THE UNIVERSITY OF CHICAGO

ROBUST MIXING

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY

MURALI KRISHNAN GANAPATHY

CHICAGO, ILLINOIS

AUGUST 2006

To my dear wife and parents

Abstract

How many times should a card shuffler shuffle to get the cards shuffled? Convergence rate questions like these are central to the theory of finite Markov Chains and arise in diverse fields including Physics, Computer Science as well as Biology. This thesis introduces two new approaches to estimating mixing times: *robust mixing time* of a Markov Chain and *Markovian product* of Markov Chains.

The "robust mixing time" of a Markov Chain is the notion of mixing time which results when the steps of the Markov Chain are interleaved with that of an oblivious adversary under reasonable assumptions on the intervening steps. We develop the basic theory of robust mixing and use it to give a simpler proof of the limitations of reversible liftings of a Markov Chain due to Chen, Lovász, and Pak (1999). We also use this framework to improve the mixing time estimate of the random-to-cyclic transposition process (a non-Markovian process) given by Peres and Revelle (2004).

The "Markovian product" of Markov Chains is like the direct product, except for a *controlling Markov Chain* which helps decide which component should be updated. Direct products as well as wreath products of Markov Chains are special cases. We show how a coupon collector type of analysis can be used to estimate the mixing times of these product chains under various distance measures. Using this, we derive L^2-mixing time estimates of a Cayley walk on *Complete Monomial Groups*, which are

only a factor 2 worse than those obtained by Schoolfield (2002) using representation theory. Together with Robust mixing time estimates, we also estimate mixing times of Markov Chains which arise in the context of Sandpile groups.

In the case of *Lamp Lighter Chains*, we are lead to estimating moment generating functions of occupancy measures of a Markov Chain. We sharpen a previous estimate, due to Peres and Revelle (2004), of how long it takes to visit all the states of a Markov Chain, and answer some questions raised by them.

Acknowledgements

I am grateful to my advisor and mentor, László Babai, for enthusiastically and patiently sharing his knowledge, for the long discussions as well as the unforgettable class lectures. To Steven Lalley, for the delightful class on Markov Chains and for helpful discussions. To Partha Niyogi, for his inspiring course on Artificial Intelligence and serving on my committee.

I would also like to thank Prasad Tetali for inviting me over to Georgia Tech, and for introducing me to several problems. To Peter Winkler, for a short but very useful discussion.

I would also like to thank my friends and colleagues here at the University, because of whom I will have fond memories of Chicago. My wife Mahalakshmi, for her endurance and support and my parents for their constant encouragement.

Table of contents

List of Figures

Chapter 1

Introduction

Markov Chains arise in diverse fields including Physics (Statistical Mechanics), Statistics (Queueing theory), Biology (Population Processes), Bio-informatics (gene prediction), Economics (Modelling Economic Growth) and Computer Science (Approximation Algorithms). One of the central questions is to get quantitative estimates on the rate of convergence of Markov Chains.

Markov chain Monte Carlo (MCMC) methods have found wide applicability. MCMC methods are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its stationary distribution. Common applications include estimating multi-dimensional integrals and Approximation algorithms. The "mixing time" of the underlying Markov Chain is a key component of the running time of the algorithms.

Google's PageRank^{TM} algorithm[1] is essentially a Markov Chain based algorithm. Consider the following random walk on the set of "all" web pages: Suppose we are currently at a page u. With probability $1-q$ we start over by moving to a page chosen

[1] as originally published

from the uniform distribution. With probability q, we follow a link out of the current page u (all links equally likely). $q = 0.85$ is the damping factor and helps handle pages without links. The stationary distribution of this Markov Chain is essentially the PageRank of all the web pages. Every once in a while, Google takes one step in this random walk and updates the current distribution.

One recent landmark in Computational Complexity Theory is the L=SL result due to Reingold [52]. It gives a log-space algorithm to decide if two vertices of an undirected graph are in the same connected component. The main ingredient in the proof is the construction of a new graph which retains the connectivity information of the old graph and the random walk on the new graph mixes in $O(\log n)$ times. The construction in turn builds on the previous work of Reingold, Vadhan, and Wigderson [53].

1.1 Approaches to estimating mixing times

There is no "right" approach to estimate the mixing times of Markov Chains. Several approaches have been suggested each with its own pros and cons. All approaches seem to have a trade off between easy applicability and tightness of the resulting estimate. We outline some of the more common approaches below. Assume that the Markov Chain has N states and that the stationary distribution is uniform.

(a) Coupling: Very sensitive to the Markov Chain but very elegant and usually provides optimal bounds. There is some "magic" involved in finding the right coupling.

(b) Spectral Gap: Involves estimating the largest non-trivial singular value of the transition matrix. For reversible chains, it determines the convergence rate

exactly. For the mixing time, there is still a $O(\log N)$-gap where N is the number of states in the chain.

(c) Conductance: Combinatorial in nature and is easy in principle and in case of Markov Chains with symmetry also easy in practice. However the resulting bound determines the mixing rate up only to a quadratic factor. Unfortunately, there are examples showing that this gap cannot be improved.

(d) Representation theory: Used for estimating L^2-mixing times of Markov Chains on groups. In principle, gives optimal results. In practice, has been applied only for Markov Chains where the generators are invariant under conjugation or close enough to being so.

(e) log-Sobolev constant: Difficult to estimate, but resulting bound only has a $O(\log \log N)$-gap for the L^2-mixing time.

(f) Entropy constant: Notoriously difficult to estimate, but resulting bound estimates the mixing time within a $O(\log \log N)$-gap.

(g) Comparison Methods: Involves comparing an unknown chain, with a known chain and deducing mixing time estimates for the unknown chain, in terms of that of the known chain. One starts by estimating the conductance or Log-Sobolev constants or spectral gap of the unknown chain in terms of that of the known chain and then infer mixing time estimates for the unknown chain. Has wide applicability. Tightness of resulting bound depends on the choice of the known chain.

(h) Decomposition Methods: Break up the states of the chain into pieces and derive mixing time bounds on the whole chain, in terms of that of the pieces and the

projection chain where the pieces are the states. Here also, we estimate the conductance or Log-Sobolev constant of the whole chain in terms of that of the pieces and the projection chain.

Other methods include Blocking Conductance, Spectral Profiling, Evolving Sets, Congestion and of course explicitly calculating the t-step transition probabilities. This thesis introduces two new approaches to estimate mixing times of Markov Chains: *robust mixing time* of a Markov Chain and *Markovian product* of Markov Chains.

1.2 Robust Mixing

The mixing time of a Markov Chain can be viewed in terms of the following two player game.

0. Let $\mathbb{P}$ be a Markov Chain with stationary distribution π.
1. *Adversary*: Picks an initial distribution μ
2. *System*: $\mu \leftarrow \mu\mathbb{P}$
3. If μ is "close enough" to π the game ends. Otherwise, go back to step 2.

The goal of the adversary is to prolong the game as much as possible. The mixing time can then be defined as the length of the game when the adversary plays his optimal strategy. If $\mathbb{P}$ is ergodic, the game will end eventually. Admittedly, the *system* has no active role in the game and the role of the *adversary* is limited. In the case of robust mixing time, the adversary plays a more active role. In the robust setting the game proceeds as follows:

0. Let $\mathbb{P}$ be a Markov Chain with stationary distribution π.

1. *Adversary*: Picks an initial distribution μ and a sequence $\{\mathbb{A}_t\}_{t>0}$ of stochastic matrices such that $\pi\mathbb{A}_t = \pi$ for all t.

2. Set $t = 0$

3. $t \leftarrow t + 1$

4. *System*: $\mu \leftarrow \mu\mathbb{P}$

5. *Adversary*: $\mu \leftarrow \mu\mathbb{A}_t$

6. If μ is "close enough" to π the game ends. Otherwise, go back to step 3.

As before the adversary's goal is to prolong the game as much as possible. The robust mixing time is then the length of the game when the adversary plays his optimal strategy. Note that the adversary is oblivious since his moves $\mathbb{A}_t$ are decided at the start of the game. Put another way, the robust mixing time of a Markov Chain is the notion of mixing time which results when the steps of the Markov Chain are interleaved with that of an oblivious adversary under reasonable assumptions on the intervening steps.

Liftings: A Markov Chain $\mathbb{P}$ is said to be a *collapsing* of $\mathbb{Q}$, if the states of $\mathbb{Q}$ can be partitioned so that each state of $\mathbb{P}$ corresponds to an element of the partition and the transition probabilities are those induced by that of $\mathbb{Q}$. We also call $\mathbb{Q}$ a *lifting* of $\mathbb{P}$. In principle, one can use a lifting to add some "inertia" to the chain which causes the chain to spread around more quickly. Thus we can hope that $\mathbb{Q}$ will mix faster than $\mathbb{P}$. Chen, Lovász, and Pak [8] showed the limitations of the speed up one can expect to gain via liftings. If we consider the robust mixing time of $\mathbb{Q}$, we can define a suitable adversarial strategy which allows the adversary to "cancel" the inertia effect. Put simply, at each step the adversary averages over the states in each element of the

partition. Hence the robust mixing time of $\mathbb{Q}$ is at least the standard mixing time of $\mathbb{P}$. This observation together with relations between the robust mixing time and standard mixing time, allow us to give a simpler (and slightly sharper) proof of the limitations of reversible liftings, i. e., when $\mathbb{Q}$ is required to be reversible.

Non-Markovian Processes: Consider the following random-to-cyclic transposition process. We start with a deck of n cards in order. At time t, we exchange the card at position t with a random card. This is not a Markov Chain as the transition probabilities are a function of t. However exchanging the card at position r with the one at time t, can be done in three steps as follows:

- Exchange cards at positions 1 and t
- Exchange cards at positions 1 and r
- Exchange cards at positions 1 and t

The middle step is the only step which involves knowledge of the random location r and more importantly does not require the knowledge of t. Hence if we surround every iteration of the top-to-random transposition Markov Chain by an appropriate adversarial strategy, we can simulate the non-Markovian random-to-cyclic process. Hence the robust mixing time of top-to-random transposition Markov Chain gives an upper bound on the mixing time of the random-to-cyclic transposition process. We use this approach to improve the mixing time estimate of the random-to-cyclic transposition process (a non-Markovian process) given by Peres and Revelle [49].

Chapter 3 develops the basic theory of robust mixing. In § 3.6 we show that many upper bounds on standard mixing time already bound robust mixing time. § 3.7 contains the result on the limitations of reversible liftings and § 4.4 shows the improved mixing time estimate of the random-to-cyclic transposition process.

1.3 Markovian products

The *Markovian product* of Markov Chains is like the direct product, except for a *controlling Markov Chain* which helps decide which component should be updated. More specifically, for $i = 1, \ldots, k$, let $\mathbb{P}_i$ be a Markov Chain on $\mathcal{X}_i$ and $\mathbb{Q}$ a Markov Chain on $\mathcal{Y}$. The Markovian product of $\mathbb{P}_1, \ldots, \mathbb{P}_k$ controlled by $\mathbb{Q}$ is a Markov Chain on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_k \times \mathcal{Y}$ which evolves as follows: Given the current state $(x_1, \ldots, x_k; y)$, we do some of the following operations:

(a) for certain values of i, update x_i using $\mathbb{P}_i$

(b) update y using $\mathbb{Q}$

We start with a random $r \in [0, 1]$. Based on r, we decide whether y should be updated. Based on y and r we select the x_i which will be updated. Having made the decisions, we do the actual updating. Direct products and Wreath products are special cases. Note that different decision and selection rules may result in different chains on the same state space with the same stationary distribution.

In Chapter 5, we show how to estimate the mixing time of a Markovian product under various measures in terms of the mixing times of its components and occupancy measures of the controlling chain. We also consider the case when the component chains are not independent and an update to one can change the state of another. Specifically, we show that if the dependency among the components is acyclic and the nature of the dependency can be described via the moves of an adversary (in robust mixing time parlance), the independent case estimate also work here, as long as one considers the robust mixing time of the components instead of their standard mixing times.

For the total variation mixing time, we need bounds on the blanket time of the controlling chain. The blanket time of a Markov Chain was introduced by Winkler and Zuckerman [65]. For L^2-mixing time, we need bounds on the moment generating function of certain occupancy measures.

Complete Monomial Groups and Sandpile Groups: In Chapter 5, we derive the required bounds when the controlling chain reaches stationarity in one step. In this case, we have a coupon collector type analysis. Using this, we derive L^2-mixing time estimates of a Cayley walk on *Complete Monomial Groups*, which are only a factor of 2 worse than those obtained by Schoolfield [56] using representation theory. Together with robust mixing time estimates, we also estimate the mixing times of Markov Chains which arise in the context of Sandpile groups.

Lamp Lighter Chains: In the case of Lamp Lighter Chains, the controlling chain doesn't have to reach stationarity in one step. Thus in order to bound the L^2 mixing times, we need estimates on moment generating functions of occupancy measures of a Markov Chain. In Chapter 6, we prove tight bounds on the moment generating functions when the component chains mix slowly, as well as a general (weaker) bound. We also sharpen a previous estimate by Peres and Revelle [49] which applies when the factor chains reach stationarity in one step. Finally, we estimate the mixing times of various Lamp Lighter chains, where the factors themselves do not mix in one step.

Chapter 2

Markov Chain Basics

This chapter develops the basic theory of Markov Chains from a Linear Algebraic point of view. We give new and simpler proofs of sub-multiplicativity.

2.1 Preliminaries

A *discrete stochastic process* on a state space $\mathcal{X}$ is a sequence of random variables $\{\mathbf{X}_t\}_{t=0}^{\infty}$ which take values in $\mathcal{X}$. It is said to be *Markovian* (or memoryless) if the conditional distribution of $\mathbf{X}_{t+1}$ given the past $\mathbf{X}_t, \mathbf{X}_{t-1}, \ldots$ depends only on $\mathbf{X}_t$.

Definition 2.1. A *Markov Kernel* $\mathcal{M}$ is a pair $(\mathcal{X}, \mathbb{P})$ where $\mathcal{X}$ is a finite *state space*, $\mathbb{P}$ is a square matrix whose rows and columns are indexed by elements of $\mathcal{X}$ and

- (Non-negativity) $\forall x, y \in \mathcal{X}, \mathbb{P}(x, y) \geq 0$
- (Stochasticity) $\forall x \in \mathcal{X}, \sum_{y \in \mathcal{X}} \mathbb{P}(x, y) = 1$

When $\mathcal{X}$ is clear from context we will denote the Markov Kernel by just $\mathbb{P}$.

Definition 2.2. A distribution π on $\mathcal{X}$ is said to *a stationary distribution* of $\mathbb{P}$ if $\pi\mathbb{P} = \pi$.

Definition 2.3. A matrix $\mathbb{P}$ is said to be *doubly stochastic* if both $\mathbb{P} \geq 0$ and both $\mathbb{P}$ and $\mathbb{P}^T$ are stochastic, i. e., $\mathbb{P}$ has non-negative entries with all row sums and column sums 1.

$\mathbb{P}(x, y)$ is the probability that the system moves from state x to state y. Given an initial distribution μ on $\mathcal{X}$, one can define a stochastic process as follows:

$$\Pr\{\mathbf{X}_0 = i\} = \mu(i) \qquad \Pr\{\mathbf{X}_{t+1} = y | \mathbf{X}_t = x, \mathbf{X}_{t-1}, \dots\} = \mathbb{P}(x, y) \tag{2.1}$$

Note that the distribution of $\mathbf{X}_{t+1}$ depends only on $\mathbf{X}_t$ and is independent of t or $\mathbf{X}_s$ for $s < t$. For $0 \leq s \leq t$,

$$\begin{aligned} \Pr\{\mathbf{X}_{t+1} = z | \mathbf{X}_s = x\} &= \sum_{y \in \mathcal{X}} \Pr\{\mathbf{X}_{t+1} = z | \mathbf{X}_t = y\} \cdot \Pr\{\mathbf{X}_t = y | \mathbf{X}_s = x\} \\ &= \sum_{y \in \mathcal{X}} \Pr\{\mathbf{X}_t = y | \mathbf{X}_s = x\} \mathbb{P}(y, z) \end{aligned} \tag{2.2}$$

Hence $\Pr\{\mathbf{X}_t = y | X_0 = x\} = \mathbb{P}^t(x, y)$. We represent a Markov Chain by a weighted graph with vertices $\mathcal{X}$ and directed edges corresponding to allowable transitions (i. e., transitions with non-zero probability). We assume the weights $w(x, y)$ are non-negative and that for each x, $w(x, \cdot) = \sum_y w(x, y) > 0$. Given the weights, the transition probabilities are given by $\mathbb{P}(x, y) = w(x, y)/w(x, \cdot)$. If $w(x, y)$ is not indicated for any edge, it is assumed to be 1. These conventions allow us to represent random walks on graphs by the graphs themselves. Figure 2.1 shows a two-state Markov Chain.

Definition 2.4. A Markov Chain $\mathbb{P}$ is said to be *irreducible* if for any two states x, y, there is a positive probability path from x to y, i. e., the graph representation of $\mathbb{P}$ is strongly connected.

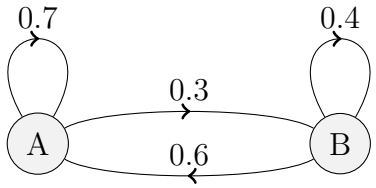


Figure 2.1: Two state Chain

Definition 2.5. A Markov Chain $\mathbb{P}$ is said to be *aperiodic* if the greatest common divisor of the lengths of all cycles in the graph representation of $\mathbb{P}$ is 1.

Definition 2.6. A Markov Chain $\mathbb{P}$ on $\mathcal{X}$ is said to be *lazy*, if $\mathbb{P}(x, x) \geq 1/2$ for all $x \in \mathcal{X}$.

Note that a lazy chain is automatically aperiodic, laziness implies cycles of length 1.

Definition 2.7. A Markov Chain $\mathbb{P}$ is said to be *ergodic* if it is irreducible and aperiodic.

The fundamental theorem on Markov Chains is the following

Theorem 2.8. *Let $\mathbb{P}$ be an ergodic Markov Chain. The following are equivalent:*

(a) All states are positive recurrent

(b) $\mathbb{P}^t$ converges to $\mathbb{A}$ where all rows of $\mathbb{A}$ are the same (equal to π say)

(c) π is the unique vector for which $\pi\mathbb{P} = \pi$.

In particular, if $\mathbb{P}$ is a finite ergodic Markov Chain, the conditions above hold, i.e. it has a unique stationary distribution π and for any initial distribution μ, $\mu\mathbb{P}^t$ converges to π.

A special class of Markov Chains which are easier to handle are the reversible Markov Chains.

Definition 2.9. A finite Markov Chain $\mathbb{P}$ is said to be *reversible* if the *detail balance* condition holds, i. e., for some positive real valued function f on $\mathcal{X}$,

$$(\forall x, y)\, f(x)\mathbb{P}(x, y) = f(y)\mathbb{P}(y, x) \tag{2.3}$$

If an ergodic Markov Chain is reversible, then $\pi(x) = f(x)/C$ is a stationary distribution of $\mathbb{P}$, where $C = \sum_x f(x)$. Hence by uniqueness of the stationary distribution, it follows that f is unique up to a multiplicative constant. We are interested in the time it takes for the t-step distribution to be close to the stationary distribution π. We start with several distance measures.

Definition 2.10. Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π and f be any (complex valued) function on $\mathcal{X}$ and $1 \le p \le \infty$. For $p \ne \infty$, define

$$\|f\|_{p,\pi} = \left(\sum_x \pi(x)|f(x)|^p\right)^{1/p} \qquad \text{and} \qquad \|f\|_{\infty,\pi} = \|f\|_\infty = \sup_x |f(x)| \tag{2.4}$$

If μ is a distribution or difference of two distributions, then define $\|\mu\|_{p,\pi} = \|f\|_{p,\pi}$, where $f(x) = \mu(x)/\pi(x)$ is the density function of μ with respect to π.

Specifically, for $p = 1$, we have $\|\mu - \nu\|_{1,\pi} = \sum_x |\mu(x) - \nu(x)|$ does not depend on π. The most commonly used distance measure is the total variation distance.

Definition 2.11. For distributions μ and ν, their *total variation distance* is given by

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2}\|\mu - \nu\|_1 \tag{2.5}$$

Lemma 2.12.

$$\|\mu - \nu\|_{\mathrm{TV}} = \max_A \{\mu(A) - \nu(A)\} = \sum_x (\mu(x) - \nu(x))^+ = \min_{V_\mu, V_\nu} \Pr\{V_\mu \ne V_\nu\} \tag{2.6}$$

where the maximum is taken over all non-trivial subsets A of $\mathcal{X}$ ($\mathcal{X}$ and $\emptyset$ are trivial), $\mu(A) = \sum_{x \in A} \mu(x)$, $\alpha^+ = \max(\alpha, 0)$ is the positive part of α and the minimum is taken over all random pairs V_μ, V_ν such that $\Pr\{V_\alpha = x\} = \alpha(x)$.

Definition 2.13. For distributions μ and ν, their *relative entropy*, also called Kullback-Leibler distance or *informational divergence*, is defined by the equation

$$\mathrm{D}(\mu||\nu) = \sum_x \mu(x) \log\left(\frac{\mu(x)}{\nu(x)}\right) \tag{2.7}$$

$\mathrm{D}(\mu||\nu) \neq \mathrm{D}(\nu||\mu)$ in general.

Definition 2.14. For two distributions μ, ν, their *separation distance* is defined via

$$\mathrm{sep}(\mu, \nu) = \max_A \left(1 - \frac{\mu(A)}{\nu(A)}\right) \tag{2.8}$$

where the maximum is taken over all non-trivial subsets A of $\mathcal{X}$. This is a one-sided L^∞ distance.

2.2 Mixing Time Definitions

We now define the various mixing times of a Markov Chain.

Definition 2.15. Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π

and μ be any initial distribution.

$$\text{(Mixing time)} \qquad \mathcal{T}(\epsilon) = \max_{\mu} \min_{t} \left\{ \|\mu\mathbb{P}^t - \pi\|_{\mathrm{TV}} \leq \epsilon \right\} \tag{2.9}$$

$$\text{(L_p Mixing time)} \qquad \mathcal{T}_p(\epsilon) = \max_{\mu} \min_{t} \left\{ \|\mu\mathbb{P}^t - \pi\|_{p,\pi} \leq \epsilon \right\} \tag{2.10}$$

$$\text{(Entropy Mixing time)} \qquad \mathcal{T}_{\mathrm{D}}(\epsilon) = \max_{\mu} \min_{t} \left\{ \mathrm{D}(\mu\mathbb{P}^t||\pi) \leq \epsilon \right\} \tag{2.11}$$

$$\text{(Filling time)} \qquad \mathcal{T}_{\mathrm{f}}(\epsilon) = \max_{\mu} \min_{t} \left\{ \mathrm{sep}(\mu\mathbb{P}^t, \pi) \leq \epsilon \right\} \tag{2.12}$$

i. e., the time it takes to reach within ϵ of the stationary distribution starting for the worst initial distribution.

When ϵ is not specified, we take $\epsilon = 1/2$, except for $\mathcal{T}$ where we take $\epsilon = 1/4$. This choice of parameters is justified by Proposition 2.26.

For $x \in \mathcal{X}$, let δ_x denote the distribution concentrated at x. Every μ can be written as a finite convex combination of the δ_x's. Linearity of matrix multiplication, and convexity of the distances imply that the worst case distribution is always a point distribution. Hence it is enough to talk about worst initial state instead of worst initial distribution.

2.3 Properties of Distance Measures

We now look at some properties of distance measures and some relations between them.

Lemma 2.16. *For distributions μ, μ_1, μ_2, ν, $0 \leq a = 1 - b \leq 1$*

(a) $\mathrm{D}(\mu||\nu) \geq 0$ *with equality iff* $\mu = \nu$

(b) (Convexity) $\mathrm{D}(a\mu_1 + b\mu_2||\nu) \leq a\,\mathrm{D}(\mu_1||\nu) + bD(\mu_2||\nu)$

(c) $\mathrm{D}(\mu||\nu) \le \log(1/\nu_*)$, *where* $\nu_* = \min_x \nu(x)$

Proof. $\phi(u) = u \log u$ is strictly convex.

(a) Jensen's inequality implies

$$\mathbb{E}_\nu[\phi(f)] \ge \phi(\mathbb{E}_\nu[f]) = \phi(1) = 0 \tag{2.13}$$

for $f(x) = \mu(x)/\nu(x)$. Hence $\mathrm{D}(\mu||\nu) \ge 0$. Strict convexity implies that f must be constant for equality to hold.

(b) Let $f(x) = \mu_1(x)/\nu(x)$ and $g(x) = \mu_2(x)/\nu(x)$. Convexity of ϕ implies

$$\mathrm{D}(a\mu_1 + b\mu_2||\nu) = \mathbb{E}_\nu[\phi(af + bg)] \le \mathbb{E}_\nu[a\phi(f) + b\phi(g)] = aD(\mu_1||\nu) + b\,\mathrm{D}(\mu_2||\nu) \tag{2.14}$$

(c) Convexity of $\mathrm{D}(\cdot||\nu)$ implies that for a fixed ν, $\mathrm{D}(\mu||\nu)$ is maximized when μ is concentrated at a single point. □

Like relative entropy, separation distance is not symmetric, but it satisfies the following

Lemma 2.17. *For distributions* μ, μ_1, μ_2, ν *and* $0 \le a = 1 - b \le 1$,

- $0 \le \mathrm{sep}(\mu, \nu) \le 1$
- $\mathrm{sep}(\mu, \nu) \le \epsilon \iff \mu \ge (1 - \epsilon)\nu$,
- (Convexity) $\mathrm{sep}(a\mu_1 + b\mu_2, \nu) \le a\,\mathrm{sep}(\mu_1, \nu) + b\,\mathrm{sep}(\mu_2, \nu)$.

The various distances we have defined can all be used to upper bound the total variation distance.

Proposition 2.18. *Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π and μ any distribution. Then the following relations hold*

(a) $p \leq q \implies \|\mu - \pi\|_{p,\pi} \leq \|\mu - \pi\|_{q,\pi}$

(b) $\|\mu - \pi\|_2 \leq 2\|\mu - \pi\|_{\mathrm{TV}}$

(c) $\|\mu - \pi\|_1 \leq \sqrt{2\,\mathrm{D}(\mu||\pi)}$

(d) $\mathrm{D}(\mu||\pi) \leq \log\left(1 + \|\mu - \pi\|_{2,\pi}^2\right)$

(e) $\|\mu - \pi\|_{\mathrm{TV}} \leq \mathrm{sep}(\mu, \pi)$

(f) $\mathrm{D}(\mu||\pi) \leq \mathrm{sep}(\mu, \pi)\log(1/\pi_*)$

where $\|\cdot\|_2$ denotes the usual Euclidean norm.

Proof. **(*a*)** Follows from Jensen's inequality

(*b*) Follows from Cauchy-Shwarz ($\|\cdot\|_2$ is the usual Euclidean distance).

(*c*) Put $f(x) = \mu(x)/\pi(x)$. $\|\mu - \pi\|_1 = \mathbb{E}_\pi(|f - 1|)$ and $\mathrm{D}(\mu||\pi) = \mathbb{E}_\pi(f \ln f)$. Put

$$h(u) = (4 + 2u)(u \log u - u + 1) - 3(u - 1)^2 \tag{2.15}$$

and observe $h(1) = h'(1) = 0$ and $h''(u) \geq 0$ for all $u \geq 0$. Hence by mean value theorem, for $u \geq 0$,

$$h(u) = h(1) + (u - 1)h'(1) + \frac{(u - 1)^2}{2}h''(u') \tag{2.16}$$

for some u' between 1 and u. Hence $h(u) \geq 0$ for $u \geq 0$. Hence

$$\begin{aligned} 3\left(\mathbb{E}_\pi\left[|f-1|\right]\right)^2 &\leq \left(\mathbb{E}_\pi\left[\sqrt{4+2f}\sqrt{f\log f - f + 1}\right]\right)^2 \\ &\leq \mathbb{E}_\pi\left[4+2f\right] \cdot \mathbb{E}_\pi\left[f\log f - f + 1\right] \\ &= 6 \cdot \mathbb{E}_\pi\left[f\log f\right] \end{aligned} \tag{2.17}$$

Hence we have $\|\mu - \pi\|_1^2 \leq 2\,\mathrm{D}(\mu||\pi)$. This inequality is called Pinsker's inequality.

(d) Put $f(x) = \mu(x)/\pi(x)$ and observe that $\mathrm{D}(\mu||\pi) = \log\left(\prod_x f(x)^{\mu(x)}\right)$. Also

$$1 + \|\mu - \pi\|_{2,\pi}^2 = 1 + \sum_x \pi(x)(f(x)-1)^2 = \sum_x \mu(x) f(x) \tag{2.18}$$

Now the result follows by the arithmetic mean-geometric mean inequality.

(e) Suppose $\mathrm{sep}(\mu,\pi) = \epsilon$. Then by Lemma 2.17, $\mu \geq (1-\epsilon)\pi$. Write $\mu = (1-\epsilon)\pi + \epsilon\nu$ for some distribution ν. Then $\|\mu - \pi\|_{\mathrm{TV}} = \epsilon\|\nu\|_{\mathrm{TV}} \leq \epsilon = \mathrm{sep}(\mu,\pi)$.

(f) Suppose $\mathrm{sep}(\mu,\pi) = \epsilon$. Lemma 2.17 implies that $\mu = (1-\epsilon)\pi + \epsilon\nu$ for an appropriate distribution ν. Lemma 2.16 implies $\mathrm{D}(\mu||\pi) \leq \epsilon\,\mathrm{D}(\nu||\pi) \leq \mathrm{sep}(\mu,\pi)\log(1/\pi_*)$. □

2.4 Examples

In this section, we will illustrate some examples of Markov Chains.

Example 2.19. (Random Walk on a Cycle) Fix $N > 2$. Here the state space is identified with $\mathbb{Z}_N$ the ring of integers modulo N. If the chain is currently at state x, then it goes to states $x-1, x$ and $x+1$ with equal probability. The transition matrix

$\mathbb{P}$ is then given by

$$\mathbb{P}(x,y) = \begin{cases} 1/3 & \text{if } y = x+1 \text{ or } y = x-1 \text{ or } y = x \\ 0 & \text{otherwise} \end{cases} \tag{2.19}$$

Figure 2.2 shows the graph for $N = 6$. Self loops ensure that the chain is aperiodic (Otherwise the chain is aperiodic only for odd N). From symmetry, it follows that the stationary distribution is uniform. Since $\mathbb{P}$ is symmetric the chain is reversible.

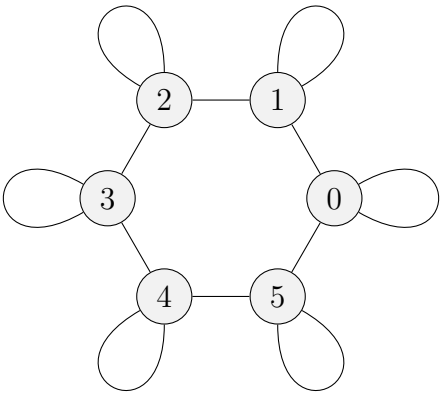


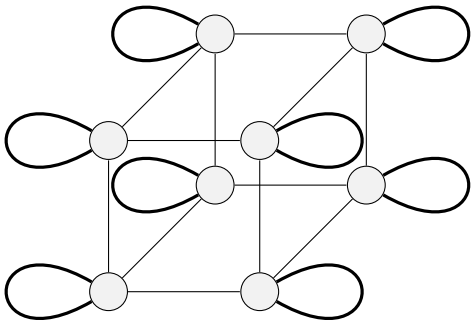
Figure 2.2: Random walk on a cycle

Example 2.20. (Random Walk on Hypercube) Fix $n > 2$. Here the state space $\mathcal{X}$ is the set of all binary vectors of length n. For two binary vectors $\vec{\mathsf{x}}$ and $\vec{\mathsf{y}}$, let $H(\vec{\mathsf{x}}, \vec{\mathsf{y}})$ denote their hamming distance, i. e., the number of coordinates where they differ. Put an edge from $\vec{\mathsf{x}}$ to $\vec{\mathsf{y}}$ if $H(\vec{\mathsf{x}}, \vec{\mathsf{y}}) = 1$ and add self loops (of weight 1/2) to ensure aperiodicity. The graph of the hypercube for $n = 3$ is shown in Figure 2.3.

The transition matrix $\mathbb{P}$ is given by

$$\mathbb{P}(\vec{\mathsf{x}}, \vec{\mathsf{y}}) = \begin{cases} 1/2n & \text{if } H(\vec{\mathsf{x}}, \vec{\mathsf{y}}) = 1 \\ 1/2 & \text{if } \vec{\mathsf{x}} = \vec{\mathsf{y}} \\ 0 & \text{otherwise} \end{cases} \tag{2.20}$$

Again by symmetry the stationary distribution is uniform and $\mathbb{P} = \mathbb{P}^T$ implies $\mathbb{P}$ is reversible. In this case, the number of states $N = 2^n$.



Thick edges have weight 1/2 while thin edges have weight $1/2n$

Figure 2.3: Random walk on hypercube

Example 2.21. (Random Walks on Graphs) Let $X = (\mathcal{X}, E, w)$ be a weighted digraph. Assume that X is strongly connected and that the greatest common divisor of all cycle lengths is 1. Then we can define a Markov Chain with state space $\mathcal{X}$ as follows. Given the current state x, we move to state y with probability $\mathbb{P}(x, y) = w(x, y)/d^+(x)$ where $d^+(x) = \sum_y w(x, y)$ is the out-degree of x. Similarly $d^-(x) = \sum_y w(y, x)$ is the in-degree of x. X is said to be *Eulerian* if all vertices have equal in and out degrees. In addition if all the out-degrees are equal, then the graph is said to be *bi-regular*. Thus undirected graphs are always Eulerian and regular undirected

graphs are bi-regular. The random walk on a bi-regular graph is reversible iff the graph is undirected.

In general the stationary distribution is not obvious from the edge weights. However if X is Eulerian, the stationary distribution is proportional to the out-degrees. A famous examples of a Markov Chain on a non-Eulerian graph is the **Winning streak graph** shown in Figure 2.4. It corresponds to the fortune of a gambler who plays a game where he wins \$1 with probability 1/2 and loses all his winnings with probability 1/2.

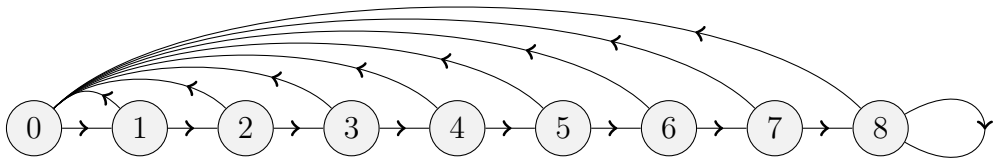


Figure 2.4: The winning streak graph

Note that the Winning streak graph is neither Eulerian nor reversible. The stationary distribution in this case falls off exponentially with $\pi(x) = 2^{-x+1}$ for $x < N$ and $\pi(N) = 2^{-N}$.

Example 2.22. (Cayley walks on Groups) Another class of examples are Cayley walks on groups. Cayley walks give a rich class of examples of Markov Chains with lots of symmetry. The random walk on the cycle as well as the walk on the hypercube can be realized as Cayley walks on appropriate groups.

We start with a group G (written multiplicatively) and a distribution P on G. Assume that the support of P (points of G where P is positive) generates G. The Markov Chain starts at $\mathsf{1}$. If the current state of the chain is $\mathsf{g} \in G$, then we pick

Random-to-top	cycles $(1, 2, \ldots, i)$
Random-transposition	transpositions (i, j)
Adjacent-transposition	adjacent transpositions $(i, i+1)$
Rudvalis Shuffle	transpositions $(1, n)$ and $(1, n-1)$

Figure 2.5: Some card shuffling schemes

$\mathsf{s} \in G$ with probability $P(\mathsf{s})$ and then move to the state $\mathsf{g} \cdot \mathsf{s}$. Because the transition matrix $\mathbb{P}$ is doubly stochastic, the stationary distribution of this chain (if ergodic) is uniform. Also the Cayley walk on G driven by P is reversible iff P is symmetric, i. e., $\forall \mathsf{s} \in G, p(\mathsf{s}) = p(\mathsf{s}^{-1})$.

Popular examples in this category are the card shuffling chains where $G = S_n$ the symmetric group on n letters. Usually P is uniform on a subset S of G which generates it. Figure 2.4 gives some examples of card shuffling chains.

2.5 Singular Value Decomposition

Now we digress a little and recall basic facts about singular value decompositions.

Definition 2.23. Let $\mathbb{A}$ be any $N \times N$-matrix. By a singular value decomposition (SVD) of $\mathbb{A}$, we mean two orthonormal bases $\{\vec{\mathsf{u}}_i\}_{i=0}^{N-1}$, $\{\vec{\mathsf{w}}_i\}_{i=0}^{N-1}$ together with scalars $\sigma_0 \geq \sigma_1 \geq \cdots \geq \sigma_{N-1} \geq 0$ which satisfy

$$(\forall 0 \leq i \leq N-1), (\mathbb{A}\vec{\mathsf{u}}_i = \sigma_i \vec{\mathsf{w}}_i) \text{ and } (A^T \vec{\mathsf{w}}_i = \sigma_i \vec{\mathsf{u}}_i) \tag{2.21}$$

A quick self contained proof that shows that every square matrix has a Singular Value Decomposition. The existence of Singular value decomposition for a matrix $\mathbb{A}$ also follows from the Spectral Theorem applied to the matrix $\mathbb{A}\mathbb{A}^T$.

Proof (adapted from [62]). We assume that A is invertible (the general case can be

handled by limiting arguments). Define subspaces L_i and unit vectors u_i as follows: Let $L_{-1} = \mathbb{R}^N$. For $i \geq 0$, assume we have defined L_{i-1}. Choose a unit vector $\vec{\mathsf{u}}_i \in L_{i-1}$ such that $\|\mathbb{A}\vec{\mathsf{u}}\|_2 \leq \|\mathbb{A}\vec{\mathsf{u}}_i\|_2$ for all unit vectors $\vec{\mathsf{u}}$ in L_{i-1} (such a $\vec{\mathsf{u}}$ exists since the closed unit ball is compact). Having defined $\vec{\mathsf{u}}_i$, let L_i be the orthogonal complement of $\vec{\mathsf{u}}_i$ in L_{i-1}. By construction we have an orthonormal basis $\vec{\mathsf{u}}_0, \ldots, \vec{\mathsf{u}}_{N-1}$ of $\mathbb{R}^N$.

Now define unit vectors $\vec{\mathsf{w}}_i$ and positive scalars σ_i so that $\mathbb{A}\vec{\mathsf{u}}_i = \sigma_i\vec{\mathsf{w}}_i$ holds for all i. Since A is invertible σ_i and $\vec{\mathsf{w}}_i$ are well defined. We now establish that $\vec{\mathsf{w}}_i$ is an orthogonal family of vectors.

We need to show that $\vec{\mathsf{w}}_i \perp \vec{\mathsf{w}}_j$ for $0 \leq i < j \leq N-1$. Fix such an i and j. When we chose the vector $\vec{\mathsf{u}}_i$ in L_{i-1}, $\vec{\mathsf{u}}_j$ was also a candidate (but was rejected). Consider the subspace spanned by $\vec{\mathsf{u}}_i$ and $\vec{\mathsf{u}}_j$ and define $f : [0, 2\pi] \to \mathbb{R}$ via

$$f(\theta) = \|\mathbb{A}(\cos\theta\vec{\mathsf{u}}_i) + \mathbb{A}(\sin\theta\vec{\mathsf{u}}_j)\|_2^2 \tag{2.22}$$

Now f is a real valued differentiable function which attains its maximum at $\theta = 0$. Hence $f'(0) = 0$, which translates to $\mathbb{A}\vec{\mathsf{u}}_i \perp \mathbb{A}\vec{\mathsf{u}}_j$. Hence $\vec{\mathsf{w}}_i \perp \vec{\mathsf{w}}_j$. This gives a decomposition of $A = U\Sigma W^T$, where U and W are unitary matrices with columns $\vec{\mathsf{u}}_i$ and $\vec{\mathsf{w}}_i$ respectively and Σ is a diagonal matrix with entries σ_i. The decomposition implies $A^T = W\Sigma U^T$ and hence $A^T\vec{\mathsf{w}}_i = \sigma_i\vec{\mathsf{u}}_i$ as well. □

We refer to Horn and Johnson [29, Chapter 3] for an introduction to SVD and its properties. Some basic facts about singular values we use without proof:

Lemma 2.24. *Let $\mathbb{A}$ be any real square matrix with singular values $\sigma_0 \geq \sigma_1 \geq \cdots \geq \sigma_{N-1} \geq 0$ and (possibly complex) eigenvalues $\rho_0, \rho_1, \ldots, \rho_{N-1}$ arranged in decreasing absolute value.*

(a) $\mathbb{A}$ *always has an SVD;* σ_i^2 *are the eigenvalues of* $\mathbb{A}\mathbb{A}^T$ *and are uniquely determined by* $\mathbb{A}$.

(b) If L *is an invariant subspace of* $\mathbb{A}$*, then the SVD of* $\mathbb{A}$ *is the direct sum of the SVD of* $\mathbb{A}$ *restricted to* L *and its orthogonal complement.*

(c) $\sigma_0(\mathbb{A})$ *is the Euclidean norm of the operator* A*; in particular if for some vector* $\vec{\mathsf{x}}$*, the subspace spanned by* $\vec{\mathsf{x}}$ *is invariant under* $\mathbb{A}$*, then* $\sigma_1(\mathbb{A})$ *is the Euclidean norm of* $\mathbb{A}$ *restricted to the orthogonal complement of* $\vec{\mathsf{x}}$.

(d) $\sigma_0(\mathbb{A}) \geq |\rho_i|$ *for all* $0 \leq i \leq N-1$.

(e) If $\mathbb{A}$ *is normal* ($\mathbb{A}^T\mathbb{A} = \mathbb{A}\mathbb{A}^T$) *then* $\sigma_i(\mathbb{A}) = |\rho_i(\mathbb{A})|$.

(f) $\lim_{t\to\infty} \sigma_i(\mathbb{A}^t)^{1/t} = |\rho_i|$

2.6 Lower bounds

In this section we address the basic results which allow us to give bounds on the mixing time of a Markov Chain. We first show sub-multiplicativity of the L_1 mixing time. The usual approach to proving sub-multiplicativity is to either use Coupling arguments or expand the entries of $\mathbb{P}^{s+t}$ in terms of the entries of $\mathbb{P}^s$ and $\mathbb{P}^t$. Our approach is very simple and proves sub-multiplicativity in all norms simultaneously. We basically show that sub-multiplicativity is built into the definition of mixing time.

Lemma 2.25. *Let* $\epsilon > 0$ *and* $\mathbb{P}$ *a stochastic matrix for which* $\pi\mathbb{P} = \pi$*. Let* $p \geq 1$ *be arbitrary and suppose that* $\|\mu\mathbb{P} - \pi\|_{p,\pi} \leq \epsilon$*, for all distributions* μ*. Then*

$$\|\mu\mathbb{P} - \pi\|_{p,\pi} \leq \epsilon \min\left(\|\mu - \pi\|_1, 1\right) \tag{2.23}$$

Proof. Let $\delta = \|\mu - \pi\|_{\mathrm{TV}}$. Put $\nu_1 = (\mu - \pi)^+/\delta$ and $\nu_2 = (\pi - \mu)^+/\delta$, where $f^+ = \max(f, 0)$ is the positive part of the function f.

Since $\|\mu - \pi\|_{\mathrm{TV}} = \sum_i (\mu(i) - \pi(i))^+$, it follows that ν_1 and ν_2 are distributions and that $\mu - \pi = \delta(\nu_1 - \nu_2)$. Hence

$$\mu\mathbb{P} - \pi = (\mu - \pi)\mathbb{P} = \delta\left((\nu_1 - \pi)\mathbb{P} - (\nu_2 - \pi)\mathbb{P}\right) \tag{2.24}$$

Therefore

$$\|\mu\mathbb{P} - \pi\|_{p,\pi} \le \delta\left(\|\nu_1\mathbb{P} - \pi\|_{p,\pi} + \|\nu_2\mathbb{P} - \pi\|_{p,\pi}\right) \le 2\delta\epsilon = \epsilon\|\mu - \pi\|_1 \tag{2.25}$$

★

Proposition 2.26. *Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π and $p \ge 1$. Then*

$$\mathcal{T}_p(\epsilon\delta) \le \mathcal{T}_p(\epsilon) + \mathcal{T}_1(\delta) \le \mathcal{T}_p(\epsilon) + \mathcal{T}_p(\delta) \tag{2.26}$$

In particular, $\mathcal{T}(\epsilon\delta/2) \le \mathcal{T}(\epsilon/2) + \mathcal{T}(\delta/2)$.

Proof. Let $t = \mathcal{T}_p(\epsilon)$ and $s = \mathcal{T}_1(\delta)$. Applying Lemma 2.25 we have

$$\|(\mu - \pi)\mathbb{P}^{s+t}\|_{p,\pi} \le \epsilon\|\mu\mathbb{P}^s - \pi\|_1 \le \epsilon\delta \tag{2.27}$$

The statement about $\mathcal{T}(\epsilon)$ follows from the fact that $\mathcal{T}(\epsilon/2) = \mathcal{T}_1(\epsilon)$. ★

Once the variation distance falls below half, it falls at an exponential rate. Also for $p > 1$, we have $\mathcal{T}_p(\epsilon/2) \le \mathcal{T}_p(1/2) + \mathcal{T}(\epsilon/2)$, i. e., except for a different burn in time, the L^p distance falls just like the total variation distance.

