Centre de Mathématiques Appliquées, École Polytechnique

Bachelor Thesis

Couplings and Poincaré inequality for Markov chains

Luca Bonengel

supervised by
Professor Giovanni Conforti

October 1, 2021

Abstract

In most applications of Markov chains, a key issue is to quantify the speed of convergence to the invariant distribution. The objective of this report is to explore two fundamental approaches to this problem. The first one, more probabilistic, is based on the notion of coupling and makes use of transport distances, in particular the Wasserstein distance. The second one, more analytical in spirit, establishes exponential convergence in an L^2 setting and leads to a Poincaré inequality. Concrete examples are given throughout the report, including the n-dimensional hypercube, in order to explore the potential applications of the proven results. In the second half of this report, a particular effort will be devoted to finding a lower bound for the Ollivier Ricci curvature of reversible Markov chains. Finally, a Markov chain Monte Carlo method will be studied to approximate invariant distributions.

Contents

1 Introduction **2**

2 Background **3**
- 2.1 Markov chains . . . 3
- 2.2 Convergence to invariant distribution . . . 5

3 Convergence using transport distances **6**
- 3.1 Optimal transport . . . 6
- 3.2 Ollivier's Ricci curvature . . . 7
- 3.3 Examples . . . 7
- 3.4 Wasserstein distance contraction . . . 11
- 3.5 Construction results . . . 12

4 Analytical approach to exponential convergence **14**
- 4.1 Averaging operator . . . 14
- 4.2 Variance contraction . . . 15
- 4.3 Poincaré inequality . . . 17

5 Ollivier's Ricci curvature bound on undirected graphs **18**
- 5.1 Unweighted graphs . . . 19
- 5.2 Weighted graphs . . . 23

6 Markov chain Monte Carlo (MCMC) **27**
- 6.1 Bias of empirical mean . . . 28
- 6.2 Variance of empirical mean . . . 29
- 6.3 Examples . . . 33

7 Conclusion **34**

1 Introduction

Markov chains are stochastic models that describe sequences of events where the probability of moving from one state to the next depends only on the present state. They are widely studied and have widespread applications, ranging from finance to biology. They are named after the Russian mathematician Andrey Markov, who studied them in depth in the early twentieth century.

Questions about rates of convergence of Markov chains are of crucial importance. Such questions are the main focus of this report. After presenting Markov chains as well as some of their important properties in the second section of this paper, their convergence will be studied using the Wasserstein distance W_1, and optimal couplings will be explored. We will firstly focus on the notion of Ollivier's Ricci curvature and see how it is directly linked to contraction rates in the Wasserstein distance. Then, a more analytical approach, leading to a Poincaré inequality, will be adopted to derive a variance contraction result. Algebraic techniques will allow us to obtain exponential contraction in an L^2 setting. Again, in this part, the link with Ollivier's Ricci curvature will be clearly demonstrated.

Reversible Markov chains can be represented as random walks on undirected graphs. As such, we will study transport plans between probability distributions on such graphs so as to derive sharp lower bounds for Ollivier's Ricci curvature. Such bounds will be given for any two neighbors x and y by finding optimal transport plans between the two probability distributions m_x and m_y. For the sake of simplicity and in order to help visualize the geometry, we will firstly work on unweighted graphs before generalizing our results to weighted graphs.

Finally, in the last part of this report, our goal will be to provide an efficient way to approximate $\pi(f) := \int f \, \mathrm{d}\pi$ for a function f defined on some space $\mathcal{X}$ using a Markov chain Monte Carlo method. Our approach will consist in constructing and simulating a Markov chain $(X_k)_{k\in\mathbb{N}}$ on $\mathcal{X}$ with stationary distribution π. The idea will be to wait for a time T_0 so that the chain gets close to its stationary distribution before estimating $\pi(f)$ on the next T steps as follows

$$\hat{\pi}(f) := \frac{1}{T} \sum_{T_0+1}^{T_0+T} f(X_k).$$

We will focus on finding an upper bound for the mean quadratic error $\mathbb{E}_x\left[|\hat{\pi}(f) - \pi(f)|^2\right]$, for some $x \in \mathcal{X}$.

In this report, we will make the assumption that the state space $\mathcal{X}$ is finite, although the main results also apply to any Polish space $(\mathcal{X}, d)$. This report is mainly based on three papers: [5, 1, 3]. Here, a special effort was made to provide much more detailed proofs and examples than in the cited publications. Moreover, several missing proofs from the cited papers are provided in this report (in particular, the proofs of 3.3.2, 4.1.2, 4.3.1, 5.2.1 and 6.2.1).

2 Background

We work throughout with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The material presented in this section largely stems from the first chapter of Norris' book [4].

2.1 Markov chains

A *Markov chain* is a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ with the Markov property, meaning that the probability of moving from one state to the next depends only on the present state and not on the previous states. In other words, as long as we have

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) > 0,$$

a Markov chain verifies

$$\mathbb{P}(X_{n+1} = x \mid X_0 = x_0, X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \mathbb{P}(X_{n+1} = x \mid X_n = x_n).$$

The possible values taken by X_i form a set $\mathcal{X}$ called the state space of the chain.

Definition 2.1.1 (Stochastic matrix)**.** *A stochastic matrix $m = (m_{xy} : x, y \in \mathcal{X})$ is a matrix with non-negative entries such that*

$$\sum_{y\in\mathcal{X}} m_{xy} = 1 \quad \textit{for any } x \in \mathcal{X}.$$

In other words, every row $m_x := (m_{xy} : y \in \mathcal{X})$ with $x \in \mathcal{X}$ of a stochastic matrix is a probability distribution.

Markov chains $(X_n)_{n\in\mathbb{N}}$ are associated with an *initial distribution* λ and a stochastic matrix, called *transition matrix* m as follows

1. X_0 has distribution λ;
2. for all $n \geq 0$, conditional on $X_n = x$, X_{n+1} has distribution m_x, independent of $X_0, \ldots, X_{n-1}$.

More explicitly, these conditions state that, for any $n \geq 0$ and any $x_0, \ldots, x_{n+1} \in \mathcal{X}$,

1. $\mathbb{P}(X_0 = x_0) = \lambda_{x_0}$;
2. $\mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0, X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = m_{x_n x_{n+1}}$.

We say that $(X_n)_{n\in\mathbb{N}}$ is $Markov(\lambda, m)$ for short.

For the sake of convenience, we will sometimes use one of the following equivalent notations

$$\mathbb{P}(X_n = y \mid X_0 = x) = \mathbb{P}_x(X_n = y) = m_{xy}^{(n)} = m_x^n(y),$$

where

$$m_{xy}^{(n)} := \sum_{z\in\mathcal{X}} m_{xz}^{(n-1)}\, m_{zy}, \quad m_x^n(y) := \sum_{z\in\mathcal{X}} m_x^{n-1}(z)\, m_z(y),$$

and of course $m^{(1)}_{xz} := m_{xz}$ and $m^1_x := m_x$. We will use the following notation for the product of a probability distribution μ on $\mathcal{X}$ by the matrix m

$$\mu * m := \sum_{x \in \mathcal{X}} \mu(x)\, m_x,$$

which corresponds to the image of μ by the Markov chain.

Definition 2.1.2 (Hitting time)**.** *The hitting time of a subset A of $\mathcal{X}$ is the random variable $H^A : \Omega \to \{0, 1, 2, \dots\} \cup \{\infty\}$ given by*

$$H^A(\omega) = \min\{n \geq 0 : X_n(\omega) \in A\}$$

where we agree that the minimum of the empty set $\varnothing$ is ∞.

Definition 2.1.3 (Irreducibility)**.** *A Markov chain is said to be irreducible if for any states $x, y \in \mathcal{X}$, there exists $n \in \mathbb{N}$ such that $m^{(n)}_{xy} > 0$.*

Definition 2.1.4 (Transience and recurrence)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with transition matrix m. We say that a state $x \in \mathcal{X}$ is recurrent if*

$$\mathbb{P}_x(X_n = x \text{ for infinitely many } n) = 1.$$

We say that $x \in \mathcal{X}$ is transient if

$$\mathbb{P}_x(X_n = x \text{ for infinitely many } n) = 0.$$

In addition, in the case of recurrence, one can add that if the *expected return time* $\mathbb{E}_x(T_x)$ is finite, x is said to be *positive recurrent*. A recurrent state which fails to have this stronger property is called *null recurrent*.

Definition 2.1.5 (Invariant distribution)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with transition matrix m. A probability distribution $\pi = (\pi_x)_{x\in\mathcal{X}}$ is said to be invariant or stationary if*

$$\pi * m = \pi.$$

Definition 2.1.6 (Reversibility)**.** *An irreducible Markov chain with stationary distribution π and transition probability matrix m verifying*

$$\pi(x)\, m_x(y) = \pi(y)\, m_y(x) \quad \text{for any } x, y \in \mathcal{X}$$

is said to be reversible.

Remark 2.1.1. *Intuitively, reversibility (2.1.6) means that the Markov chain "looks the same" regardless of whether we run it forwards or backwards in time. However, for this to be true, we must start the run from the invariant distribution.*

Definition 2.1.7 (Aperiodic)**.** *A state $x \in \mathcal{X}$ is said to be aperiodic if $m^{(n)}_{xx} > 0$ for all sufficiently large n. m is said to be aperiodic if all states $x \in \mathcal{X}$ are aperiodic.*

In our case, $\mathcal{X}$ is finite. Therefore, we know from [4] that it is impossible for all states $x \in \mathcal{X}$ to be null recurrent. There exists $x \in \mathcal{X}$ that is positive recurrent. Then, Theorem 1.7.7 of Norris' book [4] tells us that if m is irreducible, all states in $\mathcal{X}$ are positive recurrent and there exists a unique invariant distribution π. Thus, irreducible Markov chains on a finite state space $\mathcal{X}$ are positive recurrent and have a unique invariant distribution.

2.2 Convergence to invariant distribution

Now that we have introduced the necessary material and definitions, we investigate the limiting behaviour of $m_{ij}^{(n)}$ as $n \to \infty$ and its relation to the invariant distribution. An equivalent statement is provided in [4].

Theorem 2.2.1 (Convergence to equilibrium)**.** *Let m be irreducible and aperiodic, and suppose that π is the invariant distribution of m on $\mathcal{X}$. Let λ be any probability distribution on $\mathcal{X}$. Suppose that $(X_n)_{n\in\mathbb{N}}$ is $Markov(\lambda, m)$. Then,*

$$\mathbb{P}(X_n = j) \to \pi_j \text{ as } n \to \infty \text{ for all } j \in \mathcal{X}.$$

In particular,

$$m_{ij}^{(n)} \to \pi_j \text{ as } n \to \infty \text{ for all } i, j \in \mathcal{X}.$$

Proof. We use a coupling argument. Let $(Y_n)_{n\in\mathbb{N}}$ be $Markov(\pi, m)$ and independent of $(X_n)_{n\in\mathbb{N}}$. Fix a reference state b and set

$$T = \inf\{n \geq 1 : X_n = Y_n = b\}$$

Step 1. We show $\mathbb{P}(T < \infty) = 1$. The process $W_n = (X_n, Y_n)$ is a Markov chain on $\mathcal{X} \times \mathcal{X}$ with transition probabilities $\tilde{m}_{(i,k)(j,l)} = m_{ij}m_{kl}$ and initial distribution $\mu_{(i,k)} = \lambda_i\pi_k$. Since m is aperiodic, for any states $i, j, k, l \in \mathcal{X}$, we have

$$\tilde{m}_{(i,k)(j,l)}^{(n)} = m_{ij}^{(n)}m_{kl}^{(n)} > 0$$

for all sufficiently large n; so $\tilde{m}$ is irreducible. Also, $\tilde{m}$ has an invariant distribution given by $\tilde{\pi}_{(i,k)} = \pi_i\pi_k$. So, by Theorem 1.7.7 in [4], $\tilde{m}$ is positive recurrent. But T is the first passage time of W_n to (b, b) so $\mathbb{P}(T < \infty) = 1$, by Theorem 1.5.7 in [4].

Step 2. Set

$$Z_n = \begin{cases} X_n & \text{if } n < T, \\ Y_n & \text{if } n \geq T. \end{cases}$$

We want to show that $(Z_n)_{n\in\mathbb{N}}$ is $Markov(\lambda, m)$. The strong Markov property (see Theorem 1.4.2 in [4]) applies to $(W_n)_{n\in\mathbb{N}}$ at time T, so $(X_{T+n}, Y_{T+n})_{n\in\mathbb{N}}$ is $Markov(\delta_{(b,b)}, \tilde{m})$ and is independent of $(X_0, Y_0), (X_1, Y_1), \ldots, (X_{T-1}, Y_{T-1})$. By symmetry, we can replace the process $(X_{T+n}, Y_{T+n})_{n\in\mathbb{N}}$ by $(Y_{T+n}, X_{T+n})_{n\in\mathbb{N}}$ which is also $Markov(\delta_{(b,b)}, \tilde{m})$ and remains independent of $(X_0, Y_0), (X_1, Y_1), \ldots, (X_{T-1}, Y_{T-1})$. Hence, $W_n' = (Z_n, Z_n')$ is $Markov(\mu, \tilde{m})$ where

$$Z_n' = \begin{cases} Y_n & \text{if } n < T, \\ X_n & \text{if } n \geq T. \end{cases}$$

In particular, $(Z_n)_{n\in\mathbb{N}}$ is $Markov(\lambda, m)$.

Step 3. We have $\mathbb{P}(Z_n = j) = \mathbb{P}(X_n = j \text{ and } n < T) + \mathbb{P}(Y_n = j \text{ and } n \geq T)$. So, since both $(X_n)_{n\in\mathbb{N}}$ and $(Z_n)_{n\in\mathbb{N}}$ are $Markov(\lambda, m)$, we get

$$\begin{aligned} |\mathbb{P}(X_n = j) - \pi_j| &= |\mathbb{P}(Z_n = j) - \mathbb{P}(Y_n = j)| \\ &= |\mathbb{P}(X_n = j \text{ and } n < T) - \mathbb{P}(Y_n = j \text{ and } n < T)| \\ &\leq \mathbb{P}(n < T) \end{aligned}$$

and $\mathbb{P}(n < T) \to 0$ as $n \to \infty$. □

3 Convergence using transport distances

In this section, in order to study the convergence speed of Markov chains, we will use transport distances. We will focus on the Wasserstein distance, which is a common metric to measure the distance between probability distributions.

3.1 Optimal transport

To begin with, we need to introduce the optimal transport problem. The material presented in this section is adapted from the first chapter of [6].

