Mixing times of Markov chains

Justin Salez

Abstract

How many times should one shuffle a deck of 52 cards? This course is a self-contained introduction to the modern theory of mixing for Markov chains. It consists of a guided tour through the various methods for estimating mixing times, including couplings, spectral analysis, discrete geometry, and functional inequalities. Each of those tools is illustrated on a variety of examples from different contexts: interacting particle systems, card shufflings, random walks on graphs and networks, etc. A particular attention is devoted to the cutoff phenomenon, a remarkable but still mysterious phase transition in the convergence to equilibrium of certain chains.

Contents

1 Framework

This first chapter sets the stage on which we are going to perform. We start with a brief but self-contained remainder on finite Markov chains and their large-time behavior. We then introduce the total-variation distance to equilibrium, collect some of its basic properties, and use them to define three fundamental notions that will lie at the center of our attention: the relaxation time, the mixing time, and the cutoff phenomenon. Finally, we briefly present two rich classes of Markov chains which are important from both a theoretical and a practical viewpoint: random walks on graphs, and random walks on groups.

1.1 Markov chains

A Markov chain is specified by a triple $(\mathscr{X}, P, \nu)$ consisting of the following ingredients:

(i) A state space $\mathscr{X}$, which in our case will just be a finite, non-empty set.

(ii) A transition kernel $P\colon \mathscr{X} \times \mathscr{X} \to [0,1]$, which satisfies

$$\forall x \in \mathscr{X}, \quad \sum_{y \in \mathscr{X}} P(x,y) \;=\; 1. \tag{1}$$

(iii) An initial law $\nu\colon \mathscr{X} \to [0,1]$, which satisfies

$$\sum_{x \in \mathscr{X}} \nu(x) \;=\; 1. \tag{2}$$

Definition 1 (Markov chain). *A Markov chain with parameters $(\mathscr{X}, P, \nu)$ is a sequence $\mathbf{X} = (X_0, X_1, \ldots)$ of $\mathscr{X}-$valued random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ so that*

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t) \;=\; \nu(x_0)P(x_0,x_1)\cdots P(x_{t-1},x_t), \tag{3}$$

for every $t \in \mathbb{N} = \{0,1,2,\ldots\}$ and every $(x_0,\ldots,x_t) \in \mathscr{X}^{t+1}$. We write $\mathbf{X} \sim \mathrm{MC}(\mathscr{X}, P, \nu)$.

Remark 1 (Markov property). *The product form (3) asserts that the past $(X_0,\ldots,X_{t-1})$ and the future $(X_{t+1}, X_{t+2}, \ldots)$ are conditionally independent, given the present X_t.*

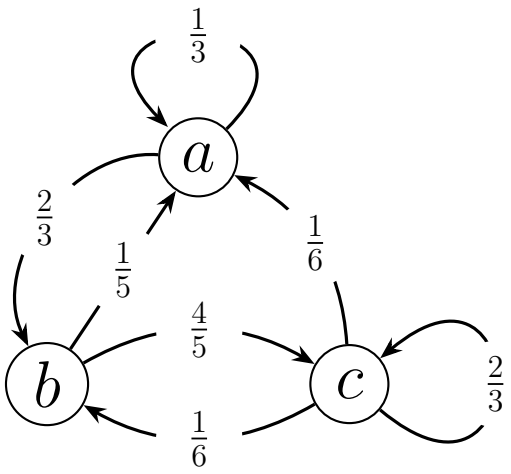

Figure 1: A particular transition kernel on $\mathscr{X} = \{a, b, c\}$, represented by its diagram.

Remark 2 (Existence and uniqueness)**.** *By Dynkin's theorem, the finite-dimensional marginals (3) determine the law of* $\mathbf{X}$ *as a stochastic process, i.e. a random variable taking values in the product space* $(\mathscr{X}, \mathscr{P}(\mathscr{X}))^{\otimes \mathbb{N}}$*. Conversely, Kolmogorov extension's theorem (or a direct construction) always ensures the existence of* $\mathbf{X} \sim \mathrm{MC}(\mathscr{X}, P, \nu)$*.*

Let $\mathbf{X} \sim \mathrm{MC}(\mathscr{X}, P, \nu)$, and let μ_t denote the law of the random variable X_t. Then it follows from the above definition that $\mu_0 = \nu$ and that for every $t \geq 1$,

$$\forall y \in \mathscr{X}, \quad \mu_t(y) \;=\; \sum_{x \in \mathscr{X}} \mu_{t-1}(x) P(x, y). \tag{4}$$

It will be convenient to view a probability distribution $\mu \in \mathcal{P}(\mathscr{X})$ as a row vector, a function $f \colon \mathscr{X} \to \mathbb{R}$ as a column vector, and a transition kernel $P \colon \mathscr{X} \times \mathscr{X} \to [0, 1]$ as a matrix. In particular, we may rewrite the above identity in the following compact way:

$$\mu_t \;=\; \mu_{t-1} P. \tag{5}$$

We shall be interested in the large-time distribution of our Markov chain, i.e., the asymptotic behavior of μ_t as $t \to \infty$. If the sequence $(\mu_t)_{t \geq 0}$ converges to some distribution $\pi \in \mathcal{P}(\mathscr{X})$, then passing to the limit in the above recursion shows that π must be stationary, i.e.

$$\pi \;=\; \pi P. \tag{6}$$

We are thus naturally led to investigate the question of existence and uniqueness of stationary distributions. In that respect, the following definition will be useful.

Definition 2 (Irreducibility)**.** *The transition kernel P is said to be irreducible if*

$$\forall (x,y) \in \mathscr{X}^2, \quad \exists t \in \mathbb{N}, \quad P^t(x,y) \;>\; 0. \tag{7}$$

Remark 3 (Graph-theoretical interpretation)**.** *The diagram of the chain is the directed graph G_P on $\mathscr{X}$ obtained by placing an edge $x \to y$ whenever $P(x,y) > 0$ (see Figure 1). The irreducibility of P simply means that G_P is strongly connected, in the sense that it contains a path from every vertex to every other. When this is not the case, we can always restrict P to a strongly connected component to obtain an irreducible kernel.*

Lemma 1 (Stationary laws)**.** *Any transition kernel P has a stationary law π. If P is irreducible, then the stationary law is unique and has full support.*

Proof. Consider the empirical mean of the first $t \geq 1$ instantaneous laws of the chain:

$$\pi_t \;:=\; \frac{1}{t}\sum_{s=0}^{t-1} \mu_s \;\in\; \mathcal{P}(\mathscr{X}). \tag{8}$$

By compactness, we can extract from the sequence $(\pi_t)_{t\geq 1}$ a subsequence that admits a limit $\pi \in \mathcal{P}(\mathscr{X})$. The latter is necessarily stationary, because for each $x \in \mathscr{X}$

$$(\pi_t P)(x) - \pi_t(x) \;=\; \frac{\mu_t(x) - \mu_0(x)}{t} \;\xrightarrow[t\to\infty]{}\; 0.$$

This shows existence. Note that the stationary equation $\pi P = \pi$ implies that the support of π is closed under the accessibility relation $x \to y$ defining the diagram G_P. In particular, if P is irreducible, then any stationary law π has full support. Finally, suppose for a contradiction that an irreducible kernel P admits two distinct stationary distributions π_1 and π_2, and consider the function $m\colon [0,1] \to [-1,1]$ defined as follows:

$$m(\theta) \;:=\; \min_{x\in\mathscr{X}} \left(\pi_1(x) - \theta\pi_2(x)\right).$$

It is clear that m is continuous, with $m(0) > 0$ (because π_1 has full support) and $m(1) < 0$ (because $\pi_1 \neq \pi_2$). Thus, there is $\theta_\star \in (0,1)$ such that $m(\theta_\star) = 0$. But then, the vector

$$\pi \colon x \quad \mapsto \quad \frac{\pi_1(x) - \theta_\star \pi_2(x)}{1 - \theta_\star}, \tag{9}$$

satisfies $\pi P = \pi$ (because so do π_1, π_2), has minimal entry equal to zero (by definition of $\theta_\star$) and has entry sum equal to 1 (because so do π_1, π_2). Thus, π is a non-fully supported stationary distribution for the irreducible kernel P, a contradiction. □

Remark 4 (Optimality)**.** *The converse is easily shown to be true: the irreducibility of P is necessary and sufficient for the invariant law π to be unique and fully supported.*

Remark 5 (Césarro mixing)**.** *By a standard compactness-uniqueness argument, the above proof shows that any irreducible Markov chain $\mathbf{X}$ satisfies the convergence*

$$\forall y \in \mathscr{X}, \quad \frac{1}{t} \sum_{s=0}^{t-1} \mathbb{P}(X_s = y) \quad \xrightarrow[t \to \infty]{} \quad \pi(y), \tag{10}$$

regardless of the chosen initial condition.

Let us note that the more natural conclusion $\mathbb{P}(X_t = y) \to \pi(y)$, which is stronger than (10), requires extra assumptions, as the following counter-example shows.

Example 1 (Periodic chain)**.** *On $\mathscr{X} = \{0,1\}$, consider the transition kernel*

$$P \quad = \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

which is irreducible with stationary law $\pi = \mathrm{Unif}(\mathscr{X})$. Since $P^2 = \mathbb{I}$ (the identity matrix on $\mathscr{X}$), we have $P^{2t} = \mathbb{I}$ and $P^{2t+1} = P$ for all $t \in \mathbb{N}$. In particular, if we start from the initial condition $\nu = \delta_0$, then the sequence $(\mu_t)_{t \geq 0}$ keeps alternating between the two Dirac masses δ_0 and δ_1, and mixing only occurs in the Césarro-mean sense (10).

In order to preclude the type of pathological periodicity displayed by Example 1, we need to strengthen the irreducibility assumption (7) by exchanging the order of quantifiers.

Definition 3 (Ergodicity). *The transition kernel P is called ergodic if*

$$\exists t \in \mathbb{N}, \quad \forall (x,y) \in \mathscr{X}^2, \quad P^t(x,y) \;>\; 0. \tag{11}$$

Remark 6 (Lazyness). *Ergodicity is strictly stronger than irreducibility. A simple way to make an irreducible kernel P ergodic consists in replacing it with its lazy version:*

$$\widehat{P} \;:=\; \frac{\mathbb{I}+P}{2}. \tag{12}$$

In other words, a fair coin is tossed at each step: if heads comes up, a transition is made according to the original kernel P, otherwise nothing happens. The resulting chain $\widehat{\mathbf{X}}$ is just a time-changed version of the original chain $\mathbf{X}$, and it retains most of its essential features, including the stationary distribution π. Obviously, any transition kernel whose diagonal entries are at least $1/2$ is the lazy version of some transition kernel.

The following fundamental result constitutes the starting point of our study. It asserts that any ergodic Markov chain mixes: as the number of iterations grows, the chain approaches the stationary distribution, regardless of the initial condition.

Theorem 1 (Convergence to equilibrium). *If P is ergodic, then we have*

$$\forall y \in \mathscr{X}, \quad \mathbb{P}(X_t = y) \;\xrightarrow[t\to\infty]{}\; \pi(y), \tag{13}$$

regardless of the initial condition. Equivalently, $P^t(x,y) \to \pi(y)$ for all $x, y \in \mathscr{X}$.

We will later see many proofs of this result, and numerous refinements. Here is a nice practical application: in order to approximately sample from a sophisticated target distribution π, all we have to do is to design an ergodic Markov chain whose stationary law is π, and let it run for *sufficiently long*. This observation is at the basis of a number of technological revolutions, such as Monte Carlo Markov Chain methods or Google's Pagerank algorithm. It motivates the following question, to which the present course is entirely devoted.

Question 1 (Speed of convergence). *How fast is the convergence to equilibrium (13) ?*

To answer this question, we first need to agree on a way to measure the distance to equilibrium, i.e., the discrepancy between the law μ_t at time t, and the stationary law π.

1.2 Distance to equilibrium

There are many natural ways to measure the distance between two probability measures μ and ν on a finite set $\mathscr{X}$: Hellinger distance for statisticians, Hilbert/L^p norms for analysts, relative entropy for physicists and information theorists, etc. Each has its own flavor and its specific list of advantages and drawbacks, making it more relevant to the study of certain chains than others. Rather than competing with each other, these distances are complementary: they are related one to another by an array of inequalities, and combining different viewpoints is often the best way to analyze a given Markov chain. We will here mainly focus on the total-variation distance, which is more natural for probabilists.

Definition 4 (Total variation)**.** *The total variation distance between μ and ν is*

$$d_{\mathrm{TV}}(\mu,\nu) \;:=\; \max_{A\subseteq\mathscr{X}} |\mu(A)-\nu(A)|\,.$$

This is clearly a distance on the set $\mathcal{P}(\mathscr{X})$ of all probability measures on $\mathscr{X}$, and we have $d_{\mathrm{TV}}(\mu,\nu)\le 1$, with equality if and only if μ and ν have disjoint supports. Let us start by collecting a list of useful alternative expressions.

Lemma 2 (Alternative expressions)**.** *For any $\mu,\nu\in\mathcal{P}(\mathscr{X})$, we also have*

$$\begin{aligned} d_{\mathrm{TV}}(\mu,\nu) &= \max_{A\subseteq\mathscr{X}} (\mu(A)-\nu(A)) \\ &= \sum_{x\in\mathscr{X}} (\mu(x)-\nu(x))_+ \\ &= 1-\sum_{x\in\mathscr{X}} \mu(x)\wedge\nu(x) \\ &= \frac{1}{2}\sum_{x\in\mathscr{X}} |\mu(x)-\nu(x)| \\ &= \frac{1}{2}\sup_{f\colon \mathscr{X}\to[-1,1]} |\mu f-\nu f|\,. \end{aligned}$$

Proof. The first identity follows from the observation that changing the set A to its comple-

ment changes the value $\mu(A) - \nu(A)$ to its opposite. For the second, note that

$$\begin{aligned} \mu(A) - \nu(A) &= \sum_{x \in A} (\mu(x) - \nu(x)) \\ &\leq \sum_{x \in A} (\mu(x) - \nu(x))_+ \\ &\leq \sum_{x \in \mathscr{X}} (\mu(x) - \nu(x))_+ , \end{aligned}$$

for all $A \subseteq \mathscr{X}$, and that those become equalities for $A = \{x \in \mathscr{X} : \mu(x) \geq \nu(x)\}$. The third and fourth claims are obtained by taking $a = \mu(x)$ and $b = \nu(x)$ in the two identities

$$\begin{aligned} (a-b)_+ &= a - a \wedge b \\ &= \frac{1}{2} (|a-b| + a - b) . \end{aligned}$$

Finally, for the last inequality, simply note that for any $f \colon \mathscr{X} \to [-1, 1]$,

$$\begin{aligned} |\mu f - \nu f| &= \sum_{x \in \mathscr{X}} (\mu(x) - \nu(x)) f(x) \\ &\leq \sum_{x \in \mathscr{X}} |\mu(x) - \nu(x)| \\ &= 2 d_{\mathrm{TV}}(\mu, \nu), \end{aligned}$$

and that there is equality in the case $f(x) = \mathrm{sign}(\mu(x) - \nu(x))$, with $\mathrm{sign} = \mathbf{1}_{\mathbb{R}_+} - \mathbf{1}_{\mathbb{R}_-}$. □

Let us now record a couple of basic but important properties of total variation distance.

Lemma 3 (Convexity and contraction)**.**

1. *The function $(\mu, \nu) \mapsto d_{\mathrm{TV}}(\mu, \nu)$ is convex on $\mathcal{P}(\mathscr{X}) \times \mathcal{P}(\mathscr{X})$.*
2. *Any transition kernel P on $\mathscr{X}$ contracts d_{TV}:*

$$\forall \mu, \nu \in \mathcal{P}(\mathscr{X}), \quad d_{\mathrm{TV}}(\mu P, \nu P) \leq d_{\mathrm{TV}}(\mu, \nu).$$

Proof. The first claim follows from the observation that $(\mu, \nu) \mapsto \mu(A) - \nu(A)$ is trivially convex for each $A \subseteq \mathscr{X}$, and that any maximum of convex functions is convex. The second follows from the last expression in Lemma 2, because if $f \in [-1, 1]^{\mathscr{X}}$, then so does Pf. □

We are now ready to introduce the main object of our study.

Definition 5 (Distance to equilibrium)**.** *The distance to equilibrium associated with an irreducible transition kernel P on $\mathscr{X}$ is the function $\mathscr{D}_P \colon \mathbb{N} \to [0,1]$ defined by*

$$\mathscr{D}_P(t) \;:=\; \max_{\nu \in \mathcal{P}(\mathscr{X})} d_{\mathrm{TV}}(\nu P^t, \pi), \tag{14}$$

where π denotes the unique stationary distribution under P.

Remark 7 (Properties)**.** *Let us make four important comments about the function $\mathscr{D}_P$.*

1. *The first item in Lemma 3 ensures that the maximum over all distributions $\nu \in \mathcal{P}(\mathscr{X})$ in (14) can be reduced to a maximum over all* extremal *distributions $(\delta_x)_{x \in \mathscr{X}}$:*

$$\mathscr{D}_P(t) \;=\; \max_{x \in \mathscr{X}} d_{\mathrm{TV}}(P^t(x,\cdot), \pi).$$

2. *The second item in Lemma 3 ensures that the function $\mathscr{D}_P$ is non-increasing.*

3. *The initial distance $\mathscr{D}_P(0)$ is close to 1 if the state space $\mathscr{X}$ is large. Indeed,*

$$\mathscr{D}_P(0) \;=\; 1 - \pi_\star \quad \text{where} \quad \pi_\star \;:=\; \min_{x \in \mathscr{X}} \pi(x) \;\leq\; \frac{1}{|\mathscr{X}|}.$$

4. *Theorem 1 asserts that $\mathscr{D}_P(t) \to 0$ as $t \to \infty$ whenever P is ergodic.*

1.3 Relaxation time and mixing time

We initiate our study of $\mathscr{D}_P$ by establishing a fundamental sub-multiplicativity property, which shows that mixing can not slow down once it has started: if s iterations suffice to reduce the distance to equilibrium to $1/4$, then ks iterations suffice to reduce it to 2^{-k}.

Lemma 4 (Sub-multiplicativity)**.** *We have $\mathscr{D}_P(t+s) \leq 2\mathscr{D}_P(t)\mathscr{D}_P(s)$ for all $s, t \in \mathbb{N}$.*

Proof. Let Π denote the "idealized" transition kernel which mixes exactly in one step, i.e.

$$\Pi(x,y) \;:=\; \pi(y). \tag{15}$$

Then, it is immediate to check that $\Pi^2 = P\Pi = \Pi P = \Pi$, so that for all $t \in \mathbb{N}$,

$$(P - \Pi)^t \;=\; \sum_{s=0}^{t} \binom{t}{s} (-1)^s \Pi^s P^{t-s} \;=\; P^t - \Pi.$$

Thus, writing $\|A\| := \max_{x\in\mathscr{X}} \sum_{y\in\mathscr{X}} |A(x,y)|$, we arrive at the important representation

$$2\mathscr{D}_P(t) \;=\; \left\|(P-\Pi)^t\right\|, \tag{16}$$

for all $t \geq 1$. The desired claim now follows from the sub-multiplicativity property

$$\|AB\| \;\leq\; \|A\|\|B\|, \tag{17}$$

which is easily checked to hold for all matrices $A, B \in \mathbb{R}^{\mathscr{X}\times\mathscr{X}}$. □

Lemma 4 asserts that the non-negative sequence $(u_t)_{t\in\mathbb{N}}$ defined by $u_t := 2\mathscr{D}(t)$ is sub-multiplicative, i.e. $u_{t+s} \leq u_t u_s$ for all $t, s \in \mathbb{N}$. By Fekete's Lemma, this implies

$$u_t^{1/t} \;\xrightarrow[t\to\infty]{}\; \inf\left\{u_s^{1/s} \colon s \geq 1\right\}. \tag{18}$$

Note that the infimum is less than 1 if P is ergodic, because we then have $\mathscr{D}_P(t) < 1/2$ for t large enough. This establishes the following notable refinement of Theorem 1.

Corollary 1 (Geometric decay)**.** *If P is ergodic, then there is $\lambda_\star(P) < 1$ such that*

$$(\mathscr{D}_P(t))^{\frac{1}{t}} \;\xrightarrow[t\to\infty]{}\; \lambda_\star(P).$$

Remark 8 (Spectral radius)**.** *In Chapter 3, we will show that the fundamental quantity*

$$\lambda_\star(P) \;=\; \inf\left\{(2\mathscr{D}_P(t))^{\frac{1}{t}} : t \geq 1\right\} \tag{19}$$

admits a simple and beautiful characterization in terms of the eigenvalues of P. For this reason, $\lambda_\star(P)$ is often called the spectral radius of the chain.

At first sight, Corollary 1 seems to bring a definitive answer to Question 1: the distance to equilibrium $\mathscr{D}_P(t)$ decays essentially like $(\lambda_\star(P))^t$ as $t \to \infty$. It is tempting to deduce that the chain is well mixed when the number of iterations is larger than the relaxation time

$$\mathrm{t}_{\mathrm{REL}}(P) \;:=\; \frac{1}{\log\left(\frac{1}{\lambda_\star(P)}\right)}, \tag{20}$$

defined so that $\lambda_\star(P) = \exp(-\frac{1}{\mathrm{t}_{\mathrm{REL}}(P)})$ (our definition differs slightly from the classical one, which we find less natural in discrete time). In fact, this intuition is wrong, and one often

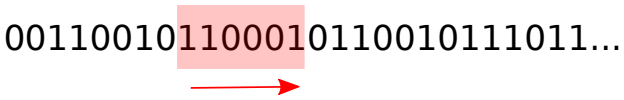

Figure 2: The "sliding window" chain with $n = 6$.

needs to wait much longer than the relaxation time before the chain even starts to mix. The reason is that the approximation

$$\mathscr{D}_P(t) \;\approx\; (\lambda_\star(P))^t \;=\; \exp\left(-\frac{t}{\mathrm{t}_{\mathrm{REL}}(P)}\right) \tag{21}$$

promised by Corollary 1 is only valid asymptotically, in a regime where t is so large that the chain is already infinitesimally close to equilibrium. Thus, the relaxation time does not answer the real question: how long do we need to wait *before* the distance to equilibrium becomes small? This naturally leads to the following definition, illustrated on Figure 3.

