [6], there exists a positive valued function h_0 on G_0 such that $\Delta_{u_0} h_0 = 0$ (the α in [6] is our $1/\operatorname{gr}(T)$ where T is the covering tree of G_0 and thus of G). The "lift" $h \equiv h_0 \circ \theta_0$ is positive u_0-harmonic on G. □

We define the usual inner product on G:

$$\langle f, g \rangle = \sum_{x \in G} f(x) g(x).$$

Let $\mathcal{E} = \{[x, y] : x \sim y\}$ denote the set of directed edges in G, and for functions $u, v : \mathcal{E} \to R$, define

$$\langle u, v\rangle = \frac{1}{2}\sum_{[x,y]\in\mathcal{E}} u([x, y])v([x, y]).$$

We remark that Δ_u is self-adjoint since A is. We write $f \asymp g$ if and only if there exists $C > 0$ such that $\frac{1}{c} < f(x)/g(x) < C$ for all x.

The following proposition is a standard fact about self-adjoint operators and appears, in a slightly less general form, in [3].

Proposition 2. *Let $\lambda \geq 0$. Then there exists $h > 0$ such that $\Delta_u h \geq \lambda h$ if and only if*

$$\inf_f \langle f, \Delta_u f\rangle/\langle f, f\rangle \geq \lambda.$$

Proof. Suppose that $\Delta_{u_0} h \geq \lambda h$ for some $h > 0$. Define

$$\nabla f([x, y]) = \alpha(x, y)f(y) - \alpha(y, x)f(x),$$

where $\alpha(x, y) = \sqrt{u_0 h(x)/h(y)}$, and $[x, y]$ is a directed edge in G. Then, the usual inner product gives, for square summable f:

$$\begin{aligned}\langle \nabla f, \nabla g\rangle &= \frac{1}{2}\sum_{[x,y]}[\alpha(x, y)f(y) - \alpha(y, x)f(x)][\alpha(x, y)g(y) - \alpha(y, x)g(x)]\\ &= \sum_x f(x)g(x)\sum_{y\sim x}\alpha(y, x)^2 - \sum_x f(x)\sum_{y\sim x}\alpha(x, y)\alpha(y, x)g(y).\end{aligned}$$

Since $\sum_{y\sim x}\alpha(y, x)^2 \leq 1 - u_0^2 + u_0^2 d(x) - \lambda$,

$$0 \leq \langle \nabla f, \nabla f\rangle \leq \langle f, \Delta_{u_0} f\rangle - \lambda\langle f, f\rangle,$$

and therefore

$$\inf_f \langle f, \Delta_{u_0} f\rangle/\langle f, f\rangle \geq \lambda.$$

Let $K \subset G$ be finite, and define Δ_K by

$$\Delta_K f = \Delta_{u_0}(\chi_K f).$$

Furthermore, let $\langle u, v\rangle_K = \sum_{x\in K} u(x)v(x)$. It is then easy to see that

$$\langle f, \Delta_K g\rangle_K = \langle \chi_K f, \Delta_{u_0}(\chi_K g)\rangle_K,$$

and thus Δ_K is self-adjoint and finite dimensional. Let $\mathcal{C}(K)$ be the space of functions supported on K, and

$$\lambda_K = \inf_{f\in\mathcal{C}(K)} \frac{\langle f, \Delta_K f\rangle_K}{\langle f, f\rangle_K}.$$

Suppose $\inf_f \langle f, \Delta_u f\rangle/\langle f, f\rangle \geq \lambda$. Then $\lambda_K \geq \lambda \geq 0$ and Δ_K is positive. Since K is finite, there exists an eigenvector f such that $\Delta_K f = \lambda_K f$. We argue that f can

be assumed to be positive on K as follows. Let h be as in Lemma 4, and define ∇ as above using this h. Then

$$\begin{aligned}\langle f, \Delta_K f\rangle_K &= \langle \chi_K f, \Delta_{u_0}(\chi_K f)\rangle \\ &= \langle \nabla(\chi_K f), \nabla(\chi_K f)\rangle \\ &\geq \langle \nabla(\chi_K |f|), \nabla(\chi_K |f|)\rangle \\ &= \langle |f|, \Delta_K |f|\rangle_K,\end{aligned}$$

where equality holds if and only if f does not change sign. However, since $\langle |f|, |f|\rangle = \langle f, f\rangle$ and $\langle |f|, \Delta_K |f|\rangle_K \geq \lambda_K$, equality indeed holds.

Let $o \in K_1 \subset K_2 \subset \cdots$, where $\bigcup K_i = G$, and define h_n to be a positive solution on K_n of $\Delta_{K_n} h_n \geq \lambda h_n$ normalized so that $h_n(o) = 1$. By Lemma 1, there exists a pointwise convergent subsequence, and the pointwise limit, h, is positive and satisfies $\Delta_{u_0} h \geq \lambda h$. □

By combining Propositions 1 and 2, we see that there is a gap between $\operatorname{cogr}(G)$ and $\operatorname{gr}(T)$ if and only if

$$\inf_f \frac{\langle f, \Delta_{u_0} f\rangle}{\langle f, f\rangle} > 0.$$

Theorem 1 then follows if we show the equivalence of this condition with a similar condition equivalent to the non-amenability of G (see [3]), namely,

$$\inf_f \frac{\langle f, \Delta_1 f\rangle}{\langle f, f\rangle} > 0.$$

With h as in Lemma 4, we indeed have this equivalence since, as in the proof of Proposition 2,

$$\begin{aligned}\langle f, \Delta_{u_0} f\rangle &= \frac{1}{2}\sum_{[x,y]} \left(\sqrt{u_o \frac{h(x)}{h(y)}} f(y) - \sqrt{u_o \frac{h(y)}{h(x)}} f(x)\right)^2 \\ &= \frac{u_0}{2}\sum_{[x,y]} h(x)h(y)\left(\frac{f(x)}{h(x)} - \frac{f(y)}{h(y)}\right)^2 \\ &\asymp \sum_{[x,y]} \left(\frac{f(x)}{h(x)} - \frac{f(y)}{h(y)}\right)^2 \\ &= \left\langle \frac{f}{h}, \Delta_1 \frac{f}{h}\right\rangle,\end{aligned}$$

and therefore Theorem 1 is proven.

Example 1. We give an example (actually a class of examples) of an amenable graph with $\operatorname{cogr}(G) < \operatorname{gr}(T)$.

Let G_0 be a non-amenable regular graph which is not a tree. Its cover is a regular tree, call it T_d. Attach an infinite chain to a vertex $o \in G_0$; call the resulting graph

G. Then G is amenable. Its cover, T, is T_d with an infinite chain attached to each point in $[o]$. Then $|S_n(o) \cap [o]|$ is the same in T and T_d, and so $\mathrm{cogr}(G_0) = \mathrm{cogr}(G)$. However, $|S_n(o)|$ is bigger in T than in T_d, so $\mathrm{gr}(G_0) \leq \mathrm{gr}(G)$. Hence, since G_0 is non-amenable and regular, $\mathrm{cogr}(G) = \mathrm{cogr}(G_0) < \mathrm{gr}(G_0) \leq \mathrm{gr}(G)$.

Remark 1. We conjecture that if G has bounded vertex degree and $\mathrm{cogr}(G) = \mathrm{gr}(T)$, then G is amenable. A counterexample would, of course, not be a cover of a finite graph and not be regular. Furthermore, from Theorem 2, $\mathrm{cogr}_\mu(G) < 1$.

3 Another cogrowth constant

For this, we assume that the universal covering tree T of G is transient (that is the random walk on T is transient). This is not a very strong condition since it is satisfied by most graphs G; for example, any G which contains two or more cycles or on which the random walk is transient. The only graphs for which this fails are recurrent graphs with at most one cycle. We shall not consider these graphs here.

