Mathematical Aspects of Mixing Times in Markov Chains

Mathematical Aspects of Mixing Times in Markov Chains

Ravi Montenegro

Department of Mathematical Sciences
University of Massachusetts Lowell
Lowell, Massachusetts 01854, USA
ravi_montenegro@uml.edu

Prasad Tetali

School of Mathematics
Georgia Institute of Technology
Atlanta, Georgia 30332, USA
tetali@math.gatech.edu

the essence of knowledge

Contents

Abstract

In the past few years we have seen a surge in the theory of finite Markov chains, by way of new techniques to bounding the convergence to stationarity. This includes functional techniques such as logarithmic Sobolev and Nash inequalities, refined spectral and entropy techniques, and isoperimetric techniques such as the average and blocking conductance and the evolving set methodology. We attempt to give a more or less self-contained treatment of some of these modern techniques, after reviewing several preliminaries. We also review classical and modern lower bounds on mixing times. There have been other important contributions to this theory such as variants on coupling techniques and decomposition methods, which are not included here; our choice was to keep the analytical methods as the theme of this presentation. We illustrate the strength of the main techniques by way of simple examples, a recent result on the Pollard Rho random walk to compute the discrete logarithm, as well as with a brief and improved analysis of the Thorp shuffle.

Introduction

Monte Carlo methods have been in use for a long time in statistical physics and other fields for sampling purposes. However, the computer scientists' novel idea [43] of reducing the problem of approximately counting the size of a large set of combinatorial objects to that of near-uniform sampling from the same set, gave the study of Markov chain Monte Carlo (MCMC) algorithms an entirely new purpose, and promptly spawned an active subtopic of research. We recall here that the work of [43] shows that in fact, under the technical assumption of so-called *self-reducibility*, approximate counting of the size of a set *in polynomial time* is feasible if and only if one is able to sample from the set with nearly uniform distribution, also in polynomial time. In terms of the finite Markov chain underlying an MCMC algorithm, the latter problem translates to designing and analyzing a Markov chain with a prescribed stationary measure, with a view (and hope) to providing rigorous estimates on the polynomial fastness of the rate of convergence to stationarity of the chain. Thus the classical subject of finite Markov chains has received much renewed interest and attention.

To add concreteness to the above story, we briefly mention as examples of large sets of combinatorial objects, the set of matchings of a given (as input) bipartite graph [39, 41], the set of proper colorings of a given graph using a fixed number of colors [32], the number of matrices having non-negative integer entries and with prescribed row and column sums [15], etc. Albeit combinatorial, a non-discrete estimation problem which received significant devotion, both by way of algorithms and analytical techniques, is that of (approximately) computing the volume of a high-dimensional convex body (see [52, 53] and references therein). There have already been some very good surveys focusing on such combinatorial, computational and statistical physics applications of finite Markov chains. For an elaboration of the above premise, and a crash course on several basic techniques, we recommend the excellent article of Jerrum [38]. Towards the end of this introduction, we provide other pointers to existing literature on this subject. However, much of the theory surveyed in this article is rather recent theoretical (analytical) development and is so far unavailable in a unified presentation. The significance of these new methods is as follows.

The rate of convergence to stationarity of a finite Markov chain is typically measured by the so-called mixing time, defined as the first time τ by which the L^1 (or more generally, L^p) distance between the distribution at time τ and the stationary distribution falls below a small threshold, such as $1/2e$. It is classical and elementary to show that the inverse spectral gap of a lazy reversible Markov chain captures the mixing time (in L^1 and L^2) up to a factor of $\log(1/\pi_*)$, where $\pi_* = \min_x \pi(x)$ denotes the smallest entry in the stationary probability (vector) π of the chain. While the more technical logarithmic Sobolev constant captures the L^2-mixing time up to a factor of $\log\log(1/\pi_*)$, it is typically much harder to bound – to mention a specific example, the exact constant is open for the 3-point space with arbitrary invariant measure; also in a few cases, the log-Sobolev constant is known not to give tight bounds on the L^1-mixing time. The main strength of the spectral profile techniques and the evolving set methodology considered in this survey seems to be that of avoiding extra penalty factors such as $\log\log(1/\pi_*)$. These extra pesky factors can indeed be non-negligible when the state space is of exponential (or worse) size in the size of the input. In the present volume, the above is illustrated with a couple of simple examples, and with the now-famous Thorp shuffle, for which an improved $O(d^{29})$ mixing time is described, building on the proof of Morris that proved the first polynomial (in d) bound of $O(d^{44})$ – here the number of cards in the deck is 2^d, and hence the state space has size $2^d!$, resulting in a $\log\log(1/\pi_*)$ factor of only $O(d)$, while a $\log(1/\pi_*)$ factor would have yielded an all too costly $O(d2^d)$.

The approach to L^2-mixing time using the spectral profile has the additional advantage of yielding known (upper) estimates on mixing time, under a log-Sobolev inequality and/or a Nash-type inequality. Thus various functional analytic approaches to mixing times can be unified with the approach of bounding the spectral profile. The one exception to this is the approach to stationarity using relative entropy; the corresponding *entropy constant* capturing the rate of decay of entropy has also been hard to estimate.

A brief history of the above development can perhaps be summarized as follows. A fundamental contribution, by way of initiating several subsequent works, was made by Lovász and Kannan in [51] in which they introduced the notion of *average conductance* to bound the total variation mixing time. This result was further strengthened and developed by Morris and Peres using the so-called *evolving sets*, where they analyze a given chain by relating it to an auxiliary (dual) chain on subsets of the states of the original chain. While this was introduced in [65] in a (martingale-based) probabilistic language, it turns out to be, retrospectively, an independent and alternative view of the notion of a Doob transform introduced and investigated by Diaconis and Fill [22]. Further refinement and generalization of the evolving sets approach was done in detail by [61]. The functional analog of some of this is done via the spectral profile, developed for the present context of finite Markov chains, in [34], while having its origins in the developments by [4] and [18] in the context of manifolds.

Besides summarizing much of the above recent developments in this exciting topic, we address some classical aspects as well. In discrete-time, much of the literature uses laziness assumptions to avoid annoying technical difficulties. While laziness is a convenient assumption, it slows down the chain by a factor of 2, which may not be desirable in practice. We take a closer look at this issue and report bounds which reflect the precise dependence on laziness. The notion of modified conductance circumvents laziness altogether, and we discuss this aspect briefly and compare it to bounds derived from the functional approach. Further details on the modified conductance and its usefulness can be found in [62]. Another issue is that of the role of reversibility (a.k.a. detailed balance conditions). We tried to pay particular attention to it, due to current trend in the direction of analyzing various nonreversible Markov chains. Although often a convenient assumption, we avoid as much as possible this additional assumption. In particular, we include a proof of the lower bound on the total variation mixing time in terms of the second eigenvalue in the general case. Besides providing upper and lower bounds for the mixing time of reversible and non-reversible chains, we report recent successes (with brief analysis) in the analysis of some non-reversible chains; see for example, the Pollard Rho random walk for the discrete logarithm problem and the Thorp shuffle.

In Chapter 1 we introduce notions of mixing times and prove the basic upper bounds on these notions using Poincaré and logarithmic Sobolev type functional constants. In Chapter 2 we move on to recent results using the spectral profile, as opposed to using simply the second eigenvalue. In Chapter 3 we review the evolving set methods. Our treatment of lower bounds on mixing times is provided in Chapter 4. We consider several examples for illustration in Chapter 5. In the penultimate chapter, we gather a few recent results together. This includes recent results on the so-called fastest mixing Markov chain problem, and a recent theorem [57] from perturbation theory of finite Markov chains; this theorem relates the stability of a stochastic matrix (subject to perturbations) to the rate of convergence to equilibrium of the matrix. We also recall here an old but not so widely known characterization of the spectral gap, which seems worth revisiting due to recent results utilizing this formulation. The Appendix contains a discussion on the relations between the distances considered in this paper, and others such as relative pointwise (L^∞) distance.

