The concentration of measure and the concentration of distance phenomena

By

Mahefa Ratsisetraina Ravelonanosy
(Ravelonanosy_Mahefa@student.ceu.edu)

24th May 2020

A THESIS PRESENTED TO CENTRAL EUROPEAN UNIVERSITY (CEU) IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF A MASTER OF SCIENCE IN MATHEMATICS AND ITS APPLICATIONS

Declaration

I, the undersigned, hereby declare that this thesis is my original work and it has not been submitted before for any other degree, part of degree or examination at Central European University Budapest or at any other university.

24th May 2020

Mahefa Ratsisetraina Ravelonanosy

Supervisor: Prof. Gabor Pete (Alfréd Rényi Institute of Mathematics and Budapest University of Technology and Economics)

Acknowledgements

I would like to express my sincere gratitude to my supervisor Prof. Gabor Pete for his motivations and insightful guidance during the completion of this research.

My sincere thanks also goes to Gergley Ódor for his assistance on the '*expander*'-part of this thesis.

I would like to offer my special thanks to Prof. Karoly Boroczky, Prof. Pal Hegedus, Elvira Kadvany, Melinda Balazs and all the CEU family for giving me the opportunity to attend the CEU Master's Program.

I would like to say thanks to all my teachers who shared their knowledges at the University of Atananarivo. Also I warmly thank my best friend who helped me too much during my undergraduate study.

Last but not the least, I would like to thank my parents, my sisters and my little brother for their spiritual support.

Abstract

In this thesis, we prove that the distance is not concentrated for the 2-dimensional integer lattice $\{o, 1, ..., n\}^2$, the two dimensional integer torus $\mathbb{Z} \times \mathbb{Z}/(n\mathbb{Z} \times n\mathbb{Z})$ and the Uniform Spanning Tree (UST) of the complete graph K_n. On the other hand, we prove that the distance is concentrated for the Hypercube $\{0, 1\}^n$, the Euclidean space $\mathbb{R}^n$ with the n-dimensional standard Gaussian measure, the unit sphere $\mathcal{S}^{n-1} \subseteq \mathbb{R}^n$ with the normalized Lebesgue measure and the ball of radius R of a non-elementary Hyperbolic group. To our knowledge, this last example has not been discussed in the literature and it is the main novel part of this work. By plotting the Kernel Density Estimations (KDE) in Python, we confirm that the distance is not concentrated for the UST of the complete graph K_n; and we only visualize that the distance is not concentrated for the UST of both the 2-dimensional integer lattice $\{0, 1, ..., n\}^2$ and the 5-dimensional integer lattice $\{0, 1, ..., n\}^5$. We present an application of the concentration of distance phenomenon for a transitive metric probability space.

Contents

1. Introduction

The concentration of measure phenomenon is a notion introduced by Vitali Milman in 1970 in his work in asymptotic geometry of Banach spaces, following the earlier works of Paul Lévy. It has many applications in various areas such as geometry, discrete mathematics and complexity theory, functional analysis and especially probability theory. It is often the behaviour of a function that depends on a large number of variables.

In Probability Theory, the theory of Large Deviations formalizes the concentration of measure phenomenon, and one of its probabilistic description is given by the strong law of large numbers (SLLN). For example, given a sequence of independent identically distributed random variables $X_1, ..., X_n, ... \sim \text{Unif}\{-1, 1\}$ and consider the function:

$$f_n(X_1, ..., X_n) := \frac{X_1 + ... + X_n}{n} \quad (n \in \mathbb{N}),$$

then the SLLN says that:

$$f_n \longrightarrow \mathbb{E}X_1 = 0 \text{ a.s},$$

meaning that when n is "large", the function f_n is highly concentrated around $\mathbb{E}X_1 = 0$. By using the Moment Generating Function method, this particular example of the concentration phenomenon can be quantified by the following concentration inequality:

$$\mathbb{P}(|f_n| \geq t) \leq 2e^{-\frac{nt^2}{2}}.$$

Here, our function f_n is the arithmetic mean of the variables $X_1, X_2, ...$ and X_n, but the point is that the phenomenon also holds for much more general functions.

The main task in this thesis is to investigate the concentration of the distance between two independent random points or the distance between a chosen 'root' and a random point for the following metric probability spaces: the Euclidean space $\mathbb{R}^n$ endowed with the n-dimensional standard Gaussian measure, the unit sphere $\mathcal{S}^{n-1} \subseteq \mathbb{R}^n$ endowed with the normalized Lebesgue measure, the 2-dimensional integer lattice $\{o, 1, ..., n\}^2$, the Hypercube $\{0, 1\}^n$, the two dimensional integer torus $\mathbb{Z} \times \mathbb{Z}/(n\mathbb{Z} \times n\mathbb{Z})$, the Uniform Spanning Tree (UST) of the complete graph K_n and the ball of radius R of a non-elementary Hyperbolic group. All those graphs are equipped with the graph distance and the normalized counting measure, and we conduct the study on a fixed Cayley graph for the Hyperbolic group.

In Chapter 2, we study the classical case of the Euclidean space $(\mathbb{R}^n, ||.||, \gamma_n)_n$, where $\gamma_n := \mathcal{N}(0_n, I_n)$ is the multivariate standard normal distribution. By combining this classical case with Haar's Theorem, see Lubotzky (2010), we prove the concentration phenomenon on the unit sphere $S^{n-1} \subseteq \mathbb{R}^n$ endowed with the Euclidean distance and the normalized Lebesgue measure.

Chapter 3 is concerned with discrete structures: we prove that in contrast with the case of the Hypercube $\{0, 1\}^n$, there is no distance concentration on the square lattice $\{0, ..., n\}^2$ as well as in the integer torus $\mathbb{Z}^2/(n\mathbb{Z})^2$.

We prove that the distance between two randomly chosen vertices x and y of the Uniform Spanning Tree $\mathcal{T}$ of K_n is not concentrated. We explain in detail the Wilson's algorithm approach for the limiting distribution of the quantity $\frac{\mathrm{d}_{\mathcal{T}}(x,y)}{\sqrt{n}}$ mentioned in Peres and Revelle (2004).

We prove the concentration of distance on transitive expanders in two different methods: the first proof follows from the spectral characterisation of expanders and the second one is derived from the combinatorial definition. Although this result should not be surprising to experts, we had thought it would have been written up for the first time in this thesis. Then Gergley Ódor drew our attention to this recent paper Roughgarden et al. (2019) of George Barmpalias, Neng Huang, Andrew Lewis-Pye, Angsheng Li, Xuechen Li, Ycheng Pan and Tim Roughgarden, which aims to characterise Finite Expected Degree (FED) *idemetric networks* and expanders are among discussed topic.

We close Chapter 3 by proving the main novel part of this work, which is the distance concentration on balls $\mathcal{B}_R$ of radius R in a non-elementary Hyperbolic group Γ. We first prove some elementary results about the balls and spheres of a d-regular tree in Theorem 3.4.3 and Theorem 3.4.4. After that, we prove Theorem 3.4.8 which says that:

$$\frac{\mathrm{d}(X,Y)}{2R} \xrightarrow{\mathbb{P}} 1,$$

where $X, Y \sim \mathrm{Unif}(\mathcal{B}_R)$ are independent, by combining the Hyperbolicity with the fact that non-elementary Hyperbolic groups have exponential growth. We close Chapter 3 with some further discussions and some open questions.

In Chapter 4, we visualize with Python the Rayleigh limit of the Uniform Spanning Tree of K_n. We also plot the Kernel Density Estimations (KDE) of the distance for the Uniform Spanning Trees of $(\mathbb{Z}/n\mathbb{Z})^5$ and $(\mathbb{Z}/n\mathbb{Z})^2$ for $n \in \{3, 7, 5\}$ and $n \in \{20, 30, 50\}$ respectively. Those other visualizations are motivated by Kenyon (2000b), Kenyon (2000a) and Lawler et al. (2011). Inspired from Roughgarden et al. (2019), we explain and visualize by simulating $\mathrm{S}_{10}(\mathbb{T}_3)$ how one can optimize space when saving a data structure that encodes the distance between all pairs of points of X_n for a large n, where (X_n, d_n, μ_n) is a transitive metric measurable space satisfying the distance concentration property.

2. Spherical measure and standard Gaussian distribution

In this chapter, we recall some properties of the n-dimensional Gaussian distribution on the Euclidean space $\mathbb{R}^n$ and we prove the corresponding concentration inequality. The concentration phenomenon on the unit sphere $\mathbb{S}^{n-1} \subseteq \mathbb{R}^n$ then follows by using Haar's Theorem on the unicity of a rotational invariant measure on the unit sphere.

2.1 Concentration of Gaussian measure

In this section, we state some properties of standard Gaussian random variables and introduce the notion of concentration of measure&distance phenomenon by considering the case of the metric measurable space $(\mathbb{R}^n, ||.||, \gamma_n)$, where $||.||$ is the usual Euclidean norm.

For $n > 1$, let γ_n denotes the n-dimensional standard Gaussian measure $\mathcal{N}(0_n, I_n)$, where 0_n is the zero-vector of $\mathbb{R}^n$ and I_n the $n \times n$ identity matrix. Let $X_1, ..., X_n, ... \sim \mathcal{N}(0,1)$ be independent identically distributed (iid) real random variables. Then $X := (X_1, ..., X_n) \sim \gamma_n$ and by the strong law of large numbers, we have:

$$\frac{||X||}{\sqrt{n}} = 1 + o(1) \;\; a.s,$$

meaning that when n is large, the Euclidean space $(\mathbb{R}^n, ||.||)$ endowed with γ_n "looks like" the n-dimensional sphere of radius $\sqrt{n}$. In order to quantify this property, we need the following lemma.

Lemma 2.1.1. For $t < \frac{1}{2}$, we have

$$m_{X_1^2}(t) := \mathbb{E}\left(\mathrm{e}^{tX_1^2}\right) = \frac{1}{\sqrt{1-2t}}.$$

Consequently,

$$\kappa_{X_1^2}(t) := \ln\left(\mathbb{E}\left(\mathrm{e}^{tX_1^2}\right)\right) = -\frac{1}{2}\ln(1-2t).$$

Proof. Since $t < \frac{1}{2}$, then $1 - 2t > 0$. So we have:

$$\begin{aligned}
\mathbb{E}(\mathrm{e}^{tX_1^2}) &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{tx^2} \mathrm{e}^{-\frac{x^2}{2}} dx = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{x^2(t-\frac{1}{2})} dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(x\sqrt{1-2t})^2}{2}} dx = \frac{1}{\sqrt{1-2t}} \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-\frac{u^2}{2}}}{\sqrt{2\pi}} du \\
&= \frac{1}{\sqrt{1-2t}}.
\end{aligned}$$

□

Given a real valued random variable χ, the function $t \longmapsto m_\chi(t) := \mathbb{E}\left(\mathrm{e}^{t\chi}\right)$ (when it exists) is called the moment generating function of χ.

The following theorem quantifies how much does the norm of an n-dimensional Gaussian vector deviate from $\sqrt{n}$.

Theorem 2.1.2. *Let $X \sim \gamma_n$ be an n-dimensional standard Gaussian vector of $\mathbb{R}^n$. Then for $0 < \epsilon < 1$, we have:*

$$\gamma_n(|||X|| - \sqrt{n}| \geq \epsilon\sqrt{n}) \leq 2\mathrm{e}^{-\frac{n\epsilon^2}{2}}.$$

Therefore

$$\frac{||X||}{\sqrt{n}} \xrightarrow{\mathbb{P}} 1.$$

Proof. We have:

$$\begin{aligned}\gamma_n\left(|||X|| - \sqrt{n}| \geq \epsilon\sqrt{n}\right) \leq\ & \gamma_n\left(||X|| \geq \sqrt{n}(1+\epsilon)\right) + \gamma_n\left(||X|| \leq \sqrt{n}(1-\epsilon)\right) \\ & \leq \gamma_n\left(||X||^2 \geq n(1+\epsilon)^2\right) + \gamma_n\left(||X||^2 \leq n(1-\epsilon)^2\right),\end{aligned}$$

since $0 < \epsilon < 1$, so that $1 - \epsilon > 0$. Hence:

$$\gamma_n\left(|||X|| - \sqrt{n}| \geq \epsilon\sqrt{n}\right) \leq \mathbb{P}\left(X_1^2 + ... + X_n^2 \geq n(1+\epsilon)^2\right) + \mathbb{P}\left(X_1^2 + ... + X_n^2 \leq n(1-\epsilon)^2\right),$$

where $X_1, ..., X_n \sim \mathcal{N}(0,1)$ are iid. We are going to give an upper bound of the two terms of the quantity in the right hand side of the inequality above by using the moment generating function.

- *For the first term:* let $t \in (0, \frac{1}{2})$. We have by Markov's inequality:

$$\begin{aligned}\mathbb{P}\left(X_1^2 + ... + X_n^2 \geq n(1+\epsilon)^2\right) &\leq \mathbb{P}\left(tX_1^2 + ... + tX_n^2 \geq tn(1+\epsilon)^2\right) \\ &\leq \mathbb{P}\left(\mathrm{e}^{tX_1^2 + ... + tX_n^2} \geq \mathrm{e}^{tn(1+\epsilon)^2}\right) \\ &\leq \frac{\mathbb{E}\left(\mathrm{e}^{tX_1^2 + ... + tX_n^2}\right)}{\mathrm{e}^{tn(1+\epsilon)^2}} = \frac{\mathbb{E}\left(\mathrm{e}^{tX_1^2}...\mathrm{e}^{tX_n^2}\right)}{\mathrm{e}^{tn(1+\epsilon)^2}} \\ &= \left(\frac{\mathbb{E}\left(\mathrm{e}^{tX_1^2}\right)}{\mathrm{e}^{t(1+\epsilon)^2}}\right)^n = \left(\frac{m_{X_1^2}(t)}{\mathrm{e}^{t(1+\epsilon)^2}}\right)^n\end{aligned}$$

since $X_1, ..., X_1$ are iid. It follows from Lemma 2.1.1 that:

$$\begin{aligned}\mathbb{P}\left(X_1^2 + ... + X_n^2 \geq n(1+\epsilon)^2\right) &\leq \mathrm{e}^{n\left(\kappa_{X_1^2}(t) - t(1+\epsilon)^2\right)} \\ &= \mathrm{e}^{n\left(-\frac{1}{2}\ln(1-2t) - t(1+\epsilon)^2\right)}\end{aligned}$$

By derivation with respect to t, the quantity $-\frac{1}{2}\ln(1-2t) - t(1+\epsilon)^2$ reaches its minima value at $t_0 = \frac{1}{2}\left(1 - \frac{1}{(1+\epsilon)^2}\right) \in (0, \frac{1}{2})$. By taking this particular value of t, we have

$$-\frac{1}{2}\ln(1-2t_0) - t_0(1+\epsilon)^2 = \ln(1+\epsilon) - \epsilon - \frac{\epsilon^2}{2}.$$

Since $1 + \epsilon \leq \mathrm{e}^\epsilon$, we have

$$\ln(1+\epsilon) \leq \epsilon,$$

i.e

$$\ln(1+\epsilon) - \epsilon \leq 0,$$

so that

$$\ln(1+\epsilon) - \epsilon - \frac{\epsilon^2}{2} \leq -\frac{\epsilon^2}{2}.$$