Fix $o \in T$, and let ∂T denote the set of all geodesic rays starting at o. Given $\eta \in T$, let T_η denote the set of all paths in ∂T which go through the vertex η. This set is called a "cone", and the set of all cones forms a topology base on ∂T. The simple random walk X_n on T, starting at o, converges a.s., in this topology, to a point X_∞ in ∂T. The distribution of X_∞ is called harmonic measure, and we write:

$$\mu(E) = P_o(X_\infty \in E).$$

Recall the resolvent kernels for the Laplacian $\Delta = D^{-1}A - I$ on T, denoted G_T^ϵ, are defined by

$$G_T^\epsilon(\xi, \eta) = \sum_{n=0}^{\infty} P_\xi(X_n = \eta)/(1-\epsilon)^{n+1}$$

and satisfy

$$(\Delta + \epsilon I)G_T^\epsilon = -I.$$

Then, the resolvent kernels cover the resolvent kernel G^ϵ on G. We base our proof of Theorem 2 on the fact that G^ϵ exists for some $\epsilon > 0$ if and only if G is not amenable (for example, see [8], and references therein).

As a first step, we prove that Green's function G_T (=G_T^0) on T and harmonic measure of cones are comparable.

Lemma 5. *For $\eta \in [o]$, $G_T(o, \eta) \asymp \mu(T_\eta)$.*

Proof. By the Markov property (twice),

$$\mu(T_\eta) = P_o(X_\infty \in T_\eta) = P_o(\exists n : X_n = \eta) P_\eta(X_\infty \in T_\eta) = G_T(o, \eta) \frac{P_\eta(X_\infty \in T_\eta)}{G_T(\eta, \eta)}.$$

The denominator of the fraction above is constant for $\eta \in [o]$ while the numerator takes on at most $d(\eta)$ values. □

Proof of Theorem 2. Suppose that G is not amenable (i.e., G^ϵ exists for some $\epsilon > 0$). Let $c = 1/(1-\epsilon)$. Note that

$$\sum_n p_T^{(n)}(o,\eta)c^{n+1} = \sum_{n\geq|\eta|} p_T^{(n)}(o,\eta)c^{n+1} \geq \sum_{n\geq|\eta|} p_T^{(n)}(o,\eta)c^{|\eta|+1} = G_T(o,\eta)c^{|\eta|+1},$$

and so, by Lemma 5,

$$\begin{aligned}\sum_n \Big(\sum_{\eta\in S_n\cap[o]} \mu(T_\eta)\Big)c^n &\asymp \sum_n \Big(\sum_{\eta\in S_n\cap[o]} G_T(o,\eta)\Big)c^n \\ &\leq \sum_{\eta\in[o]}\sum_n p^{(n)}(o,\eta)c^{n+1} \\ &= \sum_{\eta\in[o]} G_T^\epsilon(o,\eta) \\ &= G^\epsilon(o,o) < \infty,\end{aligned}$$

and thus, by (1.2), $\mathrm{cogr}_\mu(G) < 1$.

Conversely, suppose $\mathrm{cogr}_\mu(G) < 1$. Then there exists some $c > 1$ such that

$$\sum_n \sum_{\eta\in S_n\cap[o]} \mu(T_\eta)c^n < \infty,$$

which, by Lemma 5, implies

$$\sum_{\eta\in[o]} G_T(o,\eta)c^{|\eta|} < \infty.$$

Choose $\alpha \in (0,1)$ such that $G_T(o,\eta) \geq k_1 c^{-|\eta|/(1-\alpha)}$ for all $\eta \in [o]$. Then $G_T(o,\eta)^{1-\alpha} \geq k_1^{1-\alpha} c^{-|\eta|}$, and so

$$G_T(o,\eta)^\alpha \leq k_1^{\alpha-1} c^{|\eta|} G_T(o,\eta).$$

Now, choose $\epsilon > 0$ such that both G_T^ϵ and $G_T^{\epsilon'}$ exist (and are bounded), where $\epsilon' = 1-(1-\epsilon)^{1/(1-\alpha)}$. By Hölder's inequality,

$$G_T^\epsilon(o,\eta) \leq G_T(o,\eta)^\alpha G_T^{\epsilon'}(o,\eta)^{1-\alpha},$$

and so $G_T^\epsilon(o,\eta) \leq k c^{|\eta|} G_T(o,\eta)$ for some k and all $\eta \in [o]$. Therefore, there exists some $\epsilon > 0$ such that

$$G^\epsilon(o,o) = \sum_{\eta\in[o]} G_T^\epsilon(o,\eta) \leq k \sum_{\eta\in[o]} G_T(o,\eta)c^{|\eta|} < \infty.$$

□

4 The radial limit set and its Hausdorff dimension

We shall now show that a cover of a finite graph is amenable if and only if the radial limit set of $[o]$ has the highest possible Hausdorff dimension. Most of the terminology below appears in the seminal paper by Lyons [7]. The proof is based on the proof of the analogous fact for regular graphs which appeared in [10].

Proof of Theorem 3. We note that since G is a cover of a finite graph, T is also, and so it is "quasispherical". A consequence is that $\lim_{n\to\infty} |S_n|^{\frac{1}{n}}$ exists.

Fix k, and let $R' = \{\gamma \in \partial T : \gamma_{nk} \in [o] \text{ for any } n\}$. Then $R' = \partial T'$, where T' is periodic and generated by a finite tree, namely T_o, the tree of radius k in T with $\partial T_o = S_k \cap [o]$. Then, where $\mathrm{br}(T)$ denotes the branching number of T (see [7]),

$$\begin{aligned}\dim(R') &= \log(\mathrm{br}(T')) = \log(\mathrm{gr}(T')) \\ &= \liminf_{n\to\infty} \frac{1}{n} \log|S'_n| = \liminf_{n\to\infty} \frac{1}{nk} \log|S'_{nk}| \\ &\geq \frac{1}{k} \log|S'_k| = \log(|S_k \cap [o]|)^{\frac{1}{k}},\end{aligned}$$

and thus, since $R' \subset R \subset \partial T$,

$$|S_k \cap [o]|^{\frac{1}{k}} \leq e^{\dim(R')} \leq e^{\dim(R)} \leq e^{\dim(\partial T)} = \mathrm{br}(T) \leq \mathrm{gr}(T) = \lim_{n\to\infty} |S_n|^{\frac{1}{n}}.$$

Therefore, by Theorem 1, if G is amenable, then $\dim(R) = \dim(\partial T)$.

Conversely, suppose $\dim(R) = \delta \equiv \dim(\partial T)$. If there exists $\alpha < \delta$ such that $\sum_{\eta\in[o]} e^{-\alpha|\eta|} < \infty$, then

$$\inf_F \sum_{\eta\in[o]-F} e^{-\alpha|\eta|} = 0,$$

where $F \subset [o]$ is finite, and, since R is the radial limit set of any set of the form $[o] - F$, $\dim(R) \leq \alpha < \delta$. Hence, if $\dim(R) = \delta$, then, for $\alpha < \delta$,

$$\sum_n e^{-\alpha n} |S_n \cap [o]| = \sum_{\eta\in[o]} e^{-\alpha|\eta|} = \infty,$$

and thus $\limsup_{n\to\infty} |S_n \cap [0]|^{\frac{1}{n}} \geq e^{\alpha}$ for all $\alpha < \delta$. Therefore, by (1.2),

$$\mathrm{cogr}(G) \geq e^{\dim(\partial T)} = \mathrm{br}(T) = \mathrm{gr}(T) = \liminf_{n\to\infty} |S_n|^{\frac{1}{n}},$$

and, by Theorem 1, G is amenable. □

References

[1] H. Bass, The Ihara–Selberg zeta function of a tree lattice, Internat. J. Math. 3 (1992), 717–797.

[2] J. Conklin, The discrete Laplacian, applications to random walk, and inverse problems on weighted graphs, Ph.D. Thesis, University of Rochester, 1988.

