Markov Chains and Decomposition Methods

Mathematics Capstone Project

Name: Suei-Wen Chen
Supervisor: Prof. Pierre Youssef

New York University Abu Dhabi
Abu Dhabi, United Arab Emirates
April, 2023

Abstract

This paper aims to provide a self-contained discussion on the mixing time of finite Markov chains and relevant functional inequalities. Section 1 introduces the notion of discrete Markov chains along with some basic properties and examples. Section 2 discusses the convergence of Markov chains to stationarity and continuous-time chains. In section 3, two approaches (Markovian coupling and spectral analysis) to bound the mixing time of a chain are developed. Section 4 defines three functional inequalities of interest (Poincaré, log-Sobolev, and modified log-Sobolev inequalities) with motivation from continuous-time chain. The connections between mixing time and these inequalities are drawn, and the example of complete graph is given. Section 5 starts with a decomposition method for calculating the constants associated with the functional inequalities and moves on to a generalization of this method, which is the main contribution of this paper.

Acknowledgements

I would like to express my sincere gratitude to my advisor Professor Pierre Youssef for his continuous support, encouragement, and guidance throughout my undergraduate thesis. His mentorship has been instrumental in shaping the direction of my research and ensuring its completion. I would also like to extend my thanks to Professor Justin Salez for the discussion and valuable feedback on my work which helped me refine my ideas and improve the overall quality of my research. In addition, I want to express my appreciation to my family and friends for their constant academic and emotional support. Their encouragement and belief in me kept me motivated and inspired throughout the journey, and I couldn't have completed this project without them. Finally, I want to thank all the people who provided assistance, guidance, and support throughout the course of my research. Your contributions were invaluable, and I am deeply grateful for all the support that made this project possible.

Table of Contents

1 Finite Markov Chains

We start with a finite set Ω which we call the state space. An element of Ω is called a state. Intuitively, a Markov chain models a process which moves among the states $x \in \Omega$ with transition probabilities that only depend on the current position. Throughout this paper we only consider Markov chains on finite state spaces.

1.1 Discrete Markov Chains

Definition 1.1. *A stochastic process $(X_t)_{t\in\mathbb{N}_0}$ taking value in Ω is called a (finite) Markov chain if it satisfies the Markov property: for any $t \in \mathbb{N}$ and $x_0, x_1, ..., x_{t-1}, x, y \in \Omega$ such that*

$$\mathbb{P}(H_{t-1} \cap \{X_t = x\}) > 0, \text{ where } H_{t-1} = \{X_0 = x_0, X_1 = x_1, ..., X_{t-1} = x_{t-1}\}$$

we have

$$\mathbb{P}(X_{t+1} = y \mid H_{t-1} \cap \{X_t = x\}) = \mathbb{P}(X_{t+1} = y \mid X_t = x).$$

Given the Markov property, we can store all the transition probabilities in a matrix P of size $|\Omega| \times |\Omega|$, called the transition matrix of the chain, whose (i, j)-entry is defined by the probability of going from state i to state j:

$$P(i, j) := P(X_{t+1} = j \mid X_t = i).$$

Thus P is stochastic, meaning that all its entries are non-negative and every row sums to 1. For every $t \geq 0$, the row vector with elements $\mu_t(x) = \mathbb{P}(X_t = x)$ $(x \in \Omega)$ represents the probability distribution of X_t, the chain at time t. For all $y \in \Omega$, conditioning on all possible states at time t gives

$$\mu_{t+1}(y) = \sum_{x\in\Omega} \mu_t(x)P(x, y)$$

In vector form, the equality says $\mu_{t+1} = \mu_t P$ and hence $\mu_t = \mu_0 P^t$ for all $t \geq 0$. We let $\mathbb{P}_\mu$ and $\mathbb{E}_\mu$ denote the probabilities and expectations given that the starting distribution $\mu_0 = \mu$. If $\mu_0 = \delta_x$ is concentrated at some $x \in \Omega$ we write $\mathbb{P}_x$ and $\mathbb{E}_x$ for $\mathbb{P}_{\delta_x}$ and $\mathbb{E}_{\delta_x}$ respectively. Using this notation, we have

$$\mathbb{P}_x(X_t = y) = \mu_t(y) = (\delta_x P^t)(y) = P^t(x, y),$$

namely, the probability of moving from x to y using t steps is the (x, y)-entry of P^t.

The long-term behavior of Markov chains is a central theme of study. In other words, we would like to know what happens to P^t when $t \to \infty$. A fundamental idea is called the stationary or invariant measure of a chain.

Definition 1.2. *A probability distribution π on Ω is called a stationary distribution of a Markov chain P if $\pi = \pi P$, or written element-wise,*

$$\pi(y) = \sum_{x\in\Omega} \pi(x)P(x, y) \text{ for all } y \in \Omega.$$

Some common regularity assumptions for a Markov chain include reversibility, irreducibility and aperiodicity. These are desirable properties that give nice results regarding the limiting behavior of a chain.

Definition 1.3. *Let P be a Markov chain on Ω.*

1. *A measure μ on Ω is said to be reversible under the chain if for all $x, y \in \Omega$ we have $\mu(x)P(x, y) = \mu(y)P(y, x)$.*

2. *The chain is said to be irreducible if for any $x, y \in \Omega$ there exists $t > 0$ such that $P^t(x, y) > 0$.*

3. *The period of a state $x \in \Omega$ is defined to be the greatest common divisor of "all possible return times" $T(x) := \{t \geq 1 \mid P^t(x, x) > 0\}$. A state $x \in \Omega$ is aperiodic if its period is 1.*

4. *The chain is said to be the aperiodic if all states have period 1. If it is not aperiodic, it is said to be periodic.*

Note that if a measure π is reversible under P, then π after normalization is a stationary distribution of P.

Proposition 1.1. *Any two states in an irreducible chain P have the same period.*

Proof. Fix two states x and y and assume x has period d. By irreducibility, $P^r(x, y) > 0$ and $P^s(y, x) > 0$ for some $r, s > 0$. Then for any $t \in \mathbb{N}$ such that $P^t(y, y) > 0$, we have $P^{r+s+t}(x, x) > 0$ and $P^{r+s+2t}(x, x) > 0$ (start at x, reach y in r steps, stay at y after t or $2t$ steps, and then go to x in s steps). Since d divides $r + s + 2t$ and $r + s + t$, it divides their difference t and hence d divides the period of y. By symmetry, the period of y also divides the period of x, so x and y have the same period. $\square$

We thus define the period of an irreducible chain to be the period which is common to all its states. The matrix of an irreducible and aperiodic chain has the following nice property.

Proposition 1.2. *If a chain P is irreducible and aperiodic, then there exists an integer r_0 such that $P^r(x, y) > 0$ for all $x, y \in \Omega$ and $r \geq r_0$.*

Proof. We use the fact that a subset of nonnegative integers which is closed under addition and has greatest common divisor 1 contains all but finitely many nonnegative integers (See Levin & Peres[1], Lemma 1.30). By aperiodicity, $T(x)$ has greatest common divisor 1 and is closed under addition since $P^{r+s}(x, x) \geq P^r(x, x)P^s(x, x) > 0$ if $r, s \in T(x)$. So by the fact above, for every $x \in \Omega$ there exists some $t_x \in \mathbb{N}$ such that $t \geq t_x$ implies $t \in T(x)$. By irreducibility, for any $x, y \in \Omega$ we have $P^{r_{x,y}}(x, y) > 0$ for some $r_{x,y}$. Thus for $t \geq t_x + r_{x,y}$ we have $P^t(x, y) \geq P^{t-r_{x,y}}(x, x)P^{r_{x,y}}(x, y) > 0$. Therefore if $t \geq \max_{x,y\in\Omega}(t_x + r_{x,y})$ then for any $x, y \in \Omega$ we have $P^t(x, y) \geq 0$. $\square$

It turns out that irreducibility is all we need to ensure the existence and uniqueness of stationary measure for a chain. Furthermore, we have the following characterization of the stationary measure using the first return time.

Theorem 1.1. *Let X_t $(t \in \mathbb{N})$ be an irreducible Markov chain with transition matrix P. Then the chain has a unique stationary distribution π. Moreover, for any state $z \in \Omega$ we have*

$$\pi(z) = \frac{1}{\mathbb{E}_z[\tau_z^+]},$$

where $\tau_z^+ := \{t \geq 1 \mid X_t = z\}$ is the first hitting time of z, also called the first return time of z if $X_0 = z$.

Before proving this theorem we introduce some terminology. Let P be a Markov chain on a finite set Ω. Given $x, y \in \Omega$ we say that y is accessible from x (denoted by $x \to y$) if $P^r(x, y) > 0$ for some $r \in \mathbb{N}$. We say that x communicates with y (denoted by $x \leftrightarrow y$) if $[x \to y$ and $y \to x]$ or $x = y$. Then $\leftrightarrow$ defines an equivalence relation on Ω, and the equivalent classes are called communicating classes. Note that a chain is irreducible if and only if it has only one communicating class. For $x \in \Omega$ we let $[x]$ denote its class. We say that a state $x \in \Omega$ is recurrent if $\mathbb{P}_x(\tau_x^+ < \infty) = 1$ and positive recurrent if $\mathbb{E}_x[\tau_x^+] < \infty$. If $\mathbb{P}_x(\tau_x^+ = \infty) > 0$, x is called transient.

Lemma 1.1. *For any states x and y of a finite irreducible chain P, $\mathbb{E}_x[\tau_y^+] < \infty$. In particular, all states of an irreducible finite chain are positive recurrent.*

Proof. Since P is irreducible, for any $v, w \in \Omega$ there exists $r_{vw} > 0$ such that $P^{r_{vw}}(v, w) > 0$. Let

$$r := \max_{v,w\in\Omega} r_{vw} \in \mathbb{N} \text{ and } \epsilon := \min_{v,w\in\Omega} P^{r_{vw}}(v, w) > 0.$$

Then for any $z, y \in \Omega$ there exists $j \leq r$ such that $P^j(z, y) \geq \epsilon$; in other words, for any value of the chain at time t, the probability of hitting state y within the next r time units is at least ϵ. Thus for $k \in \mathbb{N}$ we have

$$1 - \mathbb{P}_x(\tau_y^+ > kr \mid \tau_y^+ > (k-1)r) \geq \epsilon \implies \mathbb{P}_x(\tau_y^+ > kr) \leq (1-\epsilon)\mathbb{P}_x(\tau_y^+ > (k-1)r)$$

and repeatedly apply this for $k-1$ in place of k yields $\mathbb{P}_x(\tau_y^+ > kr) \leq (1-\epsilon)^k$. Recall that for an $\mathbb{N}$-valued random variable Z we have

$$Z = \sum_{n=1}^{\infty} \mathbb{1}_{Z\geq n} \implies \mathbb{E}Z = \sum_{n=1}^{\infty} \mathbb{E}\mathbb{1}_{Z\geq n} = \sum_{n=1}^{\infty} \mathbb{P}(Z \geq n).$$

Therefore, using the fact that $\mathbb{P}(\tau_y^+ > \cdot)$ is non-increasing we have

$$\mathbb{E}_x[\tau_y^+] = \sum_{n=0}^{\infty} \mathbb{P}_x(\tau_y^+ \geq n) \leq \sum_{k=0}^{\infty} r\mathbb{P}_x(\tau_y^+ > kr) \leq r\sum_{k=0}^{\infty}(1-\epsilon)^k < \infty.$$

□

Proof of Theorem 1.1. Fix a state $z \in \Omega$. Intuitively, for $y \in \Omega$ the stationary distribution at y should be proportional to the number of visits to y in between visits to z. Let N_y be the number of visits of X_t to y before returning to z, namely,

$$N_y := \sum_{t=0}^{\tau_z^+-1} \mathbb{1}_{X_t=y} = \sum_{t=0}^{\infty} \mathbb{1}_{X_t=y}\mathbb{1}_{t<\tau_z^+}.$$

Define

$$\tilde{\pi}(y) := \mathbb{E}_z[N_y] = \sum_{t=0}^{\infty} \mathbb{P}(X_t = y, \tau_z^+ > t).$$

Notice that $N_y \le \tau_z^+$ and hence $\tilde{\pi}(y) \le \mathbb{E}_z[\tau_z^+] < \infty$ by Lemma 1.1. Also, $X_0, ..., X_t$ determines the event $\{\tau_z^+ > t\}$, so

$$\begin{aligned}
\sum_{x\in\Omega} \tilde{\pi}(x)P(x,y) &= \sum_{x\in\Omega}\sum_{t=0}^{\infty} \mathbb{P}(X_t = x, \tau_z^+ > t)P(x,y) \\
&= \sum_{t=0}^{\infty}\sum_{x\in\Omega} \mathbb{P}(X_t = x, X_{t+1} = y, \tau_z^+ > t) \\
&= \sum_{t=0}^{\infty} \mathbb{P}(X_{t+1} = y, \tau_z^+ > t) = \sum_{t=1}^{\infty} \mathbb{P}(X_t = y, \tau_z^+ \ge t). \\
\implies \tilde{\pi}(y) &= \sum_{x\in\Omega} \tilde{\pi}(x)P(x,y) + \mathbb{P}(X_0 = y, \tau_z^+ > 0) - \sum_{t=1}^{\infty} \mathbb{P}(X_t = y, \tau_z^+ = t) \\
&= \sum_{x\in\Omega} \tilde{\pi}(x)P(x,y) + \mathbb{P}(X_0 = y) - \mathbb{P}(X_{\tau_z^+} = y).
\end{aligned}$$

Since $X_0 = z$, $\mathbb{P}(X_0 = y) = \mathbb{P}(X_{\tau_z^+} = y)$ whether $y = z$ or not. Thus $\tilde{\pi}P = \tilde{\pi}$, and since

$$\sum_{y\in\Omega} \tilde{\pi}(y) = \mathbb{E}_z\left[\sum_{y\in\Omega} N_y\right] = \mathbb{E}_z[\tau_z^+]$$

it follows that $\pi_z := \tilde{\pi}/\mathbb{E}_z[\tau_z^+]$ defines a stationary distribution for P.

Next, we show that the stationary distribution is unique. We say that a function h on Ω is harmonic if $Ph = h$, namely $h(x) = \sum_{y\in\Omega} P(x,y)h(y)$ for all x. Note that since P is irreducible, any harmonic function h must be constant: Say h attains its maximum M at x_0. For $y \in \Omega$ such that $P(x_0,y) > 0$, if $h(y) < M$ then

$$h(x_0) = P(x_0,y)h(y) + \sum_{z\in\Omega, z\neq y} P(x_0,z)h(z) < M,$$

which is a contradiction and thus $h(y) = M$ for all y with $P(x_0,y) > 0$. Now for any $z \in \Omega$ there exists a sequence $x_0, x_1, ..., x_n = z$ with $P(x_i, x_{i+1}) > 0$ by the irreducibility assumption. The previous argument tells us $h(x_0) = h(x_1) = \cdots = h(z)$. Thus h is constant. This implies that the kernel of $P - I$ and hence that $P^t - I$ is of rank 1, so there is a unique probabilistic row vector μ such that $\mu P = \mu$.

Finally, for any $z \in \Omega$, by uniqueness of stationary distribution we have

$$\pi(z) = \pi_z(z) = \frac{\tilde{\pi}(z)}{\mathbb{E}_z\tau_z^+} = \frac{1}{\mathbb{E}_z\tau_z^+}.$$

□

In some cases, instead of using the matrix P it is more convenient to work with $Q := P - I$, called the Markov generator of the chain defined by P. Hereafter we will refer to a chain by either its transition matrix or its Markov generator.

1.2 Examples of Markov Chains

In this section we present some examples of discrete Markov chains.

Example 1.1. (Random walk on graphs) A fundamental example of Markov chain is the simple random walk on a graph. Let $G = (\Omega, E)$ be a undirected graph of size $|\Omega| = n$. The simple random walk on G is defined by $P(x, y) = 1/\deg(x)$ if $x \sim y$ (x and y are neighbors) and 0 otherwise. Let $N := \sum_{z\in\Omega} \deg(z) = 2|E|$. Then $\pi(x) := \deg(x)/N$ is reversible under P and hence is a stationary distribution of P. Note that P is irreducible if G is connected, but P need not be aperiodic.

A simply way to make any chain P aperiodic is to consider its lazy version defined by $\tilde{P} = (P + I)/2$. Note that if π is stationary under P then it is stationary under $\tilde{P}$ as well.

Example 1.2. (Random walk on the n-dimensional hypercube) The n-dimensional hypercube is a graph whose vertices are binary n-tuples $(x_1, ..., x_n) \in \mathcal{Q}^n := \{0, 1\}^n$ and, two vertices are connected by an edge if and only if they differ by exactly one coordinates. The simple random walk on $\mathcal{Q}^n$ is thus $P(x, y) = 1/n$ if $|x-y| = 1$ (the distance as vectors in $\mathbb{R}^n$) and 0 otherwise. It is clear that the uniform distribution $\pi(x) = 2^{-n}$ is stationary and that the chain is irreducible. However, the simple random walk is periodic since each state has period 2, so we often consider the lazy walk on $\mathcal{Q}^n$.

Example 1.3. (Coupon collection) A company issues n different types of coupon, each time independently and with uniform probability. A collector seeks to obtain a complete set. Let the random variable τ be the number of coupons that the collector must obtain so that his collection contains all n types. Then τ is called the coupon collector random variable.