Definition 6 (Mixing time)**.** *The mixing time of P with precision $\varepsilon \in (0,1)$ is*

$$\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P) \;:=\; \min\{t \in \mathbb{N} \colon \mathscr{D}_P(t) \le \varepsilon\}.$$

The default precision is $\varepsilon = 1/4$, in which case we write $\mathrm{t}_{\mathrm{MIX}}(P)$ instead of $\mathrm{t}_{\mathrm{MIX}}^{(1/4)}(P)$.

The value $\varepsilon = 1/4$ is standard, and sufficient in practice: any smaller precision can be achieved by just increasing $\mathrm{t}_{\mathrm{MIX}}(P)$ by a universal factor. Indeed, Lemma 18 implies

$$\forall \varepsilon \in (0,1/4), \qquad \mathrm{t}_{\mathrm{MIX}}(P) \;\le\; \mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P) \;\le\; \mathrm{t}_{\mathrm{MIX}}(P) \left\lceil \log_2 \frac{1}{\varepsilon} \right\rceil. \tag{22}$$

Thus, the fundamental quantity $\mathrm{t}_{\mathrm{MIX}}(P)$ provides a rigorous formalization of Question 1. Understanding its dependency on the underlying kernel P is a fascinating task, to which the present course is devoted. By (19), the relaxation time always provides the lower-bound

$$\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P) \;\ge\; \mathrm{t}_{\mathrm{REL}}(P) \log\left(\frac{1}{2\varepsilon}\right), \tag{23}$$

but the latter can be terribly poor, as we shall see. Let us illustrate these concepts on a toy example which is not exciting, but is simple enough to allow for explicit computations.

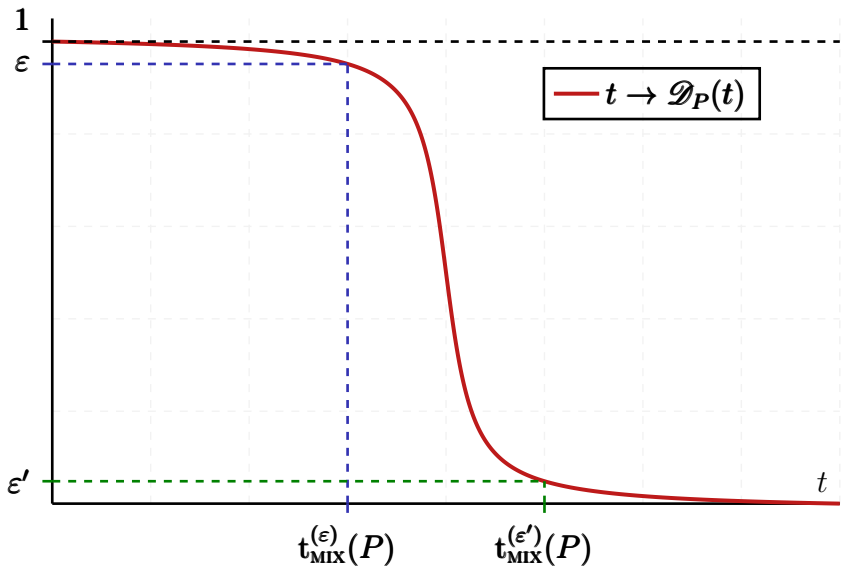

Figure 3: Distance to equilibrium and mixing times.

Example 2 (Sliding window). *On $\mathscr{X} = \{0,1\}^n$, consider the transition kernel*

$$P(x,y) \;=\; \begin{cases} \frac{1}{2} & \text{if } (y_1,\dots,y_{n-1}) = (x_2,\dots,x_n) \\ 0 & \text{else,} \end{cases}$$

which describes the content of a window of length n "sliding" along an infinite sequence of independent fair coin tosses (see Figure 2). Clearly, P is ergodic with $\pi = \mathrm{Unif}(\mathscr{X})$. For any $x = (x_1,\dots,x_n) \in \mathscr{X}$ and $t \le n$, we have $P^t(x,\cdot) = \mathrm{Unif}(A_{x,t})$, where

$$A_{x,t} \;:=\; \{y \in \{0,1\}^n \colon (y_1, y_2, \cdots, y_{n-t}) = (x_{t+1},\dots,x_n)\}\,.$$

Since $|A_{x,t}| = 2^t$, one finds $\mathscr{D}_P(t) = 1 - \left(\frac{1}{2}\right)^{n-t}$. In other words, for all $\varepsilon \in (0,1)$,

$$t_{\mathrm{MIX}}^{(\varepsilon)}(P) \;=\; \left\lceil n - \log_2\left(\frac{1}{1-\varepsilon}\right)\right\rceil.$$

In particular, $\mathscr{D}_P(t) = 0$ for all $t \ge n$. This forces $\lambda_\star(P) = 0$, hence $\mathrm{t}_{\mathrm{REL}}(P) = 0$: the relaxation time here drastically underestimates the time it takes for the chain to mix.

1.4 The cutoff phenomenon

From a practical point-of-view, estimating mixing times is particularly meaningful when the number of states becomes large. Rather than studying a fixed Markov chain, we will thus consider a *sequence* of transition kernels $(P_n)_{n\geq 1}$ whose dimensions grow with n, and investigate the order of magnitude of $\mathrm{t}_{\text{MIX}}(P_n)$ as $n \to \infty$. Recall that the particular choice $\varepsilon = 1/4$ in the definition of t_{MIX} is irrelevant: for any fixed $\varepsilon \in (0, 1/4)$,

$$\mathrm{t}^{(\varepsilon)}_{\text{MIX}}(P_n) \quad = \quad \Theta\left(\mathrm{t}_{\text{MIX}}(P_n)\right), \tag{24}$$

where the notation $a_n = \Theta(b_n)$ means that the sequence $(a_n/b_n)_{n\geq 1}$ is bounded from above and below by strictly positive constants. For many chains, we will see that an even stronger insensitivity holds: the dependency in ε disappears completely, in the following sense.

Definition 7 (Cutoff)**.** *The sequence $(P_n)_{n\geq 1}$ exhibits a cutoff phenomenon if*

$$\forall \varepsilon \in (0,1), \qquad \mathrm{t}^{(\varepsilon)}_{\text{MIX}}(P_n) \quad \sim \quad \mathrm{t}_{\text{MIX}}(P_n), \tag{25}$$

where the notation $a_n \sim b_n$ means that $a_n/b_n \to 1$ as $n \to \infty$.

In words, the number of iterations required to even slightly mix (say, $\varepsilon = 0.99$) is asymptotically the same as that needed to completely mix (say, $\varepsilon = 0.01$), at least to first order. Equivalently, the associated distance to equilibrium $\mathscr{D}_{P_n}$ undergoes a sharp phase transition, dropping abruptly from 1 to 0 around some appropriate time-scale $(t_n)_{n\geq 1}$, i.e.

$$\forall \alpha \in [0, \infty), \qquad \mathscr{D}_{P_n}\left(\lfloor \alpha t_n \rfloor\right) \quad \xrightarrow[n\to\infty]{} \quad \begin{cases} 1 & \text{if } \alpha < 1 \\ 0 & \text{if } \alpha > 1. \end{cases} \tag{26}$$

Note that this means that $\mathrm{t}^{(\varepsilon)}_{\text{MIX}}(P_n) \sim t_n$ for all $\varepsilon \in (0,1)$, hence the equivalence with Definition 7. For example, our computations for the sliding window of length n give

$$\mathrm{t}^{(\varepsilon)}_{\text{MIX}}(P_n) \quad \sim \quad n, \tag{27}$$

for any fixed $\varepsilon \in (0,1)$, providing our first instance of cutoff. Discovered in the 80's in the context of card shuffling, this remarkable phase transition has since then been established for a broad variety of chains, from random walks on certain groups to various interacting particle systems. However, the proofs of cutoff remain chain-specific, and finding a general explanation constitutes the most important open problem in the field.

Question 2. *What is the exact mechanism underlying the cutoff phenomenon?*

The following simple criterion was proposed as a possible answer in 2004.

Definition 8 (Product condition)**.** $(P_n)_{n\geq 1}$ *satisfies the product condition if*

$$\mathrm{t}_{\mathrm{MIX}}(P_n) \quad \gg \quad \mathrm{t}_{\mathrm{REL}}(P_n),$$

where the notation $a_n \gg b_n$ means that $b_n/a_n \to 0$ as $n \to \infty$.

This condition is easy to verify in practice, because it only involves a comparison of order of magnitudes, whereas Definition 7 requires determining the precise prefactor in front of mixing times. It is easy to see that the product condition is necessary for cutoff:

Lemma 5 (No cutoff without the product condition)**.** *Any sequence of transition kernels $(P_n)_{n\geq 1}$ which exhibits cutoff must satisfy the product condition.*

Proof. Suppose that $(P_n)_{n\geq 1}$ exhibits cutoff, and fix $\varepsilon \in (0, \frac{1}{2})$. We then have $t_{\mathrm{MIX}}(P_n) \sim t^{(\varepsilon)}_{\mathrm{MIX}}(P_n)$, as $n \to \infty$. On the other hand, the lower bound (23) implies

$$\mathrm{t}^{(\varepsilon)}_{\mathrm{MIX}}(P) \quad \geq \quad \mathrm{t}_{\mathrm{REL}}(P) \log\left(\frac{1}{2\varepsilon}\right).$$

Combining those two estimates, we see that

$$\limsup_{n\to\infty} \left\{ \frac{\mathrm{t}_{\mathrm{REL}}(P_n)}{\mathrm{t}_{\mathrm{MIX}}(P_n)} \right\} \quad \leq \quad \frac{1}{\ln\left(\frac{1}{2\varepsilon}\right)},$$

and the right-hand side can be made arbitrarily small by choosing ε small. □

Unfortunately, the converse statement – which is the one that would have been useful in practice – does not hold in general. Even worse, any sequence $(P_n)_{n\geq 1}$ exhibiting cutoff can be perturbed so as to produce a counter-example, as we now explain. Given an ergodic transition kernel P with stationary law π on a finite state space $\mathscr{X}$, define

$$Q \quad := \quad (1-\theta)P + \theta\Pi, \tag{28}$$

where Π is the rank-one transition kernel defined at (15), and $\theta \in (0,1)$ a parameter to be adjusted later. Note that Q is ergodic, with stationary law π. The interpretation of Q is simple: at each step, a biased coin with parameter θ is tossed: if it lands on tails, the next state is chosen according to P; otherwise, it is chosen according to the stationary law π.

Lemma 6 (Rank-one perturbations destroy cutoff)**.** *Let $(P_n)_{n\geq 1}$ be a sequence of ergodic transition kernel exhibiting cutoff, and choose $(\theta_n)_{n\geq 1}$ in $(0,1)$ so that*

$$\frac{1}{\mathrm{t}_{\mathrm{MIX}}(P_n)} \ \ll \ \theta_n \ \ll \ \frac{1}{\mathrm{t}_{\mathrm{REL}}(P_n)}.$$

Then, the sequence $(Q_n)_{n\geq 1}$ defined by the rank-one perturbation (28) satisfies

$$\mathrm{t}_{\mathrm{REL}}(Q_n) \ \sim \ \mathrm{t}_{\mathrm{REL}}(P_n), \qquad \textit{and} \qquad \mathrm{t}^{(\varepsilon)}_{\mathrm{MIX}}(Q_n) \ \sim \ \frac{1}{\theta_n}\log\left(\frac{1}{\varepsilon}\right).$$

In particular, $(Q_n)_{n\geq 1}$ satisfies the product condition, but fails to exhibit cutoff.

Proof. The impact of the rank-one perturbation (28) on the distance to equilibrium is easy to determine: we have $Q-\Pi = (1-\theta)(P-\Pi)$, so that (16) yields

$$\forall t\geq 0, \qquad \mathscr{D}_Q(t) \ = \ (1-\theta)^t \mathscr{D}_P(t).$$

Recalling Corollary 1, we deduce that $\lambda_\star(Q) = (1-\theta)\lambda_\star(P)$, i.e.

$$\mathrm{t}_{\mathrm{REL}}(Q) \ = \ \frac{\mathrm{t}_{\mathrm{REL}}(P)}{1-\mathrm{t}_{\mathrm{REL}}(P)\log(1-\theta)}.$$

Specializing these general identities to $P=P_n$ and $\theta=\theta_n$ easily leads to the claim. □

Note that when n is large, the entries of the perturbed matrix Q_n are extremely close to those of P_n. Yet, $(P_n)_{n\geq 1}$ satisfies cutoff, whereas $(Q_n)_{n\geq 1}$ does not: cutoff is a delicate and sensitive phenomenon, which can not be captured by a basic criterion such as the product condition. Nevertheless, chains that satisfy the product condition without exhibiting cutoff are regarded as *pathological* by the community, and an informal conjecture states that the product condition should correctly predict cutoff for all *reasonable* chains. Giving an honest mathematical content to this vague claim constitutes a natural open problem, which can be seen as a first step towards Question 2.

Question 3 (Reasonable?)**.** *For which chains does the product condition imply cutoff?*

Known answers include birth and death chains and random walks on trees. However, the above construction shows that the product condition will incorrectly predict cutoff for many chains, including certain random walks on groups as defined next.

1.5 Random walks on graphs and groups

Among the various chains that will be considered in this course, a particular attention is devoted to random walks on groups and graphs, which play an important role in many modern applications, from card shuffling to page-rank algorithms or the exploration of complex networks. Let us here briefly recall how these random walks are defined.

Definition 9 (Group). *A group is a pair $(\mathscr{X}, \star)$, where $\mathscr{X}$ is a set and $\star$ a binary operation on $\mathscr{X}$ satisfying the following axioms:*

(i) There is an identity element $\mathrm{id} \in \mathscr{X}$ satisfying $x \star \mathrm{id} = \mathrm{id} \star x = x$ for all $x \in \mathscr{X}$.

(ii) Every element $x \in \mathscr{X}$ admits an inverse $x^{-1} \in \mathscr{X}$ such that $x \star x^{-1} = x^{-1} \star x = \mathrm{id}$.

(iii) The operation $\star$ is associative, i.e. $(x \star y) \star z = x \star (y \star z)$ for all $(x, y, z) \in \mathscr{X}^3$.

Here are three classical examples of finite groups to which we shall come back later:

- The cyclic group $(\mathbb{Z}_n, +)$ of integers modulo n, equipped with addition modulo n.
- The hypercube $(\mathbb{Z}_2^n, +)$ of binary vectors of length n, equipped with addition mod 2.
- The symmetric group $(\mathfrak{S}_n, \circ)$ of permutations of $[n]$, equipped with composition.

Given a finite group $(\mathscr{X}, \star)$ and a probability distribution $\mu \in \mathcal{P}(\mathscr{X})$, let $(Z_t)_{t \geq 1}$ be i.i.d. random variables with law μ, and consider the process $\mathbf{X} = (X_t)_{t \geq 1}$ defined by

$$X_t \;:=\; Z_t \star \cdots \star Z_1, \tag{29}$$

with the usual convention that an empty product is the identity. Then $\mathbf{X}$ is a Markov chain on $\mathscr{X}$, called the random walk on $(\mathscr{X}, \star)$ with increment law μ. Its transition kernel is

$$\forall x, y \in \mathscr{X}, \qquad P(x, y) \;:=\; \mu(y \star x^{-1}).$$

This matrix is bi-stochastic, meaning that its transpose is also stochastic. In other words, the uniform distribution $\pi = \mathrm{Unif}(\mathscr{X})$ is stationary for P. Note that P is irreducible if and only if the incremental support $\mathrm{supp}(\mu) := \{x \in \mathscr{X} : \mu(x) > 0\}$ generates $(\mathscr{X}, \star)$, in the sense that any group element x can be written as $x = x_t \star \cdots \star x_1$ for some $t \in \mathbb{N}$ and some $x_1, \ldots, x_t \in \mathrm{supp}(\mu)$. A pleasant feature of random walks on groups is their intrinsic symmetry. In particular, the choice of the initial state is irrelevant.

Lemma 7 (Symmetry). *For random walks on groups, we have for all $x \in \mathscr{X}$ and $t \in \mathbb{N}$,*

$$d_{\mathrm{TV}}\left(P^t(x,\cdot),\pi\right) \;=\; d_{\mathrm{TV}}\left(P^t(\mathrm{id},\cdot),\pi\right).$$

Proof. By induction over $t \in \mathbb{N}$, we have $P^t(x,y) = P^t(\mathrm{id}, x^{-1} \star y)$ for all $x, y \in \mathscr{X}$. Thus,

$$\begin{aligned} d_{\mathrm{TV}}\left(P^t(x,\cdot),\pi\right) &= \frac{1}{2}\sum_{y\in\mathscr{X}}\left|P^t(\mathrm{id}, y\star x^{-1}) - \frac{1}{|\mathscr{X}|}\right| \\ &= \frac{1}{2}\sum_{y\in\mathscr{X}}\left|P^t(\mathrm{id}, y) - \frac{1}{|\mathscr{X}|}\right| \\ &= d_{\mathrm{TV}}\left(P^t(\mathrm{id},\cdot),\pi\right), \end{aligned}$$

where the second equality uses the bijective change of variables $y \mapsto y \star x$. □

Definition 10 (Graph). *A (finite, simple) graph is a pair $G = (\mathscr{X}, E)$, where*

- *$\mathscr{X}$ is a finite set whose elements are called vertices;*
- *E is a set of unordered pairs of vertices $\{x, y\}$ called edges.*

Two vertices $x, y \in \mathscr{X}$ such that $\{x, y\} \in E$ are said to be neighbors, and the number of neighbors of x is called its degree, denoted by $\deg(x)$.

If a graph $G = (\mathscr{X}, E)$ has degrees at least 1, we may define a transition kernel on $\mathscr{X}$ by

$$P(x,y) \;:=\; \begin{cases} \frac{1}{\deg(x)} & \text{if } \{x,y\} \in E \\ 0 & \text{else.} \end{cases}$$

The corresponding Markov chain is called simple random walk on G. It describes the evolution of a walker which, at each step, jumps from the current vertex to a uniformly chosen neighbor. P is irreducible if and only if G is connected, meaning that one can go from any vertex to any other by traversing a sequence of edges. Note that the formula

$$\forall x \in \mathscr{X}, \qquad \pi(x) \;:=\; \frac{\deg(x)}{2|E|},$$

defines a probability vector on $\mathscr{X}$, which satisfies the detailed balance equation

$$\forall (x,y) \in \mathscr{X}^2, \quad \pi(x)P(x,y) \;=\; \pi(y)P(y,x). \tag{30}$$

By summation over all $x \in \mathscr{X}$, this implies that π is stationary for P.

2 Probabilistic techniques

In this chapter, we introduce two simple but powerful probabilistic tools to estimate mixing times: distinguishing statistics (to obtain lower bounds), and couplings (to obtain upper bounds). These important techniques are then applied to obtain sharp estimates for the random walk on the cycle and the Ehrenfest Urn model.

2.1 Distinguishing statistics

In practice, it is often easier to obtain lower bounds on mixing times than upper bounds. The reason is that the distance to equilibrium is defined as a (double) maximum:

$$\mathscr{D}_P(t) \;=\; \max_{x\in\mathscr{X}} \max_{A\subseteq\mathscr{X}} \left|P^t(x,A)-\pi(A)\right|.$$

It readily follows from this definition that any choice of an initial state $x \in \mathscr{X}$ and a target event $A \subseteq \mathscr{X}$ provides a lower bound on the distance to equilibrium. This trivial observation will be used so often that it deserves a lemma for better visibility.

Lemma 8 (Distinguishing event)**.** *The lower bound*

$$\mathscr{D}_P(t) \;\geq\; \left|P^t(x,A)-\pi(A)\right|,$$

holds for any time $t \geq 0$, any initial state $x \in \mathscr{X}$ and any event $A \subseteq \mathscr{X}$.

To get a good bound, the pair (x, A) should of course be chosen so that $P^t(x, A)$ is abnormally large or small compared to the equilibrium value $\pi(A)$. In practice, good candidates for (x, A) are easily guessed, but estimating $P^t(x, A)-\pi(A)$ can be difficult. A simple alternative consists in computing the first and second moment of an observable $f\colon \mathscr{X} \to \mathbb{R}$ that behaves very abnormally when the chain is far from equilibrium. This formalizes as follows.

Lemma 9 (Distinguishing statistics)**.** *For any $\mu, \nu \in \mathcal{P}(\mathscr{X})$ and $f\colon \mathscr{X} \to \mathbb{R}$, we have*

$$d_{\mathrm{TV}}(\mu,\nu) \;\geq\; \frac{\delta^2}{\delta^2+\sigma^2}.$$

where $\delta = |\mu f - \nu f|$ and $\sigma^2 = 2\mathrm{Var}_\mu(f) + 2\mathrm{Var}_\nu(f)$.

Proof. Since the right-hand side is invariant under translating f by a constant, we may assume that $\mu f + \nu f = 0$, so that $(\mu f)^2 = (\nu f)^2 = \delta^2/4$. By Cauchy-Schwartz, we have

$$\begin{aligned} \delta^2 &= \left(\sum_{x \in \mathscr{X}} (\mu(x) - \nu(x)) f(x) \right)^2 \\ &\leq \left(\sum_{x \in \mathscr{X}} (\mu(x) + \nu(x)) f^2(x) \right) \left(\sum_{x \in \mathscr{X}} \frac{(\mu(x) - \nu(x))^2}{\mu(x) + \nu(x)} \right). \end{aligned}$$

The first sum is exactly equal to $(\sigma^2 + \delta^2)/2$, while the second is at most $2d_{\mathrm{TV}}(\mu, \nu)$ by the crude bound $|\mu(x) - \nu(x)| \leq \mu(x) + \nu(x)$. Rearranging yields the claim. □

Remark 9 (Concentration)**.** *The above bound says that μ and ν are far apart if the ratio σ^2/δ^2 is small. The intuition is as follows: under the measure μ, the observable f typically takes values in $\mathcal{I}_\mu := \left[\mu(f) - 10\sqrt{\mathrm{Var}_\mu(f)}, \mu(f) + 10\sqrt{\mathrm{Var}_\mu(f)}\right]$ by Chebychev's inequality, and similarly for ν. When σ^2/δ^2 is small, the two intervals $\mathcal{I}_\mu$ and $\mathcal{I}_\nu$ are disjoint, so $A := f^{-1}(\mathcal{I}_\mu)$ forms a distinguishing event showing that $d_{\mathrm{TV}}(\mu, \nu)$ is large.*

Remark 10 (Complex values)**.** *The proof readily extends to the case of a complex-valued observable $f \colon \mathscr{X} \to \mathbb{C}$, with $\mathrm{Var}_\mu(f) := \mu\left(|f - \mu f|^2\right) = \mu|f|^2 - |\mu f|^2$.*

We will illustrate the strength of those generic lower bounds on various concrete Markov chains once we have a complementary technique to obtain matching upper bounds.