Corollary 2.27. *For every $\delta < 1/2$,*

$$\mathcal{T}(\epsilon) \leq \mathcal{T}(\delta) \left\lceil \frac{\log(1/2\epsilon)}{\log(1/2\delta)} \right\rceil \tag{2.28}$$

Before we establish a lower bound for the mixing time, we need the following

Lemma 2.28. *Let $\mathbb{A}$ be a real matrix with (possibly complex) eigenvalue ρ. Then there is a non-zero real vector $\vec{\mathsf{z}}$ such that $\|\vec{\mathsf{z}}\mathbb{A}\|_2 \geq |\rho|\|\vec{\mathsf{z}}\|_2$*

Proof. Assume that $\rho \neq 0$ ($\rho = 0$ is trivial). Let $|\rho| = r$ and $\vec{\mathsf{x}} + \iota\vec{\mathsf{y}}$ be a non-zero left eigenvector corresponding to ρ. Here $\iota = \sqrt{-1}$. If $\vec{\mathsf{y}} = 0$, then ρ is real and we can take $\vec{\mathsf{z}} = \vec{\mathsf{x}}$. If $\vec{\mathsf{x}} = 0$, then again ρ is real and we can take $\vec{\mathsf{z}} = \vec{\mathsf{y}}$.

Suppose both $\vec{\mathsf{x}}$ and $\vec{\mathsf{y}}$ are non-zero and $\|\vec{\mathsf{x}}\mathbb{A}\|_2 < r\|\vec{\mathsf{x}}\|_2$ and $\|\vec{\mathsf{y}}\mathbb{A}\|_2 < r\|\vec{\mathsf{y}}\|_2$. Since $\mathbb{A}, \vec{\mathsf{x}}, \vec{\mathsf{y}}$ are real $\vec{\mathsf{x}}\mathbb{A}$ and $\iota\vec{\mathsf{y}}\mathbb{A}$ are orthogonal to each other. Hence

$$r^2\left(\|\vec{\mathsf{x}} + \iota\vec{\mathsf{y}}\|_2^2\right) = \|\vec{\mathsf{x}}\mathbb{A} + \iota\vec{\mathsf{y}}\mathbb{A}\|_2^2 = \|\vec{\mathsf{x}}\mathbb{A}\|_2^2 + \|\vec{\mathsf{y}}\mathbb{A}\|_2^2 < r^2(\|\vec{\mathsf{x}} + \iota\vec{\mathsf{y}}\|_2^2) \tag{2.29}$$

a contradiction. Hence one of $\vec{\mathsf{x}}$ or $\vec{\mathsf{y}}$ can serve as $\vec{\mathsf{z}}$. □

Proposition 2.29. *Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π. Then*

$$\mathcal{T}_1(\epsilon) \geq \frac{\log(1/\epsilon)}{\log(1/|\rho|)} \tag{2.30}$$

for any eigenvalue $\rho \neq 1$ of $\mathbb{P}$.

Proof. Let $r = |\rho|$ and $\vec{\mathsf{x}} + \iota\vec{\mathsf{y}}$ be the eigenvector corresponding to ρ. Since $\mathbb{P}$ is ergodic, $\pi(i) > 0$ for all states i. So choose $c > 0$ small enough so that $\pi + c\vec{\mathsf{x}}$ and $\pi + c\vec{\mathsf{y}}$ have all positive components. Note that since $\rho \neq 1$, the left eigenvector $\vec{\mathsf{x}} + \iota\vec{\mathsf{y}}$ is orthogonal to the right eigenvector $\vec{1}$ corresponding to the eigenvalue 1. Hence $\vec{\mathsf{x}}$ and $\vec{\mathsf{y}}$ have zero sum and therefore $\pi + c\vec{\mathsf{x}}$ and $\pi + c\vec{\mathsf{y}}$ are both valid initial distributions.

From Lemma 2.28, $\|\vec{\mathsf{z}}_t\mathbb{P}^t\|_2 \geq r^t\|\vec{\mathsf{z}}_t\|_2$ for $\vec{\mathsf{z}}_t = \vec{\mathsf{x}}$ or $\vec{\mathsf{y}}$. If $\Delta_1(t), \Delta_2(t)$ denote the L_1 distance from stationary distribution after t steps with initial distributions $\pi + c\vec{\mathsf{x}}, \pi + c\vec{\mathsf{y}}$ respectively and $\Delta(t) = \max(\Delta_1(t), \Delta_2(t))$. Since $\|\cdot\|_2 \leq \|\cdot\|_1$, we have

$$(\forall t > 0), \quad \Delta(t) \geq cr^t \tag{2.31}$$

Suppose $T = \mathcal{T}_1(\epsilon)$ and put $t = T\alpha$ for some large constant α. Then by sub-multiplicativity, we have $\Delta(t) \leq \epsilon^\alpha$. Hence we have $c^{1/\alpha}r^T \leq \epsilon$. Taking the limit as $\alpha \to \infty$ we have the result. ★

Since $\|\cdot\|_{\mathrm{TV}} = \|\cdot\|_1/2$ this translates to the following

Corollary 2.30. *Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π. Then*

$$\mathcal{T}(\epsilon) \geq \frac{\log(1/2\epsilon)}{\log(1/|\rho|)} \tag{2.32}$$

for any eigenvalue $\rho \neq 1$ of $\mathbb{P}$.

This motivates the following

Definition 2.31. Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π. Its relaxation time, denoted $\mathcal{T}_{\mathrm{rel}}$ is defined as

$$\mathcal{T}_{\mathrm{rel}} = \max_{\rho \neq 1} \frac{1}{\log(1/|\rho|)} \tag{2.33}$$

where the maximum is taken over all non-trivial eigenvalues ρ of $\mathbb{P}$. Equivalently,

$$\mathcal{T}_{\mathrm{rel}} = \min_t \max_{\rho \neq 1} \{|\rho|^t \leq 1/\mathrm{e}\} \tag{2.34}$$

where the maximum is taken over all non-trivial eigenvalues ρ of $\mathbb{P}$.

Note that unlike other measures of mixing time the relaxation time does not have a parameter associated with it. Corollary 2.30 thus shows that the relaxation time is a lower bound on the total variation mixing time. Some authors define the relaxation time to be the inverse spectral gap, i.e. $\max_{\rho \neq 1}(1-|\rho|)^{-1}$. However, for all $0 \leq \theta < 1$, we have

$$\frac{\theta}{1-\theta} \leq \frac{1}{\log(1/\theta)} \leq \frac{1}{1-\theta} = \frac{\theta}{1-\theta} + 1 \tag{2.35}$$

since $\log(1+x) \leq x$ for all $-1 < x < 1$.

We now show that the lower bound given by Proposition 2.29 is asymptotically the best possible.

Theorem 2.32. *Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π and eigenvalues $\{\lambda_i\}$. Let $r = \max_i\{|\lambda_i| \colon \lambda_i \neq 1\}$. Suppose for some $c, \delta > 0$, there exists initial distributions $\{\mu_t\}_{t>0}$ for which $\|\mu_t \mathbb{P}^t - \pi\|_2 \geq \delta c^t$. Then $c \leq r$.*

Proof. We first show that $\mathbb{P}$ is ergodic implies $r < 1$. We already know that since $\mathbb{P}$ is ergodic, 1 is a simple eigenvalue. We only need to show that $|\lambda_i| < 1$ if $\lambda_i \neq 1$. Since λ_i are the roots of a polynomial of degree $N = |\mathcal{X}|$, the only choice for $\lambda_i \neq 1$ and $|\lambda_i| = 1$ is if λ_i is an N'th root of unity. But ergodicity of $\mathbb{P}$ implies that all powers of $\mathbb{P}$ are also ergodic. But if we have a non-trivial N'th root of unity as an eigenvalue, then $\mathbb{P}^N$ will not have 1 as a simple eigenvalue.

Let $S(t) = \sigma_2(\mathbb{P}^t)$. Then $\|\mu_t \mathbb{P}^t - \pi\|_2 = \|(\mu_t - \pi)\mathbb{P}^t\|_2 \leq \sigma_2(\mathbb{P}^t)\|\mu_t - \pi\|_2 \leq 2S(t)$ since μ_t and π are distributions. Hence we have

$$c \leq \left(\frac{2}{\delta}\right)^{1/t} S(t)^{1/t} \tag{2.36}$$

Taking the limit as $t \to \infty$ we have $c \leq \limsup S(t)^{1/t}$. But Lemma 2.24 implies $\limsup S(t)^{1/t} = r$. ★

Note that this immediately implies the same result for all the L_p norms as $\|\cdot\|_2 \geq \delta\|\cdot\|_{p,\pi}$ for a suitable $\delta > 0$ which depends only on the size of the matrix and π.

2.7 L^2 Lower bounds

In this section we derive a lower bound on the L^2 mixing time, which takes all eigenvalues into account.

Definition 2.33. Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π.

- By π_* we mean the smallest stationary weight, i. e., $\pi_* = \min_x \pi(x)$
- By Π we denote the diagonal matrix with entries $\pi(x)$.
- Define $\mathbb{S}(\mathbb{P}) = \sqrt{\Pi}\mathbb{P}\sqrt{\Pi}^{-1}$

$\mathbb{S}(\mathbb{P})$ and $\mathbb{P}$ have the same eigenvalues. Hence our lower bound for $\mathcal{T}(\epsilon)$ could have been stated in terms of the eigenvalues of $\mathbb{S}(\mathbb{P})$ also. Also $\mathbb{P}$ is reversible iff $\Pi\mathbb{P} = \mathbb{P}\Pi$.

Consider what happens if we run the chain with time going backwards starting from stationary distribution. In order to do that we need to calculate

$$\Pr\{\mathbf{X}_t = x | \mathbf{X}_{t+1} = y\} = \frac{\Pr\{\mathbf{X}_t = x\}\Pr\{\mathbf{X}_{t+1} = y | \mathbf{X}_t = x\}}{\Pr\{\mathbf{X}_{t+1} = y\}} = \frac{\pi(x)\mathbb{P}(y,x)}{\pi(y)} \quad (2.37)$$

Definition 2.34. The reverse of the chain $\mathbb{P}$ denoted $\overleftarrow{\mathbb{P}}$ has the same state space $\mathcal{X}$ and transition matrix $\overleftarrow{\mathbb{P}}(x,y) = \pi(y)\mathbb{P}(y,x)/\pi(x)$, i. e., $\overleftarrow{\mathbb{P}} = \Pi^{-1}\mathbb{P}^T\Pi$

Thus $\mathbb{P}$ is reversible iff $\overleftarrow{\mathbb{P}} = \mathbb{P}$. Also $\mathbb{S}(\overleftarrow{\mathbb{P}}) = \mathbb{S}(\mathbb{P})^T$. Thus $\mathbb{P}$ is reversible iff $\mathbb{S}(\mathbb{P})$ is symmetric, in which case it is unitarily diagonalizable with real eigenvalues.

If $\mathbb{P}$ is ergodic 1 is a simple eigenvalue of $\mathbb{P}$. The right eigenvector corresponding to 1 is the constant vector while the left eigenvector corresponding to 1 is the stationary

distribution π. In case the stationary distribution was uniform the constant vector is both a left and right eigenvector of $\mathbb{P}$. In this case, we can write the vector space as a direct sum of the one dimensional space spanned by the constant vector and its orthogonal complement. This makes calculations a lot simpler. However, when π is not the uniform distribution such a clean decomposition is not possible. However, all is not lost. Consider $\mathbb{S}(\mathbb{P})$. Since $\mathbb{S}(\mathbb{P})$ and $\mathbb{P}$ have the same eigenvalues, let us look at the left and right eigenspaces of 1. Let $\sqrt{\pi}$ denote the vector whose components are $\sqrt{\pi(x)}$.

$$\sqrt{\pi}\,\mathbb{S}(\mathbb{P}) = \sqrt{\pi}\sqrt{\Pi}\mathbb{P}\sqrt{\Pi}^{-1} = \pi\mathbb{P}\sqrt{\Pi}^{-1} = \pi\sqrt{\Pi}^{-1} = \sqrt{\pi} \tag{2.38}$$

$$\mathbb{S}(\mathbb{P})\sqrt{\pi} = \sqrt{\Pi}\mathbb{P}\sqrt{\Pi}^{-1}\sqrt{\pi} = \sqrt{\Pi}\mathbb{P}\vec{1} = \sqrt{\Pi}\vec{1} = \sqrt{\pi} \tag{2.39}$$

Thus $\sqrt{\pi}$ is both the left and the right eigenvector of $\mathbb{S}(\mathbb{P})$, and we can decompose the space into the linear span of $\sqrt{\pi}$ and its orthogonal complement. We first observe that $\sigma_0(\mathbb{S}(\mathbb{P})) \leq 1$. This is because $\sigma_0(\mathbb{S}(\mathbb{P})) \leq 1$ iff $\mathbb{S}(\mathbb{P})\mathbb{S}(\mathbb{P})^T$ has its eigenvalues bounded by 1.

$$\mathbb{S}(\mathbb{P})\mathbb{S}(\mathbb{P})^T = \mathbb{S}(\mathbb{P})\,\mathbb{S}(\overleftarrow{\mathbb{P}}) = \mathbb{S}(\mathbb{P}\overleftarrow{\mathbb{P}}) \tag{2.40}$$

Since $\mathbb{P}\overleftarrow{\mathbb{P}}$ is a stochastic matrix all its eigenvalues are bounded by 1. Since $\mathbb{S}(\cdot)$ preserves eigenvalues, it follows $\sigma_0(\mathbb{S}(\mathbb{P})) \leq 1$.

Proposition 2.35. *Let $\mathbb{Q}$ be a stochastic matrix with $\pi\mathbb{Q} = \pi$. For $x \in \mathcal{X}$, let δ_x denote the initial distribution concentrated at x. Then*

$$\|\delta_x\mathbb{Q} - \pi\|_{2,\pi}^2 = \frac{(\mathbb{Q}\overleftarrow{\mathbb{Q}})(x,x)}{\pi(x)} - 1 \tag{2.41}$$

Proof. Observe that $\delta_x \mathbb{Q} = \mathbb{Q}(x, \cdot)$.

$$\begin{aligned} \|\mathbb{Q}(x,\cdot) - \pi(\cdot)\|_{2,\pi}^2 &= \sum_y \frac{(\mathbb{Q}(x,y) - \pi(y))^2}{\pi(y)} \\ &= \sum_y \frac{\mathbb{Q}(x,y)^2}{\pi(y)} - 2\mathbb{Q}(x,y) + \pi(y) \\ &= \sum_y \frac{\mathbb{Q}(x,y)\overleftarrow{\mathbb{Q}}(y,x)}{\pi(x)} - 1 = \frac{(\mathbb{Q}\overleftarrow{\mathbb{Q}})(x,x)}{\pi(x)} - 1 \end{aligned} \tag{2.42}$$

since $\pi(x)\mathbb{Q}(x,y) = \pi(y)\overleftarrow{\mathbb{Q}}(y,x)$. □

This immediately gives the following

Proposition 2.36. *Let $\mathbb{P}$ be a Markov Chain with stationary distribution π and let $1 = \rho_0, \rho_1, \ldots, \rho_{N-1}$ be the (possibly complex) eigenvalues of $\mathbb{P}$ arranged in decreasing absolute value. For $t > 0$, there exists an initial state $x \in \mathcal{X}$ for which*

$$\|\mathbb{P}^t(x,\cdot) - \pi\|_{2,\pi}^2 \geq \sum_{j=1}^{N-1} |\rho_i|^{2t} \tag{2.43}$$

Proof. Fix $t > 0$, and consider $\mathbb{Q} = \mathbb{P}^t$ with eigenvalues $\{\rho_i^t\}$. Let $x \in \mathcal{X}$ be arbitrary. Applying Proposition 2.35 to $\mathbb{Q}$, we see that

$$\|\mathbb{Q}(x,\cdot) - \pi\|_{2,\pi}^2 = \frac{(\mathbb{Q}\overleftarrow{\mathbb{Q}})(x,x)}{\pi(x)} - 1 \tag{2.44}$$

Now averaging over x,

$$\begin{aligned}\mathbb{E}_\pi\left[\|\mathbb{Q}(x,\cdot)-\pi\|_{2,\pi}^2\right] &= \mathbb{E}_\pi\left[\frac{(\mathbb{Q}\overleftarrow{\mathbb{Q}})(x,x)}{\pi(x)}\right]-1 \\ &= \mathrm{tr}(\mathbb{Q}\overleftarrow{\mathbb{Q}})-1 \\ &= \sum_{i>0}\sigma_i(\mathbb{S}(\mathbb{Q}))^2\end{aligned} \tag{2.45}$$

from (2.40). Since the singular values of a matrix majorize the modulus of the eigenvalues (see [61] for a proof), we have

$$\sum_{i\geq 0}\sigma_i(\mathbb{S}(\mathbb{Q}))^2 \geq \sum_{i\geq 0}|\rho_i(\mathbb{S}(\mathbb{Q}))|^2 = \sum_{i\geq 0}|\rho_i(\mathbb{Q})|^2 = \sum_{i\geq 0}|\rho_i(\mathbb{P})|^{2t} \tag{2.46}$$

since $\mathbb{S}(\cdot)$ is a similarity transformation. Here $\{\rho_i(\mathbb{A})\}$ are the eigenvalues of $\mathbb{A}$ arranged in decreasing absolute value.

The terms corresponding to $i=0$ are all equal to 1 in (2.46). Hence (2.45) implies the result. ★

This allows us to deduce an lower bound on the L^2 mixing time of a Markov Chain. Note that unlike Proposition 2.29, this lower bound takes the multiplicity of the eigenvalues into account.

Corollary 2.37. *Let $\mathbb{P}$ be a Markov Chain with stationary distribution π and for $0<\theta<1$, let m_θ denote the number of non-trivial eigenvalues (with multiplicity) of $\mathbb{P}$ with absolute value $\geq\theta$.*

$$\mathcal{T}_2(\mathbb{P},\epsilon) \geq \frac{(\log m_\theta)/2+\log(1/\epsilon)}{\log(1/\theta)} \tag{2.47}$$

2.8 Upper Bounds

We now turn to bounding the mixing time from above.

Proposition 2.38. *Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π. Let $\sigma_1 = \sigma_1(\mathbb{S}(\mathbb{P}))$ denote the second largest singular value of $\mathbb{S}(\mathbb{P})$. For any initial distribution μ,*

$$\|\mu\mathbb{P}^t - \pi\|_{2,\pi} \leq \sigma_1^t \|\mu - \pi\|_{2,\pi} \tag{2.48}$$

Proof. Since $\sqrt{\pi}\,\mathbb{S}(\mathbb{P}) = \mathbb{S}(\mathbb{P})\sqrt{\pi} = \sqrt{\pi}$, the subspace generated by $\sqrt{\pi}$ is invariant under $\mathbb{S}(\mathbb{P})$. Hence σ_1 is the Euclidean norm of the operator $\mathbb{S}(\mathbb{P})$ on the orthogonal complement of $\sqrt{\pi}$.

Put $\vec{\mathsf{z}} = (\mu - \pi)$ and $\vec{\mathsf{y}} = \vec{\mathsf{z}}\sqrt{\Pi}^{-1}$ and note that $\vec{\mathsf{y}}$ is orthogonal to $\sqrt{\pi}$. Hence it follows that for all $t > 0$, $\|\vec{\mathsf{y}}\,\mathbb{S}(\mathbb{P})^t\|_2 \leq \sigma_1^t \|\vec{\mathsf{y}}\|_2$. Note that

$$\vec{\mathsf{y}}\,\mathbb{S}(\mathbb{P})^t = \vec{\mathsf{y}}\,\mathbb{S}(\mathbb{P}^t) = \vec{\mathsf{z}}\mathbb{P}^t\sqrt{\Pi}^{-1} \tag{2.49}$$

and $\|\vec{\mathsf{x}}\sqrt{\Pi}^{-1}\|_2 = \|\vec{\mathsf{x}}\|_{2,\pi}$. Hence we have

$$\|\mu\mathbb{P}^t - \pi\|_{2,\pi} = \|z\mathbb{P}^t\|_{2,\pi} = \|z\mathbb{P}^t\sqrt{\Pi}^{-1}\|_2 \leq \sigma_1^t \|\vec{\mathsf{z}}\sqrt{\Pi}^{-1}\|_2 = \sigma_1^t \|\mu - \pi\|_{2,\pi} \tag{2.50}$$

□

Corollary 2.39. *Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π and $\sigma_1 = \sigma_1(\mathbb{S}(\mathbb{P}))$, then*

$$\mathcal{T}(\epsilon/2) = \mathcal{T}_1(\epsilon) \leq \mathcal{T}_2(\epsilon) \leq \frac{\log(1/\epsilon) + \log(1/\sqrt{\pi_*})}{\log(1/\sigma_1)} \tag{2.51}$$

Proof. Since the worst case initial distribution is a point distribution. The maximum value for $\|\mu - \pi\|_{2,\pi}$ is when μ is concentrated at state x for which $\pi(x) = \pi_*$. In that

case,

$$\|\mu - \pi\|_{2,\pi}^2 = \frac{(1-\pi_*)^2}{\pi_*} + (1-\pi_*) = \frac{1-\pi_*}{\pi_*} \tag{2.52}$$

This together with Proposition 2.38 gives the upper bound on $\mathcal{T}_2(\epsilon)$. Standard relations between $\|\cdot\|_{\mathrm{TV}}, \|\cdot\|_1$ and $\|\cdot\|_{2,\pi}/2$ gives the rest. □

For reversible chains (and more generally when $\mathbb{P}$ and $\overleftarrow{\mathbb{P}}$ commute), the difference between the upper bound and lower bound is not too much.

Theorem 2.40. *Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π and $\mathbb{P}\overleftarrow{\mathbb{P}} = \overleftarrow{\mathbb{P}}\mathbb{P}$. Let $1 = \rho_0, \rho_1, \ldots, \rho_{N-1}$ be the eigenvalues of $\mathbb{P}$, arranged in decreasing modulus. Then*

$$\log\left(\frac{1}{2\epsilon}\right) \leq \frac{\mathcal{T}(\epsilon)}{\mathcal{T}_{\mathrm{rel}}} \leq \log\left(\frac{1}{2\epsilon}\right) + \frac{\log\frac{1}{\pi_*}}{2} \tag{2.53}$$

Proof. Since $\mathbb{P}\overleftarrow{\mathbb{P}} = \overleftarrow{\mathbb{P}}\mathbb{P}$, $\mathbb{S}(\mathbb{P})$ is normal. Thus $\sigma_i(\mathbb{S}(\mathbb{P})) = |\rho_i(\mathbb{S}(\mathbb{P}))| = |\rho_i(\mathbb{P})$. Hence we can take $\rho = \rho_1$ in Corollary 2.30 and $\sigma_1 = |\rho_1|$ in Corollary 2.39. Combining we have the result. □

Cayley walks on abelian groups are an important class of Markov Chains where $\mathbb{P}$ commutes with $\overleftarrow{\mathbb{P}}$. For chains $\mathbb{P}$ which do not commute with $\overleftarrow{\mathbb{P}}$ there can be a substantial difference between the lower bound and the upper bound. In particular it is possible for an ergodic Markov Chain $\mathbb{P}$ to have $\sigma_1 = 1$ in which case Proposition 2.38 gives us no bound on the mixing time.

Example 2.41. (no immediate reversal walk) Let X be a d-regular graph which is aperiodic. Hence the stationary distribution is uniform. The usual random walk on X has states as the vertices. Consider the following random walk. The states are the directed edges of X, i. e., $\mathcal{X} = \{(u,v) : \{u,v\} \text{ is an edge of } X\}$ and the transition

rule is as follows:

$$\mathbb{P}((u,v),(w,x)) = \begin{cases} 1/d & \text{if } w = v \text{ and } \{v,x\} \text{ is an edge of X} \\ 0 & \text{otherwise} \end{cases} \tag{2.54}$$

The new walk is exactly like the usual random walk on X, except that at each step we remember the previous vertex as well. Since we never make use of that information in the new walk, the mixing time of the new walk and the usual walk should be the same (choosing a random edge going out of a random vertex is the same as choosing a random edge). However, $\sigma_1(\mathbb{S}(\mathbb{P})) = 1$ in this case. In fact $\sigma_\ell(\mathbb{S}(\mathbb{P})) = 1$ for $\ell = 0, \ldots, n-1$ and $|\mathcal{X}| = N = nd$. Hence it is possible for a Markov Chain to have a constant fraction of its top singular values as 1 and still mix fast.

If $\mathbb{P}$ has sufficient holding probabilities, then $\lambda_*((\mathbb{P} + \overleftarrow{\mathbb{P}})/2)$ and $\sigma_1(\mathbb{S}(\mathbb{P}))$ can be bounded in terms of each other. This was first observed by [41] and used by [20] to bound the mixing time of non-reversible Markov Chains. This allows us to estimate the mixing times of non-reversible Markov Chains with sufficient holding probabilities in terms of a related reversible Markov Chain.

Proposition 2.42. *Let $\mathbb{A}$ be a stochastic matrix and $\pi\mathbb{A} = \pi$ for a distribution π. Let $\delta > 0$ such that $\mathbb{A}(x,x) \geq \delta$. Put $\sigma_1 = \sigma_1(\mathbb{S}(\mathbb{A}))$ and $\lambda_1 = \lambda_1((\mathbb{A} + \overleftarrow{\mathbb{A}})/2)$. Then $\sigma_1^2 \leq (1-2\delta) + 2\delta\lambda_1$ and $\lambda_1 \leq \sigma_1$.*

In particular, if $\delta = 1/2$, $\sigma_1^2 \leq \lambda_1 \leq \sigma_1$.

Proof. Write $\mathbb{A} = \delta I + (1-\delta)\mathbb{A}'$ for stochastic $\mathbb{A}'$ which also satisfies $\pi\mathbb{A}' = \pi$. Let $\mathbb{B}' = \mathbb{S}(\mathbb{A}')$, $\sigma_1 = \sigma_1(\mathbb{S}(\mathbb{A}))$ and $\lambda_1 = \lambda_1((\mathbb{A} + \overleftarrow{\mathbb{A}})/2)$.

Since $\mathbb{S}$ is linear, preserves eigenvalues and satisfies $\mathbb{S}(\overleftarrow{\mathbb{C}}) = \mathbb{S}(\mathbb{C})^T$ we have $\mathbb{S}(\mathbb{A}) =$

$\delta I + (1-\delta)\mathbb{B}'$ and

$$\lambda_1((\mathbb{B}+\mathbb{B}^T)/2) = \delta + (1-\delta)\lambda_1((\mathbb{B}'+\mathbb{B}'^T)/2) \tag{2.55}$$

Since $\sqrt{\pi}$ is an invariant subspace for both I and $\mathbb{B}'$, we have the following extremal characterization of σ_1 and λ_1.

$$\sigma_1 = \max_{\vec{x}\perp\sqrt{\pi},\|\vec{x}\|_2=1} \|\mathbb{A}'\vec{x}\|_2 \quad \text{and} \quad \lambda_1 = \max_{\vec{x}\perp\sqrt{\pi},\|\vec{x}\|_2=1} \langle \mathbb{A}'\vec{x},\vec{x}\rangle \tag{2.56}$$

Since $|\langle \mathbb{A}'\vec{x},\vec{x}\rangle| \le \|\mathbb{A}'\vec{x}\|_2\|\vec{x}\|_2$, we have $\lambda_1 \le \sigma_1$.

Choose $\vec{x}\perp\sqrt{\pi}$, $\|\vec{x}\|_2 = 1$ such that $\|S(\mathbb{A})\vec{x}\|_2 = \sigma_1$. We now have

$$\begin{aligned}
\sigma_1^2 &= \|\mathbb{S}(\mathbb{A})\vec{x}\|_2^2 \\
&= \|\delta\vec{x} + (1-\delta)\mathbb{B}'\vec{x}\|_2^2 \\
&= \delta^2\|\vec{x}\|_2^2 + (1-\delta)^2\|\mathbb{B}'\vec{x}\|_2^2 + \delta(1-\delta)\langle\vec{x},\mathbb{B}'\vec{x}\rangle + \delta(1-\delta)\langle\mathbb{B}'\vec{x},\vec{x}\rangle \\
&\le \delta^2 + (1-\delta)^2 + 2\delta(1-\delta)\left\langle\vec{x},\frac{\mathbb{B}'+\mathbb{B}'^T}{2}\vec{x}\right\rangle
\end{aligned} \tag{2.57}$$

where we used the fact that $\mathbb{A}'$ is stochastic implies all singular values of $\mathbb{B}'$ are bounded above by 1.

Now using (2.55) we have

$$\sigma_1^2 \le \delta^2 + (1-\delta)^2 + 2\delta(1-\delta)\frac{\lambda_1-\delta}{1-\delta} = (1-2\delta) + 2\delta\lambda_1 \tag{2.58}$$

★

Thus if we add sufficient holding probabilities to a non-reversible Markov Chain, it cannot mix much slower than its additive symmetrization. However, usually non-reversible chains mix faster than their symmetrizations. Even in the case of reversible

chains, the gap between the lower bound and upper bound can be improved in some cases. The mixing time bound given by Proposition 2.38 only takes into account the largest non-trivial singular value of $\mathbb{S}(\mathbb{P})$. In many cases when all the singular values are known, the multiplicative overhead of $O(\log 1/\pi_*)$ in the upper bound can be reduced. This is usually the case when the Markov Chain has lots of symmetries.

Proposition 2.43. *Let $\mathbb{P}$ be an ergodic Markov Chain with uniform stationary distribution π. Let $x \in \mathcal{X}$ be arbitrary. Then for $t > 0$,*

$$\|\mathbb{P}^t(x,\cdot) - \pi\|_{2,\pi}^2 = N(\mathbb{P}^t \overleftarrow{\mathbb{P}}^t)(x,x) - 1 \tag{2.59}$$

Moreover, if $\mathbb{P}^t \overleftarrow{\mathbb{P}}^t$ has equal diagonal entries,

$$4\|\mu\mathbb{P}^t - \pi\|_{\mathrm{TV}}^2 \leq \|\mu\mathbb{P}^t - \pi\|_{2,\pi}^2 \leq \left(\mathrm{tr}\left(\mathbb{P}^t \overleftarrow{\mathbb{P}}^t\right) - 1\right) \tag{2.60}$$

where $\mathrm{tr}(\mathbb{A})$ is the trace of the matrix $\mathbb{A}$.

Proof. Applying Proposition 2.35 with $\mathbb{Q} = \mathbb{P}^t$ we get

$$\|\mathbb{P}^t(x,\cdot) - \pi\|_{2,\pi}^2 = N(\mathbb{P}^t \overleftarrow{\mathbb{P}}^t)(x,x) - 1 \tag{2.61}$$

In case the diagonal entries of $\mathbb{P}^t \overleftarrow{\mathbb{P}}^t$ are all equal, then $N(\mathbb{P}^t \overleftarrow{\mathbb{P}}^t)(x,x) = \mathrm{tr}(\mathbb{P}^t \overleftarrow{\mathbb{P}}^t)$ gives the result when $\mu = \delta_x$ for some $x \in \mathcal{X}$. By convexity the equality becomes an inequality for an arbitrary initial distribution μ. □

If $\mathbb{P}$ is the transition matrix of an ergodic Markov Chain for which $\mathbb{P}^t \overleftarrow{\mathbb{P}}^t$ has equal diagonal entries, then the stationary distribution has to be uniform. This follows from the identity:

$$\begin{aligned}\lim_{t\to\infty}(\mathbb{P}^t\overleftarrow{\mathbb{P}}^t)(x,x) &= \lim_{t\to\infty}\left\{\sum_y \mathbb{P}^t(x,y)\overleftarrow{\mathbb{P}}^t(y,x)\right\} \\ &= \sum_y \left\{\lim_{t\to\infty}\mathbb{P}^t\right\}(x,y)\cdot\left\{\lim_{t\to\infty}\overleftarrow{\mathbb{P}}^t\right\}(y,x) \\ &= \sum_y \pi(y)\pi(x) = \pi(x)\end{aligned} \tag{2.62}$$

where the last equality comes from the fact that both $\mathbb{P}^t$ and $\overleftarrow{\mathbb{P}}^t$ converge to matrix whose columns are π.

Definition 2.44. A Markov Chain $\mathbb{P}$ is said to be *vertex transitive* iff for any pair of states x_1, x_2, there is a bijection $F : \mathcal{X} \to \mathcal{X}$ which satisfies

$$\forall y_1, y_2 \in \mathcal{X}, \mathbb{P}(y_1, y_2) = \mathbb{P}(F(y_1), F(y_2)) \quad \text{and} \quad F(x_1) = x_2 \tag{2.63}$$

Note that if $\mathbb{P}$ is vertex transitive then the stationary distribution has to be uniform. Random walk on groups are the most common examples of vertex transitive Markov Chains. The most common application of Proposition 2.43 is via the following

Corollary 2.45. *Let $\mathbb{P}$ be a ergodic reversible vertex transitive Markov Chain with eigenvalues $\{\lambda_i\}_{i=0}^{N-1}$. Then for any initial distribution μ*

$$4\|\mu\mathbb{P}^t - \pi\|_{\text{TV}}^2 \le \|\mu\mathbb{P}^t - \pi\|_{2,\pi}^2 \le \sum_{i\ge 1}\lambda_i^{2t} \tag{2.64}$$

Example 2.46. Consider the lazy random walk $\mathbb{P}$ on the n-dimensional hypercube. Here $|\mathcal{X}| = 2^n$. It is easy to see that flipping 0 and 1's and permuting the coordinates leave the chain unchanged. Hence the chain is vertex transitive. By looking at the characters of the abelian group $\mathbb{Z}_2^n$, one can show that the eigenvalues of $\mathbb{P}$ are ℓ/n for $\ell = 0, \ldots, n$ and the multiplicity of $\frac{\ell}{n}$ is $\binom{n}{\ell}$. If we only consider $\lambda_1 = 1 - 1/n$,

we get $\mathcal{T}(1/4) = \Omega(n)$ and $\mathcal{T}(1/4) = O(n \log N) = O(n^2)$. But using Corollary 2.45 it follows that $\mathcal{T}(1/4) \leq \mathcal{T}_2(1/2) \leq (n \log n)/2 + O(n)$ substantially better than $O(n^2)$ bound. Since the multiplicity of $1 - 1/n$ is n, Corollary 2.37 implies $\mathcal{T}_2(1/2) \geq (n \log n)/2 - O(n)$ as well.

Even though we only have $\mathcal{T}(1/4) = \Omega(n)$, the actual answer is $\mathcal{T}(1/4) \sim \frac{n \log n}{2}$ so the upper bound is more correct than the lower bound. The lower bound of $\frac{n \log n}{2}$ for $\mathcal{T}$ can be established using a coupon-collector argument.

2.9 Relation Between Mixing Measures

In this section, we derive relations between mixing times under various distance measures. See § 2.2 for the definitions.

Proposition 2.47. *Let $\mathbb{P}$ be a Markov Chain with stationary distribution π, $\epsilon > 0$. Recall $\pi_* = \min_x \pi(x)$. Then*

(a) $1 \leq p \leq q \leq \infty \implies \mathcal{T}_p(\epsilon) \leq \mathcal{T}_q(\epsilon)$

(b) $\mathcal{T}(\epsilon) \leq \mathcal{T}_{\mathrm{D}}(2\epsilon^2)$

(c) $\mathcal{T}(\epsilon) \leq \mathcal{T}_{\mathrm{f}}(\epsilon)$

(d) $\mathcal{T}_{\mathrm{f}}(\epsilon) \leq \mathcal{T}_{\infty}(\epsilon)$

(e) $\mathcal{T}_{\mathrm{D}}(\epsilon) \leq \mathcal{T}_2(\sqrt{\epsilon})$

(f) $\mathcal{T}_{\infty}(\epsilon) \leq \mathcal{T}_1(\epsilon\pi_*) \leq \mathcal{T}_1(\epsilon)^{\frac{\log(1/\pi_*)}{\log(1/\epsilon)}}$

Proof. Results (a) – (c) are direct consequences of Proposition 2.18.

(d) This follows from the relation $\mathrm{sep}(\mu_t, \pi) \leq \|\mu_t - \pi\|_{\infty}$

(e) From Proposition 2.18 it follows that for all $t \geq 0$,

$$\mathrm{D}(\mu_t||\pi) \leq \log(1 + \|\mu_t - \pi\|_{2,\pi}^2) \leq \|\mu_t - \pi\|_{2,\pi}^2 \tag{2.65}$$

Hence $\|\mu_t - \pi\|_{2,\pi} \leq \sqrt{\epsilon}$ implies $\mathrm{D}(\mu_t||\pi) \leq \epsilon$.

(f)

$$\|\mu_t - \pi\|_\infty = \max_x \left| \frac{\mu_t(x) - \pi(x)}{\pi(x)} \right| \leq \frac{1}{\pi_*} \sum_x |\mu_t(x) - \pi(x)| = \frac{\|\mu_t - \pi\|_1}{\pi_*} \tag{2.66}$$

Hence $\|\mu_t - \pi\|_1 \leq \epsilon\pi_* \implies \|\mu_t - \pi\|_\infty \leq \epsilon$. Hence $\mathcal{T}_\infty(\epsilon) \leq \mathcal{T}_1(\epsilon\pi_*)$. Sub-multiplicativity implies $\mathcal{T}_1(\epsilon\pi_*) \leq \mathcal{T}_1(\epsilon)\log(1/\pi_*)/\log(1/\epsilon)$ □

Lemma 2.48. *For an ergodic Markov Chain $\mathbb{P}$, $\mathcal{T}_{\mathrm{f}}(\mathbb{P}) = \mathcal{T}_{\mathrm{f}}(\overleftarrow{\mathbb{P}})$ and $\mathcal{T}_\infty(\mathbb{P}) = \mathcal{T}_\infty(\overleftarrow{\mathbb{P}})$*

Proof. From the definition of $\overleftarrow{\mathbb{P}}$ we have

$$\forall t > 0, \forall x \in \mathcal{X}, \forall y \in \mathcal{X}, \frac{\overleftarrow{\mathbb{P}}^t(x,y)}{\pi(y)} = \frac{\mathbb{P}^t(y,x)}{\pi(x)} \tag{2.67}$$

$\mathcal{T}_{\mathrm{f}}(\mathbb{P}, \epsilon)$ is the smallest t for which $\forall x, y \in \mathcal{X}$, $\mathbb{P}^t(y,x) \geq (1-\epsilon)\pi(x)$ and $\mathcal{T}_\infty(\mathbb{P}, \epsilon)$ is the smallest t for which $(1-\epsilon)\pi(x) \leq \mathbb{P}^t(y,x) \leq (1+\epsilon)\pi(x)$. Hence the result. □

We now relate the filling time to the mixing time.

Lemma 2.49 (Sinclair [58, Lemma 7])**.** *Let $\mathbb{Q}$ be a stochastic matrix whose rows and columns are indexed by $\mathcal{X}$ and $\pi\mathbb{Q} = \pi$ for a distribution π. Fix $\epsilon > 0$ and suppose that for all $x \in \mathcal{X}$, $\|\mathbb{Q}(x,\cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leq \epsilon^2/2$. For $y \in \mathcal{X}$, put $Z_y = \{z \in \mathcal{X} : \mathbb{Q}(y,z) \geq (1-\epsilon)\pi(z)\}$. Then $\pi(Z_y) \geq 1 - \epsilon/2$ for all $y \in \mathcal{Y}$.*

Proof. Let $\mu_y = \mathbb{Q}(y,\cdot)$ be the distribution after one application of $\mathbb{Q}$ starting from

initial state y. Then from Lemma 2.12 we have

$$\epsilon^2/2 \geq \|\mu_y - \pi\|_{\mathrm{TV}} = \sum_{z\in\mathcal{X}} (\pi(z) - \mu_y(z))^+ \geq \sum_{z\notin Z(y)} (\pi(z) - \mu_y(z)) \geq \epsilon(1 - \pi(Z_y)) \quad (2.68)$$

Hence $\pi(Z_y) \geq 1 - \epsilon/2$. □

Proposition 2.50. *Let $\mathbb{P}$ be an ergodic Markov Chain. Then*

$$\mathcal{T}_{\mathrm{f}}(\mathbb{P}, 2\epsilon - \epsilon^2/2 - \epsilon^3/2) \leq \mathcal{T}(\mathbb{P}, \epsilon^2/2) + \mathcal{T}(\overleftarrow{\mathbb{P}}, \epsilon^2/2) \quad (2.69)$$

In particular $\mathcal{T}_{\mathrm{f}}(\mathbb{P}, 13/16) \leq \mathcal{T}(\mathbb{P}, 1/8) + \mathcal{T}(\overleftarrow{\mathbb{P}}, 1/8)$.

Proof. Let $s_1 = \mathcal{T}(\mathbb{P}, \epsilon^2)$, $s_2 = \mathcal{T}(\overleftarrow{\mathbb{P}}, \epsilon^2)$ and $t \geq s_1 + s_2$. Fix $y \in \mathcal{X}$ and let $x \in \mathcal{X}$ be an arbitrary initial state. For any $\mathcal{Y} \subseteq \mathcal{X}$, we have

$$\mathbb{P}^t(x, y) = \sum_{z\in\mathcal{X}} \mathbb{P}^{t-s_2}(x, z)\mathbb{P}^{s_2}(z, y) \geq \sum_{z\in\mathcal{Y}} \mathbb{P}^{t-s_2}(x, z)\overleftarrow{\mathbb{P}}^{s_2}(y, z)\frac{\pi(y)}{\pi(z)} \quad (2.70)$$

Define

$$A = \{z | \mathbb{P}^{t-s_2}(x, z) \geq (1 - \epsilon)\pi(z)\} \qquad \text{and} \qquad B = \{z | \overleftarrow{\mathbb{P}}^{s_2}(y, z) \geq (1 - \epsilon)\pi(z)\} \quad (2.71)$$

By Lemma 2.49 and $t - s_2 \geq s_1$ we have $\pi(A \cap B) \geq 1 - \epsilon$. Hence taking $\mathcal{Y} = A \cap B$

in (2.70), we have

$$\begin{aligned}
\mathbb{P}^t(x,y) &\geq \sum_{z\in A\cap B} \mathbb{P}^{t-s_2}(x,z)\overleftarrow{\mathbb{P}}^{s_2}(y,z)\frac{\pi(y)}{\pi(z)} \\
&\geq (1-\epsilon)\pi(y)\sum_{z\in A\cap B}\mathbb{P}^{t-s_2}(x,z) \\
&= (1-\epsilon)\pi(y)\mathbb{P}^{t-s_2}(x,A\cap B) \\
&\geq (1-\epsilon)\pi(y)\left(\pi(A\cap B) - \|\mathbb{P}^{t-s_2}(x,\cdot)-\pi\|_{\mathrm{TV}}\right) \\
&\geq \pi(y)(1-\epsilon)(1-\epsilon-\epsilon^2/2)
\end{aligned} \tag{2.72}$$

Hence for all $x, y \in \mathcal{X}$, we have $\mathbb{P}^t(x,y) \geq (1-2\epsilon+\epsilon^2/2+\epsilon^3/2)\pi(y)$. ★

Proposition 2.50 was proved by [58] for reversible chains. Our proof is an extension of the same idea to general Markov Chains. As a consequence we are able to relate the filling time to the entropy mixing time as well. We start by establishing sub-multiplicativity of $\mathcal{T}_{\mathrm{f}}(\cdot)$.

Proposition 2.51. *Let $\mathbb{P}$ be a Markov Chain with stationary distribution π. For $\epsilon, \delta > 0$, $\mathcal{T}_{\mathrm{f}}(\epsilon\delta) \leq \mathcal{T}_{\mathrm{f}}(\epsilon) + \mathcal{T}_{\mathrm{f}}(\delta)$*

Proof. Let $S = \mathcal{T}_{\mathrm{f}}(\epsilon), T = \mathcal{T}_{\mathrm{f}}(\delta)$ and μ any initial distribution. By choice of S, $\mu\mathbb{P}^S = (1-\epsilon)\pi + \epsilon\nu$ for some distribution ν_1 and hence

$$\mu\mathbb{P}^{S+T} = (1-\epsilon)\pi\mathbb{P}^T + \epsilon\nu\mathbb{P}^T = (1-\epsilon)\pi + \epsilon(1-\delta)\pi + \epsilon\delta\nu_2 \tag{2.73}$$

for some distribution ν_2. Hence $\mathrm{sep}(\mu\mathbb{P}^{S+T},\pi) \leq \epsilon\delta\,\mathrm{sep}(\nu_2,\pi) \leq \epsilon\delta$. □

Proposition 2.52. *Let $\mathbb{P}$ be a Markov Chain with stationary distribution π. Then*

$$\mathcal{T}_{\mathrm{D}}(\epsilon) \leq \mathcal{T}_{\mathrm{f}}\left(\frac{\epsilon}{\log(1/\pi_*)}\right) \leq \mathcal{T}_{\mathrm{f}}(\epsilon)\left(1+\left\lceil\frac{\log\log(1/\pi_*)}{\log(1/\epsilon)}\right\rceil\right) \tag{2.74}$$

Proof. Let $\delta = \epsilon/\log(1/\pi_*)$ and $T = \mathcal{T}_{\mathrm{f}}(\delta)$. For any initial distribution μ, we have $\mathrm{sep}(\mu\mathbb{P}^t, \pi) \leq \delta$ and hence by Proposition 2.18 $\mathrm{D}(\mu\mathbb{P}^t||\pi) \leq \delta \log(1/\pi_*) = \epsilon$. Hence $\mathcal{T}_{\mathrm{D}}(\epsilon) \leq \mathcal{T}_{\mathrm{f}}(\delta)$. The second inequality is just a restatement of Proposition 2.51. ★

Relating uniform mixing time to the L^2 mixing time is easier. We just reproduce the proof in [46].