The coupon collecting problem can be modeled as a Markov chain as follows. Let X_t denote the number of different types represented in the collector's first t coupons. Then $X_0 = 0$ and

$$\mathbb{P}(X_{t+1} = k + 1 \mid X_t = k) = \frac{n-k}{n}, \ \mathbb{P}(X_{t+1} = k \mid X_t = k) = \frac{k}{n}.$$

Then (X_t) is a Markov chain on $\{0, 1, ..., n\}$ and τ is the number of steps required for X_t to reach the state n.

Proposition 1.3. *Let τ be the coupon random variable with n types. Then*

$$\mathbb{E}[\tau] = n\sum_{k=1}^{n}\frac{1}{k} \text{ and } \mathbb{P}(\tau > \lceil n\log n + cn\rceil) \le e^{-c} \text{ for any } c > 0.$$

Proof. For $k = 0, ..., n$ let τ_k be the number of coupons collected when the collection first contains k types of coupon. Note that $\tau_k - \tau_{k-1}$ is a geometric random variable with success probability $(n - k + 1)/n$ and that $\tau = \tau_n = \sum_{k=1}^{n}(\tau_k - \tau_{k-1})$. Thus

$$\mathbb{E}[\tau] = \sum_{k=1}^{n}\mathbb{E}[\tau_k - \tau_{k-1}] = \sum_{k=1}^{n}\frac{n}{n-k+1} = n\sum_{k=1}^{n}\frac{1}{k}.$$

Now consider the event A_i that the first $\lceil n \log n + cn\rceil$ coupons collected do not contain type i. Then

$$\mathbb{P}(\tau > \lceil n\log n + cn\rceil) = \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{i=1}^{n}\mathbb{P}(A_i) = \sum_{i=1}^{n}\left(1-\frac{1}{n}\right)^{\lceil n\log n+cn\rceil}$$
$$\le n\exp\left(-\frac{n\log n + cn}{n}\right) = e^{-c}.$$

∎

2 Convergence of Markov Chains

We saw in the last section that an irreducible chain on a finite space has a unique stationary distribution. In this section we show that an irreducible and aperiodic chain converges to the stationary distribution. To make the notion of convergence of Markov chain precise, we introduce the total variation distance of probability distributions. The last subsection contains a discussion on continuous-time Markov chains, their convergence, and their relation to the underlying discrete time chain.

2.1 Total variation distance

Definition 2.1. *The total variation distance between two probability measures μ and ν on a measurable space (Ω, Σ) is defined by*

$$\|\mu - \nu\|_{TV} := \sup_{A \in \Sigma} |\mu(A) - \nu(A)|$$

In our context, Ω is a finite set and $\Sigma = 2^\Omega$, and hence

$$\|\mu - \nu\|_{TV} = \max_{A \subseteq \Omega} |\mu(A) - \nu(A)|.$$

The following proposition gives a useful characterization of total variation distance.

Proposition 2.1. $\|\mu - \nu\|_{TV} = \sum_{x \in \Omega} (\mu(x) - \nu(x))_+ = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|.$

Proof. Let $A := \{x \in \Omega \mid \mu(x) \geq \nu(x)\}$. Then for any $B \subseteq \Omega$,

$$\begin{aligned} \mu(B) - \nu(B) &= \mu(A \cap B) - \nu(A \cap B) + \mu(A^c \cap B) - \nu(A^c \cap B) \\ &\leq \mu(A \cap B) - \nu(A \cap B) \leq \mu(A) - \nu(A) \end{aligned}$$

and similarly

$$\nu(B) - \mu(B) \leq \nu(A^c) - \mu(A^c).$$

Since $\mu(A) - \nu(A) = \nu(A^c) - \mu(A^c)$ and $\mu(B) - \nu(B) = \mu(A) - \nu(A)$ if $B = A$, we know that

$$\|\mu - \nu\|_{TV} = \mu(A) - \nu(A) = \nu(A^c) - \mu(A^c).$$

Therefore

$$\|\mu - \nu\|_{TV} = \mu(A) - \nu(A) = \sum_{x \in \Omega} (\mu(x) - \nu(x))_+$$

and

$$\|\mu - \nu\|_{TV} = \frac{1}{2}(\mu(A) - \nu(A) + \nu(A^c) - \mu(A^c)) = \sum_{x \in \Omega} |\mu(x) - \nu(x)|.$$

$\square$

Corollary 2.1. *The total variation distance defines a metric on the space of probability measure on Ω.*

The total variation distance, albeit natural, may not always be the easiest metric to work with. In a later section we shall discuss the natural ℓ^2 structure induced by the stationary distribution of an irreducible chain and the relation between the two metrics.

2.2 Convergence to stationarity

We now introduce a key theorem regarding the convergence of Markov chains. The mixing time of a chain refers to the rate of such convergence.

Definition 2.2. *Given an irreducible Markov chain P with stationarity distribution π, we define its distance to stationarity by*

$$d(x;t) := \|P^t(x,\cdot) - \pi\|_{TV},\ \ d(t) := \max_{x\in\Omega} d(x;t)$$

Another related distance metric is define by

$$\overline{d}(t) := \max_{x,y\in\Omega} \|P^t(x,\cdot) - P^t(y,\cdot)\|_{TV}.$$

Definition 2.3. *The mixing time of an irreducible Markov chain is defined by the first time at which the distance to stationarity drops below a given threshold $\epsilon > 0$:*

$$t_{mix}(x;\epsilon) := \inf\{t \geq 0 \mid d(x;t) \leq \epsilon\},\ \ t_{mix}(\epsilon) := \inf\{t \geq 0 \mid d(t) \leq \epsilon\}.$$

We collect some useful properties related to the total variation distance, including the fact that $d(t)$ is non-increasing which justifies the definition of mixing time.

Proposition 2.2. *Let P be a Markov chain with stationary distribution π and let μ and ν be probability measures on Ω. For $t \geq 1$ we have the following.*

1. $d(t) = \sup_\mu \|\mu P^t - \pi\|_{TV}$ *and* $\overline{d}(t) = \sup_{\mu,\nu} \|\mu P^t - \nu P^t\|_{TV}$.
2. $\|\mu P^t - \nu P^t\|_{TV} \leq \|\mu - \nu\|_{TV}$. *In particular,* $\|\mu P^{t+1} - \pi\|_{TV} \leq \|\mu P^t - \pi\|_{TV}$, *namely advancing the chain can only move it closer to stationarity.*
3. $d(t+1) \leq d(t)$ *and* $\overline{d}(t+1) \leq \overline{d}(t)$
4. $d(t) \leq \overline{d}(t) \leq 2d(t)$.

Proof. First we prove 4. That $\overline{d}(t) \leq 2d(t)$ follows from Proposition 2.1. Since

$$\begin{aligned}\|P^t(x,\cdot) - \pi\|_{TV} = \max_{A\subseteq\Omega} |P^t(x,A) - \pi(A)| &= \max_{A\subseteq\Omega} \left| P^t(x,A) - \sum_{y\in S} \pi(y) P^t(y,A) \right| \\ &= \max_{A\subseteq\Omega} \left| \sum_{y\in\Omega} \pi(y)(P^t(x,A) - P^t(y,A)) \right| \\ &\leq \max_{A\subseteq\Omega} \sum_{y\in\Omega} \pi(y)|P^t(x,A) - P^t(y,A)| \leq \sum_{y\in\Omega} \pi(y)\overline{d}(t) = \overline{d}(t)\end{aligned}$$

for all $x \in \Omega$ we have $d(t) \leq \overline{d}(t)$.

Next we prove 1. Clearly $d(t) = \sup_{x\in\Omega} \|\delta_x P^t - \pi\|_{TV} \leq \sup_\mu \|\mu P^t - \pi\|_{TV}$. Since

$$\begin{aligned}\|\mu P^t - \pi\|_{TV} &= \frac{1}{2}\sum_{x\in\Omega} \left|(\mu P^t)(x) - \pi(x)\right| = \frac{1}{2}\sum_{x\in\Omega} \left| \sum_{y\in\Omega} \mu(y)[P^t(y,x) - \pi(x)] \right| \\ &\leq \frac{1}{2}\sum_{x\in\Omega}\sum_{y\in\Omega} \mu(y)|P^t(y,x) - \pi(x)| = \sum_{y\in\Omega} \mu(y) \left[\frac{1}{2}\sum_{x\in\Omega} |P^t(y,x) - \pi(x)| \right] \\ &= \sum_{y\in\Omega} \mu(y)\|P^t(x,\cdot) - \pi\|_{TV} \leq \sum_{y\in\Omega} \mu(y) d(t) = d(t),\end{aligned}$$

the first equality follows. Similarly, it is obvious that $\bar{d}(t) \le \sup_{\mu,\nu} \|\mu P^t - \nu P^t\|_{TV}$. Now let $y_x := \arg\max P^t(\cdot, x)$ and $z_x := \arg\min P^t(\cdot, x)$. Then for any μ and ν

$$\begin{aligned}\|\mu P^t - \nu P^t\|_{TV} &= \frac{1}{2}\sum_{x\in\Omega}\left|\sum_{y\in\Omega}\mu(y)P^t(y,x) - \sum_{z\in\Omega}\mu(z)P^t(z,x)\right| \\ &\le \frac{1}{2}\sum_{x\in\Omega}|P^t(y_x,x) - P^t(z_x,x)| \le \bar{d}(t)\end{aligned}$$

and the second equality follows.

Finally, it suffices to prove 2 since 3 is a direct consequence of the particular case in 2 in light of 1. Indeed,

$$\begin{aligned}\|\mu P^t - \nu P^t\|_{TV} &= \frac{1}{2}\sum_{x\in\Omega}\left|\sum_{y\in\Omega}(\mu(y)-\nu(y))P^t(y,x)\right| \le \frac{1}{2}\sum_{x,y\in\Omega}|\mu(y)-\nu(y)|P^t(y,x) \\ &= \frac{1}{2}\sum_{y\in\Omega}|\mu(y)-\nu(y)|\sum_{x\in\Omega}P^t(y,x) = \|\mu-\nu\|_{TV},\end{aligned}$$

and in particular $\|\mu P^{t+1} - \pi\|_{TV} = \|(\mu P^t)P - \pi P\|_{TV} \le \|\mu P^t - \pi\|_{TV}$. □

Now we introduce the main theorem of convergence, which tells us that it always makes sense to talk about the mixing time of an irreducible and aperiodic Markov chain on a finite state space.

Theorem 2.1. *Suppose P is an irreducible and aperiodic Markov chain on a finite state space Ω with stationary distribution π. Then there exists $C > 0$ and $\alpha \in (0,1)$ such that $d(t) \le C\alpha^t$.*

Proof. Since P is irreducible and aperiodic, P^r has strictly positive entries for some r by Proposition 1.2. Fix $\delta > 0$ so small that $P^r(x,y) \ge \delta\pi(y)$ for all $x, y \in \Omega$. Then we can write $P^r = (1-\theta)\Pi + \theta Q$ where $\theta = 1-\delta$, Π is the square matrix with π as rows and Q is a stochastic matrix. Next, we show by induction that for $k \ge 1$,

$$P^{rk} = (1-\theta^k)\Pi + \theta^k Q^k.$$

The base case is already verified. Assume this equality holds for k. Then

$$\begin{aligned}P^{r(k+1)} &= [(1-\theta^k)\Pi + \theta^k Q^k]P^r = (1-\theta^k)\Pi P^r + \theta^k Q^k[(1-\theta)\Pi + \theta Q] \\ &= (1-\theta^k)\Pi P^r + \theta^k(1-\theta)Q^k\Pi + \theta^{k+1}Q^{k+1} \\ &= (1-\theta^k)\Pi + \theta^k(1-\theta)\Pi + \theta^{k+1}Q^{k+1} = (1-\theta^{k+1})\Pi + \theta^{k+1}Q^{k+1}.\end{aligned}$$

Now for $t \ge 1$ write $t = rk + j$ where $k \in \mathbb{N}$, $0 \le j < r$. By the induction hypothesis

$$P^t - \Pi = [P^{rk} - \Pi]P^j = [\theta^k Q^k - \theta^k \Pi]P^j = \theta^k[Q^k P^j - \Pi].$$

Let ν_x be the x-th row of $Q^k P^j$, which is a probability distribution. For each $x \in \Omega$, summing over the absolute values of elements of $P^t - \Pi$ in the x-th row divided by 2 gives $\|P^t(x,\cdot) - \pi\|_{TV} = \theta^k\|\nu_x - \pi\|_{TV} \le \theta^k$. Therefore $d(t) \le \theta^k \le \theta^{t/r-1}$ and it suffices to take $C = 1/\theta$ and $\alpha = \theta^{1/r}$. □

2.3 Continuous-Time Chains

Discrete Markov chains has the limitation that the times at which transitions happen are natural numbers. It is sometimes more reasonable to consider continuous-time Markov chains, dynamics in which the next change happens randomly (according to a certain distribution) after the current one. In addition, continuous-time chains naturally sidestep the problem of periodicity encountered in discrete-time chains. This section provides a brief discussion on continuous-time Markov chains which will serve as the motivation for developments in later sections.

Given a transition matrix P on a finite set Ω, we define the continuous-time chain $(X_t)_{t\geq 0}$ with transition matrix P as follows. The idea is to make moves according to P based on an exponential clock of parameter $r > 0$. More precisely, let $(T_i)_{i\in\mathbb{N}}$ be i.i.d. exponential random variables with rate r and $(Z_k)_{k\in\mathbb{N}_0}$ be a Markov chain with transition matrix P independent of $(T_i)_{i\in\mathbb{N}}$. We define S_k $(k = 0, 1, 2, ...)$ as $S_0 := 0$, $S_k := \sum_{i=1}^{k} T_i$ and $X_t := Z_k$ where $S_k \leq t < S_{k+1}$. Thus S_k $(k \geq 1)$ are called transition times, the times when X_t changes occur according to Z_k.

Note that $S_k \sim \Gamma(k, r)$ follows a gamma distribution as a sum of i.i.d. exponential random variables since the characteristic function of S_k is

$$\phi_{S_k}(t) = \mathbb{E}[e^{itS_k}] = \prod_{j=1}^{k} \mathbb{E}[e^{itT_j}] = \left(\frac{r}{r - it}\right)^k .$$

Let $N_t := \max\{k \mid S_k \leq t\}$ be the number of jumps up to (and including) time t. Then $N_t = k \iff S_k \leq t < S_{k+1}$ and a simple calculation shows that N_t follows a Poisson distribution with parameter rt. Therefore, an equivalent way of defining a continuous-time chain $(X_t)_{t\geq 0}$ from a chain $(Z_k)_{k\in\mathbb{N}}$ is to take a Poisson random variable N_t (independent of Z_k) with parameter rt and set $X_t := Z_{N_t}$. By setting $r = 1$ we have a natural continuization of discrete chains.

Now we compute the heat kernel $H_t(x, y) := \mathbb{P}_x(X_t = y)$, the probability of a chain ending up at y at time t if it started at x. Notice that

$$\mathbb{P}_x(X_t = y | N_t = k) = \mathbb{P}_x(Z_k = y) = P^k(x, y)$$

and hence $H_t = \mathbb{E}Z_{N_t}$, namely

$$H_t(x, y) = \sum_{k=0}^{\infty} \mathbb{P}_x(X_t = y \mid N_t = k)\mathbb{P}_x(N_t = k) = \sum_{k=0}^{\infty} P^k(x, y)\frac{(rt)^k e^{-rt}}{k!} = e^{rtQ}(x, y).$$

If we require that $r = 1$ then $H_t = e^{tQ}$, namely $\mathbb{P}_x(X_t = y) = H_t(x, y) = e^{tQ}(x, y)$. More generally, let μ_t be the distribution of X_t given that the starting distribution is $X_0 \sim \mu$, then by the law of total probability

$$\mu_t(y) = \mathbb{P}_\mu(X_t = y) = \sum_{x\in\Omega} \mathbb{P}_\mu(X_t = y \mid X_0 = x)\mathbb{P}_\mu(X_0 = x) = \sum_{x\in\Omega} \mathbb{P}_x(X_t = y)\mu(x)$$

and therefore

$$\mu_t = \mu H_t. \tag{1}$$

Next we compute the derivative of H_t when applied to a function viewed as a column vector. For any function $f : \Omega \to \mathbb{R}$ we have

$$\begin{aligned}
\frac{d}{dt}(H_t f)(x) &= \sum_{y\in\Omega} \frac{d}{dt}[e^{tQ}(x,y)]f(y) = \sum_{y\in\Omega} \left[\sum_{k=0}^{\infty} \frac{t^k}{k!} Q^{k+1}(x,y)\right] f(y) \\
&= \sum_{y\in\Omega}\sum_{z\in\Omega}\sum_{k=0}^{\infty} \frac{t^k}{k!} Q^k(x,z)Q(z,y)f(y) = \sum_{y\in\Omega}\sum_{z\in\Omega} e^{tQ}(y,z)Q(z,x)f(y) \\
&= \sum_{y\in\Omega} (e^{tQ}Q)(x,y)f(y) = (e^{tQ}Qf)(x),
\end{aligned}$$

and hence

$$\frac{d}{dt} H_t f = H_t Q f = Q H_t f.$$

For $\mu_t = \mu H_t$, a similar calculation shows that

$$\frac{d}{dt}\mu_t = \mu Q H_t = \mu_t Q. \tag{2}$$

If φ is an eigenfunction of Q with eigenvalue $\mu := \lambda - 1$ (i.e. φ is an eigenfunction of P with eigenvalue λ), then

$$\left(\frac{d}{dt}H_t\right)\varphi = \frac{d}{dt}(H_t\varphi) = H_t Q\phi = \mu H_t \varphi$$

and hence $H_t\varphi = e^{\mu t}\varphi = e^{t(\lambda-1)}\varphi$.