Let $\mathcal{X} = \{x_1, \dots, x_n\}$ and $\mathcal{Y} = \{y_1, \dots, y_m\}$ be two discrete sets and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a cost function. Let $(\mathbf{a}, \mathbf{b}) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$ be two probability vectors. The optimal transport problem was initially formulated by **Monge** [1781] and the aim was to find a map $\mathcal{T} : \mathcal{X} \to \mathcal{Y}$ that minimizes the cost of transforming ("pushing") a discrete probability measure $\alpha = \sum_{i=1}^n \mathbf{a}_i \delta_{x_i}$ to another $\beta = \sum_{i=1}^m \mathbf{b}_i \delta_{y_i}$ with respect to a cost matrix $C = (c(x_i, y_j))_{1 \le i \le n, 1 \le j \le m}$. The map $\mathcal{T}$ must verify that

$$\text{for any } j \in \{1, \dots, m\}, \quad \mathbf{b}_j = \sum_{i : \mathcal{T}(x_i) = y_j} \mathbf{a}_i,$$

which we write in compact form as $\mathcal{T}_{\#\alpha} = \beta$. The optimal transport problem formulated by Monge can be summarised with the following formula

$$\inf_{\mathcal{T}} \left\{ \sum_i c(x_i, \mathcal{T}(x_i)) : \mathcal{T}_{\#\alpha} = \beta \right\}.$$

Note that this problem is unsolvable if $n < m$ and requires many conditions to be solvable, even if $m \le n$. For instance, if there is $j \in \{1, \dots, m\}$ such that $\mathbf{b}_j < \mathbf{a}_i$ for all $i \in \{1, \dots, n\}$, the problem is unsolvable.

The **Kantorovich** relaxation was later introduced [1942] and was able to address several of these issues. The key idea is to relax the deterministic nature of transportation. This means that a source point $x_i \in \mathcal{X}$ can be assigned to several locations in $\mathcal{Y}$. Instead of finding a map, we now look for a matrix P that describes how much "weight" travels from each input to each output. Explicitly, P_{ij} corresponds to the weight that travels from x_i to y_j. The set of admissible couplings for $(\mathbf{a}, \mathbf{b})$ is given by

$$U(\mathbf{a}, \mathbf{b}) := \left\{ P \in \mathbb{R}_+^{n \times m} : P\mathbf{1}_m = \mathbf{a}, P^\top \mathbf{1}_n = \mathbf{b} \right\}.$$

As long as the total mass on either side is equal, there exists always at least one solution. Kantorovich's optimal transport problem reads

$$L_C(\mathbf{a}, \mathbf{b}) := \min_{P \in U(\mathbf{a}, \mathbf{b})} \langle P, C \rangle = \min_{P \in U(\mathbf{a}, \mathbf{b})} \sum_{i,j} P_{ij} C_{ij}.$$

Now, we are ready to introduce the Wasserstein distance.

Definition 3.1.1 (Wasserstein distance)**.** *Let $(\mathcal{X}, d)$ be a metric space and let μ, ν be two probability measures on $\mathcal{X}$. The* ***Wasserstein distance*** *between μ and ν is*

$$\mathrm{W}_1(\mu, \nu) := \inf_{\xi \in \Pi(\mu, \nu)} \int_{(x,y) \in \mathcal{X} \times \mathcal{X}} d(x, y) \, \mathrm{d}\xi(x, y)$$

where $\Pi(\mu,\nu)$ is the set of all couplings between μ and ν, that is, the set of probability measures on $\mathcal{X}\times\mathcal{X}$ whose marginal laws are μ and ν.

Remark 3.1.1. *In our case, since $\mathcal{X}$ is finite, we have*

$$\mathrm{W}_1(\mu,\nu) = \inf_{\xi\in\Pi(\mu,\nu)} \sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} d(x,y)\,\xi(x,y).$$

In the above definition, $\mathrm{d}\xi(x,y)$ (or simply $\xi(x,y)$ in the finite case) represents the mass that travels from x to y. In our case, as $\mathcal{X}$ is finite, finding a coupling Ξ witnessing for the Wasserstein distance is equivalent to solving Kantorovich's optimal transport problem. Couplings witnessing for the Wasserstein distance are said to be *optimal*.

3.2 Ollivier's Ricci curvature

Assume we have two Markov chains associated with the same transition matrix m, located at $x\in\mathcal{X}$ and $y\in\mathcal{X}$ respectively at a given step $n\in\mathbb{N}$. We will ask whether the two measures m_x and m_y are closer or further apart than are the points x and y, in which case Ollivier's Ricci curvature introduced in [5] will be positive or negative respectively.

Definition 3.2.1 (Ollivier's Ricci curvature)**.** *Let $(\mathcal{X},d,m)$ be a metric space associated with the transition matrix of a Markov chain. Let $x,y\in\mathcal{X}$ be two distinct points.* ***Ollivier's Ricci curvature*** *of $(\mathcal{X},d,m)$ along (x,y) is*

$$\kappa(x,y) := 1 - \frac{\mathrm{W}_1(m_x,m_y)}{d(x,y)}.$$

3.3 Examples

Example 3.3.1 (n-dimensional hypercube)**.** *For some $n\in\mathbb{N}$, let $\mathcal{X}=\{0,1\}^n$ be the n-dimensional discrete hypercube equipped with the Hamming metric. Let m be the lazy random walk on $\mathcal{X}$, i.e. $m_x(x)=\frac{1}{2}$ and $m_x(y)=\frac{1}{2n}$ if y is a neighbor of x. We choose to work with the lazy random walk and not just the random walk in order to have aperiodicity. In fact we want m to verify $m_x^n(y)>0$ for any $(x,y)\in\mathcal{X}^2$ and all sufficiently large n. We want to compute Ollivier's Ricci curvature of this hypercube. Let $x,y\in\mathcal{X}$. Without loss of generality, we can assume*

$$x=(0,0,0,\dots,0) \quad \textit{and} \quad y=(1,\dots,1,0,\dots,0).$$

Let us denote by k the number of 1 bits in y. For $z\in\mathcal{X}$ and $1\le i\le n$, let us denote by z^i the neighbor of z in which the i-th bit is switched. An optimal coupling between m_x and m_y (i.e. an optimal transport plan from m_x to m_y) is as follows:

- *for $1\le i\le k$, move a mass of $\frac{1}{2n}$ from x to y^i (we have $d(x,y^i)=k-1$);*
- *move a mass of $\frac{n-k}{2n}$ from x to y (we have $d(x,y)=k$);*
- *for $1\le i\le k$, move a mass of $\frac{1}{2n}$ from x^i to y (we have $d(x^i,y)=k-1$);*

- *for $k+1 \leq i \leq n$, move a mass of $\frac{1}{2n}$ from x^i to y^i (we have $d(x^i, y^i) = k$).*

As a result, $\mathrm{W}_1(x, y)$ is at most

$$\left(k \cdot \frac{1}{2n} \cdot (k-1)\right) + \left((n-k) \cdot \frac{1}{2n} \cdot k\right) + \left(k \cdot \frac{1}{2n} \cdot (k-1)\right) + \left((n-k) \cdot \frac{1}{2n} \cdot k\right) = k - \frac{k}{n}.$$

Hence, Ollivier's Ricci curvature is at least $\frac{1}{n}$. Now, let us define f as follows

$$\begin{aligned} f : \mathcal{X} &\to \{1, \dots, n\} \\ (i_1, \dots, i_n) &\mapsto \sum_{j=1}^{n} i_j. \end{aligned}$$

This is a 1-Lipschitz function, with $f(x) = 0$ and $f(y) = k$. The expectations of f under m_x and m_y are

$$\mathbb{E}_{m_x}(f) = \sum_{i \in \mathcal{X}} f(i)\, m_x(i) = \frac{1}{2}, \quad \mathbb{E}_{m_y}(f) = \sum_{i \in \mathcal{X}} f(i)\, m_y(i) = k - \frac{k}{n} + \frac{1}{2}.$$

As a result, we can deduce from Kantorovich-Rubinstein duality, whose proof is provided in the first chapter of Villani's book [7], that $\mathrm{W}_1(x, y)$ is bounded from below by $k - \frac{k}{n}$. Thus, we get that Ollivier's Ricci curvature of $(\mathcal{X}, d, m)$ along (xy) is $\frac{1}{n}$. This is true for any $x, y \in \mathcal{X}$. Hence, Ollivier's Ricci curvature of the n-dimensional hypercube equipped with the lazy random walk is $\frac{1}{n}$.

Now, we will prove the geometric convergence in the Wasserstein distance of the lazy random walk on the n-dimensional hypercube in two different ways. We will use the coupling lemma before looking at a more general result, the Wasserstein distance contraction, in the subsequent section. To begin, let us have a look at the coupling lemma.

Lemma 3.3.1 (Coupling Lemma)**.** *Let μ and ν be two probability distributions on Ω and η be a coupling of (μ, ν). If (X, Y) is distributed according to η, we have*

$$\|\mu - \nu\|_{tv} \leq \mathbb{P}[X \neq Y],$$

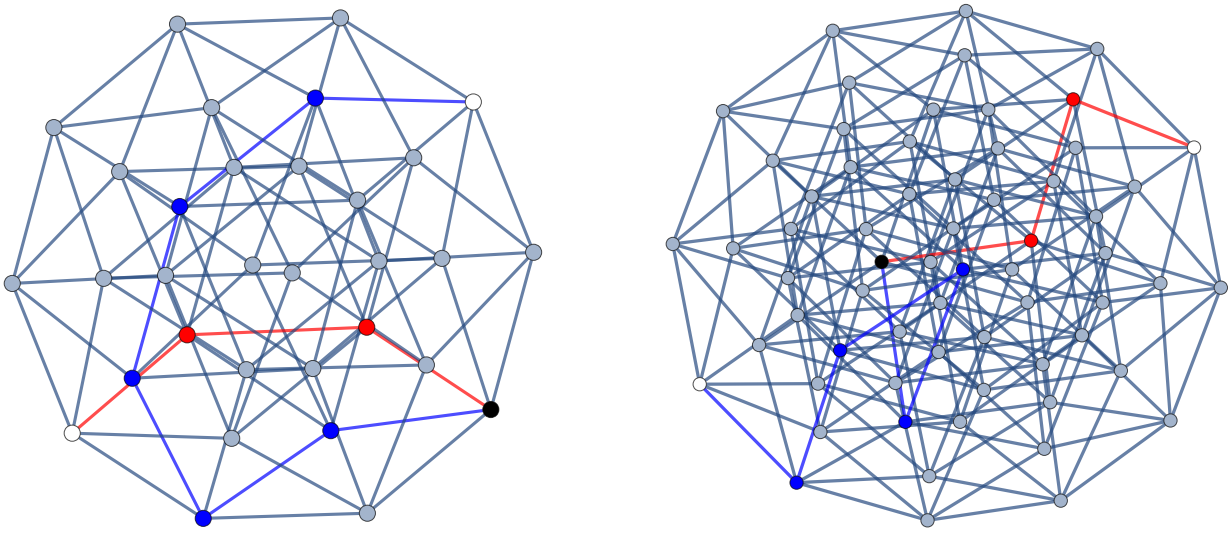
where $\|\mu - \nu\|_{tv} := \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|$ denotes the total variation distance.

Proof. For any coupling w of (μ, ν), we have that for any $z \in \Omega$, $w(z, z) \leq \min(\mu(z), \nu(z))$. As a result, we can write

$$\begin{aligned} \mathbb{P}(X \neq Y) &= 1 - \mathbb{P}(X = Y) \\ &= 1 - \sum_{z \in \Omega} w(z, z) \\ &\geq \sum_{z \in \Omega} \mu(z) - \sum_{z \in \Omega} \min(\mu(z), \nu(z)) \\ &= \sum_{z \in \Omega \,:\, \mu(z) > \nu(z)} (\mu(z) - \nu(z)) \\ &= \|\mu - \nu\|_{tv}. \end{aligned}$$

□

So as to prove the quadratic convergence of the lazy random walk on the n-dimensional hypercube, we can use a coupling $(X_t, Y_t)_{t\in\mathbb{N}}$ where at each time $t \in \mathbb{N}$, we compute the subsequent steps of the two Markov chains using the matrix formed by the transport plan described in example 3.3.1. In other words, for any $t \in \mathbb{N}$, conditional on (X_t, Y_t), in order to decide where the two Markov chains will go at time $t+1$, we use a joint distribution of (X_{t+1}, Y_{t+1}) witnessing for $\mathrm{W}_1(X_{t+1}, Y_{t+1})$. In practice, this means that at each step, we will choose a coordinate uniformly at random, and decide whether or not to switch it for the two Markov chains in order to match their coordinates. As such, we can make the two Markov chains meet as fast as possible, and then move synchronously. We give below two examples of such optimal couplings of two Markov chains starting from opposite vertices (until they meet). The initial distributions of the two Markov chains are the Dirac measures at two opposite vertices of the n-dimensional hypercube. The two figures 1a and 1b were created using Wolfram Mathematica.



(a) 5-dimensonal hypercube

(b) 6-dimensonal hypercube

Figure 1: Optimal lazy random walk couplings on hypercubes

For any two initial distributions μ and ν for $(X_t)_{t\in\mathbb{N}}$ and $(Y_t)_{t\in\mathbb{N}}$, we denote by τ_{couple} the first time the two chains meet. In other words, we write $\tau_{couple} := \min\{t \in \mathbb{N} \mid X_t = Y_t\}$. We denote by τ the first time all coordinates have been selected at least once. Clearly, we have $\|\mu * m^t - \nu * m^t\|_{tv} \leq \mathbb{P}(X_t \neq Y_t) = \mathbb{P}(t < \tau_{couple}) \leq \mathbb{P}(t < \tau)$. Moreover, we know that

$$\mathbb{P}(t < \tau) \leq \sum_{k=1}^{n} \mathbb{P}(\text{coordinate } k \text{ not selected yet}) \leq n\left(1 - \frac{1}{n}\right)^t \leq ne^{-t/n}.$$

Since the compact set $\{0,1\}^n$ is equipped with the Hamming metric, we have that for any two probability measures ψ and ϕ on $\mathcal{X}$, $\mathrm{W}_1(\psi, \phi) \leq n\|\psi - \phi\|_{tv}$. As a result, we get that for any two initial distributions μ and ν,

$$\mathrm{W}_1(\mu * m^t, \nu * m^t) \leq n\|\mu * m^t - \nu * m^t\|_{tv} \leq n\mathbb{P}(t < \tau) \leq n^2\left(1 - \frac{1}{n}\right)^t \leq n^2 e^{-t/n}.$$

Let π be the uniform distribution on $\mathcal{X}$. It is the invariant distribution since $\pi * m = \pi$. We thus also get that for any distribution μ on $\mathcal{X}$, $\mathrm{W}_1(\mu * m^t, \pi) \leq n^2 e^{-t/n}$.

Another intersting example is the discrete Ornstein Uhlenbeck process.