Proposition 2.53. *Let $\mathbb{P}$ be a Markov Chain. Then for $1/p + 1/q = 1$,*

$$\mathcal{T}_\infty(\mathbb{P}, \epsilon\delta) \leq \mathcal{T}_p(\mathbb{P}, \epsilon) + \mathcal{T}_q(\overleftarrow{\mathbb{P}}, \delta) \tag{2.75}$$

In particular $\mathcal{T}_\infty(\mathbb{P}, \epsilon\delta) \leq \mathcal{T}_2(\mathbb{P}, \epsilon) + \mathcal{T}_2(\overleftarrow{\mathbb{P}}, \delta)$

Proof. For $x, y \in \mathcal{X}$ and $s, t > 0$, we have

$$\begin{aligned} \frac{\mathbb{P}^{s+t}(x,y) - \pi(y)}{\pi(y)} &= \sum_{z\in\mathcal{X}} \pi(z) \frac{\mathbb{P}^s(x,z) - \pi(z)}{\pi(z)} \cdot \frac{\mathbb{P}^t(z,y) - \pi(y)}{\pi(y)} \\ &= \sum_{z\in\mathcal{X}} \pi(z) \frac{\mathbb{P}^s(x,z) - \pi(z)}{\pi(z)} \cdot \frac{\overleftarrow{\mathbb{P}}^t(y,z) - \pi(z)}{\pi(z)} \end{aligned} \tag{2.76}$$

since $\pi(y)\mathbb{P}^t(y,z) = \pi(z)\overleftarrow{\mathbb{P}}^t(z,y)$.

Hence Hölder's inequality gives

$$\begin{aligned} \left|\frac{\mathbb{P}^{s+t}(x,y) - \pi(y)}{\pi(y)}\right| &\leq \sum_{z\in\mathcal{X}} \pi(z) \left|\frac{\mathbb{P}^s(x,z) - \pi(z)}{\pi(z)}\right| \left|\frac{\overleftarrow{\mathbb{P}}^t(y,z) - \pi(z)}{\pi(z)}\right| \\ &\leq \|\mathbb{P}^s(x,\cdot) - \pi\|_{p,\pi} \cdot \|\overleftarrow{\mathbb{P}}^t(y,\cdot) - \pi\|_{q,\pi} \end{aligned}$$ □

2.10 Discussion

Let $\mathbb{P}$ be an ergodic Markov Chain with uniform stationary distribution π. We can lower bound the mixing time in terms of the eigenvalues of $\mathbb{P}$ and upper bound the

mixing time in terms of the singular values of $\mathbb{P}$. While these two rates are the same for reversible chains, they can be very different for non-reversible chains. One can see that asymptotically the correct mixing rate is determined by the eigenvalue and not by the singular value. This can be seen, at least for chains with enough symmetry, from Proposition 2.43 together with the fact that

$$\lim_{t\to\infty} \left(\sigma_\ell(\mathbb{P}^t)\right)^{1/t} = |\rho_\ell| \tag{2.77}$$

where ρ_ℓ are the (possibly complex) eigenvalues of $\mathbb{P}$ arranged in decreasing absolute value. In the general case this can be established by looking at the Jordan decomposition of $\mathbb{P}$.

In § 2.9 we saw that of the various mixing measures the variation mixing time, entropy mixing time and L^2-mixing time are probably the important ones, as the remaining can be bound in terms of these. While there can be a $O(\log(1/\pi_*))$-gap between the variation and L^2-mixing time, we showed that the gap between variation and entropy mixing time cannot be more than $O(\log\log(1/\pi_*))$.

Chapter 3

Robust Mixing Time

We now define the notion of Robust mixing time of a Markov Chain and develop the basic theory. Loosely speaking, the Robust mixing time of a Markov Chain $\mathbb{P}$ is the mixing time of $\mathbb{P}$ in the presence of an adversary who is allowed to modify the state of the chain in a reasonable manner.

3.1 Basic Definitions

Definition 3.1. Let $\mathbb{P}$ be the transition matrix of an irreducible Markov Chain with stationary distribution π. A stochastic matrix $\mathbb{A}$ (not necessarily irreducible) is said to be *compatible* with $\mathbb{P}$ if $\pi\mathbb{A} = \pi$.

The notion of compatibility depends only on the stationary distribution of $\mathbb{P}$.

Definition 3.2. An *adversarially modified Markov Chain* (AMMC) $\mathcal{P}$ is a pair $(\mathbb{P}, \{\mathbb{A}_t\}_{t>0})$, where $\mathbb{P}$ is the transition matrix of an irreducible Markov Chain and $\mathbb{A}_t$ is a sequence of stochastic matrices compatible with $\mathbb{P}$. Given an AMMC and an initial distribution μ_0, the AMMC process evolves as follows:

- At time $t = 0$, pick $\mathbf{X}_0 \in \mathcal{X}$ according to μ_0,
- Given $\mathbf{X}_t$, pick $\mathbf{Y}_t$ according to the distribution $\mathbb{P}(\mathbf{X}_t, \cdot)$,
- Given $\mathbf{Y}_t$, pick $\mathbf{X}_{t+1}$ according to the distribution $\mathbb{A}_t(\mathbf{Y}_t, \cdot)$

An application of $\mathbb{P}$ followed by $\mathbb{A}_t$ is called a *round*. Let μ_t and ν_t denote the distribution of $\mathbf{X}_t$ and $\mathbf{Y}_t$ respectively. Then μ_t is the distribution after t-rounds. Also we have

$$\nu_t = \mu_t \mathbb{P} \quad \text{and} \quad \mu_{t+1} = \nu_t \mathbb{A}_t \tag{3.1}$$

Definition 3.3. Let $\mathcal{P}$ be an AMMC and $1 \leq p \leq \infty$. The mixing time and L_p-mixing time of $\mathcal{P}$ are defined by the equations

$$\mathcal{T}(\mathcal{P}, \epsilon) = \max_{\mu_0} \min_t \{\|\mu_t - \pi\|_{\mathrm{TV}} \leq \epsilon\} \qquad \mathcal{T}_p(\mathcal{P}, \epsilon) = \max_{\mu_0} \min_t \{\|\mu_t - \pi\|_{p,\pi} \leq \epsilon\} \tag{3.2}$$

respectively. When ϵ is not specified, we take it to be $1/4$ for $\mathcal{T}$ and $1/2$ for $\mathcal{T}_p$.

Definition 3.4. Let $\mathbb{P}$ be an irreducible Markov Chain. An *adversarially modified version* of $\mathbb{P}$ is an AMMC $(\mathbb{P}, \{\mathbb{A}_t\}_{t>0})$.

Definition 3.5. Let $\mathbb{P}$ be an ergodic Markov Chain and $1 \leq p \leq \infty$. The *robust mixing time* and *robust L_p-mixing time* of $\mathbb{P}$ are defined by the equations

$$\mathcal{R}(\mathbb{P}, \epsilon) = \sup_{\mathcal{P}} \mathcal{T}(\mathcal{P}, \epsilon) \qquad \text{and} \qquad \mathcal{R}_p(\mathbb{P}, \epsilon) = \sup_{\mathcal{P}} \mathcal{T}_p(\mathcal{P}, \epsilon) \tag{3.3}$$

respectively, where the suprema are taken over adversarially modified versions $\mathcal{P}$ of $\mathbb{P}$. When $\mathbb{P}$ is clear from context, we drop it and when ϵ is not specified we take it to be $1/4$ for $\mathcal{R}$ and $1/2$ for $\mathcal{R}_p$.

$$\begin{pmatrix} 1-\upsilon & \upsilon & 0 \\ 1 & 0 & 0 \\ 0 & 0 & I \end{pmatrix} \tag{3.4}$$

The states are indexed starting with $y' = \arg\max_x \pi(x)$ and $y = \mathbf{Y}_t$. This adversarial choice $\mathbb{A}$ ensures $\mathbf{X}_t = y'$ if $\mathbf{Y}_t = y$. Here $\upsilon = \pi(y)/\pi(y')$

Figure 3.1: An adaptive adversary is unreasonably powerful

When we need to distinguish between the standard notion of mixing time and robust mixing time, we refer to the standard notion as "standard mixing time."

One can think of the standard mixing time of a Markov Chain as the number of (contiguous) applications of $\mathbb{P}$ required to get close to the stationary distribution starting from the worst initial distribution. In the same vein, the robust mixing time is the number of *not necessarily contiguous* applications of $\mathbb{P}$ required to get close to stationarity under reasonable assumptions on the intervening steps.

3.2 Oblivious v.s. Adaptive Adversary

Note that our adversary is *oblivious*, since the adversary is required to specify the sequence $\{\mathbb{A}_t\}_{t>0}$ in advance. An *adaptive adversary* on the other hand would be allowed to specify $\mathbb{A}_t$ after knowing the value of $\mathbf{Y}_t$. Such an adversary would be unreasonably powerful.

For e. g., let $y' = \arg\max_x \pi(x)$. If $\mathbf{Y}_t = y'$ the adversary does not do any anything. Otherwise the adversary applies the stochastic matrix given in Figure 3.2. It is easily checked that $\mathbb{A}$ is compatible with $\mathbb{P}$ and sends $\mathbf{Y}_t$ to y' with probability 1. Thus an adaptive adversary can ensure that $\mathbf{X}_t = y_0$ always. For this reason we do not consider an adaptive adversary.

3.3 Restricted Adversaries

Let $\mathcal{D}$ denote the set of stochastic matrices compatible with $\mathbb{P}$. Elements of $\mathcal{D}$ satisfy the following set of linear constraints:

- (Non-negativity) $(\forall x, y \in \mathcal{X})\,(\mathbb{A}(x,y) \geq 0)$
- (Stochasticity) $(\forall x \in \mathcal{X})\,(\sum_y \mathbb{A}(x,y) = 1)$
- (Compatibility) $(\forall y \in \mathcal{X})\,(\sum_x \pi(x)\mathbb{A}(x,y) = \pi(y))$

Thus $\mathcal{D}$ is a polytope and hence the convex hull of its vertices. We now define the robust mixing against a restricted set of adversaries.

Definition 3.6. Let $\mathbb{P}$ be an irreducible Markov Chain. A set $\mathcal{S}$ of stochastic matrices is said to be *a set of valid strategies against* $\mathbb{P}$ if it satisfies the following:

- $I \in \mathcal{S}$,
- $\mathbb{A} \in \mathcal{S} \implies \mathbb{A}$ is compatible with $\mathbb{P}$,
- $\mathcal{S}$ is closed under products and convex combinations.

$\mathcal{S}$ is said to be *symmetric* if $\mathbb{A} \in \mathcal{S} \implies \overleftarrow{\mathbb{A}} = \Pi^{-1}\mathbb{A}^T\Pi \in \mathcal{S}$ and $I \in \mathcal{S}$. Here Π is a diagonal matrix with $\Pi(x,x) = \pi(x)$.

Definition 3.7. Let $\mathbb{P}$ be an irreducible Markov Chain, $1 \leq p \leq \infty$ and $\mathcal{S}$ a valid set of strategies against $\mathbb{P}$. The *$\mathcal{S}$-robust mixing time* and *$\mathcal{S}$-robust L_p-mixing time* are defined by the equations

$$\mathcal{R}^{\mathcal{S}}(\mathbb{P},\epsilon) = \sup_{\mathcal{P}} \mathcal{T}(\mathcal{P},\epsilon) \quad \text{and} \quad \mathcal{R}_p^{\mathcal{S}}(\mathbb{P},\epsilon) = \sup_{\mathcal{P}} \mathcal{T}_p(\mathcal{P},\epsilon) \tag{3.5}$$

where $\mathcal{P} = (\mathbb{P}, \{\mathbb{A}_t\}_{t>0})$ ranges over adversarially modified versions of $\mathbb{P}$ where $\mathbb{A}_t \in \mathcal{S}$ for all t.

Setting $\mathcal{S} = \{I\}$ we recover the standard mixing time and setting $\mathcal{S} = \mathcal{D}$ gives the robust mixing time. For Markov Chains with symmetries there are other natural choices for $\mathcal{S}$ which will prove to be useful.

3.4 Finiteness and Sub-multiplicativity

In this section we classify chains $\mathbb{P}$ with finite robust mixing time and prove sub-multiplicativity of robust mixing time under various distance measures.

Theorem 3.8. *Let $\mathbb{P}$ be an ergodic Markov Chain and $\epsilon, \delta > 0$. Let $\mathcal{S}$ be a valid set of strategies against $\mathbb{P}$. Then for $1 \leq p \leq \infty$,*

$$\mathcal{R}_p^{\mathcal{S}}(\epsilon\delta) \leq \mathcal{R}_p^{\mathcal{S}}(\epsilon) + \mathcal{R}_1^{\mathcal{S}}(\delta) \leq \mathcal{R}_p^{\mathcal{S}}(\epsilon) + \mathcal{R}_p^{\mathcal{S}}(\delta) \tag{3.6}$$

Specifically, we also have $\mathcal{R}^{\mathcal{S}}(\epsilon\delta/2) \leq \mathcal{R}^{\mathcal{S}}(\epsilon/2) + \mathcal{R}^{\mathcal{S}}(\delta/2)$.

Proof. Fix $1 \leq p \leq \infty$ and let $S = \mathcal{R}_p^{\mathcal{S}}(\epsilon)$ and $T = \mathcal{R}_1^{\mathcal{S}}(\delta)$. Let $\{\mathbb{A}_t\}_{t>0}$ be any sequence of $\mathbb{P}$ compatible matrices in $\mathcal{S}$ and put $\mathbb{C} = \mathbb{P}\mathbb{A}_1\mathbb{P}\mathbb{A}_2 \ldots \mathbb{P}\mathbb{A}_T$ and $\mathbb{D} = \mathbb{P}\mathbb{A}_{T+1} \ldots \mathbb{P}\mathbb{A}_{T+S}$. If μ_t denotes the distribution of the state of the Markov Chain after t-rounds, we have $\mu_T = \mu_0\mathbb{C}$ and $\mu_{T+S} = \mu_T\mathbb{D}$. Note that $\mathbb{C}$ and $\mathbb{D}$ fix the stationary distribution of $\mathbb{P}$. By choice of S, T and Lemma 2.25, we have

$$\|\mu_{T+S} - \pi\|_{p,\pi} \leq \epsilon\|\mu_T - \pi\|_1 \leq \epsilon\delta \tag{3.7}$$

Since the $\mathbb{A}_t$ were arbitrary we have $\mathcal{R}_p^{\mathcal{S}}(\epsilon\delta) \leq \mathcal{R}_p^{\mathcal{S}}(\epsilon) + \mathcal{R}_1^{\mathcal{S}}(\delta) \leq \mathcal{R}_p^{\mathcal{S}}(\epsilon) + \mathcal{R}_p^{\mathcal{S}}(\delta)$. Using $2\|\cdot\|_{\mathrm{TV}} = \|\cdot\|_1$ and the result for $p = 1$ gives the second result. ★

Lemma 3.9. *Let $\mathbb{P}$ be an ergodic Markov Chain, $1 \leq p \leq \infty$ and $\epsilon < 1$. Let $\sigma_1 = \sigma_1(\mathbb{S}(\mathbb{P}))$ and $\mathcal{S}$ be a set of valid strategies against $\mathbb{P}$. Then*

$$\mathcal{R}_p^{\mathcal{S}}(\mathbb{P}, \epsilon) \geq \max_{\mathbb{Q}} \mathcal{T}_p(\mathbb{P}\mathbb{Q}, \epsilon) \tag{3.8}$$

where the maximum is taken over $\mathbb{Q} \in \mathcal{S}$. In particular if $\overleftarrow{\mathbb{P}} \in \mathcal{S}$

$$\mathcal{R}_p^{\mathcal{S}}(\mathbb{P}, \epsilon) \geq \max(\mathcal{T}_p(\mathbb{P}, \epsilon), \mathcal{T}_p(\mathbb{P}\overleftarrow{\mathbb{P}}, \epsilon)) \tag{3.9}$$

Proof. For $\mathbb{Q} \in \mathcal{S}$, consider the adversarial strategy of always picking $\mathbb{Q}$. The second claim follows by considering $\mathbb{Q} = I$ and $\mathbb{Q} = \overleftarrow{\mathbb{P}}$. ★

It follows that an ergodic Markov Chain $\mathbb{P}$ does not necessarily have finite robust mixing time. For example, if $\mathbb{P}$ is an ergodic Markov Chain where $\mathbb{P}\overleftarrow{\mathbb{P}}$ is not ergodic, we have $\mathcal{R}(\mathbb{P}) \geq \mathcal{T}(\mathbb{P}\overleftarrow{\mathbb{P}}) = \infty$.

Example 3.10. Let $\mathbb{P}$ denote the **no immediate reversal walk** from Example 2.41. Recall that $\mathcal{T}(\mathbb{P}) < \infty$ as long as the underlying graph is connected and non-bipartite. We now show that $\mathcal{R}(\mathbb{P}) = \infty$ by exhibiting an adversarial strategy.

Let $\mathbb{A}$ be the "reversal matrix", i.e. $\mathbb{A}$ sends the edge (u, v) to (v, u). Thus $\mathbb{A}$ is a permutation matrix and hence fixes the uniform stationary distribution of $\mathbb{P}$. Fix a vertex v of X and let μ denote the uniform distribution on edges *into* v. Then $\mu\mathbb{P}$ is the uniform distribution on the edges *out of* v. Now by choice of $\mathbb{A}$, we see that $\mu\mathbb{P}\mathbb{A}$ is once again the uniform distribution on all edges *into* v. Hence $\mathcal{R}(\mathbb{P}) = \infty$.

Example 3.11. "Bottom k to top shuffles": Consider a deck of n cards and fix $1 \leq k \leq n$. In each step, we pick a random card from the bottom k cards and move it to the top. When $k = n$, this becomes the random-to-top shuffle. When $k < n$, the

robust mixing time of this chain is infinite. Instead of calculating the singular values of this chain, it will be easier to explicitly give an adversarial strategy. Note that if $k < n$, then the top card always moves to the second position. Thus if the adversary always switches the second card with the top card, it follows that this adversarial modification of the chain will never mix.

We show below that the ergodicity of $\mathbb{P}\overleftarrow{\mathbb{P}}$ is enough to guarantee finiteness of robust mixing time.

Theorem 3.12. *Let $\mathbb{P}$ be an irreducible Markov Chain and $\sigma_1 = \sigma_1(\mathbb{S}(\mathbb{P}))$. Let μ_0 be any initial distribution of any adversarially modified version $\mathcal{P}$ of $\mathbb{P}$. Then*

$$\|\mu_t - \pi\|_{2,\pi} \leq \sigma_1^t \|\mu_0 - \pi\|_{2,\pi} \leq \sigma_1^t \sqrt{\frac{1}{\pi_*}} \tag{3.10}$$

Proof. Let $\mathcal{P} = (\mathbb{P}, \{\mathbb{A}_t\}_{t>0})$. From Proposition 2.38 it follows that each application of $\mathbb{P}$ reduces the L^2 distance to π by a factor σ_1. Thus we have $\|\nu_t - \pi\|_{2,\pi} \leq \sigma_1 \|\mu_t - \pi\|_{2,\pi}$. Another application of Proposition 2.38 for $\mathbb{A}_t$ shows that

$$\|\mu_{t+1} - \pi\|_{2,\pi} = \|\nu_t \mathbb{A}_t - \pi\|_{2,\pi} = \|(\nu_t - \pi)\mathbb{A}_t\|_{2,\pi} \leq \sigma_1(\mathbb{S}(\mathbb{A}_t))\|\nu_t - \pi\|_{2,\pi} \leq \|\nu_t - \pi\|_{2,\pi} \tag{3.11}$$

since $\mathbb{A}_t$ is stochastic implies $\sigma_1(\mathbb{S}(\mathbb{A}_t)) \leq 1$. Hence we have $\|\mu_{t+1} - \pi\|_{2,\pi} \leq \sigma_1 \|\mu_t - \pi\|_{2,\pi}$. Since $\|\cdot\|_{2,\pi}$ is convex it follows that the maximum value for $\|\mu_0 - \pi\|_{2,\pi}$ is attained when μ_0 is concentrated on a single state. ★

Theorem 3.13. *Let $\mathbb{P}$ be a Markov Chain and $\epsilon > 0$. Then*

$$\mathcal{T}_{\text{rel}}(\mathbb{P}\overleftarrow{\mathbb{P}}) \log(1/2\epsilon) \leq \mathcal{T}(\mathbb{P}\overleftarrow{\mathbb{P}}, \epsilon) \leq \mathcal{R}(\mathbb{P}, \epsilon) \leq \mathcal{R}_2(\mathbb{P}, 2\epsilon) \leq 2\mathcal{T}_{\text{rel}}(\mathbb{P}\overleftarrow{\mathbb{P}}) \left(\log(1/2\epsilon) + \frac{\log(1/\pi_*)}{2} \right) \tag{3.12}$$

In particular $\mathcal{R}(\mathbb{P})$ *is finite iff* $\mathbb{P}\overleftarrow{\mathbb{P}}$ *is irreducible.*

Proof. Lemma 3.9 and Theorem 2.40 imply

$$\mathcal{T}_{\text{rel}}(\mathbb{P}\overleftarrow{\mathbb{P}})\log(1/2\epsilon) \leq \mathcal{T}(\mathbb{P}\overleftarrow{\mathbb{P}},\epsilon) \leq \mathcal{R}(\mathbb{P},\epsilon) \tag{3.13}$$

On the other hand, for $\sigma_1 = \sigma_1(\mathbb{S}(\mathbb{P}))$

$$\begin{aligned}\sigma_1^2 = \lambda_1(\mathbb{S}(\mathbb{P})\mathbb{S}(\mathbb{P})^T) &= \lambda_1(\sqrt{\Pi}\mathbb{P}\sqrt{\Pi}^{-1}\sqrt{\Pi}^{-1}\mathbb{P}^T\sqrt{\Pi}) \\ &= \lambda_1(\sqrt{\Pi}\mathbb{P}\Pi^{-1}\mathbb{P}^T\sqrt{\Pi}) = \lambda_1(\mathbb{P}\overleftarrow{\mathbb{P}})\end{aligned} \tag{3.14}$$

since conjugation by an invertible matrix does not change the eigenvalues. Now Theorem 3.12 implies that for any initial distribution μ_0 and any adversarial modification of $\mathbb{P}$,

$$\|\mu_t - \pi\|_{2,\pi} \leq \sigma_1^t\sqrt{\frac{1}{\pi_*}} \tag{3.15}$$

for $t > 0$. Since $s = \mathcal{T}_{\text{rel}}(\mathbb{P}\overleftarrow{\mathbb{P}})$ satisfies $\lambda_1(\mathbb{P}\overleftarrow{\mathbb{P}})^s \leq 1/\text{e}$ and $\lambda_1(\mathbb{P}\overleftarrow{\mathbb{P}}) = \sigma_1^2$ we have

$$\|\mu_{2s\alpha} - \pi\|_{2,\pi} \leq \sigma_1^{2s\alpha}\sqrt{1/\pi_*} \leq \exp(-\alpha + \log(1/\pi_*)/2) \tag{3.16}$$

Setting $\alpha = \log(1/\pi_*)/2 + \log(1/2\epsilon)$ implies $\|\mu_{2s\alpha} - \pi\|_{2,\pi} \leq 2\epsilon$. ★

In particular for a reversible chain we have the following

Proposition 3.14. *Let* $\mathbb{P}$ *be a reversible Markov Chain with stationary distribution* π*. Then*

$$\begin{aligned}\mathcal{T}_{\text{rel}}(\mathbb{P})\log(1/2\epsilon) \leq \mathcal{T}(\mathbb{P},\epsilon) \leq \begin{matrix}\mathcal{R}(\mathbb{P},\epsilon) \\ \mathcal{T}_2(\mathbb{P},2\epsilon)\end{matrix} &\leq \mathcal{R}_2(\mathbb{P},2\epsilon) \\ &\leq (\mathcal{T}_{\text{rel}}(\mathbb{P})+1)\left(\log(1/2\epsilon) + \frac{\log(1/\pi_*)}{2}\right)\end{aligned} \tag{3.17}$$

Proof. Follows from Theorem 3.12, Theorem 2.40, $\mathcal{R}(\mathbb{P}, \epsilon) \geq \mathcal{T}(\mathbb{P}, \epsilon)$ (adversary applies I always) and the fact that $2\mathcal{T}_{\text{rel}}(\mathbb{P}^2) \leq \mathcal{T}_{\text{rel}}(\mathbb{P}) + 1$ (the $+1$ is to handle the case when $\mathcal{T}_{\text{rel}}(\mathbb{P})$ is odd). ★

3.5 Convexity

Unlike standard mixing time, robust mixing time enjoys the following convexity property: If $\mathbb{P}$ and $\mathbb{Q}$ are two Markov Chains with the same stationary distribution π, the robust mixing time of $(\mathbb{P} + \mathbb{Q})/2$ can be bounded in terms of the smaller of robust mixing times of $\mathbb{P}$ and $\mathbb{Q}$.

Lemma 3.15. *Let $\mathbb{P}$ be any irreducible Markov Chain and $\mathbb{S}$ a valid set of strategies against $\mathbb{P}$. If $\mathbb{Q} \in \mathbb{S}$, then $\mathcal{R}'(\mathbb{PQ}) \leq \mathcal{R}'(\mathbb{P})$ for $\mathcal{R}' \in \{\mathcal{R}^{\mathcal{S}}, \mathcal{R}_2^{\mathcal{S}}\}$.*

Proof. Let $\mathcal{P} = (\mathbb{PQ}, \{\mathbb{A}_t\}_{t>0})$ be any adversarially modified version of $\mathbb{PQ}$ where $\mathbb{A}_t \in \mathbb{S}$. Then $\mathcal{P}' = (\mathbb{P}, \{\mathbb{QA}_t\}_{t>0})$ is an adversarially modified version of $\mathbb{P}$ where $\mathbb{QA}_t \in \mathbb{S}$ since $\mathbb{Q} \in \mathbb{S}$ and $\mathbb{S}$ is closed under products. Since the mixing times of $\mathcal{P}$ and $\mathcal{P}'$ are equal, we have the result. □

We now show that the robust mixing time of a convex combination of Markov Chains can be bounded in terms of that of the participating chains.

Let $\mathbb{P}$ and $\mathbb{Q}$ be two irreducible Markov Chains with same stationary distribution π. Suppose $\mathcal{S}$ is a valid set of strategies against $\mathbb{P}$ as well as against $\mathbb{Q}$. Also assume $\mathbb{P}, \mathbb{Q} \in \mathcal{S}$. Fix $0 < a = 1 - b < 1$ and consider the chain $\mathbb{P}' = a\mathbb{P} + b\mathbb{Q}$. Let $\mathcal{P} = (\mathbb{P}', \{\mathbb{A}_t\}_{t>0})$ be any adversarial modification of $\mathbb{P}'$. Fix $S > 0$ and $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_S)$ where $\epsilon_i \in \{0, 1\}$. Define the following quantities:

- $\mathbb{P}^{(1)} = \mathbb{P}$ and $\mathbb{P}^{(0)} = \mathbb{Q}$

- $\xi(\vec{\epsilon}) = \prod_{i=1}^{S} \mathbb{P}^{(\epsilon_i)}\mathbb{A}_i$
- $H(\vec{\epsilon}) = \sum_{i=1}^{S} \epsilon_i$
- $w(\vec{\epsilon}) = \prod_{i=1}^{S} a^{\epsilon_i} b^{1-\epsilon_i} = a^{H(\vec{\epsilon})} b^{S-H(\vec{\epsilon})}$

If μ_0 is any initial distribution, and μ_S is the distribution after S-rounds, we have

$$\mu_S - \pi = \sum_{\vec{\epsilon}} w(\vec{\epsilon}) \left(\mu_0 \xi(\vec{\epsilon}) - \pi\right) \tag{3.18}$$

where the sum ranges over all 2^S choices for $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_S)$.

Lemma 3.16. *Let $\mathbb{P}$ and $\mathbb{Q}$ be ergodic Markov Chains with the same stationary distribution π. Let $\mathcal{S}$ be a valid set of strategies against both $\mathbb{P}$ and $\mathbb{Q}$ and assume that $\mathbb{P}, \mathbb{Q} \in \mathcal{S}$. Let $0 < a = 1 - b < 1$. Then for $\mathcal{R}' \in \{\mathcal{R}^{\mathcal{S}}, \mathcal{R}_2^{\mathcal{S}}\}$, $\mathcal{R}'(a\mathbb{P} + b\mathbb{Q}) \leq \mathcal{R}'(\mathbb{P}) + \mathcal{R}'(\mathbb{Q}) - 1$.*

Proof. Choose $S = \mathcal{R}^{\mathcal{S}}(\mathbb{P}) + \mathcal{R}^{\mathcal{S}}(\mathbb{Q}) - 1$. Then we have

$$\|\mu_S - \pi\|_{\mathrm{TV}} \leq \sum_{\vec{\epsilon}} w(\vec{\epsilon}) \|\mu_0 \xi(\vec{\epsilon}) - \pi\|_{\mathrm{TV}} \tag{3.19}$$

Now for each $\vec{\epsilon}$, $\xi(\vec{\epsilon})$ either contains $\geq \mathcal{R}^{\mathcal{S}}(\mathbb{P})$ occurrences of $\mathbb{P}$ or contains $\geq \mathcal{R}^{\mathcal{S}}(\mathbb{Q})$ occurrences of $\mathbb{Q}$. The remaining matrices can be considered as an adversarial choice. Hence we have $\|\mu_0 \xi(\vec{\epsilon}) - \pi\|_{\mathrm{TV}} \leq 1/4$ for all $\vec{\epsilon}$. Since $\sum_{\vec{\epsilon}} w(\vec{\epsilon}) = 1$, $\|\mu_S - \pi\|_{\mathrm{TV}} \leq 1/4$.

Similarly, taking $S = \mathcal{R}_2^{\mathcal{S}}(\mathbb{P}) + \mathcal{R}_2^{\mathcal{S}}(\mathbb{Q}) - 1$, and looking at the $\|\mu_S - \pi\|_{2,\pi}$ we get $\mathcal{R}_2^{\mathcal{S}}(a\mathbb{P} + b\mathbb{Q}) \leq \mathcal{R}_2^{\mathcal{S}}(\mathbb{P}) + \mathcal{R}_2^{\mathcal{S}}(\mathbb{Q}) - 1$. □

Now we consider the case when $\mathbb{P}$ has finite robust mixing time and $\mathbb{Q}$ may not. We start with a concentration inequality.

Proposition 3.17. *Let $S = CT/p$ for $C > 1$ and $0 < p < 1$. Let $\mathbf{Z}_1, \ldots, \mathbf{Z}_S$ be independent Bernoulli random variables with $\Pr\{\mathbf{Z}_i = 1\} = p$. Let $\mathbf{Z} = \sum_i \mathbf{Z}_i$. Then we have*

$$\Pr\{\mathbf{Z} < T\} \leq \exp\Big(-T\left((C-1) - \log C\right)\Big) \tag{3.20}$$

Proof. We use Hoeffding's inequality [28, Theorem 1], for $S - \mathbf{Z}$ to conclude

$$\Pr\{\mathbf{Z} < T\} \leq \left\{ \left(\frac{q}{q+a}\right)^{q+a} \left(\frac{p}{p-a}\right)^{p-a} \right\}^S \tag{3.21}$$

where $q = 1 - p, a = p - p/C$. After algebraic simplifications we get,

$$\Pr\{\mathbf{Z} < T\} \leq \left\{ \left(\frac{q}{1-\frac{p}{C}}\right)^{\frac{C}{p}-1} C \right\}^T \tag{3.22}$$

Taking logarithms, and using $\log(1-x) \leq -x$ for $0 < x < 1$ gives the result. □

Proposition 3.18. *Let $\mathbb{P}$ be an ergodic Markov Chain and $\mathcal{S}$ a valid set of strategies against $\mathbb{P}$. Fix $\mathbb{Q} \in \mathcal{S}$. Let $0 < a = 1 - b < 1$. Let $S = (1+\delta)\mathcal{R}^{\mathcal{S}}(\mathbb{P}, \gamma)/a$, where $\gamma > 0$ and $\delta > 0$ are arbitrary. Then*

$$\|\mu_S - \pi\|_{\mathrm{TV}} \leq \gamma + \exp\left(-\mathcal{R}^{\mathcal{S}}(\mathbb{P}, \gamma) \cdot (\delta - \log(1+\delta))\right) \tag{3.23}$$

Proof. Let $S = (1+\delta)T/a$, where $T = \mathcal{R}^{\mathcal{S}}(\mathbb{P}, \gamma)$. Write

$$\mu_S - \pi = \sum_{\vec{\epsilon}} w(\vec{\epsilon})\left(\mu_0 \xi(\vec{\epsilon}) - \pi\right) \tag{3.24}$$

Put $\mathcal{D} = \{\vec{\epsilon} : H(\vec{\epsilon}) \geq T\}$, i.e. all choices of $\vec{\epsilon}$ which resulted in $\mathbb{P}$ being used at least T times. For $\vec{\epsilon} \in \mathcal{D}$, we have $\|\mu_0\xi(\vec{\epsilon}) - \pi\|_{\mathrm{TV}} \leq \gamma$. For $\vec{\epsilon} \notin \mathcal{D}$, $\|\mu_0\xi(\vec{\epsilon}) - \pi\|_{\mathrm{TV}} \leq 1$.

From Proposition 3.17,

$$\sum_{\vec{\epsilon}\notin\mathcal{D}} w(\vec{\epsilon}) = \Pr\{H(\vec{\epsilon}) < m/(1+\delta)\} \leq \exp\left(-T(\delta - \log(1+\delta))\right) \tag{3.25}$$

Hence

$$\begin{aligned}
\|\mu_S - \pi\|_{\mathrm{TV}} &\leq \sum_{\vec{\epsilon}\in\mathcal{D}} w(\vec{\epsilon})\|\mu_0\xi(\vec{\epsilon}) - \pi\|_{\mathrm{TV}} + \sum_{\vec{\epsilon}\notin\mathcal{D}} w(\vec{\epsilon})\|\mu_0\xi(\vec{\epsilon}) - \pi\|_{\mathrm{TV}} \\
&\leq \gamma + \sum_{\vec{\epsilon}\notin\mathcal{D}} w(\vec{\epsilon}) * 1 \\
&\leq \gamma + \exp\left(-T(\delta - \log(1+\delta))\right) \qquad \bigstar
\end{aligned}$$

Corollary 3.19. *Let $\mathbb{P}$ be an ergodic Markov Chain and $\mathbb{Q}$ be compatible with $\mathbb{P}$. Let $\mathcal{S}$ be a valid set of strategies against $\mathbb{P}$. Assume $\mathbb{Q} \in \mathcal{S}$. Let $0 < a = 1 - b < 1$. Then $\mathcal{R}^{\mathcal{S}}(a\mathbb{P} + b\mathbb{Q}) \leq 2(1+\delta)\mathcal{R}^{\mathcal{S}}(\mathbb{P})/a$, as long as $2\mathcal{R}^{\mathcal{S}}(\mathbb{P})(\delta - \log(1+\delta)) \geq \log 8$. If $\mathcal{R}^{\mathcal{S}}(\mathbb{P}) \geq 11$, then δ may be taken to be $1/2$.*

Proof. Let $T = \mathcal{R}^{\mathcal{S}}(\mathbb{P})$ and $S = 2T(1+\delta)/a$. By sub-multiplicativity, we have $\mathcal{R}^{\mathcal{S}}(\mathbb{P}, \gamma) \leq 2T$ for $\gamma = 1/8$. Proposition 3.18 now gives

$$\|\mu_S - \pi\|_{\mathrm{TV}} \leq 1/8 + \exp\left(-2T(\delta - \log(1+\delta))\right) \tag{3.26}$$

If $2T(\delta - \log(1+\delta)) \geq \log 8$, we have $\|\mu_S - \pi\|_{\mathrm{TV}} \leq 1/8 + 1/8 = 1/4$ as required. □

Similarly for the $\mathcal{S}$-robust L^2-mixing time we get,

Proposition 3.20. *Let $\mathbb{P}$ be an ergodic Markov Chain and $\mathcal{S}$ a valid set of strategies against $\mathbb{P}$. Fix $\mathbb{Q} \in \mathcal{S}$. Let $0 < a = 1 - b < 1$. Let $S = (1+\delta)\mathcal{R}_2^{\mathcal{S}}(\mathbb{P}, \gamma)/a$, where*

$\gamma > 0$ and $\delta > 0$ are arbitrary. Then

$$\|\mu_S - \pi\|_{2,\pi} \leq \gamma + \exp\left(-\mathcal{R}_2^{\mathcal{S}}(\mathbb{P}, \gamma) \cdot (\delta - \log(1+\delta))\right) \cdot \sqrt{\frac{1}{\pi_*}} \tag{3.27}$$

Proof. This proof is similar to that of Proposition 3.18. Put $T = \mathcal{R}_2^{\mathcal{S}}(\mathbb{P}, \gamma)$ and $S = T(1+\delta)/a$. Put $\mathcal{D} = \{\vec{\epsilon} : H(\vec{\epsilon}) \geq T\}$. If $\vec{\epsilon} \in \mathcal{D}$, $\|\mu_0 \xi(\vec{\epsilon}) - \pi\|_{2,\pi} \leq \gamma$. For $\vec{\epsilon} \notin \mathcal{D}$, $\|\mu_0 \xi(\vec{\epsilon}) - \pi\|_{2,\pi} \leq \sqrt{1/\pi_*}$. Going along the same lines as Proposition 3.18, we have the result. ★

Corollary 3.21. *Let $\mathbb{P}$ be an ergodic Markov Chain and $\mathcal{S}$ a valid set of strategies against $\mathbb{P}$. Let $\mathbb{Q}$ be compatible with $\mathbb{P}$ and $\mathbb{Q} \in \mathcal{S}$. Let $0 < a = 1 - b < 1$. Assume that $\mathcal{R}_2^{\mathcal{S}}(\mathbb{P}) \geq \log(1/\pi_*)/2$ and $\pi_* \leq 1/16$. Then $\mathcal{R}_2^{\mathcal{S}}(a\mathbb{P} + b\mathbb{Q}) \leq 2(1+\delta)\mathcal{R}_2^{\mathcal{S}}(\mathbb{P})/a$, as long as $\mathcal{R}_2^{\mathcal{S}}(\mathbb{P})(\delta - \log(1+\delta)) \geq \log(1/\pi_*)/2$. In particular δ may be taken to be $5/2$.*

Proof. Let $T = \mathcal{R}_2^{\mathcal{S}}(\mathbb{P})$ and $S = 2T(1+\delta)/a$. By sub-multiplicativity, we have $\mathcal{R}^{\mathcal{S}}(\mathbb{P}, \gamma) \leq 2T$ for $\gamma = 1/4$. Proposition 3.20 now gives

$$\|\mu_S - \pi\|_{2,\pi} \leq 1/4 + exp\left(-2T(\delta - \log(1+\delta))\right) \cdot \sqrt{\frac{1}{\pi_*}} \tag{3.28}$$

Now put $T = \alpha \log(1/\pi_*)/2$ for $\alpha > 1$. Then we have

$$\|\mu_S - \pi\|_{2,\pi} \leq 1/4 + \pi_*^{(2\alpha(\delta - \log(1+\delta)) - 1)} \tag{3.29}$$

The second term is bounded by $1/4$ if $\delta - \log(1+\delta) \geq 1/\alpha$. ★

3.6 Upper Bounds on Robust Mixing Time

In this section we show that many methods which prove upper bounds on mixing times under various measures also give upper bounds on Robust mixing times. We start with a contraction result.

Proposition 3.22. *Let $\mathbb{A}$ be a stochastic matrix and $\pi\mathbb{A} = \pi$ for some distribution π. For any real valued function f on $\mathcal{X}$, and $1 \leq p \leq \infty$, we have*

$$\|\mathbb{A}f\|_{p,\pi} \leq \|f\|_{p,\pi} \tag{3.30}$$

This is an adaptation of [60, Theorem 3.1]

Proof. Suppose $1 \leq p < \infty$ and choose q such that $1/p+1/q = 1$. When $p = 1, q = \infty$, we consider $a^{1/q} = 1$ for any real a.

Looking at f as a column vector, we have $\|f\|_{p,\pi} = \|\Pi^{1/p} f\|_p$, where Π is the diagonal matrix with $\Pi(x, x) = \pi(x)$. Then

$$\|\mathbb{A}f\|_{p,\pi} = \|\Pi^{1/p}\mathbb{A}f\|_p = \|(\Pi^{1/p}\mathbb{A}\Pi^{-1/p})(\Pi^{1/p}f)\|_p \tag{3.31}$$

Hence it is enough to show that $\|\mathbb{B}f\|_p \leq \|f\|_p$ for any real valued function f, where $\mathbb{B} = \Pi^{1/p}\mathbb{A}\Pi^{-1/p}$. Let $g = \mathbb{B}f$. Then for any fixed $x \in \mathcal{X}$,

$$g(x) = \sum_y \left(\frac{\pi(x)}{\pi(y)}\right)^{1/p} \mathbb{A}(x, y)f(y) = \sum_y a_y b_y \tag{3.32}$$

where

$$a_y = f(y)\left(\frac{\pi(x)}{\pi(y)}\right)^{1/p} \mathbb{A}(x, y)^{1/p} \quad \text{and } b_y = \mathbb{A}(x, y)^{1/q} \tag{3.33}$$

By Hölder's inequality we have

$$\begin{aligned} |g(x)| &\leq \left(\sum_y |a_y|^p\right)^{1/p} \left(\sum_y b_y^q\right)^{1/q} \\ &= \left(\sum_y |f(y)|^p \frac{\pi(x)}{\pi(y)} \mathbb{A}(x,y)\right) \left(\sum_y \mathbb{A}(x,y)\right)^{1/q} \\ &= \sum_y |f(y)|^p \frac{\pi(x)}{\pi(y)} \mathbb{A}(x,y) \end{aligned} \tag{3.34}$$

Hence we have

$$\|g\|_p^p = \sum_x |g(x)|^p \leq \sum_y \left(\frac{|f(y)|^p}{\pi(y)} \sum_x \pi(x)\mathbb{A}(x,y)\right) = \|f\|_p^p \tag{3.35}$$

since $\pi\mathbb{A} = \pi$.

For $p = \infty$, suppose $\|f\|_{\infty,\pi} = A$. Then $|f(y)| \leq A$ for all $y \in \mathcal{X}$.

$$|g(x)| = \left|\sum_x \mathbb{A}(x,y) f(y)\right| \leq A \sum_y \mathbb{A}(x,y) = A \tag{3.36}$$

gives $\|\mathbb{A}f\|_{\infty,\pi} \leq \|f\|_{\infty,\pi}$. □

Note that we did not require that $\mathbb{A}$ is irreducible or aperiodic. This shows that the adversary cannot increase the L_p distance no matter what. Similarly we now show that the adversary cannot increase the separation or the entropy distance.

Proposition 3.23. *Let $\mathbb{A}$ be a stochastic matrix with $\pi\mathbb{A} = \pi$ for a distribution π. For any distribution μ,*

(a) $\mathrm{sep}(\mu\mathbb{A}, \pi) \leq \mathrm{sep}(\mu, \pi)$

(b) $\mathrm{D}(\mu\mathbb{A}||\pi) \leq \mathrm{D}(\mu||\pi)$

Proof. (a) Suppose $\text{sep}(\mu, \pi) = \epsilon$. Then $\mu = (1-\epsilon)\pi + \epsilon\nu$ for some distribution ν. $\mu\mathbb{A} = (1-\epsilon)\pi + \epsilon\nu\mathbb{A}$ and hence $\text{sep}(\mu\mathbb{A}, \pi) \leq \epsilon \,\text{sep}(\nu\mathbb{A}, \pi) \leq \epsilon = \text{sep}(\mu, \pi)$.

(b) This proof is along the lines of [34, Proposition 6]. Let $f \geq 0$ be any real valued function and $g = \mathbb{A}f$. Need to show that $\mathbb{E}_\pi[g \log g] \leq \mathbb{E}_\pi[f \log f]$. From convexity of $\phi(t) = t \log t$ it is easy to see that $\forall t > 0$ and $s > -t$, we have $(t+s)\log(t+s) \geq t \log t + (1 + \log t)s$.

For $x, y \in \mathcal{X}$, put $t = g(x)$ and $t + s = f(y)$ to get

$$f(y) \log f(y) \geq g(x) \log g(x) + \big(1 + \log g(x)\big)\big(f(y) - g(x)\big) \tag{3.37}$$

Multiplying both sides by $\mathbb{A}(x, y)$ and summing over y, we get

$$\sum_y \mathbb{A}(x, y) f(y) \log f(y) \geq g(x) \log g(x) \tag{3.38}$$

since $\sum_y \mathbb{A}(x, y) = 1$ and $\sum_y \mathbb{A}(x, y) f(y) = g(x)$. Multiplying by $\pi(x)$ and summing over x, we get

$$\begin{aligned} \mathbb{E}_\pi[g \log g] &\leq \sum_y \left(f(y) \log f(y) \sum_x \pi(x) \mathbb{A}(x, y) \right) \\ &= \sum_y \pi(y) f(y) \log f(y) = \mathbb{E}_\pi[f \log f] \end{aligned} \tag{3.39}$$

since $\pi\mathbb{A} = \pi$. □

Again the only property of $\mathbb{A}$, we needed was that $\mathbb{A}$ was stochastic and $\pi\mathbb{A} = \pi$. We now show that many methods of deriving upper bounds on standard mixing time actually give bounds on Robust mixing time as well.

Let $\mathbb{P}$ be a Markov Chain with stationary distribution π. Many upper bounds on mixing time can be described via the following outline:

- Let $V(\eta)$ be a "potential" function on distributions η which satisfies the following
  - $V(\eta) \geq 0$ for all distributions η,
  - $V(\eta_t) \to 0 \iff \eta_t \to \pi$, for any sequence of distributions η_t and
  - $V(\eta\mathbb{P}) \leq V(\eta)$
- Let $I(t) = V(\mu\mathbb{P}^t)$ where μ is any initial distribution. Note that $I(t)$ is a non-increasing function,
- Show that there is some non-decreasing function $G : \mathbb{R}^+ \to \mathbb{R}^+$ which satisfies $I(t) - I(t+1) \geq G(I(t))$
- Using the worst value for $I(0)$, solve for T by which $V(T) \leq 1/4$ (or some appropriate constant < 1) to get an upper bound on the mixing time of $\mathbb{P}$ with respect to V.

For e. g., $V(\eta) = \|\eta - \pi\|_{2,\pi}$ gives bounds on L^2 mixing time while $V(\eta) = \mathrm{D}(\eta||\pi)$ gives bounds on the entropy mixing time.