2.2 Couplings

In probability theory, *coupling* is a very general technique which allows one to compare two given distributions μ and ν by constructing a pair of random variables (X, Y) whose marginal distributions are μ and ν. Every such pair provides a particular *relation* between μ and ν, and the whole idea is to choose a relation that sheds useful light on μ and ν.

Definition 11 (Coupling)**.** *A coupling of two probability measures μ, ν is a pair (X, Y) of random variables defined on the same probability space, such that $X \sim \mu$ and $Y \sim \nu$.*

Of course, we can always take X and Y to be independent with respective marginals μ and ν, but this is usually not the most interesting choice: the above definition allows for X and Y to be *entangled* in an arbitrary way, and this degree of freedom can often be exploited to obtain non-trivial information about the pair (μ, ν). The following lemma is a simple but fruitful illustration of this general philosophy.

Lemma 10 (Estimating total variation distance by coupling)**.** *Fix $\mu, \nu \in \mathcal{P}(\mathscr{X})$. Then,*

$$d_{\text{TV}}(\mu, \nu) \;\leq\; \mathbb{P}(X \neq Y),$$

for any coupling (X, Y) of μ and ν. Moreover, there is a coupling which achieves equality.

Proof. If (X, Y) is a coupling of μ and ν, then for any set $A \subseteq \mathscr{X}$ we can write

$$\begin{aligned}
\mu(A) - \nu(A) &= \mathbb{P}(X \in A) - \mathbb{P}(Y \in A) \\
&\leq \mathbb{P}(X \in A) - \mathbb{P}(X \in A, Y \in A) \\
&= \mathbb{P}(X \in A, Y \notin A) \\
&\leq \mathbb{P}(X \neq Y).
\end{aligned}$$

Taking a maximum over all $A \subseteq \mathscr{X}$ establishes the first claim. Conversely, let us construct a coupling (X, Y) which achieves equality. The extreme cases $d_{\text{TV}}(\mu, \nu) = 0$ and $d_{\text{TV}}(\mu, \nu) = 1$ are trivial, so we leave them aside. By Lemma 2, we have $d_{\text{TV}}(\mu, \nu) = 1 - p$, where

$$p \;:=\; \sum_{x \in \mathscr{X}} \mu(x) \wedge \nu(x).$$

We thus want to construct a coupling (X, Y) such that $\mathbb{P}(X = Y) = p$. To do so, we define

$$(X, Y) \;:=\; \begin{cases} (Z, Z) & \text{if } B = 1 \\ (\widehat{X}, \widehat{Y}) & \text{if } B = 0, \end{cases}$$

where $B, Z, \widehat{X}, \widehat{Y}$ are independent, with $B \sim \text{Bernoulli}(p)$ and for all $x \in \mathscr{X}$,

$$\begin{aligned}
\mathbb{P}(Z = x) &= \frac{\mu(x) \wedge \nu(x)}{p} \\
\mathbb{P}(\widehat{X} = x) &= \frac{(\mu(x) - \nu(x))_+}{1 - p}, \\
\mathbb{P}(\widehat{Y} = x) &= \frac{(\nu(x) - \mu(x))_+}{1 - p}.
\end{aligned}$$

It is immediate to check that (X, Y) is a coupling of μ and ν such that $\mathbb{P}(X = Y) = p$. $\square$

Remark 11 (Variational formula for total variation distance)**.** *Lemma 10 provides yet another definition of total variation distance, to be added to the list of Lemma 2:*

$$d_{\mathrm{TV}}(\mu,\nu) \;=\; \min_{X\sim\mu, Y\sim\nu} \mathbb{P}(X\neq Y).$$

A considerable advantage of this new expression is the fact that it involves a minimum: any coupling (X,Y) of μ,ν provides an upper bound *on $d_{\mathrm{TV}}(\mu,\nu)$. The more likely the event $\{X=Y\}$ is, the better the bound will be, and the existence of a coupling achieving equality guarantees that this strategy has no intrinsic limitation. In practice however, estimating $\mathbb{P}(X=Y)$ can become difficult if the coupling is too sophisticated, and devising couplings that are both efficient and tractable is a delicate art.*

In order to use Lemma 10 to estimate the mixing time of an ergodic kernel P, we need to construct, for an appropriate time $t\in\mathbb{N}$ and for every initial state $x\in\mathscr{X}$, a coupling between $P^t(x,\cdot)$ and π which puts as much mass as possible on the diagonal set $\{(y,y)\colon y\in\mathscr{X}\}$. This strategy may seem difficult to implement at first sight, but we will now make a couple of important observations that will considerably facilitate our task.

2.3 Coalescence times

A first useful observation is that we do not need to compare $P^t(x,\cdot)$ directly with the equilibrium distribution π (which is sometimes very far from being explicit): it is enough to compare $P^t(x,\cdot)$ with the more similar measure $P^t(y,\cdot)$, for all $(x,y)\in\mathscr{X}^2$.

Lemma 11 (Mixing means forgetting)**.** *For every $t\in\mathbb{N}$, we have*

$$\frac{1}{2}\max_{(x,y)\in\mathscr{X}^2} d_{\mathrm{TV}}\left(P^t(x,\cdot),P^t(y,\cdot)\right) \;\leq\; \mathscr{D}_P(t) \;\leq\; \max_{(x,y)\in\mathscr{X}^2} d_{\mathrm{TV}}\left(P^t(x,\cdot),P^t(y,\cdot)\right).$$

Proof. By stationarity, we have $\pi=\pi P=\pi P^2=\cdots=\pi P^t=\sum_{y\in\mathscr{X}}\pi(y)P^t(y,\cdot)$. Thus, π is a convex combination of the measures $\{P^t(y,\cdot)\colon y\in\mathscr{X}\}$. Since $d_{\mathrm{TV}}(\cdot,\cdot)$ is convex (Lemma 3), we deduce that for any $\mu\in\mathcal{P}(\mathscr{X})$,

$$d_{\mathrm{TV}}\left(\mu,\pi\right) \;\leq\; \max_{y\in\mathscr{X}} d_{\mathrm{TV}}\left(\mu,P^t(y,\cdot)\right).$$

Choosing $\mu = P^t(x,\cdot)$ and then maximizing over all $x \in \mathscr{X}$ leads to the claimed upper bound. For the lower bound, we simply invoke the triangle inequality

$$d_{\text{TV}}\left(P^t(x,\cdot), P^t(y,\cdot)\right) \;\leq\; d_{\text{TV}}\left(P^t(x,\cdot), \pi\right) + d_{\text{TV}}\left(P^t(y,\cdot), \pi\right),$$

and then take a maximum over all initial conditions $(x,y) \in \mathscr{X}^2$. □

We are thus naturally led to the problem of coupling $P^t(x,\cdot)$ and $P^t(y,\cdot)$, for arbitrary $x, y \in \mathscr{X}$ and $t \in \mathbb{N}$. Recall that the instantaneous measures $P^t(x,\cdot)$ and $P^t(y,\cdot)$ were actually constructed sequentially, by iterating t times the map $\mu \mapsto \mu P$, starting from the initial measures δ_x and δ_y, respectively. In light of this sequential structure, the most natural way to couple $P^t(x,\cdot)$ and $P^t(y,\cdot)$ is to actually couple the *entire* chains, i.e. to produce a random pair $(\mathbf{X}, \mathbf{Y})$ such that $\mathbf{X} \sim \text{MC}(\mathscr{X}, P, \delta_x)$ and $\mathbf{Y} \sim \text{MC}(\mathscr{X}, P, \delta_y)$. For each $t \in \mathbb{N}$, the random pair (X_t, Y_t) is then a coupling of $P^t(x,\cdot)$ and $P^t(y,\cdot)$, so we have

$$d_{\text{TV}}\left(P^t(x,\cdot), P^t(y,\cdot)\right) \;\leq\; \mathbb{P}(X_t \neq Y_t). \tag{31}$$

Now, a very simple way to produce such a trajectorial coupling, simultaneously for all initial pairs $(x,y) \in \mathscr{X}^2$, is to use what is known as a *coupling kernel*.

Definition 12 (Coupling kernel)**.** *A coupling kernel for P is a transition kernel Q on the product space $\mathscr{X} \times \mathscr{X}$, whose marginals agree with P in the following sense:*

$$\begin{aligned}
\forall (x,y,x') \in \mathscr{X}^3, \qquad \sum_{y' \in \mathscr{X}} Q\left((x,y),(x',y')\right) &= P(x,x') \\
\forall (x,y,y') \in \mathscr{X}^3, \qquad \sum_{x' \in \mathscr{X}} Q\left((x,y),(x',y')\right) &= P(y,y').
\end{aligned}$$

Any Markov chain $(\mathbf{X}, \mathbf{Y})$ on $\mathscr{X} \times \mathscr{X}$ whose transition kernel is of this form is called a Markovian coupling for P, and the associated coalescence time is defined as

$$T \;:=\; \inf\left\{t \geq 0 \colon X_t = Y_t\right\},$$

with the usual convention $\inf \emptyset = +\infty$.

With this terminology at hands, we can now state and prove the main result of this section, which asserts that the convergence to equilibrium of a Markov chain is fast if trajectories from different starting states can be coupled so as to meet quickly.

Theorem 2 (Coalescence and mixing)**.** *Consider an arbitrary coupling kernel Q for P, and let T denote the associated coalescence time. Then,*

$$\forall t \geq 0, \qquad \mathscr{D}_P(t) \;\leq\; \max_{(x,y)\in\mathscr{X}^2} \mathbb{P}_{(x,y)}\left(T > t\right),$$

where the notation $\mathbb{P}_{(x,y)}$ indicates that the initial state is (x, y).

Proof. Let $(\mathbf{X}, \mathbf{Y})$ denote a Markov chain on $\mathscr{X}^2$ with transition kernel Q. The fact that Q is a coupling kernel ensures that for each $t \geq 1$, the conditional laws of X_t and Y_t given the past $(X_0, \ldots, X_{t-1}, Y_0, \ldots, Y_{t-1})$ are $P(X_t, \cdot)$ and $P(Y_t, \cdot)$, respectively. In particular, $\mathbf{X}$ and $\mathbf{Y}$ are Markov chains on $\mathscr{X}$ with transition kernel P. Consequently, Lemma 10 yields

$$d_{\text{TV}}\left(P^t(x,\cdot), P^t(y,\cdot)\right) \;\leq\; \mathbb{P}_{(x,y)}(X_t \neq Y_t), \tag{32}$$

for every $t \in \mathbb{N}$ and every $(x, y) \in \mathscr{X}^2$. Now, let us make the additional assumption that the diagonal set $\Delta := \{(z, z) \colon z \in \mathscr{X}\}$ is absorbing for the coupling kernel Q, in the sense that

$$\forall x \in \mathscr{X}, \qquad \sum_{y\in\mathscr{X}} Q\left((x,x),(y,y)\right) \;=\; 1. \tag{33}$$

Then almost-surely, the trajectories $\mathbf{X}$ and $\mathbf{Y}$ coincide forever after coalescing, hence

$$\mathbb{P}_{(x,y)}(X_t \neq Y_t) \;=\; \mathbb{P}_{(x,y)}(T > t).$$

Inserting this into (32) and then taking the maximum over all pairs $(x, y) \in \mathscr{X}^2$ establishes the claim (recall Lemma 11). Finally, note that if Q fails to satisfy the condition (33), then we can always replace it with the new coupling kernel

$$\widetilde{Q}\left((x,y),(x',y')\right) \;:=\; \begin{cases} Q\left((x,y),(x',y')\right) & \text{if } x \neq y \\ P(x,x') & \text{if } x = y \text{ and } x' = y' \\ 0 & \text{if } x = y \text{ and } x' \neq y'. \end{cases}$$

Since $\widetilde{Q}$ is a coupling kernel for P which satisfies (33), we obtain

$$\forall t \geq 0, \qquad \mathscr{D}_P(t) \;\leq\; \max_{(x,y)\in\mathscr{X}^2} \mathbb{P}_{(x,y)}\left(\widetilde{T} > t\right),$$

where $\widetilde{T}$ denotes the coalescence time under $\widetilde{Q}$. But $\widetilde{T}$ has the same law as T, because $\widetilde{Q}$ and Q differ only on transitions that start from the diagonal set Δ. □

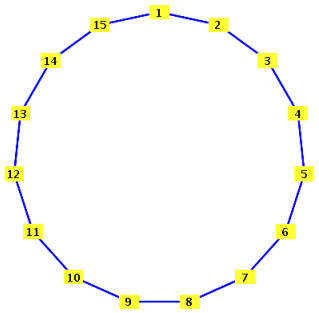

Figure 4: The $n-$cycle with $n = 15$.

Remark 12 (Product kernel)**.** *The above result provides a simple proof of the convergence to equilibrium of ergodic Markov chains (Theorem 1). Indeed, the product kernel*

$$Q\left((x,y),(x',y')\right) \;:=\; P(x,x') \times P(y,y'),$$

corresponding to the independent evolution of $\mathbf{X}$ *and* $\mathbf{Y}$*, is obviously a coupling kernel for* P*. Moreover, it is clear that the above formula* tensorizes*, in the sense that*

$$Q^t\left((x,y),(x',y')\right) \;:=\; P^t(x,x') \times P^t(y,y'),$$

for all $t \geq 0$*. In particular,* Q *is ergodic (since so is* P*), and this ensures that*

$$\forall (x,y) \in \mathscr{X}^2, \qquad \mathbb{P}_{(x,y)}\left(T < \infty\right) \;=\; 1.$$

Consequently, Theorem 2 gives $\mathscr{D}_P(t) \to 0$ *as* $t \to \infty$*, as desired.*

2.4 Application: random walk on the cycle

As a first pedagogic illustration, let us consider the lazy random walk on the n cycle. The state space is $\mathscr{X} = \mathbb{Z}/n\mathbb{Z}$, and the transition kernel is

$$P(x,y) \;=\; \begin{cases} \frac{1}{2} & \text{if } y = x \\ \frac{1}{4} & \text{if } y \in \{x-1, x+1\} \\ 0 & \text{else.} \end{cases}$$

This is the transition kernel of the lazy random walk on the $n-$cycle graph. It is also a random walk on the cyclic group $(\mathbb{Z}/n\mathbb{Z}, +)$, with increment law $\mu = \frac{1}{2}\delta_0 + \frac{1}{4}\delta_1 + \frac{1}{4}\delta_{-1}$. We

will show that the mixing time of this random walk grows quadratically with n.

Proposition 1 (Mixing time of the $n-$cycle)**.** *For lazy random walk on the $n-$cycle,*

$$\frac{n^2}{32} \;\leq\; \mathrm{t}_{\text{MIX}}(P) \;\leq\; n^2.$$

Upper bound: first attempt. The most natural way to construct a chain $\mathbf{X}$ with transition kernel P starting from a given state $x \in \mathscr{X}$ is to set $X_0 := x$ and for all $t \geq 1$,

$$X_t \;:=\; X_{t-1} + \xi_t \bmod n.$$

where $(\xi_t)_{t\geq 1}$ are i.i.d. with $\mathbb{P}(\xi_1 = 0) = \frac{1}{2}$ and $\mathbb{P}(\xi_1 = 1) = \mathbb{P}(\xi_1 = -1) = \frac{1}{4}$. In light of this, a naive way to couple $\mathbf{X}$ with a chain $\mathbf{Y}$ starting from another state $y \in \mathscr{X}$ consists in using the same increments for both chains, i.e. setting $Y_0 := y$ and for all $t \geq 1$,

$$Y_t \;:=\; Y_{t-1} + \xi_t \bmod n. \tag{34}$$

It is then clear that $(\mathbf{X}, \mathbf{Y})$ is a Markovian coupling for P, with coupling kernel

$$Q\left((x,y),(x',y')\right) \;=\; \begin{cases} \frac{1}{2} & \text{if } (x',y') = (x,y) \\ \frac{1}{4} & \text{if } (x',y') \in \{(x+1,y+1),(x-1,y-1)\} \\ 0 & \text{else.} \end{cases}$$

Unfortunately, this coupling is not smart at all: the difference $X_t - Y_t$ is preserved at each step, so the coalescence time T is a.-s. infinite, and Theorem 2 only yields $\mathscr{D}_P(t) \leq 1$!

Upper bound: second attempt. In order to "favor encounter", one could try to let the two trajectories move in opposite directions, i.e. replace (34) with

$$Y_t \;:=\; Y_{t-1} - \xi_t \bmod n.$$

Note that this remains a valid coupling, because the increment sequence $(-\xi_t)_{t\geq 1}$ has the same law as $(\xi_t)_{t\geq 1}$. The corresponding coupling kernel is then

$$Q\left((x,y),(x',y')\right) \;=\; \begin{cases} \frac{1}{2} & \text{if } (x',y') = (x,y) \\ \frac{1}{4} & \text{if } (x',y') \in \{(x+1,y-1),(x-1,y+1)\} \\ 0 & \text{else.} \end{cases}$$

Unfortunately, this choice is also problematic: when n is even and $x-y$ is odd, the difference $X_t - Y_t$ remains odd at each iteration, so that $T = \infty$ a.-s. again!

Upper bound: third attempt. A solution to this parity issue consists in letting only one of the two coordinates jump at each step, as dictated by the following coupling kernel:

$$Q\left((x,y),(x',y')\right) \;=\; \begin{cases} \frac{1}{4} & \text{if } |x'-x|+|y-y'|=1 \\ 0 & \text{else.} \end{cases}$$

The sequence of differences $(X_t - Y_t)_{t\geq 0}$ is then a simple random walk on $\mathbb{Z}/n\mathbb{Z}$ starting from $x-y$, and the coalescence time T is precisely the time it takes for this walk to hit 0. Equivalently, T is the hitting time of the set $\{0,n\}$ by a simple random walk $(W_t)_{t\geq 0}$ on $\mathbb{Z}$ starting from $W_0 = |x-y|$. The expectation of T is easily computed, for example by Doob's optional stopping theorem applied to the martingales $(W_t)_{t\geq 0}$ and $(W_t^2 - t)_{t\geq 0}$:

$$\begin{aligned} \mathbb{E}\left[W_T\right] \;&=\; \mathbb{E}[W_0] \;=\; |x-y|; \\ \mathbb{E}\left[W_T^2 - T\right] \;&=\; \mathbb{E}[W_0^2 - 0] \;=\; |x-y|^2. \end{aligned}$$

Since the random variable W_T is $\{0,n\}-$valued, we have $W_T^2 = nW_T$ and it follows that

$$\mathbb{E}[T] \;=\; (n-|x-y|)\,|x-y| \;\leq\; \frac{n^2}{4}.$$

Applying Theorem 2 and Markov's inequality, we conclude that $\mathscr{D}_P(t) \leq \frac{n^2}{4t}$, or equivalently,

$$\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P) \;\leq\; \left\lceil \frac{n^2}{4\varepsilon} \right\rceil.$$

This proves the upper bound in Proposition 1. We now turn to the lower bound.

Lower bound. Intuitively, when the number t of iterations is too small, the random walk X_t is confined in a small interval around its starting point, and hence can not be mixed. To formalize this, we use Lemma 8 with the starting state $x=0$ and the distinguishing event $A = \left[\frac{n}{4}, \frac{3n}{4}\right] \cap \mathbb{N}$. We have $\pi(A) = \frac{|A|}{|\mathscr{X}|} \geq \frac{1}{2}$ and by Chebychev inequality,

$$P^t(x,A) \;\leq\; \mathbb{P}\left(|\xi_1 + \cdots + \xi_t| \geq \frac{n}{4}\right) \;\leq\; \frac{8t}{n^2},$$

where we have used the fact that $\xi_1, \ldots, \xi_t$ are i.i.d. with mean 0 and variance $1/2$. For $t < n^2/32$ the right-hand side is less than $1/4$, and Lemma 8 yields $\mathscr{D}_P(t) > 1/4$, as desired.

2.5 Application: Ehrenfest model

Introduced by Tatiana and Paul Ehrenfest to explain the second law of thermodynamics, the Ehrenfest urns is a very simple model for the diffusion of gas molecules. Consider n labelled particles evolving among two neighboring containers as follows: at each time step, a particle is chosen uniformly at random and jumps from its current container to the other one. If one start with, say, all particles in one container, how long will it take for the gas to equilibriate? We can describe the state of the system by a binary vector $x = (x_1, \dots, x_n)$, where the variable $x_i \in \{0, 1\}$ indicates the container in which the $i-$th particle currently lies. The above diffusion is then a Markov chain with state space $\mathscr{X} := \{0, 1\}^n$ and transition kernel

$$\widetilde{P}(x, y) \;:=\; \begin{cases} \frac{1}{n} & \text{if } x, y \text{ differ by exactly one coordinate} \\ 0 & \text{else.} \end{cases}$$

This is the transition kernel of simple random walk on a well-known graph, namely the $n-$dimensional hypercube. Alternatively, $\widetilde{P}$ is the transition kernel of the random walk on the binary group $\mathbb{F}_2^n$ with increments being uniform on the set of vectors having exactly one non-zero coordinate. To avoid periodicity issues, we consider the lazy kernel $P = (I + \widetilde{P})/2$.