Example 3.3.2 (discrete Ornstein Uhlenbeck process)**.** *Let $N \in \mathbb{N}$, $\mathcal{X} = \{-N, \dots, N-1, N\}$ equipped with the absolute value. Let m be the random walk on $\mathcal{X}$ given by*

$$m_k(k) = \frac{1}{2}, \quad m_k(k+1) = \frac{1}{4} - \frac{k}{4N}, \quad m_k(k-1) = \frac{1}{4} + \frac{k}{4N}.$$

This is a lazy random walk with linear drift towards 0. The binomial distribution $\frac{1}{2^{2N}}\binom{2N}{N+k}$ is reversible for this random walk. We will show that for any two distinct $x, y \in \mathcal{X}$, one has $\kappa(x, y) = \frac{1}{2N}$. Without loss of generality, we assume $x < y$. For $z \in \mathcal{X}$, let us denote by z^- and z^+ the left and right neighbors of z, respectively. Let $k \in \mathbb{N}$, such that $d(x, y) = k$. This time, let us compare two optimal couplings between m_x and m_y. The first one can be described as follows:

- *move a mass of $\frac{1}{4} + \frac{x}{4N}$ from x^- to y^- (we have $d(x^-, y^-) = k$);*
- *move a mass of $\frac{k}{4N}$ from x to y^- (we have $d(x, y^-) = k - 1$);*
- *move a mass of $\frac{1}{2} - \frac{k}{4N}$ from x to y (we have $d(x, y) = k$);*
- *move a mass of $\frac{k}{4N}$ from x^+ to y (we have $d(x^+, y) = k - 1$);*
- *move a mass of $\frac{1}{4} - \frac{x+k}{4N}$ from x^+ to y^+ (we have $d(x^+, y^+) = k$).*

The second one is the following:

- *move a mass of $\frac{1}{4} + \frac{x}{4N}$ from x^- to y (we have $d(x^-, y) = k + 1$);*
- *move a mass of $\frac{1}{4} + \frac{y}{4N}$ from x to y^- (we have $d(x, y^-) = k - 1$);*
- *move a mass of $\frac{1}{4} - \frac{x}{4N}$ from x^+ to y (we have $d(x^+, y) = k - 1$);*
- *move a mass of $\frac{1}{4} - \frac{y}{4N}$ from x to y^+ (we have $d(x, y^+) = k + 1$).*

In both case, we get that the transport cost is $k - \frac{k}{2N}$. Using Kantorovich-Rubinstein duality as we did in the previous example (3.3.1), we can easily get that Ollivier's Ricci curvature is $\frac{1}{2N}$.

With the two couplings described in Example 3.3.2, at each step, the distance between the two Markov chains can change by at most 1. It is interesting to note that if we use the first transport plan, the two Markov chains will get closer with probability $\frac{k}{2N}$. Otherwise, they will remain at the same distance. This means that when the two chains are close to one another, the probability of meeting will be very low: $\frac{1}{2N}$ if $d(X_t, Y_t) = 1$. On the contrary, in the case of the second transport plan, the two chains will get closer with probability $\frac{1}{2} + \frac{k}{4N}$. Otherwise, with probability $\frac{1}{2} - \frac{k}{4N}$, the distance between the two chains will increase.

As a consequence of what we described in the previous paragraph, if we fix $N = 10$ and we let the two chains start from -10 and 10 respectively (i.e. $X_0 = -10$, $Y_0 = 10$), running the coupling $100,000$ times shows that it takes on average 72 steps for the two chains to meet with the first transport plan. However, in the same setting and using this time the second transport plan, running the coupling $100,000$ times shows that it takes on average 36 steps for the two chains to meet. Therefore, the choice of optimal transport plan can have a significant impact on the results when performing a simulation.

In order to find an optimal coupling for two initial probability distributions μ and ν on $\mathcal{X}$, I used linear programming techniques on Python. In particular, I used the SciPy library. However, the POT: Python Optimal Transport library was more efficient so as to get exact solutions to optimal transport problems. For instance, regarding the previous example (3.3.2), I used that library to find an optimal coupling between the two distributions displayed in Figure 2. As we can see, since the given optimal transport plan is not symmetric (as opposed to the two probability distributions), we can deduce that it is not unique. The Matplotlib library was used to create that figure.

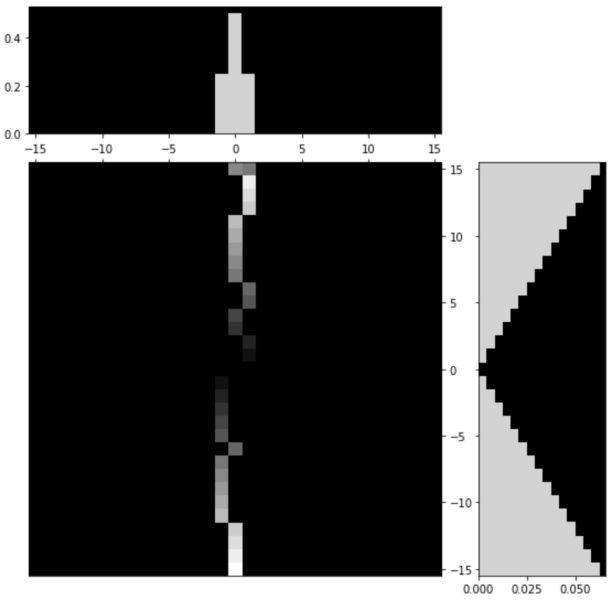


Figure 2: Coupling witnessing for W_1

3.4 Wasserstein distance contraction

Now, we will prove a more general result that will enable us to get geometric convergence directly from positive Ollivier's Ricci curvature. An equivalent statement for any metric space appears in Ollivier's paper [5].

Proposition 3.4.1 (Wasserstein distance contraction)**.** *Let $(\mathcal{X}, d, m)$ be a metric space with the transition matrix of a Markov chain. Let $\kappa \in \mathbb{R}$. We have $\kappa(x, y) \geq \kappa$ for any $x, y \in \mathcal{X}$ if and only if for any two probability distributions μ, ν on $\mathcal{X}$, one has*

$$\mathrm{W}_1(\mu * m, \nu * m) \leq (1 - \kappa)\, \mathrm{W}_1(\mu, \nu).$$

Proof. Firstly, suppose that convolution with m is contracting in the Wasserstein distance. For some $x, y \in \mathcal{X}$, let $\mu = \delta_x$ and $\nu = \delta_y$ be the Dirac measures at x and y respectively. Then, by definition, $\delta_x * m = m_x$ and likewise for y, so that

$$\mathrm{W}_1(m_x, m_y) \leq (1 - \kappa)\, \mathrm{W}_1(\delta_x, \delta_y) = (1 - \kappa) d(x, y)$$

as required. Let us now show the converse implication. For each pair $(x, y) \in \mathcal{X} \times \mathcal{X}$ let ξ_{xy} be a coupling between m_x and m_y witnessing for $\kappa(x, y) \geq \kappa$. Let Ξ be a coupling between

μ and ν witnessing for $\mathrm{W}_1(\mu,\nu)$. We know that

$$\sum_{(x,y)\in\mathcal{X}\times\mathcal{X}} \Xi(x,y)\,\xi_{xy}$$

is a coupling between $\mu * m$ and $\nu * m$. As such, we can write

$$\begin{aligned}
\mathrm{W}_1(\mu*m,\nu*m) &\leq \sum_{(x',y')\in\mathcal{X}^2} d(x',y') \sum_{(x,y)\in\mathcal{X}^2} \Xi(x,y)\,\xi_{xy}(x',y') \\
&= \sum_{(x,y)\in\mathcal{X}^2} \Xi(x,y) \sum_{(x',y')\in\mathcal{X}^2} d(x',y')\,\xi_{xy}(x',y') \\
&\leq (1-\kappa) \sum_{x,y\in\mathcal{X}^2} d(x,y)\,\Xi(x,y) \\
&= (1-\kappa)\,\mathrm{W}_1(\mu,\nu).
\end{aligned}$$

□

As an immediate consequence of this contracting property, we get the following corollary.

Corollary 3.4.1 (Wasserstein distance convergence)**.** *Suppose that $\kappa(x,y) \geq \kappa > 0$ for any two distinct $x, y \in \mathcal{X}$. Then, the Markov chain has a unique invariant distribution π. Moreover, for any probability measure μ on $\mathcal{X}$, the sequence $\mu * m^n$ tends exponentially fast to π in the Wasserstein distance. Namely*

$$\mathrm{W}_1(\mu * m^n, \pi) \leq (1-\kappa)^n\,\mathrm{W}_1(\mu,\pi).$$

Proof. Assume there exist two different invariant distributions π and π'. We have $\pi * m = \pi$ as well as $\pi' * m = \pi'$. Therefore, by iterating proposition 3.4.1, we can write

$$\mathrm{W}_1(\pi,\pi') \leq \mathrm{W}_1(\pi,\mu*m^n) + \mathrm{W}_1(\mu*m^n,\pi') \leq (1-\kappa)^n\Big(\mathrm{W}_1(\pi,\mu) + \mathrm{W}_1(\mu,\pi')\Big) \xrightarrow[n\to\infty]{} 0,$$

and we are done. □

3.5 Construction results

In this section, we describe 3 consequences of proposition 3.4.1. They are adapted from Ollivier's paper [5].

Proposition 3.5.1 (Composition)**.** *Let $(\mathcal{X},d)$ be a metric space equipped with two Markov chains m and m'. Suppose that Ollivier's Ricci curvature of m is at least κ, and that of m' is at least κ'. Then, the Ricci curvature of $m * m'$ is at least $\kappa + \kappa' - \kappa\kappa'$.*

Proof. For any $x, y \in \mathcal{X}$, we have

$$\begin{aligned}
\mathrm{W}_1(\delta_x * m * m', \delta_y * m * m') &\leq (1-\kappa')\,\mathrm{W}_1(\delta_x * m, \delta_y * m) \\
&\leq (1-\kappa')(1-\kappa)\,\mathrm{W}_1(\delta_x,\delta_y) \\
&= (1-(\kappa+\kappa'-\kappa\kappa'))d(x,y).
\end{aligned}$$

□

Proposition 3.5.2 (Superposition)**.** *Let $(\mathcal{X},d)$ be a metric space equipped with a family $\left(m^{(i)}\right)_{i\in I}$ of Markov chains. Suppose that for each $i \in I$, the Ricci curvature of $m^{(i)}$ is at least κ_i. Let $(\alpha_i)_{i\in I}$ be a family of non-negative real numbers with $\sum_{i\in I}\alpha_i = 1$. Define a Markov chain m on $\mathcal{X}$ by $m := \sum_{i\in I}\alpha_i m^{(i)}$. The Ricci curvature of m is at least $\sum_{i\in I}\alpha_i\kappa_i$.*

Proof. Let $x, y \in \mathcal{X}$ and for each i let ξ_i be a coupling between $m_x^{(i)}$ and $m_y^{(i)}$ witnessing for $\kappa(x,y) \geq \kappa_i$. Then, $\sum_{i\in I} \alpha_i \xi_i$ is a coupling between $\sum_{i\in I} \alpha_i m_x^{(i)}$ and $\sum_{i\in I} \alpha_i m_y^{(i)}$. Therefore, we have

$$\begin{aligned}
\mathrm{W}_1(m_x, m_y) &\leq \sum_{i\in I} \alpha_i \, \mathrm{W}_1\left(m_x^{(i)}, m_y^{(i)}\right) \\
&\leq \sum_{i\in I} \alpha_i (1-\kappa_i) d(x,y) \\
&= \left(1 - \sum_{i\in I} \alpha_i \kappa_i\right) d(x,y).
\end{aligned}$$

□

Proposition 3.5.3 (Tensorization)**.** *Let $((\mathcal{X}_i, d_i))_{i\in I}$ be a finite family of metric spaces and suppose that for each $i \in I$, $\mathcal{X}_i$ is equipped with a Markov chain $m^{(i)}$. Let $\mathcal{X}$ be the product of the spaces $(\mathcal{X}_i)_{i\in I}$, equipped with the distance $d := \sum_{i\in I} d_i$. Let $(\alpha_i)_{i\in I}$ be a family of non-negative real numbers such that $\sum_{i\in I} \alpha_i = 1$. Consider a Markov chain m defined for any $(x_1, \ldots, x_k) \in \mathcal{X}$ by*

$$m_{(x_1,\ldots,x_k)} := \sum_{i\in I} \alpha_i \, \delta_{x_1} \otimes \cdots \otimes m_{x_i}^{(i)} \otimes \cdots \otimes \delta_{x_k}.$$

Suppose that for each $i \in I$, the Ricci curvature of $m^{(i)}$ is at least κ_i. Then, the Ricci curvature of m is at least $\inf_{i\in I} \alpha_i \kappa_i$.

Proof. For $x \in \mathcal{X}$, let $\tilde{m}_x^{(i)}$ stand for $\delta_{x_1} \otimes \cdots \otimes m_{x_i}^{(i)} \otimes \cdots \otimes \delta_{x_k}$. Let $x = (x_i)_{i\in I}$ and $y = (y_i)_{i\in I}$ be two points in $\mathcal{X}$. Then,

$$\begin{aligned}
\mathrm{W}_1(m_x, m_y) &\leq \sum_{i\in I} \alpha_i \, \mathrm{W}_1\left(\tilde{m}_x^{(i)}, \tilde{m}_y^{(i)}\right) \\
&\leq \sum_{i\in I} \alpha_i \left(\mathrm{W}_1\left(m_{x_i}^{(i)}, m_{y_i}^{(i)}\right) + \sum_{j\neq i} d_j(x_j, y_j)\right) \\
&\leq \sum_{i\in I} \alpha_i \left((1-\kappa_i) d_i(x_i, y_i) + \sum_{j\neq i} d_j(x_j, y_j)\right) \\
&= \sum_{i\in I} \alpha_i \left(-\kappa_i d_i(x_i, y_i) + \sum_{j\in I} d_j(x_j, y_j)\right) \\
&= \sum_{i\in I} d_i(x_i, y_i) - \sum_{i\in I} \alpha_i \kappa_i d_i(x_i, y_i) \\
&\leq \left(1 - \inf_{i\in I} \alpha_i \kappa_i\right) \sum_{i\in I} d_i(x_i, y_i) \\
&= \left(1 - \inf_{i\in I} \alpha_i \kappa_i\right) d(x,y)
\end{aligned}$$

□

Remark 3.5.1. *The proposition we just proved (3.5.3) allows for a very short proof of the Ricci curvature of the lazy random walk on the n-dimensional hypercube $\{0,1\}^n$ from Example 3.3.1. In fact, since the lazy random walk on $\{0,1\}$ sends every point to the invariant ditribution $(1/2, 1/2)$, it has curvature 1. We can immediately get from the above proposition that the Ricci curvature of the lazy random walk on $\{0,1\}^n$ is at least $1/n$.*

4 Analytical approach to exponential convergence

In this section, we will work in the $L^2(\mathcal{X}, \pi)$ space, where π is the invariant distribution of m, and we will use properties of Hilbert spaces to derive a similar contraction result as in the previous section.

4.1 Averaging operator

To begin, we will introduce the following averaging operator.

Definition 4.1.1 (Avergaging operator)**.** *For $f \in L^2(\mathcal{X}, \pi)$, let the averaging operator* M *be*

$$\mathrm{M}\, f(x) := \sum_{y \in \mathcal{X}} f(y)\, m_x(y).$$

Remark 4.1.1. *We can observe that the application of the averaging operator $\mathrm{M}\, f(x)$ gives the expectation of $f(X_{t+1})$ conditional on $X_t = x$. Similarly, applying the averaging operator $n \in \mathbb{N}$ times gives*

$$\mathrm{M}^n\, f(x) = \sum_{\mathcal{X}^n} f(x_n)\, m_{x_{n-1}}(x_n) \dots m_x(x_1) = \sum_{y \in \mathcal{X}} f(y)\, m_x^n(y) = \mathbb{E}_{m_x^n}\left[f\right] = \mathbb{E}_x\left[f(X_n)\right].$$

Let us prove the Lipschitz contraction of this operator. A similar proof is provided in [5].