We mention here a few additional sources, by way of survey articles, for the interested reader. For a good overview of the basic techniques in estimating the mixing times of finite Markov chains, see [40, 38, 37]. Other updates include the tutorial lectures of [45], [69]. Also a recent manuscript of Dyer et al. [29] describes several comparison theorems for reversible as well as nonreversible Markov chains.

Acknowledgments. We thank Pietro Caputo and Eric Carlen for several helpful discussions while preparing this manuscript.

1

Basic Bounds on Mixing Times

1.1 Preliminaries: Distances and mixing times

Let $(\Omega, \mathsf{P}, \pi)$ denote a transition probability matrix (or Markov kernel) of a finite Markov chain on a finite state space Ω with a unique invariant measure π. That is

$$\mathsf{P}(x,y) \geq 0, \quad \text{for all } x, y \in \Omega, \quad \text{and} \sum_{y\in\Omega} \mathsf{P}(x,y) = 1, \quad \text{for all } x \in \Omega.$$

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

We assume throughout this paper that P is irreducible (i.e. Ω is strongly connected under P) and that π has full support (Ω). The minimal holding probability $\alpha \in [0,1]$ satisfies $\forall x \in \Omega : \mathsf{P}(x,x) \geq \alpha$, and if $\alpha \geq 1/2$ the chain is said to be lazy. If $A, B \subset \Omega$ the ergodic flow is $\mathsf{Q}(A,B) = \sum_{x\in A,\, y\in B} \pi(x)\mathsf{P}(x,y)$, while $A^c = \Omega \setminus A$ is the complement. For standard definitions and introduction to finite Markov chains, we refer the reader to [67] or [1].

It is a classical fact that if P is aperiodic then the measures $\mathsf{P}^n(x,\cdot)$ approach π as $n \to \infty$. Alternatively, let $k_n^x(y) = \mathsf{P}^n(x,y)/\pi(y)$ denote the density with respect to π at time $n \geq 0$, or simply $k_n(y)$ when the start state or the start distribution is unimportant or clear from the context. Then the density $k_n^x(y)$ converges to 1 as $n \to \infty$. A proper quantitative statement may be stated using any one of several norms. In terms of L^p-distance

$$\|k_n - 1\|_{p,\pi}^p = \sum_{y \in \Omega} |k_n(y) - 1|^p \, \pi(y) \quad 1 \le p < +\infty \,.$$

When $p = 1$ and $p = 2$ these are closely related to the total variation distance and variance, respectively, such that if μ is a probability distribution on Ω, then

$$\begin{aligned} \|\mu - \pi\|_{TV} &= \frac{1}{2} \left\| \frac{\mu}{\pi} - 1 \right\|_{1,\pi} &&= \frac{1}{2} \sum_{y \in \Omega} |\mu(y) - \pi(y)| \\ \mathrm{Var}_\pi(\mu/\pi) &= \left\| \frac{\mu}{\pi} - 1 \right\|_{2,\pi}^2 &&= \sum_{y \in \Omega} \pi(y) \left(\frac{\mu(y)}{\pi(y)} - 1 \right)^2 \end{aligned}$$

Another important measure of closeness (but not a norm) is the informational divergence,

$$\mathsf{D}(\mathsf{P}^n(x,\cdot)\|\pi) = \mathrm{Ent}_\pi(k_n^x) = \sum_{y \in \Omega} \mathsf{P}^n(x,y) \log \frac{\mathsf{P}^n(x,y)}{\pi(y)} \,,$$

where the entropy $\mathrm{Ent}_\pi(f) = \mathbb{E}_\pi f \log \frac{f}{\mathbb{E}_\pi f}$.

Each of these distances are convex, in the sense that if μ and ν are two distributions, and $s \in [0,1]$ then $dist((1-s)\mu + s\nu, \pi) \le (1-s)\, dist(\mu,\pi) + s\, dist(\nu,\pi)$. For instance, $\mathsf{D}(\mu\|\pi) = \mathrm{Ent}_\pi(\mu/\pi) = \mathbb{E}_\pi \frac{\mu}{\pi} \log \frac{\mu}{\pi}$ is convex in μ because $f \log f$ is convex. A convex distance $dist(\mu,\pi)$ satisfies the condition

$$\begin{aligned} dist(\sigma\mathsf{P}^n, \pi) &= dist\left(\sum_{x \in \Omega} \sigma(x) \mathsf{P}^n(x,\cdot), \pi \right) \\ &\le \sum_{x \in \Omega} \sigma(x) dist\left(\mathsf{P}^n(x,\cdot), \pi\right) \\ &\le \max_{x \in \Omega} dist(\mathsf{P}^n(x,\cdot),\pi)\,, \end{aligned} \tag{1.1}$$

and so distance is maximized when the initial distribution is concentrated at a point. To study the rate of convergence it then suffices to study the rate when the initial distribution is a point mass δ_x (where δ_x is 1 at point $x \in \Omega$ and 0 elsewhere; likewise, let 1_A be one only on set $A \subset \Omega$).

Definition 1.1. The total variation, relative entropy and L^2 mixing times are defined as follows.

$$\begin{aligned}
\tau(\epsilon) &= \min\{n : \ \forall x \in \Omega,\ \|\mathsf{P}^n(x,\cdot) - \pi\|_{\mathrm{TV}} \le \epsilon\} \\
\tau_{\mathsf{D}}(\epsilon) &= \min\{n : \ \forall x \in \Omega,\ \mathsf{D}(\mathsf{P}^n(x,\cdot)\|\pi) \le \epsilon\} \\
\tau_2(\epsilon) &= \min\{n : \ \forall x \in \Omega,\ \|k_n^x - 1\|_{2,\pi} \le \epsilon\}
\end{aligned}$$

One may also consider the chi-square (χ^2) distance, which is just $\mathrm{Var}(k_n^x)$ and mixes in $\tau_{\chi^2}(\epsilon) = \tau_2(\sqrt{\epsilon})$. In the Appendix it is seen that $\tau_2(\epsilon)$ usually gives a good bound on L^∞ convergence, and so for most purposes nothing stronger than L^2 mixing need be considered.

An important concept in studying Markov chains is the notion of reversibility. The time-reversal P^* is defined by the identity $\pi(x)\mathsf{P}^*(x,y) = \pi(y)\mathsf{P}(y,x)$, $x, y \in \Omega$ and is the adjoint of P in the standard inner product for $L^2(\pi)$, that is $\langle f, \mathsf{P}g\rangle_\pi = \langle \mathsf{P}^* f, g\rangle_\pi$ where

$$\langle f, g\rangle_\pi = \sum_{x\in\Omega} \pi(x) f(x) g(x)$$

and a matrix M acts on a function $f : \Omega \to \mathsf{R}$ as

$$M\,f(x) = \sum_{y\in\Omega} M(x,y) f(y)\,.$$

A useful property of the reversal is that $k_n = \mathsf{P}^* k_{n-1}$, and inductively $k_n = (\mathsf{P}^*)^n k_0$. If $\mathsf{P}^* = \mathsf{P}$ then P is said to be time-reversible, or to satisfy the detailed balance condition. Given any Markov kernel P, two natural reversible chains are the additive reversibilization $\frac{\mathsf{P}+\mathsf{P}^*}{2}$, and multiplicative reversibilization $\mathsf{P}\mathsf{P}^*$.

A straightforward way to bound the L^2-distance is to differentiate the variance. In Lemma 1.4 it will be found that $\frac{d}{dt}\mathrm{Var}(h_t) = -2\mathcal{E}(h_t, h_t)$, where $\mathcal{E}(f,g)$ denotes a Dirichlet form, as defined below, and h_t the continuous time density defined in the following section. More generally, the Dirichlet form can be used in a characterization of eigenvalues of a reversible chain (see Lemma 1.21), and to define the spectral gap and the logarithmic Sobolev type inequalities:

Definition 1.2. For $f, g : \Omega \to \mathsf{R}$, let $\mathcal{E}(f,g) = \mathcal{E}_{\mathsf{P}}(f,g)$ denote the Dirichlet form,

$$\mathcal{E}(f,g) = \langle f, (\mathsf{I}-\mathsf{P})g\rangle_\pi = \sum_{x,y} f(x)\,(g(x)-g(y))\,\mathsf{P}(x,y)\pi(x)\,.$$

If $f = g$ then

$$\mathcal{E}(f,f) = \frac{1}{2}\sum_{x,y\in\Omega}(f(x)-f(y))^2\mathsf{P}(x,y)\pi(x)\,, \tag{1.2}$$

and

$$\mathcal{E}_{\mathsf{P}}(f,f) = \mathcal{E}_{\mathsf{P}^*}(f,f) = \mathcal{E}_{\frac{\mathsf{P}+\mathsf{P}^*}{2}}(f,f)\,, \tag{1.3}$$

while if P is reversible then also $\mathcal{E}(f,g) = \mathcal{E}(g,f)$.