Therefore

$$\mathbb{P}\left(X_1^2 + ... + X_n^2 \geq n(1+\epsilon)^2\right) \leq \mathrm{e}^{n\left(\ln(1+\epsilon) - \epsilon - \frac{\epsilon^2}{2}\right)} \leq \mathrm{e}^{-\frac{n\epsilon^2}{2}}.$$

- *For the second part:* consider $t \leq 0$. Then

$$\mathbb{P}\left(X_1^2 + ... + X_n^2 \leq n(1-\epsilon)^2\right) \leq \mathbb{P}\left(tX_1^2 + ... + tX_n^2 \geq tn(1-\epsilon)^2\right).$$

By using the same techniques as in the first part, we have

$$\begin{aligned}\mathbb{P}\left(X_1^2 + ... + X_n^2 \leq n(1-\epsilon)^2\right) &\leq \left(\frac{m_{X_1^2}(t)}{\mathrm{e}^{t(1-\epsilon)^2}}\right)^n = \mathrm{e}^{n\left(\kappa_{X_1^2}(t) - t(1-\epsilon)^2\right)} \\ &= \mathrm{e}^{n\left(-\frac{1}{2}\ln(1-2t) - t(1-\epsilon)^2\right)}.\end{aligned}$$

Again by derivation with respect to t, quantity $-\frac{1}{2}\ln(1-2t) - t(1-\epsilon)^2$ reaches it minima value at $t_1 = \frac{1}{2}\left(1 - \frac{1}{(1-\epsilon)^2}\right)$ ($t_1 < 0$ because we assumed that $0 < \epsilon \leq 1$, then $\frac{1}{(1-\epsilon)^2} \geq 1$), and by taking this particular value of t, we have:

$$-\frac{1}{2}\ln(1-2t_1) - t_1(1-\epsilon)^2 = \ln(1-\epsilon) + \epsilon - \frac{\epsilon^2}{2}.$$

Since $0 \leq 1-\epsilon \leq \mathrm{e}^{-\epsilon}$, we have

$$\ln(1-\epsilon) \leq -\epsilon,$$

i.e

$$\ln(1-\epsilon) + \epsilon \leq 0,$$

so that

$$\ln(1-\epsilon) + \epsilon - \frac{\epsilon^2}{2} \leq -\frac{\epsilon^2}{2}.$$

Therefore

$$\mathbb{P}\left(X_1^2 + ... + X_n^2 \leq n(1-\epsilon)^2\right) \leq \mathrm{e}^{n\left(\ln(1-\epsilon) + \epsilon - \frac{\epsilon^2}{2}\right)} \leq \mathrm{e}^{-\frac{n\epsilon^2}{2}}.$$

Therefore,

$$\gamma_n(|||X|| - \sqrt{n}| \geq \epsilon\sqrt{n}) \leq \mathrm{e}^{-\frac{n\epsilon^2}{2}} + \mathrm{e}^{-\frac{n\epsilon^2}{2}} = 2\mathrm{e}^{-\frac{n\epsilon^2}{2}}$$

and the theorem is proved. □

Theorem 2.1.2 describes the concentration of n-dimensional standard Gaussian measure around the sphere of radius $\sqrt{n}$ of $\mathbb{R}^n$. Figure 2.1 illustrates the case where $\epsilon = \frac{1}{n^{\frac{1}{3}}}$: in this case, we have

$$\gamma_n(|||X|| - \sqrt{n}| > \sqrt{n}/n^{1/3} = n^{1/6}) \leq 2\mathrm{e}^{-\frac{n^{1/3}}{2}} \longrightarrow 0$$

as n goes to the infinity.

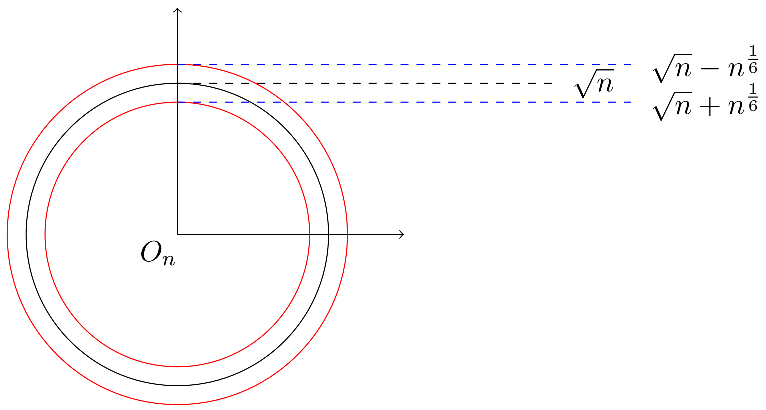

Figure 2.1: Most of the measure is contained between the two red spheres.

Now, let us investigate the concentration of the Euclidean distance in the metric measurable space $(\mathbb{R}^n, ||.||, \gamma_n)$. Let $X_1, Y_1, X_2, Y_2..., X_k, Y_k... \sim \mathcal{N}(0,1)$ be iid real random variables. For $n \in \mathbb{N}$, consider the n-coordinates standard Gaussian vectors $X := (X_1, ..., X_n), Y := (Y_1, ..., Y_n) \in \mathbb{R}^n$. We are interested in de distance $\mathrm{d}(X,Y) := ||X - Y||$.

Lemma 2.1.3. If $X_1, Y_1 \sim \mathcal{N}(0,1)$ are independent, then $X_1 - Y_1 \sim \mathcal{N}(0,2)$. Consequently, we have:

$$\frac{X - Y}{\sqrt{2}} \sim \gamma_n.$$

Proof. Since $X_1 - Y_1 = X_1 + (-Y_1)$, X_1 and Y_1 are independent and $-Y_1 \sim \mathcal{N}(0,1)$, then the distribution of $X_1 - Y_1$ is the convolution of the distribution of X_1 and Y_1. That is:

$$X_1 - Y_1 \sim \mathcal{N}(0,1) * \mathcal{N}(0,1) = \mathcal{N}(0, 2.)$$

□

The following theorem describes the concentration of distance phenomenon in the space $(\mathbb{R}, ||.||, \gamma_n)$.

Theorem 2.1.4. *Let $X, Y \sim \gamma_n$ be independent and let $\epsilon > 0$. Then:*

$$\mathbb{P}\left(|\mathrm{d}(X,Y) - \sqrt{2n}| \geq \epsilon\sqrt{n}\right) \leq 2\mathrm{e}^{-\frac{n\epsilon^2}{4}}.$$

Therefore

$$\frac{\mathrm{d}(X,Y)}{\sqrt{2n}} \xrightarrow{\mathbb{P}} 1.$$

Proof. By Lemma 2.1.3, we have $\frac{X-Y}{\sqrt{2}} \sim \gamma_n$. Since $\mathrm{d}(X,Y) = ||X - Y||$, then:

$$\begin{aligned}\mathbb{P}\left(|\mathrm{d}(X,Y) - \sqrt{2n}| \geq \epsilon\sqrt{n}\right) &= \mathbb{P}\left(|\frac{\mathrm{d}(X,Y)}{\sqrt{2}} - \sqrt{n}| \geq \frac{\epsilon}{\sqrt{2}}\sqrt{n}\right) \\ &= \gamma_n\left(|||\chi|| - \sqrt{n}| \geq \frac{\epsilon}{\sqrt{2}}\sqrt{n}\right),\end{aligned}$$

where $\chi \sim \gamma_n$. The result follows by Theorem 2.1.2.

□

2.2 Concentration of spherical measure

In this section, we are going to explain how the concentration of the Gaussian measure in section 2.1 implies the concentration of measure in the metric measurable space $(\mathrm{S}^{n-1}, ||.||, \sigma_n)$, where $\mathrm{S}^{n-1} \subseteq \mathbb{R}^n$ is the n-dimensional unit sphere and σ_n the normalised Lebesgue measure.

Consider the radial projection

$$r : X \in (\mathbb{R}^n, ||.||, \gamma_n) \longmapsto \frac{X}{||X||} \in \left(\mathrm{S}^{n-1}, ||.||, \sigma_n\right).$$

Since the probability density $f_{\gamma_n}(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{||x||^2}{2}}$ of γ_n at a point $x \in \mathbb{R}^n$ depends only on the norm of x, then γ_n is invariant under rotation, and so is its measure image by r. Since σ_n is the only rotational invariant probability measure on S^{n-1}, then it turns out that σ_n is the measure image of γ_n by the radial projection r.

Now, let $x \sim \sigma_n$ be an uniformly chosen random point in the sphere S^{n-1} and let e_n denotes the unit vector $(1, ..., 0) \in \mathrm{S}^{n-1}$. We are going to study the concentration of measure&distance phenomenon on S^{n-1} by considering the angle $\alpha := \angle(e_n, x) \in [0, \pi]$. Let us first recall the following property of real standard normal random variable.

Lemma 2.2.1. Let $X \sim \mathcal{N}(0, 1)$ and let $a > 0$. Then

$$\mathbb{P}(X \geq a) \leq \mathrm{e}^{-\frac{a^2}{2}}.$$

Proof. The result is due to the fact that the moment generating function of X is defined for $t \in \mathbb{R}$ by $m_X(t) := \mathbb{E}\left(\mathrm{e}^{tX}\right) = \mathrm{e}^{\frac{t^2}{2}}$. By applying Markov's inequality:

$$\begin{aligned} \mathbb{P}(X \geq a) &= \mathbb{P}(tX \geq ta) = \mathbb{P}\left(\mathrm{e}^{tX} \geq \mathrm{e}^{ta}\right) \\ &\leq \frac{\mathbb{E}\mathrm{e}^{tX}}{\mathrm{e}^{ta}} = \frac{\mathrm{e}^{\frac{t^2}{2}}}{\mathrm{e}^{ta}} = \mathrm{e}^{\frac{t^2}{2} - ta}, \end{aligned}$$

where $t > 0$. By choosing the particular value $t = a$ where the quantity $\frac{t^2}{2} - ta$ is minimal, we have the result. □

Theorem 2.2.2. *Let $x \sim \sigma_n$ be a random point in the n-dimensional unit sphere S^{n-1} and let $\epsilon > 0$. Then we have:*

$$\sigma_n\left(|\cos\alpha| \geq \epsilon\right) \leq 4\mathrm{e}^{-n\frac{\epsilon^2}{2}},$$

where α is the angle $\angle(e_n, x)$ (Fig. 2.2).

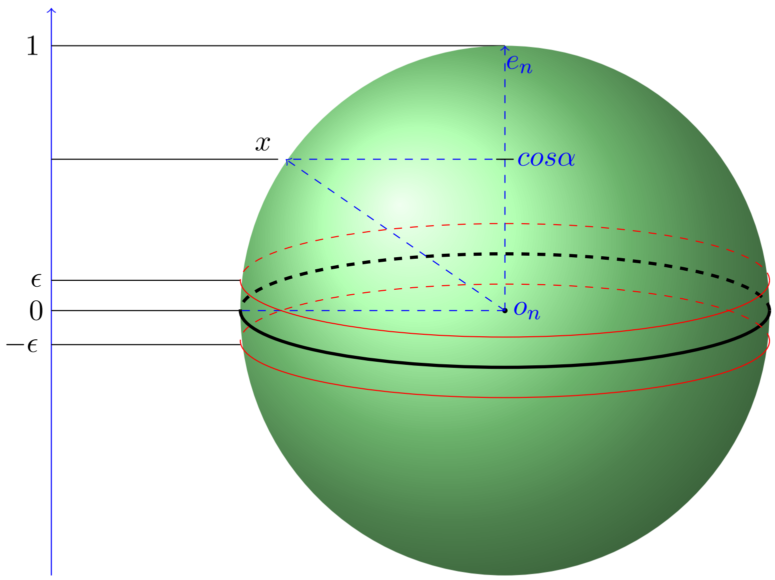

Figure 2.2: σ_n is concentrated around the equator.

Proof. By using the relation between σ_n and γ_n, we have:

$$\sigma_n\left(\cos\alpha \geq \epsilon\right) = \gamma_n\left(\cos\left(\angle(e_n, \frac{X}{||X||})\right) \geq \epsilon\right),$$

where $X = (X_1, ..., X_n)$ with $X_1, ..., X_n \sim \mathcal{N}(0,1)$ iid. Since both e_n and $\frac{X}{||X||}$ are unit vectors, we have

$$\cos\left(\angle(e_n, \frac{X}{||X||})\right) =< e_n, \frac{X}{||X||} >= \frac{X_1}{||X||},$$

where $< ., . >$ denotes the usual scalar product of $\mathbb{R}^n$.

Therefore:

$$\begin{aligned}\sigma_n\left(\cos\alpha \geq \epsilon\right) &= \gamma_n\left(\frac{X_1}{||X||} \geq \epsilon\right) = \gamma_n\left(X_1 \geq \epsilon||X||\right)\\ &= \gamma_n\left(X_1 \geq \epsilon||X||,\ ||X|| - \sqrt{n} \geq \epsilon\sqrt{n}\right) + \gamma_n\left(X_1 \geq \epsilon||X||,\ ||X|| - \sqrt{n} < \epsilon\sqrt{n}\right)\\ &\leq \gamma_n\left(||X|| - \sqrt{n} \geq \epsilon\sqrt{n}\right) + \mathbb{P}\left(X_1 \geq \epsilon(1+\epsilon)\sqrt{n}\right)\\ &\leq \mathrm{e}^{-n\frac{\epsilon^2}{2}} + \mathrm{e}^{-\frac{n\epsilon^2(1+\epsilon)^2}{2}} \leq \mathrm{e}^{-n\frac{\epsilon^2}{2}} + \mathrm{e}^{-n\frac{\epsilon^2}{2}} = 2\mathrm{e}^{-n\frac{\epsilon^2}{2}}\end{aligned}$$

by Lemma 2.2.1 and Theorem 2.1.2, and the result follows by symmetry. □

The concentration described in Theorem can be interpreted in different ways. Recall first that the group of rotation acts transitively on the unit sphere and the probability measure σ_n is invariant under rotation. Then if we pick two independent points $x, y \sim \sigma_n$ on the sphere, then there exist a rotation r such that $r(y) = e_1$. Since

$$\mathrm{d}(x, y) = \mathrm{d}(r(x), r(y)) = \mathrm{d}(e_1, r(x))$$

and Theorem says that $\mathrm{d}(e_1, r(x))$ is concentrated around $\sqrt{2}$, then $\mathrm{d}(x, y)$ is also concentrated around $\sqrt{2}$. Equivalently, since rotations preserve the scalar product, then Theorem means that the scalar product

$$< x, y >=< r(x), r(y) >=< e_1, r(x) >$$

is also concentrated around 0, i.e. two independent uniformly randomly chosen points on the unit sphere are “almost” orthogonal.

3. Discrete examples

I this chapter, we are going to study the concentration of measure and distance on a particular family of finite and connected graphs (or random graphs) $(G_n)_n$, endowed with the graph distance and the uniform probability measure on the vertices.