Proposition 4.1.1. *Let m be a Markov chain on a metric space $(\mathcal{X}, d)$. Let $\kappa \leq 1$. Then, the Ricci curvature of $\mathcal{X}$ is at least κ, if and only if, for every k-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$, the function $\mathrm{M}\, f$ is $k(1-\kappa)$-Lipschitz.*

Proof. To begin, suppose that the Ricci curvature of $\mathcal{X}$ is at least κ. Let $f : \mathcal{X} \to \mathbb{R}$ be a k-Lipschitz function. There exists a function $g : \mathcal{X} \to \mathbb{R}$ such that $f \equiv kg$. Moreover, g is 1-Lipschitz. We have, using Kantorovich-Rubinstein duality,

$$\begin{aligned}
\mathrm{M}\, f(x) - \mathrm{M}\, f(y) &= \sum_{z \in \mathcal{X}} f(z)\, m_x(z) - \sum_{z \in \mathcal{X}} f(z)\, m_y(z) \\
&= k \left(\sum_{z \in \mathcal{X}} g(z)\, m_x(z) - \sum_{z \in \mathcal{X}} g(z)\, m_y(z) \right) \\
&\leq k\, \mathrm{W}_1(m_x, m_y) \\
&\leq k(1-\kappa) d(x, y).
\end{aligned}$$

Conversely, suppose that whenever f is 1-Lipschitz, $\mathrm{M}\, f$ is $(1-\kappa)$-Lipschitz. Then, using Kantorovich-Rubinstein duality, we get

$$\begin{aligned}
\mathrm{W}_1(m_x, m_y) &= \sup_{f \text{ 1-Lipschitz}} \sum_{z \in \mathcal{X}} f(z)\, (m_x - m_y)(z) \\
&= \sup_{f \text{ 1-Lipschitz}} \mathrm{M}\, f(x) - \mathrm{M}\, f(y) \\
&\leq (1-\kappa) d(x, y).
\end{aligned}$$

□

Proposition 4.1.2. *The operator* M *is self-adjoint in* $L^2(\mathcal{X}, \pi)$ *if and only if* π *is reversible for the Markov chain.*

Proof. Assume that π is reversible for m. Then, we have

$$\pi(x)\, m_x(y) = \pi(y)\, m_y(x) \quad \text{for any } x, y \in \mathcal{X}.$$

As a result, we get

$$\begin{aligned}
\langle \mathrm{M}\, f, g \rangle_{L^2(\mathcal{X},\pi)} &= \sum_{x \in \mathcal{X}} \mathrm{M}\, f(x)\, g(x)\, \pi(x) \\
&= \sum_{x \in \mathcal{X}} g(x) \left(\sum_{y \in \mathcal{X}} f(y)\, m_x(y) \right) \pi(x) \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}} g(x)\, f(y)\, \pi(x)\, m_x(y) \\
&= \sum_{y \in \mathcal{X}} \sum_{x \in \mathcal{X}} g(x)\, f(y)\, \pi(y)\, m_y(x) \\
&= \sum_{y \in \mathcal{X}} f(y) \left(\sum_{x \in \mathcal{X}} g(x)\, m_y(x) \right) \pi(y) \\
&= \sum_{y \in \mathcal{X}} f(y)\, \mathrm{M}\, g(y)\, \pi(y) \\
&= \langle f, \mathrm{M}\, g \rangle_{L^2(\mathcal{X},\pi)}.
\end{aligned}$$

Let us now show the other implication. Fix $x_1, x_2 \in \mathcal{X}$. The two indicator functions $\mathbf{1}_{\{x_1\}}$ and $\mathbf{1}_{\{x_2\}}$ are in $L_2(\mathcal{X}, \pi)$. Therefore, using the fact that

$$\left\langle \mathbf{1}_{\{x_1\}}, \mathrm{M}\, \mathbf{1}_{\{x_2\}} \right\rangle_{L^2(\mathcal{X},\pi)} = \left\langle \mathrm{M}\, \mathbf{1}_{\{x_1\}}, \mathbf{1}_{\{x_2\}} \right\rangle_{L^2(\mathcal{X},\pi)},$$

we get

$$\pi(x_1)\, m_{x_1}(x_2) = \pi(x_2)\, m_{x_2}(x_1).$$

This is true for any $(x_1, x_2) \in \mathcal{X} \times \mathcal{X}$, which completes the proof. □

4.2 Variance contraction

Now, we want to prove the following theorem.

Theorem 4.2.1. *We assume* π *is reversible for the Markov chain transition matrix* m *and the Ricci curvature of* $\mathcal{X}$ *is at least* κ*. For any* $f \in L^2(\mathcal{X}, \pi)$ *and any* $n \in \mathbb{N}$*, we have*

$$\mathrm{Var}_\pi\left(\mathrm{M}^n\, f\right) \leq (1 - \kappa)^{2n}\, \mathrm{Var}_\pi(f).$$

Proof. Thanks to Proposition 4.1.2, we know that M is self-adjoint. Thus, there exists an orthogonal basis $(e_1, \ldots, e_N)$ of $L^2(\mathcal{X}, \pi)$ consisting of eigenvectors for M. In other words, for any $i \in \{1, \ldots, N\}$, there exists $\lambda_i \in \mathbb{R}$ such that

$$\mathrm{M}\, e_i = \lambda_i e_i.$$

Since $\mathcal{X}$ is finite, Lipschitz functions coincide with $L^2(\mathcal{X}, \pi)$ functions. We know that for all constant functions $f \in L^2(\mathcal{X}, \pi)$, $\mathrm{M}\, f = f$. As a result, there exists $i \in \{1, \dots, N\}$ such that e_i is constant, with eigenvalue 1. Without loss of generality, we can assume $i = 1$. Moreover, since $\langle e_1, e_1 \rangle_{L^2(\mathcal{X},\pi)} = 1$, we get $e_1 \equiv 1$.

Now, we will focus on the variance

$$\mathrm{Var}_\pi(f) = \|f - \mathbb{E}_\pi[f]\|^2_{L^2(\mathcal{X},\pi)} = \frac{1}{2} \sum_{\mathcal{X}\times\mathcal{X}} (f(x) - f(y))^2 \, \pi(x)\, \pi(y).$$

We can observe that for any $f \in L^2(\mathcal{X}, \pi)$,

$$\mathbb{E}_\pi[f] = \sum_{x\in\mathcal{X}} f(x)\, \pi(x) = \sum_{x\in\mathcal{X}} f(x)\, e_1(x)\, \pi(x) = \langle f, e_1 \rangle_{L^2(\mathcal{X},\pi)}.$$

Since e_1 is constant, we know that for any $i \in \{2, \dots, N\}$, $k_i := \|e_i\|_{\mathrm{Lip}} \neq 0$. Moreover, using Proposition 4.1.1, we know that $\mathrm{M}\, e_i$ is $k_i(1-\kappa)$-Lipschitz. Since $\mathrm{M}\, e_i = \lambda_i e_i$, we get that $\lambda_i \leq (1-\kappa)$ for any $i \in \{2, \dots, N\}$. As a result, for any $f \in L^2(\mathcal{X}, \pi)$ and any $n \in \mathbb{N}$, using the fact that $e_1 \equiv 1$ and that M is self-adjoint, we get

$$\begin{aligned}
\mathrm{Var}_\pi\left(\mathrm{M}^n f\right) &= \|\, \mathrm{M}^n f - \mathbb{E}_\pi\left[\mathrm{M}^n f\right] \|^2_{L^2(\mathcal{X},\pi)} \\
&= \left\| \sum_{i=1}^N \langle \mathrm{M}^n f, e_i \rangle_{L^2(\mathcal{X},\pi)}\, e_i - \langle \mathrm{M}^n f, e_1 \rangle_{L^2(\mathcal{X},\pi)} \right\|^2_{L^2(\mathcal{X},\pi)} \\
&= \left\| \sum_{i=2}^N \langle f, \mathrm{M}^n e_i \rangle_{L^2(\mathcal{X},\pi)}\, e_i \right\|^2_{L^2(\mathcal{X},\pi)} \\
&= \left\| \sum_{i=2}^N \langle f, e_i \rangle_{L^2(\mathcal{X},\pi)}\, \lambda_i^n e_i \right\|^2_{L^2(\mathcal{X},\pi)} \\
&= \sum_{i=2}^N \langle f, e_i \rangle^2_{L^2(\mathcal{X},\pi)}\, \lambda_i^{2n} \\
&\leq (1-\kappa)^{2n} \sum_{i=2}^N \langle f, e_i \rangle^2_{L^2(\mathcal{X},\pi)} \\
&= (1-\kappa)^{2n} \left\| \sum_{i=2}^N \langle f, e_i \rangle_{L^2(\mathcal{X},\pi)}\, e_i \right\|^2_{L^2(\mathcal{X},\pi)} \\
&= (1-\kappa)^{2n} \|f - \mathbb{E}_\pi[f]\|^2_{L^2(\mathcal{X},\pi)} \\
&= (1-\kappa)^{2n}\, \mathrm{Var}_\pi(f).
\end{aligned}$$

□

We can also obtain this result by working in the subspace $L^2(\mathcal{X}, \pi)/\{\text{const}\}$ of $L^2(\mathcal{X}, \pi)$ with π reversible, defined as follows

$$L^2(\mathcal{X}, \pi)/\{\text{const}\} = \left\{ f \in L^2(\mathcal{X}, \pi) \,\middle|\, \mathbb{E}_\pi[f] = 0 \right\}.$$

For any k-Lipschitz function $f \in L^2(\mathcal{X}, \pi)/\{\text{const}\}$, $\mathrm{M}^n f$ is $k(1-\kappa)^n$-Lipschitz. Now, let us define C_k as follows

$$C_k := \sup_{g\, k\text{-Lipschitz}} \mathrm{Var}_\pi(g).$$

Since $\mathcal{X}$ is finite, C_k is finite. As a result, we get that $\mathrm{Var}_\pi\left(M^n f\right) \leq (1-\kappa)^{2n} C_k$. As $f \in L^2(\mathcal{X}, \pi)/\{\text{const}\}$, we also have

$$\begin{aligned}\mathrm{Var}_\pi\left(M^n f\right) &= \langle \mathrm{M}^n f, \mathrm{M}^n f\rangle_{L^2(\mathcal{X},\pi)} \\ &= \left\langle \mathrm{M}^{n-1} f, \mathrm{M}^{n+1} f\right\rangle_{L^2(\mathcal{X},\pi)} \\ &\leq \left\langle \mathrm{M}^{n-1} f, \mathrm{M}^{n-1} f\right\rangle^{1/2}_{L^2(\mathcal{X},\pi)} \left\langle \mathrm{M}^{n+1} f, \mathrm{M}^{n+1} f\right\rangle^{1/2}_{L^2(\mathcal{X},\pi)}.\end{aligned}$$

Defining $\iota : \mathbb{N} \to \mathbb{R}$ as follows

$$\iota(n) := \ln\left((1-\kappa)^{-2n}\,\mathrm{Var}_\pi\left(M^n f\right)\right) = \ln\left(\langle \mathrm{M}^n f, \mathrm{M}^n f\rangle_{L^2(\mathcal{X},\pi)}\right) - 2n\ln(1-\kappa),$$

we get

$$\iota(n) \leq \frac{1}{2}\iota(n-1) + \frac{1}{2}\iota(n+1),$$

meaning that ι is convex. Hence, if for some $n \in \mathbb{N}$, there exists $\varepsilon > 0$ such that $\iota(n) \geq \iota(n-1) + \varepsilon$, it means that $\iota(n+1) \geq \iota(n) + \varepsilon$ and ι grows to infinity as $n \to \infty$. However, that would contradict the fact that $\mathrm{Var}_\pi\left(M^n f\right) \leq (1-\kappa)^{2n} C_k$. Therefore, we get that for any $n \in \mathbb{N}$, $\iota(n) \leq \iota(0)$, which translates into

$$\mathrm{Var}_\pi\left(M^n f\right) \leq (1-\kappa)^{2n}\,\mathrm{Var}_\pi\left(f\right).$$

Since adding a constant to f has no impact on the variance, we get that the above inequation is true for any $f \in L^2(\mathcal{X}, \pi)$.

4.3 Poincaré inequality

Using the variance contraction result we obtained in the previous section, we will now show how we can derive a Poincaré inequality. Firstly, we need to show the following variance decomposition.

Proposition 4.3.1. *We have*

$$\mathrm{Var}_\pi(f) = \sum_{x\in\mathcal{X}} \mathrm{Var}_{m_x}(f)\,\pi(x) + \mathrm{Var}_\pi(\mathrm{M}\,f).$$

Proof. We know that

$$\mathrm{Var}_\pi(f) = \mathbb{E}_\pi\left[f^2\right] - \mathbb{E}_\pi\left[f\right]^2.$$

Now, using the law of total expectation, we can get

$$\mathbb{E}_\pi\left[f^2\right] = \sum_{x\in\mathcal{X}} \left(\mathrm{Var}_{m_x}(f) + \mathbb{E}_{m_x}\left[f\right]^2\right)\pi(x).$$

As a result, using again the law of total expectation as well as the fact that for any $x \in \mathcal{X}$, $\mathbb{E}_{m_x}(f) = \mathrm{M}\,f(x)$, we get

$$\mathrm{Var}_\pi(f) = \sum_{x\in\mathcal{X}} \left(\mathrm{Var}_{m_x}(f) + \mathbb{E}_{m_x}\left[f\right]^2\right)\pi(x) - \left(\sum_{x\in\mathcal{X}} \mathbb{E}_{m_x}\left[f\right]\pi(x)\right)^2$$

$$= \sum_{x\in\mathcal{X}} \mathrm{Var}_{m_x}(f)\,\pi(x) + \sum_{x\in\mathcal{X}} \mathbb{E}_{m_x}[f]^2\,\pi(x) - \left(\sum_{x\in\mathcal{X}} \mathbb{E}_{m_x}[f]\,\pi(x)\right)^2$$
$$= \sum_{x\in\mathcal{X}} \mathrm{Var}_{m_x}(f)\,\pi(x) + \sum_{x\in\mathcal{X}} (\mathrm{M}\,f(x))^2\,\pi(x) - \left(\sum_{x\in\mathcal{X}} \mathrm{M}\,f(x)\,\pi(x)\right)^2$$
$$= \sum_{x\in\mathcal{X}} \mathrm{Var}_{m_x}(f)\,\pi(x) + \mathrm{Var}_\pi(\mathrm{M}\,f). \qquad \square$$

Corollary 4.3.1. *Suppose that the Ricci curvature of $\mathcal{X}$ is at least $\kappa > 0$ and that π is reversible for the Markov chain m. Then, the following discrete Poincaré inequality is satisfied for any $f \in L^2(\mathcal{X}, \pi)$*

$$\mathrm{Var}_\pi(f) \leq \frac{1}{\kappa(2-\kappa)} \sum_{x\in\mathcal{X}} \mathrm{Var}_{m_x}(f)\,\pi(x).$$

Proof. Using the inequality in Theorem 4.2.1 with $n = 1$, we can rewrite the equality of Proposition 4.3.1 to obtain the following

$$\begin{aligned}\sum_{x\in\mathcal{X}} \mathrm{Var}_{m_x}(f)\,\pi(x) &= \mathrm{Var}_\pi(f) - \mathrm{Var}_\pi(\mathrm{M}\,f)\\ &\geq \left(1-(1-\kappa)^2\right)\mathrm{Var}_\pi(f)\\ &= \kappa(2-\kappa)\,\mathrm{Var}_\pi(f). \qquad \square\end{aligned}$$

As we can see with the above Poincaré inequality, Ollivier's Ricci curvature again has a direct impact on the bound of $\mathrm{Var}_\pi(f)$. This again underlines its importance when studying a given Markov chain, in particular its convergence.