Proposition 2 (Mixing time of the lazy Ehrenfest urns)**.** *For any $\varepsilon \in (0, 1)$, we have*

$$\frac{(1 - o(1))}{2} n \log n \;\le\; \mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P_n) \;\le\; (1 + o(1)) n \log n$$

Proof. Given an initial state $x \in \mathscr{X}$, a convenient sequential construction of the trajectory $\mathbf{X} \sim \mathrm{MC}(\mathscr{X}, P, \delta_x)$ consists in setting $X_0 = x$ and then, for all $t \ge 1$,

$$X_t \;:=\; F(X_{t-1}, I_t, B_t),$$

where $F(x) := (x_1, \dots, x_{i-1}, b, x_{i+1}, \dots, x_n)$, and where the random variables $I_t, B_t, t \ge 1$ are independent with $B_t \sim \mathrm{Bernoulli}(1/2)$ and $I_t \sim \mathrm{Unif}(\{1, \dots, n\})$. Given another initial state $y \in \mathscr{X}$, one can then naturally couple $\mathbf{Y} \sim \mathrm{MC}(\mathscr{X}, P, \delta_y)$ with $\mathbf{X}$ by using the same update variables $(I_t, B_t)_{t \ge 1}$ for both trajectories, i.e. by setting $Y_0 = y$ and for $t \ge 1$,

$$Y_t \;:=\; F(Y_{t-1}, I_t, B_t).$$

The pair $(\mathbf{X}, \mathbf{Y})$ is then a Markovian coupling for P. By construction, at any time $t \ge 0$, the two random vectors X_t and Y_t agree on the set of coordinates $\{I_1, \dots, I_t\}$. Consequently,

the coalescence time T is at most the time it takes for all coordinates to have been hit:

$$T \;\leq\; \inf\{t \geq 0 \colon \{I_1, \ldots, I_t\} = \{1, \ldots, n\}\}.$$

(In fact, there is even equality when x and y are antipodal.) Consequently, we have

$$\begin{aligned}
\mathscr{D}_P(t) &\leq \mathbb{P}(\{I_1, \ldots, I_t\} \neq \{1, \ldots, n\}) \\
&\leq \sum_{i=1}^{n} \mathbb{P}(i \notin \{I_1, \ldots, I_t\}) \\
&= n\left(1 - \frac{1}{n}\right)^t \\
&\leq ne^{-t/n}
\end{aligned}$$

where we have successively used the union bound, the fact that $I_1, \ldots, I_t$ are independent and uniform on $\{1, \ldots, n\}$, and the convexity inequality $1 + u \leq e^u$, valid for all $u \in \mathbb{R}$. The upper bound readily follows from this. For the lower bound, we apply the method of distinguishing statistics (Lemma 9) to the observable

$$f \colon x \;\mapsto\; x_1 + \cdots + x_n,$$

which counts the number of coordinates equal to 1. Under the equilibrium law π, the coordinates are independent Bernoulli variables with parameter $1/2$, so we have

$$\begin{aligned}
\pi f &= \frac{n}{2} \\
\mathrm{Var}_\pi(f) &= \frac{n}{4}.
\end{aligned}$$

On the other hand, after t iterations starting from $x = (0, \ldots, 0)$, the coordinates are easily seen to be negatively correlated Bernoulli variables with parameter $\left(1 - (1 - \frac{1}{n})^t\right)/2$, so

$$\begin{aligned}
\mu f &= \frac{n}{2}\left(1 - \left(1 - \frac{1}{n}\right)^t\right) \\
\mathrm{Var}_\mu(f) &\leq \frac{n}{4},
\end{aligned}$$

where $\mu = P^t(0, \cdot)$. Thus, Lemma 9 yields

$$\mathscr{D}_P(t) \;\geq\; 1 - \frac{1}{1 + \frac{n}{4}\left(1 - \frac{1}{n}\right)^{2t}},$$

from which the lower bound readily follows. □

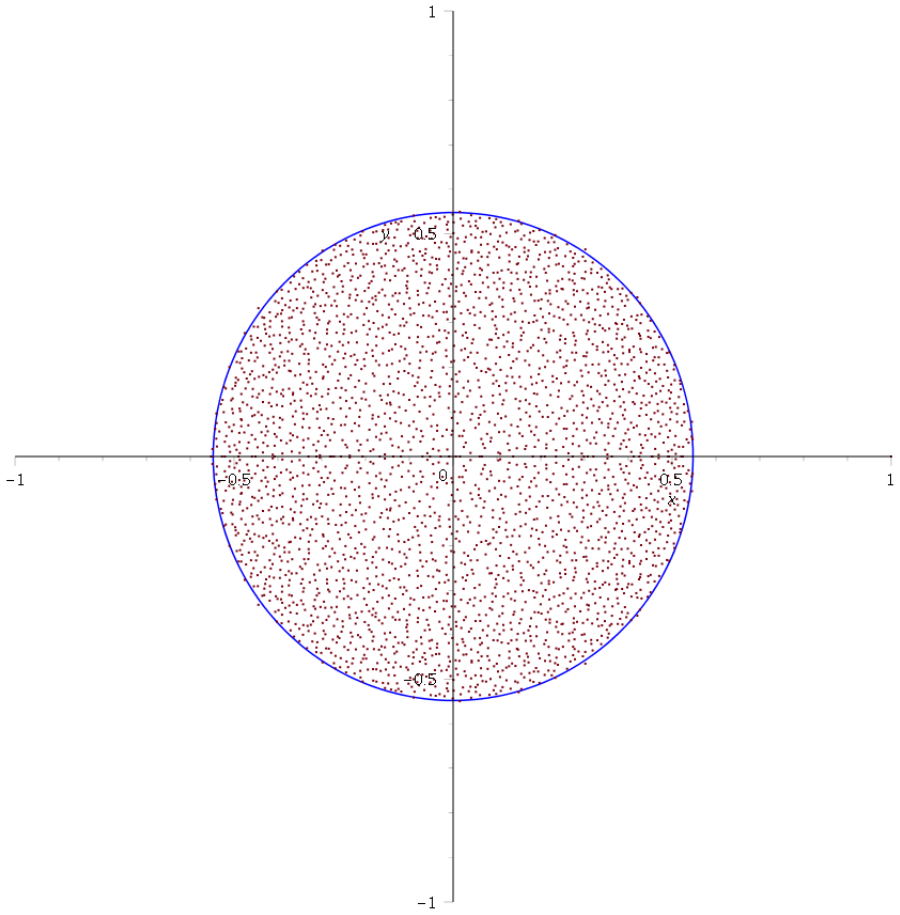

Figure 5: Eigenvalues (in red) and spectral radius (in blue) of a typical transition kernel.

3 Spectral techniques

This chapter investigates the eigenvalues and eigenvectors of transition kernels, and their relation to mixing times. This relation is particularly deep for reversible chains, to which we devote a central attention. To illustrate the strength of spectral techniques, we revisit two models that were introduced in the previous chapter and obtain considerably refined results on their convergence to equilibrium: we establish a limiting profile for random walk on the cycle, and we prove the cutoff phenomenon for random walk on the hypercube.

3.1 Spectral radius

We have seen earlier (Corollary 1) that the convergence to equilibrium of Markov chains occurs exponentially fast: more precisely, for any ergodic kernel P, the limit

$$\lambda_\star(P) \;:=\; \lim_{t\to\infty} (\mathscr{D}_P(t))^{\frac{1}{t}}$$

exists and is strictly less than 1. We will now see that this quantity admits a remarkable spectral characterization. Let us first recall some terminology. An eigenvalue of P is a root of the characteristic polynomial $\lambda \mapsto \det(P-\lambda\mathbb{I})$, i.e. a number $\lambda \in \mathbb{C}$ such that the equation

$$Pf \;=\; \lambda f \tag{35}$$

admits a non-trivial solution $f\colon \mathscr{X} \to \mathbb{C}$. Any such solution f is called an eigenfunction associated with the eigenvalue λ, and the pair (λ, f) an eigenpair of P. Since the characteristic polynomial has degree $N := |\mathscr{X}|$, P has precisely N eigenvalues, counted with algebraic multiplicities. The set of eigenvalues is called the spectrum of P, and denoted by $\mathrm{Sp}(P)$. For concreteness, Figure 5 shows a plot of the spectrum of a typical transition kernel. Here are a few elementary properties of the spectrum of any transition kernel.

Lemma 12 (Eigenvalues)**.** *Let P be any transition kernel on $\mathscr{X}$. Then,*

(i) $1 \in \mathrm{Sp}(P)$.

(ii) $\mathrm{Sp}(P) \subseteq \{\lambda \in \mathbb{C}\colon |\lambda| \le 1\}$.

(iii) If P is ergodic, then 1 is the only eigenvalue on the unit circle.

Proof. The constant function $f = \mathbf{1}$ solves the harmonic equation $Pf = f$, proving the first claim. The second follows from the fact that P contracts the $\|\cdot\|_\infty$ norm: for any $f\colon \mathscr{X} \to \mathbb{R}$,

$$\|Pf\|_\infty \;\le\; \|f\|_\infty .$$

Finally, consider an eigenpair (λ, f) with $|\lambda| = 1$. If P is ergodic, we can choose $t \in \mathbb{N}$ so that the number $\alpha := \min_{x\in\mathscr{X}} P^t(x,x)$ is strictly positive. But then, the matrix

$$Q \;:=\; \frac{P^t - \alpha\mathbb{I}}{1-\alpha}$$

is a transition kernel, of which f is an eigenfunction with eigenvalue $\frac{\lambda^t-\alpha}{1-\alpha}$. Thus, (ii) yields $|\lambda^t - \alpha| \le 1-\alpha$. Since $|\lambda^t| = 1$, this forces $\lambda^t = 1$. Replacing t with $t+1$ gives $\lambda = 1$. □

We now turn our attention to eigenfunctions.

Lemma 13 (Eigenfunctions)**.** *Let (λ, f) be an eigenpair of P. Then,*

(i) If $\lambda \neq 1$, then $\pi f = 0$.

(ii) If $\lambda = 1$ and P is irreducible, then f is constant.

Proof. Multiplying both sides of the identity $Pf = \lambda f$ by π yields $\pi f = \lambda \pi f$, proving the first claim. Now, assuming that $f = Pf$, let us prove that f is constant. Upon replacing f by its real and imaginary parts if necessary, we may assume that f is real-valued. Denoting by $A := \operatorname{argmin} f$ the set of minimizers of f, we wish to prove that $A = \mathscr{X}$. Fix an arbitrary $x \in A$, and suppose for a contradiction that there is $y \in \mathscr{X} \setminus A$. By irreducibility, we can find $t \geq 0$ such that $P^t(x, y) > 0$. Evaluating the relation $P^t f = f$ at x yields

$$\sum_{z \in \mathscr{X}} P^t(x, z) \, (f(z) - f(x)) \quad = \quad 0.$$

Since $x \in A$, each term in this sum is non-negative, so all terms must actually be zero. This is a contradiction, because $P^t(x, y) > 0$ and $f(y) > f(x)$. □

We are now ready to provide a spectral characterization of the asymptotic decay rate $\lambda_\star(P)$.

Theorem 3 (Spectral radius)**.** *We have $\lambda_\star(P) = \max\{|\lambda| \colon \lambda \in Sp(P) \setminus \{1\}\}$.*

Proof. Recall the key representation (16): writing $\|A\| := \max_{x \in \mathscr{X}} \sum_{y \in \mathscr{X}} |A(x, y)|$, we have

$$\forall t \geq 1, \qquad 2\mathscr{D}_P(t) \quad = \quad \left\| (P - \Pi)^t \right\|.$$

Since $\| \cdot \|$ is a matrix norm, Gelfand's formula asserts that for any $A \in \mathbb{C}^{\mathscr{X} \times \mathscr{X}}$,

$$\|A^t\|^{\frac{1}{t}} \quad \xrightarrow[t \to \infty]{} \quad \rho(A) := \max\{|\lambda| \colon \lambda \in \mathrm{Sp}(A)\}.$$

Thus, the claim boils down to the identity $\rho(P - \Pi) = \lambda_\star(P)$. We will actually show that

$$\mathrm{Sp}(P - \Pi) \quad = \quad \{0\} \cup \mathrm{Sp}(P) \setminus \{1\},$$

which is more than enough. Let us first prove the inclusion $\supseteq$. Clearly, $(0, \mathbf{1})$ is an eigenpair of $P - \Pi$, so $0 \in \mathrm{Sp}(P - \Pi)$. On the other hand, if (λ, f) is an eigenpair of P with $\lambda \neq 1$, then Lemma 13 forces $\pi f = 0$, i.e. $\Pi f = \mathbf{0}$. Thus, (λ, f) is also an eigenpair of $P - \Pi$. Conversely, if (λ, f) is an eigenpair of $P - \Pi$ with $\lambda \neq 0$, then the identity $(P - \Pi) f = \lambda f$ can be left-multiplied by Π to obtain $\Pi f = \mathbf{0}$, so that (λ, f) is also an eigenpair of P. Moreover, we have $\lambda \neq 1$, as otherwise Lemma 13 would imply that f is constant equal to $\pi f = 0$. □

3.2 Diagonalization of reversible kernels

Let P be an irreducible transition kernel on $\mathscr{X}$, with stationary law π. Consider the (complex) Hilbert space $\mathcal{H} = L^2_{\mathbb{C}}(\mathscr{X}, \pi)$ of all functions $f \colon \mathscr{X} \to \mathbb{C}$, with scalar product

$$\langle f, g \rangle \;:=\; \sum_{x \in \mathscr{X}} \pi(x) f(x) \overline{g(x)}.$$

As any operator on $\mathcal{H}$, P admits an adjoint $P^\star$, characterized by the duality relation

$$\forall f, g \in \mathcal{H}, \qquad \langle P^\star f, g \rangle \;=\; \langle f, Pg \rangle.$$

Choosing $f = \delta_x$ and $g = \delta_y$ yields the following explicit expression.

Definition 13 (Adjoint). *The adjoint of P is defined by*

$$\forall (x, y) \in \mathscr{X}^2, \qquad P^\star(x, y) \;=\; \frac{\pi(y) P(y, x)}{\pi(x)}.$$

Note that $P^\star$ is again a transition kernel on $\mathscr{X}$, which is irreducible and with stationary law π. Note also that $P^{\star\star} = P$. The duality $P \leftrightarrow P^\star$ can be interpreted as time reversal: if $\mathbf{X} \sim \mathrm{MC}(\mathscr{X}, P, \pi)$ and $\mathbf{X}^\star \sim \mathrm{MC}(\mathscr{X}, P^\star, \pi)$, then it is easy to check that for all $t \geq 0$,

$$(X^\star_0, \ldots, X^\star_t) \;\stackrel{d}{=}\; (X_t, \ldots, X_0)\,.$$

Definition 14 (Reversibility). *P is reversible if $P^\star = P$, or equivalently,*

$$\forall (x, y) \in \mathscr{X}^2, \qquad \pi(x) P(x, y) \;=\; \pi(y) P(y, x).$$

The above equation, called detailed balance, is satisfied by all random walks on undirected graphs, as well as many other interesting Markov chains. It is much stronger than the stationarity property $\pi P = \pi$ (which can be recovered by summing over all $x \in \mathscr{X}$), and has remarkable consequences on the mixing properties of the associated Markov chain. Indeed, the spectral theorem for self-adjoint operators can then be applied to guarantee the following.

(i) P admits $N = |\mathscr{X}|$ real eigenvalues, which can thus be ordered as follows:

$$1 \;=\; \lambda_1 \;\geq\; \lambda_2 \;\geq\; \cdots \;\geq\; \lambda_N \;\geq\; -1,$$

(ii) There is an orthonormal basis $(\phi_1, \dots, \phi_N)$ of eigenfunctions of P: for all $1 \le i \ne j \le N$,

$$P\phi_i = \lambda_i \phi_i, \qquad \|\phi_i\| = 1, \qquad \langle \phi_i, \phi_j \rangle = 0.$$

Note that, with these notations, we have $\lambda_\star(P) = \max\{\lambda_2, -\lambda_N\}$. We will always choose $\phi_1 = \mathbf{1}$ (this is indeed a unit eigenfunction associated with the eigenvalue $\lambda_1 = 1$). Such a spectral decomposition provides an explicit expression for the distribution of the chain.

Lemma 14 (Eigen-decomposition of reversible chains)**.** *If P is reversible, then*

$$\frac{P^t(x,y)}{\pi(y)} \;=\; 1 + \sum_{i=2}^{N} \lambda_i^t \phi_i(x) \overline{\phi_i(y)},$$

for all $t \in \mathbb{N}$ and all $x, y \in \mathscr{X}$.

Proof. Any function $f\colon \mathscr{X} \to \mathbb{C}$ can be decomposed over the orthonormal basis $(\phi_1, \dots, \phi_N)$:

$$f \;=\; \sum_{i=1}^{N} \langle f, \phi_i \rangle \phi_i$$

Since $\phi_1, \dots \phi_N$ are eigenfunctions of P, we deduce that for all $t \in \mathbb{N}$,

$$P^t f \;=\; \sum_{i=1}^{N} \langle f, \phi_i \rangle \lambda_i^t \phi_i.$$

Choosing $f = \delta_y/\pi(y)$ and evaluating at $x \in \mathscr{X}$, yields exactly the result. □

Remark 13 (Exponential mixing)**.** *The sum in Lemma 14 clearly behaves like $\lambda_\star^t(P)$ as $t \to \infty$, thereby providing yet another proof of the convergence to equilibrium (Theorem 1) and of its geometric refinement (Corollary 1), albeit for reversible chains.*

In order to use the exact expression given in Lemma 14, we need to have explicit access to the eigenvalues and eigenfunctions of P, which is not often the case. Fortunately, the expression can be bounded by a function of $\lambda_\star(P)$ only, yielding the following simple and general estimate on the mixing time of a reversible chain.

Theorem 4 (Mixing times of reversible chains)**.** *If P is reversible, then for all $\varepsilon \in (0,1)$,*

$$\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P) \;\leq\; \mathrm{t}_{\mathrm{REL}}(P) \left\lceil \log\left(\frac{1}{2\varepsilon\sqrt{\pi_\star}}\right)\right\rceil,$$

where we recall that $\pi_\star = \min_{x\in\mathscr{X}} \pi(x)$.

Proof. Fix $t \in \mathbb{N}$ and $x \in \mathscr{X}$. By Lemma 14, the function $y \mapsto \frac{P^t(x,y)}{\pi(y)} - 1$ has squared norm

$$\begin{aligned}
\left\| \frac{P^t(x,\cdot)}{\pi} - 1 \right\|^2 &= \sum_{i=2}^{N} \lambda_i^{2t} |\phi_i(x)|^2 \\
&\leq \lambda_\star^{2t}(P) \sum_{i=2}^{N} |\phi_i(x)|^2 \\
&= \lambda_\star^{2t}(P) \left(\frac{1}{\pi(x)} - 1\right) \\
&\leq \frac{\lambda_\star^{2t}(P)}{\pi(x)}.
\end{aligned}$$

where the third line is obtained by setting $t = 0$ and $y = x$ in Lemma 14. On the other hand, for any probability measure $\mu \in \mathcal{P}(\mathscr{X})$, the Cauchy-Schwarz inequality gives

$$d_{\mathrm{TV}}(\mu,\pi) \;=\; \frac{1}{2}\sum_{x\in\mathscr{X}} \pi(x) \left|\frac{\mu(x)}{\pi(x)} - 1\right| \;\leq\; \frac{1}{2}\left\|\frac{\mu}{\pi} - 1\right\|. \tag{36}$$

Choosing $\mu = P^t(x,\cdot)$ and combining this with the previous estimate, we conclude that

$$\mathscr{D}_P(t) \;\leq\; \frac{\lambda_\star^t(P)}{2\sqrt{\pi_\star}},$$

from which the claim readily follows. □

Remark 14 (L2 bound)**.** *The Cauchy-Schwarz inequality (36) plays a decisive role in the proof, because it connects the probabilistic quantity of interest (total-variation distance) to a much more tractable analytic quantity (Hilbert norm).*

Remark 15 (Relaxation time vs mixing time)**.** *Combining this result with the lower bound (23) (which does not require reversibility), we obtain*

$$\mathrm{t}_{\mathrm{REL}}(P)\left\lceil \log\left(\frac{1}{2\varepsilon}\right)\right\rceil \;\leq\; \mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P) \;\leq\; \mathrm{t}_{\mathrm{REL}}(P)\left\lceil \log\left(\frac{1}{2\varepsilon\sqrt{\pi_\star}}\right)\right\rceil .$$

Thus, for reversible chains, the relaxation time provides an approximation of the mixing time that is precise up to a factor which is only logarithmic in the "size" $\frac{1}{\pi_\star}$. Note that this would not be true without reversibility, as Example 2 shows.

3.3 Wilson's method

The method of distinguishing statistics (Lemma 9) provides a lower bound on the distance to equilibrium based on the first and second moments of an observable $f\colon \mathscr{X} \to \mathbb{C}$ that is expected to behave very abnormally when the chain is far from equilibrium. In a celebrated paper, Wilson obtained remarkably sharp lower bounds for several concrete examples of Markov chains by taking f to be an eigenfunction of P. This fruitful idea is now known as Wilson's method, and summarized in the following lemma. We emphasize that reversibility is not required here. In particular, (λ, f) can be complex.

Lemma 15 (Wilson's method)**.** *If (λ, f) is an eigenpair of P, then*

$$\forall t \in \mathbb{N}, \qquad \mathscr{D}_P(t) \;\geq\; \left(1 + \frac{4V}{(1-|\lambda|^2)|\lambda|^{2t}}\right)^{-1},$$

where V is the worst-case expected quadratic variation of f under P, i.e.

$$V \;:=\; \frac{1}{\|f\|_\infty^2} \max_{x\in\mathscr{X}} \left\{ \sum_{y\in\mathscr{X}} P(x,y)\,|f(y)-f(x)|^2 \right\}.$$

Proof. We may assume that $|\lambda| < 1$, since otherwise the bound is trivial. Upon dividing f by $\|f\|_\infty$ if necessary, we may further assume that $\|f\|_\infty = 1$. Now, fix $x \in \mathscr{X}$ and $t \in \mathbb{N}$. Following Wilson's idea, we estimate $d_{\mathrm{TV}}(P^t(x,\cdot), \pi)$ by applying (the complex version of) Lemma 9 to the eigenfunction f. Letting $\mathbf{X} \sim \mathrm{MC}(\mathscr{X}, P, \delta_x)$, we have

$$\mathbb{E}\left[f(X_{t+1}) | X_0, \ldots, X_t\right] \;=\; (Pf)(X_t) \;= \lambda f(X_t), \tag{37}$$

where we have first used the Markov property and then the fact that (λ, f) is an eigenpair of P. Taking expectations, we find that $\mathbb{E}[f(X_t)] = \lambda^t f(x)$. On the other hand, Lemma 13 (or sending $t \to \infty$) gives $\pi f = 0$. With the notation of Lemma 9, we thus have

$$\delta^2 \;=\; |\lambda|^{2t} |f(x)|^2.$$

Let us now estimate the variance parameter σ^2. By (37), we have

$$\mathbb{E}\left[|f(X_{t+1}) - f(X_t)|^2 \,|X_0, \ldots, X_t\right] \;=\; \mathbb{E}\left[|f(X_{t+1})|^2 \,|X_0, \ldots, X_t\right] + (1 - 2\Re(\lambda))\,|f(X_t)|^2 .$$

But the left-hand side is at most V by definition, so taking expectations gives

$$\begin{aligned} \mathbb{E}\left[|f(X_{t+1})|^2\right] &\le (2\Re(\lambda) - 1)\,\mathbb{E}\left[|f(X_t)|^2\right] + V \\ &\le |\lambda|^2 \mathbb{E}\left[f^2(X_t)\right] + V, \end{aligned}$$

because $2\Re(\lambda) \le 1 + |\lambda|^2$. Subtracting $|\mathbb{E}[f(X_{t+1})]|^2 = |\lambda|^{2t+2} |f(x)|^2$, we obtain

$$\mathrm{Var}\left(f(X_{t+1})\right) \;\le\; |\lambda|^2 \mathrm{Var}\left(f(X_t)\right) + V,$$

from which it follows inductively that

$$\mathrm{Var}\left(f(X_t)\right) \;\le\; \frac{1 - |\lambda|^{2t}}{1 - |\lambda|^2} V \;\le\; \frac{V}{1 - |\lambda|^2}.$$

Taking $t \to \infty$ shows that the same is true under the equilibrium measure π. Thus,

$$\sigma^2 \;\le\; \frac{4V}{1 - |\lambda|^2}.$$

Consequently, the complex version of Lemma 9 guarantees that

$$d_{\mathrm{TV}}\left(P^t(x, \cdot), \pi\right) \;\ge\; \left(1 + \frac{\sigma^2}{\delta^2}\right)^{-1} \;\ge\; \left(1 + \frac{4V}{(1 - |\lambda|^2)|\lambda|^{2t} |f(x)|^2}\right)^{-1},$$

and taking a maximum over $x \in \mathscr{X}$ concludes the proof (recall that $\|f\|_\infty = 1$). □

Remark 16 (Spectral radius)**.** *Choosing λ so that $|\lambda| = \lambda_\star(P)$, we obtain*

$$\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P) \;\ge\; \frac{1}{2} \mathrm{t}_{\mathrm{REL}}(P) \log\left\{\frac{(1 - |\lambda|^2)(1 - \varepsilon)}{4\varepsilon V}\right\},$$

which constitutes a considerable improvement over (23) when $V \ll 1 - |\lambda|^2$.