Finally, we recall some notation from complexity theory which will be used occasionally. Given positive functions $f, g : \mathsf{R}_+ \to \mathsf{R}_+$ we say that $f = O(g)$ if $f \le c\,g$ for some constant $c \ge 0$, while $f = \Omega(g)$ if $f \ge c\,g$ for a constant $c \ge 0$, and finally $f = \Theta(g)$ if $c_1\,g \le f \le c_2\,g$ for constants $c_1, c_2 \ge 0$. For instance, while attempting to analyze an algorithm requiring $\tau(n) = 3n^4 + n$ steps to terminate on input of size n, it might be found that $\tau(n) = O(n^5)$, or $\tau(n) = \Omega(n\log n)$, when in fact $\tau(n) = \Theta(n^4)$.

1.2 Continuous Time

Many mixing time results arise in a natural, clean fashion in the continuous time setting, and so we consider this case first. The arguments developed here will then point the way for our later consideration of discrete time results.

Let $\mathcal{L}$ denote the (discrete) Laplacian operator given by $\mathcal{L} = -(\mathsf{I}-\mathsf{P})$. Then for $t \ge 0$, $H_t = e^{t\mathcal{L}}$ represents the continuized chain [1] (or the heat kernel) corresponding to the discrete Markov kernel P. The continuized chain simply represents a Markov process $\{X_t\}_{t\ge 0}$ in Ω with initial distribution, μ_0 (say), and transition matrices

$$H_t = e^{-t(\mathsf{I}-\mathsf{P})} = \sum_{n=0}^{\infty}\frac{t^n\mathcal{L}^n}{n!} = e^{-t}\sum_{n=0}^{\infty}\frac{t^n\mathsf{P}^n}{n!}, \quad t \ge 0,$$

with the generator $\mathcal{L} = -(\mathsf{I} - \mathsf{P})$. Thus $H_t(x, y)$ denotes the probability that the rate one continuous Markov chain having started at x is at y at time t. Let $h_t^x(y) = H_t(x, y)/\pi(y)$, for each $y \in \Omega$, denote its density with respect to π at time $t \geq 0$, and $h_t(y)$ when the start state or the start distribution is unimportant or clear from the context. Also, let

$$H_t^* = e^{t\mathcal{L}^*} = \sum_{n=0}^{\infty} \frac{t^n (\mathcal{L}^*)^n}{n!}$$

be the semigroup associated to the dual $\mathcal{L}^* = -(\mathsf{I} - \mathsf{P}^*)$. The following is elementary and a useful technical fact.

Lemma 1.3. For any h_0 and all $t \geq 0$, $h_t = H_t^* h_0$. Consequently, for any $x \in \Omega$,

$$\frac{dh_t(x)}{dt} = \mathcal{L}^* h_t(x).$$

Using Lemma 1.3, the following lemma is easy to establish.

Lemma 1.4.

$$\begin{aligned} \frac{d}{dt}\mathrm{Var}(h_t) &= -2\mathcal{E}(h_t, h_t) && (1.4) \\ \frac{d}{dt}\mathrm{Ent}(h_t) &= -\mathcal{E}(h_t, \log h_t) && (1.5) \end{aligned}$$

Proof. Indeed,

$$\begin{aligned} \frac{d}{dt}\mathrm{Var}(h_t) &= \int \frac{d}{dt} h_t^2 d\pi = 2 \int h_t \mathcal{L}^* h_t d\pi \\ &= 2 \int \mathcal{L}(h_t) h_t d\pi = -2\mathcal{E}(h_t, h_t). \end{aligned}$$

$$\begin{aligned} \frac{d}{dt}\mathrm{Ent}(h_t) &= \int \frac{d}{dt} h_t \log h_t d\pi = \int (\log h_t + 1)\mathcal{L}^* h_t d\pi \\ &= \int \mathcal{L}(\log h_t) h_t d\pi = -\mathcal{E}(h_t, \log h_t). \end{aligned}$$

□

The above motivates the following definitions of the spectral gap λ and the entropy constant ρ_0.

Definition 1.5. Let $\lambda > 0$ and $\rho_0 > 0$ be the optimal constants in the inequalities:

$$\lambda \mathrm{Var}_\pi f \le \mathcal{E}(f,f), \quad \text{for all } f : \Omega \to \mathsf{R}.$$

$$\rho_0 \mathrm{Ent}_\pi f \le \mathcal{E}(f, \log f), \quad \text{for all } f : \Omega \to \mathsf{R}_+. \tag{1.6}$$

When it is necessary to specify the Markov chain K being considered then use the notation λ_{K}.

Lemma 1.21 (Courant-Fischer theorem) shows that for a reversible Markov chain, the second largest eigenvalue λ_1 (of P) satisfies the simple relation $1 - \lambda_1 = \lambda$. However, reversibility is not needed for the following result.

Corollary 1.6. Let $\pi_* = \min_{x\in\Omega} \pi(x)$. Then, in continuous time,

$$\tau_2(\epsilon) \le \frac{1}{\lambda}\left(\frac{1}{2}\log\frac{1-\pi_*}{\pi_*} + \log\frac{1}{\epsilon}\right) \tag{1.7}$$

$$\tau_{\mathsf{D}}(\epsilon) \le \frac{1}{\rho_0}\left(\log\log\frac{1}{\pi_*} + \log\frac{1}{\epsilon}\right). \tag{1.8}$$

Proof. Simply solve the differential equations,

$$\frac{d}{dt}\mathrm{Var}(h_t^x) = -2\mathcal{E}(h_t^x, h_t^x) \le -2\lambda\,\mathrm{Var}(h_t^x) \tag{1.9}$$

and

$$\frac{d}{dt}\mathrm{Ent}(h_t^x) = -\mathcal{E}(h_t^x, \log h_t^x) \le -\rho_0\,\mathrm{Ent}(h_t^x), \tag{1.10}$$

and note that $\mathrm{Var}(h_0) \le \frac{1-\pi_*}{\pi_*}$ and $\mathrm{Ent}(h_0) \le \log\frac{1}{\pi_*}$ (e.g. by equation (1.1)). □

It is worth noting here that the above functional constants λ and ρ_0 indeed capture the rate of decay of variance and relative entropy, respectively, of H_t for $t > 0$:

Proposition 1.7. If $c > 0$ then

(a) $\mathrm{Var}_\pi(H_t f) \le e^{-ct}\mathrm{Var}_\pi f$, for all f and $t > 0$, if and only if $\lambda \ge c$.

(b) $\mathrm{Ent}_\pi(H_t f) \le e^{-ct}\mathrm{Ent}_\pi f$, for all $f > 0$ and $t > 0$, if and only if $\rho_0 \ge c$.

Proof. The "if" part of the proofs follows from (1.9) and (1.10). The only if is also rather elementary and we bother only with that of part (b): Starting with the hypothesis, we may say, for every $f > 0$, and for $t > 0$,

$$\frac{1}{t}\Big(\mathrm{Ent}_\pi(H_t f) - \mathrm{Ent}_\pi f\Big) \le \frac{1}{t}\Big(e^{-ct} - 1\Big)\mathrm{Ent}_\pi f.$$

Letting $t \downarrow 0$, we get $-\mathcal{E}(f, \log f) \le -c\mathrm{Ent}_\pi f$. □

While there have been several techniques (linear-algebraic and functional-analytic) to help bound the spectral gap, the analogous problem of getting good estimates on ρ_0 seems challenging. The following inequality relating the two Dirichlet forms introduced above also motivates the study of the classical logarithmic Sobolev inequality. In practice this is a much easier quantity to bound, and moreover it will later be shown to bound the stronger L^2 mixing time, and hence L^∞ as well.