[3] J. Dodziuk, Difference equations, isoperimetric inequality and transience of certain random walks, Trans. Amer. Math. Soc. 284 (1984), 787–794.

[4] J. Dodziuk and L. Karp, Spectral and function theory for combinatorial Laplacians, in: Geometry of Random Motion (Ithaca, N.Y., 1987), Contemp. Math. 73, Amer. Math. Soc., Providence, RI, 1988, 25–40.

[5] R. I. Grigorchuk, Symmetrical random walks on discrete groups, in: Multicomponent Random Systems, Adv. Probab. Related Topics 6, Dekker, New York 1980, 285–325.

[6] M. Kotani and T. Sunada, Zeta functions of finite graphs, J. Math. Sci. Univ. Tokyo 7 (2000), 7–25.

[7] R. Lyons, Random walks and percolation on trees, Ann. Probab. 18 (1990), 931–958.

[8] S. Northshield, Cogrowth of regular graphs, Proc. Amer. Math. Soc. 116 (1992), 203–205.

[9] S. Northshield, Several proofs of Ihara's theorem, preprint 1459, IMA, University of Minnesota (1997).

[10] S. Northshield, A note on recurrence, amenability, and the universal cover of graphs, in: Random Discrete Structures (Minneapolis, MN, 1993), IMA Vol. Math. Appl. 76, Springer-Verlag, New York 1996, 199–206.

[11] H. M. Stark and A. A. Terras, Zeta functions of finite graphs and coverings, Adv. Math. 121 (1996), 124–165.

Sam Northshield, Department of Mathematics, SUNY, Plattsburgh, NY 12901, USA
E-mail: northssw@plattsburgh.edu

Total variation lower bounds for finite Markov chains: Wilson's lemma

*Laurent Saloff-Coste**

Abstract. Using results and ideas due to Persi Diaconis and to David Wilson, we discuss lower bounds for convergence in total variation of ergodic finite Markov chains. These lower bounds are based on eigenvalues and eigenfunctions.

1 Introduction

There is a large body of literature concerning quantitative estimates on the convergence of finite ergodic Markov chains. See, e.g., [1, 3, 5, 9, 10, 19, 21] and the references therein. In particular, a number of different techniques such as coupling and eigenvalue estimates are available to obtain upper bounds on the time needed to reach approximate stationarity. In this expository article based on works of Diaconis [5] and Wilson [22], we will focus on lower bounds. Note that good lower bounds are required to prove the cut-off phenomenon discussed in [2, 6, 19, 22].

Lower bounds are obtained by direct computations using a "test function" or a "test set" and, in principle, are easier than upper bounds. In practice, this is only partially true. First, finding a good "test function" is more an art than a science. Second, even after a good guess, some clever computations might be required. The goal of a good lower bound is to show that approximate stationarity has not been reached yet. This necessarily involves understanding "something" concerning the behavior of the studied Markov chain before it reaches stationarity, an often difficult task.

We now introduce some notation. Let X be a finite state space. Let K be a Markov kernel with invariant probability measure π. For $n \geq 2$, set

$$K^n(x, y) = \sum_{z \in X} K^{n-1}(x, z) K(z, y), \quad K_x^n(y) = K^n(x, y).$$

The chain K is irreducible if for each x, y there exists $n = n(x, y)$ such that $K^n(x, y) > 0$. An irreducible chain is aperiodic if there exists n such that $K^n(x, y) > 0$ for all x, y. Irreducible aperiodic chains are ergodic, i.e., satisfy $K_x^n \to \pi$ as n tends to infinity

*Research partially supported by NSF grant DMS 0102126

for some probability measure π. The measure π can be characterized as the unique invariant distribution, i.e., $\sum_x \pi(x)K(x,y) = \pi(y)$. For any probability measure μ on X, set

$$\mu(f) = \sum_x f(x)\mu(x), \quad \mathrm{Var}_\mu(f) = \sum_x |f(x) - \mu(f)|^2 \mu(x).$$

For any (signed) measure ν on X, set

$$\|\nu\|_{\mathrm{TV}} = \sup_{A \subset X} |\nu(A)|.$$

We will be mostly interested in the total variation distance $\|K_x^n - \pi\|_{\mathrm{TV}}$ between K_x^n and π. It is well-known and easy to see that

$$\|K_x^n - \pi\|_{\mathrm{TV}} = \frac{1}{2}\sum_y |K_x^n(y) - \pi(y)| = \frac{1}{2}\sum_y |[K_x^n(y)/\pi(y)] - 1|\pi(y).$$

We also set $\|f\|_2^2 = \sum |f(x)|^2\pi(x)$, $\|f\|_\infty = \max_X |f|$.

2 Lower bound in $\ell^2(\pi)$

In this section, we assume that (K, π) is reversible, i.e., satisfies

$$K(x,y)\pi(x) = K(y,x)\pi(y).$$

If we look at K as an operator acting on $\ell^2(\pi)$ by $Kf(x) = \sum_y K(x,y)f(y)$, we see that K is a self-adjoint contraction. Thus K has real eigenvalues bounded by 1 in absolute value (1 is always an eigenvalue). Let us enumerate the eigenvalues of K in non-increasing order starting from $\beta_0 = 1$ so that

$$\beta_0 \geq \beta_1 \geq \beta_2 \geq \cdots \geq \beta_{N-1} \geq -1,$$

where $N = \#X$. The chain is ergodic if and only if $1 > \beta_1$ and $\beta_{N-1} > -1$. Let ϕ_i, $0 \leq i \leq N-1$, be the associated eigenfunctions which we choose to be real valued and normalized by $\|\phi_i\|_2 = 1$ with $\phi_0 \equiv 1$.

Let $k_x^n(y) = K^n(x,y)/\pi(y)$ be the density of K_x^n w.r.t. π. Then (e.g., [19])

$$\|k_x^n - 1\|_2^2 = \sum_1^{N-1} \beta_i^{2n} |\phi_i(x)|^2.$$

Since this is a sum of non-negative terms, we can get a lower bound on $\|k_x^n - 1\|_2^2$ by keeping only the terms with $\beta_i = \beta_1$. That is,

$$\|k_x^n - 1\|_2^2 \geq \beta_1^{2n} \sum_{i:\beta_i = \beta_1} |\phi_i(x)|^2. \tag{2.1}$$

To analyse this, we use the following construction. Consider the vector space $V \subset \ell^2(\pi)$ spanned by the ϕ_i such that $\beta_i = \beta_1$. Fix x and consider the linear map $f \mapsto f(x)$ from V to $\mathbb{R}$. The kernel V_x of this map is of codimension 0 or 1. If it is of codimension 0, we are rather unlucky since our lower bound (2.1) is trivial. If the codimension is 1 then we can pick $\phi_1 = \psi_x$ to be the unique normalized function orthogonal to the kernel V_x, and the lower bound (2.1) can be written as

$$\|k_x^n - 1\|_2^2 \geq \beta_1^{2n} |\psi_x(x)|^2. \tag{2.2}$$

Thus, an obvious way to get a lower bound on $\|k_x^n - \pi\|_2$ is to focus on β_1 and find an associated normalized eigenfunction which is large at x. It is useful to note that we do not really need to work with β_1 in the argument above. Any eigenvalue could be used instead and it is possible that using a different eigenvalue would yield a better lower bound, at least in a certain range of values of n.

Before we look at a simple example, observe that Jensen's inequality gives

$$\|K_x^n - \pi\|_{\text{TV}} = \frac{1}{2} \sum_y |[k_x^n(y) - 1|\pi(y) \leq \frac{1}{2} \|k_x^n - 1\|_2.$$

Hence, as far as lower bounds are concerned, working with $\|K_x^n - \pi\|_{\text{TV}}$ is more demanding than working with $\|k_x^n - 1\|_2$. This will be clearly illustrated below.