In case of an AMMC $\mathcal{P} = (\mathbb{P}, \{\mathbb{A}_t\})$, we can do the following:

- Let μ_0 be an arbitrary initial distribution.
- V will usually also satisfy $V(\mu\mathbb{A}) \leq V(\mu)$ for all $\mathbb{A}$ compatible with $\mathbb{P}$.
- Let $I(t) = V(\mu_t)$ and $J(t) = V(\nu_t)$ where $\nu_t = \mu_t\mathbb{P}$ and $\mu_{t+1} = \nu_t\mathbb{A}_t$.
- $I(t) - J(t) \geq G(I(t))$ and $I(t+1) \leq J(t)$ imply $I(t) - I(t+1) \geq G(I(t))$ as before.

Hence the mixing time upper bound is also an upper bound on the Robust mixing time.

For example taking $V(\eta) = \|\eta - \pi\|_{2,\pi}$ shows that upper bounds on mixing times of Markov Chains using the spectral gap, conductance, spectral profiling [25], Congestion bounds for L^2-distance [43] carry over to the robust setting. In case of bounds using Log-Sobolev inequalities, the original proof of Diaconis and Saloff-Coste [11] works only in continuous time and does not show a decay after each application of $\mathbb{P}$. [34] proves smooth bounds on entropy decay in discrete time using Log-Sobolev constant of $\mathbb{P}\overleftarrow{\mathbb{P}}$. See also [45] for an alternate proof.

3.6.1 Conductance Approach

As an application of the observation in § 3.6, we derive bounds on the Robust mixing time of a Markov Chain $\mathbb{P}$ in terms of its standard mixing time. Theorem 3.13 shows that there cannot be any unconditional relation between $\mathcal{R}(\mathbb{P})$ and $\mathcal{T}(\mathbb{P})$. The simplest way is to ensure $\mathbb{P}\overleftarrow{\mathbb{P}}$ is irreducible is to add holding probabilities.

Definition 3.24. Let $\mathbb{P}$ be a Markov Chain with stationary distribution π. For $\mathcal{Y}_1, \mathcal{Y}_2 \subseteq \mathcal{X}$, define $Q(\mathcal{Y}_1, \mathcal{Y}_2) = \sum_{x \in \mathcal{Y}_1, y \in \mathcal{Y}_2} \pi(x)\mathbb{P}(x, y)$ and

$$\Phi_{\mathcal{Y}}(\mathbb{P}) = \frac{Q(\mathcal{Y}, \overline{\mathcal{Y}})}{\pi(\mathcal{Y})} \quad \textit{and} \quad \Phi(\mathbb{P}) = \min_{0 < \pi(\mathcal{Y}) < 1} \Phi_{\mathcal{Y}}(\mathbb{P}) \tag{3.40}$$

where $\overline{\mathcal{Y}} = \mathcal{X} \setminus \mathcal{Y}$.

$\Phi_{\mathcal{Y}}(\mathbb{P})$ is the conditional probability of crossing over from $\mathcal{Y}$ to $\overline{\mathcal{Y}}$ in one step at stationarity given that we are currently in $\mathcal{Y}$.

Lemma 3.25. *For $\mathcal{Y} \subseteq \mathcal{X}$, $Q(\mathcal{Y}, \overline{\mathcal{Y}}) = Q(\overline{\mathcal{Y}}, \mathcal{Y})$. In particular the minimum in Definition 3.24 is always attained for $\pi(\mathcal{Y}) \leq 1/2$*

Proof. Let $\mathcal{Y} \subseteq \mathcal{X}$ be arbitrary and $\overline{\mathcal{Y}} = \mathcal{X} \setminus \mathcal{Y}$. Since $\mathbb{P}$ is stochastic we have

$$\pi(\mathcal{Y}) = \sum_{x \in \mathcal{Y}, y \notin \mathcal{Y}} \pi(x)\mathbb{P}(x,y) + \sum_{x \in \mathcal{Y}, y \in \mathcal{Y}} \pi(x)\mathbb{P}(x,y) = Q(\mathcal{Y}, \overline{\mathcal{Y}}) + Q(\mathcal{Y}, \mathcal{Y}) \tag{3.41}$$

Since π is the stationary distribution we also have

$$\pi(\mathcal{Y}) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \pi(x)\mathbb{P}(x,y) = Q(\overline{\mathcal{Y}}, \mathcal{Y}) + Q(\mathcal{Y}, \mathcal{Y}) \tag{3.42}$$

Combining we get the result. □

Conductance was first introduced by Jerrum and Sinclair [31]. Intuitively, if we collapse the chain to two states $\mathcal{Y}$ and $\overline{\mathcal{Y}}$ and start the chain at $\mathcal{Y}$, we move over to $\overline{\mathcal{Y}}$ with probability Φ. So in order to have the right probability of being at $\overline{\mathcal{Y}}$ we need to run the chain for at least $O(1/\Phi)$ steps. This intuition was made precise in [17].

Theorem 3.26 (Dyer et al. [17, Claim 2.3])**.** *Let $\mathbb{P}$ be a (not-necessarily reversible) Markov Chain with stationary distribution π and Φ its conductance. Then*

$$\mathcal{T}(\mathbb{P}, \epsilon) \geq (1/2 - \epsilon)/\Phi(\mathbb{P}) \tag{3.43}$$

Proof. For distributions μ, ν, Proposition 3.22 implies $\|(\mu - \nu)\mathbb{P}\|_{\mathrm{TV}} \leq \|\mu - \nu\|_{\mathrm{TV}}$. Hence for $t > 0$, and initial distribution μ, $\|\mu\mathbb{P}^t - \mu\|_{\mathrm{TV}} \leq \sum_{i=0}^{t-1} \|\mu\mathbb{P}^{i+i} - \mu\mathbb{P}^i\|_{\mathrm{TV}} \leq t\|\mu\mathbb{P} - \mu\|_{\mathrm{TV}}$. Thus

$$\|\mu\mathbb{P}^t - \pi\|_{\mathrm{TV}} \geq \|\mu - \pi\|_{\mathrm{TV}} - \|\mu\mathbb{P}^t - \mu\|_{\mathrm{TV}} \geq \|\mu - \pi\|_{\mathrm{TV}} - t\|\mu\mathbb{P} - \mu\|_{\mathrm{TV}} \tag{3.44}$$

Fix $\mathcal{Y} \subseteq \mathcal{X}$ and let

$$\mu(y) = \begin{cases} \pi(y)/\pi(\mathcal{Y}) & \text{if } y \in \mathcal{Y} \\ 0 & \text{otherwise} \end{cases} \tag{3.45}$$

Now $\|\mu - \pi\|_{\mathrm{TV}} \geq \pi(\overline{\mathcal{Y}})$ and for $y \in \mathcal{Y}$, we have

$$\begin{aligned} \mu(y) - (\mu\mathbb{P})(y) &= \frac{\pi(y)}{\pi(\mathcal{Y})} - \frac{\sum_{x\in\mathcal{Y}} \mu(x)\mathbb{P}(x,y)}{\pi(\mathcal{Y})} \\ &= \frac{\sum_{x\in\mathcal{X}} \pi(x)\mathbb{P}(x,y)}{\pi(\mathcal{Y})} - \frac{\sum_{x\in\mathcal{Y}} \mu(x)\mathbb{P}(x,y)}{\pi(\mathcal{Y})} \\ &= \frac{\sum_{x\notin\mathcal{Y}} \pi(x)\mathbb{P}(x,y)}{\pi(\mathcal{Y})} \geq 0 \end{aligned} \tag{3.46}$$

If $y \notin \mathcal{Y}$, $\mu(y) - (\mu\mathbb{P})(y) \leq 0$. Hence

$$\|\mu - \mu\mathbb{P}\|_{\mathrm{TV}} = \sum_{y\in\mathcal{X}} (\mu(y) - (\mu\mathbb{P})(y))^+ = Q(\overline{\mathcal{Y}}, \mathcal{Y})/\pi(\mathcal{Y}) = \Phi_{\mathcal{Y}}(\mathbb{P}) \tag{3.47}$$

by Lemma 3.25. Thus (3.44) implies

$$\|\mu\mathbb{P}^t - \mu\|_{\mathrm{TV}} \geq \pi(\overline{\mathcal{Y}}) - t\Phi_{\overline{\mathcal{Y}}}(\mathbb{P}) \tag{3.48}$$

Now choose $\mathcal{Y}$ so that $\Phi(\mathbb{P}) = \Phi_{\mathcal{Y}}(\mathbb{P})$ and Lemma 3.25 implies $\pi(\mathcal{Y}) \leq 1/2$. Thus we have $\|\mu\mathbb{P}^t - \pi\|_{\mathrm{TV}} \geq 1/2 - t\Phi(\mathbb{P})$. □

For reversible chains Φ captures $1 - \lambda_1$ up to a quadratic factor. More precisely we have

Theorem 3.27 (Sinclair [57, Lemma 2.4])**.** *Let $\mathbb{P}$ be a reversible Markov Chain with stationary distribution π and Φ its conductance. Then $\Phi \geq 1 - \lambda_1(\mathbb{P}) \geq \Phi^2/2$.*

Putting these together we can bound the Robust mixing time in terms of standard

mixing time for lazy chains.

Theorem 3.28. *Let $\mathbb{P}$ be a lazy Markov Chain with stationary distribution π and conductance $\Phi(\mathbb{P})$. Then $\mathcal{R}(\mathbb{P}) \leq \mathcal{R}_2(\mathbb{P}) = O(\mathcal{T}(\mathbb{P})^2 \log(1/\pi_*))$*

Proof. Let $\mathbb{Q} = (\mathbb{P} + \overleftarrow{\mathbb{P}})/2$. For any $\mathcal{Y} \subseteq \mathcal{X}$, we have $Q_{\mathbb{P}}(\mathcal{Y}, \overline{\mathcal{Y}}) = Q_{\overleftarrow{\mathbb{P}}}(\overline{\mathcal{Y}}, \mathcal{Y})$ and hence $\Phi_{\mathcal{Y}}(\mathbb{P}) = \Phi_{\overline{\mathcal{Y}}}(\overleftarrow{\mathbb{P}})$. This implies $\Phi(\mathbb{P}) = \Phi(\mathbb{Q})$.

Now Theorem 3.27 implies $\lambda_1(\mathbb{Q}) \leq 1 - \Phi(\mathbb{Q})^2/2$. But since $\mathbb{P}$ is lazy, Proposition 2.42 together with $\Phi(\mathbb{P}) = \Phi(\mathbb{Q})$ implies

$$\sigma_1(\mathbb{S}(\mathbb{P}))^2 \leq \lambda_1(\mathbb{Q}) \leq 1 - \Phi(\mathbb{Q})^2/2 = 1 - \Phi(\mathbb{P})^2/2 \tag{3.49}$$

Let $\mathcal{P} = (\mathbb{P}, \{\mathbb{A}_t\})$ be any adversarially modified version of $\mathbb{P}$ and μ_0 be any initial distribution. Theorem 3.12 shows that for $t > 0$,

$$\|\mu_t - \pi\|_{2,\pi} \leq \sigma_1(\mathbb{S}(\mathbb{P}))^t \sqrt{1/\pi_*} \tag{3.50}$$

Theorem 3.26 now implies $\sigma_1(\mathbb{S}(\mathbb{P})) \leq 1 - \mathcal{T}(\mathbb{P})^{-2}/4$ which implies the result. ★

As an immediate corollary we get

Proposition 3.29. *Let $\mathbb{P}$ be a lazy Markov Chain with stationary distribution π. Then*

$$\mathcal{T}((\mathbb{P} + \overleftarrow{\mathbb{P}})/2) \leq O(\mathcal{T}(\mathbb{P})^2 \log(1/\pi_*)) \tag{3.51}$$

Proof. Let $\mathbb{Q} = (\mathbb{P} + \overleftarrow{\mathbb{P}})/2$. Corollary 3.19 and Theorem 3.28 imply

$$\mathcal{T}(\mathbb{Q}) \leq \mathcal{R}(\mathbb{Q}) = O(\mathcal{R}(\mathbb{P})) = O(\mathcal{T}(\mathbb{P})^2 \log(1/\pi_*)) \tag{3.52}$$

★

Definition 3.30. Let X be an undirected graph. An *Eulerian orientation* of X is a directed graph Y obtained by assigning an orientation to each edge of X such that for every vertex v of Y, $d^+(v) = d^-(v)$, where d^-, d^+ are the in and out-degrees of v in Y respectively.

Let Y be the Eulerian orientation of X. Proposition 3.29 implies that the lazy walk on Y cannot lead to a better than quadratic speed up when compared to the lazy walk on X. Proposition 2.42 together with Corollary 2.39 implies that the lazy walk on Y cannot be slower by more than a constant factor when compared to the lazy walk on X.

The laziness assumption on the speed up can be removed if we by pass Robust mixing and just use that X and Y have the same conductance. On the other hand, [4] shows the diameter of Y can be unbounded in the diameter of X, if we only require that X be regular.

3.6.2 log-Sobolev Approach

In this section, we apply the observation of § 3.6 to get better bounds on Robust L^2 Mixing times of lazy Markov Chains.

Definition 3.31. Let $\mathbb{P}$ be a Markov Chain with stationary distribution π. The Dirichlet form associated with $\mathbb{P}$ is defined as

$$\mathcal{E}_{\mathbb{P}}(f, g) = \langle f, (I - \mathbb{P})g \rangle_{\pi} = \sum_{x,y \in \mathcal{X}} \pi(x) f(x)(g(x) - g(y))\mathbb{P}(x, y) \tag{3.53}$$

where f, g are real valued functions on $\mathcal{X}$.

Lemma 3.32. *Let $\mathbb{P}$ be a Markov Chain and f, g real valued functions on $\mathcal{X}$.*

(a) $2\,\mathcal{E}_{\mathbb{P}}(f,f) = \sum_{x,y}(f(x)-f(y))^2\pi(x)\mathbb{P}(x,y)$

(b) $\mathcal{E}_{\mathbb{P}}(f,f) = \mathcal{E}_{\overleftarrow{\mathbb{P}}}(f,f) = \mathcal{E}_{\frac{\mathbb{P}+\overleftarrow{\mathbb{P}}}{2}}(f,f)$

(c) If $\mathbb{P}$ *is lazy, i. e.,* $\mathbb{P}(x,x) \geq 1/2$ *for all* $x \in \mathcal{X}$, $\mathcal{E}_{\mathbb{P}\overleftarrow{\mathbb{P}}}(f,f) \geq \mathcal{E}_{\mathbb{P}}(f,f)$

(d) If $\mathbb{P}$ *is reversible, then*

$$\mathcal{E}_{\mathbb{P}}(f,g) = \mathcal{E}_{\mathbb{P}}(g,f) = \frac{1}{2}\sum_{x,y}(f(x)-f(y))(g(x)-g(y))\pi(x)\mathbb{P}(x,y) \tag{3.54}$$

Definition 3.33. Let $\mathbb{P}$ be a Markov Chain with stationary distribution π. The *log-Sobolev* constant is defined as

$$\begin{aligned} \rho(\mathbb{P}) &= \min_{f,\mathcal{L}(f)\neq 0} \frac{\mathcal{E}_{\mathbb{P}}(f,f)}{\mathcal{L}(f)} \\ \text{where } \mathcal{L}(g) &= \mathbb{E}_\pi\left[g^2 \log\left(\frac{g^2}{\mathbb{E}_\pi[g^2]}\right)\right] \end{aligned} \tag{3.55}$$

The log-Sobolev constant was introduced in the context of finite Markov Chains by [11] to derive tight bounds on L^2-mixing time of (continuous time) Markov Chains. In the case of reversible chains, they showed a tight relation between the L^2-mixing times in discrete and continuous time. Rather than stating the result their result in full, we only state a consequence which we need

Theorem 3.34 (Diaconis and Saloff-Coste [11, Corollary 2.2, Theorem 3.7])**.** *Let* $\mathbb{P}$ *be a reversible Markov Chain with stationary distribution* π*. Assume that all eigenvalues of* $\mathbb{P}$ *are non-negative and* $\pi_* \leq 1/\mathrm{e}$*.*

$$\mathcal{T}_2(\mathbb{P},\epsilon) \leq \frac{1}{\rho(\mathbb{P})}\left(\log(1/\epsilon) + \frac{\log\log(1/\pi_*)}{4}\right) \tag{3.56}$$

Moreover, $\mathcal{T}_2(\mathbb{P}) \geq \frac{1}{2\rho(\mathbb{P})}$

Unfortunately the proof of the above result does not use the outline of § 3.6 and hence does not translate into a Robust mixing time bound. [34] uses the log-Sobolev constant to show a smooth entropy decay which immediately translates into a Robust mixing time bound.

Theorem 3.35 (Laurent [34, Proposition 6])**.** *Let $\mathbb{P}$ be a Markov Chain with stationary distribution π and μ any distribution.* $\mathrm{D}(\mu\mathbb{P}||\pi) \leq (1 - \rho(\mathbb{P}\overleftarrow{\mathbb{P}}))\,\mathrm{D}(\mu||\pi)$

Montenegro and Tetali [45] show a similar result for L^2-mixing time which uses the outline of § 3.6.

Theorem 3.36 (Montenegro and Tetali [45, Corollary 3.4])**.** *Let $\mathbb{P}$ be a Markov Chain with stationary distribution π.*

$$\mathcal{T}_2(\epsilon) \leq \frac{1}{\rho(\mathbb{P}\overleftarrow{\mathbb{P}})}\left(\log\log(1/\pi_*) - \log\log(1+\epsilon^2)\right) \tag{3.57}$$

Combining the above results, we have

Theorem 3.37. *Let $\mathbb{P}$ be a Markov Chain with stationary distribution π and assume $\pi_* \leq 1/\mathrm{e}$.*

$$\mathcal{T}_2(\mathbb{P}\overleftarrow{\mathbb{P}}) \leq \mathcal{R}_2(\mathbb{P}) \leq \mathcal{T}_2(\mathbb{P}\overleftarrow{\mathbb{P}}) \cdot O\left(\log\log(1/\pi_*)\right) \tag{3.58}$$

Proof. The lower bound follows from the adversarial strategy $\overleftarrow{\mathbb{P}}$. Let $\mathbb{Q} = \mathbb{P}\overleftarrow{\mathbb{P}}$ and note that the eigenvalues of $\mathbb{Q}$ are squares of the singular values of $\mathbb{S}(\mathbb{P})$ and hence are all non-negative.

Let $\rho = \rho(\mathbb{Q})$. By Theorem 3.34, $\mathcal{T}_2(\mathbb{Q}) \geq (2\rho)^{-1}$. By Theorem 3.36 together with observation in § 3.6 we have

$$\mathcal{R}_2(\mathbb{P}, \epsilon) \leq \frac{1}{\rho}\left(\log\log(1/\pi_*) - \log\log(1+\epsilon^2)\right) \tag{3.59}$$

Combining this with $1/\rho \leq 2\mathcal{T}_2(\mathbb{Q})$, we have the result. ★

This allows us to improve the estimate of Proposition 3.14 for L^2-mixing times.

Proposition 3.38. *Let $\mathbb{P}$ be a reversible lazy Markov Chain with stationary distribution π and assume $\pi_* \leq 1/\mathrm{e}$.*

$$\mathcal{T}_2(\mathbb{P}) \leq \mathcal{R}_2(\mathbb{P}) = \mathcal{T}_2(\mathbb{P}) \cdot O(\log\log(1/\pi_*)) \tag{3.60}$$

Proof. The lower bound follows by the adversary not doing anything. Let $\mathbb{Q} = \mathbb{P}\overleftarrow{\mathbb{P}} = \mathbb{P}^2$. From Theorem 3.36, we know that $\mathcal{R}_2(\mathbb{P}) = (\rho(\mathbb{Q}))^{-1}O(\log\log(1/\pi_*))$. Since $\mathbb{P}$ is lazy, Lemma 3.32 implies $\rho(\mathbb{Q}) \geq \rho(\mathbb{P})$. Hence

$$\mathcal{R}_2(\mathbb{P}) \leq \frac{1}{\rho(\mathbb{P})}O(\log\log(1/\pi_*)) \tag{3.61}$$

Since $\mathbb{P}$ is reversible, $\mathcal{T}_2(\mathbb{P}) \geq (2\rho(\mathbb{P}))^{-1}$ by Theorem 3.34. ★

3.7 Application: Liftings

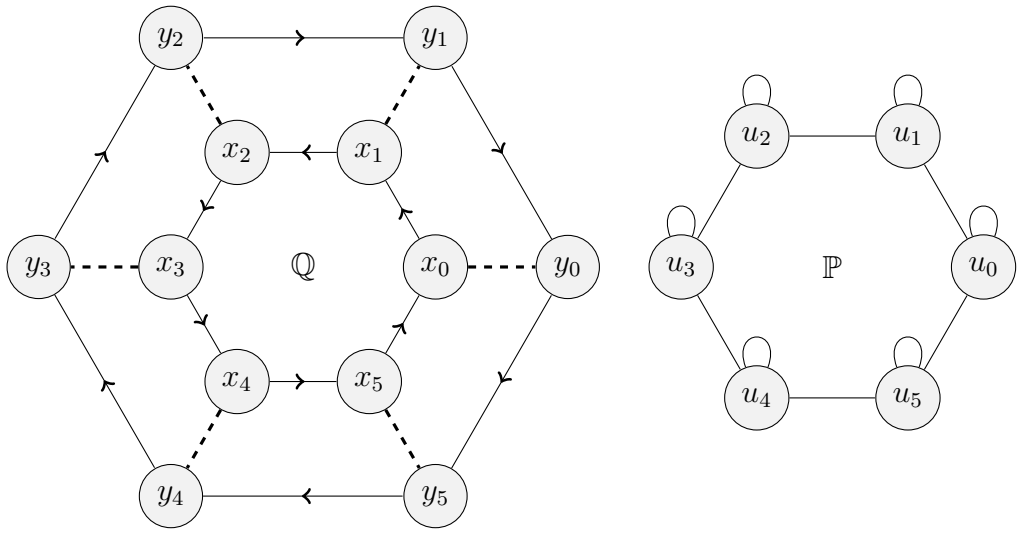
Definition 3.39. Let $\mathbb{P}$ and $\mathbb{Q}$ be Markov Chains on state spaces $\mathcal{X}$ and $\mathcal{Y}$ with stationary distributions π and μ respectively. $\mathbb{P}$ is said to be a *collapsing* of $\mathbb{Q}$ if there exists a mapping $F : \mathcal{Y} \to \mathcal{X}$ such that the following hold

- $\pi(x) = \mu(\mathcal{Y}_x)$ for all $x \in \mathcal{X}$ where $\mathcal{Y}_x = F^{-1}(x)$
- For all $x_1, x_2 \in \mathcal{X}$,

$$\mathbb{P}(x_1, x_2) = \sum_{y_1 \in \mathcal{Y}_{x_1}} \sum_{y_2 \in \mathcal{Y}_{x_2}} \mu^{x_1}(y_1)\mathbb{Q}(y_1, y_2) \tag{3.62}$$

where μ^x is the conditional distribution of $y \in \mathcal{Y}$ given $F(y) = x$, i.e. $\mu^x(y) = \mu(y)/\pi(x)$.

A *lifting* of $\mathbb{P}$ is a chain $\mathbb{Q}$ such that $\mathbb{P}$ is the collapsing of $\mathbb{Q}$.



- dashed edges of $\mathbb{Q}$ have weight $1/n$; solid edges of $\mathbb{Q}$ have weight $1 - 1/n$
- The projection map is defined via $F(x_i) = F(y_i) = u_i$
- Hence loops in $\mathbb{P}$ have weight $1/n$; remaining edges of $\mathbb{P}$ have weight $(1/2 - 1/2n)$

Figure 3.2: Double cover of cycle

Example 3.40. Figure 3.2 shows an example of lifting $\mathbb{Q}$ of $\mathbb{P}$. Assume n is odd so that there are no periodicity effects. Since $\mathbb{P}$ is just the random walk on the cycle with a small holding probability, we have $\mathcal{T}(\mathbb{P}) = \Theta(n^2)$. On the other hand, one can show that $\mathcal{T}_2(\mathbb{Q}) = O(n \log n)$ by explicitly diagonalizing $\mathbb{Q}$. Thus by lifting the Markov Chain $\mathbb{P}$ to $\mathbb{Q}$ we have decreased the mixing time by almost a quadratic factor.

Theorem 3.41 (Chen et al. [8])**.** *Let $\mathbb{P}$ be a reversible Markov Chain and $\mathbb{Q}$ a lifting of $\mathbb{P}$.*

(a) $\mathcal{T}(\mathbb{P}) = O(\max(\mathcal{T}(\mathbb{Q}), \mathcal{T}(\overleftarrow{\mathbb{Q}}))^2 \log(1/\pi_*))$

(b) If $\mathbb{Q}$ is reversible, $\mathcal{T}(\mathbb{P}) = O(\mathcal{T}(\mathbb{Q}) \log(1/\pi_*))$

(c) There is an explicit family of reversible liftings $\mathbb{Q}$ of $\mathbb{P}$ for which

$$\mathcal{T}(\mathbb{P}) = \Omega\left(\frac{\mathcal{T}(\mathbb{Q}) \log(1/\pi_*)}{\log\log(1/\pi_*)}\right) \tag{3.63}$$

where $\mu_* := \min_y \mu(y) = \Theta(\pi_*)$.

Proposition 3.42. *Let $\mathbb{Q}$ be a lifting of $\mathbb{P}$. For $x \in \mathcal{X}$, let μ^x denote the distribution μ conditioned on $F(y) = x$. Given a distribution ν on $\mathcal{X}$ put*

$$\widehat{\nu}(y) = \nu(F(y))\mu^{F(y)}(y) \tag{3.64}$$

For all $1 \leq p \leq \infty$, $\|\widehat{\nu} - \mu\|_{p,\pi} = \|\nu - \pi\|_{p,\pi}$. Also $\mathrm{D}(\widehat{\nu}||\mu) = D(\nu||\pi)$.

Proof. Let

$$f(x) = \frac{\nu(x)}{\pi(x)} \quad \text{and} \quad \widehat{f}(y) = \frac{\widehat{\nu}(x)}{\mu(x)} \tag{3.65}$$

be the density functions of $\nu - \pi$ and $\widehat{\nu} - \mu$ respectively. For $y \in \mathcal{Y}$ and $x = F(y)$ we have

$$\frac{\widehat{\nu}(y)}{\mu(y)} = \frac{\nu(x)}{\pi(x)} \tag{3.66}$$

Also since $\mathbb{Q}$ is a lifting of $\mathbb{P}$, we have

$$\sum_{F(y)=x} \mu(y) = \pi(x) \tag{3.67}$$

Now the result follows from the definition of the p-norms and $\mathrm{D}(\cdot)$. $\square$

The crucial observation is the following

Theorem 3.43. *Let $\mathbb{Q}$ be a lifting of $\mathbb{P}$ and $1 \leq p \leq \infty$. Then $\mathcal{R}_p(\mathbb{Q}) \geq \mathcal{R}_p(\mathbb{P})$.*

Proof. Fix $1 \leq p \leq \infty$. Let $\mathcal{P} = (\mathbb{P}, \{\mathbb{A}_t\})$ be an adversarial modification of $\mathbb{P}$ such that $\mathcal{T}(\mathcal{P}) = \mathcal{R}(\mathbb{P})$. We now exhibit an adversarial strategy $\{\mathbb{B}_t\}$ against $\mathbb{Q}$ so that the adversarially modified $\mathcal{Q} = (\mathbb{Q}, \{\mathbb{B}_t\})$ simulates the evolution of $\mathcal{P}$.

Let $F : \mathcal{Y} \to \mathcal{X}$ be the projection map and consider the following adversarial strategy $\mathbb{C}$:

$$\mathbb{C}(y, y') = \begin{cases} \mu^x(y') & \text{if } F(y) = F(y') = x \\ 0 & \text{otherwise} \end{cases} \tag{3.68}$$

i.e. given $y \in \mathcal{Y}$, the adversary picks a state $y' \in \mathcal{Y}$ according to the distribution μ^x where $x = F(y)$. Recall that μ^x is the conditional distribution of μ given that $F(y) = x$.

For $y' \in \mathcal{Y}$ and $x' = F(y')$, we have

$$(\mu\mathbb{C})(y') = \sum_{y \in \mathcal{Y}} \mu(y)\mathbb{C}(y, y') = \sum_{y \in \mathcal{Y}_{x'}} \mu(y)\mu^{x'}(y') = \mu^{x'}(y')\mu(\mathcal{Y}_{x'}) = \mu(y') \tag{3.69}$$

Thus $\mathbb{C}$ is compatible with $\mathbb{Q}$.

Given a distribution ν on $\mathcal{X}$, we lift it to distribution $\widehat{\nu}$ on $\mathcal{Y}$ as defined in Proposition 3.42. Also, Proposition 3.42 implies $\|\widehat{\nu} - \mu\|_{p,\pi} = \|\nu - \pi\|_{p,\pi}$.

Suppose we start with the distribution ν on $\mathcal{X}$ and $\widehat{\nu}$ on $\mathcal{Y}$. We show that applying

$\mathbb{P}$ to ν is "equivalent" to applying $\mathbb{QA}$ to $\widehat{\nu}$.

$$\begin{aligned}(\widehat{\nu}\,\mathbb{Q})(\mathcal{Y}_{x_2}) &= \sum_{y_1\in\mathcal{Y}}\sum_{y_2\in\mathcal{Y}_{x_2}} F(\nu)(y_1)\mathbb{Q}(y_1,y_2)\\ &= \sum_{x_1\in\mathcal{X}}\sum_{y_1\in\mathcal{Y}_{x_1}}\sum_{y_2\in\mathcal{Y}_{x_2}} \nu(x_1)\mu^{x_1}(y_1)\mathbb{Q}(y_1,y_2)\\ &= \sum_{x_1\in\mathcal{X}} \nu(x_1)\mathbb{P}(x_1,x_2)\\ &= \widehat{\nu\mathbb{P}}\,(x_2)\end{aligned} \tag{3.70}$$

which implies $\widehat{\nu}\,\mathbb{QC} = \widehat{\nu\mathbb{P}}$. Thus the Markov Chain $\mathbb{QC}$ simulates $\mathbb{P}$ and hence $\mathcal{R}_p(\mathbb{Q}) \geq \mathcal{T}_p(\mathbb{QC}) \geq \mathcal{T}_p(\mathbb{P})$.

To show $\mathcal{R}_p(\mathbb{Q}) \geq \mathcal{R}_p(\mathbb{P})$, we just need to lift the strategy $\{\mathbb{A}_t\}$ on $\mathcal{X}$ to $\{\mathbb{B}_t\}$ on $\mathcal{Y}$. Given $\{\mathbb{A}_t\}$ define $\mathbb{B}_t$ as follows:

$$\mathbb{B}_t(y_1,y_2) = \mu^{x_2}(y_2)\mathbb{A}_t(x_1,x_2) \tag{3.71}$$

where $x_1 = F(y_1)$ and $x_2 = F(y_2)$. First observe that if $\mathbb{A}_t = I$ then $\mathbb{B}_t = \mathbb{C}$. Also for $x_1, x_2 \in \mathcal{X}$, we have

$$\sum_{y_1\in\mathcal{Y}_{x_1}, y_2\in\mathcal{Y}_{x_2}} \mu^{x_1}(y_1)\mathbb{B}_t(y_1,y_2) = \mathbb{A}_t(x_1,x_2) \tag{3.72}$$

showing that the transitions of $\mathbb{B}_t$ and $\mathbb{A}_t$ are compatible with the projection F. It only remains to verify that for any distribution ν on $\mathcal{X}$, $\widehat{\nu}\,\mathbb{B}_t = \widehat{\nu\mathbb{A}_t}$.

For $y_2 \in \mathcal{Y}$ such that $F(y_2) = x_2$, (3.71) implies

$$(\widehat{\nu}\,\mathbb{B}_t)(y_2) = \sum_{y_1\in\mathcal{Y}} \widehat{\nu}\,(y_1)\mathbb{B}_t(y_1,y_2) = \mu^{x_2}(y_2)\sum_{y_1\in\mathcal{Y}} \nu(x_1)\mu^{x_1}(y_1)\mathbb{A}_t(x_1,x_2) \tag{3.73}$$

where $x_1 = F(y_1)$. Thus $\widehat{\nu}\,\mathbb{B}_t$ restricted to $\mathcal{Y}_{x_2} = F^{-1}(x_2)$ is proportional to μ^{x_2}. On the other hand, summing (3.72) over $x_1 \in \mathcal{X}$, shows $(\widehat{\nu}\,\mathbb{B}_t)(\mathcal{Y}_{x_2}) = (\nu\mathbb{A}_t)(x_2)$ for any $x_2 \in \mathcal{X}$. Hence $\widehat{\nu}\,\mathbb{B}_t = \widehat{\nu\mathbb{A}_t}$. It now follows from Proposition 3.42 that $\mathcal{Q}$ simulates $\mathcal{P}$ which implies $\mathcal{R}_p(\mathbb{Q}) \geq \mathcal{R}_p(\mathbb{P})$. ★

Together with relation between Robust and standard mixing times, we have

Proposition 3.44. *Let $\mathbb{Q}$ with stationary distribution μ be a lifting of $\mathbb{P}$ with stationary distribution π. Let $\mu_* = \min_{y\in\mathcal{Y}} \mu(y)$.*

(a) If $\mathbb{Q}$ is lazy, $\mathcal{T}(\mathbb{P}) \leq \mathcal{T}_2(\mathbb{P}) \leq \mathcal{T}(\mathbb{Q})^2 \log(1/\mu_*)$

(b) If $\mathbb{Q}$ is reversible, $\mathcal{T}(\mathbb{P}) \leq \mathcal{T}(\mathbb{Q}) \log(1/\mu_*)$

(c) If $\mathbb{Q}$ is reversible and lazy, $\mathcal{T}_2(\mathbb{P}) \leq \mathcal{T}_2(\mathbb{Q}) \log\log(1/\mu_*)$

Proof. By Theorem 3.43 we have $\mathcal{R}_p(\mathbb{Q}) \geq \mathcal{R}_p(\mathbb{P}) \geq \mathcal{T}_p(\mathbb{P})$ for $p = 1, 2$.

(a) Theorem 3.28 implies $\mathcal{R}_2(\mathbb{Q}) \leq \mathcal{T}(\mathbb{Q})^2 \log(1/\mu_*)$.

(b) Proposition 3.14 implies $\mathcal{R}(\mathbb{Q}) \leq \mathcal{T}(\mathbb{Q}) \log(1/\mu_*)$.

(c) Proposition 3.38 implies $\mathcal{R}_2(\mathbb{Q}) \leq \mathcal{T}_2(\mathbb{Q}) \log\log(1/\mu_*)$. ★

Note that Proposition 3.44(b) has a $\log(1/\mu_*)$ while Theorem 3.41 has $\log(1/\pi_*)$. The difference is only a constant factor if μ_* is only polynomially smaller than π_*. In potential applications, μ_* is usually only a constant factor smaller than π_*.

For the general case, the laziness requirement can be dropped as well as $\log(1/\mu_*)$ improved to $\log(1/\pi_*)$ if we do not go via the Robust mixing time and observe that the conductance of $\mathbb{Q}$ is smaller than that of $\mathbb{P}$ and use Theorem 3.26 and Theorem 3.27. See [18] for details.

When $\mathbb{Q}$ is reversible, we improve the Theorem 3.41 by analyzing the strategy $\mathbb{C}$ used in the proof of Theorem 3.43.

Theorem 3.45. *Let $\mathbb{Q}$ be a reversible lifting of $\mathbb{P}$. Assume that $\pi_* \leq 1/\mathrm{e}$. Then*

$$\mathcal{T}(\mathbb{Q}) \geq \mathcal{T}_{\mathrm{rel}}(\mathbb{P}) \log(2) \tag{3.74}$$

$$\mathcal{T}_2(\mathbb{Q}) \geq \frac{1}{2\rho(\mathbb{P})} \tag{3.75}$$

Proof. Let $\mathbb{C}$ be the stochastic matrix as defined in Theorem 3.43. Note that $\mathbb{C}$ is reversible, since it reaches stationarity in one step in each reducible component.

Let $\vec{\mathsf{v}}$ denote the eigenvector (of length $|\mathcal{X}|$) corresponding to $\lambda_*(\mathbb{P})$ and lift $\vec{\mathsf{w}}$ to

$$\vec{\widehat{\mathsf{w}}} = \sum_{x \in \mathcal{X}} \vec{\mathsf{w}}(x) \mu^x(y) \tag{3.76}$$

It is easy to see that for $x \in \mathcal{X}$, $\sum_{y \in \mathcal{Y}_x} (\vec{\widehat{\mathsf{w}}}\, \mathbb{Q})(y) = (w\mathbb{P})(x)$. Hence

$$\vec{\widehat{\mathsf{w}}}\, (\mathbb{Q}\mathbb{C})(y) = \lambda_*(\mathbb{P}) \vec{\mathsf{w}}(x) \mu^x(y) = \lambda_*(\mathbb{P}) \vec{\widehat{\mathsf{w}}}\, (y) \tag{3.77}$$

where $x = F(y)$. Thus $\lambda_*(\mathbb{Q}\mathbb{C}) \geq \lambda_*(\mathbb{P})$.

Since $\mathbb{A}$ is a contraction (it is stochastic), we have $|\lambda_*(\mathbb{Q})| \geq |\lambda_*(\mathbb{Q}\mathbb{C})| \geq |\lambda_*(\mathbb{P})|$. Hence by Theorem 2.40, $\mathcal{T}(\mathbb{Q}) = \mathcal{T}(\mathbb{Q}, 1/4) \geq \mathcal{T}_{\mathrm{rel}}(\mathbb{Q}) \log(2) \geq \mathcal{T}_{\mathrm{rel}}(\mathbb{P}) \log(2)$.

For the second result, by Theorem 3.34 it is enough to show that $\rho(\mathbb{Q}) \leq \rho(\mathbb{P})$. For $f : \mathcal{X} \to \mathbb{R}$, define $g : \mathcal{Y} \to \mathbb{R}$ via $g(y) = f(F(y))$. Then

$$\|g\|_{2,\mu}^2 = \sum_{x \in \mathcal{X}} \pi(x) \sum_{y = F(x)} \mu^x(y) g(y)^2 = \sum_{x \in \mathcal{X}} \pi(x) f(x)^2 = \|f\|_{2,\pi}^2 \tag{3.78}$$

It is easy to see that $\mathcal{L}_\mu(g) = \mathcal{L}_\pi(f)$ and since $\mathbb{Q}$ is a reversible lifting of $\mathbb{P}$, we have

$$\begin{aligned}
2\,\mathcal{E}_{\mathbb{Q}}(g,g) &= \sum_{x_1,x_2\in\mathcal{X}} \sum_{F(y_1)=x_1} \sum_{F(y_2)=x_2} \pi(x_1)\mu^{x_1}(y_1)\mathbb{Q}(y_1,y_2)(g(y_1)-g(y_2))^2 \\
&= \sum_{x_1,x_2\in\mathcal{X}} \pi(x_1)(f(x_1)-f(x_2))^2 \left(\sum_{F(y_1)=x_1} \sum_{F(y_2)=x_2} \mu^{x_1}(y_1)\mathbb{Q}(y_1,y_2) \right) \\
&= \sum_{x_1,x_2\in\mathcal{X}} \pi(x_1)(f(x_1)-f(x_2))^2\mathbb{P}(x_1,x_2) \\
&= 2\,\mathcal{E}_{\mathbb{P}}(f,f) \qquad (3.79)
\end{aligned}$$

Hence $\rho(\mathbb{Q}) \le \rho(\mathbb{P})$. ★

For many reversible chains $\mathcal{T}(\mathbb{P}) = O(\mathcal{T}_{\text{rel}}(\mathbb{P}))$. Theorem 3.45 shows that one cannot gain more than a constant factor improvement by considering a reversible lifting. Similarly if $\mathcal{T}_2(\mathbb{P}) = O(\rho(\mathbb{P})^{-1})$, one cannot gain more than a constant factor by taking a reversible lifting.

3.8 Discussion

Boyd et al. [6] considers the **fastest mixing Markov Chain on a graph** problem. Given an undirected graph X, assign weights to the edges of the graph, so that the Random walk on the resulting graph has the largest spectral gap. Robust Mixing is a complementary notion which addresses the question of how much can one slow down a Markov Chain.

Proposition 3.14 shows that for a reversible chain the adversary cannot slow it down by more than a $O(\log(1/\pi_*)$ factor. However the example in Theorem 3.41 together with Theorem 3.43 shows that this gap is nearly tight.

Chapter 4

Cayley Walks on Groups

In this chapter, we look at Markov Chains which arise from walks on finite groups.

Definition 4.1. Let G be a finite group and $P(\cdot)$ a distribution over G. The Cayley walk induced by P on G is the Markov Chain $\mathbb{P}$ with following transition matrix

$$\mathbb{P}(\mathsf{g}, \mathsf{h}) = P(\mathsf{g}^{-1}\mathsf{h}) \tag{4.1}$$

In other words, $\mathbb{P}$ takes the current group element g and right multiplies it with an element s of G chosen from the distribution P.

By a Cayley walk on G we mean a Cayley walk on G induced by some probability distribution P on G.

Definition 4.2. For a distribution P over G, the support of P, denoted $\mathrm{supp}(P)$ is defined as $\{\mathsf{h} \in G : P(\mathsf{h}) > 0\}$.

Lemma 4.3. *Let $\mathbb{P}$ be a Cayley walk on a finite group G induced by P.*

(a) $\mathbb{P}$ is irreducible iff $\mathrm{supp}(P)$ *generates G*

(b) $\mathbb{P}$ *is aperiodic iff* $\mathrm{supp}(P)$ *does not lie in a coset of some non-trivial subgroup of* G

In particular, $\mathbb{P}$ *is ergodic if* $\mathrm{supp}(P)$ *generates* G *and* $P(\mathsf{1}) > 0$*, where* $\mathsf{1}$ *is the identity element of* G*.*

Cayley walks of groups are a well studied class of Markov Chains. Since the transition matrices of Cayley walks are doubly stochastic, the stationary distribution is always uniform. The underlying group structure ensures that Cayley walks are vertex transitive and hence Corollary 2.45 applies. However estimating all the eigenvalues of the transition matrix is not always easy and usually leads to problems dealing with representation theory of G, as the invariant subspaces of $\mathbb{P}$ are the irreducible representations of G (the $|G|$-dimensional vector space is the so called *regular representation* of G).

In this chapter, we look at the Robust mixing time of Cayley walks under certain restricted adversaries.

4.1 Cayley Adversaries

Definition 4.4. Let $\mathbb{P}$ be a Cayley walk on a finite group G. A *Cayley strategy* is a sequence of doubly stochastic matrices $\{\mathbb{A}_t\}$ such that each $\mathbb{A}_t$ is the transition matrix of some Cayley walk on G. Denote by $\mathcal{C}$ the set of all Cayley strategies on G (G will be clear from context).

G. [21] showed that set of $N \times N$ doubly stochastic matrices equals the convex hull of the $N!$ permutation matrices. In case of a Cayley strategy, the choice of $\mathbb{A}_t$ is limited to the convex combinations of the N right translations, where $N = |G|$.

Definition 4.5. Let $\mathbb{P}$ be a Cayley walk on a group G. The *Cayley Robust mixing times* of $\mathbb{P}$, are defined via

$$\mathcal{R}^{\mathcal{C}}(\mathbb{P}, \epsilon) = \sup_{\mathcal{P}} \mathcal{T}(\mathbb{P}, \epsilon) \qquad \mathcal{R}^{\mathcal{C}}_p(\mathbb{P}, \epsilon) = \sup_{\mathcal{P}} \mathcal{T}_p(\mathbb{P}, \epsilon) \tag{4.2}$$

where the suprema are taken over adversarially modified versions $\mathcal{P} = (\mathbb{P}, \{\mathbb{A}_t\})$ of $\mathbb{P}$ where $\{\mathbb{A}_t\}$ is a Cayley strategy.

Since the set of Cayley strategies is *symmetric* as well as a *valid set of strategies* (see Definition 3.6), the Cayley Robust mixing time enjoys all the basic properties of Robust mixing time.

Another natural class of adversaries is the following:

Definition 4.6. Let G be a finite group. A *translation strategy* is a sequence of doubly stochastic matrices $\{\mathbb{A}_t\}$ such that each $\mathbb{A}_t$ can be written as the convex combination of two-sided translations $\eta_{\mathsf{s},\mathsf{s}'}$, where $\eta_{\mathsf{s},\mathsf{s}'}$ is the permutation matrix corresponding to the permutation $\mathsf{g} \mapsto \mathsf{sgs}'$

A *translation adversary* is one who is restricted to translation strategies.

While a Cayley adversary is only allowed right translations by elements of the group, a *translation adversary* is allowed two-sided translations. We quickly see this does not buy any additional power for the adversary.

Proposition 4.7. *Let $\mathbb{P}$ be a Cayley walk on a group G. The Cayley Robust mixing time equals the* translation Robust mixing time *under any distance measure which is invariant under permutations.*

In particular, for the L^p mixing times and entropy mixing time, a translation adversary has the same power as a Cayley adversary.

Proof. Since the set of translation strategies is a convex hull of two-sided translations, the translation Robust mixing time of $\mathbb{P}$ is the mixing time of $\mathcal{P} = (\mathbb{P}, \{\mathbb{A}_t\})$ where each $\mathbb{A}_t$ corresponds to the two-sided translation $\eta_{\ell(t),r(t)}$ for some fixed elements $\ell(t), r(t)$ of G.

Now consider the Cayley strategy $\mathcal{Q} = (\mathbb{P}, \{\mathbb{B}_t\})$ where $\mathbb{B}_t$ corresponds to the right translation by $r(t)$. If we define $L(t) = \ell(1) \cdot \ell(2) \dots \ell(t)$, it is clear that the t-round distribution of $\mathcal{P}$ and $\mathcal{Q}$ only differ by a left translation by $L(t)$. Since the stationary distribution is uniform and the distance measure is invariant under permutations, they are equal. ★

4.2 Holomorphic Adversaries

We now consider another natural set of strategies which preserve the symmetry of underlying group. In additional to allowing translations we also allow the adversary to apply automorphisms of the group.