Lemma 1.8. If $f \ge 0$ then

$$2\mathcal{E}(\sqrt{f}, \sqrt{f}) \le \mathcal{E}(f, \log f)$$

Proof. Observe that

$$a(\log a - \log b) = 2a \log\frac{\sqrt{a}}{\sqrt{b}} \ge 2a\left(1 - \frac{\sqrt{b}}{\sqrt{a}}\right) = 2\sqrt{a}\,(\sqrt{a} - \sqrt{b})$$

by the relation $\log c \ge 1 - c^{-1}$. Then

$$\begin{aligned}
\mathcal{E}(f, \log f) &= \sum_{x,y} f(x)(\log f(x) - \log f(y))\mathsf{P}(x,y)\pi(x) \\
&\ge 2\sum_{x,y} f^{1/2}(x)(f^{1/2}(x) - f^{1/2}(y))\mathsf{P}(x,y)\pi(x) \\
&= 2\mathcal{E}(\sqrt{f}, \sqrt{f})
\end{aligned}$$

□

Let $\rho_{\mathsf{P}} > 0$ denote the logarithmic Sobolev constant of P defined as follows.

Definition 1.9.

$$\rho = \rho_{\mathsf{P}} = \inf_{\mathrm{Ent} f^2 \neq 0} \frac{\mathcal{E}(f,f)}{\mathrm{Ent} f^2}.$$

Proposition 1.10. For every irreducible chain P,

$$2\rho \leq \rho_0 \leq 2\lambda\,.$$

Proof. The first inequality is immediate, using Lemma 1.8. The second follows from applying (1.6) to functions $f = 1 + \epsilon g$, for $g \in L^2(\pi)$ with $\mathbb{E}_\pi g = 0$. Assume $\epsilon \ll 1$, so that $f \geq 0$. Then using the Taylor approximation, $\log(1+\epsilon g) = \epsilon g - 1/2(\epsilon)^2 g^2 + o(\epsilon^2)$, we may write

$$\mathrm{Ent}_\pi(f) = \frac{1}{2}\epsilon^2 \pi(g^2) + o(\epsilon^2),$$

and

$$\mathcal{E}(f, \log f) = -\epsilon \mathbb{E}_\pi((\mathcal{L}g)\log(1+\epsilon g)) = \epsilon^2 \mathcal{E}(g,g) + o(\epsilon^2).$$

Thus starting from (1.6), and applying to f as above, we get

$$\epsilon^2 \mathcal{E}(g,g) \geq \frac{\rho_0}{2}\epsilon^2 \mathbb{E}_\pi g^2 + o(\epsilon^2).$$

Canceling ϵ^2 and letting $\epsilon \downarrow 0$, yields the second inequality of the proposition, since $\mathbb{E}_\pi g = 0$.

□

Remark 1.11. The relation $2\rho \leq 2\lambda$ found in the lemma can be strengthened somewhat to $\rho \leq \lambda/2$, by a direct application of the method used above. Under the additional assumption of reversibility, the inequality in Lemma 1.8 can be strengthened by a factor of 2 to match this, as explained in [25], in turn improving the above proposition to $4\rho \leq \rho_0 \leq 2\lambda$ for reversible chains.

1.3 Discrete Time

We now turn our attention to discrete time. A mixing time bound in terms of the spectral gap will be shown in a fashion similar to that in continuous time. There seems to be no discrete-time analog of the modified log-Sobolev bound on relative entropy, although in Chapter 3 a bound in terms of Evolving Set will be found. We defer consideration of the log-Sobolev constant to Section 2.1.

In discrete time we consider two approaches to mixing time, both of which are equivalent. The first approach involves operator norms, and is perhaps the more intuitive of the two methods.

Proposition 1.12. In discrete time,

$$\tau_2(\epsilon) \leq \left\lceil \frac{1}{1-\|\mathsf{P}^*\|} \log \frac{1}{\epsilon\sqrt{\pi_*}} \right\rceil ,$$

where $\mathsf{P}^*(x,y) = \frac{\pi(y)\mathsf{P}(y,x)}{\pi(x)}$, $\pi_* = \min_{x\in\Omega} \pi(x)$ and

$$\|\mathsf{P}^*\| = \sup_{f:\Omega\to\mathsf{R},\,\mathbb{E}f=0} \frac{\|\mathsf{P}^* f\|_2}{\|f\|_2} .$$

This result has appeared in mixing time literature in many equivalent forms. A few can be found in Remark 1.19 at the end of this section.

Proof. Since $k_{i+1} - 1 = \mathsf{P}^*(k_i - 1)$ and $\mathbb{E}(k_i - 1) = 0$ for all i then

$$\|k_n - 1\|_2 = \|\mathsf{P}^{*n}(k_0 - 1)\|_2 \leq \|\mathsf{P}^*\|^n \|k_0 - 1\|_2 .$$

Solving for when this expression drops to ϵ and using the approximations $\log x \leq -(1-x)$ and $\|k_0 - 1\|_2 \leq \sqrt{\frac{1-\pi_*}{\pi_*}}$ gives the result. □

A good example in which this bound has been used in practice can be found in Section 5.2, in which we discuss a recent proof that Pollard's Rho algorithm for discrete logarithm requires order $\sqrt{n}\log^3 n$ steps to detect a collision, and likely determine the discrete log.

In Proposition 1.12 the mixing bound followed almost immediately from the definition. However, there is an alternate approach to this problem which bears more of a resemblance to the continuous time result and is more convenient for showing refined bounds.

The discrete time analog of differentiating $\mathrm{Var}(h_t)$ is to take the difference $\mathrm{Var}(k_n) - \mathrm{Var}(k_{n-1})$, or more generally, $\mathrm{Var}(\mathsf{P}^* f) - \mathrm{Var}(f)$. The analog of equation (1.4) is the following lemma of Mihail [55], as formulated by Fill [30].

Lemma 1.13. Given Markov chain P and function $f : \Omega \to \mathsf{R}$, then

$$\mathrm{Var}(\mathsf{P}^* f) - \mathrm{Var}(f) = -\mathcal{E}_{\mathsf{PP}^*}(f, f) \le -\mathrm{Var}(f)\, \lambda_{\mathsf{PP}^*} \,.$$

Proof. Note that $\mathbb{E}_\pi f = \mathbb{E}_\pi (Kf)$ for any transition probability matrix K, because $\sum_x \pi(x) \mathsf{K}(x, y) = \pi(y)$. It follows that

$$\mathrm{Var}(\mathsf{P}^* f) - \mathrm{Var}(f) = \langle \mathsf{P}^* f, \mathsf{P}^* f \rangle_\pi - \langle f, f \rangle_\pi = -\langle f, (\mathsf{I} - \mathsf{PP}^*) f \rangle_\pi \,,$$

giving the equality. The inequality follows from Definition 1.5 of λ_{PP^*}. □

In Lemma 1.21 it will be found that $1 - \lambda_{\mathsf{PP}^*}$ is the largest non-trivial singular value of P.

Proceeding, we now bound mixing time of a discrete time chain.

Corollary 1.14. A discrete time Markov chain P satisfies

$$\tau_2(\epsilon) \le \left\lceil \frac{2}{\lambda_{\mathsf{PP}^*}} \log \frac{1}{\epsilon \sqrt{\pi_*}} \right\rceil \,.$$

Hence, to study mixing in discrete-time consider the multiplicative reversibilization PP^*, and in continuous-time consider the additive reversibilization $\frac{\mathsf{P}+\mathsf{P}^*}{2}$ (as $\mathcal{E}(f, f) = \mathcal{E}_{\frac{\mathsf{P}+\mathsf{P}^*}{2}}(f, f)$, and so $\lambda = \lambda_{\frac{\mathsf{P}+\mathsf{P}^*}{2}}$).