5 Ollivier's Ricci curvature bound on undirected graphs

In the previous chapter, we have seen that Ollivier's Ricci curvature is key when studying the convergence of Markov chains. With the first Lemma of this section (5.0.1), we will see that reversible Markov chains can be associated with random walks on undirected graphs. As such, we will study undirected graphs to derive lower bounds for Ollivier's Ricci curvature. We will denote by $G = (V, E)$ a finite undirected connected graph without loops and multiple edges. We will work on unweighted graphs before generalizing our results on weighted graphs. If $x, y \in V$ are neighbors, we write $x \sim y$. In the case of weighted graphs, we denote by w_{xy} the weight associated to $x, y \in V$, where $x \sim y$ (we may simply put $w_{xy} = 0$ if x and y are not neighbors, to simplify the notation). The unweighted case corresponds to $w_{xy} = 1$ whenever $x \sim y$. The degree of $x \in V$ is $d_x = \sum_{y,y\sim x} w_{xy}$. Since we work with undirected graphs, note that w_{xy} and w_{yx} are equivalent notations.

Lemma 5.0.1. *Let $(X_n)_{n\in\mathbb{N}}$ be a reversible Markov chain on a finite state space $\mathcal{X}$, with transition probability matrix m and stationary distribution π. There exists a random walk on an undirected weighted graph associated with the transition probability matrix m.*

Proof. We construct a graph $G = (\mathcal{X}, E)$, where $(x, y) \in E$ if and only if $m_x(y) > 0$. For any $(x, y) \in E$, we define its associated weight as

$$w_{xy} := \pi(x)\,m_x(y) = \pi(y)\,m_y(x).$$

With this construction, we get that for any $x \in \mathcal{X}$,

$$d_x = \sum_{y \in \mathcal{X} \,:\, (x,y) \in E} w_{xy} = \sum_{y \in \mathcal{X} \,:\, m_x(y) > 0} \pi(x)\, m_x(y) = \pi(x).$$

As such, for any $x, y \in \mathcal{X}$, $m_y(x) = \frac{w_{xy}}{\pi(y)} = \frac{w_{xy}}{d_y}$, meaning that the transition matrix associated with the random walk is exactly m. □

Now that we know that there exists a correspondence between reversible Markov chains and random walks on undirected graphs, let us focus on finding a lower bound for Ollivier's Ricci curvature using undirected graphs.

5.1 Unweighted graphs

We will prove the following theorem presented in [1]. The Markov chain used, m, puts equal weight on all neighbors. We make the assumption that $m_x(x) = 0$ for any $x \in \mathcal{X}$ in order not to have loops in the associated graph.

Theorem 5.1.1. *On an unweighted graph $G = (V, E)$, we put for any pair of neighboring vertices x, y,*

$$\#(x,y) := \text{ number of triangles which include } x, y \text{ as vertices } = \sum_{z, z\sim x, z\sim y} 1.$$

We then have

$$\kappa(x,y) \geq -\left(1 - \frac{1}{d_x} - \frac{1}{d_y} - \frac{\#(x,y)}{d_x \wedge d_y}\right)_+ - \left(1 - \frac{1}{d_x} - \frac{1}{d_y} - \frac{\#(x,y)}{d_x \vee d_y}\right)_+ + \frac{\#(x,y)}{d_x \vee d_y},$$

where $s_+ := \max(s, 0)$, $s \vee t := \max(s, t)$, $s \wedge t := \min(s, t)$. Moreover, this inequality is sharp for certain graphs.

Proof. To begin with, we assume $\#(x, y) = 0$, so that the inequality we want to prove reduces to

$$\kappa(x,y) \geq -2\left(1 - \frac{1}{d_x} - \frac{1}{d_y}\right)_+ = \begin{cases} -2 + \frac{2}{d_x} + \frac{2}{d_y}, & \text{if } d_x > 1 \text{ and } d_y > 1; \\ 0, & \text{otherwise.} \end{cases}$$

Firstly, if $d_x = 1$ (or $d_y = 1$), then we directly have $\kappa(x, y) = 0$ since all the weight of m_x (or m_y) is at y (or x). Therefore, we assume that both $d_x > 1$ and $d_y > 1$. Using Kantorovich-Rubinstein duality, we have

$$\begin{aligned} \mathrm{W}_1(m_x, m_y) &= \sup_{f, 1\text{-Lipschitz}} \left(\frac{1}{d_x} \sum_{z, z\sim x} f(z) - \frac{1}{d_y} \sum_{z', z'\sim y} f(z')\right) \\ &= \sup_{f, 1\text{-Lipschitz}} \left(\frac{1}{d_x} \sum_{z, z\sim x, z\neq y} (f(z) - f(x)) - \frac{1}{d_y} \sum_{z', z'\sim y, z'\neq x} (f(z') - f(y)) + C\right) \\ &\leq \frac{d_x - 1}{d_x} + \frac{d_y - 1}{d_y} + \left|1 - \frac{1}{d_x} - \frac{1}{d_y}\right| \end{aligned}$$

$$= 2 - \frac{1}{d_x} - \frac{1}{d_y} + \left|1 - \frac{1}{d_x} - \frac{1}{d_y}\right|$$
$$= 1 + 2\left(1 - \frac{1}{d_x} - \frac{1}{d_y}\right)_+,$$

where $C = (f(x) - f(y))\left(1 - \frac{1}{d_x} - \frac{1}{d_y}\right)$. Now, since $x \sim y$, we have $\kappa(x,y) = 1 - \mathrm{W}_1(m_x, m_y)$ and so

$$\kappa(x,y) \geq -2\left(1 - \frac{1}{d_x} - \frac{1}{d_y}\right)_+.$$

Let us now show that this lower bound can be attained. Let $f : V \to \mathbb{R}$ be a 1-Lipschitz function defined as follows for the vertices $v \in V$ of interest

$$f(v) = \begin{cases} 0, & \text{at } y\text{'s own neighbors;} \\ 1, & \text{at } y; \\ 2, & \text{at } x; \\ 3, & \text{at } x\text{'s own neighbors.} \end{cases}$$

Using Kantorovich-Rubinstein duality, we have

$$\mathrm{W}_1(m_x, m_y) \geq \frac{1}{d_x}\left(3(d_x - 1) + f(y)\right) - \frac{1}{d_y} f(x)$$
$$= 3 - \frac{2}{d_x} - \frac{2}{d_y},$$

and thus,

$$\kappa(x,y) \leq -2\left(1 - \frac{1}{d_x} - \frac{1}{d_y}\right)_+.$$

Before continuing with the case $\#(x,y) \geq 1$, we fix some notations. The vertices z that are adjacent to x or y, where $x \sim y$, are divided into three classes:

- common neighbors of x and y: $z \sim x$ and $z \sim y$;
- x's own neighbors: $z \sim x$, $z \nsim y$, $z \neq y$;
- y's own neighbors: $z \sim y$, $z \nsim x$, $z \neq x$.

We assume without loss of generality that $d_x = d_x \vee d_y$ and $d_y = d_x \wedge d_y$. In principle, our transfer plan moving m_x to m_y should be as follows.

1. Move the mass of $\frac{1}{d_x}$ from y to y's own neighbors;
2. Move a mass of $\frac{1}{d_y}$ from x's own neighbors to x;
3. Fill the remaining gaps using the mass at x's own neighbors. Filling the gap at common neighbors costs 2 and the one at y's own neighbors costs 3.

A critical point will be whether the steps 1 and 2 can be realized or not. It is easy to see that we can realize step 1 if and only if

$$1-\frac{1}{d_y}-\frac{\#(x,y)}{d_y}\geq\frac{1}{d_x}\quad\text{or}\quad A:=1-\frac{1}{d_x}-\frac{1}{d_y}-\frac{\#(x,y)}{d_x\wedge d_y}\geq 0.$$

That is, after taking off the masses at x and at the common neighbors, m_y still has at least a mass of $\frac{1}{d_x}$. Since $d_x\geq d_y$, it means that y needs at least one own neighbor. Step 2 can be realized if and only if

$$1-\frac{1}{d_x}-\frac{\#(x,y)}{d_x}\geq\frac{1}{d_y}\quad\text{or}\quad B:=1-\frac{1}{d_x}-\frac{1}{d_y}-\frac{\#(x,y)}{d_x\vee d_y}\geq 0.$$

That is, after taking off the masses at y and at the common neighbors, m_x still has enough mass to fill $\frac{1}{d_y}$. Clearly, we have $A\leq B$. We will divide the discussion into three cases according to whether the first two steps can be realized or not.

(1) Assume $0\leq A\leq B$. This means that we can adopt the above transfer plan. We get

$$\begin{aligned}\mathrm{W}_1(m_x,m_y)&\leq\frac{1}{d_x}\times 1+\frac{1}{d_y}\times 1+\left(\frac{1}{d_y}-\frac{1}{d_x}\right)\times\#(x,y)\times 2\\&\quad+\left(1-\frac{1}{d_x}-\frac{1}{d_y}-\left(\frac{1}{d_y}-\frac{1}{d_x}\right)\times\#(x,y)-\frac{1}{d_x}\times\#(x,y)\right)\times 3\\&=3-\frac{2}{d_x}-\frac{2}{d_y}-\frac{\#(x,y)}{d_y}-\frac{2\#(x,y)}{d_x},\end{aligned}$$

or in a symmetrical way,

$$\mathrm{W}_1(m_x,m_y)\leq 3-\frac{2}{d_x}-\frac{2}{d_y}-\frac{\#(x,y)}{d_x\wedge d_y}-\frac{2\#(x,y)}{d_x\vee d_y}.$$

In conclusion, we have

$$\kappa(x,y)\geq -2+\frac{2}{d_x}+\frac{2}{d_y}+\frac{\#(x,y)}{d_x\wedge d_y}+\frac{2\#(x,y)}{d_x\vee d_y}.$$

Let us now show that this lower bound can be attained. Let $f:V\to\mathbb{R}$ be a 1-Lipschitz function defined as follows for the vertices $v\in V$ of interest

$$f(v)=\begin{cases}0, & \text{at } y\text{'s own neighbors;}\\ 1, & \text{at } y \text{ and common neighbors;}\\ 2, & \text{at } x;\\ 3, & \text{at } x\text{'s own neighbors.}\end{cases}$$

Using Kantorovich-Rubinstein duality, we have

$$\mathrm{W}_1(m_x,m_y)\geq\frac{1}{d_x}\left(f(y)+3(d_x-1-\#(x,y))+\#(x,y)\right)-\frac{1}{d_y}\left(f(x)+\#(x,y)\right)$$

$$= 3 - \frac{2}{d_x} - \frac{2}{d_y} - \frac{\#(x,y)}{d_y} - \frac{2\#(x,y)}{d_x}.$$

Thus, we have

$$\kappa(x,y) \leq -2 + \frac{2}{d_x} + \frac{2}{d_y} + \frac{\#(x,y)}{d_x \wedge d_y} + \frac{2\#(x,y)}{d_x \vee d_y}.$$

(2) Assume $A < 0 \leq B$. In this case, we cannot realize step 1 but step 2 can be realized. $A < 0$ implies that y has no own neighbors. Our transfer plan should be step 2 at first. $B \geq 0$ can be rewritten as

$$1 - \frac{1}{d_y} - \frac{\#(x,y)}{d_x} \geq \frac{1}{d_x},$$

meaning that we can move the mass of $\frac{1}{d_x}$ at y for distance 1 to common neighbors. Finally, we fill the gap at common neighbors for distance 2. In a formula, this gives

$$\begin{aligned}\mathrm{W}_1(m_x, m_y) &\leq \frac{1}{d_x} \times 1 + \frac{1}{d_y} \times 1 + \left(1 - \frac{1}{d_x} - \frac{1}{d_y} - \frac{\#(x,y)}{d_x}\right) \times 2 \\ &= 2 - \frac{1}{d_x} - \frac{1}{d_y} - \frac{2\#(x,y)}{d_x},\end{aligned}$$

or in a symmetrical manner,

$$\mathrm{W}_1(m_x, m_y) \leq 2 - \frac{1}{d_x} - \frac{1}{d_y} - \frac{2\#(x,y)}{d_x \vee d_y}.$$

In conclusion, we get

$$\kappa(x,y) \geq -1 + \frac{1}{d_x} + \frac{1}{d_y} + \frac{2\#(x,y)}{d_x \vee d_y}.$$

Let us show that this lower bound can be attained. Let $f : V \to \mathbb{R}$ be a 1-Lipschitz function defined as follows for the vertices $v \in V$ of interest

$$f(v) = \begin{cases} 0, & \text{at common neighbors;} \\ 1, & \text{at } x \text{ and } y; \\ 2, & \text{at } x\text{'s own neighbors.} \end{cases}$$

Using Kantorovich-Rubinstein duality, we have

$$\begin{aligned}\mathrm{W}_1(m_x, m_y) &\geq \frac{1}{d_x}\left(f(y) + 2(d_x - 1 - \#(x,y))\right) - \frac{1}{d_y}\left(f(x)\right) \\ &= 2 - \frac{1}{d_x} - \frac{1}{d_y} - \frac{2\#(x,y)}{d_x}.\end{aligned}$$

Thus, we get

$$\kappa(x,y) \leq -1 + \frac{1}{d_x} + \frac{1}{d_y} + \frac{2\#(x,y)}{d_x \vee d_y}.$$

(3) Assume $A \leq B < 0$. In this case, the steps 1 and 2 are not applicable. Also, y has no own neighbor, and $B < 0$ implies that we can move all the mass at x's own neighbors to x at first. Then, we move the mass of $\frac{1}{d_x}$ at y for distance 1 to fill the remaining gaps. In a formula, it gives

$$\mathrm{W}_1(m_x, m_y) \leq \left(1 - \frac{\#(x,y)}{d_x}\right) \times 1 = 1 - \frac{\#(x,y)}{d_x},$$

or in a symmetrical way,

$$\mathrm{W}_1(m_x, m_y) \leq 1 - \frac{\#(x,y)}{d_x \vee d_y}.$$

In conclusion,

$$\kappa(x,y) \geq \frac{\#(x,y)}{d_x \vee d_y}.$$

Let us show that this lower bound can be attained. Let $f : V \to \mathbb{R}$ be a 1-Lipschitz function defined as follows for the vertices $v \in V$ of interest

$$f(v) = \begin{cases} 0, & \text{at } x \text{ and common neighbors;} \\ 1, & \text{at } y \text{ and } x\text{'s own neighbors.} \end{cases}$$

Now, using Kantorovich-Rubinstein duality, we have

$$\begin{aligned} \mathrm{W}_1(m_x, m_y) &\geq \frac{1}{d_x}\left(f(y) + d_x - 1 - \#(x,y)\right) - \frac{1}{d_y} \times 0 \\ &= 1 - \frac{\#(x,y)}{d_x}. \end{aligned}$$

Thus,

$$\kappa(x,y) \leq \frac{\#(x,y)}{d_x \vee d_y},$$

and we are done. □

5.2 Weighted graphs

We can now extend our last result to weighted graphs as follows. The Theorem below is applicable for any reversible transition matrix m verifying $m_{xx} = 0$ for any $x \in \mathcal{X}$.