Example 2.1 (The hypercube). Let $X = \{0, 1\}^d$ equipped with addition mod 2. Set $e_0 = (0, \ldots, 0)$ and let e_i be the binary vector of length d with a unique 1 in position i. Define a Markov kernel K by setting $K(x, y) = 0$ except if $y = x + e_i$ with $i = 0, \ldots, d$, in which case $K(x, x + e_i) = 1/(d + 1)$. This defines the simple random walk on the d-dimensional hypercube. It is not hard to see that the centred and normalized coordinate functions $f_i(x) = 1 - 2x_i$ are orthogonal eigenfunctions, all with the same eigenvalue $\beta = 1 - 2/(d + 1)$ (we write β to emphasize that we do not know, *a priori*, that this is the second largest eigenvalue). Inequality (2.1) gives

$$\|k_x^n - 1\|_2^2 \geq d \left(1 - \frac{2}{d+1}\right)^{2n}.$$

This shows that $\|k_x^n - 1\|_2$ is not small if n is less than $\frac{d}{4} \log d$, which is exactly the right order of magnitude. See [7, 5]. To see what (2.2) corresponds to, let $|x|$ be the number of 1's in x. For $x = 0$, one finds that the function ψ_0 is given by

$$\psi_0(x) = \frac{1}{\sqrt{d}} \sum_1^d f_i(x) = \frac{1}{\sqrt{d}} (d - 2|x|).$$

Indeed, $\psi_0(0) = \sqrt{d}$ is the maximum possible value at 0 for a normalized eigenfunction associated to $\beta = 1 - 2/(d + 1)$.

3 Lower bounds in total variation

This section presents a basic idea that has been used many times by Persi Diaconis and his collaborators in the study of finite Markov chains. See [5, pp. 29, 44]. Here, we do not assume that (K, π) is reversible. To bound $\|K_x^n - \pi\|_{\text{TV}}$ from below it suffices to find a good test set A on which $K_x^n(A) - \pi(A)$ is not small. More precisely, we are interested in showing that when n is not too large $K_x^n(A) - \pi(A)$ is close to it maximal possible value 1.

Any set A can be represented as $A = \{x : |\psi(x)| > s\}$ for some function ψ and it will be convenient to think of the set A in this form with $s > 0$ a parameter to be chosen later. Assuming that ψ satisfies $\pi(\psi) = 0$, we have

$$\pi(A) = \pi(|\psi| > s) \le \frac{\text{Var}_\pi(\psi)}{s^2}.$$

Next, assume that the mean of ψ under K_x^n is not too small, namely, assume that n is such that

$$|K_x^n(\psi)| \ge 2s. \tag{3.1}$$

Then

$$K_x^n(A) = 1 - K_x^n(A^c) \ge 1 - \frac{\text{Var}_{K_x^n}(\psi)}{s^2}$$

because $A^c = \{|\psi| \le s\} \subset \{|\psi - K_x^n(\psi)| \ge s\}$. Hence, under these circumstances,

$$\|K_x^n - \pi\|_{\text{TV}} \ge K_x^n(A) - \pi(A) \ge 1 - \frac{\text{Var}_\pi(\psi)}{s^2} - \frac{\text{Var}_{K_x^n}(\psi)}{s^2}.$$

In view of Section 2, a natural idea is to use an eigenfunction of K as the function ψ above. Thus, assume that we have at our disposal an eigenfunction ψ of K with associated eigenvalue β. We do not assume that ψ, β are real but we do assume that $|\beta| < 1$ and that ψ is normalized in $\ell_2(\pi)$, i.e., $\pi(|\psi|^2) = 1$. Observe that since K preserves the orthogonal complement of the constant functions in $\ell^2(\pi)$, we have $\pi(\psi) = 0$. Moreover, $K_x^n(\psi) = \beta^n \psi(x)$, and (3.1) becomes

$$|\beta|^n |\psi(x)| \ge 2s. \tag{3.2}$$

Thus, if we can bound $\text{Var}_{K_x^n}(\psi)$ by

$$\text{Var}_{K_x^n}(\psi) \le B(x)^2 \tag{3.3}$$

independently of n, we get that

$$\|K_x^n - \pi\|_{\text{TV}} \ge K_x^n(A) - \pi(A) \ge 1 - \tau$$

for all n such that

$$n \le \frac{-1}{2\log|\beta|} \log\left(\frac{\tau|\psi(x)|^2}{4(1 + B(x)^2)}\right). \tag{3.4}$$

This technique was developed by Diaconis [5] in the context of examples for which one can compute explicitly

$$\mathrm{Var}_{K_x^n}(\psi) = K_x^n(|\psi|^2) - |K_x^n(\psi)|^2$$

by expanding $|\psi|^2$ along the $\ell^2(\pi)$-orthonormal basis of the eigenfunctions $(\phi_i)_0^{N-1}$.

Example 3.1 (The hypercube). We keep the notation introduced at the end of Section 2. Then

$$\psi(x) = \psi_0(x) = \frac{1}{\sqrt{d}}(d - 2|x|)$$

is an eigenfunction with eigenvalue $1 - 2/(d+1)$. To make the computation easier, it is useful to write

$$\psi(x) = \frac{1}{\sqrt{d}} \sum_1^d (-1)^{x_i}$$

(this is the natural form of ψ from the viewpoint of representation theory). Then

$$\psi(x)^2 = \frac{1}{d}\left(d + 2 \sum_{1 \le i < j \le d} (-1)^{x_i + x_j}\right).$$

As $(-1)^{x_i+x_j}$ is an eigenfunction with eigenvalue $1 - 4/(d+1)$, this gives

$$K_0^n(\psi^2) = 1 + (d-1)\left(1 - \frac{4}{d+1}\right)^n$$

and

$$\mathrm{Var}_{K_0^n}(\psi) = 1 + (d-1)\left(1 - \frac{4}{d+1}\right)^n - d\left(1 - \frac{2}{d+1}\right)^{2n}.$$

Hence $\mathrm{Var}_{K_0^n}(\psi) \le 1$. Taking $B(x) = 1$ in (3.3) and using (3.4) gives

$$K_0^n(A) - \pi(A) \ge 1 - \tau$$

for all n such that

$$n \le \frac{-1}{2\log(1 - \frac{2}{d+1})} \log\left(\frac{d\tau}{8}\right).$$

Asymptotically, this says that $\|K_x^n - \pi\|_{\mathrm{TV}}$ is not small if n is less than $\frac{1}{4} d \log d$ which is known to be the correct cut-off time for this example. See [5, p. 29].

4 Wilson's lemma

Wilson's lemma [22, Lemma 5] gives a bound on the variance $\mathrm{Var}_{K_x^n}(\psi)$ for an eigenfunction ψ. Hence, together with the ideas presented in Section 3, it offers a way to bound total variation from below. Before presenting Wilson's result, we find it useful to discuss the same computation in the context of diffusion semigroups, e.g., Brownian motion on a compact manifold.