Proof. Recall the n-step density satisfies $k_n = (\mathsf{P}^*)^n k_0$. Then, by Lemma 1.13, $\mathrm{Var}(k_n) \le \mathrm{Var}(k_{n-1})\,(1 - \lambda_{\mathsf{PP}^*})$, and by induction

$$\mathrm{Var}(k_n) \le \mathrm{Var}(k_0)\ (1 - \lambda_{\mathsf{PP}^*})^n \,. \tag{1.11}$$

The result follows by solving for when variance drops to ϵ^2 and using the approximation $\log(1 - \lambda_{\mathsf{PP}^*}) \le -\lambda_{\mathsf{PP}^*}$. □

It is preferable to work with P instead of PP^*. Several simplifications make this possible.

Corollary 1.15. In discrete time, a Markov chain with holding probability α satisfies

$$\tau_2(\epsilon) \leq \left\lceil \frac{1}{\alpha\lambda} \log \frac{1}{\epsilon\sqrt{\pi_*}} \right\rceil .$$

For a reversible Markov chain,

$$\tau_2(\epsilon) \leq \left\lceil \frac{1}{1-\lambda_{max}} \log \frac{1}{\epsilon\sqrt{\pi_*}} \right\rceil \leq \left\lceil \frac{1}{\min\{2\alpha,\, \lambda\}} \log \frac{1}{\epsilon\sqrt{\pi_*}} \right\rceil ,$$

where $\lambda_{max} = \max\{\lambda_1,\, |\lambda_{n-1}|\}$ when $\lambda_1 = 1-\lambda$ is the largest non-trivial eigenvalue of P and $\lambda_{n-1} \geq -1$ is the smallest eigenvalue.

Proof. Observe that $\forall x \in \Omega :\ \mathsf{P}^*(x,x) = \mathsf{P}(x,x) \geq \alpha$, and so

$$\begin{aligned}\pi(x)\,\mathsf{PP}^*(x,y) &\geq \pi(x)\,\mathsf{P}(x,x)\,\mathsf{P}^*(x,y) + \pi(x)\,\mathsf{P}(x,y)\,\mathsf{P}^*(y,y) \\ &\geq \alpha\,\pi(y)\mathsf{P}(y,x) + \alpha\,\pi(x)\mathsf{P}(x,y)\,.\end{aligned}$$

It follows from Equation 1.2 that

$$\mathcal{E}_{\mathsf{PP}^*}(f,f) \geq \alpha\,\mathcal{E}(f,f) + \alpha\,\mathcal{E}(f,f) = 2\alpha\mathcal{E}(f,f)\,, \qquad (1.12)$$

and so $\lambda_{\mathsf{PP}^*} \geq 2\alpha\lambda$. The first bound then follows from Corollary 1.14.

For the reversible case, we require Lemmas 1.20 and 1.21, to be shown in the next section. Lemma 1.20 shows that P has an eigenbasis. If λ_i is an eigenvalue of P with corresponding right eigenvector v_i then $\mathsf{PP}^*\,v_i = \mathsf{P}^2\,v_i = \lambda_i^2\,v_i$, and so the eigenvalues of PP^* are just $\{\lambda_i^2\}$. By Lemma 1.21 (to be shown later) it follows that

$$\lambda_{\mathsf{PP}^*} = \lambda_{\mathsf{P}^2} = 1 - \max\{\lambda_1^2, \lambda_{n-1}^2\} = 1 - \lambda_{max}^2\,.$$

Solving equation (1.11) then gives the first reversible bound.

Finally, if P is reversible then $\frac{\lambda_{n-1}-\alpha}{1-\alpha}$ is the smallest eigenvalue of the reversible Markov chain $\frac{\mathsf{P}-\alpha\mathsf{I}}{1-\alpha}$, so Lemma 1.20 shows that $\frac{\lambda_{n-1}-\alpha}{1-\alpha} \geq -1$. Re-arranging the inequality gives the relation $-\lambda_{n-1} \leq 1-2\alpha$, and so $\lambda_{max} = \max\{\lambda_1, -\lambda_{n-1}\} \leq 1 - \min\{\lambda, 2\alpha\}$. □

Remark 1.16. In the proof above it was found that $\mathcal{E}_{\mathsf{PP}^*}(f,f) \geq 2\alpha\mathcal{E}(f,f)$, and so in particular $\lambda_{\mathsf{PP}^*} \geq 2\alpha\lambda$. Since $\mathcal{E}_{\mathsf{P}}(f,f) = \mathcal{E}_{\frac{\mathsf{P}+\mathsf{P}^*}{2}}(f,f)$ then this is a statement that the additive reversibilization can be used to lower bound the multiplicative reversibilization.

A related inequality holds in the opposite direction as well. Recall from the proof above that if P is reversible then $\lambda_{\mathsf{PP}^*} = 1 - \lambda_{max}^2$. Re-arranging this shows that

$$1 - \lambda_{max} = 1 - \sqrt{1 - \lambda_{\mathsf{PP}^*}} \geq \frac{1}{2}\lambda_{\mathsf{PP}^*}\,.$$

In Theorem 5.10 we will see that this holds, with $\lambda_{max} = \max_{\lambda_i \neq 1} |\lambda_i|$, even for non-reversible walks with complex eigenvalues.

In summary, if α is the holding probability, and P is reversible then

$$\lambda \geq \frac{1}{2}\lambda_{\mathsf{PP}^*} \geq \alpha\lambda\,,$$

while in general if $\{\lambda_i\}$ are the eigenvalues of P then

$$1 - \max_{\lambda_i \neq 1} |\lambda_i| \geq \frac{1}{2}\lambda_{\mathsf{PP}^*} \geq \alpha\lambda\,.$$

Remark 1.17. We now show that, as mentioned earlier, the two approaches to bounding mixing in this section are equivalent.

$$\begin{aligned} 1 - \lambda_{\mathsf{PP}^*} &= \sup_{f:\Omega\to\mathsf{R}} \frac{\mathrm{Var}(f) - \mathcal{E}_{\mathsf{PP}^*}(f,f)}{\mathrm{Var}(f)} \\ &= \sup_{f:\Omega\to\mathsf{R},\,\mathbb{E}f=0} \frac{\langle f,f\rangle_\pi - \langle f,(I-\mathsf{PP}^*)f\rangle_\pi}{\langle f,f\rangle_\pi} \\ &= \sup_{f:\Omega\to\mathsf{R},\,\mathbb{E}f=0} \frac{\langle \mathsf{P}^*f,\mathsf{P}^*f\rangle_\pi}{\langle f,f\rangle_\pi} = \|\mathsf{P}^*\|^2\,. \end{aligned}$$

The second supremum is equal to the first because the numerator and denominator in the first are invariant under addition of a constant to f, so it may be assumed that $\mathbb{E}f = 0$.

Our concluding remark will require knowledge of the $L^p \to L^q$ operator norm:

Definition 1.18. Suppose $T : \mathsf{R}^\Omega \to \mathsf{R}^\Omega$ is an operator taking functions $f : \Omega \to \mathsf{R}$ to other such functions. Then, given $p, q \in [1, \infty]$, let $\|T\|_{p\to q}$ be the optimal constant in the inequality

$$\|Tf\|_q \leq \|T\|_{p\to q}\, \|f\|_p\,, \quad \text{for all } f : \Omega \to \mathsf{R}.$$

Remark 1.19. It has already been seen that $\|\mathsf{P}^*\|^2 = 1-\lambda_{\mathsf{PP}^*}$. We now consider a few other equivalent forms which have appeared in mixing bounds equivalent to Proposition 1.12.

First, consider the operator norm. Let E denote the expectation operator, that is, E is a square matrix with rows all equal to π. Then $\mathsf{P}^*E = E\mathsf{P}^* = E^2 = E$ and $\|f\|_2 \geq \min_{c\in\mathsf{R}} \|f - c\|_2 = \|f - Ef\|_2$, and so

$$\frac{\|(\mathsf{P}^* - E)f\|_2}{\|f\|_2} \leq \frac{\|\mathsf{P}^*(f - Ef)\|_2}{\|f - Ef\|_2} \leq \|\mathsf{P}^*\|\,.$$

In particular, $\|\mathsf{P}^* - E\|_{2\to 2} \leq \|\mathsf{P}^*\|$. Conversely, if $\mathbb{E}f = 0$ then $(\mathsf{P}^* - E)f = \mathsf{P}^*f$, and so $\|\mathsf{P}^*\| \leq \|\mathsf{P}^* - E\|_{2\to 2}$. It follows that $\|\mathsf{P}^* - E\|_{2\to 2} = \|\mathsf{P}^*\|$.