3.1 Integer lattices and torus

Let us first give an example where the distance does not have concentration property. For a positive integer n, let $G = \{0, 1, ..., n\}^2 \subseteq (\mathbb{Z}^2, ||.||_1)$ be a square lattice of size $(n+1)^2$, endowed with the Ll^1- distance. Let $X = (x_1, x_2)$ be a uniformly chosen vertex of G, i.e. $x_1, x_2 \sim \text{Unif}\{0, 1, 2, ..., n\}$ are independent. The l^1-distance between $o = (0, 0)$ and X, which is exactly their graph distance, is given by:

$$\mathrm{d}(o, X) = x_1 + x_2.$$

Since x_1 and x_2 are independent, then the distribution of $\mathrm{d}(o, X)$ is given by the convolution

$$\text{Unif}\{0, 1, 2, ..., n\} \star \text{Unif}\{0, 1, 2, ..., n\},$$

i.e, for $k = 0, 1, ..., 2n$:

$$\mathbb{P}\left(\mathrm{d}(o, X) = k\right) = \begin{cases} \frac{k+1}{(n+1)^2} & \text{if } k \leq n, \\ \frac{2n-k+1}{(n+1)^2} & \text{if } k \geq n. \end{cases}$$

The aim of this example is to point out why $\mathrm{d}(o, X)$ does not have the concentration property. The idea is that given a convergent sequence of positive number $(u_n)_n$ (the limit can be ∞), the quotient $\frac{\mathrm{d}(o,X)}{u_n}$ cannot be close to 1 with high probability whatever the asymptotic behaviour of u_n is. This can be formalised in the following theorem.

Theorem 3.1.1. *Let $(u_n)_n$ be a convergent sequence of positive numbers (the limit can be ∞) and let $\epsilon \in (0, 1)$ be fixed. Then there exists a constant $c = c(\epsilon) \in (o, 1)$ such that:*

$$\lim_{n \longrightarrow \infty} \mathbb{P}\left(|\frac{\mathrm{d}(o, X)}{u_n} - 1| > \epsilon\right) \geq c.$$

Proof. The two events $[\mathrm{d}(o, X) > (1 + \epsilon)u_n]$ and $[\mathrm{d}(o, X) < (1 - \epsilon)u_n]$ are disjoint for all n, then:

$$\begin{aligned} \mathbb{P}\left(|\frac{\mathrm{d}(o, X)}{u_n} - 1| > \epsilon\right) &= \mathbb{P}\left(\frac{\mathrm{d}(o, X)}{u_n} - 1 > \epsilon \text{ or } \frac{\mathrm{d}(o, X)}{u_n} - 1 < -\epsilon\right) \\ &= \mathbb{P}\left[\mathrm{d}(o, X) > (1 + \epsilon)u_n \text{ or } \mathrm{d}(o, X) < (1 - \epsilon)u_n\right] \\ &= \mathbb{P}[\mathrm{d}(o, X) > (1 + \epsilon)u_n] + \mathbb{P}[\mathrm{d}(o, X) < (1 - \epsilon)u_n]. \end{aligned}$$

-If $u_n >> n$:

since $\mathrm{d}(o, X) \in \{0, 1, ..., 2n\}$, then for large value of n we have $\mathbb{P}[\mathrm{d}(o, X) > (1 + \epsilon)u_n] = 0$ and $\mathbb{P}[\mathrm{d}(o, X) < (1 - \epsilon)u_n] = 1$.

-If $u_n << n$**:**

in this case, notice that for any possible value of k, we always have:

$$\mathbb{P}[\mathrm{d}(o,X)=k] \leq \frac{1}{n+1}.$$

Then:

$$\begin{aligned}\mathbb{P}[\mathrm{d}(o,X)>(1+\epsilon)u_n] &= 1-\mathbb{P}[\mathrm{d}(o,X)\leq(1+\epsilon)u_n] \\ &= 1-\frac{\lfloor(1+\epsilon)u_n\rfloor}{n+1} \xrightarrow[n\longrightarrow\infty]{} 1,\end{aligned}$$

where $\lfloor(1+\epsilon)u_n\rfloor$ denotes the floor of $(1+\epsilon)u_n$.

-If $u_n \sim n$:

for large values of n, we have $(1-\epsilon)u_n \leq n \leq (1+\epsilon)u_n \leq 2n$. By using the formula of the distribution of $\mathrm{d}(o,X)$, we have:

$$\begin{aligned}\mathbb{P}[\mathrm{d}(o,X)>(1+\epsilon)u_n] &= \sum_{k=\lceil(1+\epsilon)u_n\rceil}^{2n}\frac{2n-k+1}{(n+1)^2} \\ &= \frac{1}{(n+1)^2}[1+2+...+(2n-\lceil(1+\epsilon)u_n\rceil+1)] \\ &\sim \frac{1}{(n+1)^2}\frac{(2n-\lceil(1+\epsilon)u_n\rceil)^2}{2} \\ &\xrightarrow[n\longrightarrow\infty]{} \frac{(1-\epsilon)^2}{2},\end{aligned}$$

where $\lceil(1+\epsilon)u_n\rceil$ denotes the Ceil of $(1+\epsilon)u_n$; and

$$\begin{aligned}\mathbb{P}[\mathrm{d}(o,X)<(1-\epsilon)u_n] &= \sum_{k=0}^{\lfloor(1-\epsilon)u_n\rfloor}\frac{k+1}{(n+1)^2} \\ &\sim \frac{1}{(n+1)^2}\frac{(\lfloor(1-\epsilon)u_n\rfloor)^2}{2} \\ &\xrightarrow[n\longrightarrow\infty]{} \frac{(1-\epsilon)^2}{2}.\end{aligned}$$

Hence, we can take $c=(1-\epsilon)^2$ and the theorem is proved. □

We can apply the non concentration phenomenon of the square lattice $\{0,1,...,n\}^2$ to the case of the integer torus $\mathbb{Z}^2/(n\mathbb{Z})^2$, which is obtained by glueing the two opposite sides of the square lattice $\{0,1,...,n\}^2$ (Fig 3.1). In this glueing, the elements $o=(0,0),(n,0),(0,n)$ and (n,n) of the lattice give one point $\bar{O}$ on the torus.

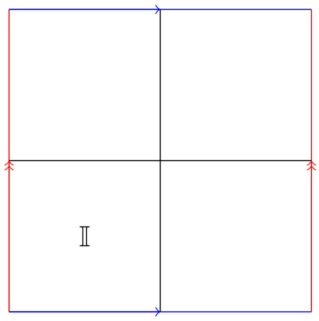

Figure 3.1: We obtain the integer torus by glueing the two pair sides of the same colour.

By symmetry, the distance between $\bar{O}$ and a uniformly chosen point in the integer torus is completely determined by the distance between o and a point $X \in \mathbb{I}$ (see Fig 3.1) of the lattice. Since the part $\mathbb{I}$ contains a positive fraction of the lattice and $\mathbb{I}$ does not have any concentration property, then the distance on the integer torus also does not have any concentration property.

In contrast with the 2-dimensional integer torus and lattice, let us consider the unit cube $\{0,1\}^n$, i.e. we let the dimension n goes to the infinity. Let $X \sim \mathrm{Unif}(\{0,1\}^n)$, then the L^1 norm of X which is exactly the L^1 distance between X and the n-dimensional zero vector $O = (0, ..., 0)$ is $\mathcal{B}(n, \frac{1}{2})$-distributed. We then have the following theorem.

Theorem 3.1.2. *Let $X \sim \mathrm{Unif}(\{0,1\}^n)$ and let $\epsilon > 0$. Then there exists $\alpha(\epsilon) > 0$ such that*

$$\mathbb{P}\left(\left|\frac{2\mathrm{d}(X,O)}{n} - 1\right| > \epsilon\right) \leq 2\mathrm{e}^{-\alpha(\epsilon)n},$$

where d denotes the L^1 distance.

Proof. Notice that $\mathbb{E}\left(\mathcal{B}(n, \frac{n}{2})\right) = \frac{1}{2}$ and the moment generating function of $\mathcal{B}(n, \frac{1}{2})$ is defined for t in a neighbourhood of 0 by

$$m(t) = \left(\frac{\mathrm{e}^t + 1}{2}\right)^n.$$

The result follows by using the same argument as in the proof of Theorem 2.1.2. □

3.2 Uniform spanning tree

Let n be a positive integer and let $\mathcal{T}_n$ denotes the set of spanning trees of the complete graph $\mathbb{K}_n$. The aim of this section is to explore the concentration of distance phenomenon on an uniformly chosen random tree of $\mathcal{T}_n$. The main result of this section is the Rayleigh limit in Theorem 3.2.5, which is mentioned in Peres and Revelle (2004).

Let $u \in V(K_n)$ and let S be a subset of $V(K_n)$. A Loop Erased Random Walk $\mathrm{LERW}(u, S)$ from u to S is a trajectory obtained by the following process: *start a simple symmetric random walk at u, remove each loops by order of appearance and stop the walk when it hits an element of S.* In particular, $\mathrm{LERW}(u, S)$ is a sub-graph of K_n which is a path without loop. For a simple symmetric random walk $X_0, X_1, ..., X_n, ...$ on K_n, we denote $\mathbb{P}_u$ the transition probability of the walk starting at u, and we adopt the following notation for the hitting times:

$$\tau_u := inf\{i : X_i = u\}$$

and

$$\tau_S := inf\{i : X_i \in S\}.$$

Now, consider the random spanning tree T of K_n obtained in the following process:

(i) give any order $x_1, x_2, ..., x_n$ on the vertices of K_n and define the graph with one vertex $T_1 = (\{x_1\}, \emptyset)$;

(ii) for any positive integer $i > 1$, define the tree T_i whose edges consist on those of T_{i-1} and $\text{LERW}(x_i, V(T_{i-1}))$.

Since we have only finitely many vertices, then the above process will end up with a random element $T \in \mathcal{T}_n$.

The process described above is called Wilson's algorithm and we have the following theorem, which is also true for any finite connected graph G.

Theorem 3.2.1. *(Wilson's algorithm). Let T be a random spanning tree of K_n obtained by Wilson's algorithm. Then:*

$$T \sim \text{Unif}(\mathcal{T}_n).$$

Proof. See Lyons and Peres (2017). □

Theorem 3.2.1 allows us to compute the distribution of the tree distance $d_T(u, v)$ between between two fixed vertices u and v of K_n where $T \sim \text{Unif}(\mathcal{T}_n)$: for any positive integer k, the quantity

$$\mathbb{P}[d_T(u, v) = k] := \frac{\#\{T \in \mathcal{T}_n : d_T(u, v) = k\}}{\#\mathcal{T}_n}$$

is exactly the same as the probability that $\text{LERW}(u, \{v\})$ has k edges. Let us first prove the following lemma:

Lemma 3.2.2. Let u, v be two vertices of K_n $(u \neq v)$ and let $T \sim \text{Unif}(\mathcal{T}_n)$. Then

$$\mathbb{P}[d_T(u, v) = 1] = \frac{2}{n}.$$

Proof. We need to compute the probability that $\{u, v\}$ is the only edge of the loop erased random walk from u to v. This means that a simple symmetric random walk starting at u hits v in the first step, or it hits another vertex $z \neq v$ but returns back to u before hitting v.

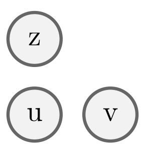

Figure 3.2: The walk stop after hitting v from u.

The probability of such event is given by the following recursion:

$$\mathbb{P}[d_T(u, v) = 1] = \mathbb{P}_u(v) + (1 - \mathbb{P}_u(v))\ \mathbb{P}_z[\tau_u < \tau_v]\ \mathbb{P}[d_T(u, v) = 1].$$

By symmetry, we have $\mathbb{P}_u(v) = \frac{1}{n-1}$ and $\mathbb{P}_z[\tau_u < \tau_v] = \frac{1}{2}$. Hence:

$$\mathbb{P}[d_T(u,v)=1] = \frac{1}{n-1} + \frac{n-2}{2(n-1)}\mathbb{P}[d_T(u,v)=1],$$

and the result follows by solving in $\mathbb{P}_z[\tau_u < \tau_v]$. □

In Lemma 3.2.2, we fix $u \neq v$ so that the event $[d_T(u,v) > 0]$ is trivial. The quantity in Lemma 3.2.2 is then same as the quantity

$$\mathbb{P}[d_T(u,v) = 1 \mid d_T(u,v) > 0].$$

We have a more general result in the following Lemma.

Lemma 3.2.3. For $i = 2, 3, ..., n-1$, we have:

$$\mathbb{P}[d_T(u,v) = i \mid d_T(u,v) > i-1] = \frac{i+1}{n},$$

where $T \sim \text{Unif}(\mathcal{T}_n)$.

Proof. Let $x_1 = u, x_2, ..., x_i$ be the first $i-th$ vertices of $LERW(u, \{v\})$. Then our condition $[d_t(u,v) > i-1]$ means that our random walk will never hit $\{x_1, ..., x_{i-1}\}$ again, i.e

$$\tau_v < \tau_{\{x_1,...,x_{i-1}\}}.$$

In this case, the event $[d_T(u,v) = i]$ means that either the walk hits v by its first step from x_i or it hits another vertex $y \notin \{x_1, ..., x_{i-1}, v\}$ and returns back to x_i before hitting v and so on.

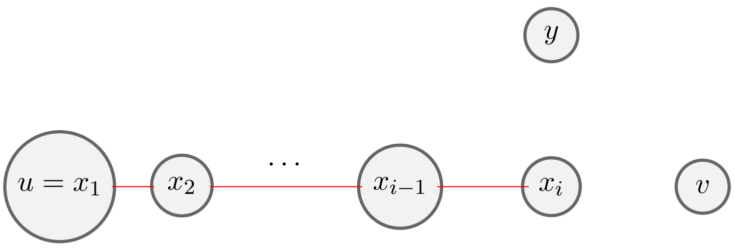

Figure 3.3: The red edges will not be removed before the walk hits v from x_i.

We then have the following recursion:

$$\mathbb{P}[d_T(u,v) = i \mid d_T(u,v) > i-1] = \mathbb{P}_{x_i}[v \mid \tau_v < \tau_{\{x_1,...,x_{i-1}\}}] + \Big(1 - \mathbb{P}_{x_i}[v \mid \tau_v < \tau_{\{x_1,...,x_{i-1}\}}]\Big) * \mathbb{P}_y[\tau_{x_i} < \tau_v \mid \tau_v < \tau_{\{x_1,...,x_{i-1}\}}] * \mathbb{P}[d_T(u,v) = i \mid d_T(u,v) > i-1].$$

Since the event that "v is reached by the first step from x_i" is a subset of the event $\tau_v < \tau_{\{x_1,...,x_{i-1}\}}$, we have by symmetry:

$$\mathbb{P}_{x_i}[v \mid \tau_v < \tau_{\{x_1,...,x_{i-1}\}}] = \frac{\mathbb{P}_{x_i}(v)}{\mathbb{P}_{x_i}[\tau_v < \tau_{\{x_1,...,x_{i-1}\}}]} = \frac{\frac{1}{n-1}}{\frac{(i-1)!}{i!}} = \frac{i}{n-1}.$$

On the other hand, we have:

$$\mathbb{P}_y[\tau_{x_i} < \tau_v \mid \tau_v < \tau_{\{x_1,\dots,x_{i-1}\}}] = \frac{\mathbb{P}_y[\tau_{x_i} < \tau_v < \tau_{\{x_1,\dots,x_{i-1}\}}]}{\mathbb{P}_y[\tau_v < \tau_{\{x_1,\dots,x_{i-1}\}}]} = \frac{\frac{(i-1)!}{(i+1)!}}{\frac{(i-1)!}{i!}} = \frac{1}{i+1}.$$

Therefore:

$$\mathbb{P}[d_T(u,v) = i \mid d_T(u,v) > i-1] = \frac{i}{n-1} + \frac{n-1-i}{n-1}\,\frac{1}{i+1}\,\mathbb{P}[d_T(u,v) = i \mid d_T(u,v) > i-1],$$

and the result follows by solving on $\mathbb{P}[d_T(u,v) = i \mid d_T(u,v) > i-1]$.