Definition 4.8. Let G be a finite group. A permutation J on G is said to be a *holomorphism* if it can be written as the composition of a right translation and an automorphism of G.

Definition 4.9. Let G be a group. A *holomorphic strategy* is a sequence $\{\mathbb{A}_t\}$ of matrices each of which can be written as the convex combination of holomorphisms of G. Denote by $\mathcal{H}$ the set of all holomorphic strategies of G (G will be clear from context).

Definition 4.10. Let $\mathbb{P}$ be a Cayley walk on a group G. The *holomorphic Robust*

mixing times of $\mathbb{P}$, are defined via

$$\mathcal{R}^{\mathcal{H}}(\mathbb{P},\epsilon) = \sup_{\mathcal{P}} \mathcal{T}(\mathbb{P},\epsilon) \qquad \mathcal{R}_p^{\mathcal{H}}(\mathbb{P},\epsilon) = \sup_{\mathcal{P}} \mathcal{T}_p(\mathbb{P},\epsilon) \tag{4.3}$$

where the suprema are taken over adversarially modified versions $\mathcal{P} = (\mathbb{P}, \{\mathbb{A}_t\})$ of $\mathbb{P}$ where $\{\mathbb{A}_t\}$ is a holomorphic strategy.

Like Cayley strategies, the set $\mathcal{H}$ of holomorphic strategies are symmetric as well as a valid set of strategies against any Cayley walk on G.

Definition 4.11. (a) For a group G, let $\mathrm{Aut}(G)$ denote the group of automorphisms of g.

(b) $f \in \mathrm{Aut}(G)$ is said to be an *inner automorphism* if $f(\mathsf{g}) = \mathsf{hgh}^{-1}$ for some $\mathsf{h} \in G$, i. e., f corresponds to a conjugation.

(c) $\mathrm{Inner}(G)$ denote the sub-group of inner automorphisms of G

(d) $\mathrm{Outer}(G) \cong \mathrm{Aut}(G)/\mathrm{Inner}(G)$ denote the group of *outer-automorphisms* of G.

Proposition 4.12. *Let $\mathbb{P}$ be a Cayley walk on a group G. If $\mathrm{Outer}(G)$ is trivial, then the Cayley Robust mixing times are the same as the holomorphic Robust mixing times.*

In particular, this is the case for the symmetric group S_n for $n \geq 5$,$n \neq 6$.

Proof. If $\mathrm{Outer}(G)$ is trivial, then all automorphisms are only conjugations. Hence a translation adversary can simulate the moves of a holomorphic adversary. By Proposition 4.7 we have the result.

Since $\mathrm{Outer}(S_n)$ is trivial for $n \geq 6$ the claim for S_n holds. ★

Definition 4.13. Let G be a group. By the *holomorph* of G, we mean the semi-direct product $\mathrm{Hol}(G) := G \rtimes \mathrm{Aut}(G)$, where $\mathrm{Aut}(G)$ acts on G naturally.

Example 4.14. Consider the lazy random walk on the hypercube. Here $G = \mathbb{Z}_2^n$. A Cayley adversary would only be allowed to flip coordinates (which are chosen arbitrarily). A holomorphic adversary on the other hand, can also apply any permutation on the coordinates, since $\mathrm{Hol}(\mathbb{Z}_2^n) = \mathbb{Z}_2^n \rtimes S_n$.

4.3 Upper Bounds on Holomorphic Robust Mixing Time

Let $\mathbb{P}$ be a Cayley walk on a finite group G and assume $|G| \geq 3$. Note that $\overleftarrow{\mathbb{P}}$ is a legal Cayley strategy. Hence from Theorem 3.37, we have

$$\mathcal{T}_2(\mathbb{P}\overleftarrow{\mathbb{P}}) \leq \mathcal{R}_2^{\mathcal{C}}(\mathbb{P}) \leq \mathcal{R}_2^{\mathcal{H}}(\mathbb{P}) \leq \mathcal{T}_2(\mathbb{P}\overleftarrow{\mathbb{P}}) \cdot O(\log\log|G|) \tag{4.4}$$

In this section we improve this result by eliminating the $O(\log\log|G|)$ factor and show $\mathcal{R}_2^{\mathcal{H}}(\mathbb{P}) \leq 2\mathcal{T}_2(\mathbb{P}\overleftarrow{\mathbb{P}})$.

We identify holomorphisms of G with the permutation they induce. This permutation representation of G is faithful, i. e., only the identity holomorphism induces the identity permutation on G. To see this consider $\alpha = \mathsf{h} \rtimes f \in \mathrm{Hol}(G)$ where $f \in \mathrm{Aut}(G)$ and suppose that α induces the identity permutation on G, i.e. α fixes all elements of G. Since f always fixes the identity element of G, we have $\mathsf{1} = \alpha(\mathsf{1}) = \mathsf{h}$. Thus α must be an automorphism of G. Now if α fixes every element of G then α is the identity automorphism.

The holomorphic adversary corresponds to $\mathrm{Hol}(G)$ and the Cayley adversary cor-

responds to $G \leq \mathrm{Hol}(G)$.

Definition 4.15. Let G be a group. A permutation J on G is said to be *G-respecting* if there exists a permutation K on G such that for all $\mathsf{h}, \mathsf{g} \in G$,

$$J(\mathsf{h})^{-1} J(\mathsf{g}) = K(\mathsf{h}^{-1}\mathsf{g}) \tag{4.5}$$

Proposition 4.16. *Let G be a group. A permutation J on G is G-respecting iff it is a holomorphism of G.*

Proof. Let $\mathcal{G}$ be the set of all G-respecting permutations of G. If $J \in \mathrm{Aut}(G)$, then J is G-respecting (take $K = J$). If J is a right translation, then J is G-respecting (take K to be identity). We now show that $\mathcal{G}$ is closed under compositions.

Suppose $J_1, J_2 \in \mathcal{G}$ and let K_1, K_2 be the corresponding permutations given by Definition 4.15. Then for $\mathsf{h}, \mathsf{g} \in G$,

$$J_1(J_2(\mathsf{h}))^{-1} \cdot J_1(J_2(\mathsf{g}))^{-1} = K_1(J_2(\mathsf{h})^{-1} J_2(\mathsf{g})) = K_1(K_2(\mathsf{h}^{-1}\mathsf{g})) \tag{4.6}$$

Hence $\mathcal{G}$ contains all the holomorphisms of G.

Now let J be any G-respecting permutation of G and set $J'(\mathsf{h}) = J(\mathsf{h})J(\mathsf{1})^{-1}$. Clearly J is a holomorphism iff J' is a holomorphism. Since J is G-respecting and $\mathcal{G}$ is closed under compositions and contains right translations, J' is also G-respecting. Moreover, $J'(\mathsf{1}) = \mathsf{1}$. If K' is the permutation on G given for J' by Definition 4.15 we have

$$J'(\mathsf{g}) = J'(\mathsf{1})^{-1} J'(\mathsf{g}) = K'(\mathsf{1}^{-1}\mathsf{g}) = K'(\mathsf{g}) \tag{4.7}$$

and now $K' = J'$ implies $J'(\mathsf{h})^{-1} J'(\mathsf{g}) = J'(\mathsf{h}^{-1}\mathsf{g})$, i. e., J' is an automorphism of G. Hence J is a holomorphism of G. ★

Definition 4.17. Let G be a group and $\mathbb{A}$ be a square matrix whose rows and columns are indexed by elements of G. $\mathbb{A}$ is said to be *G-circulant* if for some $F : G \to \mathbb{R}$,

$$\mathbb{A}(\mathsf{h}, \mathsf{g}) = F(\mathsf{h}^{-1}\mathsf{g}) \tag{4.8}$$

Note that transition matrices of Cayley walks on G are G-circulants as are the strategies of an Cayley adversary. Even though holomorphic strategies do not correspond to G-circulants they respect G-circulant matrices as shown in the following

Proposition 4.18. *(a) G-circulants are closed under products, affine combinations and taking transposes.*

(b) Let J be a G-respecting permutation and $\mathbb{A}$ be a G-circulant. Then so is $J\mathbb{A}J^T$ where by abuse of notation J also stands for the $|G| \times |G|$ matrix representing the permutation J on G.

Proof. (a) If $\mathbb{A}_i(\mathsf{h}, \mathsf{g}) = F_i(\mathsf{h}^{-1}\mathsf{g})$, then products correspond to convolution, affine combinations correspond to affine combinations and $\mathbb{A}^T$ corresponds to $F'(\mathsf{g}) = F(\mathsf{g}^{-1})$.

(b) Let K denote the permutation corresponding to J as given by Definition 4.15. Note that since J is a permutation matrix, $J^T = J^{-1}$.

$$(J^{-1}\mathbb{A}J)(\mathsf{h}, \mathsf{g}) = \mathbb{A}(J(\mathsf{h}), J(\mathsf{g})) = F(J(\mathsf{h})^{-1} \cdot J(\mathsf{g})) = F(K(\mathsf{h}^{-1}\mathsf{g})) \tag{4.9}$$

Hence the result. ★

Combining all of the above, we can show the following

Theorem 4.19. *Let $\mathbb{P}$ be a Cayley walk on a group G. Then $\mathcal{R}_2^{\mathcal{H}}(\mathbb{P}) \leq 2\mathcal{T}_2(\mathbb{P}\,\overleftarrow{\mathbb{P}})$*

Proof. Let $\mathcal{P} = (\mathbb{P}, \{\mathbb{A}_t\})$ be any adversarially modified version of $\mathbb{P}$ where $\{\mathbb{A}_t\}$ is a holomorphic strategy. By convexity and Proposition 4.16 we may assume that each $\mathbb{A}_t$ corresponds to some G-respecting permutation J_t on G.

For $t > 0$, let $\mathbb{Q}_t = \mathbb{P}\mathbb{A}_1\mathbb{P}\mathbb{A}_2 \dots \mathbb{P}\mathbb{A}_t$ so that the t-round distribution is given by $\mu\mathbb{Q}_t$ for an initial distribution μ. Note that for a Markov Chain with uniform stationary distribution, its reverse chain is just its transpose.

For $\mathsf{g} \in G$ and initial state g, applying Proposition 2.35 for $\mathbb{Q}_t$ we get

$$\|\delta_{\mathsf{g}}\mathbb{Q}_t\pi\|_{2,\pi}^2 = |G|\mathbb{Q}_t\overleftarrow{\mathbb{Q}_t}(\mathsf{g},\mathsf{g}) - 1 \tag{4.10}$$

Now evaluate $\mathbb{Q}_t\overleftarrow{\mathbb{Q}_t}$ inside out by defining

$$\mathbb{C}_{t+1} = I \qquad\qquad \mathbb{C}_k = \mathbb{P}(\mathbb{A}_k\mathbb{C}_{k+1}{\mathbb{A}_k}^T)\mathbb{P}^T \tag{4.11}$$

so that $\mathbb{C}_1 = \mathbb{Q}_t\overleftarrow{\mathbb{Q}_t}$. Clearly $\mathbb{C}_{t+1}$ is G-circulant. Assuming $\mathbb{C}_{k+1}$ is a G-circulant, Proposition 4.18 together with the fact that $\mathbb{A}_t$ is a G-respecting permutation shows that $\mathbb{A}_k\mathbb{C}_{k+1}{\mathbb{A}_k}^T$ is also G-circulant. Since $\mathbb{P}$ is G-circulant, it follows that $\mathbb{C}_k$ is G-circulant. Hence $\mathbb{Q}_t\overleftarrow{\mathbb{Q}_t}$ is G-circulant.

From (4.10) and the fact that $\mathbb{Q}_t\overleftarrow{\mathbb{Q}_t}$ is constant on the diagonal, we have

$$\|\mu\mathbb{Q}_t - \pi\|_{2,\pi}^2 \leq \operatorname{tr}(\mathbb{Q}_t\overleftarrow{\mathbb{Q}_t}) - 1 \tag{4.12}$$

for any initial distribution μ. Also,

$$\operatorname{tr}(\mathbb{Q}_t\overleftarrow{\mathbb{Q}_t}) = \sum_{i=0}^{|G|-1}\sigma_i(\mathbb{Q}_t)^2 \leq \sum_{i=0}^{|G|-1}\prod_{j=1}^{t}\sigma_i(\mathbb{P})^2\sigma_i(\mathbb{A}_j)^2 \leq \sum_{i=0}^{|G|-1}\sigma_i(\mathbb{P})^{2t} \tag{4.13}$$

since the singular values of a product is majorized the product of the singular values (see [29, Chapter 3] for details) and $\sigma_i(\mathbb{A}_j) \leq 1$ ($\mathbb{A}_j$ is a permutation). Since $\sigma_0(\mathbb{A}_j) = \sigma_0(\mathbb{P}) = 1$, the term $i = 0$ evaluates to 1. Hence for any initial distribution μ,

$$\|\mu\mathbb{Q}_t - \pi\|_{2,\pi}^2 \leq \sum_{i=1}^{|G|-1} \sigma_i(\mathbb{P})^{2t} \tag{4.14}$$

Now let $\mathbb{P}' = \mathbb{P}\overleftarrow{\mathbb{P}}$, $\mathsf{g} \in G$ and $s > 0$. Consider the s-step distribution of the Markov Chain $\mathbb{P}'$ when the initial distribution is δ_{g}. By Proposition 2.35, we have

$$\|\mathbb{P}'^s(\mathsf{g}, \cdot) - \pi\|_{2,\pi}^2 = |G|(\mathbb{P}'^s \overleftarrow{\mathbb{P}'^s})(\mathsf{g}, \mathsf{g}) - 1 \tag{4.15}$$

Since $\mathbb{P}'$ is G-circulant and doubly stochastic, we have $\|\mathbb{P}'^s(\mathsf{g}, \cdot) - \pi\|_{2,\pi}^2 = \mathrm{tr}(\mathbb{P}'^{2s}) - 1$. But the eigenvalues of $\mathbb{P}'$ are the squares of the singular values of $\mathbb{P}$ and the top eigenvalue is 1. Hence

$$\|\mathbb{P}'^s(\mathsf{g}, \cdot) - \pi\|_{2,\pi}^2 = \sum_{i=1}^{|G|-1} \sigma_i(\mathbb{P})^{4s} \tag{4.16}$$

Hence the worst case $(2s)$-round L^2-distance for $\mathcal{P}$ is bounded above the worst case s-step L^2-distance for the Markov Chain $\mathbb{P}\overleftarrow{\mathbb{P}}$. Hence the result. ★

As an immediately corollary, we have

Corollary 4.20. *Let $\mathbb{P}$ be a Cayley walk on a group G.*

$$\max(\mathcal{T}_2(\mathbb{P}), \mathcal{T}_2(\mathbb{P}\overleftarrow{\mathbb{P}})) \leq \mathcal{R}_2^{\mathcal{C}}(\mathbb{P}) \leq \mathcal{R}_2^{\mathcal{H}}(\mathbb{P}) \leq 2\mathcal{T}_2(\mathbb{P}\overleftarrow{\mathbb{P}}) \tag{4.17}$$

In particular, if $\mathbb{P}$ is reversible, $\mathcal{T}_2(\mathbb{P}) \leq \mathcal{R}_2^{\mathcal{C}}(\mathbb{P}) \leq \mathcal{R}_2^{\mathcal{H}}(\mathbb{P}) \leq \mathcal{T}_2(\mathbb{P}) + 1$.

Proof. The adversarial strategies $\mathbb{A}_t = I$ and $\mathbb{A}_t = \overleftarrow{\mathbb{P}}$ show the lower bound on $\mathcal{R}_2^{\mathcal{C}}(\mathbb{P})$. Clearly $\mathcal{R}_2^{\mathcal{C}}(\mathbb{P}) \leq \mathcal{R}_2^{\mathcal{H}}(\mathbb{P})$. Theorem 4.19 shows $\mathcal{R}_2^{\mathcal{H}}(\mathbb{P}) \leq 2\mathcal{T}_2(\mathbb{P}\overleftarrow{\mathbb{P}})$.

If $\mathbb{P}$ were reversible, we have $\mathcal{R}_2^{\mathcal{H}}(\mathbb{P}) \leq 2\mathcal{T}_2(\mathbb{P}^2) \leq \mathcal{T}_2(\mathbb{P}) + 1$, since running a chain twice as fast mixes in half the time. ★

An immediate consequence is that a convex combination of reversible Cayley walk cannot mix slower than the faster of the two.

Theorem 4.21. *Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two reversible ergodic Cayley walks on a group G and put $\mathbb{Q} = a_1\mathbb{P}_1 + a_2\mathbb{P}_2$ where $0 < a_1 = 1 - a_2 < 1$. Then assuming $\mathcal{T}_2(\mathbb{P}_i) \geq \log(|G|)/2$ for $i = 1, 2$ and $|G| \geq 16$, we have*

$$\mathcal{T}_2(\mathbb{Q}) \leq 1 + \min\left(\frac{7\mathcal{T}_2(\mathbb{P}_1)}{a_1}, \frac{7\mathcal{T}_2(\mathbb{P}_2)}{a_2}, \mathcal{T}_2(\mathbb{P}_1) + \mathcal{T}_2(\mathbb{P}_2)\right) \tag{4.18}$$

Proof. Since the $\mathbb{P}_i$ are reversible, Corollary 4.20 implies $\mathcal{R}_2^{\mathcal{H}}(\mathbb{P}_i) \leq \mathcal{T}_2(\mathbb{P}_i) + 1$. Applying Proposition 3.20 and Lemma 3.16 for $\mathcal{S} = \mathcal{H}$, we have

$$\mathcal{R}_2^{\mathcal{H}}(\mathbb{Q}) \leq \min\left(\mathcal{R}_2^{\mathcal{H}}(\mathbb{P}_1) + \mathcal{R}_2^{\mathcal{H}}(\mathbb{P}_2) - 1, 7\mathcal{R}_2^{\mathcal{H}}(\mathbb{P}_1)/p, 7\mathcal{R}_2^{\mathcal{H}}(\mathbb{P}_2)/q\right) \tag{4.19}$$

□

4.4 Application: Non-Markovian Processes

We now illustrate the power of Robust mixing by estimating the mixing time of a non-Markovian process.

Definition 4.22. Let $\mathbb{P}_{RT}$ denote the *random-to-top transposition chain* on S_n which we define below: Given n cards, each arrangement can be identified with a permutation of S_n.

At each time t, pick $1 \leq i \leq n$, uniformly at random and exchange the cards at locations i and 1. Formally, it is the Cayley walk on S_n generated by uniform distribution on the transpositions $\{(1, i) : 1 \leq i \leq n\}$.

Definition 4.23. Let $\mathbb{P}_{RC}$ denote the *random-to-cyclic transposition process* on S_n which we define below: Given n cards, each arrangement can be identified with a permutation of S_n.

At each time t, pick $1 \leq i \leq n$, uniformly at random and exchange the cards at locations i and $(t \mod n)$.

In the random-to-cyclic process at time t, we exchange the card at location t with a random card. Due to this dependence on t, the process is not Markovian. The problem of estimating the mixing time of $\mathbb{P}_{RC}$ was raised by Aldous and Diaconis [2] in 1986. Mironov [42] used this process to analyze a cryptographic system known as RC4 and showed that $\mathbb{P}_{RC}$ mixes in time $O(n \log n)$ without an estimate on the hidden constant. Mossel et al. [47] showed $\mathcal{T}(\mathbb{P}_{RC}) = \Theta(n \log n)$. They showed a lower bound of $(0.0345 + o(1))n \log n$.

For the upper bound, they considered the following generalization of $\mathbb{P}_{RC}$: At time t, we exchange the card at location L_t with a random card, where the sequence L_t is chosen by an oblivious adversary. For this chain $\mathbb{P}_{RA}$ (random-to-adversarial), they showed that $\mathcal{T}(\mathbb{P}_{RA}) \leq Cn \log n + O(n)$ giving the first explicit bound of $C \approx 4 \times 10^5$. They also observed that since $\mathbb{P}_{RA}$ can simulate $\mathbb{P}_{RT}$, $\mathcal{T}(\mathbb{P}_{RA}) \geq \mathcal{T}(\mathbb{P}_{RT})$. Since $\mathcal{T}(\mathbb{P}_{RT}) = n \log n + O(n)$ ([13, 55]) it follows $C \geq 1$.

We now reduce C to 1 by using the Robust mixing framework.

Theorem 4.24.

$$\mathcal{T}_2(\mathbb{P}_{RC}) \leq \mathcal{T}_2(\mathbb{P}_{RT}) + 1 \leq n \log n + O(n) \tag{4.20}$$

Proof. In fact, we prove the following chain of inequalities:

$$\mathcal{T}_2(\mathbb{P}_{RC}) \leq \mathcal{T}_2(\mathbb{P}_{RA}) \leq \mathcal{R}_2^{\mathcal{H}}(\mathbb{P}_{RT}) \leq \mathcal{T}_2(\mathbb{P}_{RT}) + 1 \leq n \log n + O(n) \tag{4.21}$$

For a particular choice of adversarial moves the $\mathbb{P}_{RA}$ process can simulate the $\mathbb{P}_{RC}$ process. Hence $\mathcal{T}_2(\mathbb{P}_{RC}) \leq \mathcal{T}_2(\mathbb{P}_{RA})$.

By convexity arguments, it is enough to consider the case that the adversary's choice is deterministic to estimate the mixing time of $\mathbb{P}_{RA}$. Let $\alpha_t \in \{1, \ldots, n\}$ denote an adversarial choice for time t (fixed before the process begins). We first observe that an adversarial version of $\mathbb{P}_{RT}$ can simulate $\mathbb{P}_{RA}$. For $k, r \in \{1, \ldots, n\}, (k, r) = (1, k)(1, r)(1, k)$. Hence if we let $\mathbb{A}_t$ correspond to right multiplication by $(1, \alpha_t)(1, \alpha_{t+1})$, it follows that the given adversarial modification of $\mathbb{P}_{RT}$ simulates $\mathbb{P}_{RA}$. Since the simulation was done by a Cayley adversary, we have

$$\mathcal{T}_2(\mathbb{P}_{RA}) \leq \mathcal{R}_2^{\mathcal{C}}(\mathbb{P}_{RT}) \leq \mathcal{R}_2^{\mathcal{H}}(\mathbb{P}_{RT}) \tag{4.22}$$

Since $\mathbb{P}_{RT}$ is reversible, by Corollary 4.20 we have $\mathcal{R}_2^{\mathcal{H}}(\mathbb{P}_{RT}) \leq \mathcal{T}_2(\mathbb{P}_{RT}) + 1$. But $\mathcal{T}_2(\mathbb{P}_{RT;T}) \leq n \log n + O(n)$ ([13, 55]). □

The same reductionist approach can be used to estimate the mixing times of non-Markovian processes, if we can un-entangle the non-homogenous aspect of the evolution rule from the random choices made during the evolution. We give another example below.

Definition 4.25. Fix $1 \leq k \leq n$. *k-adjacent transposition chain* has the following evolution rule. At time t, we pick a random card r and exchange it with the card at location $(r + k \mod n)$.

For $k = 1$, this is the adjacent transpositions chain. If k is not relatively prime to n, then it is easy to see that the Markov Chain is reducible. Suppose k is relatively prime to n. Then we can re-order the cards so that they appear in the order $1, k + 1, 2k + 1, \ldots$. Since k and n are relatively prime, the set $\{(ki + 1 \mod n)\}_{i=0}^{n-1}$ is

a permutation of $\{1, \ldots, n\}$. Also after this reordering the k-adjacent transposition chain, reduces to the adjacent transposition chain. Hence for fixed k, the mixing time of adjacent transposition chain is the same as that of the k-adjacent transposition chain.

Now suppose k is not fixed and varies with time, i. e., at time t, we apply the k_t-adjacent transposition rule, where k_t is an adversarially specified sequence. Now if all k_t are relatively prime to n (which holds if n is prime), then the adversary can apply the k_t-reordering before and after the adjacent transposition step. Thus we have shown that the $\{k_t\}$-adjacent transposition process cannot mix slower than the adjacent transpositions Markov Chain. This non-Markovian shuffle is related to *Shell Sort*, just like adjacent transposition is related to *Bubble Sort*.

4.5 Application: Semi-direct Products

Let G and H be groups and suppose $G \leq \mathrm{Aut}(H)$. Then we can consider their semi-direct product $H \rtimes G$. In this section, we show how we can use Robust mixing to show that semi-direct products cannot mix slower than the direct product.

We restrict ourselves to semi-direct products of two groups so that we don't get bogged down by the notation.

Proposition 4.26. *Let μ_1 and μ_2 be distributions on $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively and $\mu = \mu_1\mu_2$ denote the product distribution on $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2$. Let ν be a distribution on $\mathcal{X}$ such that*

(a) the marginal distribution along $\mathcal{X}_1$ satisfies $\|\nu(\cdot, \mathcal{X}_2) - \mu_1\|_{2,\mu_1} \leq \epsilon_1$ and

(b) $\forall x \in \mathcal{X}_1$, the conditional distribution along $\mathcal{X}_2$ satisfies $\|\nu(x, \cdot) - \mu_2\|_{2,\mu_2} \leq \epsilon_2$.

Then $1 + \|\nu - \mu\|_{2,\mu}^2 \leq (1+\epsilon_1^2)(1+\epsilon_2^2)$

Proof. Let $f(x) = \nu(x, \mathcal{X}_2)/\mu_1(x)$ denote the density of the marginal with respect to μ_1 and for $x \in \mathcal{X}_1$, let $g_x(y) = \nu(x,y)/\nu(x,\mathcal{X}_2)\mu_2(y)$ be the density of the conditional of ν along $\mathcal{X}_2$ with respect to μ_2. Then for any $x \in \mathcal{X}_1$, we have

$$1 + \|\nu(\cdot, \mathcal{X}_2) - \mu_1\|_{2,\mu_1}^2 = \|f\|_{2,\mu_1}^2 = \mathbb{E}_{\mu_1}[f^2] \tag{4.23}$$

$$1 + \|\nu(x, \cdot) - \mu_2\|_{2,\mu_2}^2 = \|g_x\|_{2,\mu_2}^2 = \mathbb{E}_{\mu_2}[g_x^2] \tag{4.24}$$

$$1 + \|\nu - \mu\|_{2,\mu}^2 = \|h\|_{2,\mu}^2 = \mathbb{E}_{\mu}[h^2] \tag{4.25}$$

where $h(x,y) = \nu(x,y)/\mu(x,y)$ is the density function of ν.

$$\mathbb{E}_{\mu}[h^2] = \mathbb{E}_{\mu_1}[f^2\, \mathbb{E}_{\mu_2}[g_x^2]] \leq (1+\epsilon_2^2)\, \mathbb{E}_{\mu_1}[f^2] = (1+\epsilon_1^2)(1+\epsilon_2^2) \tag{4.26}$$

by the law of iterated conditional expectations. ★

Qualitatively, Proposition 4.26 implies that in order for the joint distribution to be close to stationarity, it is enough for one marginal and each conditional to be close enough to stationarity.

Proposition 4.27. *Let* G_1, G_2 *be groups and* $\varphi : G_2 \to \mathrm{Aut}(G_1)$ *be a homomorphism of groups. For* $i = 1, 2$*, let* $\mathbb{P}_i$ *be a Cayley walk on* G_i*.*

Put $G = G_1 \rtimes G_2$ *and consider the following Markov Chain* $\mathbb{P}$ *on* G*.*

1. *Let the current state of the Markov Chain be* $(\mathsf{g}_1, \mathsf{g}_2) \in G$
2. *Pick* $\mathsf{s}_1 \in G_1$ *according to* $\mathbb{P}_1$ *and* $\mathsf{s}_2 \in G_2$ *according to* $\mathbb{P}_2$
3. *With probability* p_1 *move to* $(\mathsf{g}_1\mathsf{s}_1, \mathsf{g}_2)$ *and with probability* $p_2 = 1 - p_1$ *move to* $(\varphi(\mathsf{s}_2)(\mathsf{g}_1), \mathsf{g}_2)$

Fix $\epsilon > 0$ and put $T_2 = \mathcal{T}_2(\mathbb{P}_2, \epsilon)$ and $T_1 = \mathcal{R}_2^{\mathcal{H}}(\mathbb{P}_1, \epsilon)$. Assume that $T_i \geq \log(|G_i|)$. Then

$$\mathcal{T}_2(\mathbb{P}, \delta) = O\left(\frac{T_1}{p_1} + \frac{T_2}{p_2}\right) \tag{4.27}$$

where $1 + \delta^2 = (1 + \epsilon^2)^2 + o(1)$ as $\min(|G_1|, |G_2|) \to \infty$.

Proof. Fix $t > 0$ and let μ_t be the distribution after t-steps. Let $\mathbf{T}$ denote the number of times G_2 was updated, so $\mathbb{E}[\mathbf{T}] = tp_2$.

Suppose $\mathbf{T} \geq T_2$. Then the marginal distribution of μ_t along G_2 is ϵ_2 close to uniform in L^2 distance. Now condition on the complete evolution of G_2 during the $\mathbf{T}$ steps. Due to this conditioning, we know exactly how the G_1 component evolves whenever G_2 is updated.

What remains is an adversarially modified version of $\mathbb{P}_1$ on G_1 with a holomorphic adversary, since G_2 acts on G_1 via automorphisms. Hence if $\mathbf{T} \geq T_2$ and $t - \mathbf{T} \geq T_1$, we have by Proposition 4.26, $\|\mu_t - \pi\|_{2,\pi} \leq \delta$.

Suppose $\mathbf{T} \geq T_2$ and $t - \mathbf{T} < T_1$. Then we can apply Proposition 4.26 with ϵ and $\sqrt{|G_1| - 1}$ (worst case L^2 distance). Similarly for $\mathbf{T} < T_2$ and $t - \mathbf{T} \geq T_1$.

Now take $t = \alpha(T_1/p_1 + T_2/p_2)$ for some $\alpha > 1$ to be determined later. This ensures that either $\mathbf{T} \geq T_2$ or $t - \mathbf{T} \geq T_1$. By Proposition 3.17,

$$\Pr\{\mathbf{T} < T_2\} \leq \exp\left(-(\alpha - 1 - \log\alpha)T_2\right) \leq |G_2|^{\alpha - 1 - \log\alpha} \tag{4.28}$$

$$\Pr\{t - \mathbf{T} < T_1\} \leq \exp\left(-(\alpha - 1 - \log\alpha)T_1\right) \leq |G_1|^{\alpha - 1 - \log\alpha} \tag{4.29}$$

Hence for $t = \alpha(T_1/p_1 + T_2/p_2)$ and $\beta = \alpha - 1 - \log\alpha$, we have

$$\begin{aligned} 1 + \|\mu_t - \pi\|_{2,\pi}^2 &\leq (1 + \epsilon^2)^2 + |G_2|^{-\beta}(1 + \epsilon^2)|G_2| + |G_1|^{-\beta}(1 + \epsilon^2)|G_1| \\ &\leq (1 + \epsilon^2)^2 + 2\min(|G_1|, |G_2|)^{1-\beta}(1 + \epsilon^2) \end{aligned} \tag{4.30}$$

Choose $\beta = 3$ so that $\|\mu_t - \pi\|_{2,\pi} = \epsilon + o(\epsilon)$ as $|G_1|, |G_2| \to \infty$. ★

Hence the mixing time of the semi-direct product is essentially the time it takes to update each component the right number of times.

- Robust Mixing time accounted for changes to G_1 caused by updates to G_2
- In case of a direct product, we may take $T_1 = \mathcal{T}_2(\mathbb{P}_1, \epsilon)$
- The result continues to hold for constant number of factors.
- If the number of factor is unbounded, a more careful analysis is required and will be considered in § 5.3.

Chapter 5

Markovian Products

In this chapter, we generalize Proposition 4.27 in various directions and show how to bound mixing times of product chains, in terms of that of the factor chains by only using a coupon collector type of analysis. Traditionally, this type of analysis has only been done to estimate the total variation mixing time of the product chains.

In § 5.2 we consider the case when the evolution of the components are independent. Just like in the semi-direct product case, one can consider the product of k chains, where updating component i can potentially affect all components $< i$. § 5.3 deals with the dependent case.

5.1 Distributions on Product Spaces

Fix $k > 0$ and for $i = 1, \ldots, k$, let μ_i be a distribution on $\mathcal{X}_i$. Consider $\mathcal{X} = \prod_{i=1}^{k} \mathcal{X}_k$ and let $\mu = \otimes_i \mu_i$ be the product distribution on $\mathcal{X}$. Let ν be any distribution on $\mathcal{X}$.

If each component of ν is close enough to μ_i and the components have enough independence then ν is close to μ.

We now give quantitative versions of the above statement for various distance

measures. Proposition 4.26 covered the L^2-distance for $k = 2$.

Definition 5.1. Let $\{\mu_i\}_{i=1}^k, \{\mathcal{X}_i\}_{i=1}^k, \nu$ be as above. For $1 \leq i \leq k$ and fixed $y_1, \ldots, y_i$, denote by $\nu^{(i)}$ the marginal distribution of $\nu(x_1, \ldots, x_k)$ along $\mathcal{X}_i$ conditioned on $x_j = y_j$ for $1 \leq j < i$.

Proposition 5.2. *Let $k > 0$ and $\nu, \{\mu_i\}, \{\mathcal{X}_i\}$ be as above. Fix $\epsilon_1, \ldots, \epsilon_k \geq 0$. Suppose that for all $1 \leq i \leq k$ and $y_1 \in \mathcal{X}_1, \ldots, y_{i-1} \in \mathcal{X}_{i-1}$, we have*

$$\|\nu^{(i)} - \mu_i\|_{2,\mu_i} \leq \epsilon_i \tag{5.1}$$

Then $1 + \|\nu - \mu\|_{2,\mu}^2 \leq \prod_{i=1}^k (1 + \epsilon_i^2)$ where $\mu = \otimes_i \mu_i$ is the product distribution on $\mathcal{X} = \prod_i \mathcal{X}_i$.

More over, if $\nu = \otimes_i \nu_i$ where ν_i is the marginal distribution along $\mathcal{X}_i$, then

$$1 + \|\nu - \mu\|_{2,\mu}^2 = \prod_i (1 + \|\nu_i - \mu_i\|_{2,\mu_i}^2) \tag{5.2}$$

Proof. Proposition 4.26 shows the inequality for $k = 2$. The general case follows by induction on k.

Now suppose the marginals ν_i are independent and let $f_i = \nu_i/\mu_i$ by the corresponding density function. Then

$$1 + \|\nu_i - \mu_i\|_{2,\mu_i}^2 = \mathbb{E}_{\mu_i}[f_i^2] \tag{5.3}$$

Independence implies $\mathbb{E}_\mu[\prod_i f_i^2] = \prod_i \mathbb{E}_{\mu_i}[f_i^2]$. Now the result follows since

$$1 + \|\nu - \mu\|_{2,\mu}^2 = \mathbb{E}_\mu \left[\left(\frac{\nu}{\mu} \right)^2 \right] = \mathbb{E}_\mu \left[\prod_i f_i^2 \right] \tag{5.4}$$

★

Proposition 5.3. *Assume $k \geq 2$. For $i = 1, \ldots, k$, let ν_i, μ_i be distributions on $\mathcal{X}_i$. Put $\nu = \otimes_i \nu_i$ and $\mu = \otimes_i \mu_i$. In order for $\|\nu - \mu\|_{2,\mu} \leq 0.7$, we must have $\|\nu_i - \mu_i\|_{2,\mu_i} \leq \frac{1}{\sqrt{k}}$ for at least $k/2$ values of i.*

Proof. If the conclusion does not hold, we have by Proposition 5.2,

$$1 + \|\nu - \mu\|_{2,\mu}^2 \geq (1 + 1/k)^{k/2} \geq 1.5 \tag{5.5}$$

since $(1 + 1/x)^x$ increases to e as $x \to \infty$. Hence $\|\nu - \mu\|_{2,\mu} \geq \sqrt{0.5} \geq 0.7$ □

For the L^∞ norm, we have

Proposition 5.4. *Let $k > 0$ and $\nu, \{\mu_i\}, \{\mathcal{X}_i\}$ be as above. Fix $\epsilon_1, \ldots, \epsilon_k \geq 0$. Suppose that for all $1 \leq i \leq k$ and $y_1 \in \mathcal{X}_1, \ldots, y_i \in \mathcal{X}_i$, we have*

$$\|\nu^{(i)} - \mu_i\|_{\infty,\mu_i} \leq \epsilon_i \tag{5.6}$$

Then $1 + \|\nu - \mu\|_{\infty,\mu} \leq \prod_{i=1}^k (1 + \epsilon_i)$ where $\mu = \otimes_i \mu_i$ is the product distribution on $\mathcal{X} = \prod_i \mathcal{X}_i$.

Proof. We only show for the case when the marginals are independent. Let $f_i = \nu_i / \mu_i$, where ν_i is the marginal distribution of ν along $\mathcal{X}_i$. Then

$$\|\nu^{(i)} - \mu_i\|_{\infty,\mu_i} = \max_{y_i \in \mathcal{X}_i} |f_i(y_i) - 1| \tag{5.7}$$

For real numbers a, b we have $|ab + a + b| \leq |a| + |b| + |ab|$. Substituting $a = \alpha - 1, b = \beta - 1$ and rearranging, we get

$$1 + |\alpha\beta - 1| \leq (1 + |\alpha - 1|)(1 + |\beta - 1|) \tag{5.8}$$

and by induction we have

$$1 + |\prod_i f_i - 1| \leq \prod_i (1 + |f_i - 1|) \tag{5.9}$$

Taking expectations on both sides and using the independence of f_i, we have the result. In the general case, we take the expectation of the right hand side, by taking expectation over each μ_i in turn in that order. ★

Proposition 5.5. *Let $k > 0$ and $\nu, \{\mu_i\}, \{\mathcal{X}_i\}$ be as above. Fix $\epsilon_1, \ldots, \epsilon_k \geq 0$. Suppose that for all $1 \leq i \leq k$ and $y_1 \in \mathcal{X}_1, \ldots, y_i \in \mathcal{X}_i$, we have*

$$\|\nu^{(i)} - \mu_i\|_{\mathrm{TV}} \leq \epsilon_i \tag{5.10}$$

Then $1 - \|\nu - \mu\|_{\mathrm{TV}} \geq \prod_{i=1}^{k}(1 - \epsilon_i)$ where $\mu = \otimes_i \mu_i$ is the product distribution on $\mathcal{X} = \prod_i \mathcal{X}_i$.

Proof. Let η_1, η_2 be two distributions on $\mathcal{Y}$ and ω be any distribution on $\mathcal{Y} \times \mathcal{Y}$ with marginals η_1 and η_2 respectively. Then from Lemma 2.12, we have

$$1 - \|\eta_1 - \eta_2\|_{\mathrm{TV}} = \max_{\omega} \Pr\{\omega(\Delta)\} \tag{5.11}$$

where Δ is the diagonal in $\mathcal{Y}^2$.

Suppose that ν is the product of its independent marginals. Then for each i there is an ω_i on $\mathcal{X}_i^2$ such that $\Pr\{\omega_i(\Delta_i)\} = 1 - \epsilon_i$, where Δ_i is the diagonal in $\mathcal{X}_i^2$. Define ω on $\mathcal{X} \times \mathcal{X} = \prod_i \mathcal{X}_i^2$ as $\otimes_i \omega_i$. Observe that the diagonal Δ of $\mathcal{X}^2$ is the intersection of Δ_i. Hence we have

$$1 - \|\nu - \mu\|_{\mathrm{TV}} \geq \Pr\{\omega(\Delta)\} = \prod_i \Pr\{\omega_i(\Delta_i)\} = \prod_i (1 - \epsilon_i) \tag{5.12}$$

In the general case, we have a family of couplings on $\mathcal{X}_i^2$ one for each value of $y_1, y_2, \ldots, y_{i-1}$. Define the coupling on $\mathcal{X}^2$ by an appropriate product. We illustrate it for $k = 2$. Let ω_1 be a coupling on $\mathcal{X}_1^2$ so that $\Pr\{\omega_1(\Delta_1)\} = 1 - \epsilon_1$. For each $y_1 \in \mathcal{X}_1$, we have $\|\nu^{(2)} - \mu_2\|_{\text{TV}} \leq \epsilon_2$. Hence we have a coupling ω_{2,y_1} on $\mathcal{X}_2^2$ such that $\Pr\{\omega_{2,y_1}(\Delta_2)\} \geq (1 - \epsilon_2)$ for every $y_1 \in \mathcal{X}_1$. Now define ω on $\mathcal{X}_1^2 \times \mathcal{X}_2^2$ as follows:

$$\omega(y_1, y_1', y_2, y_2') = \omega_1(y_1, y_1') \cdot \omega_{2,y_1}(y_2, y_2') \tag{5.13}$$

Now

$$\begin{aligned}
\Pr_{\omega}\{y_1 = y_1', y_2 = y_2'\} &= \Pr_{\omega}\{y_1 = y_1'\} \Pr_{\omega} y_2 = y_2' | y_1 = y_1' \\
&= \Pr_{\omega}\{y_1 = y_1'\} \omega_{2,y_1}(\Delta_2) \\
&\geq (1 - \epsilon_2) \Pr_{\omega}\{y_1 = y_1'\} \\
&\geq (1 - \epsilon_2)(1 - \epsilon_1) \qquad \bigstar
\end{aligned}$$

Even though, the total variation distance for a product space cannot be written in terms of the component wise total variation distances, one can still show a result like Proposition 5.3 for total variation distance.

Definition 5.6. Given two distributions ν and μ on $\mathcal{X}$, their *Hellinger distance* is defined via

$$H(\nu, \mu) = \|\sqrt{f}\|_{2,\mu} = \sqrt{2 - 2\left(\sum_x \sqrt{\nu(x)\mu(x)}\right)} \tag{5.14}$$

where $f(x) = \nu(x)/\mu(x)$.

The Hellinger distance is the $L^2(\mu)$ norm of the square root of the density function of ν with respect to μ and lies between 0 and $\sqrt{2}$. Hellinger distance is closely related to the total variation distance and splits nicely over product measures.

Proposition 5.7. *Let ν, μ be two measures on $\mathcal{X}$. Then*

(a) $\frac{H(\nu,\mu)^2}{2} \leq \|\nu - \mu\|_{\mathrm{TV}} \leq H(\nu, \mu)$

(b) For $k > 0$, if $\nu = \otimes_{i=1}^k \nu_i$ and $\mu = \otimes_{i=1}^k \mu_i$ where ν_i and μ_i are distributions on $\mathcal{X}_i$, we have

$$1 - \frac{H(\nu,\mu)^2}{2} = \prod_{i=1}^{k} \left(1 - \frac{H(\nu_i,\mu_i)^2}{2}\right) \tag{5.15}$$

Proof. (a) See [51, Chapter 3] for a proof.

(b) Observe from Definition 5.6 that

$$\begin{aligned} 1 - \frac{H(\nu,\mu)^2}{2} &= \sum_{\vec{\mathsf{x}}} \sqrt{\nu(\vec{\mathsf{x}})\mu(\vec{\mathsf{x}})} \\ &= \sum_{\vec{\mathsf{x}}} \prod_i \sqrt{\nu_i(x_i)\mu_i(x_i)} \\ &= \prod_i \left(1 - \frac{H(\nu_i,\mu_i)^2}{2}\right) \end{aligned} \tag{5.16}$$

where $\vec{\mathsf{x}} = (x_1, \ldots, x_k)$ ranges over $\prod_i \mathcal{X}_i$.

□

Proposition 5.8. *Fix $k \geq 2$. For $i = 1, \ldots, k$, let ν_i, μ_i be distributions on $\mathcal{X}_i$. Let $\nu = \otimes_i \nu_i$ and $\mu = \otimes_i \mu_i$. In order for $\|\nu - \mu\|_{\mathrm{TV}} \leq 0.3$, we must have $\|\nu_i - \mu_i\|_{\mathrm{TV}} \leq 1/\sqrt{k}$ for at least $k/4$ values of i.*

Proof. If not, $\|\nu_i - \mu_i\|_{\mathrm{TV}} \geq 1/\sqrt{k}$ for more than $3k/4$ values of i. Proposition 5.7 implies $H(\nu_i, \mu_i) \geq 1/\sqrt{k}$ for $3k/4$ values of i. Again by Proposition 5.7, we have

$$1 - \frac{H(\nu,\mu)^2}{2} \leq \left(1 - \frac{1}{2k}\right)^{3k/4} \tag{5.17}$$

Since $(1-1/x)^x$ increases monotonically to $1/\mathrm{e}$ as $x \to \infty$, we have $1 - H(\nu,\mu)^2/2 \leq (1/\mathrm{e})^{3/8}$. A final application of Proposition 5.7 gives,

$$\|\nu - \mu\|_{\mathrm{TV}} \geq \frac{H(\nu,\mu)^2}{2} \geq 1 - \left(\frac{1}{\mathrm{e}}\right)^{3/8} \geq 0.3 \tag{5.18}$$

□

For Relative Entropy we have

Proposition 5.9. *For $i = 1, \ldots, k$, let ν_i, μ_i be distributions on $\mathcal{X}_i$. Suppose that the ν_i are independent. Put $\nu = \otimes_i \nu_i$ and $\mu = \otimes_i \mu_i$. Then*

$$\mathrm{D}(\nu||\mu) = \sum_i \mathrm{D}(\nu_i||\mu_i) \tag{5.19}$$

Proof. Let $f_i = \nu_i/\mu_i$ be the density functions of ν_i w.r.t. μ_i. Then $\mathrm{D}(\nu_i||\mu_i) = \mathbb{E}_{\mu_i}[f_i \log(f_i)]$ and $\mathrm{D}(\nu||\mu) = \mathbb{E}_\mu[f \log f]$ where $f = \prod_i f_i$. Hence

$$\begin{aligned} \mathrm{D}(\nu||\mu) &= \mathbb{E}_\mu \left[f \left(\sum_i \log f_i \right) \right] \\ &= \sum_i \mathbb{E}_\mu \left[f \log f_i \right] \\ &= \sum_i \mathbb{E}_{\mu_i} [f_i \log f_i] \\ &= \sum_i \mathrm{D}(\nu_i||\mu_i) \end{aligned} \tag{5.20}$$

since $\mathbb{E}_{\mu_i}[f_i] = 1$ and the f_i are independent. ★

Note that unlike the other distances, we needed full independence here. As an immediate corollary of Proposition 5.9, we get

Proposition 5.10. *Fix $k \geq 2$. For $i = 1, \ldots, k$, let ν_i, μ_i be distributions on $\mathcal{X}_i$. Let*

$\nu = \otimes_i \nu_i$ and $\mu = \otimes_i \mu_i$. In order for $\mathrm{D}(\nu||\mu) \leq 1/2$, we must have $\mathrm{D}(\nu_i||\mu_i) \leq 1/k$ for at least $k/2$ values of i.