It turns out that in order for a continuous-time chain on a finite space to converge, irreducibility is sufficient. (After all, the same notion of period as defined for discrete chains does not make sense in the continuous setting.)

Theorem 2.2. *Let P be an irreducible matrix (not necessarily aperiodic) and H_t be its corresponding heat kernel. Then there exists a unique probability measure π such that $\pi H_t = \pi$ for all $t \geq 0$ and there exist $C > 0$ and $\alpha \in (0,1)$ such that*

$$d^{cont}(t) := \max_{x\in\Omega} \|H_t(x,\cdot) - \pi\|_{TV} \leq C\alpha^t.$$

Proof. Let $\tilde{P} := (P + I)/2$ be the lazy version of P. Then $\pi\tilde{P} = \pi$ and the heat kernel of $\tilde{P}$ is $\tilde{H}_t = H_{t/2}$. Denote the lazy continuous chain by $(\tilde{X}_t)$ and let N_t be as defined above. Then by Theorem 2.1 there exist $C > 0$ and $\beta \in (0,1)$ such that $d(k) \leq C\beta^k$ where $d(k)$ is the distance of $\tilde{P}^k$ to stationarity. Thus

$$\begin{aligned}
\|H_t(x,\cdot) - \pi\|_{TV} &= \|\tilde{H}_{2t}(x,\cdot) - \pi\|_{TV} \\
&= \frac{1}{2}\sum_{y\in\Omega}\left|\sum_{k=0}^{\infty}[\mathbb{P}_x(\tilde{X}_{2t} = y \mid N_{2t} = k) - \pi(y)]\mathbb{P}(N_{2t} = k)\right| \\
&= \frac{1}{2}\sum_{y\in\Omega}\left|\sum_{k=0}^{\infty}[\tilde{P}^k(x,y) - \pi(y)]\mathbb{P}(N_{2t} = k)\right| \\
&\leq \frac{1}{2}\sum_{y\in\Omega}\sum_{k=0}^{\infty}\left|\tilde{P}^k(x,y) - \pi(y)\right|\mathbb{P}(N_{2t} = k) \\
&\leq \sum_{k=0}^{\infty} d(k)\mathbb{P}(N_{2t} = k) \leq \sum_{k=0}^{\infty} C\beta^k e^{-2t}\frac{(2t)^k}{k!} = Ce^{-2t(1-\beta)}
\end{aligned}$$

and therefore $d^{cont}(t) \leq C(e^{2(\beta-1)})^t$. $\square$

The mixing time of a continuous chain $t_{mix}^{cont}(\epsilon)$ and $t_{mix}^{cont}(x;\epsilon)$ is defined analogously to the discrete case (Definition 2.3), with P^t replaced by H_t. In fact, we have a stronger result that relates the distance to stationarity of a discrete chain to that of its continuous-time chain.

Theorem 2.3. *Let P be an irreducible chain, $\tilde{P} = (P+I)/2$ be the lazy version of P and H_t be its heat kernel.*

1. *Let Φ be a Poisson random variable with mean $2k$. Then*
$$\|H_k(x,\cdot) - \pi\|_{TV} \leq \left\|\tilde{P}^k(x,\cdot) - \pi\right\|_{TV} + \mathbb{P}(\Phi < k).$$

2. *Let $Y \sim B(4m, 1/2)$ be a binomial random variable and Ψ be a Poisson random variable with mean m, and let $\eta_m := \|\mathbb{P}(Y \in \cdot) - \mathbb{P}(\Psi + m \in \cdot)\|_{TV}$. Then*
$$\|\tilde{P}^{4m}(x,\cdot) - \pi\|_{TV} \leq \|H_m(x,\cdot) - \pi\|_{TV} + \eta_m.$$

Remark 2.1. By the law of large number, $\mathbb{P}(\Phi < k) \to 0$ as $k \to \infty$; in fact, $\mathbb{P}(\Phi < k) \leq (e/4)^k$. In addition, using Chebyshev's inequality and Stirling's formula one can show that $\eta_m \to 0$ as $m \to \infty$ (see Chapter 20.2 in[1]).

Proof. Let N_t be the Poisson random variable with mean 1 indicating the number of transitions of the continuous-time chain. By the law of total probability and the fact that $\|P^j(x,\cdot) - \pi\|_{TV}$ is non-increasing in j,

$$\begin{aligned}\|H_t(x,\cdot) - \pi\|_{TV} &\leq \sum_{j=0}^{k-1} \mathbb{P}(N_t = j)\|P^j(x,\cdot) - \pi\|_{TV} + \sum_{j=k}^{\infty} \mathbb{P}(N_t = j)\|P^j(x,\cdot) - \pi\|_{TV} \\ &\leq \mathbb{P}(N_t < k) + \|P^k(x,\cdot) - \pi\|_{TV}.\end{aligned}$$

Apply this inequality to $\tilde{P}$, $\tilde{H}_t = H_{t/2}$ in place of P, H_t and setting $t = 2k$ finishes the proof for 1. For 2, observe that

$$H_m P^m = \sum_{k=0}^{\infty} \mathbb{P}(\Psi + m = k) P^k, \;\; \tilde{P}^{4m} = \sum_{k=0}^{\infty} \mathbb{P}(Y = k) P^k.$$

Therefore

$$\begin{aligned}\|\tilde{P}^{4m}(x,\cdot) - \pi\|_{TV} &\leq \|H_m P^m(x,\cdot) - \pi\|_{TV} + \|\tilde{P}^{4m}(x,\cdot) - H_m P^m(x,\cdot)\|_{TV} \\ &\leq \|H_m(x,\cdot) - \pi\|_{TV} + \eta_m,\end{aligned}$$

since the chain is closer to π when moved m more steps after N_m steps. □

At first glance, it seems that Theorem 2.1 and 2.2 tell us everything we want to know about how fast the convergence occurs. However, when the state space is large, the base α is often close to 1 and gives bad upper bounds for the mixing time. In subsequent sections, we develop multiple methods to give better estimates for the mixing time. As we will see in later chapters, it is technically easier to work with continuous-time chains, and thanks to Theorem 2.3 (and Theorem 3.5) the bounds for the corresponding discrete chains will follow.

3 Coupling and Spectral Methods

In this section we introduce two common ways to bound the mixing time of a Markov chain. The method of coupling involves ”tying together” two probability measures or Markov chains by the same source of randomness, which will enable us to bound the mixing time from above in terms of the coalescence time. The spectral method applies to reversible chains and allows us to bound the mixing time in terms of the spectral gap both from above and from below.

3.1 Coupling of Markov chains

Definition 3.1. *A coupling of two probability distributions μ and ν on Ω is a pair of Ω-valued random variables (X, Y) defined on a single probability space such that $\mathbb{P}(X = x) = \mu(x)$ and $\mathbb{P}(Y = y) = \nu(y)$ for all $x, y \in \Omega$.*

It is easy to see that specifying a coupling (X, Y) of two probability measures μ and ν is equivalent to specifying a joint distribution q on $\Omega \times \Omega$ such that the two marginal distributions are μ and ν.

Any two distributions μ and ν have an independent coupling: simply take two independent random variables $X \sim \mu$, $Y \sim \nu$ and the product measure $\mu \otimes \nu$ gives an independent coupling. If $\mu = \nu$ we can take $X \sim \mu$ and $Y = X$. Then (X, Y) is a coupling consisting of two identical copies of X. It turns out that the total variation distance $\|\mu - \nu\|_{TV}$ is closely related to how two random variables X and Y in a coupling (X, Y) of μ and ν must differ.

Proposition 3.1. *Let μ and ν be two probability measures on Ω. Then*

$$\|\mu - \nu\|_{TV} = \min\{\mathbb{P}(X \neq Y) \mid (X, Y) \textit{ is a coupling of } \mu \textit{ and } \nu\}.$$

A coupling (X, Y) is called optimal if $\|\mu - \nu\|_{TV} = \mathbb{P}(X \neq Y)$. Thus any two probability measures have an optimal coupling.[2]

Proof. Let (X, Y) be a coupling of μ and ν. Then for any $A \subseteq \Omega$,

$$\begin{aligned}
|\mu(A) - \nu(A)| &= |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)| \\
&\leq |\mathbb{P}(X \in A, X = Y) - \mathbb{P}(Y \in A, X = Y)| \\
&\quad + |\mathbb{P}(X \in A, X \neq Y) - \mathbb{P}(Y \in A, , X \neq Y)| \\
&\leq 0 + \mathbb{P}(X \neq Y) = \mathbb{P}(X \neq Y).
\end{aligned}$$

This shows that $\|\mu - \nu\|_{TV} \leq \min \mathbb{P}(X \neq Y)$. Now we construct a coupling such that the equality is attained. Consider

$$p := \sum_{x \in \Omega} \min\{\mu(x), \nu(x)\} \in [0, 1]$$

and note that

$$\|\mu - \nu\|_{TV} = \sum_{x \in \Omega} (\mu(x) - \nu(x))_+ = \sum_{x \in \Omega} (\mu(x) - \min\{\mu(x), \nu(x)\}) = 1 - p.$$

Now let Z, Z_1 and Z_2 be random variables such that

- Z has distribution λ, where $\lambda(x) := \dfrac{\min\{\mu(x), \nu(x)\}}{p}$,
- Z_1 has distribution $\dfrac{(\mu(x) - \nu(x))_+}{\|\mu - \nu\|_{TV}}$, and Z_2 has distribution $\dfrac{(\nu(x) - \mu(x))_+}{\|\mu - \nu\|_{TV}}$.

Now define a coupling (X, Y) as follows. Let B be a Bernoulli random variable with parameter p. If $B = 1$ let $X = Y = Z$ and if $B = 0$ let $X = Z_1$ and $Y = Z_2$. Then

$$\mathbb{P}(X = x) = p\lambda(x) + (1-p)\frac{(\mu(x) - \nu(x))_+}{\|\mu - \nu\|_{TV}} = \mu(x)$$

and similarly $\mathbb{P}(Y = y) = \nu(y)$. Moreover, $X = Y$ if and only if $B = 1$ and hence

$$\mathbb{P}(X \neq Y) = 1 - p = \|\mu - \nu\|_{TV}.$$

Therefore (X, Y) is an optimal coupling of μ and ν. □

Definition 3.2.

1. *A coupling of Markov chains with transition matrix P is defined to be a process $(X_t, Y_t)_{t\in\mathbb{N}_0}$ such that both $(X_t)_{t\in\mathbb{N}_0}$ and $(Y_t)_{t\in\mathbb{N}_0}$ are Markov chains with transition matrix P, possibly with different starting distributions.*

2. *A Markovian coupling of two Markov chains on Ω with transition matrix P is a coupling $(X_t, Y_t)_{t\in\mathbb{N}_0}$ which is itself a Markov chain on $\Omega \times \Omega$ and satisfies*

$$\mathbb{P}(X_{t+1} = x' \mid X_t = x, Y_t = y) = P(x, x')$$

$$\mathbb{P}(Y_{t+1} = y' \mid X_t = x, Y_t = y) = P(y, y')$$

for all $x, x', y, y' \in \Omega$.

Note that any Markovian coupling $\{(X_t, Y_t)\}$ can be modified so that if $X_s = Y_s$ for some s, then $X_t = Y_t$ for all $t \geq s$.

If $\{(X_t, Y_t)\}$ is a coupling of Markov chains with starting distributions $X_0 \sim \mu$, $Y_0 \sim \nu$, then we let $\mathbb{P}_{\mu,\nu}$ denote the probability measure on the space where (X_t) and (Y_t) are both defined to make explicit the starting distributions.

Theorem 3.1. *Let P be an irreducible Markov chain on Ω with stationary distribution π. Let $\{(X_t, Y_t)\}$ be a Markovian coupling such that*

1. *$X_0 = x$ and $Y_0 = y$ for some $x, y \in \Omega$, and*

2. *if $X_s = Y_s$ for some s, then $X_t = Y_t$ for all $t \geq s$.*

Define $\tau_{couple} := \min\{s \geq 0 \mid X_t = Y_t$ for all $t \geq s\}$, the coalescence time of the two chains. Then

$$\|P^t(x, \cdot) - P^t(y, \cdot)\|_{TV} \leq \mathbb{P}_{x,y}(\tau_{couple} > t).$$

More generally, if we replace condition 1. by $X_0 \sim \mu$ and $Y_0 \sim \nu$ for some probability distributions μ and ν, then

$$\|\mu P^t - \nu P^t\|_{TV} \leq \mathbb{P}_{\mu,\nu}(\tau_{couple} > t).$$

Proof. For any $z \in \Omega$ the definition of a Markovian coupling gives

$$(\mu P^t)(z) = \mathbb{P}(X_t = z \mid X_0 \sim \mu) = \mathbb{P}_{\mu,\nu}(X_t = z)$$

$$(\nu P^t)(z) = \mathbb{P}(Y_t = z \mid Y_0 \sim \nu) = \mathbb{P}_{\mu,\nu}(Y_t = z),$$

so (X_t, Y_t) is a coupling of μP^t and νP^t. By Proposition 3.1 we have

$$\|\mu P^t - \nu P^t\|_{TV} \leq \mathbb{P}_{\mu,\nu}(X_t \neq Y_t) = \mathbb{P}_{\mu,\nu}(\tau_{couple} > t),$$

where the last equality follows from the definition of τ_{couple} and condition 2. □

Corollary 3.1. *Suppose that for each pair of states $x, y \in \Omega$ there is a Markovian coupling (X_t, Y_t) with $X_0 = x$, $Y_0 = y$ and coalescence time τ_{couple}. Then*

$$d(t) \leq \max_{x,y\in\Omega} \mathbb{P}_{x,y}(\tau_{couple} > t)$$

and therefore

$$t_{mix}(\epsilon) \leq \frac{1}{\epsilon} \max_{x,y\in\Omega} \mathbb{E}_{x,y}[\tau_{couple}].$$

Proof. The first inequality follows from taking the maximum over all pairs $x, y \in \Omega$ on both sides of the inequality $\|P^t(x,\cdot) - P^t(y,\cdot)\|_{TV} \leq \mathbb{P}_{x,y}(\tau_{couple} > t)$. By Markov's inequality, $\mathbb{P}_{x,y}(\tau_{couple} > t) \leq \mathbb{E}_{x,y}[\tau_{couple}]/t$. Setting $t = \mathbb{E}_{x,y}[\tau_{couple}]/\epsilon$ and then taking the maximum over all pairs of $x, y \in \Omega$ yields the second inequality. □

Example 3.1. (Hypercube) Note that the chain (X_t) of simple random walk on $\mathcal{Q}^n$ does not converge to the uniform distribution π on $\mathcal{Q}^n$ in the sense that $d(t)$ does not converge to 0: if $X_0 = x$, then for any odd $n \in \mathbb{N}$

$$d(t) \geq \frac{1}{2} \sum_{y\in\mathcal{Q}^n} \|P^n(x,y) - \pi(y)\|_{TV} \geq \frac{1}{2}|P^n(x,x) - \pi(x)| = 2^{-n-1}.$$

The lazy version, on the other hand, does converge to π by Theorem 2.1. To give an upper bound on the mixing time, we consider a Markovian coupling of two walks on $\mathcal{Q}^n$, with possibly different starting state, as follows. At each step, first pick a coordinate $j \in \{1, ..., n\}$ uniformly at random, choose a random bit between 0 and 1 with uniform probability, and update the state of the two chains by setting both their j-th bit to the same chosen bit. Then each of the two walks is a lazy random walk on $\mathcal{Q}^n$, and the coupling is Markovian.

Let τ be the first time when all coordinates have been selected at least once. Then the two walks agree after step τ. Note that τ is the coupon collecting random variable in Example 1.3. Thus by Corollary 3.1 and Proposition 1.3,

$$d(n \log n + cn) \leq \mathbb{P}(\tau > n \log n + cn) \leq e^{-c}$$

for any $c > 0$. Thus setting $c = \log \epsilon^{-1}$ gives $d(n \log n + (\log \epsilon^{-1})n) \leq \epsilon$ and hence

$$t_{mix}(\epsilon) \leq n \log n + (\log \epsilon^{-1})n.$$

A lower bound is given by $t_{mix}(\epsilon) \geq \frac{1}{2} n \log n - c_\epsilon n$ (see Levin & Peres, Section 7.3.1[1]). Therefore $t_{mix} = \Theta(n \log n)$.

3.2 Spectral Method

In this section we explore the relation between the mixing time of a chain and its eigenvalues. As we will see in Theorem 3.4, the mixing time of an irreducible, aperiodic and reversible chain is closely related to its second largest eigenvalue in absolute value. This kind of spectral analysis is possible thanks to the famous Perron–Frobenius theorem [3]. Before stating a partial version of the theorem and its consequences we introduce some definitions.

We say that a matrix T is positive/nonnegative (denoted by $T > 0/T \geq 0$) if all its entries are positive/nonnegative. We say that T is primitive if $T^n > 0$ for some $n \in \mathbb{N}$ and that T is irreducible if for any (i, j) location there exist $n \in \mathbb{N}$ such that $T^n(i, j) > 0$.

Theorem 3.2 (Perron-Frobenius)**.** *Let T be an irreducible matrix.*

1. *T has an eigenvalue $\lambda_{\max} \in \mathbb{R}_{>0}$ such that $|\lambda| \leq \lambda_{\max}$ for all eigenvalues $\lambda \in \mathbb{C}$.*
2. *The eigenspace associated with $\lambda_{\max}$ is one-dimensional and is spanned by a vector $x > 0$ with positive entries.*
3. *If T is primitive then $|\lambda| < \lambda_{\max}$ for all eigenvalues $\lambda \neq \lambda_{\max}$.*

Applying this theorem to transition matrices we get the following.