□

Lemma 3.2.2 and 3.2.3 allows us to compute the the distribution of $d_T(u,v)$, where u,v are fixed vertices and $T \sim \mathrm{uni}(\mathcal{T}_n)$.

Theorem 3.2.4. *Let u,v be two vertices of K_n and let $T \sim \mathrm{unif}(\mathcal{T}_n)$. Then for $k = 1,2,\dots,n-1$:*

$$\mathbb{P}[d_T(u,v) = k] = \frac{k+1}{n}\prod_{i=1}^{k-1}\left(1 - \frac{i+1}{n}\right).$$

Proof. We have

$$\begin{aligned}\mathbb{P}[d_T(u,v) = k] &= \mathbb{P}\left[d_T(u,v) = k \mid d_T(u,v) > k-1\right]\prod_{i=1}^{k-1}\mathbb{P}[d_T(u,v) > i \mid d_T(u,v) > i-1] \\ &= \mathbb{P}\left[d_T(u,v) = k \mid d_T(u,v) > k-1\right]\prod_{i=1}^{k-1}\left(1 - \mathbb{P}[d_T(u,v) = i \mid d_T(u,v) > i-1]\right),\end{aligned}$$

and the result follows by using the results of Lemma 3.2.2 and Lemma 3.2.3.

□

Let us now use the formula in Theorem 3.2.4 to compute the limit in distribution of $\frac{d_T(u,v)}{\sqrt{n}}$ as n goes to the infinity. Consider the function defined for $t \in \mathbb{R}$ by $f(t) = t\mathrm{e}^{-\frac{t^2}{2}}\mathbb{I}_{t\geq 0}$. We have

$$\int_{\mathbb{R}} f(t)dt = \left[1 - \mathrm{e}^{-\frac{t^2}{2}}\right]_{t=0}^{t=\infty} = 1.$$

Hence f is the density with respect to the Lebesgue measure of a probability measure on $\mathbb{R}$. This probability measure with density f is called the Rayleigh distribution.

Theorem 3.2.5. *Let u,v be two vertices of K_n and let $T \sim \mathrm{unif}(\mathcal{T}_n)$. Then the random variable $\frac{d_T(u,v)}{\sqrt{n}}$ converges in distribution to the Rayleigh distribution when n goes to the infinity.*

Proof. Fix $t \geq 0$. Since $d_T(u,v) \in \{1,2,...,n-1\}$, then $\frac{d_T(u,v)}{\sqrt{n}} \in \{\frac{1}{\sqrt{n}}, \frac{2}{\sqrt{n}}, ..., \frac{n-1}{\sqrt{n}}\}$ and $\mathbb{P}\left(\frac{d_T(u,v)}{\sqrt{n}} = \frac{k}{\sqrt{n}}\right) = \mathbb{P}[d_T(u,v) = k]$ for $k = 1,2,...,n-1$. So we have:

$$\begin{aligned}
\mathbb{P}(\frac{d_T(u,v)}{\sqrt{n}} \leq t) &= \sum_{\frac{k}{\sqrt{n}} \leq t} \frac{k+1}{n} \prod_{i=1}^{k-1} \left(1 - \frac{i+1}{n}\right) \\
&= \sum_{\frac{k}{\sqrt{n}} \leq t} \frac{k+1}{n} \exp\left(\sum_{i=1}^{k-1} \ln\left(1 - \frac{i+1}{n}\right)\right) \\
&= \sum_{\frac{k}{\sqrt{n}} \leq t} \frac{k}{n} \exp\left(\sum_{i=1}^{k-1} \ln\left(1 - \frac{i+1}{n}\right)\right) + \sum_{\frac{k}{\sqrt{n}} \leq t} \frac{1}{n} \exp\left(\sum_{i=1}^{k-1} \ln\left(1 - \frac{i+1}{n}\right)\right).
\end{aligned}$$

If $\frac{k}{\sqrt{n}} \leq t$, then $\frac{i+1}{n} \leq \frac{k}{n} \leq \frac{t}{\sqrt{n}}$ for $i = 1,...,k-1$. Since t is fixed, then $\frac{i+1}{n}$ converges to zero when n goes to the infinity. Hence, $\ln(1 - \frac{i+1}{n}) = -\frac{i+1}{n} + (i+1)o(\frac{1}{n})$ for $i = 1,...,k-1$, and we have:

$$\begin{aligned}
\sum_{i=1}^{k-1} \ln\left(1 - \frac{i+1}{n}\right) &= -\sum_{i=1}^{k-1} \frac{i+1}{n} + \sum_{i=1}^{k-1} o(\frac{i+1}{n}) \\
&= -\frac{(k-1)(k+2)}{2n} + o\left(\frac{(k-1)(k+2)}{2n}\right) \\
&= -\frac{k^2}{2n} - \frac{k}{2n} + \frac{1}{n} + o(\frac{k^2}{2n}) \\
&= -\frac{k^2}{2n} + o(1),
\end{aligned}$$

because $\frac{k}{\sqrt{n}} < t$ implies $\frac{k^2}{2n} \leq \frac{t^2}{2}$ and $\frac{k}{2n} \leq \frac{t}{\sqrt{n}}$ (t is fixed).

Therefore,

$$\begin{aligned}
\lim_{n \longrightarrow \infty} \mathbb{P}(\frac{d_T(u,v)}{\sqrt{n}} \leq t) &= \lim_{n \longrightarrow \infty} \sum_{\frac{k}{\sqrt{n}} \leq t} \frac{k}{n} \mathrm{e}^{-\frac{k^2}{2n}} + \lim_{n \longrightarrow \infty} \sum_{\frac{k}{\sqrt{n}} \leq t} \frac{1}{n} \mathrm{e}^{-\frac{k^2}{2n}} \\
&= \lim_{n \longrightarrow \infty} \sum_{\frac{k}{\sqrt{n}} \leq t} \frac{1}{\sqrt{n}} \left(\frac{k}{\sqrt{n}}\right) \mathrm{e}^{-\frac{1}{2}(\frac{k}{\sqrt{n}})^2} + \lim_{n \longrightarrow \infty} \frac{\sum_{\frac{t}{\sqrt{n}} \leq t} \frac{1}{\sqrt{n}} \mathrm{e}^{-\frac{1}{2}\left(\frac{k}{\sqrt{n}}\right)^2}}{\sqrt{n}}.
\end{aligned}$$

Let us compute separately the two terms in the right.

- We have

$$\lim_{n \longrightarrow \infty} \sum_{\frac{k}{\sqrt{n}} \leq t} \frac{1}{\sqrt{n}} \left(\frac{k}{\sqrt{n}}\right) \mathrm{e}^{-\frac{1}{2}(\frac{k}{\sqrt{n}})^2} = \int_0^t x \mathrm{e}^{-\frac{x^2}{2}} \, dx.$$

  Indeed, the sum

$$\sum_{\frac{k}{\sqrt{n}} \leq t} \frac{1}{\sqrt{n}} \left(\frac{k}{\sqrt{n}}\right) \mathrm{e}^{-\frac{1}{2}(\frac{k}{\sqrt{n}})^2}$$

is exactly the Darboux sum of the function $xe^{-\frac{x^2}{2}}$ with respect to the partition

$$[o,t] = [0, \frac{1}{\sqrt{n}}[\cup[\frac{1}{\sqrt{n}}, \frac{2}{\sqrt{n}}[\cup ... \cup [\frac{k(t)}{\sqrt{n}}, t]$$

of the interval $[0,t]$, and the length of a part of this partition is at most $1/\sqrt{n}$.

- Likewise, we have

$$\lim_{n\longrightarrow\infty} \sum_{\frac{t}{\sqrt{n}}\leq t} \frac{1}{\sqrt{n}} e^{-\frac{1}{2}\left(\frac{k}{\sqrt{n}}\right)^2} = \int_0^t e^{-\frac{x^2}{2}} dx,$$

by considering the same partition as before and the sum

$$\sum_{\frac{t}{\sqrt{n}}\leq t} \frac{1}{\sqrt{n}} e^{-\frac{1}{2}\left(\frac{k}{\sqrt{n}}\right)^2}$$

is exactly the Darboux sum of the function $e^{-x^2/2}$.

Hence,

$$\lim_{n\longrightarrow\infty} \frac{\sum_{\frac{t}{\sqrt{n}}\leq t} \frac{1}{\sqrt{n}} e^{-\frac{1}{2}\left(\frac{k}{\sqrt{n}}\right)^2}}{\sqrt{n}} = 0.$$

Therefore,

$$\lim_{n\longrightarrow\infty} \mathbb{P}(\frac{d_T(u,v)}{\sqrt{n}} \leq t) = \int_0^t xe^{-\frac{x^2}{2}} dx = \int_{-\infty}^t xe^{-\frac{x^2}{2}} \mathbb{I}_{x>0} dx,$$

and the integral in the right hand side is exactly the cumulative distribution function of the Rayleigh distribution. □

The result of Theorem 3.2.5 is also true in the case where $(u,v) \sim \text{unif}(V(K_n)^2)$ and $T\text{unif}(\mathcal{T}_n)$. In this case, we have:

$$\begin{aligned}
\mathbb{P}\left(\frac{d_T(u,v)}{\sqrt{n}} \leq t\right) &= \mathbb{P}\left(\frac{d_T(u,v)}{\sqrt{n}} \leq t \mid u \neq v\right) \mathbb{P}[u \neq v] + \mathbb{P}\left(\frac{d_T(u,v)}{\sqrt{n}} \leq t \mid u = v\right) \mathbb{P}[u = v] \\
&= \mathbb{P}\left(\frac{d_T(u,v)}{\sqrt{n}} \leq t \mid u \neq v\right) \frac{n^2-n}{n^2} + \mathbb{P}\left(\frac{d_T(u,v)}{\sqrt{n}} \leq t \mid u = v\right) \frac{1}{n} \\
&\sim \mathbb{P}\left(\frac{d_T(u,v)}{\sqrt{n}} \leq t \mid u \neq v\right),
\end{aligned}$$

and we obtain the same limit as in Theorem 3.2.5. Hence, we can conclude this section with the following corollary:

Corollary 3.2.6. Let $(u,v) \sim \text{unif}(V(K_n)^2)$ and let $T \sim \text{unif}(\mathcal{T}_n)$. Then the quantity

$$\frac{d_T(u,v)}{\sqrt{n}}$$

converges in distribution to the Rayleigh distribution.

3.3 Concentration and expander graphs

Expander graphs are graphs with high connectivity property in the sense that in order to make the graphs disconnected, one needs to cut a "large" number of edges. In geometric point of view, expander graphs are considered as graphs satisfying a "good" isoperimetric inequality, meaning that the order of the "surface" of any ball is at least the same as the order of its "volume". In this section, we are going to show that the distance in such kind of graphs has a concentration property. An explicit construction of a family of expander graphs is given in Lubotzky (2010).

Let us first start by giving a formal definition what expander graphs and some of their spectral characterisations.

Definition 3.3.1. Fix $d \in \mathbb{N}$ and $c > 0$. A graph $G\,(V(G), E(G))$ of size $n = |V(G)|$ is called an (n, d, c)-*expander* if the following are satisfied:

(i) $deg(v) \leq d$ for all $v \in V(G)$,

(ii) $|\partial_{\mathrm{v}}^{\mathrm{out}} A| \geq c|A|$ for any $A \subseteq V(G)$ such that $|A| \leq \frac{n}{2}$. Or equivalently,

$$|\{v \in V(G) : \mathrm{d}(v, A) \leq 1\}| \geq (1 + c)|A|$$

for $A \subseteq V(G)$ such that $|A| \leq \frac{n}{2}$.

The constant c is called the *Cheeger constant* of G. If in addition G is transitive, we say that G is a *transitive expander*.

Now, let us describe how the second largest eigenvalue (or the spectral gap) of a connected graph makes it a good expander. Let $G = (V(G), E(G))$ be a connected graph with n vertices, and assume that $\deg(x) \leq d$ for all $x \in V(G)$ for some fixed d. Let P denotes the transition probability of a simple symmetric random walk on G i.e.

$$P_{x,y} = \frac{\mathbb{I}_{\{(x,y) \in E(G)\}}}{\deg(x)}.$$

The probability measure π defined for $x \in V(G)$ by $\pi(x) = \frac{\deg(x)}{2|E(G)|}$ is the stationary distribution of P.

Since

$$\pi(x) P_{x,y} = \frac{\mathbb{I}_{\{(x,y) \in E(G)\}}}{2|E(G)|} = \pi(y) P_{y,x}$$

for $x, y \in E(G)$, then π is also reversible.

Let $\lambda_2(G)$ denote the second largest (in absolute value) eigenvalue of P, and define the spectral gap:

$$\lambda(G) := 1 - \lambda_2(G)$$

Given a real valued function f on $V(G)$, the Dirichlet energy of f is defined by:

$$\mathcal{E}(f) := \frac{1}{2} \sum_{x,y \in V(G)} \left(f(x) - f(y)\right)^2 \pi(x) P_{x,y}.$$

Recall the variational version of the spectral gap $\lambda(G)$, see Levin and Peres (2017):

$$\lambda(G) = \inf_{\mathbb{E}_\pi f=0} \left\{ \frac{\mathcal{E}(f)}{\text{Var}_\pi f} : \text{Var}_\pi \text{f} \neq 0 \right\},$$

where $\mathbb{E}_\pi f$ and $Var_\pi f$ denotes respectively the expectation and the variance of f with respect to the probability measure π.

Now we have enough material to prove the following proposition.

Proposition 3.3.2. Let G be as above and let $A \subseteq V(G)$ such that $0 \neq |A| \leq \frac{n}{2}$. Then

$$\frac{\lambda(G)}{4d^2}|A| \leq |\partial_V^{\text{out}} A|,$$

where $\partial_V^{\text{out}} A$ denotes the outer vertex boundary of A.

In another words, Proposition 3.3.2 says that graphs with large spectral gap are good expanders.

Proof. Let $a = |A|$ and let $E(A, A^c)$ denotes the set of edges between A and A^c. Consider the function

$$f := \frac{n}{a}\mathbb{I}_{\mathbb{A}} - \frac{n}{n-a}\mathbb{I}_{A^c}$$

defined on $V(G)$.