4.1 Continuous computations

Recall the following old idea (see, e.g., [4, 16]). Consider a semigroup of operators $(H_t)_{t>0}$. Then $H_t[|f|^2]-|H_t f|^2$ is the difference of the values of $s \mapsto H_s\big[|H_{t-s} f|^2\big]$ at $s=0$ and $s=t$. Hence, we can write

$$H_t\big[|f|^2\big] - |H_t f|^2 = \int_0^t \partial_s\big(H_s\big[|H_{t-s} f|^2\big]\big)ds. \tag{4.1}$$

Assume now that H_t is the heat diffusion semigroup of a compact manifold M. The main point in making this assumption is that the infinitesimal generator $-L$ is then a differential second order operator for which the chain rule applies in the form

$$LG(f) = G'(f)Lf - G''(f)|\nabla f|^2,$$

where f is a smooth real function on M, $G : \mathbb{R} \to \mathbb{R}$, and $|\nabla f|^2$ is the square of length of the gradient of f. In fact the special case of this formula corresponding to $G(t) = t^2$ provides an intrinsic definition of the length of the gradient in terms of the operator L (e.g., see [4, 16] and the references therein), namely,

$$2|\nabla f|^2 = 2fLf - Lf^2. \tag{4.2}$$

Under these circumstances, the derivative inside the integral in (4.1) is easy to compute. Setting $H_{t-s} f = F_s$ for ease, we have

$$\begin{aligned} \partial_s\big(H_s[H_{t-s} f]^2\big) &= -LH_s[F_s]^2 + 2H_s[F_s LF_s] \\ &= -H_s L[F_s]^2 + 2H_s[F_s LF_s] \\ &= 2H_s|\nabla F_s|^2 = 2H_s|\nabla H_{t-s} f|^2. \end{aligned} \tag{4.3}$$

The third line follows from (4.2). Now,

$$H_t[f^2] - [H_t f]^2 = 2\int_0^t H_s|\nabla H_{t-s} f|^2 ds.$$

If $H_t(x, dy) = H_t^x(dy)$ denotes the transition kernel of the Markov semigroup H_t, then the left-hand side, evaluated at a point x, is exactly the variance of f under the probability measure H_t^x. When $f = \phi$ is a real eigenfunction of L with real eigenvalue

$\lambda > 0$ ($L\phi = \lambda\phi$), we have $H_{t-s}\phi = e^{-(t-s)\lambda}\phi$, and

$$H_t[\phi^2] - [H_t\phi]^2 = 2\int_0^t e^{-2(t-s)\lambda} H_s|\nabla\phi|^2 ds \leq \frac{\|\nabla\phi\|_\infty^2}{\lambda}.$$

Thus $\mathrm{Var}_{H_t^x}(\phi) \leq \|\nabla\phi\|_\infty^2/\lambda$, which is a version of Wilson inequality in [22, Lemma 5]. See [20] for a complete treatment and applications in the diffusion setting.

4.2 Discrete computations

Replacing derivatives by difference operators can turn straightforward computations into messy ones. In such cases, it is often a good idea to perform the computations involving differences by following closely the analogous computations involving derivatives. For a given finite reversible Markov chain (K, π) consider the discrete Laplacian $L = I - K$ and set

$$|\nabla f(x)|^2 = \frac{1}{2}\sum_y |f(x) - f(y)|^2 K(x, y).$$

Then we have

$$\langle (I - K)f, f\rangle_\pi = \sum_x (f(x) - Kf(x))f(x)\pi(x) = \sum_x |\nabla f(x)|^2\pi(x).$$

To connect with Wilson's notation in [22, Lemma 5], observe that if we let ξ_n be the position of our Markov chain at time n, then

$$2|\nabla f(x)|^2 = \mathbb{E}(|f(\xi_{n+1}) - f(\xi_n)|^2/\xi_n = x)$$

(which is of course independent of n). Thus, in Wilson's notation from [22, Lemma 5], R can be taken to be $2\|\nabla\phi\|_\infty^2$.

Lemma 4.1. *For any finite reversible Markov chain, and any real function f, we have*

$$Lf^2 = 2fLf - 2|\nabla f|^2 \tag{4.4}$$

and

$$f^2 - (Kf)^2 = 2f(I - K)f - [(I - K)f]^2 = 2fLf - (Lf)^2. \tag{4.5}$$

Proof. For (4.4) (the analog of (4.2)), write

$$\begin{aligned} Lf^2(x) = (I - K)f^2(x) &= \sum_y K(x, y)[f^2(x) - f^2(y)] \\ &= \sum_y K(x, y)[f(x) + f(y)][f(x) - f(y)] \\ &= \sum_y K(x, y)[2f(x) + f(y) - f(x)][f(x) - f(y)] = \end{aligned}$$

$$= 2f(x)(I-K)f(x) - \sum_y |f(x)-f(y)|^2 K(x,y)$$
$$= 2f(x)Lf(x) - 2|\nabla f(x)|^2.$$

For (4.5) which is analogous to $\partial_t (H_t f)^2 = 2H_t f \partial_t H_t f$, write

$$f^2 - (Kf)^2 = (f+Kf)(f-Kf)$$
$$= (2f + Kf - f)(f-Kf) = 2f(I-K)f - [(I-K)f]^2.$$

This finishes the proof of Lemma 4.1. □

Using these tools we can express the variance of a function under K_x^n.

Lemma 4.2. *For any finite reversible Markov chain, and any real function f, we have*

$$\mathrm{Var}_{K_x^n}(f) = K_x^n(f^2) - (K_x^n(f))^2$$
$$= 2\sum_0^{n-1} K^\ell \big[|\nabla K^{n-\ell-1} f|^2 - \frac{1}{2}((I-K)K^{n-\ell-1}f)^2 \big](x).$$

In particular,

$$\mathrm{Var}_{K_x^n}(f) = K_x^n(f^2) - (K_x^n(f))^2 \le 2\sum_0^{n-1} K^\ell [|\nabla K^{n-\ell-1} f|^2].$$

Proof. For ease, set $K^{n-\ell-1} f = F_\ell$. Then, using (4.4) and (4.5), write

$$K^n(f^2) - (K^n f)^2 = \sum_0^{n-1} K^{\ell+1}(K^{n-\ell-1}f)^2 - K^\ell (K^{n-\ell} f)^2$$
$$= \sum_0^{n-1} K^\ell (K-I)F_\ell^2 + K^\ell (F_\ell^2 - (KF_\ell)^2)$$
$$= \sum_0^{n-1} K^\ell \big(2F_\ell (K-I)F_\ell + 2|\nabla F_\ell|^2 - 2F_\ell (K-I)F_\ell - ((I-K)F_\ell)^2 \big)$$
$$= 2\sum_0^{n-1} K^\ell \big(|\nabla F_\ell|^2 - \frac{1}{2}((I-K)F_\ell)^2 \big).$$

This proves the desired equality. □

We now specialize to the case of an eigenfunction ϕ.

Lemma 4.3. *Let (K, π) be a reversible finite Markov chain. Assume ϕ is a real eigenfunction of K with eigenvalue β and set $\lambda = 1 - \beta$. Then*

$$\mathrm{Var}_{K_x^n}(\phi) \le \frac{2\|\nabla\phi\|_\infty^2}{\lambda(2-\lambda)}$$

(the statement is empty if $\beta = 1$ or $\beta = -1$).

Proof. As ϕ is an eigenfunction, we have $|\nabla K^{n-\ell-1}\phi|^2 = \beta^{2(n-\ell-1)}|\nabla\phi|^2$. Thus, by Lemma 4.2,

$$\mathrm{Var}_{K_x^n}(\phi) \le 2\sum_0^{n-1} \beta^{2(n-\ell-1)}\|\nabla\phi\|_\infty^2.$$

Here we used that K^ℓ contracts ℓ^∞. This gives

$$\mathrm{Var}_{K_x^n}(\phi) \le 2\frac{1-\beta^{2n}}{1-\beta^2}\|\nabla\phi\|_\infty^2 \le \frac{2\|\nabla\phi\|_\infty^2}{\lambda(2-\lambda)},$$

which is the desired bound. □

Remark 4.4. Lemma 4.3 is the main part of what we refer to as *Wilson's lemma* in the title. See also Theorem 4.7 below. The statement in Lemma 4.3 is slightly different from the statement obtained by Wilson in [22, Lemma 5], although, for all practical purposes, the differences are irrelevant. Wilson's statement in the present notation is that

$$\mathrm{Var}_{K_x^n}(\phi) \le \frac{\|\nabla\phi\|_\infty^2}{\lambda}$$

assuming $0 < \lambda \le 2 - \sqrt{2}$. The differences are due purely to the different treatment of "undesirable " terms that show up in "discrete time". In the diffusion setting or in the continuous time finite Markov chain setting (see below), Wilson's computation and the present variation give precisely the same result.