Theorem 3.3. *Let P be a transition matrix. Then*

1. *If $\lambda \in \mathbb{C}$ is an eigenvalue then $|\lambda| \leq 1$*
2. *If P is irreducible then the eigenspace associated with 1 is one-dimensional and is generated by $(1, 1, ..., 1)^T$.*
3. *If P is irreducible and aperiodic then -1 is not an eigenvalue.*

Proof. For any function $f : \Omega \to \mathbb{C}$ we have

$$\begin{aligned}\|Pf\|_\infty = \max_{x\in\Omega}\left|\sum_{y\in\Omega} P(x, y)f(y)\right| &\leq \max_{x\in\Omega}\sum_{y\in\Omega} P(x, y)|f(y)| \\ &\leq \|f\|_\infty \max_{x\in\Omega}\sum_{y\in\Omega} P(x, y) = \|f\|_\infty.\end{aligned}$$

Thus if φ is an eigenfunction with eigenvalue λ then

$$\|P\varphi\|_\infty = \|\lambda\varphi\|_\infty = |\lambda|\|\varphi\|_\infty \leq \|\varphi\|_\infty \implies |\lambda| \leq 1.$$

Note that 1 is indeed an eigenvalue of P with eigenfunction $(1, ..., 1)^T$ since P is stochastic. Thus $1 = \lambda_{\max}$ and its eigenspace is one-dimensional. Finally, if P is irreducible and aperiodic then P is primitive by Proposition 1.2, so $|\lambda| < 1$ and in particular $\lambda \neq -1$. □

Let $|\Omega| = n$ and π be the stationary measure of a chain P. We consider the natural inner product $\langle f, g\rangle_\pi$ and the usual p-norm $\|f\|_p$ ($p \in [1, \infty]$) of real-valued functions on Ω defined by

$$\langle f, g\rangle_\pi := \sum_{x\in\Omega} f(x)g(x)\pi(x) \text{ and } \|f\|_p := \left(\sum_{x\in\Omega} |f(x)|^p\pi(x)\right)^{1/p}.$$

It is then natural to consider the ℓ^p distance to stationarity and compare it to the total variance distance. The densities of transition measure with respect to the stationary servers this purpose.

Definition 3.3. *Let P be an irreducible chain with stationary measure π and H_t be its heat kernel. Define for $k \in \mathbb{N}$ and $t \geq 0$*

$$p^k(x,\cdot) = p_x^k(\cdot) := \frac{P^k(x,\cdot)}{\pi(\cdot)},\ h_t(x,\cdot) = h_t^x(\cdot) := \frac{H_t(x,\cdot)}{\pi(\cdot)},$$

the density of $P^k(x,\cdot)$ and $H_t(x,\cdot)$ with respect to π.

Note that $H_t(x,y) = e^{-t}\sum_{k=0}^{\infty}\frac{P^k(x,y)}{k!}$ and hence $h_t(x,y) = e^{-t}\sum_{k=0}^{\infty}\frac{p^k(x,y)}{k!}$.

Definition 3.4. *For $p \geq 1$ we define the ℓ^p distance to stationarity by*

$$d_p(x;t) := \left\|p_x^t - 1\right\|_p,\ d_p(t) := \max_{x\in\Omega} d_p(x;t)$$

Lemma 3.1. *Assume that P is irreducible with stationary π. Then for $p \in [1,\infty]$ and $x \in \Omega$ we have $d(x;t) = (1/2)d_1(x;t)$ and $d(x;t) \leq (1/2)d_p(x;t)$.*

Proof. By Proposition 2.1 we have

$$\|P^t(x,\cdot) - \pi\|_{TV} = \frac{1}{2}\sum_{y\in\Omega}|P^t(x,y) - \pi(y)| = \frac{1}{2}\sum_{y\in\Omega}\pi(y)\left|p_x^t(y) - 1\right| = \frac{1}{2}\left\|p_x^t - 1\right\|_1.$$

By Hölder's inequality, $\|p_x^t - 1\|_1 \leq \|p_x^t - 1\|_p$ and the assertions follow. $\square$

The case $p = 2$ is particularly useful because the associated inner product $\langle\cdot,\cdot\rangle_\pi$ enables us to decompose a function using an orthonormal basis when the chain is reversible, according to the following Proposition.

Proposition 3.2. *Assume that π is reversible with respect to P. Then*

1. *All the eigenvalues are real and $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n \geq -1$ where $n = |\Omega|$*
2. *There exists orthonormal eigenfunctions $f_1, .., f_n$ with respect to $\langle\cdot,\cdot\rangle_\pi$ and $f_1 = (1,1,...,1)^T$.*
3. *$p^t(x,\cdot) = P^t(x,\cdot)/\pi(\cdot)$ can be decomposed as*

$$p^t(x,y) = \frac{P^t(x,y)}{\pi(y)} = \sum_{j=1}^{n} f_j(x)f_j(y)\lambda_j^t. \tag{3}$$

Remark 3.1. It follows from the definition of H_t that $e^{\lambda_j - 1}$ $(j = 1,...,n)$ are eigenvalues of H_t with the same set of eigenfunctions $f_1,...,f_n$ as P, and therefore

$$h_t(x,y) = \frac{H_t(x,y)}{\pi(y)} = \sum_{j=1}^{n} f_j(x)f_j(y)e^{(\lambda_j-1)t}. \tag{4}$$

In addition, using the decomposition we find that $\delta_x = \sum_{j=1}^n f_j(x)\pi(x)f_j$ and $\pi(x) = \delta_x(x)\pi(x) = \sum_{j=1}^n f_j(x)^2\pi(x)^2$ and therefore

$$\pi(x)^{-1} = \sum_{j=1}^{n} f_j(x)^2. \tag{5}$$

Proof. Let D_π be the diagonal matrix with entries $\pi(x)$. By reversibility, $A := D_\pi^{1/2} P D_\pi^{-1/2}$ is symmetric since $A(x,y) = \sqrt{\pi(x)/\pi(y)}P(x,y)$. By the spectral theorem, $A = \Phi\Lambda\Phi^{-1}$ for some $\Phi = [\phi_1 \cdots \phi_n]$ orthonormal with respect to the standard inner product $\langle\cdot,\cdot\rangle$ on $\mathbb{R}^n$ and some real matrix $\Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$ where $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n \geq -1$ by Theorem 3.3. Thus $P = (D_\pi^{-1/2}\Phi)\Lambda(D_\pi^{-1/2}\Phi)^{-1}$ and 1. follows. Now let $f_j := D_\pi^{-1/2}\varphi_j$. Then $f_1, ..., f_n$ are orthonormal since $\langle f_i, f_j\rangle_\pi = \langle\varphi_i, \varphi_j\rangle$, and $f_1 = (1, ..., 1)$ since P is stochastic. With this orthonormal basis we can write $\delta_y = \sum_{j=1}^n \langle\delta_y, f_j\rangle_\pi f_j = \sum_{j=1}^n f_j(y)\pi(y)f_j$ and hence

$$P^t(x,y) = (P^t\delta_y)(x) = \left(\sum_{j=1}^n f_j(y)\pi(y)P^t f_j\right)(x) = \sum_{j=1}^n f_j(y)\pi(y)\lambda_j^t f_j(x).$$

□

Definition 3.5. *Suppose P is irreducible and aperiodic. Then the eigenvalues of P satisfy $\lambda_1 = 1$ and $|\lambda| < \lambda_1$ if $\lambda \neq \lambda_1$. Let $\lambda_* := \max\{|\lambda| : \lambda \neq 1\}$. We call $\gamma_* := 1 - \lambda_*$ the absolute spectral gap of P, $\gamma := 1 - \lambda_2$ the spectral gap of P, and $t_{rel} := 1/\gamma_*$ the relaxation time. Note that $\gamma \in \mathbb{R}$ if P is reversible.*

Proposition 3.3. *Let (P, π) be reversible and let $C_x := (1 - \pi(x))/\pi(x)$ for $x \in \Omega$. Then*

$$\|p_x^t - 1\|_2^2 = \sum_{j=2}^n f_j(x)^2\lambda_j^{2t} \leq C_x\lambda_*^{2t} = C_x e^{2t\log(1-\gamma_*)} \leq C_x e^{-2\gamma_* t} \tag{6}$$

$$\|h_t^x - 1\|_2^2 = \sum_{j=2}^n f_j(x)^2 e^{2(\lambda_j - 1)t} \leq C_x e^{2(\lambda_2-1)t} = C_x e^{-2\gamma t} \tag{7}$$

Proof. This follows directly from Equations (3), (4) and (5), the orthonormality of $f_1, ..., f_n$ and the inequality $\log(1-x) \leq -x$ for $0 \leq x \leq 1$. □

The following theorem relates the mixing time and the (absolute) spectral gap (or relaxation time) of a chain.

Theorem 3.4. *Assume that P is irreducible, aperiodic and reversible with respect to π. Then for $0 < \epsilon < 1$ and $x \in \Omega$ we have*

$$t_{mix}(x;\epsilon) \leq \frac{1}{\gamma_*}\log\left(\frac{1}{2\epsilon}\sqrt{\frac{1-\pi(x)}{\pi(x)}}\right), t_{mix}^{(cont)}(x;\epsilon) \leq \frac{1}{\gamma}\log\left(\frac{1}{2\epsilon}\sqrt{\frac{1-\pi(x)}{\pi(x)}}\right),$$

and for any eigenvalue $\lambda \neq 1$,

$$t_{mix}(\epsilon) \geq \left(\frac{1}{1-|\lambda|} - 1\right)\log\frac{1}{2\epsilon} \text{ and in particular } t_{mix}(\epsilon) \geq (t_{rel} - 1)\log\frac{1}{2\epsilon}.$$

Proof. The upper bounds follow from Lemma 3.1 and Proposition 3.3. For the lower bounds, let $\lambda \neq 1$ be an eigenvalue with eigenfunction f. Then $\mathbb{E}_\pi f = \langle f, f_1\rangle_\pi = 0$. Let $x \in \Omega$ such that $f(x) = \|f\|_\infty$. Then

$$|\lambda^t f(x)| = |P^t f(x)| = \left|\sum_{y\in\Omega}[P^t(x,y) - \pi(y)]f(y)\right| \leq 2\|f\|_\infty d(t) \implies |\lambda|^t \leq 2d(t)$$

and hence

$$|\lambda|^{t_{mix}(\epsilon)} \leq 2d(t) \leq 2\epsilon \implies \log \frac{1}{2\epsilon} \leq t_{mix}(\epsilon) \log \frac{1}{|\lambda|} \leq t_{mix}(\epsilon) \left(\frac{1}{|\lambda|} - 1 \right).$$

□

With the reversibility assumption, the ℓ^2 distances to stationarity of the discrete and continuous-time chain bound each other.

Theorem 3.5. *Let (P, π) be reversible and $\lambda_- := \max\{0, -\lambda_n\}$. Then*

$$\|h_t^x - 1\|_2^2 \leq \frac{e^{2-t}}{\pi(x)} + \|p_x^{\lfloor t/2 \rfloor} - 1\|_2^2 \tag{8}$$

$$\|p_x^K - 1\|_2^2 \leq \lambda_-^{2k} \left(1 + \|h_{k'}^x - 1\|_2^2\right) + \|h_K^x - 1\|_2^2 \tag{9}$$

where $K = k + k' + 1$.

Proof. Write $\beta_j = 1 - \lambda_j$. Then $\lambda_j^K = e^{K \log(1-\beta_j)}$ for $K \in \mathbb{N}$. Using $\log(1-x) \geq -2x$ for $0 \leq x \leq 1/2$ we have $\lambda_j^{2K} \geq e^{-4K\beta_j}$ for $\lambda_j \geq \frac{1}{2}$, so

$$\begin{aligned}
\|p_x^K - 1\|_2^2 &= \sum_{j: \frac{1}{2} \leq \lambda_j < 1} \lambda_j^{2K} f_j(x)^2 + \sum_{j: \lambda_j < \frac{1}{2}} \lambda_j^{2K} f_j(x)^2 \geq \sum_{j: \frac{1}{2} \leq \lambda_j < 1} e^{-4K\beta_j} f_j(x)^2 \\
\implies \|h_{2K}^x - 1\|_2^2 &= \sum_{j=2}^n e^{-4K\beta_j} f_j(x)^2 \leq \|p_x^K - 1\|_2^2 + \sum_{j: \lambda_j < \frac{1}{2}} e^{-4K\beta_j} f_j(x)^2 \\
&\leq \|p_x^K - 1\|_2^2 + e^{-2K} \sum_{j=1}^n f_j(x)^2 = \|p_x^K - 1\|_2^2 + \frac{e^{-2K}}{\pi(x)}.
\end{aligned}$$

The first equation follows by setting $K = \lfloor \frac{t}{2} \rfloor$ for any $t > 0$. Next, note that

$$\sum_{j: \lambda_j < 0} \lambda_j^{2k'+2} f_j(x)^2 \leq - \sum_{j: \lambda_j < 0} \lambda_j^{2k'+1} f_j(x)^2 \leq \sum_{j: \lambda_j > 0} \lambda_j^{2k'+1} f_j(x)^2 \leq \sum_{j: \lambda_j > 0} \lambda_j^{2k'} f_j(x)^2$$

where the second inequality follows from expanding $p^{2k'+1}(x, x) \geq 0$. Finally, since $\lambda_j^{2K} = e^{2K \log(1-\beta_j)} \leq e^{-2K\beta_j}$ we have

$$\begin{aligned}
\|p_x^K - 1\|_2^2 &= \sum_{j: \lambda_j < 0} \lambda_j^{2K} f_j(x)^2 + \sum_{j: 0 < \lambda_j < 1} \lambda_j^{2K} f_j(x)^2 \\
&\leq \lambda_-^{2k} \sum_{j: \lambda_j < 0} \lambda_j^{2k'+2} f_j(x)^2 + \sum_{j: 0 < \lambda_j < 1} e^{-2K\beta_j} f_j(x)^2 \\
&\leq \lambda_-^{2k} \sum_{j: \lambda_j > 0} \lambda_j^{2k'} f_j(x)^2 + \sum_{j=2}^n e^{-2K\beta_j} f_j(x)^2 \\
&\leq \lambda_-^{2k} \left(1 + \|h_{k'}^x - 1\|_2^2\right) + \|h_K^x - 1\|_2^2.
\end{aligned}$$

□

4 Functional Inequalities

One powerful technique to upper bound the mixing time of a reversible Markov chain is through functional inequalities of Dirichlet forms. Intuitively speaking, we seek to control the global variation of random variables in the form of variance or entropy via their local variation called Dirichlet forms which encodes information about the Markov generator. Although many of our discussions may apply to non-reversible chains[4] [5], we restrict our attention to reversible chains henceforth.

4.1 Variance, Entropy and Dirichlet Form

We consider two versions of global variation of a random variable.

Definition 4.1. *Let π be a probability measure on Ω.*

1. *For $f : \Omega \to \mathbb{R}$ define the variance of f with respect to π by*

$$Var_\pi(f) := \mathbb{E}_\pi[(f - \mathbb{E}_\pi f)^2] = \mathbb{E}_\pi[f^2] - (\mathbb{E}_\pi f)^2$$

2. *For $f : \Omega \to (0, \infty)$ define the entropy of f with respect to π by*

$$Ent_\pi(f) := \mathbb{E}_\pi\left[f \log \frac{f}{\mathbb{E}_\pi f}\right] = \mathbb{E}_\pi[f \log f] - \mathbb{E}_\pi f \log \mathbb{E}_\pi f.$$

Lemma 4.1. *For $f : \Omega \to \mathbb{R}$ we have*

$$Var_\pi(f) = \frac{1}{2}\sum_{x\in\Omega}\sum_{y\in\Omega}\pi(x)\pi(y)(f(x) - f(y))^2.$$

Proof. Let g be an independent copy of f. Then

$$\begin{aligned}
\mathrm{Var}_\pi(f) &= \frac{1}{2}\mathrm{Var}(f-g) = \frac{1}{2}\mathbb{E}(f-g)^2 = \frac{1}{2}(\mathbb{E}f^2 + \mathbb{E}g^2 - 2\mathbb{E}f\mathbb{E}g)\\
&= \frac{1}{2}\left[\sum_{x\in\Omega}\pi(x)f^2(x) + \sum_{y\in\Omega}\pi(y)g^2(x) - 2\left[\sum_{x\in\Omega}\pi(x)f(x)\right]\left[\sum_{y\in\Omega}\pi(y)g(y)\right]\right]\\
&= \frac{1}{2}\sum_{x\in\Omega}\sum_{y\in\Omega}\pi(x)\pi(y)(f(x)-g(y))^2 = \frac{1}{2}\sum_{x\in\Omega}\sum_{y\in\Omega}\pi(x)\pi(y)(f(x)-f(y))^2.
\end{aligned}$$

□

Next we define the Dirichlet form which encodes the average local (co-)variation of two random variables while considering the Markov generator.

Definition 4.2. *Let $Q = P - I$ be a reversible Markov generator with stationary measure π. For $f, g : \Omega \to \mathbb{R}$ define the Dirichlet form*

$$\mathcal{E}_Q(f, g) := \langle -Qf, g\rangle_\pi = -\sum_{x\in\Omega}(Qf)(x)g(x)\pi(x).$$

When the chain in question is clear we write $\mathcal{E}(f, g)$ in place of $\mathcal{E}_Q(f, g)$.