3.4 Application: limit profile for the cycle

Let us illustrate our spectral techniques by applying them to the lazy random walk on the cycle $\mathscr{X} = \mathbb{Z}/n\mathbb{Z}$, whose transition kernel P_n acts on functions $f\colon \mathscr{X} \to \mathbb{C}$ as follows:

$$\forall x \in \mathscr{X}, \qquad (P_n f)(x) \;=\; \frac{f(x)}{2} + \frac{f(x+1)}{4} + \frac{f(x-1)}{4}.$$

In particular, for any $1 \leq k \leq n$, the function $\phi_k\colon x \mapsto \exp\left(\frac{2i\pi kx}{n}\right)$ is an eigenfunction of P_n with eigenvalue $\lambda_k = \frac{1+\cos\left(\frac{2\pi k}{n}\right)}{2}$. Moreover, for $1 \leq k, \ell \leq n$, we have

$$\frac{1}{n}\sum_{x\in\mathscr{X}} e^{\frac{2i\pi(k-\ell)x}{n}} \;=\; \begin{cases} 1 & \text{if } k = \ell \\ 0 & \text{else,} \end{cases}$$

showing that $(\phi_1, \ldots, \phi_n)$ is an orthonormal basis of $\mathbb{C}^{\mathscr{X}}$. In particular, $\lambda_\star(P_n) = \frac{1+\cos(\frac{2\pi}{n})}{2}$. Using $1 - \cos(h) \sim \frac{h^2}{2}$ and $\ln(1+h) \sim h$ for $h \ll 1$, we obtain the asymptotic estimate

$$\mathrm{t}_{\mathrm{REL}}(P_n) \;\sim\; \frac{n^2}{\pi^2}. \tag{38}$$

Using only this information, Remark 15 already gives $\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P_n) = \Omega(n^2)$ and $\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P_n) = O(n^2 \ln n)$. Of course, we already know from Chapter 2 that the lower bound is sharp. Moreover, cutoff can not occur, because the product condition is not satisfied. This is confirmed by the following refined result, which uses the entire spectral decomposition of P to conclude that the rescaled distance to equilibrium converges to a smoothly decreasing function (hence, not a step function) displayed on Figure 3.4.

Theorem 5 (Limit profile for random walk on the cycle)**.** *For any $\alpha > 0$, we have*

$$\mathscr{d}_{P_n}(\alpha n^2) \;\xrightarrow[n\to\infty]{}\; \Psi(\alpha) := \int_0^1 \left| \sum_{k=1}^{\infty} e^{-\alpha\pi^2 k^2} \cos\left(2\pi k u\right) \right| \mathrm{d}u.$$

In other words, $\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P_n) \sim \Psi^{-1}(\varepsilon) n^2$ as $n \to \infty$, for any fixed $\varepsilon \in (0,1)$.

Proof. By symmetry (Lemma 7), we can choose the initial state to be 0. Our starting point is the following integral representation, which follows from the definition:

$$\mathscr{d}_{P_n}(t) \;=\; \frac{1}{2}\int_0^1 \left|1 - nP_n^t(0, \lfloor un \rfloor)\right| \mathrm{d}u. \tag{39}$$

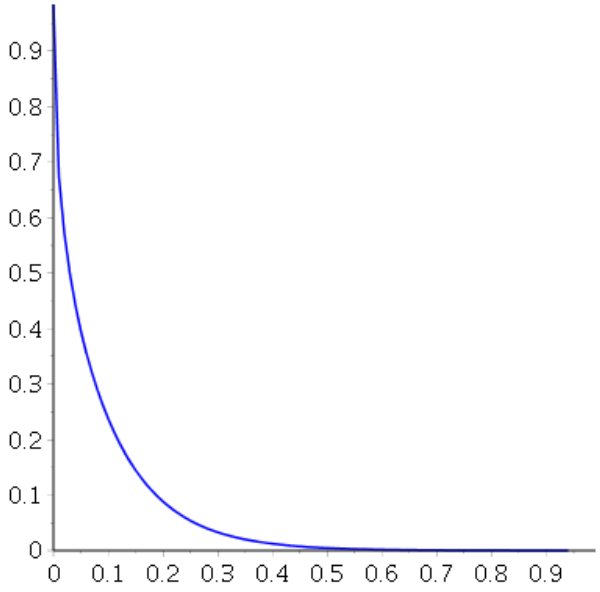

Figure 6: Plot of the limit profile $\Psi\colon (0,\infty) \to (0,1)$ appearing in Theorem 5: the convergence to equilibrium of random walk on the cycle occurs very gradually (no cutoff).

To estimate the integrand, we use the spectral decomposition given in Lemma 14:

$$nP^t(x,y) \;=\; \sum_{k=-\lfloor n/2\rfloor}^{\lceil n/2\rceil-1} \left(\frac{1+\cos\left(\frac{2\pi k}{n}\right)}{2}\right)^t \cos\left(\frac{2\pi k(x-y)}{n}\right).$$

We choose $x=0$, $y=\lfloor un\rfloor$ and $t=t_n=\lceil \alpha n^2\rceil$. As $n\to\infty$, we have for all $k\in\mathbb{Z}$,

$$\left(\frac{1+\cos\left(\frac{2\pi k}{n}\right)}{2}\right)^{t_n} \cos\left(\frac{2\pi k\lfloor un\rfloor}{n}\right) \;\xrightarrow[n\to\infty]{}\; e^{-\alpha\pi^2k^2}\cos\left(2\pi ku\right).$$

On the other hand, since $\frac{1+\cos(a\pi)}{2}\le 1-a^2\le e^{-a^2}$ for all $a\in[-1,1]$, we have the domination

$$\left|\left(\frac{1+\cos\left(\frac{2\pi k}{n}\right)}{2}\right)^{t_n} \cos\left(\frac{2\pi k\lfloor un\rfloor}{n}\right)\right| \;\le\; e^{-4\alpha k^2}.$$

Since the right-hand side is summable in k, we can safely conclude that

$$nP_n^{t_n}(0,\lfloor un\rfloor) \;\xrightarrow[n\to\infty]{}\; \sum_{k\in\mathbb{Z}} e^{-\alpha\pi^2k^2}\cos\left(2\pi ku\right). \tag{40}$$

Moreover, the above domination shows that the left-hand side is bounded uniformly in n et u, so we can pass to the limit in the integral representation (39) to obtain

$$\mathscr{D}_{P_n}(t_n) \quad \xrightarrow[n\to\infty]{} \quad \frac{1}{2}\int_0^1 \left| 1 - \sum_{k\in\mathbb{Z}} e^{-\alpha\pi^2 k^2} \cos(2\pi k u) \right| du.$$

We conclude by noting that the $k = 0$ term is 1, and that the other are even in k. □

Remark 17 (Local CLT)**.** *Our random walk has the representation* $X_t = \xi_1 + \cdots + \xi_t$ mod n*, where* $(\xi_t)_{t\geq 1}$ *are i.i.d., centered and with variance* $\frac{1}{2}$*. Thus, the CLT yields*

$$\mathbb{P}\left(X_{\lceil \alpha n^2 \rceil} \in [an, bn]\right) \quad \xrightarrow[n\to\infty]{} \quad \int_a^b f_\alpha(u) du,$$

for $0 \leq a \leq b \leq 1$*, where* f_α *is the density of* $\mathcal{N}\left(0, \frac{\alpha}{2}\right)$ mod 1*. In view of (40), we have*

$$f_\alpha(u) \quad = \quad \sum_{k\in\mathbb{Z}} e^{-\alpha\pi^2 k^2} \cos(2\pi k u)\,.$$

Therefore, the convergence (40) constitutes a very precise local refinement of the above CLT, where the the macroscopic interval $[an, bn]$ *is replaced by a singleton!*

3.5 Application: cutoff for the hypercube

As a second illustration, let us come back to the random walk on the hypercube $\mathscr{X} = \{0,1\}^n$ (also known as the Ehrenfest model). For any $f\colon \mathscr{X} \to \mathbb{C}$ and $x = (x_1, \ldots, x_n) \in \mathscr{X}$,

$$(P_n f)(x) \quad = \quad \frac{1}{2n}\sum_{i=1}^n \left(f(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n) + f(x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n)\right).$$

In particular, for any fixed set $I \subseteq \{1, \ldots, n\}$, the observable $\phi_I\colon x \mapsto (-1)^{\sum_{i\in I} x_i}$ is an eigenfunction of P_n with eigenvalue $\lambda_I = 1 - \frac{|I|}{n}$. Since $\phi_I\overline{\phi_J} = \phi_{I\Delta J}$, we have

$$\frac{1}{|\mathscr{X}|}\sum_{x\in\mathscr{X}} \phi_I(x)\overline{\phi_J(x)} \quad = \quad \begin{cases} 1 & \text{if } I = J \\ 0 & \text{else,} \end{cases}$$

so that the 2^n eigenfunctions $(f_I)_{I\subseteq[n]}$ form an orthonormal basis of $\mathbb{C}^{\mathscr{X}}$. In particular, the spectral radius is $\lambda_\star(P_n) = 1 - \frac{1}{n}$, which yields the asymptotic estimate

$$\mathrm{t}_{\mathrm{REL}}(P_n) \quad \sim \quad n.$$

Thus, Remark 15 gives $\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P_n) = \Omega(n)$ and $\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P_n) = O(n^2)$. However, both estimates can be considerably refined if we use the entire spectral decomposition of P_n:

Theorem 6 (Cutoff for random walk on the hypercube)**.** *For fixed $\alpha \geq 0$, we have*

$$\mathscr{D}_{P_n}(\alpha n \ln n) \xrightarrow[n\to\infty]{} \begin{cases} 1 & \text{if } \alpha < \frac{1}{2}; \\ 0 & \text{if } \alpha > \frac{1}{2}. \end{cases}$$

In other words, $\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P_n) \sim \frac{n \ln n}{2}$ as $n \to \infty$, for any fixed precision $\varepsilon \in (0,1)$.

Proof. Since the eigenfunctions $(\phi_I)_{I \subseteq [n]}$ take values in $\{-1, 1\}$, the L2 bound (36) yields

$$\begin{aligned} 4d_{\mathrm{TV}}^2\left(P^t(x,\cdot), \pi\right) &\leq \left\| \frac{P^t(x,\cdot)}{\pi} - 1 \right\|^2 \\ &= \sum_{\emptyset \neq I \subseteq [n]} \lambda_I^{2t} |\phi_I(x)|^2 \\ &= \sum_{k=1}^{n} \binom{n}{k} \left(1 - \frac{k}{n}\right)^{2t} \\ &\leq \sum_{k=1}^{n} \binom{n}{k} \exp\left(-\frac{2kt}{n}\right) \\ &\leq \left(1 + e^{-\frac{2t}{n}}\right)^n - 1 \\ &\leq e^{ne^{-\frac{2t}{n}}} - 1. \end{aligned}$$

This suffices to establish the second half of the theorem (case $\alpha > \frac{1}{2}$). For the first half, we apply Wilson's method (Lemma 15) to the eigenpair (λ, f), where $\lambda = 1 - \frac{1}{n}$ and

$$f(x) := \sum_{i=1}^{n} \phi_{\{i\}} = \sum_{i=1}^{n} (1 - 2x_i).$$

Since modifying a coordinate of x changes $f(x)$ by ± 2, we have (taking lazyness into account),

$$\forall x \in \mathscr{X}, \qquad \sum_{y \in \mathscr{X}} P(x,y) \left|f(y) - f(x)\right|^2 = 2,$$

and $\|f_\infty\| = n$. Thus, Lemma 15 applies with $\lambda = 1 - 1/n$ and $V = 2/n^2$, yielding

$$\mathscr{D}_{P_n}(t_n) \geq \left(1 + \frac{4V}{(1 - |\lambda|^2)|\lambda|^{2t_n}}\right)^{-1} = \left(1 + \frac{1}{n^{1-2\alpha+o(1)}}\right)^{-1},$$

when $t_n \sim \alpha n \log n$ with fixed $\alpha \in (0, \infty)$. In particular, $\mathscr{D}_{P_n}(t_n) \to 1$ when $\alpha < 1/2$. □

4 Geometric techniques

We have seen that a transition kernel P is irreducible if and only if its associated diagram G_P is connected, in the sense that it contains a path from any vertex to any other. In light of this, it is natural to suspect an intimate relation between the mixing behavior of P and the geometry of G_P. Formalizing this intuition is precisely the purpose of this chapter.

4.1 Volume, degree, diameter

Any transition kernel P on a finite state space $\mathscr{X}$ naturally induces a directed graph G_P, called the diagram of the chain: its vertex set is $\mathscr{X}$ and its edge set is

$$E \;:=\; \big\{(x,y) \in \mathscr{X}^2 \colon x \neq y \;\&\; P(x,y) > 0\big\}\,.$$

In this graph, a path of length $t \in \mathbb{N}$ is a sequence of $t+1$ vertices $(x_0, \ldots, x_t)$ such that (x_{s-1}, x_s) is an edge for each $1 \leq s \leq t$. The distance from a vertex x to a vertex y is defined as the minimum length of a path that start at x and ends at y. More concisely,

$$\mathrm{dist}(x,y) \;:=\; \min\big\{t \geq 0 \colon P^t(x,y) > 0\big\}\,. \tag{41}$$

The function $\mathrm{dist} \colon \mathscr{X} \times \mathscr{X} \to [0,\infty]$ is not necessarily symmetric, but it always satisfies the two other axioms that define a distance, namely:

(i) Separation: $\mathrm{dist}(x,y) = 0 \iff x = y$ for all $x, y \in \mathscr{X}$;

(ii) Triangle inequality: $\mathrm{dist}(x,z) \leq \mathrm{dist}(x,y) + \mathrm{dist}(y,z)$ for all $x, y, z \in \mathscr{X}$.

We may then consider the (forward) ball of radius $t \in \mathbb{N}$ centered at $x \in \mathscr{X}$:

$$\mathscr{B}(x,t) \;:=\; \{y \in \mathscr{X} \colon \mathrm{dist}(x,y) \leq t\}.$$

Understanding how the volume of these balls grows with t is a natural geometric question. We therefore introduce a function $\mathrm{vol}_P \colon \mathbb{N} \to [0,1]$, called the volume growth of the chain:

$$\mathrm{vol}_P(t) \;:=\; \min_{x \in \mathscr{X}} \pi\left(\mathscr{B}(x,t)\right).$$

A basic observation is that $\mathrm{vol}_P(t)$ has to be large for the chain to be mixed at time t.

Lemma 16 (Volume bound). *We have $\mathscr{D}_P(t) \geq 1 - \mathrm{vol}_P(t)$ for all $t \in \mathbb{N}$.*

Proof. Simply apply the distinguishing event method (Lemma 8) to the pair (x, A), where x is any state that realizes the minimum in the definition of $\mathrm{vol}_P(t)$, and $A = \mathscr{B}(x,t)$. □

We now present two useful consequences of this result, which are easy to apply in practice. Recall that the degree of a vertex $x \in \mathscr{X}$ is the number of vertices at distance 1 from x:

$$\deg(x) \;:=\; \#\{y \in \mathscr{X} : \mathrm{dist}(x,y) = 1\}.$$

Of particular interest is the maximum degree $\deg(P) := \max_{x \in \mathscr{X}} \deg(x)$.

Corollary 2 (Degree bound). *For all $\varepsilon \in (0,1)$, one has*

$$\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P) \;\geq\; \begin{cases} \left\lfloor \frac{\log\left(\frac{1-\varepsilon}{\max \pi}\right)}{\log \deg(P)} \right\rfloor & \text{if } \deg(P) \geq 2 \\ \left\lceil \frac{1-\varepsilon}{\max \pi} \right\rceil - 1 & \text{if } \deg(P) = 1 \end{cases}$$

Proof. For any $x \in \mathscr{X}$ and $t \in \mathbb{N}$, we have $\pi(\mathscr{B}(x,t)) \leq (\max \pi) \times |\mathscr{B}(x,t)|$ and

$$|\mathscr{B}(x,t)| \;\leq\; 1 + \deg(P) + \cdots + (\deg(P))^t .$$

When $\deg(P) \geq 2$, this geometric sum is less than $(\deg(P))^{t+1}$, hence

$$\mathrm{vol}_P(t) \;<\; (\max \pi) \times (\deg(P))^{t+1} .$$

On the other hand, when $\deg(P) = 1$, the geometric sum is $t+1$, so we obtain

$$\mathrm{vol}_P(t) \;\leq\; (\max \pi) \times (t+1).$$

To conclude, take $t := \mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P)$ and note that $\mathrm{vol}_P(t) \geq 1 - \varepsilon$, by Lemma 16. □

Example 3 (Sliding window). *Consider the chain in Example 2: each state has degree 2, so $\deg(P) = 2$. Since π is the uniform law on $\{0,1\}^n$, Corollary 2 yields*

$$\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P) \;\geq\; \left\lfloor n - \log_2\left(\frac{1}{1-\varepsilon}\right) \right\rfloor .$$

This is off by only 1, the exact value of $\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P)$ being obtained by replacing $\lfloor \cdot \rfloor$ with $\lceil \cdot \rceil$.

Example 4 (Riffle shuffle). *One of the most standard methods for shuffling a deck of n cards consists in repeating the following two-step procedure:*

(i) cut the deck into two (possibly unequal) "halves";

(ii) interleave cards from the two halves to produce a new deck.

How many times shall one repeat this procedure for the deck to be well mixed? To formalize this question, let us identify each card with a unique label $i \in [n]$ and represent a deck of cards by a permutation $\sigma \in \mathfrak{S}_n$, where $\sigma(i)$ indicates the label of the $i-th$ top card in the deck. Then, the above procedure transforms σ into a new permutation of the form $\sigma' = \sigma \circ \gamma_I$, where the set $I \subseteq [n]$ indicates the positions to which the top "half" gets relocated, and where γ_I is the permutation that takes values $1, 2, \ldots, |I|$ (in this order) on I and $|I|+1, \ldots, n$ (in this order) on $n \setminus I$. If we choose the subset $I \subseteq [n]$ at random according to some prescribed distribution (e.g., uniform) and repeat this procedure independently at each step, we obtain a well-defined random walk on the symmetric group $\mathfrak{S}_n$, whose kernel is denoted by P_n. Since there at most 2^n possible choices for the subset $I \subseteq [n]$, we have $\deg(P_n) \leq 2^n$, so Corollary 2 yields

$$\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P_n) \;\geq\; \frac{1}{n}\log_2((1-\varepsilon)n!) \sim \log_2 n,$$

for any fixed $\varepsilon \in (0,1)$. This general bound happens to be remarkably sharp: indeed, when I is uniform, the sequence $(P_n)_{n\geq 1}$ is known to exhibit cutoff at time $\frac{3}{2}\log_2(n)$!

Our second application of Lemma 16 involves the radius of the chain, defined as the smallest integer $t \in \mathbb{N}$ such that any two balls of radius t intersect:

$$\mathrm{rad}(P) \;:=\; \min\left\{t \geq 0 \colon \forall (x,y) \in \mathscr{X}^2,\ \mathscr{B}(x,t) \cap \mathscr{B}(y,t) \neq \emptyset\right\}.$$

Corollary 3 (Radius bound). *For any $\varepsilon \in (0, \frac{1}{2})$, we have $\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P) \geq \mathrm{rad}(P)$.*

Proof. For $t = \mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P)$ with $\varepsilon < 1/2$, Lemma 16 yields $\mathrm{vol}_P(t) > \frac{1}{2}$. Since two events of probability more than $1/2$ must intersect, the result follows. □

Example 5 (Sliding window)**.** *Consider the kernel P of Example 2. The balls of radius $n-1$ around the states $(0,\dots,0)$ and $(1,\dots,1)$ are disjoint, because the former consists of all binary words of length n that start with a 0, and the latter those with a 1. Thus,* $\mathrm{rad}(P) \geq n$ *(there is in fact equality), and Corollary 3 gives*

$$\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P) \;\geq\; n,$$

for all $\varepsilon < \frac{1}{2}$. This is in fact the correct answer, as we have seen in Example 2.

Note that when the edge set E is symmetric, we have

$$\mathrm{rad}(P) \;=\; \left\lceil \frac{\mathrm{diam}(P)}{2} \right\rceil,$$

where $\mathrm{diam}(P) := \max_{x,y} \mathrm{dist}(x,y)$ denotes the diameter. In general however, the radius may be significantly smaller than the diameter, as shown in the following example.

Example 6 (Greasy ladder)**.** *On $\mathscr{X} = \{1,2,\dots,n\}$, consider the transition kernel*

$$P(x,y) \;:=\; \begin{cases} \frac{1}{2} & \text{if } y = 1 \text{ or } y = (x+1)\wedge n \\ 0 & \text{else.} \end{cases}$$

The latter represents the evolution of a climber on a greasy ladder, where each step has a chance $1/2$ to result in an abrupt fall. Clearly, $\mathrm{diam}(P) = n-1$*. However,* $\mathrm{rad}(P) = 1$ *because $P(x,1) > 0$ for all $x \in \mathscr{X}$. Thus, Corollary 3 yields the seemingly poor bound*

$$\forall \varepsilon \in \left(0, \frac{1}{2}\right), \quad \mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P) \;\geq\; 1.$$

This is in fact sharp. Indeed, consider the obvious coupling where falls occur simultaneously in both chains: at each step, coalescence occurs with chance at least a half, so Theorem 2 yields $\mathscr{D}_P(t) \leq 2^{-t}$. In particular, $\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P) \leq 2$ for $\varepsilon = 1/4$.