It may seem more intuitive to work with P instead of P^*. In fact, both cases are the same.

$$\begin{aligned}
\|\mathsf{P}^* - E\|_{2\to 2} &= \sup_{\|f\|_2=1} \|(\mathsf{P}^* - E)f\|_2 \\
&= \sup_{\|f\|_2=1} \sup_{\|g\|_2=1} |\langle(\mathsf{P}^* - E)f, g\rangle_\pi| \\
&= \sup_{\|f\|_2=1} \sup_{\|g\|_2=1} |\langle f, (\mathsf{P} - E)g\rangle_\pi| \\
&= \sup_{\|g\|_2=1} \sup_{\|f\|_2=1} |\langle(\mathsf{P} - E)g, f\rangle_\pi| \\
&= \|\mathsf{P} - E\|_{2\to 2}.
\end{aligned}$$

The second equality is just L^p duality, $\|f\|_p = \sup_{\|g\|_q=1} |\langle f, g\rangle_\pi|$ when $1/p + 1/q = 1$ (which is just an extension of the dot product property that $\|f\|_2 = f \cdot \frac{f}{\|f\|_2} = \max_{\|g\|_2=1} f \cdot g$).

Some authors have worked with complex valued functions. Note that if $f : \Omega \to \mathbb{C}$ and T is a real valued square matrix then

$$\begin{aligned}
\|Tf\|_2^2 &= \|T(\mathrm{Re}f)\|_2^2 + \|T(\mathrm{Im}f)\|_2^2 \\
&\leq \|T\|_{2\to 2}^2 \left(\|\mathrm{Re}f\|_2^2 + \|\mathrm{Im}f\|_2^2\right) \\
&= \|T\|_{2\to 2}^2 \|f\|_2^2\,,
\end{aligned}$$

and so $\|T\|_{2\to 2}$ would be the same, even if defined over complex valued functions.

In summary,

$$\begin{aligned}\|\mathsf{P}^*\| &= \|\mathsf{P}^* - E\|_{2\to 2} = \|\mathsf{P} - E\|_{2\to 2} = \|\mathsf{P}\| \\ &= \sup_{f:\Omega\to\mathbb{C}} \frac{\|(\mathsf{P}-E)f\|_2}{\|f\|_2} = \sup_{f:\Omega\to\mathbb{C},\,\mathbb{E}f=0} \frac{\|\mathsf{P}f\|_2}{\|f\|_2}\,.\end{aligned}$$

1.4 Does Reversibility Matter?

Many mixing results were originally shown only in the context of a reversible Markov chain. In this survey we are able to avoid this requirement in most cases. However, there are still circumstances under which reversible and non-reversible chains behave differently, and not just as an artifact of the analysis. In this section we discuss these differences, and also prove a few classical lemmas about reversible chains which were used in the discrete time results given above, and which explain why the reversibility assumption is helpful.

The difference between reversible and non-reversible results is most apparent when upper and lower bounds on distances are given. Let

$$d(n) = \max_x \|\mathsf{P}^n(x,\cdot) - \pi(\cdot)\|_{\mathrm{TV}}$$

denote the worst variation distance after n steps. Then, combining the above work, the lower bounds of Theorem 4.9, and recalling that $\lambda_{max} = \max\{\lambda_1, |\lambda_{n-1}|\}$, we have

$$\begin{array}{l} \textit{if } \mathsf{P} \textit{ is reversible}: \\ \qquad\qquad \frac{1}{2}\lambda_{max}^n \quad\le\quad d(n) \quad\le\quad \frac{1}{2}\lambda_{max}^n\sqrt{\frac{1-\pi_*}{\pi_*}} \\ \textit{if } \mathsf{P} \textit{ is non-reversible}: \\ \qquad \frac{1}{2}\max_{i>0}|\lambda_i|^n \quad\le\quad d(n) \quad\le\quad \frac{1}{2}\left(\sqrt{1-\lambda_{\mathsf{PP}^*}}\right)^n\sqrt{\frac{1-\pi_*}{\pi_*}} \end{array}$$

In particular, in the reversible case the variation distance is determined, up to a multiplicative factor, by the size of the largest magnitude eigenvalue. The rapid mixing property is then entirely characterized by whether $1-\lambda_{max}$ is polynomially large or not.

In contrast, Example 5.2 gives a convergent non-reversible chain with complex eigenvalues such that $\max_{\lambda_i\neq 1}|\lambda_i| = 1/\sqrt{2}$ and $\lambda_{\mathsf{PP}^*} = 0$.

The non-reversible lower bound given above then converges to zero with n, as it should, while the upper bound is constant and useless.

This is not to say that the situation is hopeless in the non-reversible case. If the chain is lazy then it will be found that

$$\frac{1-2\epsilon}{\tilde{\Phi}} \leq \tau(\epsilon) \leq \frac{2}{\tilde{\Phi}^2} \log \frac{1}{\epsilon\sqrt{\pi_*}}\,,$$

where the conductance is given by

$$\tilde{\Phi} = \min_{A \subset \Omega} \frac{\mathsf{Q}(A, A^c)}{\pi(A)\pi(A^c)}\,.$$

It follows that for a lazy, non-reversible chain, the mixing time is determined, up to squaring and a multiplicative factor, by the conductance $\tilde{\Phi}$, a much weaker relation than was available in the reversible case, and with no guarantees possible if the minimal holding probability α is 0. Nevertheless, both upper and lower bounds are necessary. For instance, Example 5.5 considers two lazy walks, both on a pair of cycles, and both with identical conductance and spectral gap, $\tilde{\Phi} = 1/100n$ and $\lambda = O(1/n^2)$, and yet the reversible walk mixes in $\Theta(n^2)$, while the non-reversible walk mixes in $\Theta(n)$.

Another case in which reversibility will play a key role is comparison of mixing times. This is a method by which the mixing time of a Markov chain P can be bounded by instead studying a similar, but easier to analyze chain $\hat{\mathsf{P}}$. For many Markov chains this is the only way known to bound mixing time. In Theorem 4.17 we find that good comparison is possible if P and $\hat{\mathsf{P}}$ are reversible, while if P is non-reversible then there is a slight worsening, but if $\hat{\mathsf{P}}$ is non-reversible then the comparison is much worse. Example 5.4 shows that even for walks as simple as those on a cycle $\mathbb{Z}/n\mathbb{Z}$ each of these three cases is necessary, and not just an artifact of the method of proof.

The main reason why a reversible Markov chain is better behaved is that it has a complete real valued spectral decomposition, and because the spectral gap λ is exactly related to eigenvalues of the reversible chain. For the sake of completeness, and because they are occasionally used in this survey, we now show these classical properties.

The Perron-Frobenius theorem states that a reversible Markov chain has a complete spectral decomposition into real valued eigenvalues and eigenvectors, and that these have magnitude at most 1.

Lemma 1.20. If P is reversible and irreducible on state space of size $|\Omega| = n$, then it has a complete spectrum of real eigenvalues with magnitudes at most one, that is

$$1 = \lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{n-1} \geq -1\,.$$

A non-reversible chain may have complex valued eigenvalues, as in Example 5.2.

Proof. Let$(\sqrt{\pi}) = diag(\sqrt{\pi(1)}, \sqrt{\pi(2)}, \ldots, \sqrt{\pi(n)})$ denote the diagonal matrix with entries drawn from $\sqrt{\pi(\cdot)}$. The matrix $M = (\sqrt{\pi})\,\mathsf{P}(\sqrt{\pi})^{-1}$ is a symmetric matrix because

$$M(x,y) = \sqrt{\frac{\pi(x)}{\pi(y)}}\mathsf{P}(x,y) = \frac{\pi(x)\mathsf{P}(x,y)}{\sqrt{\pi(x)\pi(y)}} = \frac{\pi(y)\mathsf{P}(y,x)}{\sqrt{\pi(x)\pi(y)}} = M(y,x)\,.$$

It follows from the spectral theorem that since P is similar to a symmetric real matrix then it has a real valued eigenbasis.