Remark 4.5. Let us also comment on the denominator $\lambda(2-\lambda) = 1 - \beta^2$. This is 0 if and only if either $\beta = 1$ or $\beta = -1$. Of course, if $\beta = 1$ and the chain is *irreducible* then $\mathrm{Var}_{K_x^n}(\phi) = 0$ since ϕ is constant. However, the proof above does not use irreducibility. In the reducible case one can of course have $\mathrm{Var}_{K_x^n}(\phi) \ne 0$ and $\nabla\phi = 0$. Dividing by $(2-\lambda)$ seems to be more of an accident.

The following is stated mostly for curiosity although it will be useful to obtain cleaner statements later on.

Lemma 4.6. *Assume ϕ is a normalized eigenfunction of K with eigenvalue $\beta < 1$ and set $\lambda = 1 - \beta$. Then $\lambda^{-1}\|\nabla\phi\|_\infty^2 \ge 1$.*

Proof. Since $(I - K)\phi = \lambda\phi$, we have

$$1 = \sum_x |\phi(x)|^2\pi(x) = \frac{1}{\lambda}\langle (I-K)\phi, \phi\rangle_\pi = \frac{1}{\lambda}\sum_x |\nabla\phi(x)|^2\pi(x) \le \frac{\|\nabla\phi\|_\infty^2}{\lambda}. \quad \square$$

From this lemma it follows that, under the hypothesis of Lemma 4.3 and assuming that ϕ is normalized, we have

$$1 + \mathrm{Var}_{K_x^n}(\phi) \le \frac{(4-\lambda)\|\nabla\phi\|_\infty^2}{\lambda(2-\lambda)}.$$

The next statement combines Diaconis' lower bound technique of Section 3 and Wilson's variance bound to get a lower bound in total variation. Observe that when applying this result we do not need to care about the normalization of ϕ.

Theorem 4.7 (Wilson [22, Lemma 5]). *Let (K, π) be a finite reversible Markov chain. Let $\beta \in (-1, 1)$ be an eigenvalue of K with associated real eigenfunction ϕ, and set $\lambda = 1 - \beta$. Then*

$$\|K_x^n - \pi\|_{\mathrm{TV}} \ge 1 - \tau \tag{4.6}$$

for all n such that

$$n \le \frac{-1}{2\log(1-\lambda)} \log\left(\frac{\tau(2-\lambda)\lambda\phi(x)^2}{4(4-\lambda)\|\nabla\phi\|_\infty^2}\right).$$

In particular, if $\beta \in (0, 1)$ then (4.6) *holds for all*

$$n \le \frac{-1}{2\log(1-\lambda)} \log\left(\frac{\tau\lambda\phi(x)^2}{12\|\nabla\phi\|_\infty^2}\right).$$

Example 4.8 (The hypercube). The eigenfunction (not normalized) $\phi(x) = d - 2|x| = \sum_1^d (1 - 2x_i)$ has eigenvalue $\beta = 1 - 2/(d+1)$ and satisfies

$$\|\phi\|_\infty^2 = d^2, \qquad \|\nabla\phi\|_\infty^2 = \frac{d}{d+1}.$$

Thus $\|K_x^n - \pi\|_{TV} \ge 1 - \tau$ for all n such that

$$n \le \frac{-1}{2\log(1 - \frac{2}{d+1})} \log\left(\frac{\tau d}{12}\right).$$

Observe that this is essentially the same result as obtained earlier in Section 3 by a more detailed variance computation and that it is sharp since the walk on the hypercube has a cut-off at time $\frac{1}{4}d\log d$, see [5, 6].

4.3 Non-reversible chains

Here we discuss some generalizations of Theorem 4.7. First we address the question of reversibility (Wilson does not assume reversibility in his paper). In fact, every single

line in Sections 2.1, 2.2 above is correct assuming only that π is an invariant probability measure of K *and that all functions and eigenvalues appearing are real*. In Wilson's work, this is implicitly assumed. Somewhat interestingly, things also work for possibly complex eigenfunctions and eigenvalues. Thus let K be an irreducible aperiodic chain (this is not really needed, but otherwise the result is not very interesting!) with stationary probability π. Keep the definitions

$$|\nabla f(x)|^2 = \frac{1}{2}\sum_y |f(x) - f(y)|^2 K(x, y), \qquad L = I - K.$$

Let $\Re(z)$ be the real part of $z \in \mathbb{C}$, and set $\langle f, g\rangle_\pi = \sum f(x)\overline{g(x)}\pi(x)$.

Lemma 4.9. *Let K be a finite Markov chain with invariant probability measure π. Then, for any complex valued function f, we have*

$$\Re\left(\langle (I - K)f, f\rangle_\pi\right) = \sum_x |\nabla f(x)|^2 \pi(x)\ , \tag{4.7}$$

$$L|f|^2 = 2\Re\left(fL\overline{f}\right) - 2|\nabla f|^2, \tag{4.8}$$

and

$$|f|^2 - |Kf|^2 = 2\Re\left(fL\overline{f}\right) - |Lf|^2. \tag{4.9}$$

Proof. For (4.7), see [19]. For (4.8), write

$$\begin{aligned} L|f|^2(x) &= (I - K)|f|^2(x) = \sum_y K(x, y)[|f|^2(x) - |f|^2(y)] \\ &= \sum_y K(x, y)\Re([f(x) + f(y)][\overline{f}(x) - \overline{f}(y)]) \\ &= \Re\Big(\sum_y K(x, y)[2f(x) + f(y) - f(x)][\overline{f}(x) - \overline{f}(y)]\Big) \\ &= 2\Re(f(x)L\overline{f}(x)) - 2|\nabla f(x)|^2. \end{aligned}$$

The proof of (4.9) is similar. □

Lemma 4.10. *Let K be a finite Markov chain with invariant probability measure π. Then, for any complex valued function f, we have*

$$\begin{aligned} \mathrm{Var}_{K_x^n}(f) &= K_x^n(|f|^2) - |K_x^n(f)|^2 \\ &= 2\sum_0^{n-1} K^\ell\Big[|\nabla K^{n-\ell-1} f|^2 - \frac{1}{2}|(I - K)K^{n-\ell-1} f|^2\Big](x). \end{aligned}$$

Proof. Proceed as in the proof of Lemma 4.2, putting modulus signs in various places and using (4.8)–(4.9) instead of (4.4)–(4.5). □

Now, we easily obtain the following version of Lemma 4.3.

Lemma 4.11. *Assume ϕ is an eigenfunction (not necessarily real) of K with eigenvalue (not necessarily real) β. Then*

$$\mathrm{Var}_{K_x^n}(\phi) \le \frac{2\|\nabla\phi\|_\infty^2}{1-|\beta|^2}.$$

Lemma 4.12. *Assume ϕ is a normalized eigenfunction (not necessarily real) of K with eigenvalue (not necessarily real) β, $\Re(\beta) < 1$. Then $(1-\Re(\beta))^{-1}\|\nabla\phi\|_\infty^2 \ge 1$.*

Proof. Since $(I-K)\phi = (1-\beta)\phi$, we have

$$(1-\beta)\|\phi\|_2^2 = \langle (I-K)\phi, \phi\rangle_\pi = \Re(\langle (I-K)\phi, \phi\rangle_\pi) + \frac{1}{2}(\overline{\beta}-\beta)\|\phi\|_2^2.$$

As $\|\phi\|_2 = 1$, we obtain $1-\Re(\beta) = \Re\left(\langle (I-K)\phi, \phi\rangle_\pi\right)$, and (4.7) implies $1-\Re(\beta) \le \|\nabla\phi\|_\infty^2$ as desired. □

Finally, we obtain the following variation on Theorem 4.7.