Lemma 4.2. *For $f, g : \Omega \to \mathbb{R}$ we have*

$$\begin{aligned}\mathcal{E}(f,g) &= \frac{1}{2}\sum_{x\in\Omega}\sum_{y\in\Omega}\pi(x)Q(x,y)(g(x)-f(y))^2 \\ &= \frac{1}{2}\sum_{x\in\Omega}\sum_{y\in\Omega}\pi(x)Q(x,y)(f(x)-f(y))(g(x)-g(y))\end{aligned}$$

Proof. By definition,

$$\mathcal{E}(f,g) = -\sum_{x\in\Omega}\pi(x)(Qf)(x)g(x) = -\sum_{x\in\Omega}\sum_{y\in\Omega}\pi(x)Q(x,y)f(y)g(x).$$

The two formulas follow by applying the reversibility of Q to exchange x and y when appropriate and noting that each row of Q sums to 0. □

We draw a connection between total variation distance and entropy using Pinsker's inequality.

Definition 4.3. *Let μ_1 and μ_2 be two probability distributions on a discrete measurable space $(\Omega, 2^\Omega)$ such that $\mu_1 \ll \mu_2$. The Kullback-Leibler divergence or relative entropy from μ_2 to μ_1 is defined as*

$$D_{KL}(\mu_1\|\mu_2) := \mathbb{E}_{\mu_1}\left[\log\frac{d\mu_1}{d\mu_2}\right] = \sum_{x\in\Omega}\mu_1(x)\log\frac{\mu_1(x)}{\mu_2(x)},$$

where we adopt the convention that $0\log 0 = 0$.

Note that $D_{KL}(\mu_1\|\mu_2) \geq 0$ by Jensen's inequality. This quantity measures how μ_1 is different from the reference measure μ_2. In Information theory, μ_1 often represents observations (data) while μ_2 represents a model (theory) of how μ_1 should be distributed. Thus we are interested in how far the transition measure μ_t at time t is away from the stationary π.

Proposition 4.1 (Pinsker's inequality)**.** *Let μ_1 and μ_2 be two probability distributions. Then*

$$\|\mu_1 - \mu_2\|_{TV} \leq \sqrt{\frac{1}{2}D_{KL}(\mu_1\|\mu_2)}.$$

Proof. See Lemma 2.5 in Tsybakov[6]. □

It is clear that for a transition measure $\mu_t := \mu_0 P^t$ (discrete) or $\mu_t := \mu_0 H_t$ (continuous-time) with stationary measure π and density p_t we have

$$\mathrm{Ent}_\pi p_t = D_{KL}(\mu_t\|\pi) \implies \|\mu_t - \pi\|_{TV} \leq \sqrt{\frac{1}{2}\mathrm{Ent}_\pi p_t} \tag{10}$$

and by the proof of Lemma 3.1 we have

$$\|p_t - 1\|_2^2 = \mathrm{Var}_\pi p_t \implies \|\mu_t - \pi\|_{TV} \leq \frac{1}{2}\sqrt{\mathrm{Var}_\pi p_t}\,. \tag{11}$$

4.2 Inequalities and Constants

In this section, we introduce the Poincaré constant, log-Sobolev constant, and modified log-Sobolev constant of a Markov generator, which provide fundamental control over the behavior of the chain and in particular can be used to bound the mixing time. Unless otherwise stated, P will denote the transition matrix of a reversible chain and $Q = P - I$ will denote the associated Markov generator.

Definition 4.4. *Let (Q, π) be a reversible Markov chain.*

1. *The Poincaré constant of Q is defined as*

$$\lambda(Q) := \sup\{\lambda > 0 : \mathcal{E}_Q(f, f) \geq \lambda \, Var_\pi(f) \text{ for all } f : \Omega \to \mathbb{R}\}.$$

2. *The modified log-Sobolev constant of Q is defined as*

$$\alpha(Q) := \sup\{\alpha > 0 : \mathcal{E}_Q(f, \log f) \geq \alpha Ent_\pi(f) \text{ for all } f : \Omega \to \mathbb{R}_{>0}\}.$$

3. *The log-Sobolev constant of Q is defined as*

$$\rho(Q) := \sup\{\rho > 0 : \mathcal{E}_\pi(\sqrt{f}, \sqrt{f}) \geq \rho Ent_\pi(f) \text{ for all } f : \Omega \to \mathbb{R}_{>0}\}.$$

The inequalities defining λ, α, ρ are called the Poincaré inequality, the modified log-Sobolev inequality (MLSI) and the log-Sobolev inequality (LSI) respectively.

Note that $\mathrm{Var}(f + c) = \mathrm{Var} f$ and $\mathcal{E}_Q(f + c, f + c) = \mathcal{E}_Q(f, f)$ for constant c and hence in the definition of $\lambda(Q)$ we can require that $f > 0$ since Ω is finite.

Continuous-time chain may explain the motivation behind these functional inequalities[7] [5]. Recall from Equation (1) that $\mu_t := \mu H_t$ is the distribution of X_t given that the starting distribution is μ. Then the rate of change of the variance/entropy of $f_t := d\mu_t/d\pi$ is associate with the respective Dirichlet forms.

Proposition 4.2. *Let (Q, π) be a reversible continuous-time Markov chain. For an initial distribution μ on Ω define $f_t := d\mu H_t/d\pi = \mu H_t/\pi$ for $t \geq 0$. Then*

$$\frac{d}{dt} Var_\pi(f_t) = -2\mathcal{E}_Q(f_t, f_t), \ \ \frac{d}{dt} Ent_\pi(f_t) = -\mathcal{E}_Q(f_t, \log f_t).$$

As a result,

$$Var_\pi(f_t) \leq (Var_\pi f_0)\, e^{-2\lambda(Q)t}, \ \ Ent_\pi(f_t) \leq (Ent_\pi f_0)\, e^{-\alpha(Q)t}.$$

Proof. Note that we regard f_t as a column vector the $f_t' = Q f_t$ since (Q, π) is reversible. By definition of f_t we see that $\mathbb{E}_\pi[f_t] = 1$. Thus

$$\frac{d}{dt}\mathrm{Var}_\pi(f_t) = \frac{d}{dt}\mathbb{E}_\pi[f_t^2] = \sum_{y \in \Omega} \pi(y) \frac{d}{dt} f_t^2(y) = 2 \sum_{y \in \Omega} \pi(y) f_t(y) f_t'(y) = -2\mathcal{E}_Q(f_t, f_t)$$

and, using the fact that each row of Q sums to 0,

$$\begin{aligned} \frac{d}{dt}\mathrm{Ent}_\pi(f_t) &= \frac{d}{dt}\mathbb{E}_\pi[f_t \log f_t] = \sum_{y \in \Omega} \pi(y) \frac{d}{dt}[f_t(y) \log f_t(y)] \\ &= \sum_{y \in \Omega} \pi(y)[1 + \log f_t(y)] f_t'(y) = -\mathcal{E}_Q(f_t, \log f_t). \end{aligned}$$

Thus

$$\frac{d}{dt}\mathrm{Var}_\pi(f_t) = -2\mathcal{E}_\pi(f_t, f_t) \leq -2\lambda(Q)\mathrm{Var}_\pi(f_t) \implies \frac{d}{dt}\log\mathrm{Var}_\pi(f_t) \leq -2\lambda(Q)$$
$$\implies \log\mathrm{Var}_\pi(f_t) - \log\mathrm{Var}_\pi(f_0) \leq -2\lambda(Q)t \implies \mathrm{Var}_\pi(f_t) \leq \mathrm{Var}_\pi(f_0)e^{-2\lambda(Q)t}$$

and

$$\frac{d}{dt}\mathrm{Ent}_\pi(f_t) = -\mathcal{E}_\pi(f_t, \log f_t) \leq -\alpha(Q)\mathrm{Ent}_\pi(f_t) \implies \mathrm{Ent}_\pi(f_t) \leq \mathrm{Ent}_\pi(f_0)e^{-\alpha(Q)t},$$

as required. □

We see that $\lambda(Q)$ and $\alpha(Q)$ are defined to control the exponential decay of variance and entropy of f_t in light of the differential equations they satisfy. The log-Sobolev inequality also has an analogous result, but we introduce a lemma first.

Lemma 4.3. [4] *Under a reversible chain, for any $f > 0$ we have*

$$\mathcal{E}(\log f, f) \geq 4\mathcal{E}(\sqrt{f}, \sqrt{f}).$$

Proof. For $0 < b < a$, applying Hölder's inequality on the probability space $[a, b]$ equipped with the normalized Lebesgue measure gives

$$\left(\frac{\sqrt{a}-\sqrt{b}}{a-b}\right)^2 = \left(\frac{1}{2(a-b)}\int_b^a \frac{dt}{\sqrt{t}}\right)^2 \leq \frac{1}{4(a-b)}\int_b^a \frac{dt}{t} = \frac{1}{4}\frac{\log a - \log b}{a-b}.$$

By symmetry, for all $a, b > 0$ we have $4(\sqrt{a}-\sqrt{b})^2 \leq (\log a - \log b)(a-b)$. The inequality then follows from the two expressions in Lemma 4.2 □

Proposition 4.3. [4] *With the same setting as in Proposition 4.2 we have*

$$Ent_\pi(f_t) \leq Ent_\pi(f_0)e^{-4\rho(Q)t}.$$

Proof. By Proposition 4.2 and Lemma 4.3,

$$\frac{d}{dt}\mathrm{Ent}_\pi(f_t) = -\mathcal{E}(f_t, \log f_t) \leq -4\mathcal{E}(\sqrt{f_t}, \sqrt{f_t}) \leq -4\rho(Q)\mathrm{Ent}_\pi(f_t),$$

where the last inequality follows from the definition of $\rho(Q)$. □

With Propositions 4.2 and 4.3 the mixing time of continuous chains can be easily bounded from above, provided that the constant in question are known. In many applications, however, computing these constants is very difficult.

Theorem 4.1. *Let (Q, π) be a continuous-time reversible chain on a finite set Ω. Then for $x \in \Omega$ and $\epsilon > 0$ we have*

$$t_{mix}^{cont}(x;\epsilon) \leq \frac{1}{2\lambda(Q)}\left(\log\frac{1-\pi(x)}{\pi(x)} + \log\frac{1}{4\epsilon^2}\right) \tag{12}$$

$$t_{mix}^{cont}(x;\epsilon) \leq \frac{1}{\alpha(Q)}\left(\log\log\frac{1}{\pi(x)} + \log\frac{1}{2\epsilon^2}\right) \tag{13}$$

$$t_{mix}^{cont}(x;\epsilon) \leq \frac{1}{4\rho(Q)}\left(\log\log\frac{1}{\pi(x)} + \log\frac{1}{2\epsilon^2}\right) \tag{14}$$

Proof. For $x \in \Omega$, setting $f_0 = \delta_x/\pi$ we have $\mathbb{E}_\pi f_0 = 1$ and hence

$$\mathrm{Var}_\pi f_0 = \mathbb{E}_\pi[f_0^2] - 1 = \frac{1}{\pi(x)} - 1, \ \mathrm{Ent}_\pi f_0 = \mathbb{E}_\pi\left[f_0 \log f_0\right] = \log\frac{1}{\pi(x)}.$$

By Equations (10), (11) and Propositions 4.2, 4.3,

$$\|H_t(x,\cdot) - \pi\|_{TV} \leq \frac{1}{2}\sqrt{\mathrm{Var}_\pi f_t} \leq \frac{1}{2}\sqrt{\frac{1-\pi(x)}{\pi(x)}e^{-2\lambda(Q)t}}$$

$$\|H_t(x,\cdot) - \pi\|_{TV} \leq \sqrt{\frac{1}{2}\mathrm{Ent}_\pi f_t} \leq \sqrt{\frac{e^{-\alpha(Q)t}}{2}\log\frac{1}{\pi(x)}}$$

$$\|H_t(x,\cdot) - \pi\|_{TV} \leq \sqrt{\frac{1}{2}\mathrm{Ent}_\pi f_t} \leq \sqrt{\frac{e^{-4\rho(Q)t}}{2}\log\frac{1}{\pi(x)}}$$

Solving for the right hand side of the three expression $\leq \epsilon$ and recall the definition of $t_{mix}(x;\epsilon)$ we get the results. □

Of course, to get the bound for the whole chain we take the maximum across $x \in \Omega$. Observe that the dependence on π for LSI and MLSI is much better than that for Poincaré inequality. The constants in the denominator also form a hierarchy.

Theorem 4.2. *For a reversible chain, the Poincaré constant λ, modified log-Sobolev constant α and log-Sobolev constant ρ satisfy* $0 \leq 4\rho \leq \alpha \leq 2\lambda$.

Proof. $\rho \geq 0$ by definition and $4\rho \leq \alpha$ follows from Lemma 4.3. To show $\alpha \leq 2\lambda$, fix any $f : \Omega \to \mathbb{R}$. By the definition of α, for any $c > 0$ we have

$$\mathrm{Ent}\left[e^{cf}\right] = \mathbb{E}\left[cfe^{cf}\right] - \mathbb{E}\left[e^{cf}\right]\log\mathbb{E}\left[e^{cf}\right] \leq \alpha\mathcal{E}(cf, e^{cf})$$

The linearity of $\mathcal{E}(\cdot,\cdot)$ gives

$$\mathcal{E}(cf, e^{cf}) = \mathcal{E}(cf, 1 + cf + o(c)) = c^2\mathcal{E}(f,f) + o(c^2)$$

while the Taylor expansion of e^x gives

$$\mathbb{E}\left[cfe^{cf}\right] = \mathbb{E}[cf + c^2f^2 + o(c^2)] = c\mathbb{E}[f] + c^2\mathbb{E}[f^2] + o(c^2)$$

and

$$\mathbb{E}\left[e^{cf}\right] = \sum_{k=0}^{\infty}\frac{c^k}{k!}\mathbb{E}[f^k] = 1 + c^2\mathbb{E}[f^2]/2 + o(c^2).$$

Using $\log(1+x) = x - x^2/2 + o(x)$ for $|x| < 1$, we see that for c sufficiently small,

$$\mathbb{E}\left[e^{cf}\right]\log\mathbb{E}\left[e^{cf}\right] = c\mathbb{E}[f] + \frac{c^2}{2}\left(\mathbb{E}[f^2] - (\mathbb{E}[f])^2\right) + o(c^2)$$

and hence

$$\mathrm{Ent}\left[e^{cf}\right] = \frac{c^2}{2}\mathrm{Var}f + o(c^2) \leq \alpha c^2\mathcal{E}(f,f) + o(c^2).$$

Therefore $\mathrm{Var}f \leq 2\alpha\mathcal{E}(f,f) + o(c^2)/c^2$ and letting $c \searrow 0$ finishes the proof. □

It turns out that the spectral gap is the Poincaré constant for a reversible chain and therefore controls the mixing time. Thus the bound given in Theorem 4.1 is in accordance with Theorem 3.4.

Theorem 4.3. *If γ is the spectral gap of a reversible chain (Q, π), then*

$$\gamma = \min_{\substack{f:\Omega\to\mathbb{R} \\ \mathbb{E}_\pi f=0, \|f\|_2=1}} \mathcal{E}(f,f) = \min_{\substack{f:\Omega\to\mathbb{R} \\ \mathbb{E}_\pi f=0, \|f\|_2\neq 0}} \frac{\mathcal{E}(f,f)}{\|f\|_2} = \min_{\substack{f:\Omega\to\mathbb{R} \\ \|f\|_2\neq 0}} \frac{\mathcal{E}(f,f)}{Var_\pi f},$$

and the minimum is attained by $f = f_2$ in Proposition 3.2.

Proof. By scaling and shifting it suffices to prove the first equality. The condition $\mathbb{E}_\pi f = 0$ means that $f \perp f_1$, so $f = \sum_{j=2}^n \langle f, f_j\rangle_\pi f_j$. The condition $\|f\|_2 = 1$ implies

$$\begin{aligned}
\mathcal{E}(f,f) = \langle -Qf, f\rangle_\pi &= \left\langle \sum_{j=2}^n (1-\lambda_j)\langle f, f_j\rangle_\pi f_j, \sum_{j=2}^n \langle f, f_j\rangle_\pi f_j \right\rangle_\pi \\
&= \sum_{j=2}^n (1-\lambda_j)\langle f, f_j\rangle_\pi^2 \le (1-\lambda_2)\sum_{j=2}^n \langle f, f_j\rangle_\pi^2 = (1-\lambda_2)\|f\|_2^2 = \gamma,
\end{aligned}$$

and note that the inequality above is actually an equality when $f = f_2$. □

4.3 Examples

Example 4.1. We first compute λ, α and ρ for two-state chains, which are the easiest non-degenerate chains. Note that a two-state chain is necessarily reversible.

Theorem 4.4. *For any $a, b > 0$, if the Markov generator is of the form*

$$Q = \begin{bmatrix} -a & a \\ b & -b \end{bmatrix},$$

then $\lambda(Q) = a + b$, $\alpha(Q) \in [a + b + 2\sqrt{ab}, 2(a + b)]$, and

$$\rho(Q) = \begin{cases} \frac{a-b}{\log a - \log b} & \text{if } a \neq b, \\ a & \text{if } a = b. \end{cases}$$

Proof. Note that the stationary distribution is given by $\pi = \begin{bmatrix} \frac{b}{a+b} & \frac{a}{a+b} \end{bmatrix}$. Using Lemmas 4.1 and 4.2 we see that $\lambda(Q) = a + b$. Let $\Omega = \{x, y\}$ denote the state space of Q where $Q(x, y) = a$.