5.2 Markovian Product of Markov Chains

Definition 5.11. Fix $k > 0$ and let

- $\mathbb{P}_1, \ldots, \mathbb{P}_k$ be Markov Chains on $\mathcal{X}_1, \ldots, \mathcal{X}_k$ with stationary distributions $\mu_1, \ldots, \mu_k$,
- $\mathcal{X}' = \prod_i \mathcal{X}_i$, $\mu = \otimes_i \mu_i$,
- $\mathbb{Q}$ be a Markov Chain on $\mathcal{Y}$ with stationary distribution π,
- $F : \mathcal{Y} \times [0,1] \to 2^{[k]}$, where $[k] = \{1, \ldots, k\}$.
- $G : [0,1] \to \{0,1\}$.

Define a Markov Chain $\mathbb{P}$ on $\mathcal{X} = \mathcal{X}' \times \mathcal{Y}$ which evolves as follows.

(a) Let the current state be $(x_1, \ldots, x_k; y)$

(b) Pick $r \in [0,1]$ uniformly at random

(c) If $G(r) = 1$, choose y' from the distribution $\mathbb{Q}(y, \cdot)$ using r and any additional randomness. Otherwise, let $y' = y$.

(d) If $i \in F(y, r)$, choose x_i' from $\mathbb{P}_i(x_i, \cdot)$ without using r. Otherwise let $x_i' = x_i$.

(e) Move to the state $(x_1', \ldots, x_k'; y')$

- $\mathbb{P}$ is said to be an *independent Markovian product of* $\mathbb{P}_1, \ldots, \mathbb{P}_k$ *controlled by* $\mathbb{Q}$.
- $\mathbb{Q}$ is said to be the *controller*, F the *selector function* and G the *decider function*.

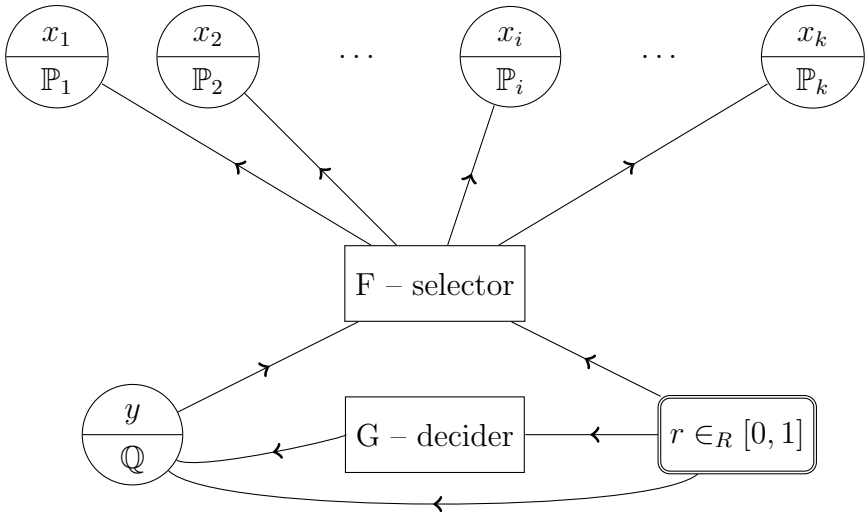


Figure 5.1: Markovian Product of Markov Chains

- By the *configuration part* we mean the projection of the state on $\mathcal{X}'$.
- By the *controller part* we mean the $\mathcal{Y}$ component of the state.

Some observations are in order

- Each application of $\mathbb{P}_i$ uses its own independent random choices.
- Applications of $\mathbb{Q}$ may reuse r, but the distribution of y' must be $\mathbb{Q}(y, \cdot)$.
- The random number r only influences $F(y, r)$, $G(r)$ and possibly y'.
- Since the evolution of the components are independent the stationary distribution is the product distribution.
- Varying F and G gives different chains on the same state space with the same stationary distribution.

We start with a few examples.

Example 5.12. (Direct product) Consider the following choices

- $\mathcal{Y} = \{1, \ldots, k\}$, π arbitrary
- $\mathbb{Q}(y_1, y_2) = \pi(y_2)$
- $G(r) = 1$
- $F(y, r) = \{y\}$

Now what we get is the direct product of $\mathbb{P}_1, \ldots, \mathbb{P}_k$ where at each time, we update component i with probability $\pi(i)$. Technically, we have a lag of one unit of time, since value of y decides which component gets updated during the next time period.

Example 5.13. (Lazy Lamplighter chains) Here we take

- $\mathcal{Y} = \{1, \ldots, k\}$
- $\mathbb{Q}$ any ergodic Markov Chain on $\mathcal{Y}$ with stationary distribution π
- $G(r) = 1$ iff $r \geq 1/2$
- $F(y, r) = \{y\}$ if $r < 1/2$, $\{\}$ otherwise
- $\mathcal{X}_i = \mathbb{Z}_2$ for $1 \leq i \leq k$
- $\mathbb{P}_i$ sets $x_i = 0/1$ with probability $1/2$ each

What we get now is a variant of the Lamplighter chain considered in [49]. Here we have a lamp lighter moving around $\mathcal{Y}$ with each vertex of $\mathcal{Y}$ having a two-state lamp.

A state of the chain, consists of the configuration of the lamps together with the position of the lamplighter. At each time step, the lamplighter moves with probability $1/2$ and randomizes the state of the current lamp with probability $1/2$.

By changing F and G suitably, we can also make it so that the lamplighter randomizes the current lamp *and* moves on $\mathcal{Y}$. This was the version considered in [49]

In both examples above r only influenced F and G. We now see an example where it influences y' as well.

Example 5.14. Complete Monomial Groups

- $\mathcal{Y} = S_k$ the symmetric group on k letters.
- $\mathbb{Q}$ is the random transposition chain.
- Interpret r as a pair $i, j \in \{1, \ldots, k\}$ where both i and j are uniform and independent
- $G((\mathsf{i}, \mathsf{j})) = 1$ always
- Let $\mathsf{y}' = \mathsf{y} \cdot (\mathsf{i}, \mathsf{j})$ where (i, j) is the transposition in S_k. Observe that y' has the distribution $\mathbb{Q}(\mathsf{y}, \cdot)$
- $\mathcal{X}_i = \{1, \ldots, m\}$ for some m for all $1 \leq i \leq k$
- $\mathbb{P}_i$ reaches its stationary distribution in one step
- $F(y, (\mathsf{i}, \mathsf{j})) = \{i, j\}$

Here we have a deck of k cards and on each card is a number between 1 and m. At each step, we pick two cards at random exchange their locations and randomize the number written on both of them. We use $\prod_i \mathcal{X}_i$ to represent the number on each card (identified by its position at time $t = 0$) and $\mathcal{Y}$ to represent the current ordering (relative to initial ordering). [56] gives tight estimates of the L^2 mixing times of this chain using Fourier analysis to estimate the eigenvalues of this chain.

We now show how to infer bounds on various mixing times of the Markovian product using just the knowledge of the mixing times of the factor chains and some in depth knowledge about the controller chain.

Definition 5.15. Consider a Markovian product of $\mathbb{P}_1, \ldots, \mathbb{P}_k$ with controller $\mathbb{Q}$ and selector and decider functions F and G. Let $\gamma : \{1, \ldots, k\} \to \mathbb{Z}^+$ be given.

For $t > 0$, consider a t-step run of the product chain and define the following random variables

$$\mathbf{N}_i(t) = \ \# \text{ of times component } i \text{ was updated till time } t \tag{5.21}$$

$$\mathbf{Z}_{t,\gamma} = |\{1 \leq i \leq k : \mathbf{N}_i(t) < \gamma(i)\}| \tag{5.22}$$

where we include time 0 and exclude time t. Thus $\mathbf{Z}_{t,\gamma}$ is the number of components i which have been updated $< \gamma(i)$ times. The function γ is called the *target* function.

When $\mathcal{Y} = \{1, \ldots, k\}$ and $F(y, r) = \{y\}$, $\mathbf{Z}_{t,\gamma}$ becomes the number of states of $\mathcal{Y}$ which have not been visited the appropriate number of times. Further specializing to $\mathbb{Q}(y_1, y_2) = 1/|\mathcal{Y}|$ and $\gamma(y) = 1$ gives the standard coupon collector problem.

Theorem 5.16. *Consider a Markovian product of chains $\mathbb{P}_1, \ldots, \mathbb{P}_k$ with controller $\mathbb{Q}$. Fix $\epsilon_1, \ldots, \epsilon_k \geq 0$. For $t > 0$, let ν_t be the distribution of the configuration (ignoring the controller component) at time t and $\mu = \otimes_i \mu_i$ be its stationary distribution.*

Finally put $M = \max_i 1/\mu_{i*}$ *where* $\mu_{i*} = \min_{x_i \in \mathcal{X}_i} \mu_i(x_i)$. *Then*

$$1 - \|\nu_t - \mu\|_{\mathrm{TV}} \geq \Pr\{\mathbf{Z}_{t,\gamma} = 0\} \cdot \left(\prod_i (1 - \epsilon_i) \right) \quad \textit{where } \gamma(i) \geq \mathcal{T}(\mathbb{P}_i, \epsilon_i) \tag{5.23}$$

$$\mathrm{D}(\nu_t || \mu) \leq \sum_i \epsilon_i + \mathbb{E}\left[\mathbf{Z}_{t,\gamma} \log(M)\right] \quad \textit{where } \gamma(i) \geq \mathcal{T}_{\mathrm{D}}(\mathbb{P}_i, \epsilon_i) \tag{5.24}$$

$$1 + \|\nu_t - \mu\|_{2,\mu}^2 \leq \mathbb{E}\left[M^{\mathbf{Z}_{t,\gamma}}\right] \cdot \left(\prod_i (1 + \epsilon_i^2) \right) \quad \textit{where } \gamma(i) \geq \mathcal{T}_2(\mathbb{P}_i, \epsilon_i) \tag{5.25}$$

$$1 + \|\nu_t - \mu\|_{\infty,\mu} \leq \mathbb{E}\left[M^{\mathbf{Z}_{t,\gamma}}\right] \cdot \left(\prod_i (1 + \epsilon_i) \right) \quad \textit{where } \gamma(i) \geq \mathcal{T}_\infty(\mathbb{P}_i, \epsilon_i) \tag{5.26}$$

Proof. Fix $t > 0$ and let $\nu_t = (\nu^{(1)}, \ldots, \nu^{(k)})$. From the definition of the product chain, it is clear that the $\nu^{(i)}$ are all independent. Let *dist* be one of total variation, relative entropy, L^2 or L^∞ distance measures and γ be the corresponding target function.

If $\mathbf{N}_i(t) \geq \gamma(i)$, then we know that $dist(\nu^{(i)}, \mu_i) \leq \epsilon_i$. If $\mathbf{N}_i(t) < \gamma(i)$, we can take $dist(\nu^{(i)}, \mu_i) \leq \alpha$, where α is worst value for the distance measure. Now condition on $\mathbf{Z}_{t,\gamma} = z$.

For total variation norm, we see by Proposition 5.5 that $1 - \|\nu_t - \mu\|_{\mathrm{TV}} \geq \prod_i (1 - \epsilon_i)$ if $z = 0$. When $z > 0$, we assume pessimistically that $\|\nu_t - \mu\|_{\mathrm{TV}} = 1$. This gives (5.23).

For relative entropy norm, the worst value of entropy distance for any component is $\log(M)$. Thus for a given value of z, we see that $\mathrm{D}(\nu^{(i)} || \mu_i) \leq \epsilon_i$ for all but z components for which it can be bounded by $\log(M)$. Thus for fixed z, Proposition 5.9 implies

$$\mathrm{D}(\nu_t || \mu) \leq \sum_i \epsilon_i + z \log(M) \tag{5.27}$$

which implies (5.24).

For the L^2 norm, the worst value of L^2 distance for any component is $\sqrt{M - 1}$.

For a given z, $1+\|\nu^{(i)}-\mu_i\|^2_{2,\mu_i} \leq 1+\epsilon_i^2$ for all but z components for which $1+\|\nu^{(j)}-\mu_j\|^2_{2,\mu_j} \leq M$. Thus for fixed z, Proposition 5.2 implies

$$1+\|\nu_t-\mu\|^2_{2,\mu} \leq M^z \prod_i (1+\epsilon_i^2) \tag{5.28}$$

which implies (5.25).

A similar analysis for L^∞ norm using the worst value of $M-1$ and applying Proposition 5.4 implies

$$1+\|\nu_t-\mu\|_{\infty,\mu} \leq M^z \prod_i (1+\epsilon_i) \tag{5.29}$$

which gives (5.26). ★

Thus in order for the configuration part of the Markovian product to mix it is enough for each component to updated the requisite number of times. Different distance measures impose different penalties if the component has not been updated as often as one would have liked. Pessimistically, we have taken the penalty to be the maximum possible distance.

A special mention should be made of the total variation bound. In many cases, one can obtain non-trivial estimates of $\|\nu_t-\mu\|_{\mathrm{TV}}$ even if $\mathbf{Z}_{t,\gamma} > 0$. However, for this one would need additional information about the chain $\mathbb{Q}$.

For example, suppose $\mathbb{Q}$ were the complete graph with self loops (so we have a coupon collector problem). Here if $\mathbf{Z}_{t,\gamma} = z > 0$, it is known that the actual set of vertices which have not been visited is equally likely to any set of z vertices not containing the initial vertex. This information can be used to get non-trivial estimates for $\|\nu_t-\mu\|_{\mathrm{TV}}$ when $z \leq \sqrt{k}$ and there by reduce the time needed to establish mixing.

In other words, our bound would be based on the time it takes to visit all vertices of the complete graph, where as it is enough to visit all but $\sqrt{k}$ vertices to show mixing. This usually saves a factor 2 in the mixing time.

We finish this section with a lower bound.

Theorem 5.17. *Consider a Markovian product of chains $\mathbb{P}_1, \ldots, \mathbb{P}_k$ with controller $\mathbb{Q}$. Fix $\epsilon \geq 0$ and for $t > 0$, let ν_t be the distribution of the configuration (ignoring the controller component) at time t and $\mu = \otimes_i$ be its stationary distribution. Then*

$$1 - \|\nu_t - \mu\|_{\mathrm{TV}} \geq \mathbb{E}\left[\left(1 - \frac{\epsilon^2}{2}\right)^{\mathbf{Z}_{t,\gamma}}\right] \qquad \text{where } \gamma(i) \leq \mathcal{T}(\mathbb{P}_i, \epsilon) \tag{5.30}$$

$$\mathrm{D}(\nu_t||\mu) \geq \mathbb{E}\left[\epsilon \mathbf{Z}_{t,\gamma}\right] \qquad \text{where } \gamma(i) \leq \mathcal{T}_{\mathrm{D}}(\mathbb{P}_i, \epsilon) \tag{5.31}$$

$$1 + \|\nu_t - \mu\|_{2,\mu}^2 \geq \mathbb{E}\left[(1+\epsilon^2)^{\mathbf{Z}_{t,\gamma}}\right] \qquad \text{where } \gamma(i) \leq \mathcal{T}_2(\mathbb{P}_i, \epsilon) \tag{5.32}$$

Proof. Let $\nu_t = (\nu^{(1)}, \ldots, \nu^{(k)})$. As in the upper bound proof, we condition on $\mathbf{Z}_{t,\gamma} = z$. If component i has not been updated $\gamma(i)$ times, then the distance along that component is at least ϵ. If component i has been updated $\geq \gamma(i)$ times, we assume pessimistically that the distance is already 0 along that component.

In case of entropy, Proposition 5.9 implies that if z components have not been updated the requisite number of times, $\mathrm{D}(\nu_t||\mu) \geq \epsilon z$. This gives (5.31).

In case of L^2-mixing, Proposition 5.2 implies that if z components have not been updated the requisite number of times, $1 + \|\nu_t - \mu\|_{2,\mu}^2 \geq (1+\epsilon^2)^z$. This gives (5.32).

In case of total variation distance, Proposition 5.7 implies that for each component i which has not been updated the requisite number of times, $1 - H(\nu^{(i)}, \mu_i)^2/2 \geq 1 - \epsilon^2/2$ and hence again by Proposition 5.7,

$$1 - \|\nu_t - \mu\|_{\mathrm{TV}} \geq 1 - \frac{H(\nu_t, \mu)^2}{2} \geq \left(1 - \frac{\epsilon^2}{2}\right)^z \tag{5.33}$$

Taking expectation over $\mathbf{Z}_{t,\gamma} = z$ we get (5.30). ★

5.3 Dependent Components

In this section we extend Theorem 5.16 to Markovian products of $\mathbb{P}_1, \ldots, \mathbb{P}_k$ controlled by $\mathbb{Q}$ where the evolution of $\mathbb{P}_i$ are not necessarily independent of each other.

Definition 5.18. Consider a Markovian product of $\mathbb{P}_1, \ldots, \mathbb{P}_k$ controlled by $\mathbb{Q}$ where the evolution of $\mathbb{P}_i$ are not necessarily independent of each other, i. e., an update of the i'th component may result in changes to other components. In such a case, the selector function F must also specify the order in which the components are to be updated, as this may influence the final state.

Recall that a state of the product chain is represented by $(x_1, \ldots, x_k; y)$ where $x_i \in \mathcal{X}_i$ and $y \in \mathcal{Y}$.

- The dependence is said to be *acyclic* if the evolution of each $\mathbb{P}_i$ is independent of all $\mathbb{P}_j$ for $j > i$.

- The dependence is said to be *compatible* if changes to x_i caused by an update of some other component is via a stochastic matrix on $\mathcal{X}_i$ which is compatible with the stationary distribution μ_i of $\mathbb{P}_i$.

- For $i = 1, \ldots, k$, let $\mathcal{S}_i$ be a valid set of strategies against $\mathbb{P}_i$ (see Definition 3.6). The dependence is said to be *compatible with* $(\mathcal{S}_1, \ldots, \mathcal{S}_k)$ if the dependence is compatible and for all i changes to x_i caused by update of some other component is via a stochastic matrix in $\mathcal{S}_i$.

- The Markovian product is said be $(\mathcal{S}_1, \ldots, \mathcal{S}_k)$-*dependent* if the dependence is acyclic and is compatible with $(\mathcal{S}_1, \ldots, \mathcal{S}_k)$.

For a $(\mathcal{S}_1, \ldots, \mathcal{S}_k)$-dependent Markovian product, the compatibility condition ensures that the stationary distribution is the product distribution.

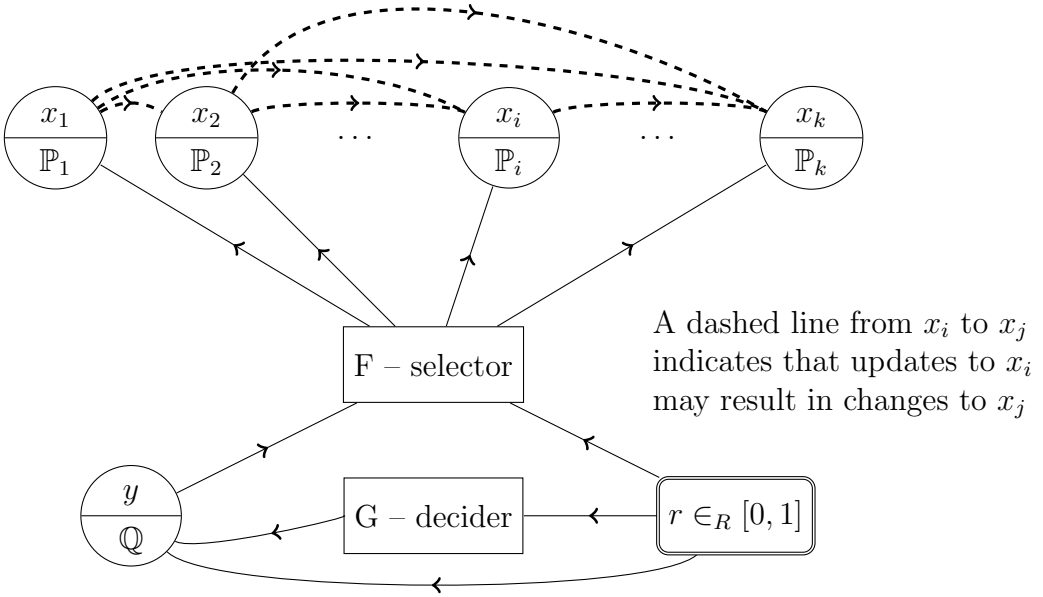


Figure 5.2: $(\mathcal{S}_1, \ldots, \mathcal{S}_k)$-dependent Markovian Product of Markov Chains

We consider an illuminating example.

Definition 5.19. Let G be a group. A chain of sub-groups $G = G_0 \geq G_1 \geq \cdots \geq G_k = \{1\}$ is said to be a *normal chain in* G if each G_i is normal in G.

Example 5.20. Let $G = G_0 \geq G_1 \geq \cdots \geq G_k = \{1\}$ be a normal chain in G and $S_i \subseteq G_i$ generate G_i/G_{i+1}. Now consider the following Markov Chain on G.

(a) Suppose the current state is g

(b) Pick $0 \leq i \leq k-1$ uniformly at random

(c) Pick $\mathsf{s} \in S_i$ uniformly at random

(d) Move to gs

Let $R_i \subseteq G_i$ be a set of coset-representatives of G_{i+1} in G_i, i. e., $R_i \cdot G_{i+1} = G_i$ and $|R_i| = |G_i|/|G_{i+1}|$. Then each element of G can be written uniquely in the form $\mathsf{r}_0\mathsf{r}_1 \dots \mathsf{r}_{k-1}$ where $\mathsf{r}_i \in R_i$.

We can look at this Markov Chain on G as a Markov Chain on the product $R_0 \times R_1 \times \cdots \times R_{k-1}$ with component chains being the Cayley walk induced by S_i on G_i/G_{i+1} and controller being the complete graph on k vertices.

If the current state is $\mathsf{g} = \mathsf{r}_0\mathsf{r}_1 \dots \mathsf{r}_{k-1}$ and we pick $\mathsf{s}_j \in S_j$. Then $\mathsf{gs}_j = \mathsf{r'}_0\mathsf{r'}_1 \dots \mathsf{r'}_{k-1}$ where

$$\mathsf{r'}_i = \begin{cases} \mathsf{r}_i & \text{if } i < j \\ \mathsf{r}_j\mathsf{s}_j & \text{if } i = j \\ \mathsf{s}_j^{-1}\mathsf{r}_i\mathsf{s}_j & \text{if } i > j \end{cases} \tag{5.34}$$

Thus an update of R_j can only cause higher components to change. Moreover in this example the change is caused by an automorphism of G_i/G_{i+1}. Hence this is an example of *acyclic dependency which is compatible with a holomorphic adversary.*

A special case of a *normal chain* is obtained by a sequence of iterated semi-direct products. For example, if G_2 acts on G_1 and G_3 acts on $G_1 \rtimes G_2$, we can consider the normal chain $(G_1 \rtimes G_2) \rtimes G_3 \geq G_1 \rtimes G_2 \geq G_1$. Thus the normal chain example is a generalization of the semi-direct product considered in Proposition 4.27

Theorem 5.21. *Consider a $(\mathcal{S}_1, \dots, \mathcal{S}_k)$-dependent Markovian product of Markov Chains $\mathbb{P}_1, \dots, \mathbb{P}_k$ with controller $\mathbb{Q}$. Fix $\epsilon_1, \dots, \epsilon_k \geq 0$. For $t > 0$, let ν_t be the distribution of the configuration part (ignoring the controller component) at time t and $\mu = \mu_1 \dots \mu_k$ be its stationary distribution. Finally put $M = \max_i 1/\mu_{i*}$ where*

$\mu_{i*} = \min_{x_i \in \mathcal{X}_i} \mu_i(x_i)$. *Then*

$$1 - \|\nu_t - \mu\|_{\mathrm{TV}} \geq \Pr\{\mathbf{Z}_{t,\gamma} = 0\} \cdot \left(\prod_i (1 - \epsilon_i) \right) \quad \textit{where } \gamma(i) \geq \mathcal{R}^{\mathcal{S}_i}(\mathbb{P}_i, \epsilon_i) \tag{5.35}$$

$$1 + \|\nu_t - \mu\|_{2,\mu}^2 \leq \mathbb{E}\left[M^{\mathbf{Z}_{t,\gamma}}\right] \cdot \left(\prod_i (1 + \epsilon_i^2) \right) \quad \textit{where } \gamma(i) \geq \mathcal{R}_2^{\mathcal{S}_i}(\mathbb{P}_i, \epsilon_i) \tag{5.36}$$

$$1 + \|\nu_t - \mu\|_{\infty,\mu} \leq \mathbb{E}\left[M^{\mathbf{Z}_{t,\gamma}}\right] \cdot \left(\prod_i (1 + \epsilon_i) \right) \quad \textit{where } \gamma(i) \geq \mathcal{R}_\infty^{\mathcal{S}_i}(\mathbb{P}_i, \epsilon_i) \tag{5.37}$$

Proof. We only prove for L^2-mixing time. The others follow similarly. Fix $t > 0$ and let ν denote the distribution of the configuration part at time t. Write $\nu = (\nu_1, \nu_2, \ldots, \nu_k)$ and note that ν_i are not necessarily independent.

Let $\mathbf{Y}_i$ denote the number of times component i has been updated (changes to a component as a result of an update of another component does not count), and let $\mathbf{Z}$ denote the number of components i for which $\mathbf{Y}_i < \gamma(i)$.

First condition on the value of $\mathbf{Z} = z$. Then condition on the choice of r and the evolution of the controller part up till time t. So now, we know which component was updated during which time step, but we do not know the random choices used during the component updates.

If $\mathbf{Y}_1 \geq \gamma(1)$, then we know $\|\nu_1 - \mu_1\|_{2,\mu_1} \leq \epsilon_1$. Otherwise, $\|\nu_1 - \mu_1\|_{2,\mu_1} \leq \sqrt{M-1}$. Here ν_1 is the marginal distribution of ν along $\mathcal{X}_1$. Note the acyclic dependency implies that the evolution of $\mathcal{X}_1$ component is independent of the others. Put $\delta_1 = \epsilon_1$ or $\sqrt{M-1}$ depending on whether $\mathbf{Y}_1 \geq \gamma(1)$ or not.

Now condition the distribution on all the random choices made during the time steps when $\mathcal{X}_1$ was updated. After this conditioning, the evolution of $\mathcal{X}_2$ is just an adversarially modified $\mathbb{P}_2$ since we have conditioned on the complete evolution of $\mathcal{X}_1$ and $\mathcal{X}_2$ is independent of the evolution of $\mathcal{X}_3$ and higher. Thus $\|\nu_2 - \mu_2\|_{2,\mu_2} \leq \delta_2$

where $\delta_2 = \epsilon_2$ or $\sqrt{M-1}$ depending on whether $\mathbf{Y}_2 \geq \gamma(2)$ or not. Note that here ν_2 is marginal distribution of ν along $\mathcal{X}_2$ conditioned on x_1.

Proceeding in this manner, conditioning on each component in turn, we have $\|\nu_i - \mu_i\|_{2,\mu_i} \leq \delta_i$ where $\delta_i = \epsilon_i$ or $\sqrt{M-1}$ depending on whether $\mathbf{Y}_i \geq \gamma(i)$ or not. Here ν_i is the marginal distribution of ν along $\mathcal{X}_i$ conditioned on $\nu_1, \ldots, \nu_{i-1}$.

Hence Proposition 5.2 implies

$$1 + \|\nu - \mu\|_{2,\mu}^2 \leq \prod_i (1 + \delta_i^2) \leq M^z \prod_i (1 + \epsilon_i^2) \tag{5.38}$$

Since this conclusion only depends on the value of $\mathbf{Z}$, we can now take expectation over the value of $\mathbf{Z}$ and all the other things we conditioned on and conclude

$$1 + \|\nu - \mu\|_{2,\mu}^2 \leq \mathbb{E}\left[M^{\mathbf{Z}}\right] \prod_i (1 + \epsilon_i^2) \tag{5.39}$$

★

We are unable to prove for $\mathcal{T}_{\mathrm{D}}(\cdot)$ because Proposition 5.9 requires complete independence. Note that in case multiple components are to be updated in a single time step, F specifies the order in which they are to be updated. Our proof does not depend on the order in which the components are updated.

5.4 Coupon Collector Revisited

In this section, we give estimates on various quantities related to the distribution of $\mathbf{Z}_{t,\gamma}$ in case $\mathbb{Q}(y_1, y_2) = \pi(y_2)$, i. e., $\mathbb{Q}$ reaches stationarity in one step. This is essentially the coupon collector problem where each coupon has a different probability of being chosen and we wish to collect $\gamma(i)$ coupons of type i.

We start with a concentration inequality for Multinomial distributions.

Theorem 5.22 (Coupon Collector Variant)**.** *Suppose we have n types of coupons and we pick coupons with repetition. The probability that we pick a coupon of type i is $\pi(i)$, where π is a distribution on $\{1,\dots,n\}$. Let $\gamma : \{1,\dots,n\} \to \mathbb{Z}^+$ be a target function, i.e. we would like to collect $\gamma(i)$ coupons of type i. For $t > 0$, let $\mathbf{Z}_{t,\gamma}$ denote the number of coupon types i for which we have $< \gamma(i)$ coupons after having collected t coupons.*

Let $\hat{\gamma} = \max_i \gamma(i)/\pi(i)$ and $\gamma_* = \min_i \gamma(i)$. Suppose that with probability $\geq 1 - 1/K$, we want to pick $\gamma(i)$ coupons of type i for all i, where $K > 1$ is a confidence parameter. Then it is enough to take $t = \hat{\gamma}(1 + \delta + o(1))$ where

$$\delta = \begin{cases} \frac{\log(Kn)}{\gamma_*} + \log\left(\frac{\log(Kn)}{\gamma_*}\right) - \log(\gamma_*) & \text{if } \gamma_* = o(\log(Kn)) \\ \delta' & \text{if } \gamma_* \sim C\log(Kn) \text{ and } \delta' - \log(1+\delta') = C \\ o(1) & \text{if } \gamma_* = \omega(\log(Kn)) \end{cases} \tag{5.40}$$

In particular, if $\gamma_* = \Omega(\log(Kn))$, $t = O(\hat{\gamma})$. Also for this choice of t, and any $\theta > 0$, we have

$$\mathbb{E}\left[\mathbf{Z}_{t,\gamma}\right] \leq 1/K \quad \text{and} \quad \mathbb{E}\left[\theta^{\mathbf{Z}_{t,\gamma}}\right] \leq \exp(\theta/K) \tag{5.41}$$

Proof. Fix $t = (1+\delta)\hat{\gamma}$ for some $\delta > 0$ to be determined later. Let $\mathbf{Y}_i$ be the number of coupons of type i collected, so that $\mathbb{E}[\mathbf{Y}_i] \geq (1+\delta)\gamma(i)$. Then by Proposition 3.17, we have

$$\Pr\{\mathbf{Y}_i < \gamma(i)\} \leq \exp\big(-\gamma(i)(\delta - \log(1+\delta))\big) \tag{5.42}$$

and hence

$$\Pr\{\mathbf{Z}_{t,\gamma} > 0\} \leq \mathbb{E}[\mathbf{Z}_{t,\gamma}] \leq n\exp\big(-\gamma_*(\delta - \log(1+\delta))\big) \tag{5.43}$$

Now we choose δ so that $\delta - \log(1+\delta) \geq \log(Kn)/\gamma_*$ to ensure $\Pr\{\mathbf{Z}_{t,\gamma} > 0\} \leq 1/K$.

For this choice of t, we have $\Pr\{A_i\} \leq (Kn)^{-1}$ where A_i is the event $\{\mathbf{Y}_i < \gamma(i)\}$. We now derive bounds on probabilities of z-wise intersections of the A_i.

Let $\mathcal{A} \subseteq \{1, \ldots, n\}$ be arbitrary and let $\pi(\mathcal{A}) = \sum_{i \in \mathcal{A}} \pi(i)$. Let $a_1, \ldots, a_n$ be arbitrary subject to the condition $a_i < \gamma(i)$ and put $a_{\mathcal{A}} = \sum_{i \in \mathcal{A}} a_i$ and let B be the event $\{(\forall i \in \mathcal{A})\mathbf{Y}_i = a_i\}$.

Finally fix $1 \leq j \leq n$ such that $j \notin \mathcal{A}$. We now show that

$$\Pr\{Y_j < \gamma(j)|B\} \leq \frac{1}{Kn} \tag{5.44}$$

To see that, observe that the distribution of $\mathbf{Y}_j$ conditional on B is binomial with parameters $t' = t - a_{\mathcal{A}}$ and $p' = \pi(j)/(1 - \pi(\mathcal{A}))$.

$$\mathbb{E}[\mathbf{Y}_j|B] = p't' = \frac{\pi(j)(t - a_{\mathcal{A}})}{1 - \pi(\mathcal{A})} \geq \frac{\pi(j)(t - t\pi(\mathcal{A}))}{1 - \pi(\mathcal{A})} \tag{5.45}$$

since $t \geq \hat{\gamma}$ implies $a_i \leq \gamma(i) \leq t\pi(i)$. Hence we have

$$\mathbb{E}[\mathbf{Y}_j|B] \geq \pi(j)t = \mathbb{E}[\mathbf{Y}_j] \geq (1 + \delta)\gamma(j) \tag{5.46}$$

Thus applying Proposition 3.17 to $(\mathbf{Y}_j|B)$ we get

$$\Pr\{\mathbf{Y}_j \leq \gamma(j)|B\} \leq \exp\Big(-\gamma(j)\big(\delta - \log(1 + \delta)\big)\Big) \leq \frac{1}{Kn} \tag{5.47}$$

by choice of δ. Hence we have

$$\Pr\{\mathbf{Y}_j \leq \gamma(j), B\} \leq \frac{\Pr\{B\}}{Kn} \tag{5.48}$$

By induction on the size of $\mathcal{A}$, and summing over all values of $a_i \leq \gamma(i)$, we have

for any $\mathcal{B} \subseteq \{1, \ldots, n\}$,

$$\Pr\{(\forall i \in \mathcal{B})\mathbf{Y}_i \leq \gamma(i)\} \leq (Kn)^{-|\mathcal{B}|} \tag{5.49}$$

Hence

$$\Pr\{\mathbf{Z}_{t,\gamma} \geq z\} \leq \binom{n}{z}\left(\frac{1}{Kn}\right)^z \leq \frac{1}{K^z z!} \tag{5.50}$$

Hence for $\theta > 0$, we have

$$\mathbb{E}\left[\theta^{\mathbf{Z}_{t,\gamma}}\right] \leq \sum_{z=0}^{n} \theta^z \Pr\{\mathbf{Z}_{t,\gamma} \geq z\} \leq \sum_{z=0}^{\infty} \frac{\theta^z}{K^z z!} = \exp(\theta/K) \tag{5.51}$$

★

5.5 Application: Complete Monomial Group

We now estimate the total variation and L^2-mixing times of walks on complete monomial groups. The complete monomial group is the semi-directed product $G^n \rtimes S_n$ where S_n acts on G^n by permutation. Here G is an arbitrary group. In the context of Markov Chains G could be any set with a Markov Chain on it.

Proposition 5.23. ***(Generalized hypercube)*** *Let π be any distribution on $\mathcal{X}$ and $\mathbb{P}'(x, y) = \pi(y)$ reach stationary distribution in one step. For $n > 0$, consider the chain $\mathbb{P}$ on $\mathcal{X}^n$ which picks a coordinate at random and randomizes it. Then*

$$\mathcal{T}(\mathbb{P}) \leq n \log n + O(n) \tag{5.52}$$

$$\mathcal{T}_{\mathrm{D}}(\mathbb{P}) \leq n \log n + n \log\log |\mathcal{X}| + O(n) \tag{5.53}$$

$$\mathcal{T}_2(\mathbb{P}) \leq n \log n + n \log |\mathcal{X}| + O(n) \tag{5.54}$$

Proof. In this case, we take the controller to be complete graph on n vertices with

self-loops. Since $\mathcal{T}(\mathbb{P}', 0) = 1$ and $\mathcal{T}_2(\mathbb{P}', 0) = 1$, we take $\gamma(i) = 1$ for all i. Fix $t > 0$ to be determined later and let $\mathbf{Z}$ be the number of components which have not been updated by time t.

For mixing in total variation norm, Theorem 5.16 implies it is enough to take t large enough so that

$$\Pr\{\mathbf{Z} = 0\}(1-0)^n = \Pr\{\mathbf{Z} = 0\} \tag{5.55}$$

is close to 1.

Let $K > 0$ be a confidence parameter which we will fix later and choose $t = n\log(Kn)(1+o(1))$ from Theorem 5.22 so that $\Pr\{\mathbf{Z} = 0\} \geq 1 - 1/K$. Thus in total variation norm the chain mixes in time $\leq n\log(Kn) \sim n\log n$ by taking $K = 4$ say.

For L^2 mixing time, we need t large enough so that $\mathbb{E}[|\mathcal{X}|^{\mathbf{Z}}] - 1$ is small. Again by Theorem 5.22 for $t \sim n\log(Kn)$ we have

$$1 + \|\nu_t - \mu\|^2_{2,\mu} \leq \exp(|\mathcal{X}|/K) \tag{5.56}$$

Thus in this case we can take $K = 3|\mathcal{X}|$, to get the L^2 distance down to $\exp(1/3) - 1 \leq 0.4$. So the L^2 mixing time we get is $n\log(3|\mathcal{X}|n) = n\log n + n\log(|\mathcal{X}|) + O(n)$.

Similarly for the entropy mixing time, we take $K = 3\log(|\mathcal{X}|)$, giving $\mathcal{T}_{\mathrm{D}}(\mathbb{P}) \leq n\log(n\log|\mathcal{X}|)$. ★

For total variation mixing, the correct answer is a factor of 2 smaller than our estimate. This is because in order to mix it is only necessary to visit all but $\sqrt{n}$ states while we took the time it takes to visit all states.

Proposition 5.24. *Let $\mathbb{P}$ denote the Markov Chain on the Complete Monomial Group*

considered in Example 5.14. Then

$$\mathcal{T}(\mathbb{P}) \leq n \log n + O(n) \tag{5.57}$$

$$\mathcal{T}_{\mathrm{D}}(\mathbb{P}) \leq n \log n + \frac{n}{2} \log \log |\mathcal{X}| + O(n) \tag{5.58}$$

$$\mathcal{T}_2(\mathbb{P}) \leq n \log n + \frac{n}{2} \log |\mathcal{X}| + O(n) \tag{5.59}$$

Proof. Here we have n identical copies of $\mathcal{X}$ controlled by the random transposition chain on S_n. If we just look at the configuration part and ignore the controller part, it evolves just like the generalized hypercube chain considered in Proposition 5.23 except at twice the speed, since each step of the Complete Monomial Group chain updates two components (not necessarily different).

After the configuration part is mixed, we can wait another $(n \log n)/2 + O(n)$ steps for the S_n component to mix. This is because the random transposition chain mixes in time $(n \log n)/2 + O(n)$ in all three measures.

After this time, the whole chain has mixed. Hence the results. ★

Using representation theory and explicit calculation of eigenvalues, [56] shows $\mathcal{T}(\mathbb{P}) \leq (n \log n)/2 + O(n)$ and $\mathcal{T}_2(\mathbb{P}) \leq (n \log n)/2 + \frac{n}{4} \log(|\mathcal{X}| - 1) + O(n)$, only a factor of 2 better than our bound. We can match the bound given in [56] if one shows the following:

- For the configuration part to mix, it is enough to visit update all but $\sqrt{n}$ components (for variation as well as L^2 mixing).
- In case of the complete monomial group, the mixing time of the whole chain is the maximum of the mixing times of the two parts.

When the Markov Chain on $\{1, \ldots, n\}$ does not reach the stationary distribution

in $O(1)$ steps, the techniques of [56] become too complex to handle, as the underlying representations would depend on the exact chain on $\{1, \ldots, n\}$. Our approach can handle those cases as well. In order to do that, we need the following

Proposition 5.25. *Let $\mathbb{P}$ be a Markov Chain on $\mathcal{X}$ with stationary distribution π. Fix $n > 0$ and consider the Markov Chain $\mathbb{Q} = \mathbb{P}^{\otimes n}$ on $\mathcal{X}^n$ where at each step we pick a random component and apply $\mathbb{P}$ to it. Let $\epsilon > 0$. Then*

$$\mathcal{T}(\mathbb{Q}) = O(n(T + \log(3n))) \qquad \text{where } T = \mathcal{T}\left(\mathbb{P}, \frac{\epsilon}{3n}\right) \tag{5.60}$$

$$\mathcal{T}_{\mathrm{D}}(\mathbb{Q}) = O(n(T + \log(3n \log|\mathcal{X}|))) \qquad \text{where } T = \mathcal{T}_{\mathrm{D}}\left(\mathbb{P}, \frac{\epsilon}{3n}\right) \tag{5.61}$$

$$\mathcal{T}_2(\mathbb{Q}) = O(n(T + \log(3n|\mathcal{X}|))) \qquad \text{where } T = \mathcal{T}_2\left(\mathbb{P}, \frac{\epsilon}{\sqrt{3n}}\right) \tag{5.62}$$

Proof. For $t > 0$, let ν_t be the distribution at time t and $\mu = \pi^{\otimes n}$ be the stationary distribution. We only prove for L^2 mixing time, others follow similarly.

Let $T = \mathcal{T}_2(\mathbb{P}, \epsilon/\sqrt{3n})$ and $\mathbf{Z}_t$ denote the number of components which have not been updated $T' = T + K$ times till time t. We take $K = \log(3n|\mathcal{X}|)$ to be the confidence parameter. Note that $T' = \Omega(\log(Kn))$.

Theorem 5.22 with confidence parameter $K = 3|\mathcal{X}|$ implies

$$\mathbb{E}\left[|\mathcal{X}|^{\mathbf{Z}_t}\right] \leq \exp(1/3) \tag{5.63}$$

for $t = O(nT')$. Now Theorem 5.16 with $t = O(nT')$ gives

$$1 + \|\nu_t - \mu\|_{2,\mu}^2 \leq \mathbb{E}\left[|\mathcal{X}|^{\mathbf{Z}_t}\right] \cdot \left(1 + \frac{\epsilon^2}{3n}\right)^n \sim \exp\left(\frac{1+\epsilon^2}{3}\right) \tag{5.64}$$

For the total variation and entropy distances we need to take the confidence parameters to be 3 and $3\log(|\mathcal{X}|)$ respectively. ★

Proposition 5.26. *Let $m, n > 0$ and let $\mathbb{Q}$ be the Markov Chain on $\mathbb{Z}_m^n$ which selects a component at random and performs a nearest neighbor random walk. Then $\mathcal{T}_2(\mathbb{Q}) = O(m^2 n \log n)$ and $\mathcal{T}(\mathbb{Q}) = \Omega(m^2 n \log n)$.*

Proof. **(Upper bound)** Apply Proposition 5.25 with $\mathbb{P} = \mathbb{Z}_m$ and use $\mathcal{T}_2(\mathbb{P}, \epsilon/\sqrt{3n}) = \Theta(m^2 \log n)$ for a small constant ϵ. Proposition 5.25 now implies $\mathcal{T}_2(\mathbb{Q}) = O(n(m^2 \log n) + \log(3nm)) = O(m^2 n \log n)$.

Note that the extra $\log n$ factor comes because of the exponential growth in the L^2 distance as a function of the number of components and not because of the coupon collector analysis as in the case for the hypercube.

(Lower bound) Proposition 5.8 implies that at least $n/4$ components must be $1/\sqrt{n}$ close to uniform in total variation distance for the product chain to be $1/4$ close to uniform. Since the spectral gap for the nearest neighbor random walk on $\mathbb{Z}_m$ is $\Theta(1/m^2)$, we need the $n/4$ components to be updated $\Omega(m^2 \log n)$ times each for that component to be $1/\sqrt{n}$ close to uniform. Hence $\mathcal{T}(\mathbb{Q}) = \Omega(m^2 n \log n)$. ★

[55, Theorem 8.10] proves $\mathcal{T}_2(\mathbb{Q}) = \Theta(m^2 n \log n)$ by going to continuous time and using all the eigenvalues of the product chain. As a corollary of Proposition 5.26, we can give bounds on mixing time of wreath product $\mathbb{Z}_m \wr S_n$ where we use the simple random walk on $\mathbb{Z}_m$.

Corollary 5.27. *Let $\mathbb{P}'$ be the simple random walk on $\mathbb{Z}_m$ and $\mathbb{Q}$ be the random transposition walk on S_n. Let $\mathbb{P}$ be the natural random walk on the wreath product $\mathbb{Z}_m \wr S_n$ induced by $\mathbb{P}'$ and $\mathbb{Q}$ where S_n acts on $\mathbb{Z}_m^n$ by coordinate permutation.*

$$\mathcal{T}_2(\mathbb{P}) = O(m^2 n \log n) \quad \text{and } \mathcal{T}(\mathbb{P}) = \Omega(m^2 n \log n) \tag{5.65}$$

Proof. Fix $\epsilon > 0$. Let $\mathbb{P}'^{\otimes n}$ denote the walk on $\mathbb{Z}_m^n$ considered in Proposition 5.26.