Theorem 4.13. *Let K be a finite Markov chain with stationary measure π. Let β be an eigenvalue of K with associated eigenfunction ϕ (possibly complex). Then*

$$\|K_x^n - \pi\|_{\mathrm{TV}} \ge 1-\tau$$

for all n such that

$$n \le \frac{-1}{2\log|\beta|}\log\left(\frac{\tau(1-|\beta|^2)|\phi(x)|^2}{4(2+|\beta|)\|\nabla\phi\|_\infty^2}\right).$$

We have not used the fact that the state space is finite, and this theorem holds true for ergodic Markov chains on countable state spaces as long as $\|\nabla\phi\|_\infty$ is finite.

4.4 Continuous time

Consider a countable state space X, and let Q be matrix indexed by $X \times X$ and such that

$$\forall\, x \in X,\ 0 \le Q(x,x) < +\infty,\ \ \forall\, x \ne y,\ \ Q(x,y) \ge 0,\ \ \forall\, x \in X,\ \sum_y Q(x,y) = 0.$$

We assume that Q is irreducible, non-explosive (see [17, §2.7]) and admits an invariant probability measure π, i.e., a probability measure such that

$$\forall\, y \in X, \qquad \sum_x \pi(x)Q(x,y) = 0.$$

Consider the minimal non-negative Markovian semigroup $H_t = e^{-tL}$ associated with the infinitesimal generator

$$-Lf = Qf = \sum_y Q(\cdot, y) f(y)$$

defined originally on finitely supported function. This operator may or may not have $\ell^2(\pi)$ eigenvalues but its $\ell^2(\pi)$ spectrum is contained in the half plane $\Re(z) \geq 0$ (because H_t is a contraction on $\ell^2(\pi)$). Let H_t^x be the distribution of the associated Markov process at time $t > 0$, starting at x. Set

$$|\nabla f(x)|^2 = \frac{1}{2} \sum_y |f(x) - f(y)|^2 Q(x, y).$$

Theorem 4.14. *Assume that λ is an eigenvalue of L with associated eigenfunction ϕ (possibly complex). Assume that $\|\nabla \phi\|_\infty < +\infty$. Then*

$$\| H_t^x - \pi \|_{\text{TV}} \geq 1 - \tau$$

for all t such that

$$t \leq \frac{1}{2\Re(\lambda)} \log \left(\frac{\tau \Re(\lambda) |\phi(x)|^2}{8 \|\nabla \phi\|_\infty^2} \right).$$

Proof. Write

$$H_t(|\phi|^2) - |H_t \phi|^2 = \int_0^t \partial_t [H_s(|H_{t-s}\phi|^2)] ds.$$

As in (4.8), we have $L|F|^2(x) = 2\Re\left(F(x) L\overline{F}(x)\right) - 2|\nabla F(x)|^2$ for any function F. Hence, computing as in (4.3) gives $\partial_t [H_s(|H_{t-s}\phi|^2)] = 2H_s |\nabla H_{t-s}\phi|^2$. As ϕ is an eigenfunction, we have $|\nabla H_{t-s}\phi|^2 = e^{-2\Re(\lambda)(t-s)} |\nabla \phi|^2$. Thus

$$\text{Var}_{H_t^x}(\phi) = H_t(|\phi|^2) - |H_t \phi|^2(x) \leq \frac{\|\nabla \phi\|_\infty^2}{\Re(\lambda)}.$$

Repeating Diaconis' argument of Section 3 in this context yields the announced result. □

5 Two examples on the symmetric group

This section illustrates Theorem 4.7 by looking at random transposition and random adjacent transposition on the symmetric group. Random adjacent transposition is one of the examples treated in [22]. Durrett [12] uses Wilson's technique of Theorem 4.7 to study further examples on the symmetric group that are related to problems in genetics. In unpublished work [23], Wilson has applied an interesting variant of his

technique to the following shuffling mechanism: move the top card to the bottom or second to bottom. This was proposed long ago by Rudvalis as possibly the slowest shuffle, see [5, p. 90]. Hildebrand proved in his thesis [15] by a coupling argument that order $n^3 \log n$ such shuffles suffice to mix up a deck of n cards. Wilson shows that order $n^3 \log n$ such shuffles are also necessary. Other examples analysed by Wilson in [23] are various versions of the shuffle that either transpose the top two cards or move the top card to the bottom (see [9]). In several of these examples the eigenvalues and eigenfunctions are complex.

Example 5.1 (Random transposition). Let $X = S_n$ be the symmetric group on n letters. Let $\tau_{i,j}$ be the element in S_n transposing i and j, $1 \le i < j < n$. Set

$$K(x, x\tau_{i,j}) = \frac{2}{n^2}, \qquad K(x, x) = \frac{1}{n},$$

and $K(x, y) = 0$ if $y \neq x$ and $y \neq x\tau_{i,j}$ for some $i < j$. If we interpret this chain as a shuffling mechanism, it corresponds to the following scheme: picture the deck of cards neatly displayed in a row on a table, face down. At each step, let the left and right hands each pick a card independently uniformly at random. If both hands pick the same card, do nothing. Otherwise, switch the two cards. This is an ergodic chain which is reversible with respect to the uniform distribution. An analysis of this chain can be found in [5, 11].

Consider the function

$$\phi = \#\text{ fixed points} - 1.$$

We claim this is an eigenfunction with eigenvalue

$$\beta = 1 - 2/n.$$

This is well known (e.g., see [5, 11]) and can be seen as follows. Consider the (disjoint) cycle structure of a permutation x. Assume x has k fixed points, ℓ two cycles (i.e., transpositions), etc. Then there are $k(k-1)/2$ transpositions that will decrease the number of fixed points by 2, ℓ transpositions that will increase the number of fixed points by 2, $k(n-k)$ transpositions that will decrease the number of fixed points by 1, $n - k - 2\ell$ transpositions that will increase the number of fixed points by 1. This gives

$$\begin{aligned} K\phi(x) &= (k-1) + \frac{2}{n^2}\left(-2\frac{k(k-1)}{2} + 2\ell - k(n-k) + (n-k-2\ell)\right) \\ &= \left(1 - \frac{n}{2}\right)(k-1) = \left(1 - \frac{n}{2}\right)\phi(x). \end{aligned}$$

Obviously

$$\|\phi\|_\infty^2 = (n-1)^2, \quad \|\nabla\phi\|_\infty^2 \le \frac{2(n-1)}{n}.$$

Thus

$$\frac{(1-\beta)\|\phi\|_\infty^2}{\|\nabla\phi\|_\infty^2} \geq n-1.$$

Using this and $\beta = 1-2/n$ in Theorem 4.7 gives

$$\|K_x^j - \pi\|_{\text{TV}} \geq 1-\tau$$

for all

$$j \leq \frac{-1}{2\log(1-2/n)} \log \frac{\tau(n-1)}{12}.$$

For fixed τ, the right hand side is asymptotic to $\frac{1}{4}n\log n$ which is off by a factor of $1/2$ since this chain has a cut-off at time $\frac{1}{2}n\log n$. This is one case where a direct computation of the variance does help. See [5, 6]. A similar analysis works for transpose top and random showing that at least $\frac{1}{2}n\log n$ steps are needed to reach approximate stationarity in this case. As for random transposition, this is off by a factor of $1/2$ since transpose random and top has a cut-off at time $n\log n$.

Example 5.2 (Random adjacent transposition). Let $X = S_n$ be the symmetric group on n letters as above. Set

$$K(x, x\tau_{i,i+1}) = \frac{1}{n}, \qquad K(x,x) = \frac{1}{n}, \quad 1 \leq i < n,$$

and $K(x,y) = 0$ if $y \neq x$ and $y \neq x\tau_{i,i+1}$ for some i. Thus, this chain moves by picking an adjacent transposition uniformly at random. The limit distribution is uniform.