- By reversibility of the probability distribution π, we have:

$$\begin{aligned} \mathcal{E}(f) &= \sum_{e \in E(A,A^c)} \left(\frac{n}{a} + \frac{n}{n-a} \right)^2 \frac{1}{2|E(G)|} \\ &= \left(\frac{n}{a} + \frac{n}{n-a} \right)^2 \frac{|E(A,A^c)|}{2|E(G)|} \\ &\leq \left(\frac{n}{a} + \frac{n}{n-a} \right)^2 d\frac{|\partial_V^{\text{out}} A|}{2|E(G)|} \end{aligned}$$

- By definition, we have

$$\mathbb{E}_\pi f = \frac{n}{a}\pi(A) - \frac{n}{n-a}\pi(A^c).$$

- On the other hand, we have $f^2 = \frac{n^2}{a^2}\mathbb{I}_A + \frac{n^2}{(n-a)^2}\mathbb{I}_{A^c}$; then:

$$\mathbb{E}_\pi f^2 = \frac{n^2}{a^2}\pi(A) + \frac{n^2}{(n-a)^2}\pi(A^c).$$

Since $\pi(A^c) = 1 - \pi(A)$, we have:

$$\begin{aligned} \text{Var}_\pi(f) &= \mathbb{E}_\pi f^2 - (\mathbb{E}_\pi f)^2 \\ &= \frac{n^2}{a^2}\pi(A)\pi(A^c) + \frac{n^2}{(n-a)^2}\pi(A^c)\pi(A) + 2\frac{n^2}{a(n-a)}\pi(A)\pi(A^c) \\ &= \left(\frac{n}{a} + \frac{n}{n-a} \right)^2 \pi(A)\pi(A^c) \\ &= \left(\frac{n}{a} + \frac{n}{n-a} \right)^2 \frac{\sum_{v\in A} \deg(v)}{2|E(G)|} \frac{\sum_{v\in A^C} \deg(v)}{2|E(G)|} \\ &\geq \left(\frac{n}{a} + \frac{n}{n-a} \right)^2 \frac{a(n-a)}{4|E(G)|^2}, \end{aligned}$$

since G is connected.

Therefore, by the variational version of the spectral gap:

$$\begin{aligned}\lambda(G) &\leq d\frac{2|E(G)|}{a(n-a)}|\partial_V^{\text{out}} A| \\ &\leq \frac{4d^2}{a}|\partial_V^{\text{out}} A| = 4d^2\frac{|\partial_V^{\text{out}} A|}{|A|},\end{aligned}$$

since $\deg(v) \leq d$ for $v \in V(G)$ and $a = |A| \leq \frac{n}{2}$

□

By definition, any finite and connected graph is an expander graph in a trivial way. Surprisingly, the existence of a family of expanders with the same d and c has been proved and explicit construction has been done.

Definition 3.3.3. (Family of expanders) Let $d, c > 0$. A family of connected and finite graphs $(G_n)_n$ is called and family of $(|G_n|, d, c)$ expanders if:

- $\lim\limits_{n \longrightarrow \infty} |V(G_n)| = \infty$,
- $\deg(x) \leq d$ for all n and for all $x \in V(G_n)$,
- for all n and for any $A \subseteq V(G_n)$ such that $|A| \leq \frac{|G_n|}{2}$, we have
$$|\partial_{\text{v}}^{\text{out}} A| \geq c|A|;$$
or equivalently,
$$|\{v \in V(G) : \text{d}(v, A) \leq 1\}| \geq (1 + c)|A|$$
for $A \subseteq V(G_n)$ such that $|A| \leq \frac{|G_n|}{2}$.

By Proposition 3.3.2, the last point can be replaced by:

$$\lambda(G_n) > c$$

for all n.

In the rest of this section, we are going to prove concentration phenomenon on a family of expanders of the same d and c; let us first reset some notations. Given a graph $G = G\,(V(G), E(G))$ and $A \subseteq V(G)$, let

$$\partial_{\text{v}}^{\text{out}} A := \{v \in A^{\text{C}} : \text{d}(v, A) = 1\}$$

denotes the outer vertex boundary of A. For $v \in V(G)$ and $r \geq 0$, let

$$\text{B}_r(v) := \{x \in V(G) : \text{d}(x, v) \leq r\}$$

denotes the ball of radius r centred at v.

Let us now prove the concentration of distance on transitive expander graphs in two different methods.

- Concentration resulting from the variational version of the spectral gap

Let G be a transitive (n, d, c)-expander and fix $o \in V(G)$. Then the stationary distribution π is exactly the uniform distribution and each vertex has degree d. Consider the function:

$$f : x \in V(G) \longmapsto \mathrm{d}(o, x),$$

that sends x to the graph distance between x and o. We have the following lemma:

Lemma 3.3.4. Let $\mathbb{E}(f)$ denotes the expected value of f with respect to the distribution $\pi = \mathrm{Unif}(V(G))$. Then:

$$\frac{1}{\mathbb{E}(f)} = o(1)$$

Proof. We have:

$$\begin{aligned}
\mathbb{E}(f) &= \sum_{k\geq 0} \mathbb{P}(f \geq k) \\
&= \sum_{k} \left(\sum_{f(x)\geq k} \mathbb{P}(x) \right) = \sum_{k} \left(\sum_{f(x)\geq k} \frac{1}{n} \right) = \sum_{k\geq 0} \frac{n - |\mathcal{B}_o(k)|}{n} \\
&\geq \sum_{k\geq 0} \frac{n - d(d-1)^{k-1}}{n},
\end{aligned}$$

since each vertices have degree d so we have $|\mathcal{B}_o(k)| \leq d(d-1)^k$ for the cardinality of the ball of radius k centered at o.

Since our expanders have bounded degree, then f cannot be bounded. Hence there exists a sequence $(c_n)_n$ of positive integers such that $1 << c_n << \ln(n)$ and

$$\mathbb{E}(f) = \sum_{k\geq 0} \mathbb{P}(f \geq k) \geq \sum_{k=0}^{c_n} \mathbb{P}(f \geq k).$$

Therefore,

$$\begin{aligned}
\mathbb{E}(f) &\geq \sum_{k=0}^{c_n} \frac{n - d(d-1)^{k-1}}{n} = c_n + 1 - \sum_{k=0}^{c_n} \frac{d(d-1)^{k-1}}{n} \\
&= c_n + 1 - \frac{d}{n} \sum_{k=0}^{c_n} (d-1)^{k-1} \\
&= c_n + 1 - \frac{(d-1)^{c_n} - 1}{n} \longrightarrow \infty \text{ when } n \longrightarrow \infty
\end{aligned}$$

because $1 << c_n << \ln(n)$. The result then follows by taking the inverse.

□

The following proposition describes the concentration of distance phenomenon resulting from the variational version of the spectral gap for a family of expanders.

Proposition 3.3.5. The quantity $\frac{f}{\mathbb{E}(f)}$ converges in probability to 1 as n goes to the infinity.

Proof. By triangular inequality and by definition of P and π, we have:

$$\begin{aligned}\mathcal{E}(f) &= \frac{1}{2}\sum_{x,y\in V(G)}[\mathrm{d}(o,x)-\mathrm{d}(o,y)]^2\pi(x)P_{x,y}\\ &\leq \frac{1}{2}\sum_{x,y\in V(G)}\mathrm{d}(x,y)^2\pi(x)P_{x,y}\\ &= \frac{1}{2}\sum_{(x,y)\in E(G)}\pi(x)P_{x,y} = \frac{1}{2}\sum_{(x,y)\in E(G)}\frac{1}{dn}\\ &= \frac{1}{2}\frac{1}{dn}2|E(G)| = \frac{1}{2}\frac{dn}{dn} = \frac{1}{2}\end{aligned}$$

Hence, by using $f-\mathbb{E}f$ in the variational version of the spectral gap $\lambda(G)$:

$$\mathrm{Var}f \leq \frac{1}{2\lambda(G)}.$$

Therefore, for a fixed $\epsilon>0$, Chebyshev's inequality and Lemma 3.3.4 give:

$$\begin{aligned}\mathbb{P}\left(|\frac{f}{\mathbb{E}f}-1|\geq\epsilon\right) &= \mathbb{P}\left(|f-\mathbb{E}f|\geq\epsilon\mathbb{E}f\right)\\ &\leq \frac{\mathrm{Var}f}{\epsilon^2[\mathbb{E}f]^2} \leq \frac{1}{2\lambda(G)\epsilon^2[\mathbb{E}f]^2} = o(1)\end{aligned}$$

and the proposition is proved.

□

-Concentration resulting from the combinatorial definition of expanders

Let G be a transitive (n,d,c)-expander and fix $o\in V(G)$. Since we have finite number of vertices, there exists $r_0\geq 0$ such that $|\mathrm{B}_{r_0}(o)|\leq\frac{n}{2}$ and $|\mathrm{B}_{(r_0+1)}(o)|>\frac{n}{2}$. Our goal is to show that $\mathrm{d}(0,x)$ is concentrated around this r_0.

Lemma 3.3.6. Let $i\geq 1$. Then:

$$\mathrm{B}^{\mathrm{C}}_{(r_0+i-1)}(0) = \mathrm{B}^{\mathrm{C}}_{(r_0+i)}(0)\bigcup\partial^{\mathrm{out}}_{\mathrm{v}}\mathrm{B}^{\mathrm{C}}_{(r_0+i)}(0)$$

Proof. We have:

$$\begin{aligned}x\in\mathrm{B}^{\mathrm{C}}_{(r_0+i-1)}(0) &\iff \mathrm{d}(0,x)>r_0+i-1 \iff \mathrm{d}(0,x)\geq r_0+i\\ &\iff \mathrm{d}(o,x)=r_0+i \;\text{ or }\; \mathrm{d}(0,x)>r_0+i.\end{aligned}$$

The result follows by definition: $\mathrm{d}(0,x)>r_0+i$ means that $x\in\mathrm{B}^{\mathrm{C}}_{(r_0+i)}(0)$, and $\mathrm{d}(o,x)=r_0+i$ means that x belongs to the inner vertex-boundary of $\in\mathrm{B}_{(r_0+i)}(0)$ which is exactly the outer vertex-boundary of $\in\mathrm{B}^{\mathrm{C}}_{(r_0+i)}(0)$. □

The following corollary follows directly from the definition of expander and Lemma 3.3.6.

Corollary 3.3.7. Let G be an (n,d,c)- expander and let $0\leq k<r_0$. Then:

(i) $|\mathrm{B}_{(r_0-k)}(0)|\leq\frac{n}{2(1+c)^k}$,

(ii) $|\mathrm{B}^{\mathrm{C}}_{(r_0+k)}(0)| \le \frac{n}{2(1+c)^{k-1}}$.

Proof. (i) For $i \ge 0$, we have:

$$\{v \in V(G) : \mathrm{d}\,(v, \mathrm{B}_{i-1}(0)) \le 1\} = \mathrm{B}_i(0).$$

Then, by definition of expander graph and by definition of r_0 :

$$\begin{aligned}\frac{n}{2} &\ge |\mathrm{B}_{r_0}(0)| \ge (1+c)|\mathrm{B}_{(r_0-1)}(0)| \\ &\ge (1+c)^2|\mathrm{B}_{(r_0-2)}(0)| \ge ... \ge (1+c)^i|\mathrm{B}_{(r_0-i)}(0)| \ge ... \\ &\ge (1+c)^k|\mathrm{B}_{(r_0-k)}(0)|.\end{aligned}$$

Therefore:

$$|\mathrm{B}_{(r_0-k)}(0)| \le \frac{n}{2(1+c)^k}.$$

(ii) For $i = 1, 2, ..., k$, we have by Lemma 3.3.6:

$$\left\{v : \mathrm{d}\left(v, \mathrm{B}^{\mathrm{C}}_{(r_0+i)}(0)\right) \le 1\right\} = \mathrm{B}^{\mathrm{C}}_{(r_0+i-1)}(0).$$

Then, by definition of expander and by definition of r_0 :

$$\begin{aligned}\frac{n}{2} &\ge |\mathrm{B}^{\mathrm{C}}_{(r_0+1)}(0)| \ge (1+c)|\mathrm{B}^{\mathrm{C}}_{(r_0+2)}(0)| \ge ... \ge (1+c)^{j-1}|\mathrm{B}^{\mathrm{C}}_{(r_0+j)}(0)| \\ &\ge ... \\ &\ge (1+c)^{k-1}|\mathrm{B}^{\mathrm{C}}_{(r_0+k)}(0)|.\end{aligned}$$

So ve have:

$$|\mathrm{B}^{\mathrm{C}}_{(r_0+k)}(0)| \le \frac{n}{2(1+c)^{k-1}}.$$

□

Now we are ready to quantify the concentration of distance phenomenon for transitive expander graphs: given tow uniformly randomly chosen vertices $u, v \in V(G)$, there exists an element $g \in Aut(G)$ such that $g(u) = 0$. Hence, instead of taking two uniformly random vertices, we only take one and set the other to 0.

Theorem 3.3.8. *Let G be a transitive (n, d, c)-expander graph and fix $0 \in V(G)$. Let $r_0 \ge 0$ such that $|\mathrm{B}_{r_0}(o)| \le \frac{n}{2}$ and $|\mathrm{B}_{(r_0+1)}(o)| > \frac{n}{2}$ and let v be an uniformly chosen random vertices of G. Then for any $k > 0$:*

$$\mathbb{P}\left(|\mathrm{d}(0, v) - r_0| > k\right) \le (1+c)\mathrm{e}^{-k\ln(1+c)}.$$

As a result, the quantity $\frac{\mathrm{d}(0,v)}{r_0}$ converges to 1 in probability as $n \longrightarrow \infty$.

Proof. We have that:

$$\begin{aligned}\{v : |\mathrm{d}(0, v) - r_0| > k\} &= \{v : \mathrm{d}(0, v) > r_0 + k\} \bigcup \{v : \mathrm{d}(0, v) < r_0 - k\} \\ &\subseteq \mathrm{B}^{\mathrm{C}}_{(r_0+k)}(0) \bigcup \mathrm{B}_{(r_0-k)}(0).\end{aligned}$$

Hence, by Corollary 3.3.7:

$$\mathbb{P}\left(|\mathrm{d}(0,v) - r_0| > k\right) \leq \frac{|\mathrm{B}^{\mathrm{C}}_{(r_0+k)}(0)| + |\mathrm{B}_{(r_0-k)}(0)|}{n}$$
$$\leq \frac{1}{(1+c)^{k-1}} = (1+c)\mathrm{e}^{-n\ln(1+c)}.$$

□

We did not compute the exact value of r_o, but we only know its existence. We can have more information about r_0 by using the proof of the first part of Corollary 3.3.7:

$$|\mathrm{B}_{r_0}(0)| \geq (1+c)^{r_0};$$

hence we must have $(1+c)^{r_0} \leq \frac{n}{2}$, i.e. $r_0 \leq \frac{\ln\frac{n}{2}}{\ln(1+c)}$.

3.4 Concentration on Gromov Hyperbolic groups and graphs

In this section, we fix a Hyperbolic group Γ and one of its undirected Cayley graph G. Then there exists a constant $\delta > 0$ such that G, endowed with the graph metric, is δ-Hyperbolic.

We are going to study the concentration of distance phenomenon on the sequence of spheres $[\mathcal{S}_R(I)]_R$ and on the sequence of balls $[\mathcal{B}_R(I)]_R$ of G, where I is the neutral element of Γ.

Let us start by recalling some useful notions on Gromov Hyperbolic metric spaces and Gromov Hyperbolic groups, all of them can be found in Druţu and Kapovich (2018).