Theorem 5.2.1. *On a weighted graph $G = (V, E)$, we have for any pair of neighboring vertices x, y,*

$$\begin{aligned} \kappa(x,y) \geq &- \left(1 - \frac{w_{xy}}{d_x} - \frac{w_{xy}}{d_y} - \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \vee \frac{w_{yz}}{d_y}\right)_+ \\ &- \left(1 - \frac{w_{xy}}{d_x} - \frac{w_{xy}}{d_y} - \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y}\right)_+ \\ &+ \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y}. \end{aligned}$$

Moreover, this inequality is sharp for certain graphs.

Proof. Similarly to the proof of the previous Theorem (5.1.1), we need to introduce the following two terms:

$$A := 1 - \frac{w_{xy}}{d_x} - \frac{w_{xy}}{d_y} - \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \vee \frac{w_{yz}}{d_y},$$

and

$$B := 1 - \frac{w_{xy}}{d_x} - \frac{w_{xy}}{d_y} - \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y}.$$

Let us also introduce two sets S_x and S_y, defined as follows

$$S_x := \left\{ z \in V \,\middle|\, z \sim x, z \sim y, \frac{w_{xz}}{d_x} > \frac{w_{yz}}{d_y} \right\},$$

and

$$S_y := \left\{ z \in V \,\middle|\, z \sim x, z \sim y, \frac{w_{xz}}{d_x} \leq \frac{w_{yz}}{d_y} \right\}.$$

Note that S_x and S_y form a partition of the common neighbors of x and y. Our transfer plan moving m_x to m_y should be similar to the previous proof. We will divide the discussion into three cases as follows.

(1) Assume $0 \leq A \leq B$. The fact that $0 \leq A$ gives us

$$\frac{w_{xy}}{d_x} + \sum_{z, z\sim x, z\sim y} \left(\frac{w_{xz}}{d_x} \vee \frac{w_{yz}}{d_y} - \frac{w_{yz}}{d_y} \right) \leq 1 - \frac{w_{xy}}{d_y} - \sum_{z, z\sim x, z\sim y} \frac{w_{yz}}{d_y},$$

meaning that we can transport the mass of $\frac{w_{xy}}{d_x}$ at y as well as all the extra mass

$$C_x := \sum_{z, z\sim x, z\sim y} \left(\frac{w_{xz}}{d_x} \vee \frac{w_{yz}}{d_y} - \frac{w_{yz}}{d_y} \right) = \sum_{z, z\sim x, z\sim y} \left(\frac{w_{xz}}{d_x} - \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y} \right)$$

at the common neighbors $z \in S_x$ to y's own neighbors for distance 1 and 2 respectively. Then, using the mass at x's own neighbors, we can fill the gap of $\frac{w_{xy}}{d_y}$ at x for distance 1. Subsequently, we can fill the gap of

$$C_y := \sum_{z, z\sim x, z\sim y} \left(\frac{w_{xz}}{d_x} \vee \frac{w_{yz}}{d_y} - \frac{w_{xz}}{d_x} \right) = \sum_{z, z\sim x, z\sim y} \left(\frac{w_{yz}}{d_y} - \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y} \right)$$

at the common neighbors $z \in S_y$ for distance 2 as well as the remaining gap at y's own neighbors for distance 3. Using the fact that

$$C_x + C_y = \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \vee \frac{w_{yz}}{d_y} - \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y},$$

we obtain

$$\mathrm{W}_1(m_x, m_y) \leq \frac{w_{xy}}{d_x} \times 1 + \frac{w_{xy}}{d_y} \times 1 + C_x \times 2 + C_y \times 2$$

$$
\begin{aligned}
&+\left(1-\frac{w_{xy}}{d_x}-\frac{w_{xy}}{d_y}-C_x-C_y-\sum_{z,z\sim x,z\sim y}\frac{w_{xz}}{d_x}\wedge\frac{w_{yz}}{d_y}\right)\times 3\\
&=3-\frac{2w_{xy}}{d_x}-\frac{2w_{xy}}{d_y}-C_x-C_y-3\sum_{z,z\sim x,z\sim y}\frac{w_{xz}}{d_x}\wedge\frac{w_{yz}}{d_y}\\
&=3-\frac{2w_{xy}}{d_x}-\frac{2w_{xy}}{d_y}-\sum_{z,z\sim x,z\sim y}\frac{w_{xz}}{d_x}\vee\frac{w_{yz}}{d_y}-2\sum_{z,z\sim x,z\sim y}\frac{w_{xz}}{d_x}\wedge\frac{w_{yz}}{d_y}.
\end{aligned}
$$

This gives us

$$
\kappa(x,y)\geq -2+\frac{2w_{xy}}{d_x}+\frac{2w_{xy}}{d_y}+\sum_{z,z\sim x,z\sim y}\frac{w_{xz}}{d_x}\vee\frac{w_{yz}}{d_y}+2\sum_{z,z\sim x,z\sim y}\frac{w_{xz}}{d_x}\wedge\frac{w_{yz}}{d_y}.
$$

Let us show that this lower bound can be attained. Let $f:V\to\mathbb{R}$ be a 1-Lipschitz function defined as follows for the vertices $v\in V$ of interest

$$
f(v)=\begin{cases}0, & \text{at } y\text{'s own neighbors;}\\ 1, & \text{at } y \text{ and common neighbors } z\in S_y;\\ 2, & \text{at } x \text{ and common neighbors } z\in S_x;\\ 3, & \text{at } x\text{'s own neighbors.}\end{cases}
$$

Using Kantorovich-Rubinstein duality, we have

$$
\begin{aligned}
\mathrm{W}_1(m_x,m_y)\geq\ &\frac{w_{xy}}{d_x}+\sum_{z\in S_y}\frac{w_{xz}}{d_x}\wedge\frac{w_{yz}}{d_y}+2\times\left(C_x+\sum_{z\in S_x}\frac{w_{xz}}{d_x}\wedge\frac{w_{yz}}{d_y}\right)\\
&+3\times\left(1-\frac{w_{xy}}{d_x}-C_x-\sum_{z,z\sim x,z\sim y}\frac{w_{xz}}{d_x}\wedge\frac{w_{yz}}{d_y}\right)\\
&-\frac{2w_{xy}}{d_y}-2\sum_{z\in S_x}\frac{w_{xz}}{d_x}\wedge\frac{w_{yz}}{d_y}-\sum_{z\in S_y}\frac{w_{xz}}{d_x}\wedge\frac{w_{yz}}{d_y}-C_y\\
=\ &3-\frac{2w_{xy}}{d_x}-\frac{2w_{xy}}{d_y}-C_x-C_y-3\sum_{z,z\sim x,z\sim y}\frac{w_{xz}}{d_x}\wedge\frac{w_{yz}}{d_y}\\
=\ &3-\frac{2w_{xy}}{d_x}-\frac{2w_{xy}}{d_y}-\sum_{z,z\sim x,z\sim y}\frac{w_{xz}}{d_x}\vee\frac{w_{yz}}{d_y}-2\sum_{z,z\sim x,z\sim y}\frac{w_{xz}}{d_x}\wedge\frac{w_{yz}}{d_y},
\end{aligned}
$$

and we are done.

(2) Assume $A<0\leq B$. The fact that $A<0$ gives us

$$
1-\frac{w_{xy}}{d_y}-\sum_{z,z\sim x,z\sim y}\frac{w_{yz}}{d_y}<\frac{w_{xy}}{d_x}+\sum_{z,z\sim x,z\sim y}\left(\frac{w_{xz}}{d_x}\vee\frac{w_{yz}}{d_y}-\frac{w_{yz}}{d_y}\right),
$$

meaning that the gap at y's own neighbors can be entirely filled using the mass of $\frac{w_{xy}}{d_x}$ at y for distance 1 and the mass of C_x at the common neighbors $z\in S_x$ for distance 2. Then, the fact that $0\leq B$ gives us

$$
\frac{w_{xy}}{d_x}\leq 1-\frac{w_{xy}}{d_y}-\sum_{z,z\sim x,z\sim y}\frac{w_{xz}}{d_x}\wedge\frac{w_{yz}}{d_y},
$$

meaning that combined, the gap of C_y at the common neighbors $z \in S_y$ and the gap at y's own neighbors are big enough to welcome the mass of $\frac{w_{xy}}{d_x}$ from y. Hence, no parts of the mass of $\frac{w_{xy}}{d_x}$ at y needs to be sent to x to fill the gap of $\frac{w_{xy}}{d_y}$.

In short, the mass of $\frac{w_{xy}}{d_x}$ at y can be transported to m_y for distance 1 and the gap of $\frac{w_{xy}}{d_y}$ at x can be filled for distance 1 separately. Then all the remaining gaps of m_y can be filled for distance 2. This gives us

$$\begin{aligned}\mathrm{W}_1(m_x, m_y) &\leq \frac{w_{xy}}{d_x} \times 1 + \frac{w_{xy}}{d_y} \times 1 + \left(1 - \frac{w_{xy}}{d_x} - \frac{w_{xy}}{d_y} - \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y}\right) \times 2 \\ &= 2 - \frac{w_{xy}}{d_x} - \frac{w_{xy}}{d_y} - 2 \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y}.\end{aligned}$$

As a result,

$$\kappa(x, y) \geq -1 + \frac{w_{xy}}{d_x} + \frac{w_{xy}}{d_y} + 2 \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y}.$$

Now, let us show that this lower bound can be attained. Let $f : V \to \mathbb{R}$ be a 1-Lipschitz function defined as follows for the vertices $v \in V$ of interest

$$f(v) = \begin{cases} 0, & \text{at } y\text{'s own neighbors and common neighbors } z \in S_y; \\ 1, & \text{at } x \text{ and } y; \\ 2, & \text{at } x\text{'s own neighbors and common neighbors } z \in S_x. \end{cases}$$

Using Kantorovich-Rubinstein duality, we have

$$\begin{aligned}\mathrm{W}_1(m_x, m_y) \geq{}& \frac{w_{xy}}{d_x} + 2 \times \left(C_x + \sum_{z \in S_x} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y}\right) \\ &+ 2 \times \left(1 - \frac{w_{xy}}{d_x} - C_x - \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y}\right) \\ &- \frac{w_{xy}}{d_y} - 2 \sum_{z \in S_x} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y} \\ ={}& 2 - \frac{w_{xy}}{d_x} - \frac{w_{xy}}{d_y} - 2 \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y}\end{aligned}$$

and we are done.

(3) Assume $A \leq B < 0$. The fact that $B < 0$ gives us

$$1 - \frac{w_{xy}}{d_y} - \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y} < \frac{w_{xy}}{d_x},$$

meaning that the gap at y's own neighbors and the gap of C_y at the common neighbors $z \in S_y$ can be entirely filled using only the mass of $\frac{w_{xy}}{d_x}$ from y, for distance 1. Then, the mass of $\frac{w_{xy}}{d_y}$ at x remains to be filled, which can be done for distance 1. As a result, we get

$$\mathrm{W}_1(m_x, m_y) \leq 1 - \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y}.$$

Hence, we obtain

$$\kappa(x,y) \geq \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y}.$$

Let us now show that this lower bound can be attained. Let $f : V \to \mathbb{R}$ be a 1-Lipschitz function defined as follows for the vertices $v \in V$ of interest

$$f(v) = \begin{cases} 0, & \text{at } x, y\text{'s own neighbors and common neighbors } z \in S_y; \\ 1, & \text{at } y, x\text{'s own neighbors and common neighbors } z \in S_x. \end{cases}$$

Using Kantorovich-Rubinstein duality, we have

$$\begin{aligned} \mathrm{W}_1(m_x, m_y) &\geq \frac{w_{xy}}{d_x} + C_x + \sum_{z\in S_x} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y} \\ &\quad + \left(1 - \frac{w_{xy}}{d_x} - C_x - \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y}\right) - \sum_{z\in S_x} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y} \\ &= 1 - \sum_{z, z\sim x, z\sim y} \frac{w_{xz}}{d_x} \wedge \frac{w_{yz}}{d_y}. \end{aligned}$$

This concludes the proof. □

It is interesting to note that with the two Theorems 5.1.1 and 5.2.1, Ollivier's Ricci curvature is equal to the given lower bound if the shortest path between any distinct pair of neighbors $(z_1, z_2) \in \{v \in V \mid v \sim x \text{ or } v \sim y\}^2$ goes through x or y (x and y if z_1 is one of x's own neighbors and z_2 is one of y's own neighbors, or the other way around). In fact, we have to make that assumption when constructing the 1-Lipschitz function to prove that the inequality is sharp using Kantorovich-Rubinstein duality. However, if it is not the case, the lower bound can be quite poor. Take the example of the random walk on the n-dimensional cube that puts equal weight on all neighbors, i.e., with the notation used in Example 3.3.1, for any $x \in \{0,1\}^n$ and any $i \in \{1, \dots, n\}$, $m_x(x^i) = 1/n$. The Ricci curvature of this random walk is 0. However, the lower bound given by Theorem 5.1.1 is $-2 + 4/n \xrightarrow[n\to\infty]{} -2$.