Consider the case $a = b$ and hence $\pi(x) = \pi(y) = 1/2$. Let $f : \Omega \to \mathbb{R}_{>0}$. Since $\mathcal{E}(\sqrt{f/\mathbb{E}f}, \sqrt{f/\mathbb{E}f}) = \mathcal{E}(\sqrt{f}, \sqrt{f})/\mathbb{E}f$ and $\mathrm{Ent}(f/\mathbb{E}f) = \mathrm{Ent}(f)/\mathbb{E}f$ we may assume that $\mathbb{E}f = 1$, i.e. $f(x) = 1 + s$, $f(y) = 1 - s$ for some $-1 < s < 1$. Thus

$$\mathcal{E}(\sqrt{f}, \sqrt{f}) = \frac{a}{2}\left(\sqrt{1+s} - \sqrt{1-s}\right)^2, \ \mathrm{Ent} f = \frac{1}{2}[(1+s)\log(1+s) + (1-s)\log(1-s)].$$

Let $g(s) = \left(\sqrt{1+s} - \sqrt{1-s}\right)^2$ and $h(s) = (1 + s)\log(1 + s) + (1 - s)\log(1 - s)$ for $-1 < s < 1$. Then

$$g'(s) = \frac{2s}{1-s^2}, \ g''(s) = \frac{2}{(1-s^2)^{3/2}}, \ h'(s) = \log\frac{1+s}{1-s}, \ h''(s) = \frac{2}{1-s^2}.$$

Notice that $g(0) = h(0) = g'(0) = h'(0) = 0$, $g''(0) = h''(0) = 2$ and $g''(s) > h''(s)$ for $s \neq 0$. Thus by L'Hôpital's rule the infimum of g/h is $g''(0)/h''(0) = 1$ and therefore $\rho = a$. If $a \neq b$, a tedious calculus argument with numerical insights (see Diaconis & Saloff-Coste[4]) shows that $\rho = (a-b)/(\log a - \log b)$.

By Theorem 4.2, $\alpha \leq 2\lambda = 2(a+b)$. Next we show that $\alpha \geq a + b + 2\sqrt{ab}$. Since $\mathcal{E}(cf, \log cf)/\mathrm{Ent}[cf] = \mathcal{E}(f, \log f)/\mathrm{Ent}f$ we may assume that $f(x) = t$ and $f(y) = 1$ for some $t > 0$. Let

$$F(t) := \mathcal{E}(f, \log f) - \alpha \mathrm{Ent}f = \frac{ab}{a+b}(t-1)\log t - \alpha\left[\frac{bt\log t}{a+b} - \frac{a+bt}{a+b}\log\frac{a+bt}{a+b}\right].$$

Then

$$F'(t) = \frac{ab}{a+b}\left[\log t + 1 - \frac{1}{t}\right] + \frac{\alpha b}{a+b}\log\frac{a+bt}{(a+b)t}$$

$$F''(t) = \frac{ab}{a+b}\left[\frac{t+1}{t^2} - \frac{\alpha}{t(a+bt)}\right] = \frac{ab}{a+b}\frac{bt^2 + (a+b-\alpha)t + a}{t^2(a+bt)}.$$

Note that $F''(t) \geq 0$ for all $t > 0$ if $(a+b-\alpha)^2 - 4ab \geq 0$ and in particular if $\alpha \geq a+b+2\sqrt{ab}$. Since $F(1) = F'(1) = 0$, it follows that $a+b+2\sqrt{ab}$ satisfies the MLSI and hence is a lower bound for α. □

Example 4.2. (Complete graph) Let Ω be a complete graph of n vertices on which π is a probability distribution supported on the entire Ω. Consider the chain $P(x,y) = \pi(y)$ which is reversible under π. Observe that for this chain we have $\mathcal{E}(f,g) = \mathbb{E}[fg] - \mathbb{E}[f]\mathbb{E}[g]$.

The Poincaré constant is trivially 1 as $\mathcal{E}(f,f) = \mathrm{Var}f$ for any $f : \Omega \to \mathbb{R}$. By Jensen's inequality, for $f > 0$ we have

$$\mathcal{E}(f, \log f) = \mathbb{E}[f\log f] - \mathbb{E}[f]\mathbb{E}[\log f] \geq \mathbb{E}[f\log f] - \mathbb{E}[f]\log\mathbb{E}[f] = \mathrm{Ent}f,$$

so the MLS constant satisfies $1 \leq \alpha \leq 2\lambda = 2$. The log-Sobolev constant ρ is given by the following theorem. (See Appendix A in Diaconis & Saloff-Coste[4] for proof).

Theorem 4.5. *Let π be a probability on Ω with $\pi_* = \min_{x\in\Omega}\pi(x) > 0$ and define the chain $P(x,y) := \pi(y)$. Then its log-Sobolev constant satisfies*

$$\rho = \frac{1-2\pi_*}{\log(1/\pi_* - 1)}$$

Corollary 4.1. *Let Q be an irreducible Markov generator on a finite set. Then*

$$\rho(Q) \geq \frac{(1-2\pi_*)\lambda(Q)}{\log(1/\pi_* - 1)}$$

Proof. It suffices to observe that $\mathcal{E}_P(f,f) = \mathrm{Var}f$ for the chain $P(x,y) := \pi(y)$ and apply Theorem 4.5 and the definition of $\lambda(Q)$. □

Alternatively, we can bound ρ from below using the following Theorem.

Theorem 4.6. *There are universal constants $0 < c < 1$ and $C > 0$ satisfying the following property. Let (Q,π) be any reversible Markov chain on Ω. If $K > 0$ is such that f satisfies LSI (respectively MSLI) with constant K for all functions f on Ω bounded below by c with $\mathbb{E}_\pi f = 1$, then $\alpha(Q) \geq CK$ (resp. $\rho(Q) \geq CK$).*

Proof. A proof for MLSI was given by Tikhomirov and Youssef[8], but the same argument applies to LSI as well. The idea is to define for any nonconstant positive function f with $\mathbb{E}_\pi f = 1$ an auxiliary function g by

$$g(x) = \begin{cases} \max(f(x), c) & \text{if } f(x) \leq 1 \\ b + (1-b)f(x) & \text{if } f(x) > 1, \end{cases}$$

where $b \in [0,1)$ is chosen so that $\mathbb{E}_\pi g = 1$. By choosing c sufficiently small, one can find $C > 0$ such that $\mathrm{Ent}_\pi g \geq C\mathrm{Ent}_\pi f$, where the choice of c and the corresponding C are universal across all positive functions f with $\mathbb{E}_\pi f = 1$. It remains to observe that for $x, y \in \Omega$ such that $f(x) \geq f(y)$ we have

$$g(x) - g(y) \leq f(x) - f(y), \sqrt{g(x)} - \sqrt{g(y)} \leq \sqrt{f(x)} - \sqrt{f(y)}, \text{ and } \frac{g(x)}{g(y)} \leq \frac{f(x)}{f(y)}$$

and thus $\mathcal{E}_Q(g, \log g) \leq \mathcal{E}_Q(f, \log f)$, $\mathcal{E}_Q(\sqrt{g}, \sqrt{g}) \leq \mathcal{E}_Q(\sqrt{f}, \sqrt{f})$. Therefore

$$\begin{aligned} CK\mathrm{Ent}f &\leq K\mathrm{Ent}g \leq \mathcal{E}_Q(g, \log g) \leq \mathcal{E}_Q(f, \log f) \\ CK\mathrm{Ent}f &\leq K\mathrm{Ent}g \leq \mathcal{E}_Q(\sqrt{g}, \sqrt{g}) \leq \mathcal{E}_Q(\sqrt{f}, \sqrt{f}) \end{aligned}$$

and the results follow. □

Now let $0 < c < 1$ and $C > 0$ be given by Theorem 4.6. Suppose f is bounded below by c with $\mathbb{E}_\pi f = 1$. Let X and Y be independent Ω-valued random variables with distribution π. Then

$$\mathcal{E}(f, \log f) = \frac{1}{2}\mathbb{E}\left[(f(X) - f(Y))\log\frac{f(X)}{f(Y)}\right] = \frac{1}{2}\mathbb{E}\left[(\sqrt{f(X)} - \sqrt{f(Y)})^2 H(Z)\right]$$

where

$$H(t) = \frac{\sqrt{t}+1}{\sqrt{t}-1}\log t, \ Z = \frac{f(X)}{f(Y)}$$

is increasing for $t \geq 1$, attains its minimum at $t = 1$ and satisfies $H(t) = H(1/t)$. By assumption $c \leq f \leq 1/\pi_*$, so $c\pi_* \leq Z \leq (c\pi_*)^{-1}$ and hence

$$\mathrm{Ent}f \leq \mathcal{E}(f, \log f) \leq H(c\pi_*)\mathrm{Var}\sqrt{f} = H(c\pi_*)\mathcal{E}(\sqrt{f}, \sqrt{f}).$$

Therefore, by Theorem 4.6 we have

$$\rho \geq \frac{C}{H(c\pi_*)} = \frac{C(1 - \sqrt{c\pi_*})}{(1 + \sqrt{c\pi_*})\log\frac{1}{c\pi_*}}.$$

5 Decomposition of Markov Chains

Many Markov chains have structures that can be naturally viewed as a collection of several Markov chains on subsets of the state space. If we decompose the state space Ω of a chain Q into classes, we can define the associated restriction chains and projection chain in a natural way, the former being the chains within each class while the latter being the chain among classes. As it turns out, with some assumptions the Poincaré, log-Sobolev and modified log-Sobolev constants of Q can be estimated from those of its projection and restriction chains. As we saw in the previous section, computing λ, α and ρ is very difficult in general. The merit of this type of decomposition methods is that the projection and restriction chains are often simpler to study, and in some cases when the chain has a recursive structure we can solve for lower bounds of the three constants.

5.1 A Method via Coupling of Classes

Here we introduce a result by Lu & Yau[9] following the notation in Hermon and Salez[10]. Let Q be an irreducible Markov generator reversible with respect to the stationary measure π on a finite state space Ω. Suppose we are given a partition on the state space $\Omega = \coprod_{i\in I} \Omega_i$ where Ω_i $(i \in I)$ are disjoint. The projection chain is the chain with state space I and the Markov generator $\hat{Q}$ defined as follows: for $i, j \in I$

$$\hat{Q}(i,j) = \frac{1}{\hat{\pi}(i)} \sum_{x\in\Omega_i} \sum_{y\in\Omega_j} \pi(x)Q(x,y) \tag{15}$$

where $\hat{\pi}(i) = \sum_{x\in\Omega_i} \pi(x)$ is a probability distribution on I which is reversible under $\hat{Q}$. The restriction chain on Ω_i is the chain with state space Ω_i and Markov generator Q_i, where $Q_i(x,y) = Q(x,y)$ if $x \neq y$ and $Q(x,x)$ are such that each row sums to 0. Thus $\pi_i := \pi/\hat{\pi}(i)$ is a probability distribution on Ω_i which is reversible under Q_i.

Now suppose that we are given couplings κ_{ij} of π_i and π_j for $(i,j) \in I^2$ with $\hat{Q}(i,j) > 0$. We define the quality of the couplings by

$$\chi := \min_{\substack{(x,y,i,j) \text{ s.t.} \\ \text{denominator}>0}} \frac{\pi(x)Q(x,y)}{\hat{\pi}(i)\hat{Q}(i,j)\kappa_{ij}(x,y)}. \tag{16}$$

Theorem 5.1 (Recursive functional inequalities)**.** *Given an irreducible reversible chain Q on a finite set $\Omega = \coprod_{i\in I} \Omega_i$. Suppose we are given couplings κ_{ij} of π_i, π_j for $i, j \in I$ with $\hat{Q}(i,j) > 0$ and let χ be their quality. Then*

$$\lambda(Q) \geq \min\{\chi\lambda(\hat{Q}), \min_{i\in I} \lambda(Q_i)\}$$

$$\alpha(Q) \geq \min\{\chi\alpha(\hat{Q}), \min_{i\in I} \alpha(Q_i)\}$$

$$\rho(Q) \geq \min\{\chi\rho(\hat{Q}), \min_{i\in I} \rho(Q_i)\}$$

Remark 5.1. This result is practical only when one can find couplings with $\chi > 0$, and it is necessary that for all i, j such that $\hat{Q}(i,j) > 0$ and for all $x \in \Omega_i$ and $y \in \Omega_j$ we have $\kappa_{ij}(x,y) > 0 \implies Q(x,y) > 0$.

Proof. Let $f : \Omega \to (0, \infty)$. By Lemma 4.2, the Dirichlet forms associated with the three constants are

$$\mathcal{L}_\pi(f) := \frac{1}{2} \sum_{(x,y)\in\Omega^2} \pi(x)Q(x,y)\Psi(f(x), f(y)), \text{ where}$$

$$\Psi(u,v) = \begin{cases} (u-v)^2 & \text{for Poincare} \\ (u-v)(\log u - \log v) & \text{for modified log-Sobolev} \\ (\sqrt{u} - \sqrt{v})^2 & \text{for log-Sobolev} \end{cases}$$

and write

$$\mathcal{R}_\pi(f) := \begin{cases} \mathrm{Var}_\pi(f) & \text{for Poincare} \\ \mathrm{Ent}_\pi(f) & \text{for modified log-Sobolev and log-Sobolev.} \end{cases}$$

Define $\hat{f} : I \to (0, \infty)$ by $\hat{f}(i) := \mathbb{E}_{\pi_i} f$. Writing $A(x,y) := Q(x,y)\Psi(f(x), f(y))$,

$$\begin{aligned}
\mathcal{L}_\pi(f) &= \frac{1}{2} \sum_{i,j\in I} \sum_{x\in\Omega_i} \sum_{y\in\Omega_j} \pi(x)A(x,y) \\
&= \frac{1}{2} \sum_{i\in I} \sum_{x,y\in\Omega_i} \pi(x)A(x,y) + \frac{1}{2} \sum_{i\in I} \sum_{j\neq i} \sum_{x\in\Omega_i} \sum_{y\in\Omega_j} \pi(x)A(x,y) \\
&= \frac{1}{2} \sum_{i\in I} \hat{\pi}(i) \sum_{x,y\in\Omega_i} \pi_i(x)A(x,y) + \frac{1}{2} \sum_{i\in I} \sum_{j\neq i} \sum_{x\in\Omega_i} \sum_{y\in\Omega_j} \pi(x)A(x,y) \\
&= \sum_{i\in I} \hat{\pi}(i)\mathcal{L}_{\pi_i}(f) + \frac{1}{2} \sum_{i\in I} \sum_{j\neq i} \sum_{x\in\Omega_i} \sum_{y\in\Omega_j} \pi(x)Q(x,y)\Psi(f(x), f(y)).
\end{aligned}$$

For the variance, by Lemma 4.1 we have

$$\mathrm{Var}_\pi(f) = \frac{1}{2} \sum_{i,j\in I} \sum_{\substack{x\in\Omega_i \\ y\in\Omega_j}} \pi(x)\pi(y)[(f(x) - \hat{f}(i)) + (\hat{f}(i) - \hat{f}(j)) + (\hat{f}(j) - f(y))]^2.$$

Expanding the square, we find that the cross terms involving $f(x) - \hat{f}(i)$ or $\hat{f}(j) - f(y)$ are all 0 by the definition of $\hat{f}$. Hence

$$\begin{aligned}
\mathrm{Var}_\pi(f) &= \frac{1}{2} \sum_{i,j\in I} \sum_{\substack{x\in\Omega_i \\ y\in\Omega_j}} \pi(x)\pi(y)[(f(x) - \hat{f}(i))^2 + (\hat{f}(j) - f(y))^2 + (\hat{f}(i) - \hat{f}(j))^2] \\
&= \sum_{i,j\in I} \sum_{\substack{x\in\Omega_i \\ y\in\Omega_j}} \pi(x)\pi(y)(f(x) - \hat{f}(i))^2 + \frac{1}{2} \sum_{i,j\in I} \hat{\pi}(i)\hat{\pi}(j)(\hat{f}(i) - \hat{f}(j))^2 \\
&= \sum_{i,j\in I} \hat{\pi}(i)\hat{\pi}(j) \sum_{x\in\Omega_i} \pi_i(x)(f(x) - \hat{f}(i))^2 + \mathrm{Var}_{\hat{\pi}}(\hat{f}) \\
&= \sum_{i\in I} \hat{\pi}(i)\mathrm{Var}_{\pi_i}(f) + \mathrm{Var}_{\hat{\pi}}(\hat{f})
\end{aligned}$$

Similarly, for the entropy we have

$$\begin{aligned}
\mathrm{Ent}_\pi f &= \mathbb{E}_\pi[f\log f] - (\mathbb{E}_\pi f)\log \mathbb{E}_\pi f \\
&= \sum_{i\in I}\hat\pi(i)\sum_{x\in\Omega_i}\pi_i(x)f(x)[\log f(x) - \log \mathbb{E}_\pi f] \\
&= \sum_{i\in I}\hat\pi(i)\mathrm{Ent}_{\pi_i} f + \sum_{i\in I}\sum_{x\in\Omega_i}\pi(x)f(x)[\log \mathbb{E}_{\pi_i} f - \log \mathbb{E}_\pi f] \\
&= \sum_{i\in I}\hat\pi(i)\mathrm{Ent}_{\pi_i} f + \sum_{i\in I}\hat\pi(i)\hat f(i)[\log \hat f(i) - \log \mathbb{E}_{\hat\pi}\hat f] \\
&= \sum_{i\in I}\hat\pi(i)\mathrm{Ent}_{\pi_i} f + \mathrm{Ent}_{\hat\pi}\hat f.
\end{aligned}$$