Definition 3.4.1. Let (X, d) be a geodesic metric space and let $\delta > 0$ be fixed. We say that a geodesic triangle $\mathcal{T}$ of X is δ-slim if for any ordering (S_1, S_2, S_3) of its sides and for any $x \in S_1$, we have:

$$\mathrm{d}(x, S_2 \cup S_3) \leq \delta.$$

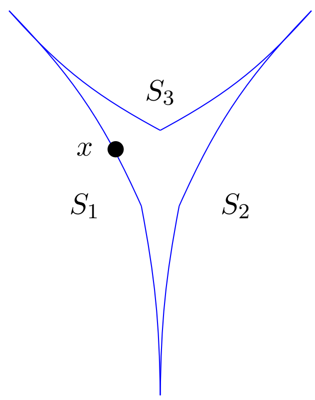

Figure 3.4: $x \in S_1$ is within distance less than δ from $S_2 \cup S_3$

We say that (X, d) is a Hyperbolic metric space if there exists $\delta > 0$ such that any geodesic triangle $\mathcal{T}$ on X is δ-slim; the metric space (X, d) is then called δ-Hyperbolic.

Now, let Γ be a finitely generated group; then we can have a Cayley graph of G with respect to a generating set S of G. So we also have the following definition for Γ.

Definition 3.4.2. We say that Γ is Hyperbolic if the Cayley graph of Γ with respect to a generating set S of Γ is a Hyperbolic metric space with the graph distance.

If in addition Γ is a finite extension of $\mathbb{Z}$, then we say that Γ is an elementary Hyperbolic group.

Note that a Cayley graph of Γ depends on a generating set S, however it is proven that Hyperbolicity is a geometric property of the group, i.e Γ is Hyperbolic if and only if any Cayley graph of Γ is a Hyperbolic metric space.

Balls and spheres in the d-regular tree $\mathbb{T}_d$

An example of 0-Hyperbolic metric space is given by an infinite d-regular tree $\mathbb{T}_d$ where d is a positive integer.

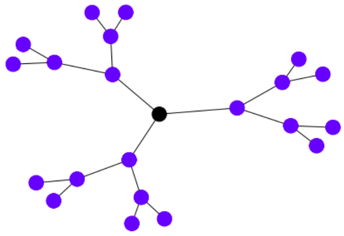

Figure 3.5: Ball $B_3(\mathbb{T}_3)$ of radius 3 of $\mathbb{T}_3$

Fix a node 0 of $\mathbb{T}_d$ and given a positive integer n, then we can compute directly the distribution of the distance $\mathrm{d}(0, X)$ by counting, where X is uniformly chosen in the ball of radius n centered at 0. We have the following data:

k	$\#\{X \in B_n(\mathbb{T}_d) : \ n - \mathrm{d}(0, X) = k\}$
0	$d(d-1)^{n-1}$
1	$d(d-1)^{n-2}$
...	...
i	$d(d-1)^{n-(i+1)}$
...	...
$n-2$	d(d-1)
$n-1$	d
n	1
Total	$1 + d\frac{(d-1)^n - 1}{d-2}$

For $i \in \{0, 1, ..., n\}$, we have:

$$\mathbb{P}(n - \mathrm{d}(0, X) = i) = \frac{d(d-1)^{n-(i+1)}}{1 + d\frac{(d-1)^n - 1}{d-2}} \longrightarrow \frac{d-2}{(d-1)^{i+1}}$$

when n goes to the infinity.

Since $\sum_{i\geq 0} \frac{d-2}{(d-1)^{i+1}} = 1$, we have a limit in distribution for the random variable $n - \mathrm{d}(0,X)$. Furthermore, we have the following concentration inequality:

Theorem 3.4.3. *For $k \in \mathbb{N}$, we have:*

$$\mathbb{P}(n - \rho(0,X) \geq k) \leq 2d\mathrm{e}^{-(k+1)}.$$

Hence, we also have the limit in probability:

$$\frac{\mathrm{d}(o,X)}{n} \xrightarrow{\mathbb{P}} 1.$$

Proof. We have:

$$\begin{aligned}
\mathbb{P}(n - \mathrm{d}(0,X) \geq k) &= \sum_{i=k}^{n} \frac{d(d-1)^{n-(i+1)}}{1 + d\frac{(d-1)^n - 1}{d-2}} \\
&\leq \sum_{i=k}^{n} \frac{d(d-1)^{n-(i+1)}}{d\frac{(d-1)^n - 1}{d-2}} = (d-2)\sum_{i=1}^{n} \frac{(d-1)^{n-(i+1)}}{(d-1)^n - 1} \\
&= (d-2)\frac{(d-1)^n}{(d-1)^n - 1}\sum_{i=k}^{n} \frac{1}{(d-1)^{i+1}} \leq 2(d-2)\left(\frac{1}{(d-1)^{k+1}} \frac{1 - \frac{1}{(d-1)^{n-(k+1)}}}{1 - \frac{1}{d-1}}\right) \\
&\leq 2(d-2)\frac{1}{1 - \frac{1}{d-1}} \frac{1}{(d-1)^{k+1}} = 2(d-1)\frac{1}{(d-1)^{k+1}} \\
&= 2(d-1)\mathrm{e}^{-(k+1)\ln(d-1)} \leq 2d\mathrm{e}^{-(k+1)}
\end{aligned}$$

□

Likewise, let us fix a node v_0 in the sphere of radius n $\mathcal{S}_n(\mathbb{T}_d)$, endowed with the induced graph metric and the uniform probability measure. Let X be an uniformly chosen random point on $\mathcal{S}_n(\mathbb{T}_d)$. We have $\mathrm{d}(v_0,X) \in \{2k : k = 0,...,n\}$ and like in the case of $\mathcal{B}_n(\mathbb{T}_d)$, the following data describes the distribution of $2n - \mathrm{d}(v_0,X)$:

k	$\#\{X \in \mathcal{S}_n(\mathbb{T}_d) : 2n - \mathrm{d}(v_0,X) = k\}$
0	$(d-1)^n$
2	$(d-2)(d-1)^{n-2}$
...	...
$2k$	$(d-2)(d-1)^{n-(k+1)}$
...	...
$2(n-2)$	$(d-2)(d-1)$
$2(n-1)$	$(d-2)$
$2n$	1
Total	$d(d-1)^{n-1}$

Theorem 3.4.4. *Let X be a uniformly randomly chosen vertex of $\mathcal{S}_n(\mathbb{T}_d)$. Then:*

$$\mathbb{P}\left(2n - \mathrm{d}(v_0,X) \geq 2k\right) \leq \mathrm{e}^{-(k-1)\ln(d-1)},$$

where $v_0 \in \mathcal{S}_n(\mathbb{T}_d)$ is fixed. Hence

$$\frac{\mathrm{d}(v_0,X)}{2n} \xrightarrow{\mathbb{P}} 1.$$

Proof. For $i = 1, ..., n-1$, we have $\mathbb{P}\left(2n - \mathrm{d}(0, X) = 2i\right) = \frac{d-2}{d(d-1)^i}$. Hence, for $k > 0$, we have:

$$\begin{aligned} \mathbb{P}\left(2n - \mathrm{d}(v_0, X) \geq 2k\right) &= \sum_{i=k}^{n} \frac{d-2}{d(d-1)^i} \\ &= \frac{d-2}{d(d-1)^k} \sum_{i=0}^{n-k} \frac{1}{(d-1)^i} \leq \frac{d-2}{d(d-1)^k} \sum_{i=0}^{\infty} \frac{1}{(d-1)^i} \\ &= \frac{1}{(d-1)^{k-1}} = \mathrm{e}^{-(k-1)\ln(d-1)}. \end{aligned}$$

□

Non-elementary Hyperbolic groups and Transitive Hyperbolic graphs

Let us now come back to the general notion of Hyperbolicity. The point here is that balls or spheres on Non-elementary Hyperbolic (we consider the metric on a fixed Cayley graph) groups have some common properties with those of d-regular trees: both the sphere and the ball have exponential growth. Furthermore, both Non-elementary Hyperbolic groups and Transitive Hyperbolic graphs are non-amenable.

As specified in the beginning of this section, we have a Non-elementary Hyperbolic group Γ and we fix an undirected Cayley graph G of Γ which is δ-Hyperbolic. Since, G is a Cayley graph, then G is also transitive.

We denote by I the identity element of Γ. For $X, Y \in \Gamma = V(G)$ and for $R > 0$, we use the following notations:

- $[X, Y]$: a geodesic joining X and Y,

- $\mathcal{B}_R(X)$: ball of radius R centered at X; we only write $\mathcal{B}_R$ if $X = I$. Note that we always have $|\mathcal{B}_R| \leq |\mathcal{B}_k|\ |\mathcal{B}_{R-k}|$ for $0 \leq k \leq R$ for non-elementary Hyperbolic groups and for hyperbolic transitive graphs with exponential growth.

- $\mathcal{S}_R(X)$: sphere of radius R centered at X; and again we omit X if $X = I$,

- $f(x) << g(x)$ (or equivalently $f(x) = o(g(x))$ means $\lim\limits_{x \longrightarrow \infty} \frac{f(x)}{g(x)} = 0$,

- Let $X, Y \in \mathcal{B}_R$. A point $u = u_{XY} \in \mathcal{B}_R$ is called a δ-*almost-common-point* of X and Y if there exist two geodesics $[I, X]$ and $[I, Y]$ such that

$$\mathcal{B}_{\frac{\delta}{2}}(u) \cap [I, X] \neq \emptyset \text{ and } \mathcal{B}_{\frac{\delta}{2}}(u) \cap [I, Y] \neq \emptyset.$$

We denote by s_{XY} the maximum of all $s \in [0, R]$ such that the sphere $\mathcal{S}_s$ contains a δ-*almost-common-point* of X and Y (Figure 3.6).

We are going to prove that the distance between two uniformly chosen random points $X, Y \in \mathcal{B}_R$ (or $\mathcal{S}_R$) satisfies

$$\frac{\mathrm{d}(X, Y)}{2R} \xrightarrow{\mathbb{P}} 1.$$

Let us first prove the following Lemma.

Lemma 3.4.5. Let $0 < c < 1$ be a constant and let $R > 0$. Then

$$\sum_{k=cR}^{R} |\mathcal{B}_k||\mathcal{B}_{R-k}|^2 << |\mathcal{B}_R|^2. \quad .$$

Proof. The idea of the proof here is to give an upper bound for $|\mathcal{B}_k||\mathcal{B}_{R-k}|$ where $k \in [cR, R]$. Given R, we can always find $r = r_R \in [cR, R]$ such that

$$|\mathcal{B}_k||\mathcal{B}_{R-k}| \leq |\mathcal{B}_r||\mathcal{B}_{R-r}|$$

for all $k \in [cR, R]$.

Due to the growth rate, we have: $|\mathcal{B}_R| \leq |\mathcal{B}_k|\ |\mathcal{B}_{R-k}|$ for $cR \leq k \leq R$. Hence, Fekete's sub additive Lemma implies that:

$$\lim_{R\longrightarrow\infty} \frac{\ln|\mathcal{B}_R|}{R} = \mathrm{Inf}_R\left(\frac{\ln|\mathcal{B}_R|}{R}\right) = \mu > 0.$$

Hence there exists a sequence of positive numbers $(\epsilon_R)_R$ converging to 0 such that $\frac{\ln|\mathcal{B}_R|}{R} = \mu + \epsilon_R$ for all R; i.e:

$$|\mathcal{B}_R| = \mathrm{e}^{R[\mu+\epsilon_R]}$$

for all R.

Since since $\epsilon_R \geq 0$, we have:

$$\begin{aligned}
|\mathcal{B}_k|\ |\mathcal{B}_{R-k}| &= \mathrm{e}^{k[\mu+\epsilon_k]}\ \mathrm{e}^{(R-k)[\mu+\epsilon_{R-k}]} \\
&= \mathrm{e}^{R\mu}\ \mathrm{e}^{k\epsilon_k+(R-k)\epsilon_{R-k}} \\
&\leq |\mathcal{B}_R|\ \ \exp[k\epsilon_k + (R-k)\epsilon_{R-k} - R\epsilon_R]
\end{aligned}$$

By taking $r = r(R) \in [cR, R]$ such that:

$$\max\{k\epsilon_k + (R-k)\epsilon_{R-k} - R\epsilon_R :\ k \in [cR, R]\} = r\epsilon_r + (R-r)\epsilon_{R-r} - R\epsilon_R,$$

we have:

$$\begin{aligned}
\sum_{k=cR}^{R} |\mathcal{B}_k||\mathcal{B}_{R-k}|^2 &\leq |\mathcal{B}_R| \exp\left[r\epsilon_r + (R-r)\epsilon_{R-r} - R\epsilon_R\right] \sum_{k=cR}^{R} |\mathcal{B}_{R-k}| \\
&\leq \alpha \exp\left[r\epsilon_r + (R-r)\epsilon_{R-r} - R\epsilon_R\right]\ |\mathcal{B}_{R(1-c)}|\ |\mathcal{B}_R|
\end{aligned}$$

where the constant α is from the fact that G has exponential growth.

Claim: $\lim_{R\longrightarrow\infty} \frac{\exp[r\epsilon_r+(R-r)\epsilon_{R-r}-R\epsilon_R]\ |\mathcal{B}_{R-cR}|}{|\mathcal{B}_R|} = 0.$

Indeed, since $(\epsilon_i)_i$ converges to 0 and since $r \leq R$, we have

$$r\epsilon_r + (R-r)\epsilon_{R-r} - R\epsilon_R\ = o(R)$$

and

$$|\mathcal{B}_{R-cR}| = \exp[(R-cR)\mu + o(R)].$$

Hence:

$$\frac{\exp\left[r\epsilon_r + (R-r)\epsilon_{R-r} - R\epsilon_R\right]\ |\mathcal{B}_{R-cR}|}{|\mathcal{B}_R|} = \frac{1}{\exp[\mathrm{R}\mu - (\mathrm{R}-\mathrm{cR})\mu - \mathrm{o(R)}]}$$
$$= \frac{1}{\exp[\mathrm{R}\mu(\mathrm{c} - \mathrm{o}(1))]} \longrightarrow 0$$

when R goes to the infinity since $c > 0$.

Therefore we have $\sum_{k=cR}^{R} |\mathcal{B}_k||\mathcal{B}_{R-k}|^2 << |\mathcal{B}_R|^2$ and the Lemma is proved.

□

We have the following proposition for δ-*almost-common-points*.

Proposition 3.4.6. Let R be positive integer and let $0 < c < 1$ be a constant. Then with the above notations, we have:

$$\#\{(X,Y) \in \mathcal{B}_R \times \mathcal{B}_R :\ s_{XY} > cR\}\ <<\ |\mathcal{B}_R|^2.$$

Proof. Assume that (X,Y) has a δ-*almost-common-points* u with $k = \mathrm{d}(I,u) \geq cR$. Then there are at most $|\mathcal{B}_k|$ possibilities for such u. By definition of u, there exists $u_X, u_Y \in \mathcal{B}_{\frac{\delta}{2}}(u)$ such that $u_X \in [I,X]$ and $u_Y \in [I,Y]$ (Figure 3.6).