6 Markov chain Monte Carlo (MCMC)

In this chapter, our goal is to approximate the integral

$$\pi(f) := \sum_{x\in\mathcal{X}} f(x)\,\pi(x).$$

The technique we will use is described in [3]. Our approach will consist in constructing and simulating an irreducible and aperiodic Markov chain $(X_k)_{k\in\mathbb{N}}$ on $\mathcal{X}$ with stationary distribution π. We will first wait for a time $T_0 \geq 0$ (the *burn-in*) so that the chain gets close to its stationary distribution π. Then, we will estimate $\pi(f)$ on the next $T \geq 1$ steps of the trajectory, with T large enough as follows

$$\hat{\pi}(f) := \frac{1}{T} \sum_{k=T_0+1}^{T_0+T} f(X_k).$$

Our goal is to establish an upper bound for the error $|\hat{\pi}(f) - \pi(f)|$, which will provide good deviation estimates and confidence intervals for $\pi(f)$. We assume that Ollivier's Ricci curvature is positive, i.e. there exists $\kappa > 0$ such that

$$\mathrm{W}_1(m_x, m_y) \leq (1-\kappa)\, d(x,y)$$

for any $x, y \in \mathcal{X}$. Let us introduce the *diffusion constant* $\sigma(x)$ of the Markov chain at point $x \in \mathcal{X}$, which controls the size of the steps, defined by

$$\sigma(x) := \left(\frac{1}{2} \sum_{(y,z) \in \mathcal{X} \times \mathcal{X}} d(y,z)\, m_x(y)\, m_x(z) \right)^{1/2}.$$

Let the *local dimension* n_x at point $x \in \mathcal{X}$ be given by

$$n_x := \inf_{f \text{ 1-Lipschitz}} \frac{\sum_{\mathcal{X}\times\mathcal{X}} d(y,z)\, m_x(y)\, m_x(z)}{\sum_{\mathcal{X}\times\mathcal{X}} |f(y) - f(z)|\, m_x(y)\, m_x(z)} \geq 1.$$

Finally, we will denote by $\|\cdot\|_{\mathrm{Lip}}$ the usual *Lipschitz seminorm* of a function f on $\mathcal{X}$, defined as follows

$$\|f\|_{\mathrm{Lip}} := \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x,y)}.$$

Our goal is to bound the **mean quadratic error**

$$\mathbb{E}_x \left[|\hat{\pi}(f) - \pi(f)|^2 \right].$$

given any starting point $x \in \mathcal{X}$ for the Markov chain. There are two contributions to this error, a *variance* part, controlling how $\hat{\pi}(f)$ differs between two independent runs both starting at x, and a *bias* part, which is the difference between $\pi(f)$ and the average value of $\hat{\pi}(f)$, starting at x. Namely, the mean quadratic error decomposes as the sum of the squared bias plus the variance as follows

$$\begin{aligned}
\mathbb{E}_x \left[|\hat{\pi}(f) - \pi(f)|^2 \right] &= \mathbb{E}_x \left[\hat{\pi}(f)^2 \right] - \mathbb{E}_x \left[2\hat{\pi}(f)\pi(f) \right] + \mathbb{E}_x \left[\pi(f)^2 \right] \\
&= \mathbb{E}_x \left[\hat{\pi}(f) \right]^2 - 2\pi(f)\mathbb{E}_x \left[\hat{\pi}(f) \right] + \pi(f)^2 + \mathbb{E}_x \left[\hat{\pi}(f)^2 \right] - \mathbb{E}_x \left[\hat{\pi}(f) \right]^2 \\
&= \left| \mathbb{E}_x \left[\hat{\pi}(f) \right] - \pi(f) \right|^2 + \mathrm{Var}_x\, \hat{\pi}(f).
\end{aligned}$$

As we will see, these two terms have different behaviors depending on T_0 and T. For instance, the bias is expected to decrease exponentially fast as the burn-in period T_0 increases, whereas if T is fixed, the variance term does not vanish as $T_0 \to \infty$.

6.1 Bias of empirical mean

To begin with, we will control the bias term, which depends on the starting point of the Markov chain.

Proposition 6.1.1 (Bias of empirical mean)**.** *For any Lipschitz function $f : \mathcal{X} \to \mathbb{R}$, we have the upper bound on the bias*

$$\left|\mathbb{E}_x\left[\hat{\pi}(f)\right] - \pi(f)\right| \leq \frac{(1-\kappa)^{T_0+1}}{\kappa T} E(x)\, \|f\|_{\mathrm{Lip}},$$

where $E(x) := \sum_{y\in\mathcal{X}} d(x,y)\,\pi(y)$ is the eccentricity at point $x \in \mathcal{X}$.

Proof. Let f be a Lipschitz function. We know from Proposition 4.1.1 that for any $k \in \mathbb{N}$, $\mathrm{M}^k f$ is $\left((1-\kappa)^k \|f\|_{\mathrm{Lip}}\right)$-Lipschitz. Hence, by invariance of π, we have

$$\begin{aligned}
\left|\mathbb{E}_x\left[\hat{\pi}(f)\right] - \pi(f)\right| &= \left|\frac{1}{T}\sum_{k=T_0+1}^{T_0+T} \mathrm{M}^k f(x) - \sum_{y\in\mathcal{X}} f(y)\,\pi(y)\right| \\
&= \frac{1}{T}\left|\sum_{k=T_0+1}^{T_0+T}\sum_{y\in\mathcal{X}} \left(\mathrm{M}^k f(x) - \mathrm{M}^k f(y)\right)\pi(y)\right| \\
&\leq \frac{1}{T}\sum_{k=T_0+1}^{T_0+T} (1-\kappa)^k \|f\|_{\mathrm{Lip}} \sum_{y\in\mathcal{X}} d(x,y)\,\pi(y) \\
&\leq \frac{(1-\kappa)^{T_0+1}}{\kappa T} E(x)\, \|f\|_{\mathrm{Lip}},
\end{aligned}$$

and we are done. □

6.2 Variance of empirical mean

Now that we have found a way to bound the bias term, we will focus on the variance term. Before stating our Theorem, we need the following preliminary lemma.

Lemma 6.2.1. *For any $N \in \mathbb{N}^*$ and any Lipschitz function f on $\mathcal{X}$, we have*

$$\mathcal{M}^N f := \mathrm{M}^N\left(f^2\right) - \left(\mathrm{M}^N f\right)^2 \leq \|f\|_{\mathrm{Lip}}^2 \sum_{k=0}^{N-1} (1-\kappa)^{2(N-1-k)}\, \mathrm{M}^k\left(\frac{\sigma^2}{n}\right).$$

Proof. We will do a proof by induction. By definition, for any Lipschitz function f on $\mathcal{X}$ and any $x \in \mathcal{X}$, we have

$$\mathcal{M}f(x) = \mathrm{M}\left(f^2\right)(x) - (\mathrm{M}\, f(x))^2 = \mathbb{E}_{m_x}\left[f^2\right] - \mathbb{E}_{m_x}\left[f\right]^2 = \mathrm{Var}_{m_x}\left(f\right) \leq \|f\|_{\mathrm{Lip}}^2 \frac{\sigma^2(x)}{n_x}.$$

Moreover, since $\mathrm{M}^N f$ is $\left((1-\kappa)^N \|f\|_{\mathrm{Lip}}\right)$-Lipschitz, we have

$$\mathcal{M}(\mathrm{M}^N f)(x) = \mathbb{E}_{m_x}\left[\left(\mathrm{M}^N f\right)^2\right] - \mathbb{E}_{m_x}\left[\mathrm{M}^N f\right]^2 = \mathrm{Var}_{m_x}\left(\mathrm{M}^N f\right) \leq \|f\|_{\mathrm{Lip}}^2 (1-\kappa)^{2N} \frac{\sigma^2(x)}{n_x}.$$

Now, assume that for some $N \in \mathbb{N}^*$, the inequality is verified. We have, for any Lipschitz function f on $\mathcal{X}$ and any $x \in \mathcal{X}$,

$$\mathcal{M}^{N+1} f(x) = \sum_{y\in\mathcal{X}}\sum_{z\in\mathcal{X}} f^2(z)\, m_y^N(z)\, m_x(y) - \left(\sum_{y\in\mathcal{X}}\sum_{z\in\mathcal{X}} f(z)\, m_y^N(z)\, m_x(y)\right)^2$$

$$
\begin{aligned}
&= \sum_{y\in\mathcal{X}}\sum_{z\in\mathcal{X}} f^2(z)\, m_y^N(z)\, m_x(y) - \sum_{y\in\mathcal{X}}\left(\sum_{z\in\mathcal{X}} f(z)\, m_y^N(z)\right)^2 m_x(y) \\
&\quad + \sum_{y\in\mathcal{X}}\left(\sum_{z\in\mathcal{X}} f(z)\, m_y^N(z)\right)^2 m_x(y) - \left(\sum_{y\in\mathcal{X}}\sum_{z\in\mathcal{X}} f(z)\, m_y^N(z)\, m_x(y)\right)^2 \\
&= \sum_{y\in\mathcal{X}}\left(\sum_{z\in\mathcal{X}} f^2(z)\, m_y^N(z) - \left(\sum_{z\in\mathcal{X}} f(z)\, m_y^N(z)\right)^2\right) m_x(y) \\
&\quad + \mathbb{E}_{m_x}\left[\left(\mathrm{M}^N f\right)^2\right] - \mathbb{E}_{m_x}\left[\mathrm{M}^N f\right]^2 \\
&= \sum_{y\in\mathcal{X}}\left(\mathrm{M}^N\left(f^2\right)(y) - \left(\mathrm{M}^N f(y)\right)^2\right) m_x(y) + \mathrm{Var}_{m_x}\left(\mathrm{M}^N f\right) \\
&\leq \sum_{y\in\mathcal{X}}\left(\|f\|_{\mathrm{Lip}}^2 \sum_{k=0}^{N-1}(1-\kappa)^{2(N-1-k)}\,\mathrm{M}^k\left(\frac{\sigma^2}{n}\right)(y)\right) m_x(y) + \mathrm{Var}_{m_x}\left(\mathrm{M}^N f\right) \\
&= \|f\|_{\mathrm{Lip}}^2 \sum_{k=0}^{N-1}(1-\kappa)^{2(N-1-k)} \sum_{y\in\mathcal{X}} \mathrm{M}^k\left(\frac{\sigma^2}{n}\right)(y)\, m_x(y) + \mathrm{Var}_{m_x}\left(\mathrm{M}^N f\right) \\
&\leq \|f\|_{\mathrm{Lip}}^2 \sum_{k=0}^{N-1}(1-\kappa)^{2(N-1-k)}\,\mathrm{M}^{k+1}\left(\frac{\sigma^2}{n}\right)(x) + \|f\|_{\mathrm{Lip}}^2(1-\kappa)^{2N}\left(\frac{\sigma^2}{n}\right)(x) \\
&= \|f\|_{\mathrm{Lip}}^2 \sum_{k=0}^{N}(1-\kappa)^{2(N-k)}\,\mathrm{M}^k\left(\frac{\sigma^2}{n}\right)(x),
\end{aligned}
$$

and we are done. □

Now, given a Lipshitz function f on $\mathcal{X}$, we will consider the following functional defined for any $x_T \in \mathcal{X}$ by

$$
f_{x_1,\dots,x_{T-1}}(x_T) := \frac{1}{T}\sum_{k=1}^{T} f(x_k),
$$

the coordinates $x_1,\dots,x_{T-1}$ being fixed. The function $f_{x_1,\dots,x_{T-1}}$ is $(\|f\|_{\mathrm{Lip}}/T)$-Lipschitz, hence $(\|f\|_{\mathrm{Lip}}/\kappa T)$-Lipschitz since $\kappa \leq 1$. Moreover, for each $k \in \{T-1, T-2, \dots, 2\}$, the conditional expectation of $\hat{\pi}(f)$ knowing $X_1 = x_1, \dots, X_k = x_k$ can be written in terms of a downward induction as follows

$$
f_{x_1,\dots,x_{k-1}}(x_k) := \sum_{x_{k+1}\in\mathcal{X}} f_{x_1,\dots,x_k}(x_{k+1})\, m_{x_k}(x_{k+1}),
$$

and

$$
f_{\varnothing}(x_1) := \sum_{x_2\in\mathcal{X}} f_{x_1}(x_2)\, m_{x_1}(x_2).
$$

Before moving on to the main Theorem of this part, we will need the following preliminary lemma, adapted from the first step of the proof of Lemma 3.2 in [2].

Lemma 6.2.2. *The functions $f_{x_1,\dots,x_{k-1}}$ for $k \in \{2,\dots,T\}$ and $f_{\varnothing}$ verify*

$$
\|f_{x_1,\dots,x_{k-1}}\|_{\mathrm{Lip}} \leq s_k \|f_{x_1,\dots,x_{T-1}}\|_{\mathrm{Lip}} \leq \frac{\|f\|_{\mathrm{Lip}}}{\kappa T} \quad \textit{and} \quad \|f_{\varnothing}\|_{\mathrm{Lip}} \leq s_1 \|f_{x_1,\dots,x_{T-1}}\|_{\mathrm{Lip}} \leq \frac{\|f\|_{\mathrm{Lip}}}{\kappa T},
$$

where $s_k = \sum_{j=0}^{T-k}(1-\kappa)^j$.

Proof. We will prove this statement by a downward recursive argument on k. Since $s_T = 1$, the property is trivially true for $k = T$. Assume now that the above inequation is satisfied for some $k \in \{2, \dots, T\}$. Firstly, letting $x_1, \dots, x_{k-2}, y, z, x_k \in \mathcal{X}$, we have

$$\begin{aligned}
|f_{x_1,\dots,x_{k-2},y}(x_k) - f_{x_1,\dots,x_{k-2},z}(x_k)| &= \Big| \sum_{\mathcal{X}^{T-k}} f_{x_1,\dots,y,x_k,\dots,x_{T-1}}(x_T)\, m_{x_{T-1}}(x_T) \dots m_{x_k}(x_{k+1}) \\
&\qquad - \sum_{\mathcal{X}^{T-k}} f_{x_1,\dots,z,x_k,\dots,x_{T-1}}(x_T)\, m_{x_{T-1}}(x_T) \dots m_{x_k}(x_{k+1}) \Big| \\
&\leq \|f_{x_1,\dots,x_{T-1}}\|_{\mathrm{Lip}}\, d(y,z) \sum_{\mathcal{X}^{T-k}} m_{x_{T-1}}(x_T) \dots m_{x_k}(x_{k+1}) \\
&= \|f_{x_1,\dots,x_{T-1}}\|_{\mathrm{Lip}}\, d(y,z),
\end{aligned}$$

from which follows the inequality

$$\sup_{x_1,\dots,x_{k-2},x_k \in \mathcal{X}} \|f_{x_1,\dots,x_{k-2},\cdot}(x_k)\|_{\mathrm{Lip}} \leq \|f_{x_1,\dots,x_{T-1}}\|_{\mathrm{Lip}}.$$

Now, let us show that the inequality is satisfied at $k-1$. Let $x_1, \dots, x_{k-2}, y, z \in \mathcal{X}$, using Proposition 4.1.1 in the second inequality below, we have

$$\begin{aligned}
|f_{x_1,\dots,x_{k-2}}(y) - f_{x_1,\dots,x_{k-2}}(z)| &\leq \Big| \sum_{x_k \in \mathcal{X}} f_{x_1,\dots,x_{k-2},y}(x_k)\, (m_y(x_k) - m_z(x_k)) \Big| \\
&\qquad + \sum_{x_k \in \mathcal{X}} \Big| f_{x_1,\dots,x_{k-2},y}(x_k) - f_{x_1,\dots,x_{k-2},z}(x_k) \Big| m_z(x_k) \\
&\leq (1-\kappa) \|f_{x_1,\dots,x_{k-2},y}\|_{\mathrm{Lip}}\, d(y,z) \\
&\qquad + \sum_{x_k \in \mathcal{X}} \|f_{x_1,\dots,x_{k-2},\cdot}(x_k)\|_{\mathrm{Lip}}\, d(y,z)\, m_z(x_k) \\
&\leq (s_k(1-\kappa) + 1)\, \|f_{x_1,\dots,x_{T-1}}\|_{\mathrm{Lip}}\, d(y,z) \\
&= s_{k-1} \|f_{x_1,\dots,x_{T-1}}\|_{\mathrm{Lip}}\, d(y,z)
\end{aligned}$$

where in the last inequality we used the induction hypothesis at the step k. Therefore, we obtain the inequality

$$\|f_{x_1,\dots,x_{k-2}}\|_{\mathrm{Lip}} \leq s_{k-1} \|f_{x_1,\dots,x_{T-1}}\|_{\mathrm{Lip}}.$$

The parameters $x_1, \dots, x_{k-2}$ being arbitrary, the inequality is established at step $k-1$, hence in full generality. Since $\|f_{x_1,\dots,x_{T-1}}\|_{\mathrm{Lip}} \leq \|f\|_{\mathrm{Lip}}/T$ and as for any $k \in \{1, \dots, T\}$ we have $s_k \leq 1/\kappa$, we get that for all $k \in \{2, \dots, T\}$,

$$\|f_{x_1,\dots,x_{k-1}}\|_{\mathrm{Lip}} \leq s_k \|f_{x_1,\dots,x_{T-1}}\|_{\mathrm{Lip}} \leq \frac{\|f\|_{\mathrm{Lip}}}{\kappa T},$$

and

$$\|f_\varnothing\|_{\mathrm{Lip}} \leq s_1 \|f_{x_1,\dots,x_{T-1}}\|_{\mathrm{Lip}} \leq \frac{\|f\|_{\mathrm{Lip}}}{\kappa T}.$$

□

Now, we are ready to introduce the following theorem, giving an upper bound of the variance term.