In summary,

$$\begin{aligned}
\mathcal{R}_\pi(f) &= \sum_{i\in I}\hat\pi(i)\mathcal{R}_{\pi_i} f + \mathcal{R}_{\hat\pi}\hat f \qquad (17)\\
\mathcal{L}_\pi(f) &= \sum_{i\in I}\hat\pi(i)\mathcal{L}_{\pi_i}(f) + \frac{1}{2}\sum_{i\in I}\sum_{j\neq i}\sum_{x\in\Omega_i}\sum_{y\in\Omega_j}\pi(x)Q(x,y)\Psi(f(x),f(y)).
\end{aligned}$$

We do a term-by-term comparison between the two. Since κ_{ij} is a coupling of π_i and π_j we have $(\hat f(i),\hat f(j)) = \mathbb{E}_{\kappa_{ij}}F$ where $F(x,y) = (f(x),f(y))$. Since Ψ is convex in all three cases, Jensen's inequality tells us that

$$\sum_{x\in\Omega_i}\sum_{y\in\Omega_j}\kappa_{ij}(x,y)\Psi(f(x),f(y)) \geq \Psi(\hat f(i),\hat f(j)),$$

and since $\chi\hat\pi(i)\hat Q(i,j)\kappa_{ij}(x,y) \leq \pi(x)Q(x,y)$ for $i\neq j$ by the definition of χ, multiplying the previous inequality by $\chi\hat\pi(i)\hat Q(i,j)$ yields

$$\sum_{x\in\Omega_i}\sum_{y\in\Omega_j}\pi(x)Q(x,y)\Psi(f(x),f(y)) \geq \chi\hat\pi(i)\hat Q(i,j)\Psi(\hat f(i),\hat f(j))$$

and therefore

$$\begin{aligned}
\mathcal{L}_\pi(f) &\geq \sum_{i\in I}\hat\pi(i)\mathcal{L}_{\pi_i}(f) + \frac{1}{2}\sum_{i\in I}\sum_{j\neq i}\chi\hat\pi(i)\hat Q(i,j)\Psi(\hat f(i),\hat f(j)) \\
&= \sum_{i\in I}\hat\pi(i)\mathcal{L}_{\pi_i}(f) + \chi\mathcal{L}_{\hat\pi}(\hat f). \qquad (18)
\end{aligned}$$

By (17) and (18), for all $\hat\tau,\tau_i\geq 0$ such that $\mathcal{L}_{\hat\pi}(\hat f)\geq\hat\tau\mathcal{R}_{\hat\pi}(\hat f)$ and $\mathcal{L}_{\pi_i}(f)\geq\tau_i\mathcal{R}_{\pi_i}(f)$ for all $i\in I$, we have

$$\mathcal{L}_\pi(f) \geq \sum_{i\in I}\hat\pi(i)\tau_i\mathcal{R}_{\pi_i}(f) + \chi\hat\tau\mathcal{R}_{\hat\pi}(\hat f) \geq \min\{\chi\hat\tau, \min_{i\in I}\tau_i\}\mathcal{R}_\pi(f),$$

which finishes the proof. □

Example 5.1. (Hypercube) Let $\Omega := \{0,1\}^n$ be the n-dimensional hypercube and $\Omega_i := \{x\in\Omega \mid x_n = i\}$ be the set of vertices whose last coordinate is i $(i = 0,1)$.

Then $\Omega = \Omega_0 \cup \Omega_1$ is a partition of Ω. For $x, y \in \Omega$ we write $x \sim y$ if x and y differ by exactly one coordinate. For $0 < r \leq 1/n$ define the chain $Q := Q[n, r]$ by

$$Q(x,y) = \begin{cases} -nr & \text{if } x = y; \\ r & \text{if } x \sim y; \\ 0 & \text{otherwise,} \end{cases}$$

which is reversible under the stationary measure $\pi(x) = 2^{-n}$ $(x \in \Omega)$. It is clear that $\hat{\pi}(0) = \hat{\pi}(1) = 1/2$, $\pi_i(x) = 1/2^{n-1}$ and the projection chain is

$$\hat{Q}(0,1) = \hat{Q}(1,0) = r = -\hat{Q}(0,0) = -\hat{Q}(1,1)$$

In this example we only need couplings κ_{01} and κ_{10}. For $(x, y) \in \Omega_0 \times \Omega_1$ we define $\kappa_{01}(x, y) = 1/2^{n-1}$ if $x \sim y$ and 0 otherwise, and $\kappa_{10}(y, x) := \kappa_{01}(x, y)$. Thus $\chi = 1$. By Theorem 4.4, $\lambda(\hat{Q}) = 2r$ and $\rho(\hat{Q}) = r$. Notice that the restriction chains satisfy

$$(Q[n,r])_i = Q[n-1,r], \ \ i = 0, 1$$

so we can repeatedly apply Theorem 5.1 and get

$$\lambda(Q[n,r]) \geq \min\{2r, \lambda(Q[1,r])\} = 2r, \ \rho(Q[n,r]) \geq \min\{r, \rho(Q[1,r])\} = r$$

In terms of mixing time, $\lambda(Q)$ gives $t_{mix} = O(n^2)$ while $\rho(Q)$ gives $t_{mix} = O(n \log n)$. As we noted in Example 3.1, the log-Sobolev constant gives the correct order of mixing time.

5.2 A Generalization

We generalize the aforementioned partition technique as follows. Let Q be a reversible chain on Ω and I be a finite set. To each state $x \in \Omega$ we assign a probability distribution $a(x) := (a_i(x))_{i \in I}$. In the previous discussion, $a(x) = \delta_i$ for some $i \in I$, but now we allow $a(x)$ to be any stochastic vector. We hope to find natural generalizations of the restriction and the projection chains so that we can obtain a recursive result similar to Theorem 5.1.

We start by rewriting the projection chain. For $i, j \in I$, using the old notation we have

$$\hat{Q}(i,j) = \frac{1}{\hat{\pi}(i)} \sum_{x \in \Omega_i} \sum_{y \in \Omega_j} \pi(x) Q(x,y) = \frac{1}{\sum_{x \in \Omega} a_i(x)\pi(x)} \sum_{x,y \in \Omega} a_i(x) a_j(y) \pi(x) Q(x,y).$$

Let $\Lambda_i := \{x \in \Omega \mid a_i(x) > 0\}$. We now redefine

$$\hat{\pi}(i) := \sum_{x \in \Lambda_i} a_i(x)\pi(x),$$

for $i \neq j$

$$\hat{Q}(i,j) := \frac{1}{\hat{\pi}(i)} \sum_{x \in \Lambda_i} \sum_{y \in \Lambda_j, y \neq x} a_i(x) a_j(y) \pi(x) Q(x,y),$$

and finally define $\hat{Q}(i,i)$ such that each row of $\hat{Q}$ sums to 0. We verify that $\hat{Q}(i,j)$ is well-defined: for $i \in I$ we have

$$
\begin{aligned}
\sum_{j\in I, j\neq i} \hat{Q}(i,j) &= \frac{1}{\hat{\pi}(i)} \sum_{x\in\Omega} a_i(x)\pi(x) \sum_{y\in\Omega, y\neq x} \sum_{j\in I, j\neq i} a_j(y)Q(x,y) \\
&= \frac{1}{\hat{\pi}(i)} \sum_{x\in\Omega} a_i(x)\pi(x) \sum_{y\in\Omega, y\neq x} (1-a_i(y))Q(x,y) \\
&\leq \frac{1}{\hat{\pi}(i)} \sum_{x\in\Omega} a_i(x)\pi(x) \sum_{y\in\Omega, y\neq x} Q(x,y) \leq \frac{1}{\hat{\pi}(i)} \sum_{x\in\Omega} a_i(x)\pi(x) = 1.
\end{aligned}
$$

As for the restriction chains, for $i \in I$ and $x, y \in \Lambda_i$ define

$$
\pi_i(x) := \frac{a_i(x)\pi(x)}{\hat{\pi}(i)},
$$

$Q_i(x,y) := a_i(y)Q(x,y)$ if $x \neq y$ and define $Q_i(x,x)$ such that every row sums to 0. It is easily seen that $\hat{Q}$ is reversible under the probability distribution $\hat{\pi}$ and that Q_i is reversible under the probability distribution π_i. As before, for $i \in I$ we define

$$
\hat{f}(i) := \mathbb{E}_{\pi_i} f = \sum_{x\in\Lambda_i} f(x)\pi_i(x).
$$

Next, we calculate the variance, entropy and Dirichlet forms. For any $f : \Omega \to \mathbb{R}$,

$$
\begin{aligned}
&\mathrm{Var} f = \frac{1}{2} \sum_{x,y\in\Omega} \pi(x)\pi(y)(f(x)-f(y))^2 \\
&= \frac{1}{2} \sum_{i,j\in I} \sum_{x,y\in\Omega} a_i(x)a_j(y)\pi(x)\pi(y)[(f(x)-\hat{f}(i)) + (\hat{f}(i)-\hat{f}(j)) + (\hat{f}(j)-f(y))]^2.
\end{aligned}
$$

The cross term given by $(f(x)-\hat{f}(i))$ and $(\hat{f}(i)-\hat{f}(j))$ when expanding the square is

$$
\begin{aligned}
&\frac{1}{2} \sum_{i,j\in I} \sum_{x,y\in\Omega} a_i(x)a_j(y)\pi(x)\pi(y)(f(x)-\hat{f}(i))(\hat{f}(i)-\hat{f}(j)) \\
=&\frac{1}{2} \sum_{i,j\in I} (\hat{f}(i)-\hat{f}(j)) \left[\sum_{y\in\Omega} a_j(y)\pi(y)\right] \sum_{x\in\Omega} a_i(x)\pi(x)(f(x)-\hat{f}(i)) \\
=&\frac{1}{2} \sum_{i,j\in I} (\hat{f}(i)-\hat{f}(j))\hat{\pi}(i)\hat{\pi}(j)\mathbb{E}_{\pi_i}[f-\hat{f}(i)] = 0.
\end{aligned}
$$

Similarly, the term given by $(\hat{f}(i)-\hat{f}(j))$ and $(\hat{f}(j)-f(y))$ is also 0. Finally,

$$
\begin{aligned}
&\frac{1}{2} \sum_{i,j\in I} \sum_{x,y\in\Omega} a_i(x)a_j(y)\pi(x)\pi(y)(f(x)-\hat{f}(i))(\hat{f}(j)-f(y)) \\
=&\frac{1}{2} \sum_{i,j\in I} \hat{\pi}(i)\hat{\pi}(j)\mathbb{E}_{\pi_i}[f-\hat{f}(i)]\mathbb{E}_{\pi_j}[\hat{f}(j)-f] = 0.
\end{aligned}
$$

Thus, $\mathrm{Var}_\pi f$ is equal to

$$\begin{aligned}
&\frac{1}{2}\sum_{i,j\in I}\sum_{x,y\in\Omega} a_i(x)a_j(y)\pi(x)\pi(y)[(f(x)-\hat{f}(i))^2+(\hat{f}(i)-\hat{f}(j))^2+(\hat{f}(j)-f(y))^2] \\
&=\sum_{i\in I}\sum_{x\in\Omega} a_i(x)\pi(x)(f(x)-\hat{f}(i))^2+\frac{1}{2}\sum_{i,j\in I}\hat{\pi}(i)\hat{\pi}(j)(\hat{f}(i)-\hat{f}(j))^2 \\
&=\sum_{i\in I}\hat{\pi}(i)\mathrm{Var}_{\pi_i}f+\mathrm{Var}_{\hat{\pi}}\hat{f}
\end{aligned}$$

As for the entropy, for $f:\Omega\to(0,\infty)$

$$\begin{aligned}
\mathrm{Ent}_\pi f &= \mathbb{E}_\pi[f(\log f-\log\mathbb{E}_\pi f)] = \sum_{x\in\Omega}\sum_{i\in I} a_i(x)\pi(x)f(x)\left[\log f(x)-\log\mathbb{E}_\pi f\right] \\
&= \sum_{i\in I}\hat{\pi}(i)\sum_{x\in\Lambda_i}\pi_i(x)f(x)\left[\log f(x)-\log\mathbb{E}_\pi f\right] \\
&= \sum_{i\in I}\hat{\pi}(i)\mathrm{Ent}_{\pi_i}f+\sum_{i\in I}\hat{\pi}(i)\sum_{x\in\Lambda_i}\pi_i(x)f(x)\left[\log\mathbb{E}_{\pi_i}f-\log\mathbb{E}_\pi f\right] \\
&= \sum_{i\in I}\hat{\pi}(i)\mathrm{Ent}_{\pi_i}f+\sum_{i\in I}\hat{\pi}(i)\hat{f}(i)\left[\log\hat{f}(i)-\log\mathbb{E}_{\hat{\pi}}\hat{f}\right] \\
&= \sum_{i\in I}\hat{\pi}(i)\mathrm{Ent}_{\pi_i}f+\mathrm{Ent}_{\hat{\pi}}\hat{f}.
\end{aligned}$$

The Dirichlet forms are

$$\begin{aligned}
\mathcal{L}_\pi f =&\frac{1}{2}\sum_{i,j\in I}\sum_{x,y\in\Omega} a_i(x)a_j(y)\pi(x)Q(x,y)\Psi(f(x),f(y)) \\
=&\frac{1}{2}\sum_{i\in I}\sum_{x,y\in\Omega} a_i(x)\pi(x)a_i(y)Q(x,y)\Psi(f(x),f(y))+ \\
&\frac{1}{2}\sum_{i\in I}\sum_{j\neq i}\sum_{x,y\in\Omega} a_i(x)a_j(y)\pi(x)Q(x,y)\Psi(f(x),f(y))
\end{aligned}$$

The first term is

$$\frac{1}{2}\sum_{i\in I}\hat{\pi}(i)\sum_{x,y\in\Lambda_i}\pi_i(x)Q_i(x,y)\Psi(f(x),f(y))=\sum_{i\in I}\hat{\pi}(i)\mathcal{L}_{\pi_i}f.$$

Now suppose that for $(i,j)\in I^2$ with $\hat{Q}(i,j)>0$ we are given couplings κ_{ij} of π_i and π_j. We define the quality of these couplings to be

$$\chi:=\min_{\substack{(x,y,i,j)\text{ s.t.}\\ \text{denominator}>0\\ x\neq y}}\frac{a_i(x)a_j(y)\pi(x)Q(x,y)}{\hat{\pi}(i)\hat{Q}(i,j)\kappa_{ij}(x,y)}=\min_{\substack{(x,y,i,j)\text{ s.t.}\\ \text{denominator}>0\\ x\neq y}}\frac{\pi_i(x)a_j(y)Q(x,y)}{\hat{Q}(i,j)\kappa_{ij}(x,y)},$$

where $x\in\Lambda_i$, $y\in\Lambda_j$, $x\neq y$ and thus $a_i(x)a_j(y)\neq 0$ since π_i and π_j are the marginals of κ_{ij}. Again, for these couplings to be useful (i.e. $\chi>0$), we must have $\kappa_{ij}(x,y)>0 \implies Q(x,y)>0$ for all $x\in\Lambda_i$, $y\in\Lambda_j$, $x\neq y$ where $\hat{Q}(i,j)>0$. By Jensen's inequality and the convexity of Ψ,

$$\sum_{x\in\Lambda_i}\sum_{y\in\Lambda_j}\kappa_{ij}(x,y)\Psi(f(x),f(y))\geq\Psi(\hat{f}(i),\hat{f}(j)),$$

and hence for any $(x, y, i, j) \in \Lambda_i \times \Lambda_j \times I^2$ with $x \neq y$ and $Q(i, j) > 0$ we have

$$a_i(x)a_j(y)\pi(x)Q(x,y)\Psi(f(x),f(y)) \geq \chi\hat{\pi}(i)\hat{Q}(i,j)\kappa_{ij}(x,y)\Psi(f(x),f(y)).$$

This holds true even for $x = y$ and (i, j) such that $\hat{Q}(i, j) \leq 0$, so

$$\begin{aligned}
&\sum_{x,y\in\Omega} a_i(x)a_j(y)\pi(x)Q(x,y)\Psi(f(x),f(y)) \geq \chi\hat{\pi}(i)\hat{Q}(i,j)\Psi(\hat{f}(i),\hat{f}(j)) \\
\implies &\frac{1}{2}\sum_{i\in I}\sum_{j\neq i}\sum_{x,y\in\Omega} a_i(x)a_j(y)\pi(x)Q(x,y)\Psi(f(x),f(y)) \\
&\geq \frac{1}{2}\sum_{i,j\in I}\chi\hat{\pi}(i)\hat{Q}(i,j)\Psi(\hat{f}(i),\hat{f}(j)) = \chi\mathcal{L}_{\hat{\pi}}\hat{f}.
\end{aligned}$$

In summary, we have recovered (17), (18) and hence Theorem 5.1 but in a more general setting.