The elementary bounds presented in the previous section can not be expected to be sharp in all situations, because the parameters $\deg(P)$ and $\mathrm{rad}(P)$ only depend on the structure of the graph G_P, and not on the precise transition probabilities. We will now introduce a more sophisticated parameter called the *conductance*, which provides more accurate lower bounds on mixing times by taking the precise transition probabilities into account.

4.2 Conductance

We turn G_P into a weighted graph by defining the weight of a pair $(x, y) \in E$ as follows:

$$\vec{\pi}(x, y) \;:=\; \pi(x)P(x, y). \tag{42}$$

By the Ergodic Theorem, this quantity represents the asymptotic proportion of time that the edge (x, y) is traversed by the chain. Note that the formula (42) extends to a probability measure on $\mathscr{X}^2$ whose first and second marginals are equal to π:

$$\forall x \in \mathscr{X}, \qquad \sum_{y\in\mathscr{X}} \vec{\pi}(x, y) \;=\; \sum_{y\in\mathscr{X}} \vec{\pi}(y, x) \;=\; \pi(x).$$

We will measure the surface of a set $A \subseteq \mathscr{X}$ by the quantity $\vec{\pi}(A \times A^c)$, and compare it with the volume $\pi(A)$. The ratio of those two quantities is called the conductance.

Definition 15 (Conductance)**.** *The conductance of a set $\emptyset \neq A \subseteq \mathscr{X}$ is the ratio*

$$\Phi(A) \;:=\; \frac{\vec{\pi}(A \times A^c)}{\pi(A)}.$$

The conductance of the chain is the quantity

$$\Phi(P) \;:=\; \min\left\{\Phi(A)\colon \emptyset \neq A \subseteq \mathscr{X}, \pi(A) \leq \frac{1}{2}\right\}.$$

The conductance of a set measures the facility for the walk to escape from it. Indeed, letting $\mathbf{X} \sim \mathrm{MC}(\mathscr{X}, P, \pi)$ denote a stationary chain with transition kernel P, we have for any $t \in \mathbb{N}$,

$$\mathbb{P}\left(X_{t+1} \notin A | X_t \in A\right) \;=\; \frac{\mathbb{P}(X_t \in A, X_{t+1} \notin A)}{\mathbb{P}(X_t \in A)} \;=\; \Phi(A).$$

Thus, a set A with small conductance constitutes a "bottleneck" in which the walk is likely to remain "trapped" for a long time. In particular, if that set misses a significant portion of the state space ($\pi(A) \leq 1/2$), then mixing should take long. Here is a rigorous confirmation.

Lemma 17 (Conductance bound)**.** *We always have* $\mathrm{t}_{\mathrm{MIX}}(P) \geq \left\lceil \frac{1}{4\Phi(P)} \right\rceil$.

Proof. Consider a stationary chain $\mathbf{X} \sim \mathrm{MC}(\mathscr{X}, P, \pi)$. Then for any $A \subseteq \mathscr{X}$ and $t \in \mathbb{N}$,

$$\{X_0 \in A, X_t \notin A\} \;\subseteq\; \bigcup_{s=1}^{t} \{X_{s-1} \in A, X_s \notin A\}.$$

Taking probabilities, we deduce that

$$\sum_{x\in A}\pi(x)P^t(x,A^c) \;\leq\; t\vec{\pi}(A\times A^c).$$

On the other hand, we have $P^t(x,A^c)\geq\pi(A^c)-\mathscr{D}_P(t)$ by Lemma 9. Thus,

$$\pi(A^c)-\mathscr{D}_P(t) \;\leq\; t\Phi(A).$$

To conclude, choose a set A realizing the definition of $\Phi(P)$ and set $t=\mathrm{t}_{\text{MIX}}(P)$. □

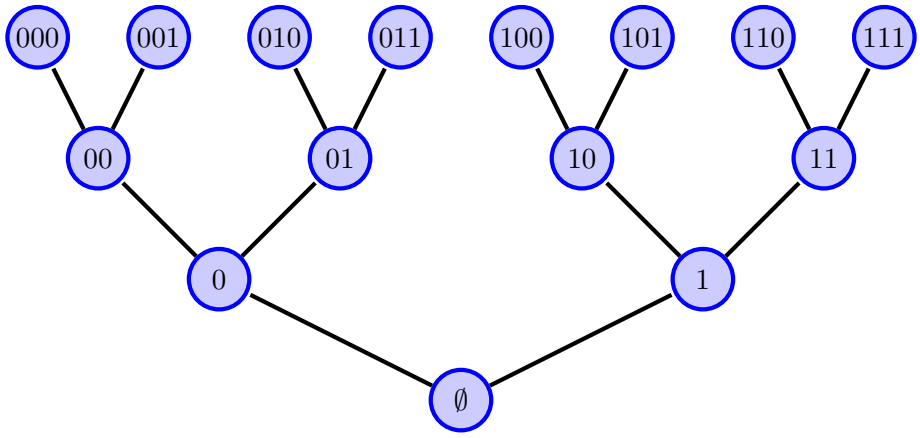

Figure 7: The binary tree of height $n=3$.

Example 7 (Random walk on a binary tree)**.** *Consider the lazy simple random walk on the binary tree of height n (see Figure 7). This is the graph $G=(V,E)$, where*

- *$V=\bigcup_{k=0}^n\{0,1\}^k$ consists of all binary words of length at most n*
- *Two words form an edge if one is obtained from the other by deleting the last letter.*

Consider the "left subtree", i.e. the set $A\subseteq V$ of all words that start with a 0. Note that $\pi(A)=\frac{1}{2}-\frac{1}{2|E|}$, and that $A\times A^c$ contains a single edge. Thus,

$$\Phi(P) \;\leq\; \Phi(A) \;=\; \frac{1}{2|E|-2}.$$

Thus, Lemma 17 gives $\mathrm{t}_{\mathrm{MIX}}(P) \geq \left\lceil \frac{|E|-1}{2} \right\rceil = 2^n - 1$. *This is in fact the correct order of magnitude as* $n \to \infty$, *as can be shown by coupling.*

In order to identify the worst bottleneck of a chain, the following remark may be helpful.

Remark 18 (Connected bottlenecks)**.** *Let* $A \subseteq \mathscr{X}$ *be any set realizing the definition of* $\Phi(P)$, *and suppose that* A *is disconnected, in the sense that it can be partitioned into two proper subsets* A_1, A_2 *with* $(A_1 \times A_2) \cap E = (A_2 \times A_1) \cap E = \emptyset$. *Then, we can write*

$$\Phi(A) \;=\; \frac{\vec{\pi}(A_1 \times A_1^c) + \vec{\pi}(A_2 \times A_2^c)}{\pi(A_1) + \pi(A_2)} \;=\; \frac{\pi(A_1)\Phi(A_1) + \pi(A_2)\Phi(A_2)}{\pi(A_1) + \pi(A_2)},$$

so A_1, A_2 *must also minimize* Φ. *Iterating this procedure eventually produces connected minimizers. Thus, the definition of* $\Phi(P)$ *can safely be restricted to connected sets.*

Remark 19 (Time-reversal)**.** *The measure* $\vec{\pi}$ *is not symmetric unless* P *is reversible. Nevertheless, we may use the fact that* $\vec{\pi}$ *has equal marginals to write, for any* $A \subseteq \mathscr{X}$

$$\begin{aligned} \vec{\pi}(A \times A^c) &= \vec{\pi}(A \times \mathscr{X}) - \vec{\pi}(A \times A) \\ &= \vec{\pi}(\mathscr{X} \times A) - \vec{\pi}(A \times A) \\ &= \vec{\pi}(A^c \times A). \end{aligned}$$

It follows that $\Phi(A)$ *is not modified if* P *is replaced with* $P^\star$ *or with* $(P + P^\star)/2$. *Thus,*

$$\Phi(P) \;=\; \Phi(P^\star) \;=\; \Phi\left(\frac{P + P^\star}{2}\right).$$

Remark 20 (Reversibility on trees)**.** *The identity* $\vec{\pi}(A \times A^c) = \vec{\pi}(A^c \times A)$ *has the following interesting consequence. Let* P *be any transition kernel supported on a tree: removing any edge* (x, y) *partitions* $\mathscr{X}$ *into two connected components* A_x *and* A_y, *and*

$$\vec{\pi}(x, y) \;=\; \vec{\pi}(A_x \times A_y) \;=\; \vec{\pi}(A_y \times A_x) \;=\; \vec{\pi}(y, x).$$

Thus, any transition kernel supported on a tree is reversible.

4.3 Curvature

The geometric methods described so far only provide lower bounds. In the present section, we introduce a fundamental geometric notion that will provide powerful upper bounds on mixing times. Our starting point is the observation that the total-variation distance is "blind" to the geometry of the state space: we have $d_{\mathrm{TV}}(\delta_x, \delta_y) = 1$ for any $x \neq y \in \mathscr{X}$, regardless of how close x is to y. A simple but far-reaching idea consists in replacing total-variation with the following geometric quantity.

Definition 16 (Wasserstein distance)**.** *Let μ, ν be two probability measures on $\mathscr{X}$. The Wasserstein (or transportation) distance from μ to ν is the quantity*

$$\mathcal{W}(\mu, \nu) \quad := \quad \inf_{X \sim \mu, Y \sim \nu} \mathbb{E}[\mathrm{dist}(X, Y)],$$

where the infimum is take over all possible couplings (X, Y) of μ and ν.

Let us make a couple of important comments before proceeding further.

Remark 21 (Dirac masses)**.** *For any $x, y \in \mathscr{X}$, we trivially have*

$$\mathcal{W}(\delta_x, \delta_y) \quad = \quad \mathrm{dist}(x, y).$$

Thus, $\mathcal{W}(\cdot, \cdot)$ extends the function $\mathrm{dist}(\cdot, \cdot)$ from points to probability measures.

Remark 22 (Optimal coupling)**.** *$\mathcal{W}(\mu, \nu)$ is the infimum of the continuous functional*

$$p \quad \mapsto \quad \sum_{x, y \in \mathscr{X}} \mathrm{dist}(x, y) p(x, y)$$

over the compact (and convex) set of all coupling distributions of μ and ν:

$$\mathcal{C}(\mu, \nu) \quad := \quad \left\{ p \in [0,1]^{\mathscr{X} \times \mathscr{X}} : \forall x \in \mathscr{X}, \sum_{z \in \mathscr{X}} p(x, z) = \mu(x), \sum_{z \in \mathscr{X}} p(z, x) = \nu(x) \right\}.$$

In particular, this infimum is attained. Thus, there is always a coupling (X, Y) that attains the minimum in Definition 16: we call it an optimal coupling from μ to ν.

The function $\mathcal{W}$ is not symmetric in general, because dist is not. However, this is the only obstruction for $\mathcal{W}$ to be a nice distance on $\mathcal{P}(\mathscr{X})$, as the next lemma shows.

Lemma 18 (Properties)**.** *The Wasserstein distance $\mathcal{W}$ is convex and satisfies the separation axiom and the triangle inequality. It is symmetric if and only if* dist *is.*

Proof of convexity. Fix $\mu, \mu', \nu, \nu' \in \mathcal{P}(\mathscr{X})$ and $\theta \in [0,1]$. Let p be the law of an optimal coupling form μ to μ', and let q be the law of an optimal coupling from ν to ν'. Then $r := \theta p + (1-\theta)q$ is the law of a coupling from $\theta\mu + (1-\theta)\nu$ to $\theta\mu' + (1-\theta)\nu'$, so

$$\begin{aligned}\mathcal{W}\left(\theta\mu + (1-\theta)\nu, \theta\mu' + (1-\theta)\nu'\right) &\leq \sum_{(x,y)\in\mathscr{X}} r(x,y)\mathrm{dist}(x,y) \\ &= \sum_{(x,y)\in\mathscr{X}} \left(\theta p(x,y) + (1-\theta)q(x,y)\right)\mathrm{dist}(x,y) \\ &= \theta\mathcal{W}\left(\mu,\mu'\right) + (1-\theta)\mathcal{W}\left(\nu,\nu'\right).\end{aligned}$$

This proves that $\mathcal{W}$ is convex on $\mathcal{P}(\mathscr{X}) \times \mathcal{P}(\mathscr{X})$. □

Proof of the separation axiom. Let $\mu, \nu \in \mathcal{P}(\mathscr{X})$ be such that $\mathcal{W}(\mu,\nu) = 0$. By Remark 22, we can find a coupling (X,Y) of μ and ν such that $\mathbb{E}[\mathrm{dist}(X,Y)] = 0$. This forces $\mathrm{dist}(X,Y) = 0$ a.-s., because $\mathrm{dist}(\cdot,\cdot)$ is non-negative. Since $\mathrm{dist}(\cdot,\cdot)$ moreover satisfies the separation axiom, we deduce that $X = Y$ a.-s., hence in distribution. Thus, $\mu = \nu$. □

Proof of the triangle inequality. Fix $\lambda, \mu, \nu \in \mathcal{P}(\mathscr{X})$. Write p (resp. q) for the law of an optimal coupling from λ to μ (resp. μ to ν). Consider a random triple (X,Y,Z) with law

$$\forall (x,y,z) \in \mathscr{X}^3, \quad \mathbb{P}(X=x, Y=y, Z=z) \;=\; \frac{p(x,y)q(y,z)}{\mu(y)},$$

this ratio being interpreted as 0 if the denominator (hence also the numerator) is zero. Summing over all $z \in \mathscr{X}$ shows that (X,Y) has law p, and summing over all $x \in \mathscr{X}$ shows that (Y,Z) has law q. In particular, (X,Z) is a coupling of λ and ν, so we have

$$\begin{aligned}\mathcal{W}(\lambda,\nu) &\leq \mathbb{E}[\mathrm{dist}(X,Z)] \\ &\leq \mathbb{E}[\mathrm{dist}(X,Y) + \mathrm{dist}(Y,Z)] \\ &= \mathcal{W}(\lambda,\mu) + \mathcal{W}(\mu,\nu),\end{aligned}$$

where we have used the triangle inequality for $\mathrm{dist}(\cdot,\cdot)$, and the optimality of p and q. □

Proof of symmetry. It is clear from the Definition 16 that $\mathcal{W}(\cdot,\cdot)$ is symmetric whenever $\mathrm{dist}(\cdot,\cdot)$ is. The converse readily follows from Remark 21. □

Remark 23 (Robustness). *The above proofs remain valid for any function* $\mathrm{dist}\colon \mathscr{X}^2 \to \mathbb{R}_+$ *satisfying the separation axiom and the triangle inequality. Thus, the Wasserstein distance is a very general tool that "lifts" any distance on* $\mathscr{X}$ *to a distance on* $\mathcal{P}(\mathscr{X})$. *The choice* $\mathrm{dist}(x,y) := \mathbf{1}_{x \neq y}$ *gives rise to the total-variation distance, by Remark 11.*

We now show that the Wasserstein distance controls the total variation distance.

Lemma 19 (Wasserstein bound). *For any* $\mu, \nu \in \mathcal{P}(\mathscr{X})$, *we have*

$$d_{\mathrm{TV}}(\mu,\nu) \;\leq\; \mathcal{W}(\mu,\nu).$$

Proof. The inequality $\mathbf{1}_{x\neq y} \leq \mathrm{dist}(x,y)$ trivially holds for all $(x,y) \in \mathscr{X}^2$. Integrating this against the law of an optimal coupling from μ to ν concludes the proof. □

Thus, any upper bound on the Wasserstein distance is also an upper bound on the total-variation distance. The interest of the Wasserstein distance is that it can be efficiently controlled by exploiting the geometry of the state space, as we will now see. The curvature of a Markov chain measures the amount by which Wasserstein distances are contracted under the action of the underlying transition kernel P.

Definition 17 (Curvature). *The curvature* $\kappa(P)$ *is the largest* $\kappa \in \mathbb{R}$ *such that*

$$\forall \mu, \nu \in \mathcal{P}(\mathscr{X}), \qquad \mathcal{W}(\mu P, \nu P) \;\leq\; e^{-\kappa}\mathcal{W}(\mu,\nu).$$

This global definition seems far too delicate for practical use. Fortunately, a pleasant feature of curvature is that it admits a simple, local characterization.

Lemma 20 (Local characterization of curvature). *We have*

$$e^{-\kappa(P)} \;=\; \max_{(x,y)\in E} \mathcal{W}\left(P(x,\cdot), P(y,\cdot)\right).$$

Proof. Setting $\rho := \max_{(x,y)\in E} \mathcal{W}\left(P(x,\cdot), P(y,\cdot)\right)$, we will establish the inequality

$$\forall \mu, \nu \in \mathcal{P}(\mathscr{X}), \qquad \mathcal{W}(\mu P, \nu P) \;\leq\; \rho\mathcal{W}(\mu,\nu). \tag{43}$$

Fix $x, y \in \mathscr{X}$, and let $(\sigma_0, \ldots, \sigma_t)$ be a shortest path from x to y, i.e.

$$t = \mathrm{dist}(x,y), \quad \sigma_0 = x, \quad \sigma_t = y, \quad \text{and} \quad (\sigma_{s-1}, \sigma_s) \in E \text{ for } 1 \leq s \leq t.$$

Using the triangle inequality for $\mathcal{W}$ (Lemma 18) and the definition of ρ, we have

$$\mathcal{W}\left(P(x,\cdot),P(y,\cdot)\right) \;\leq\; \sum_{s=1}^{t}\mathcal{W}\left(P(x_{s-1},\cdot),P(x_s,\cdot)\right) \;\leq\; \rho t \;=\; \rho\mathrm{dist}(x,y). \tag{44}$$

This establishes (43) in the special case $(\mu,\nu)=(\delta_x,\delta_y)$. For the general case, let p be the law of an optimal coupling from μ to ν, and observe that

$$(\mu P,\nu P) \;=\; \sum_{(x,y)\in\mathscr{X}^2} p(x,y)\left(P(x,\cdot),P(y,\cdot)\right).$$

Since $\mathcal{W}$ is convex (Lemma 18), we immediately deduce that

$$\begin{aligned}\mathcal{W}\left(\mu P,\nu P\right) &\leq \sum_{(x,y)\in\mathscr{X}^2} p(x,y)\mathcal{W}\left(P(x,\cdot),P(y,\cdot)\right)\\ &\leq \rho\sum_{(x,y)\in\mathscr{X}^2} p(x,y)\mathrm{dist}(x,y)\\ &= \rho\mathcal{W}(\mu,\nu),\end{aligned}$$

where the second line uses (44) and the third the optimality of p. Thus, (43) is established. Conversely, note that (43) is an equality when $(\mu,\nu)=(\delta_x,\delta_y)$ with $(x,y)\in E$ achieving the maximum in the definition of ρ. Thus, ρ is in fact the smallest constant for which (43) holds. Comparing this with Definition 17, we conclude that $\rho=e^{-\kappa(P)}$. □

The interest of curvature is contained in the following result.

Theorem 7 (Curvature bound)**.** *If $\kappa(P)>0$, then*

$$\begin{aligned}\mathrm{t}_{\mathrm{REL}}(P) &\leq \frac{1}{\kappa(P)},\\ \mathrm{t}^{(\varepsilon)}_{\mathrm{MIX}}(P) &\leq \frac{1}{\kappa(P)}\log\left(\frac{\mathrm{diam}(P)}{\varepsilon}\right).\end{aligned}$$

Proof. Using the definition of $\kappa(P)$ and an immediate induction over $t\in\mathbb{N}$, we have

$$\forall\mu,\nu\in\mathcal{P}(\mathscr{X}),\qquad \mathcal{W}(\mu P^t,\nu P^t) \;\leq\; \mathcal{W}(\mu,\nu)e^{-\kappa(P)t}.$$

Combining this with the crude bound $\mathcal{W}(\cdot,\cdot)\leq\mathrm{diam}(P)$ and Lemma 19, we obtain

$$d_{\mathrm{TV}}\left(\mu P^t,\nu P^t\right) \;\leq\; e^{-\kappa(P)t}\mathrm{diam}(P).$$

Finally, choosing $\nu = \pi, \mu = \delta_x$ and maximizing over $x \in \mathscr{X}$ yields

$$\mathscr{D}_P(t) \;\leq\; \mathrm{diam}(P) e^{-\kappa(P)t},$$

for all $t \in \mathbb{N}$. The first claim is obtained by sending $t \to \infty$ (recall that $\mathscr{D}_P^{1/t}(t) \to e^{-1/\mathrm{t_{REL}}(P)}$), and the second by choosing $t = \lceil \frac{1}{\kappa(P)} \log \frac{\mathrm{diam}(P)}{\varepsilon} \rceil$. □

Example 8 (Hypercube)**.** *Consider the lazy random walk on the $n-$dimensional hypercube. Fix two neighboring states x, y, and consider the coupling (X, Y) of $P(x, \cdot)$ and $P(y, \cdot)$ that updates the same coordinate using the same Bernoulli variable. Then,*

$$\mathcal{W}\left(P(x,\cdot), P(y,\cdot)\right) \;\leq\; \mathbb{E}[\mathrm{dist}(X,Y)] \;=\; 1 - \frac{1}{n}.$$

Since this holds for all $(x, y) \in E$, we deduce that

$$\kappa(P) \;\geq\; -\log\left(1 - \frac{1}{n}\right) \;\geq\; \frac{1}{n}.$$

Thus, Theorem 7 gives $\mathrm{t_{REL}}(P) \leq n$ and $\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P) \leq n \log\left(\frac{n}{\varepsilon}\right)$. A comparison with the results of Section 3.5 shows that those estimates are remarkably sharp. Interestingly, the bound $\kappa(P) \leq \frac{1}{\mathrm{t_{REL}}(P)} = -\log(1 - \frac{1}{n})$ shows that the first inequality in (45) is an equality.

4.4 Application: phase transition in the Curie-Weiss model

In this section, we demonstrate the strength of the above methods by establishing a dynamical phase transition for one of the most fundamental statistical physics models: the mean-field Ising ferromagnet. The latter describes the evolution of n particles (called "spins"), each being in one of two possible states ("plus" or "minus"). Each particle has a tendency to align its state with those of the other particles, and the strength of this interaction is controlled by a parameter $\beta \geq 0$ (the "inverse temperature"). The precise model is as follows.