In this eigenbasis, suppose v is a left eigenvector w a right eigenvector, with corresponding eigenvalues λ_v and λ_w. Then,

$$\lambda_v\, v\, w = (v\mathsf{P})w = v(\mathsf{P}w) = \lambda_w v\, w\,. \tag{1.13}$$

In particular, if $\lambda_v \neq \lambda_w$ then v and w are orthogonal. A special case of this is the eigenvalue 1 with right eigenvector 1, as then if eigenvalue $\lambda_i \neq 1$ has left eigenvector v_i then $\sum_x v_i(x) = v1 = 0$. Hence, for ϵ sufficiently small $\sigma = \pi + \epsilon\, v_i$ is a probability distribution. However, the n step distribution is given by

$$\sigma\mathsf{P}^n = (\pi + \epsilon\, v_i)\mathsf{P}^n = \pi + \epsilon\lambda_i^n\, v_i\,,$$

and since $\sum_x v_i(x) = 0$ then v_i has a negative entry, and so if $|\lambda_i| > 1$ then $\sigma\mathsf{P}^n$ will have a negative entry for sufficiently large n, contradicting the fact that $\sigma\mathsf{P}^n$ is a probability distribution. □

The Courant-Fischer theorem shows the connection between eigenvalues and Dirichlet forms for a reversible Markov chain.

Lemma 1.21. In a reversible Markov chain the second largest eigenvalue λ_1, and smallest eigenvalue λ_{n-1} satisfy

$$\begin{aligned} 1-\lambda_1 &= \inf_{\mathrm{Var}(f)\neq 0} \frac{\mathcal{E}(f,f)}{\mathrm{Var}(f)} = \lambda\,, \\ 1+\lambda_{n-1} &= \inf_{\mathrm{Var}(f)\neq 0} \frac{\mathcal{F}(f,f)}{\mathrm{Var}(f)}\,, \end{aligned}$$

where

$$\mathcal{F}(f,f) = \langle f, (\mathsf{I}+\mathsf{P})f\rangle_\pi = \frac{1}{2}\sum_{x,y\in\Omega} (f(x)+f(y))^2 \mathsf{P}(x,y)\pi(x)\,.$$

In Section 5.3 we will find that the relation $1-\lambda_1=\lambda$ becomes an inequality $1-\mathrm{Re}\lambda_i \geq \lambda$ in the non-reversible case.

Proof. The numerator and denominator in the infinum are invariant under adding a constant to f, so it may be assumed that $\mathbb{E}f=0$, that is, $\langle f,1\rangle_\pi = 0$.

Let $\{v_i\}$ be a set of right eigenvectors of P forming an orthonormal eigenbasis for R^Ω, with $v_0=1$. Given $f:\,\Omega\to\mathsf{R}$ then $f=\sum c_i\,v_i$ with $c_i=\langle f,v_i\rangle_\pi$, and so

$$\begin{aligned} \mathcal{E}(f,f) &= \langle f,(\mathsf{I}-\mathsf{P})f\rangle_\pi = \sum_{i,j\in\Omega} c_i c_j \langle v_i,(\mathsf{I}-\mathsf{P})v_j\rangle_\pi \\ &= \sum_{i\in\Omega} c_i^2(1-\lambda_i) \geq \sum_{i\in\Omega} c_i^2(1-\lambda_1) \end{aligned}$$

with an equality when $f=v_1$. Also,

$$\begin{aligned} \mathrm{Var}(f) &= \langle f,f\rangle_\pi - \langle f,1\rangle_\pi \\ &= \left\langle \sum_{i\in\Omega} c_i v_i, \sum_{j\in\Omega} c_j v_j \right\rangle_\pi - \left\langle \sum_{i\in\Omega} c_i v_i, 1 \right\rangle_\pi \\ &= \sum_{i\in\Omega} c_i^2 \end{aligned}$$

and so

$$\frac{\mathcal{E}(f,f)}{\mathrm{Var}(f)} \geq 1-\lambda_1$$

with an equality when $f = v_1$. The result then follows.

The same argument, but with $\mathsf{I} + \mathsf{P}$ instead of $\mathsf{I} - \mathsf{P}$ gives the result for λ_{n-1}. □

2

Advanced Functional Techniques

The relation between functional constants and mixing time bounds was studied in Chapter 1. In this section it is shown that information on functions of large variance, or on functions with small support, can be exploited to show better mixing time bounds.

The argument is simple. Recall that $\frac{d}{dt}\mathrm{Var}(h_t) = -2\mathcal{E}(h_t, h_t)$. If $\mathcal{E}(f, f) \geq G(\mathrm{Var}(f))$ for some $G : \mathsf{R}_+ \to \mathsf{R}_+$ and $f : \Omega \to \mathsf{R}_+$ with $\mathbb{E}f = 1$, then it follows that $\frac{d}{dt}\mathrm{Var}(h_t) = -2\mathcal{E}(h_t, h_t) \leq -2\,G(\mathrm{Var}(h_t))$. With a change of variables to $I = \mathrm{Var}(h_t)$, this becomes $\frac{dI}{dt} \leq -2\,G(I)$, and it follows that

$$\tau_2(\epsilon) = \int_0^{\tau_2(\epsilon)} 1\,dt \leq \int_{\mathrm{Var}(h_0)}^{\epsilon^2} \frac{dI}{-2G(I)}\,. \tag{2.1}$$

If one makes the obvious choice of $G(r) = \lambda r$, then this is just the bound of the previous chapter. More generally, in this chapter we derive such functions G in terms of the log-Sobolev constant, Nash inequalities, spectral profile, or via comparison to another Markov chain.

With minimal modifications the argument applies in discrete-time as well. First, replace $\frac{d}{dt}\mathrm{Var}(h_t) = -2\mathcal{E}(h_t, h_t)$ with $\mathrm{Var}(k_n) - \mathrm{Var}(k_{n-1}) = -\mathcal{E}_{\mathsf{PP}^*}(k_n, k_n)$. Then, if $\mathcal{E}_{\mathsf{PP}^*}(f, f) \geq G_{\mathsf{PP}^*}(\mathrm{Var}(f))$, and both $I(n) = \mathrm{Var}(k_n)$ and $G_{\mathsf{PP}^*}(r)$ are non-decreasing, the piecewise linear extension of $I(n)$ to $t \in \mathsf{R}_+$ will satisfy

$$\frac{dI}{dt} \leq -G_{\mathsf{PP}^*}(I)\,.$$

At integer t, the derivative can be taken from either right or left. It follows that

$$\tau_2(\epsilon) = \int_0^{\tau_2(\epsilon)} 1\,dt \leq \left\lceil \int_{\mathrm{Var}(h_0)}^{\epsilon^2} \frac{dI}{-G_{\mathsf{PP}^*}(I)} \right\rceil. \tag{2.2}$$

In terms of $G(r)$, by Equation (1.12),

$$\mathcal{E}_{\mathsf{PP}^*}(f,f) \geq 2\alpha\mathcal{E}(f,f) \geq 2\alpha G(\mathrm{Var}(f))\,,$$

and so we may take $G_{\mathsf{PP}^*}(r) = 2\alpha G(r)$.

2.1 Log-Sobolev and Nash Inequalities

Some of the best bounds on L^2 mixing times were shown by use of the log-Sobolev constant (see Definition 1.9), a method developed in the finite Markov chain setting by Diaconis and Saloff-Coste [25]. One example of this is a walk on a class of matroids.