If we wait for time $\mathcal{T}_2(\mathbb{P}_n, \epsilon) + \mathcal{T}_2(\mathbb{Q}, \epsilon)$, by Proposition 5.2 the L^2 distance from stationarity is bounded by 2ϵ. Now the upper bound follows from Proposition 5.26 and the fact that $\mathcal{T}_2(\mathbb{Q}) = O(n \log n)$. The lower bound follows from the corresponding lower bound on $\mathcal{T}_1(\mathbb{P}_n)$. ★

5.6 Application: Sandpile Chains

We use the techniques of this chapter together with Robust mixing to get bounds on L^2 mixing time of the Sandpile Chain of some simple graphs. The Abelian Sandpile model is used to model Self-organized criticality, one motivating example being how adding grains of sand to an existing heap affects the structure of the heap. A closely related notion is that of a Chip Firing Game. See [9, 10, 37] for more information.

Definition 5.28. Let X be a directed graph with a sink s. Assume that every vertex has a path to the sink.

(a) A *configuration* is an assignment f of non-negative integers to each vertex except the sink. We think of it as the amount of sand at vertex v.

(b) A configuration f is said to be *stable* if for all vertices v, $f(v) < d^+(v)$, where $d^+(v)$ is the out-degree of v.

(c) If a configuration is unstable, it will *topple* as follows: Choose a v for which $f(v) \geq d^+(v)$. Reduce $f(v)$ by $d^+(v)$ and increase $f(w)$ by 1 for all out-neighbors w of v, i. e., v distributes $d^+(v)$ of its sand grains to each of its out-neighbors.

(d) A stable configuration f is said to be *reachable* from configuration g, if f can be obtained from g by adding some grains to some vertices and then toppling.

(e) A stable configuration is said to be *recurrent* if it is reachable from any configuration

Note that sand grains reaching the sink are lost for ever and is only way for the sand grains to leave the system. Every unstable configuration can be reduced to a stable configuration by a sequence of topplings, since every vertex v has a path to the sink s. Dhar [9] shows that the order in which the vertices are toppled is immaterial and hence the stable configuration produced by toppling an unstable configuration is unique.

One remarkable property of the recurrent configurations is that they form an abelian group.

Theorem 5.29 (Dhar [9])**.** *Let X be a directed graph with a sink s. Assume that every vertex has a directed path to the sink.*

- *The set of all recurrent configurations form an abelian group.*
- *The size of the group equals the number of spanning trees of X with root s.*

Since the groups are abelian we will write them additively and $\mathsf{0}$ would represent the identity of the group.

Definition 5.30. Let X be a directed graph with sink s. The *sandpile chain* $\mathrm{SP}(X)$ is defined as follows:

(a) The states of $\mathrm{SP}(X)$ are the recurrent configurations.

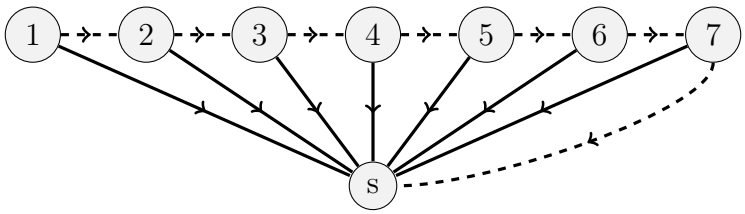
(b) Let f be the current configuration

(c) Pick a vertex v of X (except the sink) at random

(d) With probability 1/2 do nothing (ensures aperiodicity)

(e) With probability 1/2 move to g obtained by adding one grain to f at v, and toppling till it stabilizes.

We now estimate the mixing times of the sandpile chain of some simple graphs.



Solid edges have weight $d-1$. Dashed edges have unit weight.

Figure 5.3: d out-regular di-path with sink

Proposition 5.31. *Fix d, n. Let $P_{n,d}$ denote an n-vertex directed path together with a sink s. For every vertex $v \neq s$, add edges to s so that the out-degree of every vertex (except sink) is d. Then $\mathrm{SP}(P_{n,d})$ is the lazy Cayley walk on the group $\mathbb{Z}_{d^n}$ with generators $1, d, \ldots, d^{n-1}$.*

Proof. The set of all stable configurations has size d^n (grains at each vertex $< d$). Every configuration can be reached from every other configuration. This is seen by observing that the all zero configuration is reached from every configuration (keep adding sand grains and push all the grains to the right till they fall off), and every configuration is reachable from the all zero configuration (add appropriate number of sand grains at appropriate vertex so there is no toppling).

Hence all stable configurations are recurrent. For $i = 1, \ldots, n$, let e_i denote the generator corresponding to the vertex i. Then the toppling rules imply $d\mathsf{e}_i = \mathsf{e}_{i+1}$ for $i = 1, \ldots, n-1$ and $d\mathsf{e}_n = 0$. Hence the result. ★

Proposition 5.32. $\mathcal{T}(\mathrm{SP}(P_{n,d})) = \Omega(nd^2)$, $\mathcal{T}_2(\mathrm{SP}(P_{n,d})) = \Omega(nd^2 \log n)$

Proof. Since the walk is lazy, we see that the eigenvalues of $\mathrm{SP}(P_{n,d})$ are given by $(1+\lambda_\ell)/2$, where

$$\lambda_\ell = \frac{1}{n}\left(\exp\left(\frac{2\pi\iota\ell d^0}{N}\right) + \exp\left(\frac{2\pi\iota\ell d^1}{N}\right) + \cdots + \exp\left(\frac{2\pi\iota\ell d^{n-1}}{N}\right)\right) \tag{5.66}$$

where $N = d^n$ and $\iota = \sqrt{-1}$. Getting good estimates on $|\lambda_\ell|$ seems to be difficult. Note that $\lambda_0 = 1$ and $\lambda_\ell = \lambda_{d\ell}$. In particular, λ_1 has multiplicity n.

λ_1 is the average of n complex numbers on the unit circle at angles $\theta, d\theta, d^2\theta, \ldots, d^{n-1}\theta$ where $\theta = 6.28/N$, where 6.28 is "2*pi". The largest of these is $d^{n-1}\theta = 6.28/d$. Hence

$$1 - \left|\frac{1+\lambda_1}{2}\right| \le 1 - \mathrm{re}\left(\frac{1+\lambda_1}{2}\right) \le \frac{1-\cos(6.28/d)}{2n} \le \frac{(6.28)^2}{2nd^2} \tag{5.67}$$

By Proposition 2.29 we have $\mathcal{T}(\mathrm{SP}(P_{n,d})) = \Omega(nd^2)$. Since the multiplicity of λ_1 is n, Corollary 2.37 implies $\mathcal{T}_2(\mathrm{SP}(P_{n,d})) = \Omega(nd^2 \log n)$. ★

Theorem 5.33. $\mathcal{T}_2(\mathrm{SP}(P_{n,d})) = O(nd^2 \log n)$

Proof. Instead of looking at this chain as a walk on $\mathbb{Z}_{d^n}$, we look at it as a walk on $\mathbb{Z}_d^n$. Identify $\mathbb{Z}_{d^n}$ with the set of all n-digit d-ary numbers. $\mathrm{SP}(P_{n,d})$ picks a digit at random and increments it with probability half. When a digit overflows, the carry gets propagated to the next significant digit and the carry out of the most significant digit is lost.

Let $\mathbb{P}$ be the random walk on $\mathbb{Z}_d$ which at each step increments by one with probability 1/2 and does nothing with probability 1/2. Consider n copies of $\mathbb{P}$ and consider the Markov Chain where we pick a random component and update it using $\mathbb{P}$. This gives a random walk on $\mathbb{Z}_d^n$ but is not quite the same as $\mathrm{SP}(P_{n,d})$. Let us now modify the walk on $\mathbb{Z}_d^n$ so that whenever a component finds itself going from $d-1$ to 0, it updates the next significant digit by incrementing it (and if it over

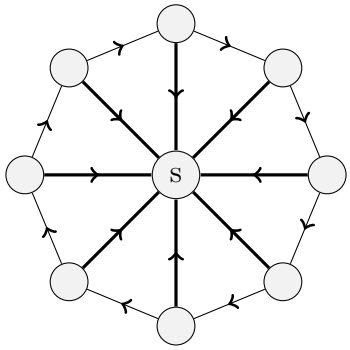
flows update the next one and so on). Since incrementing a digit can be done by a holomorphic adversary, what we have now is a *holomorphic dependent* Markovian product of n-copies of $\mathbb{P}$ with the complete graph on n-vertices as the controller. We do not have any cyclic dependencies since the overflow from the most significant digit is lost.

We now apply Theorem 5.21 to bound the mixing time. By Theorem 4.19 and Proposition 2.42 we have

$$\mathcal{R}_2^{\mathcal{H}}(\mathbb{P}) \leq 2\mathcal{T}_2(\mathbb{P}\overleftarrow{\mathbb{P}}) = O(\mathcal{T}_2((\mathbb{P} + \overleftarrow{\mathbb{P}})/2)) \tag{5.68}$$

Since $(\mathbb{P} + \overleftarrow{\mathbb{P}})/2$ is just the symmetric random walk on $\mathbb{Z}_d$ with some holding probability, we see that $\mathcal{R}_2^{\mathcal{H}}(\mathbb{P}) = O(d^2)$. Applying Proposition 5.25 just like in Proposition 5.26, we have $\mathcal{T}_2(\mathrm{SP}(P_{n,d})) = O(nd^2 \log n)$. ★

Note that there is a $O(\log n)$ gap between the lower and upper bounds for total variation mixing time of $\mathrm{SP}(P_{n,d})$.



Thin edges have unit weight.
Thick edges have weight $d - 1$.

Figure 5.4: d out-regular di-cycle with sink

Theorem 5.34. *Let $C_{n,d}$ be the n-vertex di-cycle shown in Figure 5.4 and* $\mathrm{SP}(C_{n,d})$ *be its sandpile chain. Then* $\mathcal{T}_1(\mathrm{SP}(C_{n,d})) = \Omega(nd^2)$ *and* $\mathcal{T}_2(\mathrm{SP}(C_{n,d})) = \Theta(nd^2 \log n)$

Proof. $\mathrm{SP}(C_{n,d})$ is very similar to $\mathrm{SP}(P_{n,d})$. Now, the all zero configuration is not recurrent so the state space now has $N := d^n - 1$ elements. Thus $\mathrm{SP}(C_{n,d})$ is the lazy Cayley walk on the group $\mathbb{Z}_{d^n-1}$ with generators, $1, d, d^2, \ldots, d^{n-1}$.

In terms of a n-digits d-ary number, the chain is like before except that the carry from the most significant digit comes back to the least significant digit.

Unlike $\mathrm{SP}(P_{n,d})$, the dependency among the different digits is cyclic. However we can still show the same upper bound by directly comparing the eigenvalues of $\mathrm{SP}(C_{n,d})$ with that of $\mathrm{SP}(P_{n,d})$ and using the mixing time bound for $\mathrm{SP}(P_{n,d})$. The lower bound arguments of $P_{n,d}$ apply here as well. ★

Theorem 5.35. *Let K_n denote the $(n-1)$-vertex undirected complete graph with self loops together with a sink s. Add a directed edge from every vertex to s, so K_n is n out-regular. Let* $\mathrm{SP}(K_n)$ *denote the sandpile chain on K_n. Then* $\mathcal{T}_2(\mathrm{SP}(K_n)) = \Theta(n^3 \log n)$ *and* $\mathcal{T}(\mathrm{SP}(K_n)) = \Omega(n^3)$.

Proof. For $i = 1, \ldots, n$, let e_i denote the group element corresponding addition of a grain at vertex i. The toppling rule at vertex i gives the relation

$$n \cdot \mathsf{e}_i = \mathsf{e}_1 + \cdots + \mathsf{e}_{n-1} \tag{5.69}$$

Adding these over various i, gives $\mathsf{e}_1 + \cdots + \mathsf{e}_{n-1} = 0$. Hence we have $n-2$ generators $\mathsf{e}_1, \ldots, \mathsf{e}_{n-2}$ each of order n. Theorem 5.29 implies that the number of recurrent configurations equals the number of spanning trees of K_n rooted at the sink, which is n^{n-2}. Hence there cannot be any more relations among $\mathsf{e}_1, \ldots, \mathsf{e}_{n-2}$.

Thus $\mathrm{SP}(K_n)$ is the lazy Cayley walk on $\mathbb{Z}_n^{n-2}$ with generators $\mathsf{e}_1, \ldots, \mathsf{e}_{n-2}$ and $\mathsf{e}_{n-1} = -(\mathsf{e}_1 + \cdots + \mathsf{e}_{n-2})$. Let $\mathbb{P}$ be the lazy Cayley walk on $\mathbb{Z}_n^{n-2}$ with generators $\mathsf{e}_1, \ldots, \mathsf{e}_{n-2}$, i. e., we ignore the last generator for now. Since the components now evolve independently and each component has mixing time $\Theta(n^2)$, Proposition 5.25 implies $\mathcal{T}_2(\mathbb{P}) = O(n^3 \log n)$. Now

$$\mathrm{SP}(K_n) = \left(\frac{n-2}{n-1}\right)\mathbb{P} + \left(\frac{1}{n-1}\right)\mathbb{Q} \tag{5.70}$$

where $\mathbb{Q}$ corresponds to the generator $-(\mathsf{e}_1 + \mathsf{e}_2 + \cdots + \mathsf{e}_{n-2})$. Since $\mathbb{Q}$ can be implemented by a Cayley adversary, Proposition 3.20 and Corollary 4.20 imply

$$\mathcal{T}_2(\mathrm{SP}(K_n)) \leq \mathcal{R}_2^{\mathcal{C}}(\mathrm{SP}(K_n)) = O(\mathcal{R}_2^{\mathcal{C}}(\mathbb{P})) = O(\mathcal{T}_2(\mathbb{P})) = O(n^3 \log n) \tag{5.71}$$

For the lower bound, we calculate the largest non-trivial eigenvalue of $\mathrm{SP}(K_n)$. Consider the vector

$$\vec{\alpha}(x_1, \ldots, x_{n-2}) = \eta^{x_1} \tag{5.72}$$

where $(x_1, \ldots, x_{n-2}) \in \mathbb{Z}_n^{n-2}$ indexes the coordinates of the vector $\vec{\alpha}$ and η is a primitive n'th root of unity. The effect of the generator e_1 is to multiply $\vec{\alpha}$ by η and the generators $\mathsf{e}_2, \ldots, \mathsf{e}_{n-2}$ do not move $\vec{\alpha}$. Hence the effect of $-(\mathsf{e}_1 + \cdots + \mathsf{e}_{n-2})$ is to multiply by $\overline{\eta} = \eta^{-1}$. Since $\mathrm{SP}(K_n)$ is a lazy walk, we see that $\vec{\alpha}$ is an eigenvector of $\mathrm{SP}(K_n)$ corresponding to the eigenvalue $(1 + \lambda_1)/2$, where

$$\lambda_1 = \frac{n-3}{n-1} + \frac{\eta}{n-1} + \frac{\overline{\eta}}{n-1} = 1 - \frac{2}{n-1}\left(1 - \cos\left(\frac{6.28}{n}\right)\right) \tag{5.73}$$

giving a spectral gap of $O(n^{-3})$. Also, by symmetry this eigenvalue has multiplicity $n-2$. Hence Corollary 2.30 and Corollary 2.37 gives $\mathcal{T}(\mathrm{SP}(K_n)) = \Omega(n^3)$ and

$\mathcal{T}_2(\mathrm{SP}(K_n)) = \Omega(n^3 \log n)$ respectively. ★

Chapter 6

Visiting States of a Markov Chain

So far, we have derived mixing time bounds for Markovian product chains (independent or otherwise) when the controlling Markov Chain reaches its stationary distribution in one step. In order to facilitate further applications, we derive some estimates of moment generating functions like those in Definition 5.15.

Notation: In this chapter, by $\mathbb{E}_y[\mathbf{X}]$ we mean the expected value of the random variable $\mathbf{X}$ when the Markov Chain under consideration has initial state y.

Definition 6.1. Let $\mathbb{P}$ be an ergodic Markov Chain on $\mathcal{X}$ with stationary distribution π. Let $\gamma : \mathcal{X} \to \mathbb{Z}^+$ be given. For $t > 0$, consider a t-step run of $\mathbb{P}$ and define the following random variables

$$\mathbf{N}_x(t) = \ \#\text{ of times } x \text{ was visited till time } t \tag{6.1}$$

$$\mathbf{Z}_{t,\gamma} = |\{x \in \mathcal{X} : \mathbf{N}_x(t) < \gamma(x)\}| \tag{6.2}$$

where we include time 0 and exclude time t. Thus $\mathbf{Z}_{t,\gamma}$ is the number of states in $\mathcal{X}$ which have been visited $< \gamma(x)$ times.

- When γ is clear from context, we drop the γ.
- When γ is the constant function C, by abuse of notation, we denote the random variable as $\mathbf{Z}_{t,C}$.
- We call γ the *target* function and define

$$\hat{\gamma} = \max_{x \in \mathcal{X}} \frac{\gamma(x)}{\pi(x)} \qquad \gamma^* = \max_x \gamma(x) \qquad \gamma_* = \min_x \gamma(x) \tag{6.3}$$

This is the same as Definition 5.15 if the components of a product chain are in one-to-one correspondence with the states of the controlling chain. Note that $\hat{\gamma}$ is the length of time one must run the chain starting from the stationary distribution, so that each state can be expected to be visited the right number of times.

6.1 Occupancy Measures

We recall a few concepts associated with occupancy measures of Markov Chains and derive bounds on t for which $\Pr\{\mathbf{Z}_{t,\gamma} > 0\}$ is small.

Definition 6.2. (Hitting time) Let $\mathbb{P}$ be an ergodic Markov Chain on $\mathcal{X}$ with stationary distribution π. For a state $x \in \mathcal{X}$, and $\ell > 0$, the ℓ-hitting time is defined via

$$\mathbf{H}_x^\ell = \min_t \{\mathbf{N}_x(t) \geq \ell\} \tag{6.4}$$

An application of Strong Law of Large Numbers, shows that $\mathbb{E}_x[\mathbf{H}_x^\ell] = \ell\, \mathbb{E}_x[\mathbf{H}_x^1] = \ell/\pi(x)$.

Definition 6.3. (Maximal Hitting time) Let $\mathbb{P}$ be an ergodic Markov Chain on

$\mathcal{X}$ with stationary distribution π. The maximal hitting time is defined as

$$H = \max_{x,y} \mathbb{E}_y[\mathbf{H}_x^1] \tag{6.5}$$

Definition 6.4. (Cover time) The *Cover time* of a Markov Chain $\mathbb{P}$ is the expected time to visit all the states at least once. Formally,

$$C = \max_{y \in \mathcal{X}} \mathbb{E}_y[\mathbf{C}] \quad \text{where} \quad \mathbf{C} = \max_x \mathbf{H}_x^1 \tag{6.6}$$

Definition 6.5. (Blanket time) Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π. For $0 < \delta < 1$, let

$$\mathbf{B}_\delta = \min_t \{(\forall x)\mathbf{N}_x(t) > \delta\pi(x)t\} \tag{6.7}$$

and the blanket time $B_\delta = \max_y \mathbb{E}_y[\mathbf{B}_\delta]$

Intuitively, the blanket time is amount of time, we need to run the chain, so that the observed frequency of visits is representative of the stationary distribution. The following inequalities connect these quantities

Theorem 6.6. *Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π.*

$$H \leq C \leq B_{1-\delta} = O\left(\frac{H \log |\mathcal{X}|}{\delta^2}\right) \tag{6.8}$$

Also $B_{1/\sqrt{2}} = O(C(\log\log|\mathcal{X}|)^2)$.

Proof. $H \leq C \leq B_{1-\delta}$ follows from definition. Winkler and Zuckerman [65] show $B_{1-\delta} = O\left(\frac{H \log |\mathcal{X}|}{\delta^2}\right)$. Kahn, Kim, Lovász, and Vu [33] show $B_\delta = O(C(\log\log|\mathcal{X}|)^2)$.

□

[65] conjectures that for constant δ, $B_\delta = O(C)$. For total variation bounds on mixing times of product chains, the quantity of interest is the time t for which $\Pr\{\mathbf{Z}_{t,\gamma} = 0\}$ is bounded away from 0. Note that this can be related to the blanket time via

Proposition 6.7. *Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π. Let $\gamma : \mathcal{X} \to \mathbb{Z}^+$ be a target function and $K \geq 2.5$ be a confidence parameter. Then $\Pr\{\mathbf{Z}_{t',\gamma} > 0\} \leq 1/4K$ for $t' = 4K \max(B_{1/2}, \hat{\gamma})$.*

Proof. Suppose $\hat{\gamma} \leq 4B_{1/2}$. By Markov's inequality we have for $t' = 4KB_{1/2}$

$$\Pr\{(\exists x), \mathbf{N}_x(t') \leq t'\pi(x)/2\} < 1/4K \tag{6.9}$$

$4KB_{1/2} \geq K\hat{\gamma}$ implies $t'\pi(x)/2 \geq \gamma(x)$ since $K \geq 2$. Hence $\Pr\{\mathbf{Z}_{t',\gamma} > 0\} \leq 1/4K$.

Now suppose $\hat{\gamma} = \alpha B_{1/2}$ for some $\alpha \geq 4$. Now divide $t' = 4K\hat{\gamma} = 4K\alpha B_{1/2}$ into $2K\alpha$ intervals of size $2B_{1/2}$ each. For each $1 \leq i \leq 2K\alpha$, let A_i denote the event that during interval i, every state x was visited at least $B_{1/2}\pi(x)$ times. By Markov's inequality, we have $\Pr\{A_i\} \geq 1/2$.

Let $\mathbf{Y}$ denote the number of events $\{A_i\}$ which happen. Since $\mathbf{Y}$ is the sum of $2K\alpha$ independent Bernoulli random variables with success probability $\geq 1/2$, Proposition 3.17 implies

$$\Pr\{\mathbf{Y} < \alpha\} \leq \exp(-\alpha(K - 1 - \log K)) \tag{6.10}$$

If $\mathbf{Y} \geq \alpha$, each state $x \in \mathcal{X}$ is visited at least $\alpha B_{1/2}\pi(x) = \hat{\gamma}\pi(x) \geq \gamma(x)$ times. Thus $\Pr\{\mathbf{Z}_{t',\gamma} > 0\} \leq \Pr\{\mathbf{Y} < \alpha\}$. Since $\alpha \geq 4$, (6.10) implies the result since for $K \geq 2.5$, we have $\exp(-4(K - 1 - \log K)) \leq 1/4K$. □

A useful special case is

Corollary 6.8. *Let $\mathbb{P}$ be an ergodic Markov Chain with uniform stationary distribution π and target function γ. Suppose that the maximal hitting time, $H = O(|\mathcal{X}|)$. Then for $t' = O(|\mathcal{X}|(\gamma^* + \log|\mathcal{X}|))$, we have $\Pr\{\mathbf{Z}_{t',\gamma} > 0\} < 1/4$.*

Proof. Assume without loss of generality that $\gamma^* = \Omega(\log|\mathcal{X}|)$. Otherwise take $\gamma' = \max(\gamma, \log|\mathcal{X}|)$.

Since $H = O(|\mathcal{X}|)$, Theorem 6.6 implies $B_{1/2} = O(|\mathcal{X}|\log|\mathcal{X}|)$. $\gamma^* = \Omega(\log|\mathcal{X}|)$ implies $\hat{\gamma} = \gamma^*|\mathcal{X}| = \Omega(B_{1/2})$. Now Proposition 6.7 gives the result. ★

In order to bound the expected value and moment generating function of $\mathbf{Z}_{t,\gamma}$, we need strong bounds on $\Pr\{\mathbf{Z}_{t,\gamma} = z\}$ for various values of z. We start with one such result, using the techniques of [65, Theorem 1].

Proposition 6.9. *Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π and γ a target function. Let b and c be positive integers. Then for any $x \in \mathcal{X}$, and any initial distribution of the Markov Chain,*

$$\Pr\left\{\mathbf{H}_x^{bc} \geq 15\left(\frac{bc}{\pi(x)} + bH\right)\right\} \leq 3^{-b} \tag{6.11}$$

where H is the maximal hitting time of the chain.

The strong law of large numbers implies that starting from any initial distribution, the expected time to visit x, bc-times is $\leq (bc)/\pi(x) + H$, where H is the maximal hitting time. The above result shows that if we wait for an additional bH time, the probability of not visiting x the right number of times decays exponentially in b.

Proof. Fix $x \in \mathcal{X}$, $b, c \geq 1$ and let $y \in \mathcal{X}$ be arbitrary.

$$\mathbb{E}_x[\mathbf{H}_x^c] = \frac{c}{\pi(x)} \quad \text{and} \quad \mathbb{E}_y[\mathbf{H}_x^c] \leq \frac{c-1}{\pi(x)} + H \tag{6.12}$$

Put $\alpha = (c-1)/\pi(x) + H$. Markov's inequality implies $\Pr_y\{\mathbf{H}_x^c \geq \mathrm{e}\alpha\} \leq 1/\mathrm{e}$. Since y was arbitrary, we have for integer $q \geq 1$,

$$\Pr_z\{\mathbf{H}_x^c \geq q\mathrm{e}\alpha\} \leq \mathrm{e}^{-q} \tag{6.13}$$

for any $z \in \mathcal{X}$. Fix z for the rest of the proof and put $\mathbf{W} = \mathbf{H}_x^c/(\mathrm{e}\alpha)$.

$$\begin{aligned} \mathbb{E}_z[\mathrm{e}^{\mathbf{W}/2}] &= \int_0^\infty \Pr_z\{\mathrm{e}^{\mathbf{W}/2} \geq r\} \\ &\leq 1 + \int_1^\infty \Pr_z\{\mathbf{W} \geq \lfloor 2\log r \rfloor\} \\ &\leq 1 + \int_1^\infty \mathrm{e} r^{-2} \leq 1 + \mathrm{e} \end{aligned} \tag{6.14}$$

Let $\mathbf{W}' = \mathbf{W}_1 + \cdots + \mathbf{W}_b$ be the sum of b independent copies of $\mathbf{W}$. For $p \geq 0$, we have

$$\Pr_z\{\mathbf{W}' \geq bp\} = \Pr_z\{\exp(\mathbf{W}'/2) \geq \exp(bp/2)\} \leq \frac{\mathbb{E}_z[\mathrm{e}^{\mathbf{W}'/2}]}{\mathrm{e}^{bp/2}} \leq \left(\frac{1+\mathrm{e}}{\mathrm{e}^{p/2}}\right)^b \tag{6.15}$$

Since the sum of b independent copies of $\mathbf{H}_x^c$ is $\mathbf{H}_x^{bc}$, taking $p = 5$ we have

$$\Pr_z\{\mathbf{H}_x^{bc} \geq 5\mathrm{e}b\alpha\} \leq 3^{-b} \tag{6.16}$$

Since z was arbitrary, we have the result. □

This gives good bounds on how large t must be to get $\mathbb{E}[Z_{t,\gamma}]$ small.

Proposition 6.10. *Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π and target function γ. Fix $\theta > 0$ and a confidence parameter $K > 1$. Suppose that*

$\gamma_* \geq \log(K\theta|\mathcal{X}|)$. *Then for* $t' = 15(\hat{\gamma} + \log(K\theta|\mathcal{X}|)H)$, *we have*

$$\mathbb{E}[\theta \mathbf{Z}_{t',\gamma}] \leq 1/K \tag{6.17}$$

where H *is the maximal hitting time of* $\mathbb{P}$.

Proof. Fix $x \in \mathcal{X}$ and apply Proposition 6.9 with $b = \log(K\theta|\mathcal{X}|)$ and $bc = \gamma(x)$ to get

$$\Pr\{A_x\} \leq \frac{1}{K\theta|\mathcal{X}|} \tag{6.18}$$

where A_x is the event $\{\mathbf{H}_x^{\gamma(x)} \geq t'\}$. Now $\mathbf{Z}_{t',\gamma}$ is just the sum of $|\mathcal{X}|$ random variables correspond to the events $\{A_x\}_{x\in\mathcal{X}}$. Hence we have

$$\mathbb{E}[\theta \mathbf{Z}_{t',\gamma}] = \sum_{x\in\mathcal{X}} \Pr\{A_x\} \leq \frac{1}{K} \tag{6.19}$$

□

As before if γ is not large enough we consider $\gamma'(x) = \max(\gamma(x), \log(K\theta|\mathcal{X}|))$. Specializing to a common case,

Corollary 6.11. *Let* $\mathbb{P}$ *be an ergodic Markov Chain with uniform stationary distribution* π *and target function* γ. *Suppose that the maximal hitting time* $H = O(|\mathcal{X}|)$. *Fix* $\theta > 1$ *and a confidence parameter* $K > 1$. *Suppose* $\gamma_* = \Omega(\log(K\theta|\mathcal{X}|))$. *Then for* $t' = O(\hat{\gamma})$, *we have* $\mathbb{E}[\theta \mathbf{Z}_{t',\gamma}] < 1/K$.

6.2 Moment Generating Functions

In this section we derive bounds on t, so that the moment generating function of $\mathbf{Z}_{t,\gamma}$ is close to 1. Unlike for the case of $\Pr\{\mathbf{Z}_{t,\gamma} > 0\}$ and $\mathbb{E}[\mathbf{Z}_{t,\gamma}]$, the picture here is not entirely satisfactory when it comes to $\mathbb{E}[\theta^{\mathbf{Z}_{t,\gamma}}]$.

When γ is huge, we show that Strong Law of Large numbers takes over and we can get tight results. For arbitrary γ, we give an upper bound on how large t must be for $\mathbb{E}[\theta^{\mathbf{Z}_{t,\gamma}}]$ to be close to 1. When $\gamma = 1$, [49] gives lower and upper bounds on how large t must be. By a careful analysis of their proof, we improve their upper bound to match the lower bound enabling us to give tight bounds on L^2-mixing times of Lamp Lighter chains. Implicit in [49, Theorem 1.4] is the following

Theorem 6.12. *Let $\mathbb{P}$ be a reversible ergodic Markov Chain with uniform stationary distribution π. Then $\mathbb{E}[2^{\mathbf{Z}_{t',1}}] \leq 2$ implies $t' = \Omega(|\mathcal{X}|(\mathcal{T}_{\text{rel}}(\mathbb{P}) + \log(|\mathcal{X}|)))$.*

We now consider the case when γ is huge.

Theorem 6.13. *Let $\mathbb{P}$ be a Markov Chain with stationary distribution π and γ a target function. Let $K > 1$ be a confidence parameter. If $\gamma_* \geq |\mathcal{X}| \log(K\theta)$, then for $t' = 15(\hat{\gamma} + |\mathcal{X}| \log(K\theta)H)$, we have*

$$\mathbb{E}[\theta^{\mathbf{Z}_{t',\gamma}}] \leq 1 + K^{-|\mathcal{X}|} \tag{6.20}$$

where H is the maximal hitting time.

Proof. Apply Proposition 6.9 with $bc = \gamma(x)$ and $b = |\mathcal{X}| \log(K\theta)$, to get

$$\Pr\left\{\mathbf{H}_x^{\gamma(x)} \geq 15\left(\frac{\gamma(x)}{\pi(x)} + |\mathcal{X}| \log(K\theta)H\right)\right\} \leq 3^{-|\mathcal{X}| \log(K\theta)} \leq (K\theta)^{-|\mathcal{X}|} \tag{6.21}$$

Hence for $t' = 15(\hat{\gamma} + |\mathcal{X}| \log(K\theta)H)$, we have

$$\mathbb{E}[\theta^{\mathbf{Z}_{t',\gamma}}] \leq 1 + \theta^{|\mathcal{X}|} \Pr\{\mathbf{Z}_{t',\gamma} > 0\} \leq 1 + K^{-|\mathcal{X}|} \tag{6.22}$$

★

Thus for fixed θ and large enough γ, it is enough to take $t = O(\hat{\gamma})$ for $\mathbb{E}[\theta^{\mathbf{Z}_{t,\gamma}}]$ to be close to 1.

Corollary 6.14. *Let $\mathbb{P}$ be an ergodic Markov Chain with uniform stationary distribution π with target function γ and maximal hitting time $H = O(|\mathcal{X}|)$. For $\theta > 1$ and a confidence parameter $K > 1$, suppose $\gamma_* = \Omega(|\mathcal{X}| \log(K\theta))$. Then for $t' = O(\hat{\gamma})$, we have $\mathbb{E}[\theta^{\mathbf{Z}_{t',\gamma}}] \leq 1 + K^{-|\mathcal{X}|}$.*

Theorem 6.13 is also useful for chains which mix slowly, where $H = \omega(|\mathcal{X}|)$.

Theorem 6.15. *Let $\mathbb{P}$ be an ergodic Markov Chain with stationary distribution π and target function γ. Let $s = \mathcal{T}_{\mathrm{f}}(\mathbb{P}, 1/2)$ be the filling time of $\mathbb{P}$. Fix $\theta > 1$ and $K > 1$ a confidence parameter. Assume that $\gamma_* \geq \log(K\theta|\mathcal{X}|)$. Then for $t' = O(\hat{\gamma}s)$, we have*

$$\mathbb{E}\left[\theta^{\mathbf{Z}_{t',\gamma}}\right] \leq \exp(1/K) + (K|\mathcal{X}|)^{-|\mathcal{X}|} \tag{6.23}$$

Proof. Let $\{\mathbf{X}_t\}_{t>0}$ be the states visited by the Markov Chain. We start by constructing a stopping time $\mathbf{T}$ for which $\mathbf{X}_{\mathbf{T}}$ has distribution exactly π.

(Stopping rule) Choose k from a geometric distribution with success probability $1/2$ and set $\mathbf{T} = ks$. since $s = \mathcal{T}_{\mathrm{f}}(\mathbb{P}, 1/2)$ it follows that $\mathbf{X}_{\mathbf{T}}$ has distribution exactly π. See [1] for a proof. Note that $\mathbb{E}[\mathbf{T}] = 2s$.

(Sub-sample) Now define $\{\mathbf{T}_i\}_{i=0}^{\infty}$, so that $\mathbf{T}_0 = 0$ and $\mathbf{T}_{i+1} - \mathbf{T}_i$ are independently distributed as $\mathbf{T}$. Finally put $\mathbf{Y}_i = \mathbf{X}_{\mathbf{T}_i}$.

By construction of $\mathbf{Y}_i$, it follows that $\mathbf{Y}_i$ are all independent and distributed as π, leading us back to the coupon collector case. Taking the confidence parameter as $K\theta$ in Theorem 5.22, we have for $t'' = C\hat{\gamma}$, where $C > 1$,

$$\mathbb{E}[\theta^{\mathbf{Z}'_{t'',\gamma}}] \leq \exp(1/K) \tag{6.24}$$

where $\mathbf{Z}'$ is the number of states which have not been visited the appropriate number of times for the chain $\{\mathbf{Y}_i\}$. Since $\mathbf{Y}_i$ is a sub-sample of $\mathbf{X}_t$, we have

$$\mathbb{E}[\theta^{\mathbf{Z}_{t,\gamma}} | \mathbf{T}_{t''} \leq t] \leq \exp(1/K) \tag{6.25}$$

It remains to show that for $t = O(t''s)$, $\Pr\{\mathbf{T}_{t''} > t\}$ is very small. Suppose we toss a fair coin every s steps. The number of heads we have seen is exactly the number of samples we have in the $\{\mathbf{Y}_i\}$ chain. Let $t = 2C't''s$ where $C' > 1$ to be determined later. Proposition 3.17 implies

$$\Pr\{\mathbf{T}_{t''} > t\} \leq \exp(-t''(C' - 1 - \log C')) \leq \exp(-\hat{\gamma}) \tag{6.26}$$

for $C' = 4$. Since $\hat{\gamma} \geq \gamma_*/(\pi_*) \geq |\mathcal{X}| \log(K\theta|\mathcal{X}|)$, (6.25) and (6.26) imply that for $t = 8t''s$, we have

$$\begin{aligned}
\mathbb{E}[\theta^{\mathbf{Z}_{t,\gamma}}] &\leq \mathbb{E}[\theta^{\mathbf{Z}_{t,\gamma}} | \mathbf{T}_{t''} \leq t] + \Pr\{\mathbf{T}_{t''} > t\}\theta^{|\mathcal{X}|} \\
&\leq \exp(1/K) + \left(\frac{1}{K\theta|\mathcal{X}|}\right)^{|\mathcal{X}|} \theta^{|\mathcal{X}|} \\
&\leq \exp(1/K) + \left(\frac{1}{K|\mathcal{X}|}\right)^{|\mathcal{X}|} \qquad \bigstar
\end{aligned}$$

If γ_* is not large enough, we can consider $\gamma'(x) = \max(\gamma(x), \log(K\theta|\mathcal{X}|)$. Note that this proof is quite wasteful, because it looks at only about $(1/s)$-fraction of the states actually visited. This bought us independence which allowed us to apply the coupon collector result. Also when $\mathbb{P}$ has uniform stationary distribution and $\gamma = 1$, Theorem 6.15 gives $t = O(|\mathcal{X}| \cdot \log(K\theta|\mathcal{X}|) \cdot \mathcal{T}(\mathbb{P}))$ where as the right answer is $t = O(|\mathcal{X}| \cdot (\log(K\theta|\mathcal{X}|) + \mathcal{T}_{\text{rel}}(\mathbb{P})))$.

6.3 Visiting All States Once

We now consider the case where

- $\gamma = 1$,
- π is uniform, and
- the maximal hitting time $H = O(|\mathcal{X}|)$

[49] showed that under these assumptions, it is enough to take $t = O(|\mathcal{X}|(\mathcal{T}(\mathbb{P}) + \log|\mathcal{X}|))$ to get $\mathbb{E}[2^{\mathbf{Z}_{t,1}}]$ close to 1. Note that under these assumptions, Theorem 6.12 shows we must take $t = \Omega(|\mathcal{X}|(\mathcal{T}_{\text{rel}}(\mathbb{P}) + \log|\mathcal{X}|))$. By a more careful analysis of the proof in [49], we improve the upper bound to match the lower bound.

We start with a claim which is implicit in the proof of [49, Lemma 5.1]. Since our definition of relaxation time does not match the one used in [49], the final expressions are slightly different. See the discussion after Definition 2.31 for the difference between the two definitions.

Lemma 6.16. *Let $\mathbb{P}$ be a Markov Chain with uniform stationary distribution π and f a real valued function on $\mathcal{X}$. Let $\mathbf{X}_t$ denote the state of the chain at time t. Then*

$$\text{Cov}_\pi(f(\mathbf{X}_1), f(\mathbf{X}_t)) \leq \sigma_1^t \,\text{Var}_\pi(f(X_1)) \tag{6.27}$$

where σ_1 is the second largest singular value of $\mathbb{P}$. In particular if $\mathbb{P}$ is reversible, $\sigma_1 = \lambda_*(\mathbb{P})$.

Proof. Let $g = \mathbb{P}^t f$, so that $\text{Cov}_\pi(f(\mathbf{X}_1), f(\mathbf{X}_t)) = \text{Cov}_\pi(f, g)$. and E denote the

expectation operator. Observe that $E\mathbb{P} = \mathbb{P}E = E$ implies $(\mathbb{P} - E)^t = \mathbb{P}^t - E$.

$$\begin{aligned}
\mathrm{Cov}_\pi(f, g) &= \mathbb{E}_\pi[(f - Ef)(g - Eg)] \\
&= \mathbb{E}_\pi[(f - Ef)(\mathbb{P}^t f - E\mathbb{P}^t f)] \\
&= \mathbb{E}_\pi[(f - Ef)((\mathbb{P} - E)^t f)] \\
&\leq \sqrt{\mathrm{Var}_\pi(f - Ef)}\sigma_1^t \sqrt{\mathrm{Var}_\pi(f)} \\
&= \sigma_1^t \,\mathrm{Var}_\pi(f)
\end{aligned} \tag{6.28}$$

since the operator norm of $\mathbb{P} - E$ is σ_1 and $\mathrm{Var}_\pi(f - c) = \mathrm{Var}_\pi(f)$ for all c. ★

The following lemma shows that if we have lots of unvisited states, we should be able to visit some of them relatively quickly, if the chain admits long excursions.

Lemma 6.17. *Let $\mathbb{P}$ be a reversible Markov Chain on $\mathcal{X}$. Let $\{\mathbf{X}_t\}$ denote the state of the Markov Chain at time t and assume that the initial distribution μ_0 of $\mathbf{X}_0$ satisfies $\mu_0 \geq \pi/2$. Let $\mathbf{T}^+(x)$ denote the return time to x and assume there are $\epsilon, \delta > 0$ for which*

$$\Pr_x(\mathbf{T}_x^+ \geq \epsilon|\mathcal{X}|) \geq \delta > 0 \tag{6.29}$$

for all $x \in \mathcal{X}$. Let $\mathcal{Y} \subset \mathcal{X}$ and assume $|\mathcal{Y}| \geq \mathcal{T}_{\mathrm{rel}} + 1$. Then the probability of hitting at least $\delta\epsilon|\mathcal{Y}|/4$ elements of $\mathcal{Y}$ by time $C\delta^{-2}|\mathcal{X}|(\mathcal{T}_{\mathrm{rel}} + 1)/|\mathcal{Y}|$ is at least $1/2$, where $\mathcal{T}_{\mathrm{rel}}$ is the relaxation time of $\mathbb{P}$ and $C \geq 16$ is an absolute constant.

Proof. Let $r = \epsilon|\mathcal{X}|(\mathcal{T}_{\mathrm{rel}} + 1)/|\mathcal{Y}|$. For $1 \leq i \leq r$, let $\mathbf{I}_i$ be an indicator random variable for the event $\{\mathbf{X}_i \in \mathcal{Y}\}$ and $\mathbf{J}_i$ for the event $\{\mathbf{X}_i \in \mathcal{Y}\}$ and $\{\mathbf{X}_j \neq \mathbf{X}_i\}$ for $i < j \leq r$. Finally let $\mathbf{J} = \sum_i \mathbf{J}_i$ and $\mathbf{I} = \sum_i \mathbf{I}_i$.

$\mathbf{J}$ is the number of distinct elements of $\mathcal{Y}$ which are visited in the time interval $[1, r]$. Also $\Pr\{\mathbf{J}_i = 1|\mathbf{I}_i = 1\} \geq \delta$ since $r \leq \epsilon|\mathcal{X}|$. This together with $\mathbb{E}[\mathbf{I}_i] \geq \pi(\mathcal{Y})/2$

gives

$$\mathbb{E}[\mathbf{J}] \geq \delta\,\mathbb{E}[\mathbf{I}] \geq \delta r\pi(\mathcal{Y})/2 = \frac{\delta\epsilon(\mathcal{T}_{\text{rel}}+1)}{2} \tag{6.30}$$

We first conclude $\Pr\{\mathbf{J} \geq \epsilon\delta(\mathcal{T}_{\text{rel}}+1)/4\}$ is bounded away from 0 by bounding $\mathbb{E}[\mathbf{I}^2]$. Fix $1 \leq i < j \leq r$. From Lemma 6.16 we have

$$\sum_{j>i} \operatorname{Cov}(\mathbf{I}_i, \mathbf{I}_j) \leq \sum_{j\geq i} \lambda_*^{j-i} \operatorname{Var}(\mathbf{I}_i) \leq \frac{\lambda_* \operatorname{Var}(\mathbf{I}_i)}{1-\lambda_*} = \mathcal{T}_{\text{rel}} \operatorname{Var}(\mathbf{I}_i) \leq \mathcal{T}_{\text{rel}}\,\mathbb{E}[\mathbf{I}_i] \tag{6.31}$$

since $\mathbf{I}_i$ is an indicator random variable. Now we have

$$\begin{aligned} \mathbb{E}[\mathbf{I}^2] &\leq \sum_i \operatorname{Var}(\mathbf{I}_i) + 2\sum_{j>i} \operatorname{Cov}(\mathbf{I}_i, \mathbf{I}_j) \\ &\leq \sum_i (2\mathcal{T}_{\text{rel}}+1)\,\mathbb{E}[\mathbf{I}_i] \\ &= (2\mathcal{T}_{\text{rel}}+1)\,\mathbb{E}[\mathbf{I}] \end{aligned} \tag{6.32}$$

Since $\mathbf{J}_i \leq \mathbf{I}_i$ we have

$$\mathbb{E}[\mathbf{J}^2] \leq \mathbb{E}[\mathbf{I}^2] \leq (2\mathcal{T}_{\text{rel}}+1)\,\mathbb{E}[\mathbf{I}] \leq (2\mathcal{T}_{\text{rel}}+1)\,\mathbb{E}[\mathbf{J}]/\delta \tag{6.33}$$

since $\mathbb{E}[\mathbf{J}_i] \geq \delta\,\mathbb{E}[\mathbf{I}_i]$. Hence using (6.30) we have

$$\mathbb{E}[J^2] \leq \frac{4}{\delta^2\epsilon}\,\mathbb{E}[J]^2 \tag{6.34}$$

Now let χ be the indicator for the event $\{\mathbf{J} \geq \mathbb{E}[\mathbf{J}]/2\}$. Then $\mathbb{E}[\mathbf{J}(1-\chi)] \leq \mathbb{E}[\mathbf{J}]/2$ and hence $\mathbb{E}[\mathbf{J}\chi] \geq \mathbb{E}[\mathbf{J}]/2$. Now by Cauchy Shwartz, we have $\mathbb{E}[\mathbf{J}\chi]^2 \leq \mathbb{E}[\mathbf{J}^2]\,\mathbb{E}[\chi^2]$. Hence

$$\Pr\{\mathbf{J} \geq \mathbb{E}[\mathbf{J}]/2\} = \mathbb{E}[\chi^2] \geq \frac{\mathbb{E}[\mathbf{J}\chi]^2}{\mathbb{E}[\mathbf{J}^2]} \geq \frac{(\mathbb{E}[\mathbf{J}]/2)^2}{\mathbb{E}[\mathbf{J}^2]} \geq \frac{\delta^2\epsilon}{16} \tag{6.35}$$

Thus in a trial of length $r = \epsilon|\mathcal{X}|(\mathcal{T}_{\text{rel}} + 1)/|\mathcal{Y}|$, the probability that we do not pick up $\delta\epsilon|\mathcal{Y}|/4$ elements of $\mathcal{Y}$ is less than $1 - \delta^2\epsilon/16$. Hence if we repeat this for $C\delta^{-2}/\epsilon$ intervals of length r each, we can reduce the probability of failure to less than $1/2$. Note that $C \geq 16$ here. □

Suppose that the initial distribution $\mu \geq \pi/2$. As long as the the current set of unvisited states $\mathcal{Y}$ is large ($\geq \mathcal{T}_{\text{rel}}+1$) we can apply Lemma 6.17 to show that we visit $\Omega(\mathcal{T}_{\text{rel}} + 1)$ new states within time $O(|\mathcal{X}|(\mathcal{T}_{\text{rel}} + 1)/|\mathcal{Y}|)$ with probability $\geq 1/2$. The next lemma establishes the assumption of Lemma 6.17 and handles the case when $\mathcal{Y}$ is small.