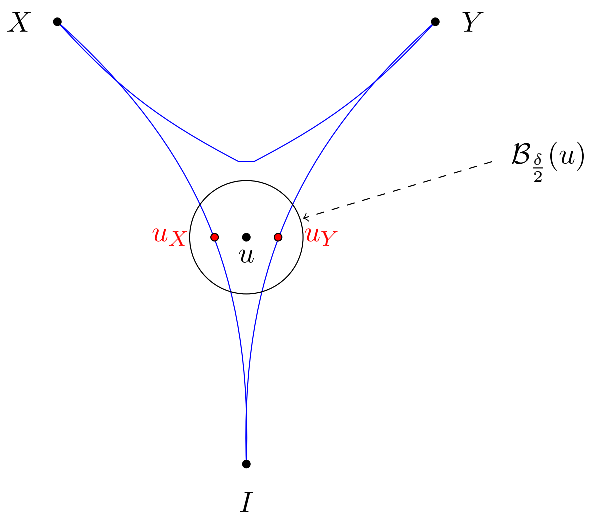

Figure 3.6: $\mathrm{d}(I,u) = s_{XY}$

By the triangular inequality, we have:

$$\mathrm{d}(I,u_X) \geq \mathrm{d}(I,u) - \mathrm{d}(u,u_X) \geq k - \frac{\delta}{2},$$

and

$$\mathrm{d}(I,u_Y) \geq \mathrm{d}(I,u) - \mathrm{d}(u,u_Y) \geq k - \frac{\delta}{2},$$

because $\mathrm{d}(u,u_X) \leq \frac{\delta}{2}$ and $\mathrm{d}(u,u_Y) \leq \frac{\delta}{2}$.

Since $u_X \in [I, X]$ and $u_Y \in [I, Y]$ we have:

$$d(u_X, X) = d(I, X) - d(I, u_X) \leq R - k + \frac{\delta}{2}$$

and

$$d(u_Y, Y) = d(I, Y) - d(I, u_Y) \leq R - k + \frac{\delta}{2}.$$

By transitivity, we have the following possibilities given u:

(i) there are at most $|\mathcal{B}_{\frac{\delta}{2}}(u)| = |\mathcal{B}_{\frac{\delta}{2}}|$ possibilities for both u_X and v_X,

(ii) there are at most $|\mathrm{B}_{R-k+\frac{\delta}{2}}(u_X)| = |\mathrm{B}_{R-k+\frac{\delta}{2}}(u_Y)| = |\mathrm{B}_{R-k+\frac{\delta}{2}}|$ possibilities for both X and Y.

In sum, we have the following bound:

$$\#\{(X, Y) \in \mathcal{B}_R \times \mathcal{B}_R : \ s_{XY} > cR\} \leq |\mathcal{B}_{\frac{\delta}{2}}|^2 \sum_{k=cR}^{R} |\mathcal{B}_k||\mathcal{B}_{R-k+\frac{\delta}{2}}|^2$$

By Lemma 3.4.5, we have:

$$|\mathcal{B}_{\frac{\delta}{2}}|^2 \sum_{k=cR}^{R} |\mathcal{B}_k||\mathcal{B}_{R-k+\frac{\delta}{2}}|^2 << |\mathcal{B}_{R+\frac{\delta}{2}}|^2.$$

Since δ is a constant, we have the result due to the fact that G has exponential growth.

□

We have the following corollary:

Corollary 3.4.7. Let X, Y be uniformly chosen random points of $\mathcal{B}_R$ and let $0 < c < 1$. Then *with high probability*, there exist two geodesics $[I, X]$, $[I, Y]$ and a point $v \in [I, X]$ located within distance cR from I such that $d(v, [I, Y]) > \delta$.

Proof. By Proposition 3.4.6, if r is the highest possible distance between I of a δ-*almost-common-point* of any geodesics $[I, X]$ and $[I, Y]$, then we have $r \leq cR$ with high probability. Then, with high probability, a point $v \in [I, X]$ within distance cR from I has distance more than δ from $[I, Y]$. □

The following Theorem combines δ-Hyperbolicity, Proposition 3.4.6 and Corollary 3.4.7.

Theorem 3.4.8. *Let X, Y be uniformly chosen random points of $\mathcal{B}_R$ and let $0 < c < 1$. Then:*

$$\lim_{R \longrightarrow \infty} \mathbb{P}[2R - 2cR - 2\delta \leq d(X, Y) \leq 2R] \longrightarrow 1,$$

i.e.

$$\lim_{R \longrightarrow \infty} \mathbb{P}\left[1 - \frac{d(X, Y)}{2R} \leq c + \frac{2\delta}{2R}\right] \longrightarrow 1.$$

Proof. With high probability, we have $[I, X]$, $[I, Y]$ and $v \in [I, X]$ defined as in Corollary 3.4.7. By Hyperbolicity, there exist a node $u = u_\delta \in [X, Y]$ such that $\mathrm{d}(v, u) \leq \delta$ ($[X, Y]$ is any geodesic). By triangular inequality, we have:

$$\begin{aligned} \mathrm{d}(X, u) &\geq \mathrm{d}(X, I) - \mathrm{d}(I, u) \\ &\geq \mathrm{d}(X, I) - \mathrm{d}(I, v) - \mathrm{d}(v, u) \\ &\geq R - cR - \delta; \end{aligned}$$

likewise, we have:

$$\begin{aligned} \mathrm{d}(Y, u) &\geq \mathrm{d}(Y, I) - \mathrm{d}(I, u) \\ &\geq \mathrm{d}(Y, I) - \mathrm{d}(I, v) - \mathrm{d}(v, u) \\ &\geq R - cR - \delta, \end{aligned}$$

where C is a constant. Since $u \in [X, Y]$, then

$$\mathrm{d}(X, Y) = \mathrm{d}(Y, u) + \mathrm{d}(X, u) \geq 2R - 2cR - 2\delta$$

with high probability. □

Since G has exponential growth, we can find a constant C such that $|\mathcal{B}_R \leq C|\mathcal{S}_R|$ for some all R, then the last result in Theorem 3.4.8 is also true for $\mathcal{S}_R$.

3.5 Final remark and open questions

We proved that the balls (or spheres) of non elementary Hyperbolic groups have distance concentration property by using the the 'Hyperbolicity' and the 'exponential growth'. One can now asks how is it in the gap between Hyperbolicity and non-Hyperbolicity or the gap between exponential growth and sub-exponential growth. We then can ask the following open questions:

- Are there transitive graphs of exponential growth, but non-hyperbolic, where the distance in the balls or spheres is still concentrated?
- Are there transitive graphs of exponential growth, but non-hyperbolic, where the distance in the balls or spheres is non-concentrated?
- Is it true that there is always no concentration for the spheres or balls on transitive graphs of sub-exponential growth?

For example, the Descartes product $\mathbb{T}_k \times \mathbb{Z}$ and the Diestel-Leader graph $\mathrm{DL}(2, 2)$ (Lyons and Peres (2017)) are transitive graphs of exponential growth but they are not Hyperbolic.

4. Visualizations and Application

In Chapter 3, we investigated the concentration of distance on several examples of graphs. In this chapter, we visualize the Rayleigh limit in Theorem 3.2.5 with python, and we try to explore similar phenomena for UST of other graphs via visualization. We end this chapter by presenting an example of application of the concentration of distance.

From Figure 4.3 and 4.4, we can conclude that the distance in the UST of $(\mathbb{Z}/n\mathbb{Z})^2$ and $(\mathbb{Z}/n\mathbb{Z})^5$ are not concentrated. Then a question can be asked here whether it is true that the distance is always non-concentrated in the UST of any transitive graph?

4.1 Visualization in Python

From the previous chapter, we showed the Rayleigh limit for the distance in the UST of the complete graph K_n. We are going to visualize this result in this chapter and we try to explore similar phenomena on other networks. We generate the corresponding data and plot the approximated Probability Density Function in Python.

By Wilson's algorithm, the output of the processes (i) and (ii) below have the same distribution given a connected Graph G:

(i) take two random elements $u, v \in V(G)$ uniformly at a random, take a UST $\mathcal{T}$ of G and take the distance $\mathrm{d}_{\mathcal{T}}(u, v)$.

(ii) take two random $u, v \in V(G)$ uniformly at a random, run a Loop Erased Random Walk from u to v and we then have a random path $\mathrm{LERW}(u, v)$, then take the length of $\mathrm{LERW}(u, v)$.

The Python code below load all the usual packages of this chapter. The Python function $\mathrm{LERW}(G, u, v)$ is an implementation of the process (ii) discussed above.

```
"""@author: Mahefa"""
# usefull package
import networkx as nx
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import random

def LERW(G,u,v):
    #start a LERW at u and stop when hitting v
    path = [u]


    # we continue updating the path until v becomes its last element
    while v != path[len(path)-1]:

        # take a random neighbour of the last element of the path
        x = random.choice(list(nx.Graph.neighbors(G,path[len(path)-1])))
```

```
    #if the neighbour above is already in the path (i.e. a loop is obtained),
        #then reduce the path up to that node
        if x in path:
            path = path[0:path.index(x)+1]

        # if the random neighbour is not yet in the path, then append it
        else:
            path = path + [x]

    return(len(path)-1) # path is a sequence of nodes
```

-Uniform Spanning Tree of the complete graph K_n

For the complete graph K_n, let us consider the cases $n = 50$, $n = 100$ and $n = 1000$. The following code generates 400 independent trials of $\frac{\text{Length}[\text{LERW}(G,u,v)]}{\sqrt{n}}$, where LERW is the function in the code above. The output of the plot is shown in figure 4.1.

```
x= np.linspace(0,5,400) # number of data to generate
# we do the visualisation for n= 50, 100 and 1000
K_50 = nx.complete_graph(50)
K_100 = nx.complete_graph(100)
K_1000 = nx.complete_graph(1000)
# take 400 independent trials for each graph
# take two random nodes and save the length of a LERW between them

#n=50
data_50 = np.array([LERW(K_50,random.choice(list(K_50.node())),
random.choice(list(K_50.node()))) for i in range(400)])/np.sqrt(50)

#n=100
data_100 = np.array([LERW(K_100,random.choice(list(K_100.node())),
random.choice(list(K_100.node()))) for i in range(400)])/np.sqrt(100)

#n=1000
data_1000 = np.array([LERW(K_1000,random.choice(list(K_1000.node())),
random.choice(list(K_1000.node()))) for i in range(400)])/np.sqrt(1000)




#Plot with the Rayleigh Probability Density Function
fig, axes = plt.subplots(ncols=2, nrows=1, figsize = (20,5))#construct subplots
ax1, ax2= axes.ravel() # name of each subplot

#histogram
ax1.hist(data_50,bins=50,color='orange', label = r'$K_{50}$')
ax1.hist(data_100,bins=50,color='green', label = r'$K_{100}$')
ax1.hist(data_1000,bins=50, color='red', label = r'$K_{1000}$')
ax1.set_xlabel(r'$\frac{LERW}{\sqrt{n}}$')
```

```
ax1.set_title('Histogram')
ax1.legend()

#estimated density vs Rayleigh density
ax2.plot(x,x*np.exp(-x**2/2), label = 'Rayleigh PDF')
pd.Series(data_50).plot.kde(label = r'$K_{50}$')
pd.Series(data_100).plot.kde(label = r'$K_{100}$')
pd.Series(data_1000).plot.kde(label = r'$K_{1000}$')
ax2.set_xlabel(r'$\frac{LERW}{\sqrt{n}}$')
ax2.set_xlim([0,6])
ax2.legend()
plt.show()
```

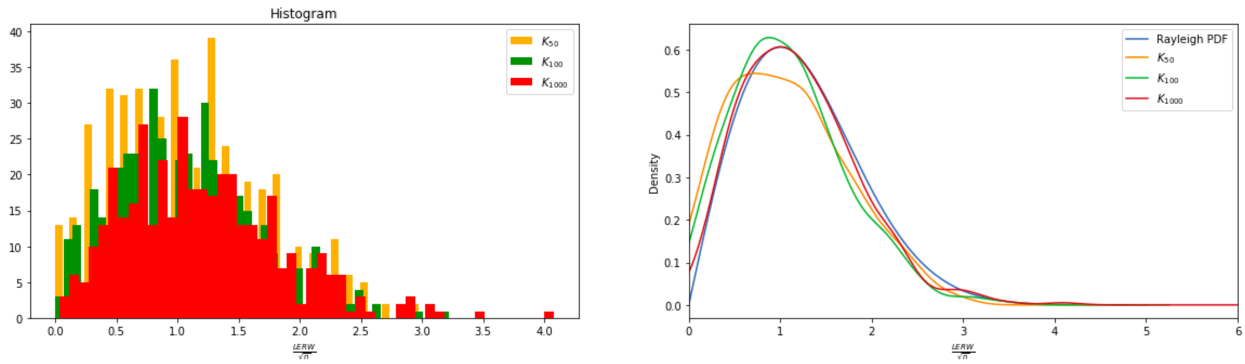

Figure 4.1: Histogram and approximated density

Figure 4.1 suggests that the approximated densities converge to the Rayleigh density. This result is proved in Theorem 3.2.5 of Chapter 3 and it is mentioned in Peres and Revelle (2004).

-Uniform Spanning Tree of the square lattice $G = (\mathbb{Z}/n\mathbb{Z})^2$ with $n = 20, 30, 50$

Note that all we do in this section is just visualisations and and there are no mathematical proofs: we just take conclusions from the observed data.

Let us first proceed as in the UST of K_n, where we divide the length of an LERW by the square root of the size of our graph, here $|(\mathbb{Z}/n\mathbb{Z})^2| \sim n^2$. The following code generates 400 independent trials of $\frac{\text{Length}[\text{LERW}(G,u,v)]}{\sqrt{n^2}}$.

```
#Generate Data
data_20 = np.array([LERW(G_20,random.choice(list(G_20.node())),
     random.choice(list(G_20.node()))) for i in range(400)])/20

data_30 = np.array([LERW(G_30,random.choice(list(G_30.node())),
     random.choice(list(G_30.node()))) for i in range(400)])/30

data_50 = np.array([LERW(G_50,random.choice(list(G_50.node())),
     random.choice(list(G_50.node()))) for i in range(400)])/50

#Plot estimated density
```

```
fig, axes = plt.subplots(ncols=2, nrows=1, figsize = (20,5)) # construct subplots
ax1, ax2= axes.ravel() # name of each subplot

#histogram
ax1.hist(data_20, bins=100, label = r'$d=2, n=20$')
ax1.hist(data_30,bins=100, color='orange',label=r'$d=2,n=30$')
ax1.hist(data_50, bins=100, color='green',label=r'$d=2,n=50$')
ax1.set_xlabel(r'$\frac{LERW}{\sqrt{n^2}}$')
ax1.set_title('Histogram')
ax1.legend()

#estimated density
pd.Series(data_20).plot.kde(label = r'$d=2, n=20$')
pd.Series(data_30).plot.kde(label = r'$d=2,n=30$')
pd.Series(data_50).plot.kde(label = r'$d=2,n=50$')
ax2.legend()
ax2.set_xlim([0,4.5])
ax2.set_xlabel(r'$\frac{LERW}{\sqrt{n^2}}$')
plt.show()
```

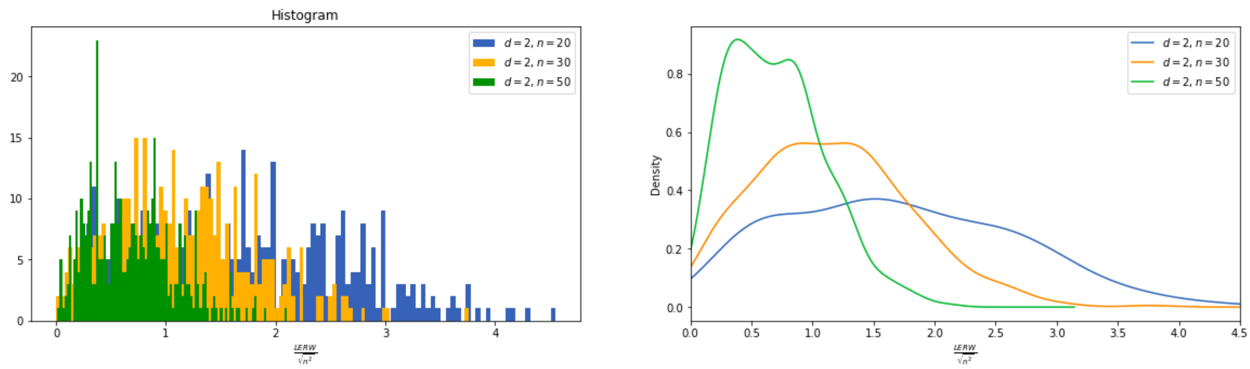

Figure 4.2: Histogram and approximated density

The result in Figure 4.2 suggests that there is no limiting curve for the densities of $\frac{\text{Length}[\text{LERW}(G,u,v)]}{\sqrt{n^2}}$ and maybe this is because we did not chose the right denominator.