The system can be represented by a vector $x = (x_1, \ldots, x_n) \in \{-1, +1\}^n$, where $x_i = +1$ (resp. $x_i = -1$) indicates that the $i-$th particle is in the "plus" (resp. "minus") state. At each step, the vector x is randomly modified as follows: a coordinate $i \in [n]$ is selected uniformly at random, and its current value x_i is replaced by $+1$ or -1 with respective

probabilities $\psi(+s)$ and $\psi(-s)$, where

$$s := \frac{\beta}{n} \sum_{j \in [n] \setminus \{i\}} x_j \quad \text{and} \quad \psi(s) := \frac{e^s}{e^s + e^{-s}}.$$

Note that $\psi(s) + \psi(-s) = 1$ for any $s \in \mathbb{R}$, as required. Note also that $\psi(s)$ increases from 0 to 1 as s ranges from $-\infty$ to $+\infty$. Thus, the new state of the $i-$th particle is likely to be "plus" if s is a large positive number, and "minus" if s is a large negative number. Formally, we have defined a Markov chain on $\mathscr{X} = \{-1, +1\}^n$ with transition kernel

$$P_n(x, y) := \begin{cases} \frac{1}{n} \sum_{i=1}^n \psi\left(\frac{\beta}{n} x_i \sum_{j \in [n] \setminus \{i\}} x_j\right) & \text{if } y = x \\ \frac{1}{n} \psi\left(-\frac{\beta}{n} x_i \sum_{j \in [n] \setminus \{i\}} x_j\right) & \text{if } y = (x_1, \ldots, x_{i-1}, -x_i, x_{i+1}, \ldots, x_n) \\ 0 & \text{else.} \end{cases}$$

The fact that $\psi > 0$ ensures that this transition kernel is ergodic. Moreover, it is easily checked to be reversible with respect to the measure

$$\pi(x) := \frac{1}{Z(\beta)} \exp\left[\frac{\beta}{2n}\left(\sum_{i=1}^n x_i\right)^2\right],$$

where $Z(\beta)$ denotes the appropriate normalizing constant. How does the mixing time of this chain depend on the interaction parameter β? In the easy case $\beta = 0$ (no interaction), we recover the random walk on the hypercube, which has mixing time $\mathrm{t}_{\text{MIX}}(P_n) = \Theta(n \log n)$. On the other hand, in the limit where $\beta \to +\infty$ (strong interaction), the selected particle will systematically adopt the majority state, so the chain will need an infinite amount of time to move from $(-1, \ldots, -1)$ to $(+1, \ldots, +1)$. For "intermediate" values of β, it is then tempting to believe that the mixing time will simply interpolate between those two extreme situations, in a gradual way. In fact, the mixing time changes dramatically from $O(n \log n)$ (fast-mixing regime) to $\exp(\Omega(n))$ (slow-mixing regime) as β passes the critical value 1.

Theorem 8 (Dynamic phase transition for the mean-field Ising model)**.**

1. *For any fixed $\beta < 1$, we have $\mathrm{t}_{\text{MIX}}(P_n) \leq (1 + o(1)) \frac{n \log n}{1 - \beta}$ as $n \to \infty$ (fast mixing).*
2. *For any fixed $\beta > 1$, we have $\mathrm{t}_{\text{MIX}}(P_n) = \exp(\Omega(n))$ (exponentially slow mixing).*

Proof of fast mixing when $\beta < 1$. In light of Theorem 7, it suffices to prove that

$$\kappa(P_n) \geq \frac{1 - \beta}{n},$$

which we now do. Let I and U be independent with $I \sim \mathrm{Unif}(\{1,\dots,n\})$ and $U \sim \mathrm{Unif}([0,1])$. Starting from a fixed state $x = (x_1,\dots,x_n) \in \mathscr{X}$, one can construct a random state $X = (X_1,\dots,X_n)$ with law $P(x,\cdot)$ by setting for each $i \in [n]$,

$$X_i \;:=\; \begin{cases} x_i & \text{if } I \neq i \\ +1 & \text{if } I = i \text{ and } U \leq \psi\left(\frac{\beta}{n}\sum_{j\in[n]\setminus\{i\}} x_j\right) \\ -1 & \text{if } I = i \text{ and } U > \psi\left(\frac{\beta}{n}\sum_{j\in[n]\setminus\{i\}} x_j\right). \end{cases}$$

Now, consider a state $y \in \mathscr{X}$ which differs from x by a single coordinate, say $x_k = -1$ and $y_k = +1$. Then, the coupling (X,Y) of $P(x,\cdot), P(y,\cdot)$ that uses the same pair (U,I) gives

$$\mathrm{dist}(X,Y) \;=\; \begin{cases} 0 & \text{if } I = k \\ 2 & \text{if } I \neq k \text{ and } \psi\left(\frac{\beta}{n}\sum_{j\in[n]\setminus\{I\}} x_j\right) \leq U < \psi\left(\frac{\beta}{n}\sum_{j\in[n]\setminus\{I\}} y_j\right) \\ 1 & \text{else.} \end{cases}$$

But $\sum_{j\neq I} y_j - \sum_{j\neq I} x_j \leq 2$ and $\|\psi'\|_\infty \leq \frac{1}{2}$, so $\psi\left(\frac{\beta}{n}\sum_{j\in[n]\setminus\{I\}} y_j\right) - \psi\left(\frac{\beta}{n}\sum_{j\in[n]\setminus\{I\}} x_j\right) \leq \frac{\beta}{n}$. Thus, the second case occurs with probability at most β/n, and we deduce that

$$\mathbb{E}[\mathrm{dist}(X,Y)] \;\leq\; \left(1-\frac{1}{n}\right)\left(1+\frac{\beta}{n}\right) \;\leq\; e^{\frac{\beta-1}{n}}.$$

This shows that $\kappa(P_n) \geq \frac{1-\beta}{n}$, as desired. □

Proof of slow mixing when $\beta > 1$. Let us consider the event of negative magnetization:

$$A \;:=\; \left\{x \in \mathscr{X} : \sum_{i=1}^n x_i < 0\right\}.$$

The symmetry property $\pi(x) = \pi(-x)$ ensures that $\pi(A) \leq \frac{1}{2}$, so that $\Phi(P) \leq \Phi(A)$. By the conductance bound (Lemma 17), we only have to show that $\Phi(A) = \exp(-\Omega(n))$. For $0 \leq k \leq n$, let A_k consist of all configurations with k "plus" and $n-k$ "minus" states:

$$A_k \;:=\; \left\{x \in \mathscr{X} : \sum_{i=1}^n x_i = 2k - n\right\}.$$

Since at most one coordinate is modified at each step, the only way for the chain to jump from A to A^c is to actually jump from $A_{\lceil n/2\rceil - 1}$ to $A_{\lceil n/2\rceil}$. Thus,

$$\vec{\pi}(A \times A^c) \;=\; \vec{\pi}(A_{\lceil n/2\rceil-1} \times A_{\lceil n/2\rceil}) \;\leq\; \pi(A_{\lceil n/2\rceil}),$$

where the inequality follows from the fact that the second marginal of $\vec{\pi}$ is π. On the other hand, we have $\pi(A) \geq \max_{k<\lceil n/2 \rceil} \pi(A_k)$ and for $0 \leq k \leq n$, $\pi(A_k) = a_k/Z(\beta)$, where

$$a_k \;:=\; \binom{n}{k} \exp\left[\frac{\beta}{2n}(2k-n)^2\right].$$

Consequently, $\phi(A) \leq \min_{k<\lceil n/2 \rceil} \frac{a_{\lceil n/2 \rceil}}{a_k}$. To see that this ratio is exponentially small in n, fix $\theta \in (0,1)$ and observe that when $k = k(n) \sim \theta n$ as $n \to \infty$, we have

$$\frac{1}{n}\log a_{k(n)} \;\xrightarrow[n\to\infty]{}\; f(\theta) := \frac{\beta}{2}(2\theta-1)^2 - \theta\log\theta - (1-\theta)\log(1-\theta).$$

Thus, our task boils down to showing the existence of $\theta < \frac{1}{2}$ so that $f(\theta) > f(1/2)$. But

$$\begin{aligned} f'(\theta) &= 2\beta(2\theta-1) + \log(1-\theta) - \log\theta \\ f''(\theta) &= 4\beta - \frac{1}{\theta(1-\theta)}. \end{aligned}$$

Thus, $f'(1/2) = 0$ and $f''(1/2) = 4(\beta-1)$, so $f(\frac{1}{2})$ is a strict local minimum when $\beta > 1$. $\square$

4.5 Carne-Varopoulos bound

The volume bound (Lemma 16) and its useful consequences (Corollaries 2 and 3) relied on a crude observation: after t steps, the chain is *necessarily* at distance at most t from its starting point. In many cases however, this best-case scenario is rather optimistic, compared to the *typical* displacement of the chain. For example, a simple random walk $\mathbf{X} = (X_t)_{t\geq 0}$ on $\mathbb{Z}$ has asymptotic speed zero by the strong law of large number, and it is the statistical fluctuations that really drive the motion, as quantified by the Central Limit Theorem:

$$\frac{\mathrm{dist}(X_0, X_t)}{\sqrt{t}} \;\xrightarrow[t\to\infty]{d}\; |Z|, \quad \text{where} \quad Z \sim \mathcal{N}(0,1).$$

Consequently, after t steps, the walk typically lies at distance $\Theta(\sqrt{t})$ from its starting point, rather than t as we used in our volume bound. In light of this, it is natural to hope for a quadratic improvement over the diameter lower bound (Corollary 3) for "diffusive" chains. Note that this intuition is correct for the random walk on the $n-$cycle, for which we have seen that $\mathrm{t}_{\text{MIX}}(P_n) \asymp n^2 \asymp \mathrm{diam}(P_n)^2$. With those preliminary observations in mind, let us now state a remarkable inequality which compares the transition kernel of *any* reversible chain to that of the simple random walk on $\mathbb{Z}$.

Theorem 9 (Carne-Varopoulos estimate)**.** *For any reversible kernel P, we have*

$$P^t(x,y) \;\leq\; \sqrt{\frac{\pi(y)}{\pi(x)}}\,\mathbb{P}\left(|X_t| \geq \mathrm{dist}(x,y)\right),$$

where $\mathbf{X} = (X_t)_{t\geq 0}$ denotes the simple random walk on $\mathbb{Z}$. In particular,

$$P^t(x,y) \;\leq\; 2\sqrt{\frac{\pi(y)}{\pi(x)}}\exp\left\{-\frac{(\mathrm{dist}(x,y))^2}{2t}\right\},$$

for all $x, y \in \mathscr{X}$ and $t \in \mathbb{N}$.

Proof. The proof uses the Chebychev polynomials $(q_k)_{k\geq 0}$, which are defined by the recursion

$$q_{k+1}(z) \;:=\; 2zq_k(z) - q_{k-1}(z) \qquad (k \geq 1),$$

with initial conditions $q_0(z) = 1$ and $q_1(z) = z$. The trigonometric identity

$$2\cos(k\theta)\cos(\theta) \;=\; \cos\left((k+1)\theta\right) + \cos\left((k-1)\theta\right),$$

shows that $q_k(\cos\theta) = \cos(k\theta)$ for all $\theta \in \mathbb{R}$ and $k \in \mathbb{N}$. Now, observe that for all $t \in \mathbb{N}$,

$$(\cos\theta)^t \;=\; \left(\frac{e^{i\theta}+e^{-i\theta}}{2}\right)^t \;=\; \mathbb{E}\left[e^{i\theta X_t}\right] \;=\; \mathbb{E}\left[\cos\left(\theta X_t\right)\right] \;=\; \sum_{k=0}^{t}\mathbb{P}\left(|X_t| = k\right)q_k(\cos\theta).$$

Since this is true for all $\theta \in \mathbb{R}$, we must have the polynomial identity

$$z^t \;=\; \sum_{k=0}^{t}\mathbb{P}\left(|X_t| = k\right)q_k(z).$$

Let us apply this polynomial identity to the matrix P, and evaluate the $(x,y)-$entry:

$$\begin{aligned} P^t(x,y) \;&=\; \sum_{k=0}^{t}\mathbb{P}\left(|X_t| = k\right)q_k(P)(x,y) \\ &=\; \sum_{k=\mathrm{dist}(x,y)}^{t}\mathbb{P}\left(|X_t| = k\right)q_k(P)(x,y), \end{aligned}$$

where we have observed that $I(x,y) = P(x,y) = \cdots = P^k(x,y) = 0$ for $k < \mathrm{dist}(x,y)$, so that $q_k(P)(x,y) = 0$ (since q_k has degree k). To conclude, it remains to show that

$$q_k(P)(x,y) \;\leq\; \sqrt{\frac{\pi(y)}{\pi(x)}}, \tag{45}$$

for all $k \geq 0$, and this is where we use reversibility: $q_k(P)$ is a self-adjoint operator with spectrum $\{q_k(\lambda) \colon \lambda \in \mathrm{Sp}(P)\} \subseteq q_k([-1,1]) \subseteq [-1,1]$, where the last inclusion follows from the identity $q_k(\cos\theta) = \cos(k\theta)$. Thus, $q_k(P)$ is a contraction, which means that

$$|\langle q_k(P) f, g\rangle| \leq \|f\|\|g\|, \tag{46}$$

for all observables $f, g \colon \mathscr{X} \to \mathbb{C}$. Taking $f = \delta_y$ and $g = \delta_x$ yields exactly (45). The second claim follows from a classical application of Markov's inequality: for $d, \lambda > 0$,

$$\mathbb{P}\left(X_t \geq d\right) \;=\; \mathbb{P}\left(e^{\lambda X_t} \geq e^{\lambda d}\right) \;\leq\; e^{-\lambda d}\mathbb{E}\left[e^{\lambda X_t}\right] \;=\; e^{-\lambda d}\left(\frac{e^{\lambda} + e^{-\lambda}}{2}\right)^t \;\leq\; e^{-\lambda d + \frac{\lambda^2 t}{2}}.$$

The right-hand side is minimized for $\lambda = \frac{d}{t}$, in which case it is equal to $e^{-\frac{d^2}{2t}}$. This of course also applies to $-X_t$, and combining the two estimates concludes the proof. □

Corollary 4 (Diffusive bound)**.** *For lazy simple random walk on any $N-$vertex graph,*

$$\mathrm{t}_{\mathrm{MIX}}^{(\varepsilon)}(P) \;\geq\; \frac{(\mathrm{diam}(P))^2}{16 \ln N}.$$

for all $\varepsilon \in (0, \frac{1}{2})$, provided N is large enough.

Proof. Set $d = \lceil \mathrm{diam}(P)/2 \rceil$, so that $\mathrm{diam}(P) > 2(d-1)$: this means that we can find two disjoint balls of radius $d-1$. Thus, there is $x \in \mathscr{X}$ such that $A = \mathscr{B}(x, d-1)$ satisfies

$$\pi(A) \;\leq\; \frac{1}{2}$$

But the elements of A^c are at distance $d \geq \mathrm{diam}(P)/2$ from x, so Theorem 9 ensures that

$$P^t(x, A^c) \;\leq\; 2N^{\frac{3}{2}} e^{-\frac{\mathrm{diam}^2(P)}{8t}},$$

where we have used the crude estimates $|A^c| \leq N$ and $\frac{\pi(y)}{\pi(x)} = \frac{\deg(y)}{\deg(x)} \leq N$. We conclude that

$$\mathscr{D}_P(t) \;\geq\; \frac{1}{2} - 2n^{\frac{3}{2}} e^{-\frac{\mathrm{diam}^2(P)}{8t}}.$$

Thus, as long as $t \leq \frac{(\mathrm{diam} P)^2}{16 \ln N}$, we have $\mathscr{D}_P(t) \geq \frac{1}{2} - \frac{2}{\sqrt{N}}$, concluding the proof. □

5 Variational techniques

For an ergodic transition kernel P with stationary law π on a state space $\mathscr{X}$, the convergence to equilibrium $\mathscr{D}_P(t) \to 0$ as $t \to \infty$ can be equivalently formulated as follows:

$$\forall x \in \mathscr{X}, \quad (P^t f)(x) \xrightarrow[t\to\infty]{} \pi f,$$

for all $f\colon \mathscr{X} \to \mathbb{R}$. In words, observables become constant under the repeated action of P. Equivalently, the variance $\mathrm{Var}(f) = \pi(f^2) - \pi^2(f)$ decays under the repeated action of P:

$$\mathrm{Var}(P^t f) \xrightarrow[t\to\infty]{} 0. \tag{47}$$

This naturally raises the following two questions:

1. At what speed does the convergence (47) take place?
2. What are the consequences in terms of mixing times?

To answer those questions, we introduce a fundamental object: the *Dirichlet form*.

5.1 Dirichlet form and Poincaré constant

The Dirichlet form is a quadratic form on the Hilbert space $L^2(\mathscr{X}, \pi)$ that measures the expected quadratic variation of observables under a transition of the stationary chain.

Definition 18 (Dirichlet form)**.** *The Dirichlet form is the quadratic form defined by*

$$\begin{aligned} \mathscr{E}_P(f) &= \frac{1}{2}\mathbb{E}_\pi\left[(f(X_1) - f(X_0))^2\right] \\ &= \frac{1}{2}\sum_{x,y\in\mathscr{X}} \vec{\pi}(x,y)\,(f(y) - f(x))^2 \\ &= \langle (\mathbb{I} - P)f, f\rangle \end{aligned}$$

for any observable $f\colon \mathscr{X} \to \mathbb{R}$, where $\mathbb{E}_\pi$ denotes expectation under $\mathrm{MC}(\mathscr{X}, P, \pi)$.

Remark 24 (Rank-one case)**.** *It is instructive to consider the "ideal chain" $P = \Pi$*

defined in (15), which mixes exactly in a single step. Since $\vec{\pi}(x,y) = \pi(x)\pi(y)$, we have

$$\begin{aligned} \mathscr{E}_{\Pi}(f) &= \frac{1}{2} \sum_{x,y \in \mathscr{X}} \pi(x)\pi(y) \left(f(x) - f(y)\right)^2 \\ &= \pi(f^2) - \pi^2(f) \\ &= \mathrm{Var}(f). \end{aligned}$$

A natural way to quantify the variational behavior of P consists in comparing its Dirichlet form with that of the ideal chain Π. This leads to the following fundamental definition.

Definition 19 (Poincaré constant)**.** *The Poincaré constant of the chain is the quantity*

$$\gamma(P) := \inf \left\{ \frac{\mathscr{E}_P(f)}{\mathrm{Var}(f)}, \quad f \colon \mathscr{X} \to \mathbb{R} \text{ is not constant} \right\}.$$

Since the ratio $\mathscr{E}_P(f)/\mathrm{Var}(f)$ is invariant under translation and scaling, we also have

$$\gamma(P) = \inf \left\{ \mathscr{E}_P(f) \colon \|f\| = 1, \pi f = 0 \right\},$$

which shows that the infimum is actually attained.

Remark 25 (Time reversal)**.** *Replacing P by $P^\star$ changes $\vec{\pi}(x,y)$ to $\vec{\pi}(y,x)$, so we have*

$$\mathscr{E}_P(f) = \mathscr{E}_{P^\star}(f) = \mathscr{E}_{\frac{P+P^\star}{2}}(f),$$

for all observables $f \colon \mathscr{X} \to \mathbb{R}$. In particular, we deduce that

$$\gamma(P) = \gamma(P^\star) = \gamma\left(\frac{P+P^\star}{2}\right).$$

In words, the Dirichlet form and the Poincaré constant are invariant under time reversal.

The Poincaré constant happens to enjoy a simple spectral interpretation.

Lemma 21 (Spectral interpretation of the Poincaré constant)**.** *We always have*

$$\gamma(P) = 1 - \lambda_2\left(\frac{P+P^\star}{2}\right),$$

where $\lambda_2(Q)$ denotes the second largest eigenvalue of a self-adjoint transition matrix Q.

In particular, if P is lazy and reversible, then

$$\gamma(P) \;=\; 1 - \lambda_\star(P).$$

Proof. First consider the case where P is reversible. By decomposing the observable f in our orthonormal basis $(\phi_1, \dots, \phi_N)$ of eigenfunctions, one finds

$$\langle (I-P)f, f\rangle \;=\; \sum_{k=2}^{N} (1-\lambda_k)\,|\langle f, \phi_k\rangle|^2 .$$

On the other hand, since $\phi_1 \equiv 1$, we have $\langle f, \phi_1\rangle = \pi f$, so that

$$\mathrm{Var}(f) \;=\; \sum_{k=2}^{N} |\langle f, \phi_k\rangle|^2 .$$

It readily follows that $\mathscr{E}_P(f) \geq (1-\lambda_2(P))\mathrm{Var}(f)$, with equality when $f = \phi_2$. Thus,

$$\gamma(P) \;=\; 1 - \lambda_2(P),$$

which establishes the claim when P is reversible. The general case is obtained by replacing P with $(P + P^\star)/2$, which is always reversible and has the same Poincaré constant (Remark 25). Finally, when P is lazy and reversible, we have $\mathrm{Sp}(P) \subseteq [0,1]$, so $\lambda_\star(P) = \lambda_2(P)$. □

Remark 26 (Range of $\gamma(P)$)**.** *The above result shows that we always have $\gamma(P) \in [0,2]$, and even that $\gamma(P) \in [0,1]$ in the case where P is lazy.*

We can now answer the first question raised at the beginning of this chapter.

Lemma 22 (Variational contraction)**.** *For all $f\colon \mathscr{X} \to \mathbb{R}$, we have*

$$\mathrm{Var}(Pf) \;\leq\; [1 - \gamma(P^\star P)]\,\mathrm{Var}(f).$$

Moreover, if P is lazy, then $\gamma(P^\star P) \geq \gamma(P)$.