Theorem 6.2.1 (Variance of empirical means)**.** *We have*

$$\operatorname{Var}_x \hat{\pi}(f) \leq \begin{cases} \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa T} \sup_{x \in \mathcal{X}} \frac{\sigma^2(x)}{n_x \kappa}, & \text{if } T_0 = 0; \\ \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa T} \left(1 + \frac{1}{\kappa T}\right) \sup_{x \in \mathcal{X}} \frac{\sigma^2(x)}{n_x \kappa}, & \text{otherwise.} \end{cases}$$

Proof. To obtain that bound, we will use the inequalities given in Lemma 6.2.2 as well as the variance bound provided by Lemma 6.2.1 with $f_{x_1,\dots,x_{k-1}}$ and $N = 1$, successively for $k = T, T-1, \dots, 2$ as follows

$$\begin{aligned}
\mathbb{E}_x\left[\hat{\pi}(f)^2\right] &= \sum_{\mathcal{X}^T} f_{x_1,\dots,x_{T-1}}^2(x_T)\, m_{x_{T-1}}(x_T) \dots m_{x_1}(x_2)\, m_x^{T_0+1}(x_1) \\
&= \sum_{\mathcal{X}^{T-1}} \left(\sum_{\mathcal{X}} f_{x_1,\dots,x_{T-1}}^2(x_T)\, m_{x_{T-1}}(x_T) \right) m_{x_{T-2}}(x_{T-1}) \dots m_x^{T_0+1}(x_1) \\
&= \sum_{\mathcal{X}^{T-1}} \mathrm{M}\left(f_{x_1,\dots,x_{T-1}}^2\right)(x_{T-1})\, m_{x_{T-2}}(x_{T-1}) \dots m_x^{T_0+1}(x_1) \\
&\leq \sum_{\mathcal{X}^{T-1}} \left(\mathrm{M}\, f_{x_1,\dots,x_{T-1}}(x_{T-1})\right)^2 m_{x_{T-2}}(x_{T-1}) \dots m_x^{T_0+1}(x_1) \\
&\quad + \sum_{\mathcal{X}^{T-1}} \|f_{x_1,\dots,x_{T-1}}\|_{\mathrm{Lip}}^2 \left(\frac{\sigma^2}{n}\right)(x_{T-1})\, m_{x_{T-2}}(x_{T-1}) \dots m_x^{T_0+1}(x_1) \\
&\leq \sum_{\mathcal{X}^{T-1}} f_{x_1,\dots,x_{T-2}}^2(x_{T-1})\, m_{x_{T-2}}(x_{T-1}) \dots m_x^{T_0+1}(x_1) \\
&\quad + \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa^2 T^2} \mathrm{M}^{T_0+T-1}\left(\frac{\sigma^2}{n}\right)(x) \\
&\leq \sum_{\mathcal{X}^{T-2}} f_{x_1,\dots,x_{T-3}}^2(x_{T-2})\, m_{x_{T-3}}(x_{T-2}) \dots m_x^{T_0+1}(x_1) \\
&\quad + \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa^2 T^2} \left(\mathrm{M}^{T_0+T-2}\left(\frac{\sigma^2}{n}\right)(x) + \mathrm{M}^{T_0+T-1}\left(\frac{\sigma^2}{n}\right)(x)\right) \\
&\vdots \\
&\leq \sum_{\mathcal{X}} f_{\varnothing}^2(x_1)\, m_x^{T_0+1}(x_1) + \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa^2 T^2} \sum_{k=T_0+1}^{T_0+T-1} \mathrm{M}^k\left(\frac{\sigma^2}{n}\right)(x) \\
&= \mathrm{M}^{T_0+1}\left(f_{\varnothing}^2\right)(x) + \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa^2 T^2} \sum_{k=T_0+1}^{T_0+T-1} \mathrm{M}^k\left(\frac{\sigma^2}{n}\right)(x) \\
&\leq \left(\mathrm{M}^{T_0+1} f_{\varnothing}(x)\right)^2 + \|f_{\varnothing}\|_{\mathrm{Lip}}^2 \sum_{k=0}^{T_0} (1-\kappa)^{2(T_0-k)}\, \mathrm{M}^k\left(\frac{\sigma^2}{n}\right)(x) \\
&\quad + \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa^2 T^2} \sum_{k=T_0+1}^{T_0+T-1} \mathrm{M}^k\left(\frac{\sigma^2}{n}\right)(x) \\
&\leq \left(\sum_{\mathcal{X}} f_{\varnothing}(x_1)\, m_x^{T_0+1}(x_1)\right)^2 + \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa^2 T^2} \sum_{k=0}^{T_0} (1-\kappa)^{2(T_0-k)}\, \mathrm{M}^k\left(\frac{\sigma^2}{n}\right)(x)
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa^2 T^2} \sum_{k=T_0+1}^{T_0+T-1} \mathrm{M}^k \left(\frac{\sigma^2}{n}\right)(x) \\
&= \left(\sum_{\mathcal{X}^T} f_{x_1,\dots,x_{T-1}}(x_T)\, m_{x_{T-1}}(x_T) \dots m_{x_1}(x_2)\, m_x^{T_0+1}(x_1) \right)^2 \\
&+ \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa^2 T^2} \left(\sum_{k=0}^{T_0} (1-\kappa)^{2(T_0-k)}\, \mathrm{M}^k \left(\frac{\sigma^2}{n}\right)(x) + \sum_{k=T_0+1}^{T_0+T-1} \mathrm{M}^k \left(\frac{\sigma^2}{n}\right)(x) \right) \\
&= \mathbb{E}_x\left[\hat{\pi}(f)\right]^2 + \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa^2 T^2} \left(\sum_{k=0}^{T_0} (1-\kappa)^{2(T_0-k)}\, \mathrm{M}^k \left(\frac{\sigma^2}{n}\right)(x) + \sum_{k=T_0+1}^{T_0+T-1} \mathrm{M}^k \left(\frac{\sigma^2}{n}\right)(x) \right),
\end{aligned}
$$

where in the end, we use the variance bound of Lemma 6.2.1 with $f_\varnothing$ and $N = T_0+1$. Hence, we finally get

$$
\begin{aligned}
\mathrm{Var}_x\, \hat{\pi}(f) &\leq \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa^2 T^2} \left(\sum_{k=0}^{T_0} (1-\kappa)^{2(T_0-k)}\, \mathrm{M}^k \left(\frac{\sigma^2}{n}\right)(x) + \sum_{k=T_0+1}^{T_0+T-1} \mathrm{M}^k \left(\frac{\sigma^2}{n}\right)(x) \right) \\
&\leq \begin{cases} \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa T} \sup_{x\in\mathcal{X}} \frac{\sigma^2(x)}{n_x \kappa}, & \text{if } T_0 = 0; \\ \frac{\|f\|_{\mathrm{Lip}}^2}{\kappa T} \left(1 + \frac{1}{\kappa T}\right) \sup_{x\in\mathcal{X}} \frac{\sigma^2(x)}{n_x \kappa}, & \text{otherwise.} \end{cases}
\end{aligned}
$$

and we are done. □

6.3 Examples

In order to illustrate the previous results, we will give two examples.

Example 6.3.1 (n-dimensional hypercube)**.** *Let $\mathcal{X} = \{0,1\}^N$. Again, we will use the Hamming metric on $\mathcal{X}$. The Markov chain we shall consider is the same lazy random walk m as in Example 3.3.1. We have shown previously that $\kappa = \frac{1}{N}$. Clearly, for any $x \in \mathcal{X}$, we have $E(x) = N/2$. Our bias estimate for a Lipschitz function f is thus*

$$
\left|\mathbb{E}_x\left[\hat{\pi}(f)\right] - \pi(f)\right| \leq \frac{N^2 (1-1/N)^{T_0+1}}{2T} \|f\|_{\mathrm{Lip}} \leq \frac{N^2}{2T} e^{-T_0/N} \|f\|_{\mathrm{Lip}}.
$$

So taking $T_0 \approx 2N \ln N$ is enough to ensure small bias. Regarding the variance estimate, we need to find an upper bound for $\frac{\sigma^2(x)}{n_x}$. This fraction corresponds to the maximal variance under m_x of a 1-Lipschitz function. Let $f : \mathcal{X} \to \mathbb{R}$ be a 1-Lipschitz function. As the variance does not depend on constants, we can assume $f(x) = 0$. For all neighbors y of x, we have $|f(y)| \leq 1$. As a result, we get

$$
\mathrm{Var}_{m_x}(f) = \mathbb{E}_{m_x}\left[f^2\right] - \mathbb{E}_{m_x}\left[f\right]^2 \leq \mathbb{E}_{m_x}\left[f^2\right] \leq 1 - m_x(x) = \frac{1}{2}.
$$

This upper bound can be achieved. In fact, if x has an even number of neighbors with half of them taking the value 1 under f and the other half taking the value -1, we have $\mathbb{E}_{m_x}[f] = 0$ with moreover $\mathbb{E}_{m_x}[f^2] = k \cdot \frac{1}{2k} \cdot 1^2 = \frac{1}{2}$. We finally get the following variance estimate

$$
\mathrm{Var}_x\, \hat{\pi}(f) \leq \begin{cases} \frac{N^2}{2T} \|f\|_{\mathrm{Lip}}^2, & \textit{if } T_0 = 0; \\ \frac{N^2}{2T} \left(1 + N/T\right) \|f\|_{\mathrm{Lip}}^2, & \textit{otherwise.} \end{cases}
$$

As such, with the function f equal to the proportion of "1" bits in the sequence (this function is $1/N$-Lipschitz), taking $T \approx N$ will yield a variance around $1/N$.

Example 6.3.2 (discrete Ornstein Uhlenbeck process)**.** *We work in the same setting as Example 3.3.2. We have $\kappa = 1/2N$ and for any $x \in \{-N, \dots, N-1, N\}$, $E(x) \leq N$. Our bias estimate for a Lipschitz function f is thus*

$$|\mathbb{E}_x[\hat{\pi}(f)] - \pi(f)| \leq \frac{2N^2(1-1/2N)^{T_0+1}}{T}\|f\|_{\text{Lip}} \leq \frac{2N^2}{T}e^{-T_0/2N}\|f\|_{\text{Lip}}.$$

We should this time take $T_0 \approx 4N \ln 2N$ to get a similarly small bias as in the previous example (6.3.1). As regards the variance estimate, we choose a 1-Lipschitz function f verifying $f(x) = 0$. We have

$$\text{Var}_{m_x}(f) = \mathbb{E}_{m_x}\left[f^2\right] - \mathbb{E}_{m_x}[f]^2 \leq \mathbb{E}_{m_x}\left[f^2\right] \leq 1 - m_x(x) = \frac{1}{2},$$

and this upper bound can be achieved at $x = 0$. We thus get the following variance estimate

$$\text{Var}_x\,\hat{\pi}(f) \leq \begin{cases} \frac{2N^2}{T}\|f\|_{\text{Lip}}^2, & \text{if } T_0 = 0; \\ \frac{2N^2}{T}\left(1 + 2N/T\right)\|f\|_{\text{Lip}}^2, & \text{otherwise.} \end{cases}$$

Similarly to the previous example, choosing a $1/N$-Lipschitz function with $T \approx 2N$ will yield a variance around $1/N$.

7 Conclusion

In this report, we firstly went through a short introduction to discrete-time Markov chains and saw some of their main properties (e.g. irreducibility, reversibility). Moreover, we provided a proof of convergence to the invariant distribution.

We then stressed the importance of Ollivier's Ricci curvature when studying the convergence speed of Markov chains. We studied couplings, especially those witnessing for the Wasserstein distance W_1, i.e. the optimal ones. Interestingly, we saw with Example 3.3.2 that choosing the right optimal coupling is not negligible if we want to make two Markov chains meet as fast as possible. We proved that Ollivier's Ricci curvature dictates the exponential convergence speed of the Wasserstein distance contraction as well as the variance contraction. We were finally able to derive a Poincaré inequality.

In the 5th section, we focused on reversible Markov chains and in particular their correspondence with random walks on undirected graphs. We managed to derive sharp lower bounds for Ollivier's Ricci curvature, firstly for unweighted graphs before generalizing our results to weighted graphs. Moreover, we explained how to know if the bounds given by the two Theorems 5.1.1 and 5.2.1 are close to the actual Ricci curvature or not.

Finally, in the last, more applied, part of this report, we focused on a Markov chain Monte Carlo method to approximate $\pi(f)$ with $\hat{\pi}(f)$. We studied the mean quadratic error and explained with two examples how to correctly choose the burn-in T_0 as well we the trajectory time T to get a good approximation.

Acknowledgements

My first thanks are for my Bachelor Thesis supervisor, Professor Giovanni Conforti. Thank you for your guidance, for the many conversations we had and for the crucial support you gave me throughout the project. These two months have been particularly exciting for me and have given me a better understanding of what I like about mathematics. So thank you for helping me find my way to what I like. I would also like to thank Polytechnique and in particular the CMAP department for hosting me during my Thesis. Last but not least, thank you to my family for their continued support.

References

[1] J. Jost and S. Liu. Ollivier's ricci curvature, local clustering and curvature-dimension inequalities on graphs. *Discrete & Computational Geometry*, 51:300–322, 2014.

[2] Aldéric Joulin. A new Poisson-type deviation inequality for Markov jump processes with positive Wasserstein curvature. *Bernoulli*, 15(2):532 – 549, 2009.

[3] Aldéric Joulin and Yann Ollivier. Curvature, concentration and error estimates for markov chain monte carlo. *The Annals of Probability*, 38(6), Nov 2010.

[4] J. R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

[5] Yann Ollivier. Ricci curvature of markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

[6] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11 (5-6):355–602, 2019.

[7] C. Villani. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003.

Statement of Academic Integrity Regarding Plagiarism

I, the undersigned Luca Bonengel, hereby certify on my honor that:

1. The results presented in this report are the product of my own work.
2. I am the original creator of this report.
3. I have not used sources or results from third parties without clearly stating thus and referencing them according to the recommended rules for providing bibliographic information.

I hereby declare that this work contains no plagiarized material.

October 1, 2021.