If we restrict attention to the moves of the Ace of Spade (or any other fixed card), we see that it performs a random walk on the integer segment $\{1,\ldots,n\}$ with holding $1/2$ at the extremity, and where moves are occurring at random (geometric) times with rate $2/n$. Ignoring the random holding times, the eigenfunctions of such a random walk on $\{1,\ldots,n\}$ are known. For instance,

$$v(j) = \cos\left(\frac{\pi(j-1/2)}{n}\right)$$

is an eigenfunction [13, p. 436] with eigenvalue $\cos\pi/n$. Let $1 \leq \ell \leq n$ denote the values of the cards. Let $\ell(x)$ be the position of card ℓ in the permutation x. Consider the function

$$v_\ell(x) = v \circ \ell(x).$$

Let us compute Kv_ℓ. It is not hard to check that

$$Kv_\ell = \left(\frac{n-2}{n} + \frac{2}{n}\cos\frac{\pi}{n}\right) v_\ell = \left(1 - \frac{2}{n}\left(1-\cos\frac{\pi}{n}\right)\right) v_\ell.$$

Since this holds for all ℓ, we get n eigenfunctions, all associated with the same eigenvalue

$$\beta = 1 - \frac{2}{n}\left(1 - \cos\frac{\pi}{n}\right) \le 1 - \frac{\pi^2}{n^3}.$$

Of course, any sum of these eigenfunctions is another eigenfunction as long as it is not identically zero. Set $a_\ell = v_\ell(\ell) = \cos(\pi(\ell - 1/2)/n)$ and

$$\phi = \sum_{\ell=1}^{n} a_\ell v_\ell.$$

What we need is a lower bound on $\|\phi\|_\infty$ and an upper bound on $\|\nabla\phi\|_\infty$. By construction, the function ϕ attains its maximum at the identity e, and we have

$$\begin{aligned}\phi(e) &= \sum_1^n a_\ell^2 = \sum_1^n |\cos\frac{\pi(\ell - 1/2)}{n}|^2 \\ &= n\left(\int_0^1 |\cos\pi x|^2 dx + o(1)\right) = n(1 + o(1)).\end{aligned}$$

Next, we estimate the gradient. First observe that

$$|v_\ell(x\tau_{i,i+1}) - v_\ell(x)| = \begin{cases} 0 & \text{if } \ell \notin x^{-1}(\{i, i+1\}), \\ \left|\cos\frac{\pi(i+1/2)}{n} - \cos\frac{\pi(i-1/2)}{n}\right| & \text{if } \ell \in x^{-1}(\{i, i+1\}). \end{cases}$$

As

$$\cos\frac{\pi(i + 1/2)}{n} - \cos\frac{\pi(i - 1/2)}{n} = -2\sin\frac{\pi i}{n}\ \sin\frac{\pi}{2n},$$

we obtain

$$|v_\ell(x\tau_{i,i+1}) - v_\ell(x)| \le \frac{\pi}{n},$$

and

$$\begin{aligned}|\nabla\phi(x)|^2 &= \frac{1}{2n}\sum_{i=1}^n |\sum_{\ell=1}^n a_\ell(v_\ell(x\tau_{i,i+1}) - v_\ell(x))|^2 \\ &= \frac{1}{2n}\sum_{i=1}^n |\sum_{\ell\in x^{-1}(\{i,i+1\})} a_\ell(v_\ell(x\tau_{i,i+1}) - v_\ell(x))|^2 \\ &\le \frac{2\pi^2}{n^3}\sum_{i=1}^n a_i^2 = \frac{2\pi^2}{n^3}\phi(e).\end{aligned}$$

Hence,

$$\frac{(1-\beta)\|\phi\|_\infty^2}{\|\nabla\phi\|_\infty^2} \ge \frac{\phi(e)}{2} = \frac{n}{2}(1 + o(1)).$$

This yields

$$\|K_x^j - \pi\|_{\text{TV}} \geq 1 - \tau$$

for

$$j \leq (1 - o(1))\frac{n^3}{\pi^2}\log n.$$

It is known that order $n^3 \log n$ random adjacent transpositions suffice to mix up a deck of n cards. See [8, 22].

References

[1] D. Aldous, Random walks on finite groups and rapidly mixing Markov chains, in: Séminaire de Probabilités, XVII, Lecture Notes in Math. 986, Springer-Verlag, Berlin 1983, 243–297.

[2] D. Aldous and P. Diaconis, Strong uniform times and finite random walks, Adv. in Appl. Math. 8 (1987), 69–97.

[3] D. Aldous and J. Fill, preliminary version of a book on finite Markov chains available electronically at http://www.stat.berkeley.edu/users/aldous (2000).

[4] D. Bakry, L'hypercontractivité et son utilisation en théorie des semigroupes, in: Lectures on Probability Theory (Saint-Flour, 1992), Lecture Notes in Math. 1581, Springer-Verlag, Berlin 1994, 1–114.

[5] P. Diaconis, Group Representations in Probability and Statistics, Institute of Mathematical Statistics Lecture Notes–Monograph Series, 11, Institute of Mathematical Statistics, Hayward CA 1988.

[6] P. Diaconis, The cutoff phenomenon in finite Markov chains, Proc. Nat. Acad. Sci. U.S.A. 93 (1996), 1659–1664.

[7] P. Diaconis, R. Graham and J. Morrison, Asymptotic analysis of a random walk on a hypercube with many dimensions, Random Structures Algorithms 1 (1990), 51–72.

[8] P. Diaconis and L. Saloff-Coste, Comparison techniques for random walk on finite groups, Ann. Probab. 21 (1993), 2131–2156.

[9] P. Diaconis and L. Saloff-Coste, Random walks on finite groups: a survey of analytic techniques, in: Probability Measures on Groups and Related Structures, XI (Oberwolfach, 1994), World Sci. Publishing, River Edge, NJ, 1995, 44–75.

[10] P. Diaconis and L. Saloff-Coste, What do we know about the Metropolis algorithm? J. Comput. System Sci. 57 (1998), 20–36.

[11] P. Diaconis and M. Shahshahani, Generating a random permutation with random transpositions, Z. Wahrsch. Verw. Gebiete 57 (1981), 159–179.

[12] R. Durrett, Shuffling chromosomes, J. Theor. Probab. 16 (2003), 725–750.

[13] W. Feller, An Introduction to Probability Theory and its Applications. Vol. I. Third edition, Wiley, New York 1968.

[14] M. Fukushima, Y. Oshima and M. Takeda, Dirichlet forms and Symmetric Markov processes, de Gruyter Stud. Math. 19, Walter de Gruyter, Berlin 1994.

[15] M. Hildebrand, Rates of convergence of some random processes on finite groups, Ph.D. thesis, Department of Mathematics, Harvard University, 1990.

[16] M. Ledoux, The geometry of Markov diffusion generators, Annales Fac. Sci. Toulouse (6) 9 (2000), 305–366.

[17] J. Norris, Markov Chains, Cambridge Series in Statistical and Probabilistic Mathematics 2, Cambridge University Press, Cambridge 1997.

[18] L. Saloff-Coste, Precise estimates on the rate at which certain diffusions tend to equilibrium, Math. Z. 217 (1994), 641–677.

[19] L. Saloff-Coste, Lectures on finite Markov chains, in: Lectures on Probability Theory and Statistics (Saint-Flour, 1996), Lecture Notes in Math. 1665, Springer-Verlag, Berlin 1997, 301–413.

[20] L. Saloff-Coste, On the convergence to equilibrium for Brownian motion on compact simple Lie groups, preprint (2002).

[21] A. Sinclair, Algorithms for Random Generation and Counting. A Markov Chain Approach, Progr. Theoret. Comput. Sci., Birkhäuser, Boston, MA, 1993.

[22] D. Wilson, Mixing times of lozenge tiling and card shuffling Markov chains, Ann. Appl. Probab. 14 (2004), 274–325.

[23] D. Wilson, Mixing time of the Rudvalis shuffle, Electron. Comm. Probab. 8 (2003), 77–85.

Laurent Saloff-Coste, Department of Mathematics, Cornell University, Ithaca, NY 14853, USA
E-mail: lsc@math.cornell.edu