Proof. Using the equality $\pi f = \pi P f$, and the definition of the adjoint $P_\star$, we have

$$\begin{aligned}
\mathrm{Var}(f) - \mathrm{Var}(Pf) &= \langle f, f\rangle - \langle Pf, Pf\rangle \\
&= \langle f, f\rangle - \langle f, P^\star P f\rangle \\
&= \langle (I - P^\star P) f, f\rangle \\
&= \mathscr{E}_{P^\star P}(f) \\
&\geq \gamma(P^\star P)\mathrm{Var}(f),
\end{aligned}$$

and the first claim readily follows. Now, if P is lazy, then for all $x, y \in \mathscr{X}$, we have

$$\begin{aligned}\pi(x)P^\star P(x,y) &= \pi(x)\sum_{z\in S} P_\star(x,z)P(z,y)\\ &\geq \pi(x)\left(P(x,x)P(x,y)+P_\star(x,y)P(y,y)\right)\\ &\geq \frac{1}{2}\left(\pi(x)P(x,y)+\pi(y)P(y,x)\right).\end{aligned}$$

Multiplying by $\frac{1}{2}\left(f(x)-f(y)\right)^2$ and then summing over all $x, y \in \mathscr{X}$, we obtain

$$\mathscr{E}_{P^\star P}(f) \;\geq\; \mathscr{E}_P(f).$$

Since this is true for all $f\colon \mathscr{X} \to \mathbb{R}$, we can safely conclude that $\gamma(P^\star P) \geq \gamma(P)$. □

The above lemma shows that the variance of any observable decays exponentially fast under the repeated action of P, with rate $1/\gamma(P^\star P)$. This answers the first question raised at the beginning of the chapter. The following result answers the second question, by showing that $1/\gamma(P)$ plays the role of a relaxation time, without requiring reversibility.

Theorem 10 (Poincaré bound)**.** *If P is lazy, then for all $\varepsilon \in (0,1)$,*

$$\mathrm{t}_{\mathrm{REL}}(P) \leq \frac{2}{\gamma(P)} \quad \textit{and} \quad \mathrm{t}^{(\varepsilon)}_{\mathrm{MIX}}(P) \;\leq\; \left\lceil \frac{2}{\gamma(P)} \log \frac{1}{2\varepsilon\sqrt{\pi_\star}} \right\rceil.$$

Proof. Our starting point is the Cauchy-Schwarz bound, which we recall here:

$$d_{\mathrm{TV}}\left(P^t(x,\cdot),\pi\right) \;\leq\; \frac{1}{2}\left\| \frac{P^t(x,\cdot)}{\pi(\cdot)} - 1 \right\|.$$

In the reversible case, the existence of an orthonormal basis of eigenfunctions of P had allowed us to prove exponential decay of the right-hand side. Without reversibility, we no longer have an orthonormal basis of eigenfunctions at our disposal, but we can write

$$\frac{P^t(x,y)}{\pi(y)} \;=\; \frac{P^{\star t}(y,x)}{\pi(x)} \;=\; P^{\star t} f_x(y),$$

where we have introduced the observable $f_x\colon y \mapsto \frac{\delta_x(y)}{\pi(x)}$. Since $\pi P^{\star t} f_x = \pi f_x = 1$, we see that

$$\left\| \frac{P^t(x,\cdot)}{\pi(\cdot)} - 1 \right\|^2 \;=\; \mathrm{Var}\left(P^{\star t} f_x\right),$$

and the right-hand side can be estimated using Lemma 22 and Remark 25:

$$\begin{aligned}\mathrm{Var}\left(P^{\star t} f_x\right) &\leq \mathrm{Var}\left(f_x\right)\left(1-\gamma(P^\star)\right)^t \\ &= \left(\frac{1}{\pi(x)}-1\right)\left(1-\gamma(P)\right)^t \\ &\leq \frac{e^{-\gamma(P)t}}{\pi_\star},\end{aligned}$$

Puttings things together leads to the following conclusion, which implies the two claims:

$$\forall t \in \mathbb{N}, \quad \mathscr{D}_P(t) \leq \frac{e^{-\frac{\gamma(P)t}{2}}}{2\sqrt{\pi_\star}}.$$

□

Remark 27 (Comparison with Theorem 4)**.** *A careful inspection reveals that the above argument actually holds with $\gamma(PP^\star)$ instead of $\gamma(P)$, without the need for lazyness. When P is reversible, we have $1-\gamma(PP^\star) = \lambda_\star^2(P)$, and we recover Theorem 4.*

5.2 Cheeger inequalities

In a previous chapter, we have used the geometric notion of conductance to provide a good lower bound on mixing times. As we will now see, this quantity also gives a two-sided control on the Poincaré constant. This result is fundamental, because it creates a bridge between a geometric notion (conductance) and a spectral one (the Poincaré constant).

Theorem 11 (Cheeger's inequalities)**.** *For any transition kernel P, we have*

$$\frac{\Phi^2(P)}{2} \leq \gamma(P) \leq 2\Phi(P).$$

Proof. In view of Remarks 25 and 19, we may suppose that P is reversible. Fix $A \subseteq \mathscr{X}$. The observable $f = \mathbf{1}_A$ satisfies $\mathrm{Var}(f) = \pi(A)\pi(A^c)$ and $\mathscr{E}_P(f) = \vec{\pi}(A \times A^c)$. Consequently,

$$\Phi(A) = \pi(A^c)\frac{\mathscr{E}_P(f)}{\mathrm{Var}(f)}.$$

The claimed upper bound follows immediately. For the lower bound, we can assume that $\lambda_2(P) \geq 0$, because $\Phi(P) \in [0,1]$. Consider a non-negative observable $f\colon \mathscr{X} \to \mathbb{R}_+$ with

$\pi(f=0) \geq \frac{1}{2}$. For each $t \in \mathbb{R}_+$, we may choose $A = \{f > t\}$ in the definition of $\Phi(P)$ to get

$$\Phi(P)\pi(f>t) \;\leq\; \sum_{x,y\in\mathscr{X}} \vec{\pi}(x,y)\mathbf{1}_{(f(y)\leq t<f(x))}.$$

Integrating w.r.t. t and interchanging the sum and integral, we obtain

$$\Phi(P)\pi f \;\leq\; \frac{1}{2}\sum_{x,y\in\mathscr{X}} \vec{\pi}(x,y)\,|f(y)-f(x)|\,.$$

We now replace f by f^2, and apply the Cauchy-Schwarz inequality:

$$\Phi^2(P)\|f\|^4 \;\leq\; \frac{1}{4}\left(\sum_{x,y\in\mathscr{X}} \vec{\pi}(x,y)\,(f(y)-f(x))^2\right)\left(\sum_{x,y\in\mathscr{X}} \vec{\pi}(x,y)\,(f(y)+f(x))^2\right).$$

Expanding the squares, we see that the right-hand side simplifies to $\|f\|^4 - \langle f, Pf\rangle^2$, so that

$$\Phi^2(P)\|f\|^4 \;\leq\; \|f\|^4 - \langle f, Pf\rangle^2. \tag{48}$$

To conclude, we would like to take $f = \phi_2$, but our initial assumption $\pi(f=0) \geq \frac{1}{2}$ has no reason to be satisfied. Let us instead choose $f = \max(\phi_2, 0)$, which verifies the assumption upon changing ϕ_2 to -ϕ_2 if necessary. Since $f \geq 0$ and $f \geq \phi_2$, we have $Pf \geq 0$ and $Pf \geq \lambda_2\phi_2$, so $Pf \geq \lambda_2 f$. Thus, $\langle f, Pf\rangle \geq \lambda_2\|f\|^2$, and (48) easily implies the claim. □

By combining Theorems 10 and 11, we obtain the following important upper bound.

Corollary 5 (Conductance upper-bound)**.** *For any lazy kernel P and any $\varepsilon \in (0,1)$,*

$$\mathrm{t}^{(\varepsilon)}_{\mathrm{MIX}}(P) \;\leq\; \left\lceil \frac{4}{\Phi^2(P)} \ln \frac{1}{2\varepsilon\sqrt{\pi_\star}} \right\rceil.$$

Example 9 (Hypercube)**.** *Consider lazy random walk on the hypercube. The set $A := \{x \in \{0,1\}^n \colon x_1 = 0\}$ satisfies $\pi(A) = \frac{1}{2}$ and $\vec{\pi}(A \times A^c) = \frac{1}{4n}$, so that $\Phi(A) = \frac{1}{2n}$. We deduce that $\Phi(P) \leq \frac{1}{2n}$. On the other hand, we know that $\gamma(P) = 1 - \lambda_\star(P) = \frac{1}{n}$, so that there is equality in Cheeger's upper bound. In particular,*

$$\Phi(P) \;=\; \frac{1}{2n}.$$

Example 10 (Cycle). *On $\mathbb{Z}/n\mathbb{Z}$, any set A of size $k \leq n/2$ contains at least two boundary edges, so $\Phi(A) \geq \frac{1}{2k}$. Moreover, there is equality when $A = \{1, \ldots, k\}$. Thus,*

$$\Phi(P) \quad = \quad \frac{1}{2\lfloor n/2 \rfloor}.$$

We know that $\gamma(P) = 1 - \lambda_\star(P) \sim \frac{\pi^2}{n^2}$, so Cheeger's lower bound is sharp up to prefactors.

Example 11 (Universal bound). *For lazy random walk on an undirected graph $G = (V, E)$, the crude bound $\phi(P) \geq 2\vec{\pi}_\star$ gives $\mathrm{t}_{\mathrm{REL}}(P) \leq 8|E|^2$, hence $\mathrm{t}_{\mathrm{MIX}}(P) = \mathcal{O}(|E|^2 \log |E|)$.*

Example 12 (Expanders). *A sequence of graphs $(G_n)_{n \geq 1}$ is an expander sequence if*

1. *the number of vertices diverges;*
2. *the degrees are uniformly bounded;*
3. *the conductances are uniformly bounded from below.*

The lazy random walk on G_n satisfies $\mathrm{t}_{\mathrm{MIX}}(P_n) = \Theta(\log |G_n|)$, by Corollaries 2 and 5.

5.3 Comparison principle

The Poincaré constant $\gamma(P)$ was defined by comparing the Dirichlet form of P to that of the idealized kernel Π. Replacing the latter by an arbitrary kernel Q is the starting point of a very powerful comparison theory for Markov chains: suppose P is a sophisticated chain, whose mixing time is delicate to estimate directly, and consider a much simpler chain Q which has the same stationary law π. Then, one can *transfer* quantitative results from Q to P, by paying a "price" $\gamma(P \colon Q)$ that depends on how *close* Q is to P.

Definition 20 (Comparison constant). *Given two irreducible transition kernels P, Q on*

the same state space $\mathscr{X}$, their comparison constant is defined as

$$\gamma(P:Q) \;:=\; \inf\left\{\frac{\mathscr{E}_P(f)}{\mathscr{E}_Q(f)},\quad f\colon \mathscr{X}\to\mathbb{R} \text{ is not constant}\right\}.$$

Note in particular that $\gamma(P:\Pi)=\gamma(P)$, by Remark 24. The motivation behind this definition is contained in the following elementary but fruitful observation.

Lemma 23 (Comparison principle)**.** *If P and Q have the same stationary law, then*

$$\gamma(P) \;\geq\; \gamma(Q)\gamma(P:Q).$$

Thus, any lower bound on $\gamma(Q)$ yields a lower bound on $\gamma(P)$, at a price $\gamma(P:Q)$.

Proof. For any non-constant observable $f\colon \mathscr{X}\to\mathbb{R}$, we have by definition

$$\frac{\mathscr{E}_P(f)}{\mathrm{Var}(f)} \;=\; \frac{\mathscr{E}_P(f)}{\mathscr{E}_Q(f)}\times\frac{\mathscr{E}_Q(f)}{\mathrm{Var}(f)} \;\geq\; \gamma(P\colon Q)\times\frac{\mathscr{E}_Q(f)}{\mathrm{Var}(f)}.$$

Taking an infimum over all possible choices for f concludes the proof. □

Remark 28 (Extension)**.** *The Courant-Fischer-Weyl min-max principle expresses the $k-$th largest eigenvalue of any compact self-adjoint operator A on a Hilbert space $\mathcal{H}$ as*

$$\lambda_k(A) \;=\; \max_{\dim(F)=k}\min_{f\in F}\frac{\langle Af,f\rangle}{\langle f,f\rangle},$$

where the maximum ranges over all $k-$dimensional subspaces $F\subseteq\mathcal{H}$. Applying this to $A=\mathbb{I}-\frac{P+P^\star}{2}$ on $\mathcal{H}=L^2(\mathscr{X},\pi)$, we obtain the global comparison principle

$$1-\lambda_k\left(\frac{P+P^\star}{2}\right) \;\geq\; \gamma(P\colon Q)\left(1-\lambda_k\left(\frac{Q+Q^\star}{2}\right)\right),$$

for all $1\leq k\leq|\mathscr{X}|$, of which the above lemma is only the special case $k=2$.

5.4 Distinguished paths

We now present a very robust technique to establish a lower bound on the comparison constant $\gamma(P\colon Q)$ between two irreducible chains P and Q on the same state space $\mathscr{X}$.

Theorem 12 (Distinguished paths)**.** *Let P, Q be kernels on $\mathscr{X}$ with supports E_P, E_Q and weights $\vec{\pi}_P, \vec{\pi}_Q$. For each $(x,y) \in E_Q$, let $\Upsilon_{x,y}$ be a path from x to y in G_P. Then,*

$$\frac{1}{\gamma(P\colon Q)} \;\le\; \max_{e \in E_P} c(e), \quad \textit{where} \quad c(e) := \frac{1}{\vec{\pi}_P(e)} \sum_{(x,y)\in E_Q} \vec{\pi}_Q(x,y) |\Upsilon x, y| \mathbf{1}_{(e \in \Upsilon_{x,y})}.$$

Here $|\Upsilon|$ denotes the length of the path Υ, and $e \in \Upsilon$ means that e is traversed by Υ.

Proof. Fix an observable $f\colon \mathscr{X} \to \mathbb{R}$. Writing $\nabla f(x,y) := f(y) - f(x)$, we have

$$\mathscr{E}_Q(f) \;=\; \frac{1}{2} \sum_{(x,y)\in E_Q} \vec{\pi}_Q(x,y) \left|\nabla f(x,y)\right|^2.$$

Now for each $(x,y) \in E_Q$, we may use the fact that $\Upsilon_{x,y}$ is a path from x to y to write

$$\begin{aligned} \left|\nabla f(x,y)\right|^2 \;&=\; \left| \sum_{e \in \Upsilon_{x,y}} \nabla f(e) \right|^2 \\ &\le\; |\Upsilon_{x,y}| \sum_{e \in \Upsilon_{x,y}} \left|\nabla f(e)\right|^2, \end{aligned}$$

by the Cauchy-Schwarz inequality. Inserting above and re-arranging, we obtain

$$\begin{aligned} \mathscr{E}_Q(f) \;&\le\; \frac{1}{2} \sum_{e \in E_P} \vec{\pi}_P(e) \left|\nabla f(e)\right| c(e) \\ &\le\; \mathscr{E}_P(f) \max_{e \in E_P} c(e). \end{aligned}$$

Since this inequality holds for all observables $f\colon \mathscr{X} \to \mathbb{R}$, the result follows. □

The simplest and most natural choice for Q is the ideal matrix Π that mixes in one step:

Corollary 6 (Rank-one case $Q = \Pi$)**.** *Let P be a transition kernel on $\mathscr{X}$ and for each $(x,y) \in \mathscr{X} \times \mathscr{X}$, let $\Upsilon_{x,y}$ be a path from x to y in G_P. Then,*

$$\frac{1}{\gamma(P)} \;\le\; \max_{e \in E_P} c(e), \quad \textit{where} \quad c(e) := \frac{1}{\vec{\pi}(e)} \sum_{(x,y)\in \mathscr{X}\times\mathscr{X}} \pi(x)\pi(y) |\Upsilon_{x,y}| \mathbf{1}_{(e \in \Upsilon x,y)}.$$

Remark 29 (Congestion). *The quantity $c(e)$ is called the congestion induced at the edge e by the collection of paths $(\Upsilon_{x,y})_{(x,y)\in E_Q}$. The challenge lies in choosing paths that will make the maximal congestion $\max_{e\in E_P} c(e)$ as low as possible. Note that we must have*

$$\begin{aligned}\max_{e\in E_P} c(e) &\geq \sum_{e\in E_P} \vec{\pi}_P(e)c(e) \\ &= \sum_{(x,y)\in E_Q} \vec{\pi}_Q(x,y)|\Upsilon_{x,y}|^2 \\ &\geq \sum_{(x,y)\in E_Q} \vec{\pi}_Q(x,y)\mathrm{dist}^2(x,y).\end{aligned}$$

Thus, the best congestion we can hope for is the average quadratic distance (in G_P) between two consecutive states of $\mathrm{MC}(\mathscr{X}, Q, \pi)$. To achieve this optimum (at least approximately), we need our paths $(\Upsilon_{x,y})_{(x,y)\in E_Q}$ to be (close to) shortest paths in G_P, and the resulting congestion to be (close to) uniform across all edges, as in the next example.

Example 13 (Square grid). *Let P_n be the transition matrix for lazy simple random walk on a $n \times n$ grid: the state space is $\mathscr{X} = [n] \times [n]$, and two vertices $x = (x_1, x_2)$ and $y = (y_1, y_2)$ are neighbors if $|x_1 - x_2| + |y_1 - y_2| = 1$. For $x, y \in \mathscr{X}$, consider the path $\Upsilon_{x,y}$ of minimum length that goes from x to y first horizontally, then vertically. The congestion induced at any edge is easily seen to be of order n^2, so Corollary 6 gives*

$$\mathrm{t}_{\mathrm{REL}}(P_n) = \mathcal{O}(n^2),$$

which is actually the correct order of magnitude.

Example 14 (Universal bound). *One can always choose $\Upsilon_{x,y}$ to be a shortest path from x to y, and use the crude bound $|\Upsilon_{x,y}| \leq \mathrm{diam}(P)$ to deduce that*

$$\frac{1}{\gamma(P)} \leq \frac{\mathrm{diam}(P)}{\vec{\pi}_\star}.$$

In particular, for lazy simple random walk on a graph $G = (V, E)$, we have

$$\begin{aligned} \mathrm{t}_{\mathrm{REL}}(P) &= \mathcal{O}(\mathrm{diam}(G)|E|) \\ \mathrm{t}_{\mathrm{MIX}}(P) &= \mathcal{O}(\mathrm{diam}(G)|E| \log |E|). \end{aligned}$$

The two inequalities appearing in Remark 29 show that our distinguished paths $(\Upsilon_{x,y})_{(x,y)\in E_Q}$ should not only be as *short* as possible, but also as *spread-out* as possible across the graph, so that the congestion is fairly balanced among the edges. A simple but powerful idea for reducing the imbalance consists in letting the paths $(\Upsilon_{x,y})_{(x,y)\in E_Q}$ be *random*.

Theorem 13 (Random paths)**.** *Let P, Q be kernels on $\mathscr{X}$ with supports E_P, E_Q and weights $\vec{\pi}_P, \vec{\pi}_Q$. For $(x, y) \in E_Q$, let $\Upsilon_{x,y}$ be a random path from x to y in G_P. Then,*

$$\frac{1}{\gamma(P\colon Q)} \leq \max_{e\in E_P} c(e), \quad \textit{where} \quad c(e) := \frac{1}{\vec{\pi}_P(e)} \sum_{(x,y)\in E_Q} \vec{\pi}_Q(x, y)\mathbb{E}\left[|\Upsilon_{x,y}|\mathbf{1}_{(e\in\Upsilon_{x,y})}\right].$$

Proof. The argument is exactly the same as in the case of deterministic paths, except that we take expectations just before the very last inequality. □

Here is an example that illustrates the advantage of random paths over deterministic ones.

Example 15 (Lazy random walk on $K_{2,n}$)**.** *The graph $K_{2,n}$ has vertex set $\mathscr{X} = \{L, R\} \cup [n]$, two vertices being neighbors if one is in $\{L, R\}$ and the other in $[n]$. Any deterministic path $\Upsilon_{L,R}$ from L to R contributes to the congestion of $e \in \Upsilon_{L,R}$ by at least*

$$c(e) \geq \frac{1}{\vec{\pi}(e)}\pi(L)\pi(R)|\Upsilon_{L,R}| = n.$$

On the other hand, when the distinguished path $\Upsilon_{x,y}$ is chosen uniformly at random over all shortest paths from x to y, the congestion is only $c(e) = 2 - \frac{1}{n}$ for every edge e.

We conclude with an application to chains with a high amount of symmetry.

Definition 21 (Transitivity)**.** *An automorphism of P is a permutation ϕ on $\mathscr{X}$ so that*

$$\forall x, y \in \mathscr{X}, \qquad P\left(\phi(x), \phi(y)\right) = P(x, y).$$

The transition kernel P is called transitive if for any two states $x, x' \in \mathscr{X}$, there is an

automorphism of P that takes x to x'. It is called arc-transitive if for any two edges $(x,y),(x',y') \in E$, there is an automorphism that takes x to x' and y to y'.

Intuitively, a chain is transitive if "all states play the same role". Examples include all random walks on groups (take $\phi(z) = z \star x^{-1} \star x'$). Arc-transitivity is the stronger requirement that "all edges play the same role". Examples include the lazy simple random walk on the hypercube, the cycle, and more generally discrete tori of the form $\mathbb{Z}_n^d$ for any $d, n \in \mathbb{N}$.

Corollary 7 (Poincaré inequality for symmetric chains)**.**

(i) If P is transitive, then $\frac{1}{\gamma(P)} \leq \frac{\mathrm{diam}^2(P)}{p_\star}$, where $p_\star = \min_{(x,y)\in E} P(x,y)$.

(ii) If P is arc-transitive, then $\frac{1}{\gamma(P)} \leq \frac{\mathrm{diam}^2(P)}{1-\alpha}$, where $\alpha = \min_{x\in\mathscr{X}} P(x,x)$.

Proof. We use the methods of random paths to compare P with the rank-one matrix $Q = \Pi$. For each $x, y \in \mathscr{X}$, we take $\Upsilon_{x,y}$ to be a uniformly chosen shortest path between x and y. In view of Remark 29, we know that the resulting congestion satisfies

$$\sum_{e\in E} \vec{\pi}(e)c(e) \;=\; \sum_{x,y\in\mathscr{X}} \pi(x)\pi(y)\mathrm{dist}^2(x,y) \;\leq\; \mathrm{diam}^2(P).$$

When P is arc-transitive, the quantities $\vec{\pi}(e)$ and $c(e)$ do not depend on $e \in E$. Consequently, the left-hand side equals $\vec{\pi}(E) \times \max_{e\in E} c(e)$, and the second bound follows because $\vec{\pi}(E) = 1 - \sum_{x\in\mathscr{X}} \vec{\pi}(x,x) = 1 - \alpha$. If P is only transitive, then we can write

$$\sum_{e\in E} \vec{\pi}(e)c(e) \;=\; \sum_{x\in\mathscr{X}} \pi(x) \sum_{y\neq x} P(x,y)c(x,y) \;\geq\; p_\star \max_{e\in E} c(e),$$

because the quantity $\pi(x)\sum_{y\neq x} P(x,y)c(x,y)$ does not depend on x. □