Lemma 6.18. *Let $\mathbb{P}$ be a Markov Chain with uniform stationary distribution π and maximal hitting time H. For $x \in \mathcal{X}$, let $\mathbf{T}_x^+$ denote the expected length of the return time to x.*

(a)

$$\min_x \Pr_x\Big\{\mathbf{T}_x^+ \geq \frac{|\mathcal{X}|}{2}\Big\} \geq \frac{|\mathcal{X}|}{2H} \tag{6.36}$$

where $\Pr_x\{\cdot\}$ refers to the probability of $\{\cdot\}$ if the initial state of the Markov Chain is x.

(b) For any $\mathcal{Y} \subseteq \mathcal{X}$, with probability $\geq 1/2$, we visit at least $|\mathcal{Y}|/2$ elements of $\mathcal{Y}$ by time $4H$.

Proof. (a) Since the stationary distribution is uniform, $\mathbb{E}_x[\mathbf{T}_x^+] = 1/\pi(x) = |\mathcal{X}|$. If after $|\mathcal{X}|/2$ steps we have not yet returned to x, and are currently at state y, then we expect to visit x within another H steps. Hence

$$|\mathcal{X}| = \mathbb{E}_x[\mathbf{T}_x^+] \leq \Pr\{\mathbf{T}_x^+ \leq |\mathcal{X}|/2\}|\mathcal{X}|/2 + \Pr\{\mathbf{T}_x^+ \geq |\mathcal{X}|/2\}H \tag{6.37}$$

Rearranging terms, we get the result.

(b) Fix $x \in \mathcal{Y}$ and let $\mathbf{H}_x$ denote the time when x is visited for the first time. $\mathbb{E}[\mathbf{H}_x] \leq H$. Thus by time $4H$ we visit x with probability $\geq 3/4$. Let $\mathbf{Y}$ denote the number of elements of $\mathcal{Y}$ which have been visited by time $4H$. Then $\mathbb{E}[\mathbf{Y}] \geq 3|\mathcal{Y}|/4$. For $q = \Pr\{\mathbf{Y} \geq |\mathcal{Y}|/2\}$, we have

$$3|\mathcal{Y}|/4 \leq \mathbb{E}[\mathbf{Y}] \leq (1-q)\frac{|\mathcal{Y}|}{2} + q|\mathcal{Y}| \tag{6.38}$$

Solving gives $q \geq 1/2$. □

Now for our sharpening of [49, Lemma 5.4].

Theorem 6.19. *Let $\mathbb{P}$ be a reversible Markov Chain with initial distribution $\mu \geq \pi/2$. Let $H \leq K|\mathcal{X}|$ for a constant $K \geq 1$ and $|\mathcal{X}| \geq 2$. Let $\theta \geq 2$ be arbitrary. There exists a universal constant c such that for all $a, b > 0$ we have*

$$\mathbb{E}[\theta^{\mathbf{Z}_t}] \leq 1 + \delta + \delta^2 + \delta^9 \tag{6.39}$$

where

- $\mathbf{Z}_t$ *is the number of unvisited states at time t,*
- $t \geq t' := cK^2|\mathcal{X}|\Big(2(1+a)(\mathcal{T}_{\text{rel}}+1)\log\theta + (1+b)\log|\mathcal{X}|\Big)$, *and*
- $\delta = \theta^{-(1+2a)(\mathcal{T}_{\text{rel}}+1)}|\mathcal{X}|^{-b}$

In particular when $a = 1/2$,$b = 0$, we have $\mathbb{E}[\theta^{\mathbf{Z}_t}] < 1.32$.

Proof. Let $r = \lfloor |\mathcal{X}|/(\mathcal{T}_{\text{rel}}+1) \rfloor$ and for $i = 0, 1, \ldots, r-1$, let $k_i = |\mathcal{X}| - i(\mathcal{T}_{\text{rel}}+1)$ and for $i = r$, $k_r = 0$. Let $\mathbf{Z}_t$ denote the number of unvisited states at time t and

define stopping times $\mathbf{T}_i$ via

$$\mathbf{T}_i = \min_t \{\mathbf{Z}_t = k_i\} \tag{6.40}$$

From Lemma 6.17 and Lemma 6.18 it follows that $\mathbf{T}_i - \mathbf{T}_{i-1}$ is stochastically dominated by $\mathbf{Y}_i = \kappa \mathbf{G}_i / k_i$ where $\mathbf{G}_i$ is geometric with mean 2 and $\kappa = C \cdot K^2 |\mathcal{X}| (\mathcal{T}_{\text{rel}} + 1)$ and $C \geq 16$ is a universal constant.

Fix $t > 0$, $i < r$ and $\beta > 0$ be arbitrary, Then

$$\begin{aligned} \Pr\{\mathbf{T}_i \geq t\} = \Pr\left\{\sum_{j=1}^{i} \mathbf{Y}_i \geq t\right\} &\leq \Pr\left\{\sum_{j=1}^{i} \frac{\kappa}{k_i} \mathbf{G}_i \geq t\right\} \\ &\leq \exp(-t\beta)\, \mathbb{E}\left[\sum_{j=1}^{i} \frac{\kappa\beta}{k_i} \mathbf{G}_i\right] \leq \exp(-t\beta) \prod_{j=1}^{i} \mathbb{E}\left[\frac{\kappa\beta}{k_i} \mathbf{G}_i\right] \end{aligned} \tag{6.41}$$

Choose $\beta = k_i / 3\kappa$ so that $\kappa\beta \leq k_j/3$ for all $j \leq i$. For $\alpha \leq 1/3$, $\mathbb{E}[\alpha \mathbf{G}_j] \leq \exp(3\alpha)$. This gives

$$\mathbb{E}\left[\frac{\kappa\beta}{k_j} \mathbf{G}_j\right] \leq \exp(k_i / k_j) \tag{6.42}$$

Hence we have

$$\Pr\{\mathbf{T}_i \geq t\} \leq \exp\left(-t\frac{k_i}{3\kappa} + \sum_{j=1}^{i} \frac{k_i}{k_j}\right) \leq \exp\left(-t\frac{k_i}{3\kappa} + \frac{k_i}{(\mathcal{T}_{\text{rel}} + 1)} \log |\mathcal{X}|\right) \tag{6.43}$$

For $i = r$, Lemma 6.18 implies $(\mathbf{T}_r - \mathbf{T}_{r-1})$ is stochastically dominated by the sum of $\ell = \log_2(2(\mathcal{T}_{\text{rel}} + 1))$ independent geometric random variables with mean $4K|\mathcal{X}|$ each. For $t > 0$, $\Pr\{\mathbf{T}_r - \mathbf{T}_{r-1} \geq t\}$ is the probability that after t independent coin tosses with success probability $(4K|\mathcal{X}|)^{-1}$, we have fewer than ℓ successes. Now we

can apply Proposition 3.17 to conclude

$$\Pr\{\mathbf{T}_r - \mathbf{T}_{r-1} \geq t\} \leq \exp\left(\ell - \frac{t}{4K|\mathcal{X}|} + \ell \log\left(\frac{t}{4K|\mathcal{X}|}\right)\right) \tag{6.44}$$

Breaking the values of $\mathbf{Z}_t$ into intervals of size $(\mathcal{T}_{\text{rel}} + 1)$ we have

$$\mathbb{E}[\theta^{\mathbf{Z}_t}] \leq 1 + \sum_{i=0}^{r} \theta^{k_i + (\mathcal{T}_{\text{rel}}+1)} \Pr\{\mathbf{Z}_t \geq k_i\} = 1 + \sum_{i=0}^{r} \theta^{k_i + (\mathcal{T}_{\text{rel}}+1)} \Pr\{\mathbf{T}_i \geq t\} \tag{6.45}$$

For $i < r$, $k_i \geq (\mathcal{T}_{\text{rel}} + 1)$ and hence

$$\begin{aligned}\theta^{k_i + (\mathcal{T}_{\text{rel}}+1)} \Pr\{\mathbf{T}_i \geq t\} &\leq \exp\left((k_i + (\mathcal{T}_{\text{rel}} + 1)) \log\theta - t\frac{k_i}{3\kappa} + \frac{k_i}{(\mathcal{T}_{\text{rel}} + 1)} \log|\mathcal{X}|\right) \\ &\leq \exp\left(2k_i \log\theta + \frac{k_i}{(\mathcal{T}_{\text{rel}} + 1)} \log|\mathcal{X}| - t\frac{k_i}{3\kappa}\right)\end{aligned} \tag{6.46}$$

When $i = r$, $0 = k_r \leq (\mathcal{T}_{\text{rel}} + 1) \leq k_{r-1}$ and hence

$$\begin{aligned}\theta^{k_r + (\mathcal{T}_{\text{rel}}+1)} \Pr\{\mathbf{T}_r \geq t\} &\leq \theta^{(\mathcal{T}_{\text{rel}}+1)} \big(\Pr\{\mathbf{T}_{r-1} \geq t/2\} + \Pr\{\mathbf{T}_r - \mathbf{T}_{r-1} \geq t/2\}\big) \\ &\leq \theta^{(\mathcal{T}_{\text{rel}}+1)} \exp\left(-\frac{t}{2}\frac{k_{r-1}}{3\kappa} + \frac{k_{r-1}}{(\mathcal{T}_{\text{rel}} + 1)} \log|\mathcal{X}|\right) \\ &\quad + \theta^{(\mathcal{T}_{\text{rel}}+1)} \exp\left(\ell - \frac{t}{8K|\mathcal{X}|} + \ell \log\left(\frac{t}{8K|\mathcal{X}|}\right)\right)\end{aligned} \tag{6.47}$$

where $\ell = \log_2(2(\mathcal{T}_{\text{rel}} + 1))$.

Let $t' = 6CK^2|\mathcal{X}|\big(2(1+a)(\mathcal{T}_{\text{rel}} + 1)\log\theta + (1+b)\log|\mathcal{X}|\big)$ for any $a, b > 0$ and hence take $c = 6C$. We now show that for $t \geq t'$, $\mathbb{E}[\theta^{|\mathbf{S}_{t'}|}] - 1$ is small.

Recall $\kappa = CK^2|\mathcal{X}|(\mathcal{T}_{\text{rel}}+1)$, hence

$$\frac{t'}{3\kappa} = 4(1+a)\log\theta + 2(1+b)\frac{\log|\mathcal{X}|}{(\mathcal{T}_{\text{rel}}+1)} \tag{6.48}$$

$$\frac{t'}{8K|\mathcal{X}|} = \frac{3CK}{4}\Big(2(1+a)(\mathcal{T}_{\text{rel}}+1)\log\theta + (1+b)\log|\mathcal{X}|\Big) \tag{6.49}$$

$$\ell' := \log\left(\frac{t'}{8K|\mathcal{X}|}\right) \geq \log\left(\frac{3(1+a)CK\log\theta\cdot 2(\mathcal{T}_{\text{rel}}+1)}{4}\right) \geq \log(6\cdot(2(\mathcal{T}_{\text{rel}}+1))) \tag{6.50}$$

since $C \geq 16, K \geq 1, \theta \geq 2$.

Put $\eta = \theta^{-(1+2a)(\mathcal{T}_{\text{rel}}+1)}$. For $t \geq t'$ and $i < r$, we have $k_i \geq (r-i)(\mathcal{T}_{\text{rel}}+1)$. This reduces (6.46) to

$$\begin{aligned}\theta^{k_i+(\mathcal{T}_{\text{rel}}+1)}\Pr\{\mathbf{T}_i \geq t'\} &\leq \exp\left(2k_i\log\theta + \frac{k_i}{(\mathcal{T}_{\text{rel}}+1)}\log|\mathcal{X}|\right.\\ &\qquad\left.-4(1+a)k_i\log\theta - 2(1+b)\frac{k_i}{(\mathcal{T}_{\text{rel}}+1)}\log|\mathcal{X}|\right)\\ &= \theta^{-(2+4a)k_i}|\mathcal{X}|^{-(1+2b)k_i/(\mathcal{T}_{\text{rel}}+1)}\\ &\leq \eta^{2(r-i)}|\mathcal{X}|^{-(1+2b)(r-i)}\end{aligned} \tag{6.51}$$

On the other hand, (6.47) reduces to

$$\begin{aligned}\theta^{(\mathcal{T}_{\text{rel}}+1)}\Pr\{\mathbf{T}_i \geq t'\} &\leq \theta^{(\mathcal{T}_{\text{rel}}+1)}\exp\Big(-2(1+a)k_{r-1}\log\theta\\ &\qquad-(1+b)\frac{k_{r-1}}{(\mathcal{T}_{\text{rel}}+1)}\log\mathcal{X} + \frac{k_{r-1}}{(\mathcal{T}_{\text{rel}}+1)}\log\mathcal{X}\Big)\\ &\quad+\theta^{(\mathcal{T}_{\text{rel}}+1)}\exp\left(\ell - \exp(\ell') + \ell\ell'\right)\\ &\leq \exp\left(-(1+2a)k_{r-1}\log\theta - b\frac{k_{r-1}}{(\mathcal{T}_{\text{rel}}+1)}\log|\mathcal{X}|\right)\\ &\quad+\exp\Big((\mathcal{T}_{\text{rel}}+1)\log\theta + (1+\ell')(\ell'-\log(6)) - \exp(\ell')\Big)\end{aligned} \tag{6.52}$$

using $\ell = \log_2(2(\mathcal{T}_{\text{rel}}+1))$ and $\ell' \geq \ell + \log(6)$. Since $f(z) = \exp(z)/4 - (1+z)(z - \log(6)) \geq 0$ for all $z \geq 0$, and $k_{r-1} \geq (\mathcal{T}_{\text{rel}}+1)$, we now have

$$\begin{aligned}
\theta^{(\mathcal{T}_{\text{rel}}+1)} \Pr\{\mathbf{T}_i \geq t'\} &\leq \eta|\mathcal{X}|^{-b} + \exp\Big((\mathcal{T}_{\text{rel}}+1)\log\theta \\
&\qquad - \frac{9CK}{16}\left(2(1+a)(\mathcal{T}_{\text{rel}}+1)\log\theta + (1+b)\log|\mathcal{X}|\right)\Big) \\
&\leq \eta|\mathcal{X}|^{-b} + \exp\Big(-9(1+2a)(\mathcal{T}_{\text{rel}}+1)\log\theta - 9(1+b)\log|\mathcal{X}|\Big) \\
&\leq \eta|\mathcal{X}|^{-b} + \eta^9|\mathcal{X}|^{-9b} \qquad (6.53)
\end{aligned}$$

using $C \geq 16$, $K \geq 1$. Combining (6.53) and (6.51) we have

$$\begin{aligned}
\mathbb{E}[\theta^{\mathbf{Z}_{t'}}] &\leq 1 + \left(\sum_{i=0}^{r-1} \eta^{2(r-i)}|\mathcal{X}|^{-(1+2b)(r-i)}\right) + \eta|\mathcal{X}|^{-b} + \eta^9|\mathcal{X}|^{-9b} \\
&\leq 1 + \frac{\eta^2|\mathcal{X}|^{-(1+2b)}}{1-\eta^2|\mathcal{X}|^{-(1+2b)}} + \eta|\mathcal{X}|^{-b} + \eta^9|\mathcal{X}|^{-9b} \qquad (6.54) \\
&\leq 1 + \eta^2|\mathcal{X}|^{-2b} + \eta|\mathcal{X}|^{-b} + \eta^9|\mathcal{X}|^{-9b}
\end{aligned}$$

since $\eta \leq 1, |\mathcal{X}| \geq 2$ implies $1 - \eta^2|\mathcal{X}|^{-(1+2b)} \geq 1/|\mathcal{X}|$. □

Our improvement over [49] is from Lemma 6.17. [49] ensured that the starting distribution $\mu_0 \geq \pi/2$ by running the chain for $O(\mathcal{T}(\mathbb{P}))$ steps in Lemma 6.17. Thus in each application of Lemma 6.17 they incurred an overhead of $O(\mathcal{T}(\mathbb{P}))$ steps. We assume $\mu_0 \geq \pi/2$, and ensure its validity by running the chain for $O(\mathcal{T}(\mathbb{P}))$ steps once and for all.

We now have four results giving bounds on how large t should be for the moment generating function of $\mathbf{Z}_{t,\gamma}$ to be close to 1.

- Theorem 6.13 gives optimal bounds when γ is huge,
- Theorem 6.19 gives optimal bounds when $\gamma = 1$ under the additional assump-

tion that the maximal hitting time $H = O(|\mathcal{X}|)$,

- Theorem 5.22 gives bounds when the chain mixes in one step,
- Theorem 6.15 doesn't make any assumptions, but the bounds are not always optimal.

6.4 Application: Lamp Lighter Chains

We now estimate mixing times of Lamp Lighter chains. We start by defining the Lamp Lighter chains, we consider

Definition 6.20. Let $\mathbb{Q}'$ and $\mathbb{Q}$ Markov Chains on $\mathcal{Y}'$ and $\mathcal{Y}$ respectively. Consider the following Markov Chain $\mathbb{P}$ on $\mathcal{Y}'^{\mathcal{Y}} \times \mathcal{Y}$:

(a) The states of the chain are of the form (f, y) where f is a function from $\mathcal{Y}$ to $\mathcal{Y}'$ and $y \in \mathcal{Y}$.

(b) At each time step, update y with probability $1/2$ according to $\mathbb{Q}$ and do not update f.

(c) With the remaining probability, update $f(y)$ (no where else) according to $\mathbb{Q}'$ and do not update y.

We call this *a wreath product* of $\mathbb{Q}'$ and $\mathbb{Q}$ and denote it $\mathbb{Q}' \wr \mathbb{Q}$.

Note that if $\mathbb{Q}'$ and $\mathbb{Q}$ are reversible, the product $\mathbb{Q}' \wr \mathbb{Q}$ is reversible. As shown in Example 5.13 this wreath product can be realized as a Markovian product of $|\mathcal{Y}|$ copies of $\mathbb{Q}'$ controlled by $\mathbb{Q}$ for an appropriate choice of the selector and decider functions.

When $\mathbb{Q}'$ is the natural chain on $\mathbb{Z}_2$ we recover the result in [49], but for a slightly different set of generators.

Proposition 6.21. *Let $\mathbb{Q}'$ be the chain on $\mathcal{Y}' := \mathbb{Z}_2$ which mixes in one step, and $\mathbb{Q}$ on $\mathcal{Y}$ be arbitrary reversible chain. Let $\mathbb{P} = \mathbb{Q}' \wr \mathbb{Q}$ and $\mathbf{Z}_t$ denote the number of states of $\mathcal{Y}$ which have not been visited by a random walk on $\mathcal{Y}$ of length t.*

(a) $\mathcal{T}(\mathbb{P}) = O\Big(C(\mathbb{Q}) + \mathcal{T}(\mathbb{Q})\Big)$

(b) $\mathcal{T}_2(\mathbb{P}) = O\Big(|\mathcal{Y}|(\mathcal{T}_{\text{rel}}(\mathbb{Q}) + \log|\mathcal{Y}|) + \mathcal{T}_2(\mathbb{Q})\Big)$

where $C(\mathbb{Q})$ and $\mathcal{T}_{\text{rel}}(\mathbb{Q})$ are the cover time and relaxation time of $\mathbb{Q}$ respectively.

Proof. Since $\mathcal{T}(\mathbb{Q}', 0) = \mathcal{T}_2(\mathbb{Q}', 0) = 1$, we take the target function γ to be the constant function 1. Let $\mathbf{Z}_t$ denote the number of states of $\mathcal{Y}$ which have not been visited by the controller (a.k.a lamp lighter).

(a) Apply Theorem 5.16 to conclude that configuration part is mixed by time $\inf_t \Pr\{\mathbf{Z}_t > 0\} \le 1/4$. But $\Pr\{\mathbf{Z}_t > 0\} \le 1/4$ for $t = 4C(\mathbb{Q})$. Thus by time $4C(\mathbb{Q})$ the configuration part has been randomized. Another $O(\mathcal{T}(\mathbb{Q}))$ steps ensures that the controller part is also randomized.

(b) Apply Theorem 5.16 to conclude that configuration part is mixed by time $\inf_t \mathbb{E}[2^{\mathbf{Z}_t}] < 1.25$. First run the chain for $O(\mathcal{T}(\mathbb{Q}))$ steps, so that the distribution of the controller's current position is at least $1/2$ its stationary distribution. Now Theorem 6.19 implies

$$\mathbb{E}[2^{\mathbf{Z}_t}] < 1.25 \tag{6.55}$$

for $t = O(\mathcal{T}(\mathbb{Q})) + O(|\mathcal{Y}|(\mathcal{T}_{\text{rel}}(\mathbb{Q}) + \log|\mathcal{Y}|))$. Now run the chain for additional $\mathcal{T}_2(\mathbb{Q})$ steps to randomize the lamplighter's position. ★

Instead of looking at two-state lamps, we can look at n-state lamps. For concreteness we consider the lamplighter (a.k.a controller) to be a d-dimensional torus for constant d and the lamp states to be the random walk on the circle.

Proposition 6.22. *Let $\mathbb{Q}'$ denote the simple random walk on $\mathbb{Z}_n$ and $\mathbb{Q}$ the simple random walk on $\mathbb{Z}_n^d$, where $d \geq 2$. Let $\mathbb{P} = \mathbb{Q}' \wr \mathbb{Q}$. Then*

(a) $\mathcal{T}(\mathbb{P}) = \Theta(n^{d+2} \log n)$

(b) $\mathcal{T}_2(\mathbb{P}) = O(n^{d+4} \log n)$

(c) $\mathcal{T}_2(\mathbb{P}) = O(n^{2d} \log n)$

Proof. We have n^d copies of $\mathbb{Z}_n$ controlled by a random walk on $\mathbb{Z}_n^d$. By Proposition 5.8, in order to mix in total variation distance, $\Omega(n^d)$ copies of $\mathbb{Z}_n$ must be $\Omega((d \log n)^{-1})$ close to stationarity. Since the relaxation time of $\mathbb{Z}_n$ is $\Theta(n^2)$, each of the $\Omega(n^d)$ copies must be updated $\Omega(n^2 \log n)$ times. Hence $\mathcal{T}(\mathbb{P}) = \Omega(n^{d+2} \log n)$.

On the other hand, if each state of $\mathbb{Z}_n^d$ is visited $O(n^2 \log n)$ times, with high probability Theorem 5.16 implies that the controller part of the state is mixed in total variation norm. Since the maximal hitting time for $\mathbb{Z}_n^d$ is $\Theta(n^d)$, Theorem 6.6 implies its blanket time is $O(n^d \log n)$ and now Proposition 6.7 implies that the probability that we have not visited all the states of $\mathbb{Z}_n^d$ by time $O(n^{d+2} \log n)$ is less than $1/10$. Now run the chain for an additional $O(n^d)$ time to randomize the controller part. Hence $\mathcal{T}(\mathbb{P}) = O(n^{d+2} \log n)$.

For the L^2-mixing bound: Theorem 5.16 implies we need to find t, for which $\mathbb{E}[n^{\mathbf{Z}_t}]$ is close to 1, where the target function γ is the constant $O(n^2 \log n)$. Applying Theorem 6.13 with the target function as $O(n^d \log n)$, gives $\mathbb{E}[n^{\mathbf{Z}_t}]$ is small for $t = O(n^{2d} \log n)$. On the other hand, applying Theorem 6.15 shows that $\mathbb{E}[n^{\mathbf{Z}_t}]$ is small

for $t = O(n^{d+2} \log n \cdot s)$, where $s = \mathcal{T}_{\mathrm{f}}(\mathbb{Z}_n^d)$. Since the random walk on $\mathbb{Z}_n^d$ is reversible, Proposition 2.50 implies $s = \Theta(n^2)$. ★

Thus for $d = 2$, we see that the variation mixing time and the L^2 mixing time are of the same order. Even though we have a $O(n^2)$ gap between the variation and L^2 mixing times for $d \geq 4$, we believe that the correct answer for the L^2-mixing times are $O(n^{d+2} \log n)$.

Wreath products $\mathbb{P} = \mathbb{Q}' \wr \mathbb{Z}_n$ was considered by [27] and [59] when $\mathbb{Q}'$ mixes completely in one step. [27] shows $\mathcal{T}(\mathbb{P}) = O(n^2)$ and [59] shows $\mathcal{T}_2(\mathbb{P}) = O(n^3)$. Our bounds give the same order of magnitude as those obtained by [27] and [59].

Proposition 6.23. *Let $\mathbb{Q}'$ be the natural walk on $\mathbb{Z}_2$ and $\mathbb{Q}$ the random walk on $\mathbb{Z}_n$. Consider $\mathbb{P} = \mathbb{Q}' \wr \mathbb{Z}_n$*

(a) $\mathcal{T}(\mathbb{P}) = \Theta(n^2)$

(b) $\mathcal{T}_2(\mathbb{P}) = \Theta(n^3)$

Proof. Since the controller part requires $\Omega(n^2)$ time to mix we have the lower bound. Since $\mathbb{Q}'$ mixes in one step, here we have $\gamma = 1$. [65] shows that the blanket time for $\mathbb{Z}_n$ is $O(n^2)$. Hence Proposition 6.7 together with Theorem 5.16 shows $\mathcal{T}(\mathbb{P}) = O(n^2)$.

For the L^2 upper bound: We apply Theorem 6.13 to see that when $t = O(n^3)$, the moment generating function of the number of unvisited states in $\mathbb{Z}_n$ becomes close to 1. Here we use the fact that the maximal hitting time of $\mathbb{Z}_n$ is $O(n^2)$.

For the L^2 lower bound: For $t > 0$, let $\mathbf{Z}_t$ denote the number states of $\mathbb{Z}_n$ which have been visited. Applying Theorem 5.17 with $\epsilon = 1$ and $\gamma(i) = 1$, we see that the t-step L^2-distance is at least $\sqrt{\mathbb{E}[2^{\mathbf{Z}_t}] - 1}$. Now Theorem 6.12 implies $\mathbb{E}[2^{\mathbf{Z}_t}] \geq 2$ unless $t = \Omega(|\mathbb{Z}_n| \mathcal{T}_{\mathrm{rel}}(\mathbb{Q})) = \Omega(n^3)$. ★

We now consider the case when $\mathbb{Q}'$ does not mix in one step. For concreteness we take $\mathbb{Q}'$ to be the random walk on $\mathbb{Z}_n^d$.

Theorem 6.24. *Let $\mathbb{Q}'$ be the random walk on $\mathbb{Z}_n^d$ and $\mathbb{Q}$ be the random walk on $\mathbb{Z}_n$. Consider $\mathbb{P} = \mathbb{Q}' \wr \mathbb{Q}$.*

(a) $\mathcal{T}(\mathbb{P}) = \Omega(n^3 \log n)$

(b) $\mathcal{T}_2(\mathbb{P}) = O(n^3 \log n)$

Proof. By Proposition 5.8, in order to mix in total variation distance, $\Omega(n)$ copies of $\mathbb{Z}_n^d$ must be $\Omega(1/\sqrt{n})$ close to stationarity. Since the relaxation time of $\mathbb{Q}'$ is $\Theta(n^2)$, we need to update $\Omega(n)$ copies $\Omega(n^2 \log n)$ times each, giving $\mathcal{T}(\mathbb{P}) = \Omega(n^3 \log n)$.

Take γ to be the constant function $O(n^2 \log n)$ and let $\mathbf{Z}_t$ be the number of states of $\mathbb{Z}_n$ which have not been visited γ-times, by time t. By Theorem 5.16, we need to find t such that $\mathbb{E}[n^{d\mathbf{Z}_t}]$ is close to 1. Applying Theorem 6.13, we see that it is enough to take $t = O(n^3 \log n)$ since the maximal hitting time of the simple random walk on $\mathbb{Z}_n$ is $\Theta(n^2)$. ★

Chapter 7

Open Questions

We conclude with a few questions.

7.1 Robust Mixing

Just how powerful is the adversary? Let $\mathbb{P}$ be a Markov Chain with stationary distribution π. How large can $\mathcal{R}(\mathbb{P})/\mathcal{T}(\mathbb{P}\overleftarrow{\mathbb{P}})$ be? Theorem 3.13 shows that it can be at most $O(\log(1/\pi_*))$ and Theorem 3.43 and Theorem 3.41 show that this is almost tight. However, the example in Theorem 3.41 has π_* exponentially small in the number of states of the chain. Can we improve the $O(\log(1/\pi_*))$ gap under additional assumptions on the stationary distribution?

Question 7.1. Let $\mathbb{P}$ be a Markov Chain on $\mathcal{X}$ with stationary distribution π. Is is true that $\mathcal{R}(\mathbb{P}) = O(\mathcal{T}(\mathbb{P}\overleftarrow{\mathbb{P}}))$ when the stationary distribution is uniform?

More specifically, if $\mathbb{P}$ is a Cayley walk on a group G, is it true that $\mathcal{R}(\mathbb{P}) = \mathcal{T}(\mathbb{P}\overleftarrow{\mathbb{P}})$?

What strategies can the adversary use to slow down the mixing? In this thesis,

we encountered two types of strategies:

(a) Reverse: Simulate the evolution of $\mathbb{P}\overleftarrow{\mathbb{P}}$,

(b) Collapse: Simulate the evolution of $\mathbb{Q}$ where $\mathbb{Q}$ is obtained from $\mathbb{P}$ by lumping certain states of $\mathbb{P}$ together

Question 7.2. Is there an optimal strategy for the adversary which is a combination of reversing and collapsing?

Question 7.3. Even though the adversarial strategy $\{\mathbb{A}_t\}$ is allowed to be time dependent, in all our examples the optimal adversarial strategy has been time homogenous, i.e. all the $\mathbb{A}_t$ are equal. Is this always the case? i. e., is it true that for some absolute constant C,

$$\mathcal{R}(\mathbb{P}) \leq C \cdot \max_{\mathbb{A}} \mathcal{T}(\mathbb{P}\mathbb{A}) \tag{7.1}$$

where the maximum is taken over $\mathbb{A}$ compatible with $\mathbb{P}$.

For the L^2-mixing time, Theorem 4.19 addresses the case when $\mathbb{P}$ is a Cayley walk and the adversary is restricted to respecting the group structure. If this were true, then one can make sense of the robust mixing time in continuous time as well.

7.2 Sandpile Chains

Let $\mathbb{P} = \mathrm{SP}(P_{n,d})$ denote the sandpile chain on the directed path. Theorem 5.33 and Proposition 5.32 show that $\mathcal{T}_2(\mathbb{P}) = \Theta(nd^2 \log n)$ and $\mathcal{T}(\mathbb{P}) = \Omega(nd^2)$. Can we close this gap?

Question 7.4. What is the exact order of magnitude of $\mathcal{T}(\mathrm{SP}(P_{n,d}))$?

If we use the Markov Chain on $\mathbb{Z}_d^n$ as a guide, one would expect $\mathcal{T}(\mathrm{SP}(P_{n,d})) = \Theta(nd^2 \log n)$. On the other hand, it is not clear that every digit needs to be selected at least once in order for the chain to mix. For example, if we update every alternate component $O(d^3)$ times each as well as the least significant digit $O(d^2)$ times, the chain has mixed, because of the way carries work.

Question 7.5. Let $\mathbb{P}$ be a Markov Chain on the wreath product $\mathbb{Z}_n \wr \mathbb{Z}_n^d$ for $d \geq 3$. Proposition 6.22 shows $\mathcal{T}_2(\mathbb{P}) = \Omega(n^{d+2} \log n)$ and $\mathcal{T}_2(\mathbb{P}) = O(n^{\min(d+4,2d)} \log n)$. What is the exact order of magnitude of $\mathcal{T}_2(\mathbb{P})$?

One way to close this gap would be to get good estimates on the moment generating function of occupancy measures of the random walk on $\mathbb{Z}_n^d$. Note that if the bound in (7.2) is true, we would have $\mathcal{T}_2(\mathbb{Z}_n \wr \mathbb{Z}_n^d) = \Theta(n^{d+2} \log n)$.

7.3 Occupancy Measures

Let $\mathbb{P}$ be a Markov Chain with stationary distribution π and $\gamma : \mathcal{X} \to \mathbb{Z}$ a target function. For $t > 0$, let $\mathbf{Z}_t$ denote the number of states $x \in \mathcal{X}$ which have been visited $< \gamma(x)$ times by time t.

Theorem 6.13, Theorem 6.19 and Theorem 5.22 give bounds on how large must t be for the moment generating function of $\mathbf{Z}_t$ to be close to 1, each assuming something about $\mathbb{P}$ and/or γ. Theorem 6.15 on the other hand doesn't make any assumptions but gives weak bounds.

Question 7.6. Is there a tight bound on how large t must be for the moment generating function of $\mathbf{Z}_t$ to be close to 1?

More specifically, suppose $\mathbb{P}$ is a reversible Markov Chain on $\mathcal{X}$ with uniform stationary distribution π and maximal hitting time $H = O(|\mathcal{X}|)$ and γ is any target

function. Is it true that for $\mathbb{E}[\theta^{\mathbf{Z}_t}]$ to be close to 1, it is enough to take

$$t = O\Big(|\mathcal{X}| \cdot (\gamma^* + \mathcal{T}_{\mathrm{rel}}(\mathbb{P}) \log \theta + \log(\theta|\mathcal{X}|))\Big) \tag{7.2}$$

Note that this bound reduces to the right value for those cases where tight bounds are known: $\mathcal{T}_{\mathrm{rel}}(\mathbb{P}) = 0$ is Theorem 5.22, $\gamma^* = 1$ is Theorem 6.19 and $\gamma_* \geq |\mathcal{X}| \log(\theta|\mathcal{X}|)$ is Theorem 6.13.

References

[1] David Aldous. Some inequalities for reversible Markov chains. *J. London Math. Soc. (2)*, 25(3):564–576, 1982.

[2] David Aldous and Persi Diaconis. Shuffling cards and stopping times. *The American Mathematical Monthly*, 93(5):333–348, May 1986.

[3] David Aldous and James Fill. Reversible markov chains and random walks on graphs, 200x. URL `http://www.stat.berkeley.edu/users/aldous/RWG/book.html`.

[4] László Babai. On the diameter of eulerian orientations of graphs. In *Proceedings of the SODA 2006*, 2006.

[5] Stephen Boyd, Persi Diaconis, Pablo Parrilo, and Lin Xiao. Symmetry analysis of reversible markov chains, 2003. URL `citeseer.ist.psu.edu/boyd03symmetry.html`.

[6] Stephen Boyd, Persi Diaconis, and Lin Xiao. Fastest mixing markov chain on a graph. *SIAM Review*, 46(4):667–689, 2004.

[7] Russ Bubley and Martin Dyer. Path coupling: A technique for proving rapid mixing in markov chains. In *Proceedings of the 38th Annual Symposium on Foundations of Computer Science*, page 223, Washington, DC, USA, 1997. IEEE Computer Society. ISBN 0-8186-8197-7.

[8] Fang Chen, László Lovász, and Igor Pak. Lifting markov chains to speed up mixing. In *Proceedings of the 31st annual ACM symposium on Theory of computing*, pages 275–281, New York, NY, USA, 1999. ACM Press. ISBN 1-58113-067-8. URL `http://doi.acm.org/10.1145/301250.301315`.

[9] Deepak Dhar. Self-organized critical state of sandpile automaton models. *Physical Review Letters*, 64(14):1613–1616, April 1990.

[10] Deepak Dhar, Philippe Ruelle, Siddhartha Sen, and D. N. Verma. Algebraic aspects of abelian sandpile models. *J.PHYS.A*, 28:805, 1995. URL `http://www.citebase.org/cgi-bin/citations?id=oai:arXiv.org:cond-mat/9408022`.

[11] Perci Diaconis and Laurent Saloff-Coste. Logarithmic sobolev inequalities for finite markov chains. *The Annals of Applied Probability*, 6(3):695–750, 1996.

[12] Perci Diaconis, James Allen Fill, and Jim Pitman. Analysis of top to random shuffles. *Combinatorics, Probability and Computing*, 1:135–155, 1992.

[13] Persi Diaconis. *Group Representations in Probability and Statistics*, volume 11 of *Lecture Notes – Monograph Series*. Institute of Mathematical Statistics, 1998.

[14] Persi Diaconis and Laurent Saloff-Coste. Comparison theorems for reversible markov chains. *Annals of Applied Probability*, 3(3):696–730, Aug 1993.

[15] Persi Diaconis and Mehrdad Shahshahani. Generating a random permutation with random transpositions. *Z. Wahrscheinlichkeitstheorie*, 57:159–179, 1981.

[16] Persi Diaconis, Susan Holmes, and Radford Neal. Analysis of a nonreversible markov chain sampler. *Annals of Applied Probability*, 10(3):726–752, 2000.

[17] Martin Dyer, Alan Frieze, and Mark Jerrum. On counting independent sets in sparse graphs. *SIAM Journal of Computing*, 33:1527–1541, 2002.

[18] Martin Dyer, Leslie Goldberg, Mark Jerrum, and Russell Martin. Markov chain comparison. *Probab. Surveys*, 3:89–111, 2006.

[19] Ky Fan and Alan Hoffman. Some metric inqualities in the space of matrices. *Proceedings of the American Mathematical Society*, 6:111–116, 1955.

[20] James Allen Fill. Eigenvalue bounds on convergence to stationarity for nonreversible markov chains with an application to the exclusion process. *Annals of Applied Probability*, 1(1):62–87, February 1991.

[21] Birkhoff G. Three observations on linear algebra. *Univ. Nac. Tucumán. Rev. Ser. A*, 5:147–151, 1946.

[22] Alison Gibbs and Francis Su. On choosing and bounding probability metrics. *Intl. Stat. Rev.*, 7(3):419–435, 2002.

[23] Sharad Goel. Analysis of top to bottom shuffles. *Annals of Applied Probability*, 16, February 2006. URL `http://www.stanford.edu/~scgoel/`.

[24] Sharad Goel. *Estimating Mixing Times: Techniques and Applications*. PhD thesis, Cornell University, 2005. URL `http://hdl.handle.net/1813/1500`.

[25] Sharad Goel, Ravi Montenegro, and Prasad Tetali. Mixing time bounds via the spectral profile. *Electronic Journal of Probability*, 11:1–26, 2006. URL `http://www.math.washington.edu/~ejpecp/EjpVol11/paper1.abs.html`.

[26] Gorenstein. *Finite Groups*. Chelsea, NY, 2 edition, 1980.

[27] Olle Häggström and Johan Jonasson. Rates of convergence for lamplighter processes. *Stochastic Process. Appl.*, 67:227–249, 1997. URL `http://dx.doi.org/10.1016/S0304-4149(97)00007-0`.

[28] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal*, pages 13–30, March 1963.

[29] Roger Horn and Charles Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[30] Mark Jerrum and Alistair Sinclair. Approximating the permanent. *SIAM J. Comput.*, 18:1149–1178, 1989.

[31] Mark Jerrum and Alistair Sinclair. Conductance and the rapid mixing property for markov chains: the approximation of the permanent resolved. In *20th Annual ACM Symposium on Theory of Computing*, pages 235–243, 1988.

[32] Johan Jonasson. Biased random-to-top shuffling. *Annals of Applied Probability*, 16(2), May 2006.

[33] Jeff Kahn, Jeong Han Kim, László Lovász, and Van Vu. The cover time, the blanket time, and the matthews bound. In *41st Annual Symposium on Foundations of Computer Science*, pages 467–475, 2000.

[34] Miclo Laurent. Remarques sur l'hypercontractivité et l'évolution de l'entropie pour des chaînes de markov finies. *Séminaire de probabilités de Strasbourg*, 31: 136–167, 1997. URL `http://www.numdam.org/item?id=SPS_1997__31__136_0`.

[35] Tom Leighton and Satish Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *Journal of ACM*, 46(6): 787–832, 1999. URL `http://doi.acm.org/10.1145/331524.331526`.

[36] László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random Struct. Algorithms*, 4(4):359–412, 1993.

[37] László Lovász and Peter Winkler. Mixing of random walks and other diffusions on a graph. In Walker, editor, *Surveys in Combinatorics*, number 187 in London Mathematical Society Lecture Note Series. Cambridge University Press, 1993.

[38] László Lovász and Peter Winkler. Reversal of markov chains and the forget time. *Comb. Probab. Comput.*, 7(2):189–204, 1998. ISSN 0963-5483. URL `http://dx.doi.org/10.1017/S0963548397003349`.

[39] László Lovász and Peter Winkler. Mixing times. In *Microsurveys in Discrete Probability*, DIMACS Series in Discrete Math. and Theor. Comp. Sci., pages 85–133. AMS, 1998.

[40] Albert Marshall and Ingram Olkin. *Inequalities: Theory of Majorization and its Applications*. Academic, New York, 1979.

[41] Milena Mihail. Conductance and convergence of markov chains: a combinatorial treatment of expanders. In *30th Symposium on Foundations of Computer Science*, pages 526–531. IEEE Computer Society Press, 1989.

[42] Ilya Mironov. (Not so) random shuffles of RC4. In *Crypto '02*, pages 304–319, 2002.

[43] Ravi Montenegro. Duality and evolving set bounds on mixing times, 2006. URL `http://www.ravimontenegro.com/research/evosets.pdf`.

[44] Ravi Montenegro. Convergence of markov chains: Lecture notes, 2004. URL `http://ravimontenegro.com/Archives/8843/index.html`.

[45] Ravi Montenegro and Prasad Tetali. *Mathematical Aspects of Mixing Times in Markov Chains*, volume 1 of *Foundations and Trends in Theoretical Computer Science*. now Publishers Inc., 2005. ISBN 1-933019-29-8.

[46] Ben Morris and Yuval Peres. Evolving sets and mixing. In *Proceedings of the 35th annual ACM symposium on Theory of computing*, pages 279–286, New York, NY, USA, 2003. ACM Press. ISBN 1-58113-674-9. URL `http://doi.acm.org/10.1145/780542.780585`.

[47] Elchanan Mossel, Yuval Peres, and Alistair Sinclair. Shuffling by semi-random transpositions, 2004. URL `http://www.citebase.org/cgi-bin/citations?id=oai:arXiv.org:math/0404438`.

[48] James Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

[49] Yuval Peres and David Revelle. Mixing times for random walks on finite lamplighter groups. *Electronic Journal of Probability*, 9:825–845, 2004. URL `http://www.math.washington.edu/~ejpecp/EjpVol9/paper26.abs.html`.

[50] Hazel Perfect and Leon Mirsky. Spectral properties of doubly stochastic matrices. *Monatsh. Math.*, 69:35–57, 1965.

[51] Pollard. Asymptotia, 200x. URL `http://www.stat.yale.edu/~pollard/Asymptotia/`.

[52] Omer Reingold. Undirected st-connectivity in log-space. *Electronic Colloquium on Computational Complexity (ECCC)*, 2004. URL `http://eccc.hpi-web.de/eccc-reports/2004/TR04-094/index.html`.

[53] Omer Reingold, Salil Vadhan, and Avi Wigderson. Entropy waves, the zig-zag graph product, and new constant degree expanders. *Annals of Mathematics*, 155 (1):157–187, 2002.

[54] Omer Reingold, Luca Trevisan, and Salil Vadhan. Pseudorandom walks in biregular graphs and the RL vs. L problem. Technical Report TR05-022, ECCC, February 2005. URL `http://eccc.uni-trier.de/eccc-reports/2005/TR05-022/index.html`.

[55] Laurent Saloff-Coste. *Probability on discrete structures*, volume 110 of *Encyclopaedia Math. Sci*, chapter 5, pages 263–346. Springer, Berlin, 2004. URL `http://www.math.cornell.edu/~lsc/chap5.pdf`.

[56] Clyde Schoolfield. Random walks on wreath products of groups. *Journal of Theoretical Probability*, 15(3):667–693, July 2002. URL `http://dx.doi.org/10.1023/A:1016219932004`.

[57] Alistair Sinclair. *Algorithms for Random Generation and Counting: A Markov Chain Approach*. Birkhäuser, Boston, 1993.

[58] Alistair Sinclair. Improved bounds for mixing rates of markov chains and multicommodity flow. *Combinatorics, Probability and Computing*, 1:351–370, 1992.

[59] Jay-Calvin Uyemura-Reyes. *Random walk, semi-direct products, and card shuffling*. PhD thesis, Stanford University, 2002.

[60] David Vere-Jones. Ergodic properties of nonnegative matrices - II. *Pacific Journal of Mathematics*, 26(3):601–620, 1968.

[61] Hermann Weyl. Inequalities between two kinds of eigenvalues of a linear transformation. In *Proceedings of National Academy of Sciences of the United States of America*, volume 35, pages 408–411, July 1949.

[62] Todd Will. Introduction to singular value decomposition, 1999. URL `http://www.uwlax.edu/faculty/will/svd/svd/index.html`.

[63] David Bruce Wilson. Mixing time of the rudvalis shuffle. *Electron. Comm. Probab.*, 8:77–85, 2003.

[64] David Bruce Wilson. Mixing times of lozenge tiling and card shuffling markov chains. *Ann. Appl. Probab.*, 14:274–325, 2004.

[65] Peter Winkler and David Zuckerman. Multiple cover time. *Random Structures and Algorithms*, 9(4):403–411, 1996. URL `citeseer.ist.psu.edu/139.html`.