Instead of dividing with $\sqrt{n^2}$, let us divide with the coefficient $n^{\frac{5}{4}}$, see Kenyon (2000a), Kenyon (2000b) and Lawler et al. (2011).

```
data_20 = np.array([LERW(G_20,random.choice(list(G_20.node())),
    random.choice(list(G_20.node()))) for i in range(400)])/20**(5/4)

data_30 = np.array([LERW(G_30,random.choice(list(G_30.node())),
    random.choice(list(G_30.node()))) for i in range(400)])/30**(5/4)

data_50 = np.array([LERW(G_50,random.choice(list(G_50.node())),
    random.choice(list(G_50.node()))) for i in range(400)])/50**(5/4)
```

```
#Plot
fig , axes = plt.subplots(ncols=2, nrows=1, figsize = (20,5)) # construct subplots
ax1, ax2= axes.ravel() # name of each subplot

#histogram
ax1.hist(data_20, bins=100, color='orange', label=r'$d=2,␣n=20$')
ax1.hist(data_30,bins=100, color='green', label=r'$d=2,n=30$')
ax1.hist(data_50, bins=100, color='red', label=r'$d=2,n=50$')
ax1.set_xlabel(r'$\frac{LERW}{n^{\frac{5}{4}}}$')
ax1.set_title('Histogram')
ax1.legend()
#approximated density
ax2.plot(x,x*np.exp(-x**2/2), label = 'Rayleigh␣PDF')
pd.Series(data_20).plot.kde(label = r'$d=2,␣n=20$')
pd.Series(data_30).plot.kde(label = r'$d=2,n=30$')
pd.Series(data_50).plot.kde(label = r'$d=2,n=50$')
ax2.set_xlim([0,4])
ax2.set_xlabel(r'$\frac{LERW}{n^{\frac{5}{4}}}$')
ax2.legend()
plt.show()
```

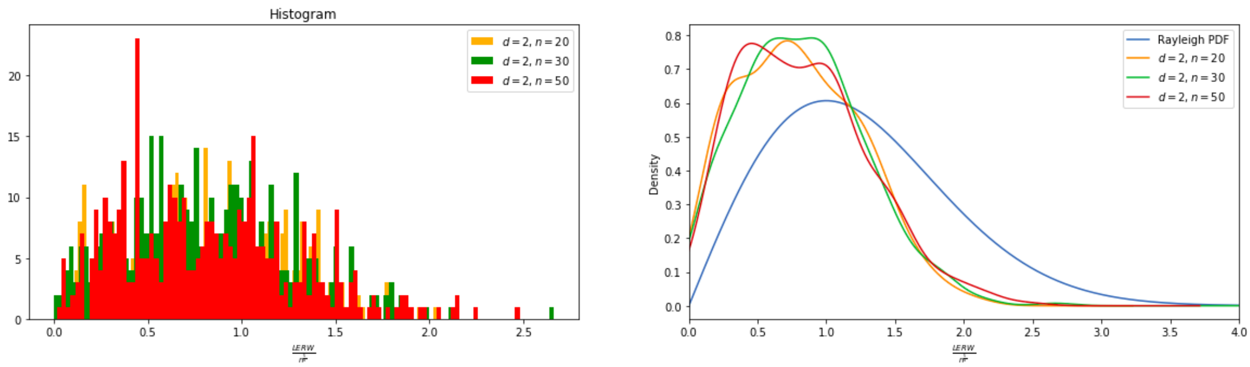

Figure 4.3: Histogram and approximated density

The output in Figure 4.3 suggests that there is a limiting curve for the estimated density distribution, and the limit in distribution is not the Rayleigh.

-Uniform Spanning Tree of the square lattice $G = (\mathbb{Z}/n\mathbb{Z})^5$ with $n = 3, 5, 7$

Let us proceed exactly like in the UST of K_n. We have $|(\mathbb{Z}/n\mathbb{Z})^5| = n^5$ and we are going to simulate the estimated density of $\frac{\text{Length}[\text{LERW}(G,u,v)]}{\sqrt{n^5}}$ for $n = 3, 5$ and 7 in the following code. The coefficient $\sqrt{n^5}$ in the denominator is from Peres and Revelle (2004).

```
G_3 = nx.grid_graph(dim=[3,3, 3,3,3], periodic = False) # (Z/3Z)^5
G_5 = nx.grid_graph(dim=[5,5, 5,5,5], periodic = False) # (Z/5Z)^5
G_7 = nx.grid_graph(dim=[7,7, 7,7,7], periodic = False) # (Z/7Z)^5

# Data for UST of the discrete torus Z_n^d with d=5 and n=4,5,7
```

```
data_3 = np.array([LERW(G_3,random.choice(list(G_3.node())),
    random.choice(list(G_3.node()))) for i in range(400)])/3**(5/2) # n=3

data_5 = np.array([LERW(G_5,random.choice(list(G_5.node())),
    random.choice(list(G_5.node()))) for i in range(400)])/5**(5/2) # n=5

data_7 = np.array([LERW(G_7,random.choice(list(G_7.node())),
    random.choice(list(G_7.node()))) for i in range(400)])/7**(5/2) # n=7


x= np.linspace(0,5,400)
fig, axes = plt.subplots(ncols=2, nrows=1, figsize = (20,5))#construct subplots
ax1, ax2= axes.ravel() # name of each subplot

#histogram
ax1.hist(data_2, bins=100,color='orange', label = r'$d=5, n=3$')
ax1.hist(data_3,bins=100, color='green',label=r'$d=5,n=5$')
ax1.hist(data_5, bins=100, color='red',label=r'$d=5,n=7$')
ax1.set_xlabel(r'$\frac{LERW}{\sqrt{n^5}}$')
ax1.set_title('Histogram')
ax1.legend()


#Plot the corresponding approximated Probability Density Function
ax2.plot(x,x*np.exp(-x**2/2), label = 'Rayleigh PDF')
pd.Series(data_3).plot.kde(label = r'$d=5,\; n=3$')
pd.Series(data_5).plot.kde(label = r'$d=5,\;n=5$')
pd.Series(data_7).plot.kde(label = r'$d=5,\; n=7$')
ax2.set_xlabel(r'$\frac{LERW}{n^{5/2}}$')
ax2.set_xlim([0,6])
ax2.legend()
plt.show()
```

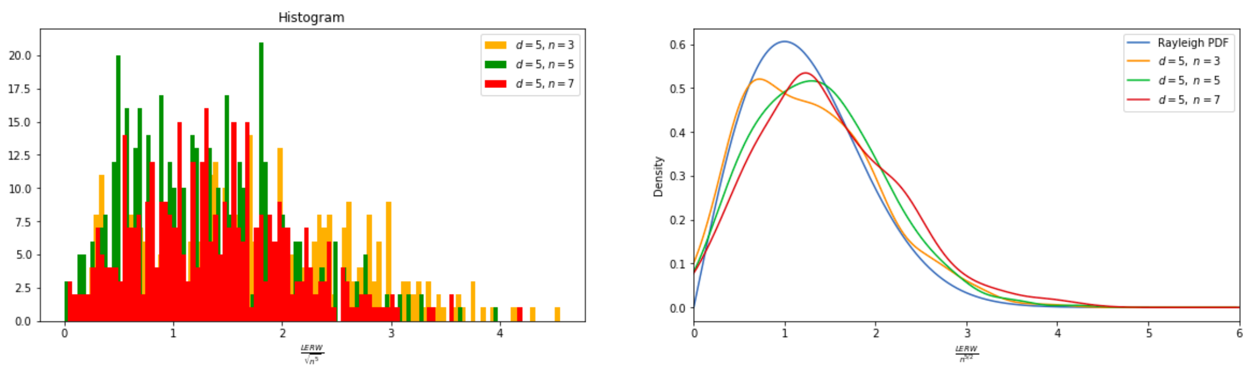

Figure 4.4: The limiting curve should be around the Rayleigh density

The output in Figure 4.4 suggests that there is a limiting curve, and that limiting curve is around the Rayleigh density.

4.2 Single source distance vs all pairs distance in transitive metric spaces

The idea in this section is from the second part of Roughgarden et al. (2019). We say that a metric space X is transitive if the natural action of its automorphism group is transitive. An example of transitive metric space is given by the sphere $\mathcal{S}_n(\mathbb{T}_d)$ endowed with the metric induced by the graph structure of $B_n(\mathbb{T}_d)$.

Now, consider the second result of Section 3.4: fix an element $o \in \mathcal{S}_n(\mathbb{T}_d)$ and let $x, y \in \mathcal{S}_n(\mathbb{T}_d)$ be uniformly chosen. Since $\mathcal{S}_n(\mathbb{T}_d)$ is transitive, then there exists $\rho \in \mathrm{Aut}(\mathcal{S}_n(\mathbb{T}_d))$ such that $\rho(x) = o$. Therefore, By Theorem 3.4.4, all the quantities

$$\frac{\mathrm{d}(o,x)}{2n}, \ \frac{\mathrm{d}(o,y)}{2n} \ \text{and} \ \frac{\mathrm{d}(x,y)}{2n} = \frac{\mathrm{d}(o,\rho(y))}{2n}$$

converge in probability to 1 as n goes to the infinity.

Hence, we also have the convergence:

$$\frac{d(o,x)+d(o,y)}{d(x,y)} = \frac{\frac{d(o,x)}{2n} + \frac{d(o,x)}{2n}}{\frac{d(x,y)}{2n}} \longrightarrow 2 \ \text{ in probability as } n \longrightarrow \infty.$$

Hence, saving a data structure that encodes the family of all pairs distance $[\mathrm{d}(x,y)]_{x,y\in X}$ can be reduced in saving a data structure of a family of single source distance $[\mathrm{d}(o,y)]_{x\in X}$.

The following Python code computes some samples of the stretch for $\mathcal{S}_{10}(\mathbb{T}_3)$ and plots the corresponding histogram and the approximated distribution. The plot is shown in Figure 4.5

```
#create the ball of radius n (n=2) in a d-regular tree (3<=d)
#by the function ball_d_r
def ball_d_r(degree=3, radius=2):
    while degree > 2:
          T = nx.disjoint_union(nx.balanced_tree(degree-1, radius),
          nx.balanced_tree(degree-1, radius-1))

          roots = [i for i in list(T.nodes()) if T.degree(i)==degree-1]
          T.add_edge(roots[0], roots[1])
          return(T)

# single source distance and all pairs distance for the sphere
def distance(T):
    #Shpere consists on nodes of degree 1
    leaf = [i for i in T.nodes() if nx.degree(T,i)==1]

    o = leaf[0] # the sphere is transitive, hence we can fix one node
    u = np.random.choice(leaf)
    v = np.random.choice(leaf) # we need u!=v in order to see the strech
```

```
    d_0 = nx.shortest_path_length(T, o,u) + nx.shortest_path_length(T, o, v)
    d = nx.shortest_path_length(T,u,v)
    return([d_0,d])


#generate data for visualization: case d=3 and r=10
T = ball_d_r(3,10)
Dist = [distance(T) for i in range(3*2**9)]
D_O = np.array([L[0] for L in Dist])
D = np.array([L[1] for L in Dist])
stretch = []# we avoid division by 0
for i in range(len(D)):
    if D[i]==0:
        stretch.append(0)
    else:
        stretch.append(D_O[i]/D[i])

freq = dict(Counter(strech))#frequence

# Plot
fig, axes = plt.subplots(ncols=2, nrows=1,
            figsize = (15,5)) # construct subplots
ax1, ax2= axes.ravel() # name of each subplot

ax1.scatter(freq.keys(), freq.values(),
           marker='o', label = 'Frequency␣of␣the␣stretch')
ax1.set_xlabel('stretch')
ax1.set_xlim([0,5])
ax1.set_ylabel('frequency')
ax1.legend()

pd.Series(strech).plot.kde( color = 'black',
        label = 'Estimated␣density␣of␣the␣stretch')
ax2.axvline(2, color = 'red', label = 'x=2')
ax2.set_xlim([0,5])
ax2.set_xlabel('stretch')
ax2.legend()

plt.show()
```

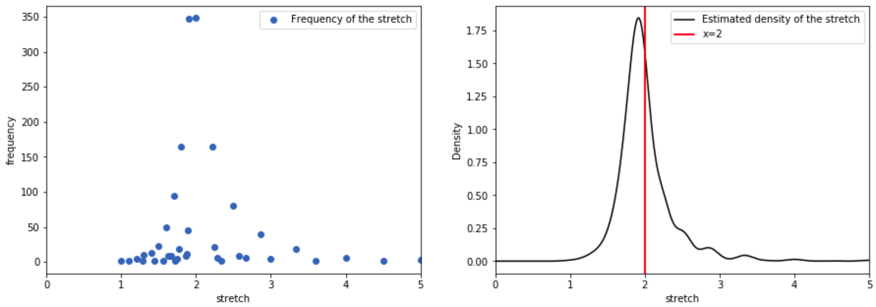

Figure 4.5: Frequency and approximated density of the stretch

References

Druţu, C. and Kapovich, M. *Geometric group theory*, volume 63. American Mathematical Soc., 2018.

Kenyon, R. The asymptotic determinant of the discrete laplacian. *arXiv preprint math-ph/0011042*, 2000a.

Kenyon, R. Long-range properties of spanning trees. *Journal of Mathematical Physics*, 41(3): 1338–1363, 2000b.

Lawler, G. F., Schramm, O., and Werner, W. Conformal invariance of planar loop-erased random walks and uniform spanning trees. In *Selected Works of Oded Schramm*, pages 931–987. Springer, 2011.

Levin, D. A. and Peres, Y. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Lubotzky, A. *Discrete groups, expanding graphs and invariant measures.* Springer Science & Business Media, 2010.

Lyons, R. and Peres, Y. *Probability on trees and networks*, volume 42. Cambridge University Press, 2017.

Peres, Y. and Revelle, D. Scaling limits of the uniform spanning tree and loop-erased random walk on finite graphs. *arXiv preprint math/0410430*, 2004.

Roughgarden, T., Huang, N., Lewis-Pye, Andrew, Li, Angsheng, Li, Xuechen, Pan, Yicheng, and Barmpalias, G. The idemetric property: when most distances are (almost) the same. *Proceedings of the Royal Society A*, 475(2222):20180283, 2019.