Theorem 5.2. *Let Ω be a finite set in which each $x \in \Omega$ is assigned a stochastic vector $a(x) = (a_i(x))_{i\in I}$ of dimension $|I| < \infty$. Suppose that Q is a reversible Markov generator on Ω. With the new definitions of $\hat{\pi}$, π_i, $\hat{Q}$, Q_i and the assumption that there exist couplings of π_i and π_j for $i, j \in I$ with $\hat{Q}(i, j) > 0$, we have*

$$\begin{aligned}
\lambda(Q) &\geq \min\{\chi\lambda(\hat{Q}), \min_{i\in I}\lambda(Q_i)\}, \\
\alpha(Q) &\geq \min\{\chi\alpha(\hat{Q}), \min_{i\in I}\alpha(Q_i)\}, \\
\rho(Q) &\geq \min\{\chi\rho(\hat{Q}), \min_{i\in I}\rho(Q_i)\}.
\end{aligned}$$

In general, the classes Λ_i ($i \in I$) are overlapping, in the sense that there may be $i \neq j$ in I such that $a_i(x)a_j(x) > 0$ for some $x \in \Omega$. This allows for greater flexibility when choosing the partition.

Example 5.2. Consider a three-state chain on $\Omega = \{x, y, z\}$ with Markov generator

$$Q = \begin{bmatrix} -p & p & 0 \\ q & -q-r & r \\ 0 & s & -s \end{bmatrix}$$

where $0 < p, q, r, s < 1$, $q + r \leq 1$ and the columns/rows are in the order of x, y, z. Note that there is no partition that will allow us to use Theorem 5.1: Nontrivial partitions include $\{x\} \cup \{y, z\}$, $\{x, y\} \cup \{z\}$ and $\{x, z\} \cup \{y\}$. The first two cases cannot possibly have couplings with $\chi > 0$ because $Q(x, z) = Q(z, x) = 0$, while in the last case the restriction on $\{x, z\}$ is not irreducible.

Now consider $a(x) = (1, 0)$, $a(y) = (\theta, 1-\theta)$ and $a(z) = (0, 1)$. Then $\Lambda_1 = \{x, y\}$ and $\Lambda_2 = \{y, z\}$. In order to apply Theorem 5.2 we need to find a coupling κ_{12} on $\Lambda_1 \times \Lambda_2$ (and then set $\kappa_{21}(w, v) = \kappa_{12}(v, w)$) with $\chi > 0$. Since $Q(x, z) = 0$ we must require that $\kappa_{12}(x, z) = 0$, and since π_1 and π_2 are the marginals of κ_{12} it follows that $\kappa_{12}(x, y) = \pi_1(x)$ and $\kappa_{12}(y, z) = \pi_2(z)$. For κ_{12} to be well-defined we must require that $\kappa_{12}(y, y) = 1 - \kappa_{12}(x, y) - \kappa_{12}(y, z) \geq 0$. Straightforward calculation shows that this is possible if and only if $sp \leq 4qr$, and in this case we can choose θ to be in the range

$$\theta \in \left[\frac{1}{2} - \sqrt{\frac{1}{4} - \frac{qr}{sp}}, \frac{1}{2} + \sqrt{\frac{1}{4} - \frac{qr}{sp}}\right],$$

and in particular $\theta = 1/2$ is always possible. Therefore, using Theorem 5.2 one can bound the constants of Q from below in terms of those of $\hat{Q}$, Q_1 and Q_2, which are two-state chains with parameters p, q, r, s, θ and hence Theorem 4.4 applies. This example may be not interesting as one can solve for its eigenvalues and obtain $\lambda(Q)$ directly. Nevertheless it shows how the method could be applied.

Example 5.3. (Hypercube with one additional point) We now consider a chain on $\Omega = \mathcal{Q}^n \cup \{z\}$, the n-dimensional cube with one additional point. We write $v \sim w$ to denote that $v, w \in \mathcal{Q}^n$ differ by exactly one coordinate, and we separate $\mathcal{Q}$ by the last coordinate into $\Omega_1 = \{x_1, ..., x_{2^{n-1}}\}$ and $\Omega_2 = \{y_1, ..., y_{2^{n-1}}\}$ such that $x_i \sim y_i$ for all $i = 1, ..., 2^{n-1}$.

For positive numbers r, ϵ and η such that $nr + \epsilon \leq 1$ and $2^n \eta \leq 1$, define the chain $Q = Q[n, r, \epsilon, \eta]$ by

$$\begin{cases} Q(x_i, y_i) = Q(y_i, x_i) = r \\ Q(x_i, z) = Q(y_i, z) = \epsilon \\ Q(z, x_i) = Q(z, y_i) = \eta \\ Q(v, w) = 0 \text{ otherwise if } v \neq w \end{cases}$$

with diagonal entries $Q(v, v)$ defined so that each row sums to 0. The conditions $nr+\epsilon \leq 1$ and $2^n \eta \leq 1$ are to make sure that Q is well-defined. To find the reversible stationary distribution π we note that $\pi(x) = \pi(y) =: \pi_0$ for all $x \in \Omega_1$ and $y \in \Omega_2$ and that $\pi(z) = \frac{\epsilon}{\eta}\pi_0$. Solving for $\sum_v \pi(v) = 1$ gives

$$\pi_0 = \frac{1}{2^n + \epsilon/\eta}.$$

It is possible to apply Theorem 5.1 by using the partition $\Omega_1' := \Omega_1 \cup \Omega_2$ and $\Omega_2' := \{z\}$. There is only one possible coupling $\kappa(x, z) = 2^{-n}$ $(x \in \Omega_1')$ for which $\chi = 1$ and the projection chain is $\hat{Q}(1, 2) = \epsilon$, $\hat{Q}(2, 1) = \eta$. Note that Q_1 was the chain considered in Example 5.1, so we have

$$\lambda(Q) \geq \min\{2r, \epsilon + \eta\}, \rho(Q) \geq \begin{cases} \min\{r, \epsilon\} & \text{if } \epsilon = \eta, \\ \min\{r, \frac{\eta - \epsilon}{\log \eta - \log \epsilon}\} & \text{if } \epsilon \neq \eta. \end{cases} \tag{19}$$

Now we attempt to use Theorem 5.2 in a nontrivial manner (i.e. classes overlap). It is natural to split z evenly between the two classes by defining the weight vectors to be

$$a(x_i) = (1, 0), a(y_i) = (0, 1), a(z) = (1/2, 1/2)$$

and hence $\Lambda_1 = \Omega_1 \cup \{z\}$, $\Lambda_2 = \Omega_2 \cup \{z\}$. Calculation shows that $\hat{\pi}(1) = \hat{\pi}(2) = 1/2$ and thus

$$\pi_1(x_i) = \pi_2(y_i) = 2\pi_0, \ \pi_1(z) = \pi_2(z) = \frac{\epsilon}{\eta}\pi_0.$$

In addition, the projection chain is

$$\hat{Q}(1, 2) = 2^n(r + \epsilon)\pi_0 = \hat{Q}(2, 1)$$

and the restriction chains are

$$\begin{cases} Q_1(x_i, x_j) = Q(x_i, x_j), Q_2(y_i, y_j) = Q(y_i, y_j) \\ Q_1(z, x_i) = Q_2(z, y_i) = \eta \\ Q_1(x_i, z) = Q_2(y_i, z) = \epsilon/2. \end{cases}$$

This gives rise to the recursive relation

$$(Q[n,r,\epsilon,\eta])_k = Q[n-1,r,\epsilon/2,\eta], \quad k=1,2. \tag{20}$$

Note that $Q[n-1,r,\epsilon/2,\eta]$ is well-defined since $(n-1)r+\epsilon/2 \le 1$ and $2^{n-1}\eta \le 1$.

Next we find a coupling $\kappa = \kappa_{12}$ of π_1 and π_2, and by symmetry we will set $\kappa_{21}(w,v) := \kappa(v,w)$ for $v \in \Lambda_1$ and $w \in \Lambda_2$. As in Example 5.1 we seek to pair together $x_i \in \Omega_1$ and $y_j \in \Omega_2$ by setting $\kappa(x_i,y_i) = p$ and $\kappa(x_i,y_j) = 0$ if $i \ne j$ for some appropriate $p \ge 0$. Since the marginals of κ are π_1 and π_2, we must have

$$\kappa(z,y_i) = \kappa(x_i,z) = 2\pi_0 - p,$$

so $p \le 2\pi_0$. Finally, $\sum_i \kappa(z,y_i) + \kappa(z,z) = \pi_1(z)$ forces

$$\kappa(z,z) = 2^{n-1}(p-2\pi_0) + \frac{\epsilon}{\eta}\pi_0 \ge 0 \tag{21}$$

and hence

$$\max\left\{0, \frac{1}{2^{n-1}}\left(2^n - \frac{\epsilon}{\eta}\right)\right\} \le p/\pi_0 \le 2. \tag{22}$$

Thus any p satisfying (22) gives rise to a coupling κ whose quality is

$$\chi = \min_{\substack{(v,w)\in\Lambda_1\times\Lambda_2\\ v\ne w,\, \kappa(v,w)>0}} \frac{a_1(v)a_2(w)\pi(v)Q(v,w)}{\hat{\pi}(1)\hat{Q}(1,2)\kappa(v,w)} = \frac{1}{2^n(r+\epsilon)} \min\left\{\frac{2r}{p}, \frac{\epsilon}{2\pi_0 - p}\right\}.$$

Theorem 4.4 and 5.2 tell us

$$\begin{aligned}
\lambda(Q[n,r,\epsilon,\eta]) &\ge \min\left\{\frac{4r\pi_0}{p}, \frac{2\epsilon\pi_0}{2\pi_0 - p}, \lambda(Q[n-1,r,\epsilon/2,\eta])\right\} \\
&= \min\left\{\frac{4r}{p/\pi_0}, \frac{2\epsilon}{2-p/\pi_0}, \lambda(Q[n-1,r,\epsilon/2,\eta])\right\}, \qquad (23)\\
\rho(Q[n,r,\epsilon,\eta]) &\ge \min\left\{\frac{2r}{p/\pi_0}, \frac{\epsilon}{2-p/\pi_0}, \rho(Q[n-1,r,\epsilon/2,\eta])\right\}. \qquad (24)
\end{aligned}$$

where p can be chosen in the range (22) to maximize the expressions based on the values of r, ϵ and η.

Now we address the base case $n = 1$ where $Q = Q[1,r,\epsilon,\eta]$ is a chain on three elements $\{x_1, y_1, z\}$. The stationary distribution is

$$\pi(x_1) = \pi(y_1) = \frac{\eta}{\epsilon + 2\eta}, \; \pi(z) = \frac{\epsilon}{\epsilon+2\eta}$$

The characteristic polynomial of Q is

$$ch(t) = \det(Q - tI) = -t(t+2r+\epsilon)(t+2\eta+\epsilon)$$

and hence by Theorem 4.3 $\lambda(Q) = \min\{2r+\epsilon, 2\eta+\epsilon\} = 2\min(r,\eta)+\epsilon$ and a lower bound for $\rho(Q)$ is given by Corollary 4.1. Thus

$$\lambda(Q[1,r,\epsilon/2^{n-1},\eta]) = 2\min(r,\eta) + \epsilon/2^{n-1}$$

and repeatedly applying (23), (24) while choosing the optimal p at each step yields the result. We do not go down this path further as the application of Theorem 5.1 is arguably more conceptually appropriate in this case. Next we study an example in which Theorem 5.1 may not apply.

Example 5.4. Let G be a finite set on which an undirected connected graph is defined. We consider a graph consisting of two copies of G but with some points glued together. More precisely, let H be a subset of G such that if $x \in H$ then $N_G(x) \cap H = \varnothing$, where $N_G(x)$ is the set of neighbors of x in G. Define an equivalence relation $\equiv$ on $G \times \{1,2\}$ by $(x,i) \equiv (y,j)$ if and only if $(x,i) = (y,j)$ or $x = y \in H$. Let $\mathcal{G} := {}^{G \times \{1,2\}}\!\big/_{\equiv}$. We identify the elements of H with their classes and write $G_i = \{(x,i) \in \mathcal{G} \mid x \in G \setminus H\}$. Thus $\mathcal{G} = G_1 \cup G_2 \cup H$. The edges of $\mathcal{G}$ are defined by the edges of the two copies of G while connecting $(x,1)$ and $(x,2)$ for all $x \in G \setminus H$. Write $v \sim w$ if v and w are neighbors in $\mathcal{G}$.

Now we consider the simple random walk on $\mathcal{G}$. Let $d_G(x)$ and $d_{\mathcal{G}}(x)$ denote the degree of x in G and $\mathcal{G}$ respectively, and let

$$D_G := \sum_{x \in G} d_G(x), \quad D_{\mathcal{G}} := \sum_{x \in \mathcal{G}} d_{\mathcal{G}}(x).$$

Then the chain in question is $Q(x,y) = 1/d_{\mathcal{G}}(x)$ if $x \sim y$ and $\pi(x) = d_{\mathcal{G}}(x)/D_{\mathcal{G}}$, where

$$d_{\mathcal{G}}(x) = \begin{cases} 2d_G(x) & \text{if } x \in H, \\ d_G(x) + 1 & \text{otherwise.} \end{cases}$$

Thus $D_{\mathcal{G}} = 2D_G + 2\#(G \setminus H)$ where $\#$ denotes the cardinality of a set. Assign weight vectors of $x \in \mathcal{G}$ to be $a(x) = (1/2, 1/2)$ if $x \in H$, $a(x) = (1,0)$ if $x \in G_1$ and $a(x) = (0,1)$ if $x \in G_2$. Thus $\Lambda_i = G_i \cup H$ and $\hat{\pi}(i) = 1/2$ by symmetry $(i = 1,2)$. Note that $\pi_1(x) = \pi_2(x) = \pi(x)$ for $x \in H$ and $\pi_1(x) = \pi_2(y) = 2\pi(x)$ if $x \in G_1$ and $y \in G_2$ correspond to the same element in G. Hence the probability measure $\kappa := \kappa_{12}$ on $\Lambda_1 \times \Lambda_2$ defined by

$$\kappa(x,y) = \begin{cases} \pi(x) & \text{if } x = y \in H \\ 2\pi(x) & \text{if } x = (v,1) \text{ and } y = (v,2) \text{ for some } v \in G \setminus H \\ 0 & \text{otherwise} \end{cases}$$

is a coupling of π_1 and π_2. Again, by symmetry we set $\kappa_{21}(y,x) = \kappa(x,y)$. Note that $\pi(x)Q(x,y) = \mathbb{1}_{x \sim y}/D_{\mathcal{G}}$, so the projection chain is

$$\begin{aligned} \hat{Q}(2,1) = \hat{Q}(1,2) &= \frac{1}{\hat{\pi}(1)} \sum_{x \in \Lambda_1} \sum_{y \in \Lambda_2 \setminus \{x\}} a_1(x) a_2(y) \pi(x) Q(x,y) \\ &= \frac{2}{D_{\mathcal{G}}} \left(\sum_{x \in H} \frac{1}{2} d_G(x) + \sum_{y \in H} \frac{1}{2} d_G(y) + \#(G \setminus H) \right) \\ &= \frac{2}{D_{\mathcal{G}}} \left(\sum_{x \in H} d_G(x) + \#(G \setminus H) \right). \end{aligned}$$

Recall that the quality of the couplings are

$$\chi = \min \frac{a_1(x) a_2(y) \pi(x) Q(x,y)}{\hat{\pi}(1) \hat{Q}(1,2) \kappa(x,y)}$$

where the minimum is taken for $x \neq y$ with $\kappa(x,y) > 0$ and therefore $x = (v,1)$, $y = (v,2)$ for some $v \in G \setminus H$. It follows that

$$\chi = \min_{x \in G_1 \cup G_2} \frac{1/D_{\mathcal{G}}}{\frac{1}{2}\hat{Q}(1,2) \cdot 2d_{\mathcal{G}}(x)/D_{\mathcal{G}}} = \frac{1}{\hat{Q}(1,2)} \frac{1}{\max_{v \in G \setminus H} d_G(v) + 1}.$$

The restriction chains $Q_i(x, y)$ on G are largely the same as $Q(x, y)$, except that $Q_i(x, y) = \frac{1}{2}Q(x, y)$ if $y \in H$. We conclude that

$$\lambda(Q) \geq \min \left\{ \frac{1}{\max_{v \in G \setminus H} d_G(v) + 1}, \lambda(Q_i) \right\}$$

and similarly for $\alpha(Q)$ and $\rho(Q)$. This procedure is a generalization of what we did in Example 5.1, which is the case when $G = \{0, 1\}^{n-1}$ and $H = \varnothing$. However, in general when $H \neq \varnothing$, there is no (obvious) way to choose a partition of Ω for Theorem 5.1 such that the projection and restriction chains all stay irreducible while ensuring that the quality χ is positive.

Bibliography

[1] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times, second edition*. 2017.

[2] N. Berestycki, "Mixing times of markov chains: Techniques and examples," 2019.

[3] "The perron-frobenius theorem," 2007.

[4] P. Diaconis and L. Saloff-Coste, "Logarithmic sobolev inequalities for finite markov chains," *The Annals of Applied Probability*, August 1996.

[5] S. G. Bobkov and P. Tetali, "Modified logarithmic sobolev inequalities in discrete settings," *Journal of Theoretical Probability*, April 2006.

[6] A. B. Tsybakov, "Introduction to nonparametric estimation," 2009.

[7] "University of waterloo, cs 860: Eigenvalues and polynomials," April 5, 2022.

[8] K. Tikhomirov and P. Youssef, "Regularized modified log-sobolev inequalities, and comparison of markov chains," 2022.

[9] S. L. Lu and H.-T. Yau, "Spectral gap and logarithmic sobolev inequality for kawasaki and glauber dynamics," *Communications in Mathematical Physics*, 1993.

[10] J. Hermon and J. Salez, "Modified log-sobolev inequalities for strong-rayleigh measures," February 2020.