Example 2.1. A matroid $\mathcal{M}$ is given by a ground set $E(\mathcal{M})$ with $|E(\mathcal{M})| = n$ and a collection of bases $\mathcal{B}(\mathcal{M}) \subseteq 2^{E(\mathcal{M})}$. The bases $\mathcal{B}(\mathcal{M})$ must all have the same cardinality r, and $\forall X, Y \in \mathcal{B}(\mathcal{M})$, $\forall e \in X$, $\exists f \in Y : X \cup \{f\} \setminus \{e\} \in \mathcal{B}(\mathcal{M})$. One choice of a Markov chain on matroids is, given state $X \in \mathcal{B}(\mathcal{M})$ half the time do nothing, and otherwise choose $e \in X$, $f \in E(\mathcal{M})$ and transition to state $X - e + f$ if this is also a basis.

Jerrum and Son [42] found that $\lambda = 1/rn$ for this walk on the class of matroids known as balanced matroids. There are at most $C(n,r) \leq n^r$ bases, and so Corollary 1.15 implies mixing in time $\tau_2(\epsilon) = O\left(rn(r\log n + \log(1/\epsilon))\right)$. However, initially the density k_n has high variance, and so if the spectral gap gives an overly pessimistic lower bound on the Dirichlet form $\mathcal{E}(f,f)$ when f has high variance then Equation (2.2) suggests the mixing time may be faster. This is in fact the case, as the authors of [42] find by use of the method of log-Sobolev inequalities, which we develop below, to sharpen this bound to $\tau_2(\epsilon) = O\left(rn(\log r + \log\log n + \log(1/\epsilon))\right)$.

Equations (2.1) and (2.2) will bound mixing time in terms of the log-Sobolev constant if we can show a relation between the Dirichlet form $\mathcal{E}(f,f)$ and a function of the variance $\mathrm{Var}(f)$. The following lemma establishes this connection.

Lemma 2.2. If f is non-negative then

$$\mathrm{Ent}(f^2) \geq \mathbb{E}f^2 \log \frac{\mathbb{E}f^2}{(\mathbb{E}f)^2},$$

and in particular, if $\mathbb{E}f = 1$ then

$$\mathcal{E}(f,f) \geq \rho \mathrm{Ent}(f^2) \geq \rho\,(1 + \mathrm{Var}(f)) \log(1 + \mathrm{Var}(f)).$$

Proof. By definition

$$\mathrm{Ent}(f^2) = \mathbb{E}f^2 \log \frac{f^2}{\mathbb{E}f^2} = 2\mathbb{E}f^2 \log \frac{f\mathbb{E}f}{\mathbb{E}f^2} + \mathbb{E}f^2 \log \frac{\mathbb{E}f^2}{(\mathbb{E}f)^2}$$

The first term drops out by applying the approximation $\log x \geq 1 - 1/x$.

$$\mathbb{E}f^2 \log \frac{f\mathbb{E}f}{\mathbb{E}f^2} \geq \mathbb{E}f^2 \left(1 - \frac{\mathbb{E}f^2}{f\mathbb{E}f}\right) = \mathbb{E}f^2 - \mathbb{E}f^2 \frac{\mathbb{E}f}{\mathbb{E}f} = 0$$

□

Those familiar with cross-entropy might prefer to rewrite this proof as follows. Noting that the cross-entropy $H(f,g) = \mathbb{E}f \log \frac{f}{g} \geq 0$ for densities f, g, the proof is just the statement

$$\mathrm{Ent}(f^2) = 2\mathbb{E}f^2\, H\left(\frac{f^2}{\mathbb{E}f^2}, \frac{f}{\mathbb{E}f}\right) + \mathbb{E}f^2 \log \frac{\mathbb{E}f^2}{(\mathbb{E}f)^2} \geq \mathbb{E}f^2 \log \frac{\mathbb{E}f^2}{(\mathbb{E}f)^2}.$$

Diaconis and Saloff-Coste also showed that the Nash-inequality can be used to study L^2 mixing [26]. A Nash-inequality is a tool often used to show that when the variance of the density is extremely high then the walk converges even faster than predicted by the log-Sobolev constant.

This often helps get rid of small terms such as a double logarithm. For instance, in the case of balanced matroids considered in Example 2.1 one might hope for a Nash inequality to improve the log-Sobolev mixing time bound to $\tau_2(\epsilon) = O\left(rn(\log r + \log(1/\epsilon))\right)$. In Examples 2.7 and 2.13 we consider the well known exclusion process on the grid $\mathbb{Z}^d/L\mathbb{Z}^d$, and for this walk it might be possible for a Nash inequality to supplement the log-Sobolev result and improve the bound by an even larger $O(d/\log d)$. Unfortunately, however, Nash inequalities are notoriously difficult to establish, and in neither of these two cases has the necessary Nash inequality been established.

We now show that the Dirichlet form can also be lower bounded in terms of variance by using a Nash inequality.

Lemma 2.3. Given a Nash Inequality

$$\|f\|_2^{2+1/D} \le C\left[\mathcal{E}(f,f) + \frac{1}{T}\|f\|_2^2\right]\|f\|_1^{1/D}$$

which holds for every function $f : \Omega \to \mathsf{R}$ and some constants C, D, $T \in \mathsf{R}_+$, then whenever $f \ge 0$ and $\mathbb{E}f = 1$ then

$$\mathcal{E}(f,f) \ge (1+\mathrm{Var}(f))\left(\frac{(1+\mathrm{Var}(f))^{1/D}}{C} - \frac{1}{T}\right)$$

Proof. The Nash inequality can be rewritten as

$$\mathcal{E}(f,f) \ge \|f\|_2^2\left(\frac{1}{C}\left(\frac{\|f\|_2}{\|f\|_1}\right)^{1/D} - \frac{1}{T}\right)$$

However, $\|f\|_1 = \mathbb{E}|f| = 1$, and $\mathrm{Var}(f) = \|f\|_2^2 - 1$, giving the result. □

Mixing time bounds follow immediately from Equations 2.1 and 2.2.

Corollary 2.4. Given the spectral gap λ and the log-Sobolev constant ρ and/or a Nash inequality with $DC \ge T$ and $D \ge 2$, and given $\epsilon \le 2$, then the continuous time Markov chain satisfies

$$\tau_2(\epsilon) \;\le\; \frac{1}{2\rho}\log\log\frac{1}{\pi_*} + \frac{1}{\lambda}\left(\frac{1}{4} + \log\frac{1}{\epsilon}\right)$$

$$
\begin{aligned}
\tau_2(\epsilon) &\leq T + \frac{1}{\lambda}\left(\frac{D}{2}\log\frac{DC}{T} + \log\frac{1}{\epsilon}\right) \\
\tau_2(\epsilon) &\leq T + \frac{1}{2\rho}\log\log\left(\frac{DC}{T}\right)^D + \frac{1}{\lambda}\left(\frac{1}{4} + \log\frac{1}{\epsilon}\right)
\end{aligned}
$$

Upper bounds for the discrete time Markov chain are a factor of 2 larger when Nash, log-Sobolev and spectral gap are computed in terms of the chain PP^*, while when computed for P with holding probability α then they are a factor α^{-1} larger.

Proof. Apply Equation (2.1) with the log-Sobolev bound of Lemma 2.2 when $\mathrm{Var}(h_t) \geq 4$, and the spectral gap bound $\mathcal{E}(f,f) \geq \lambda \mathrm{Var}(f)$ when $\mathrm{Var}(h_t) < 4$, to obtain

$$
\begin{aligned}
\tau_2(\epsilon) &\leq \int_{\mathrm{Var}(h_0)}^{4} \frac{dI}{-2\rho(1+I)\log(1+I)} + \int_4^{\epsilon^2} \frac{dI}{-2\lambda\, I} \\
&= \frac{1}{-2\rho}\left(\log\log(1+4) - \log\log(1+\mathrm{Var}(h_0))\right) + \frac{1}{-2\lambda}\log\frac{\epsilon^2}{4}
\end{aligned}
$$

Simplify this by $\mathrm{Var}(h_0) \leq \frac{1-\pi_*}{\pi_*}$, and apply $\rho \leq \lambda/2$ to the $\log\log(5)$ term.

For the second mixing bound use the Nash bound of Lemma 2.3 when $\mathrm{Var}(h_t) \geq (DC/T)^D - 1$, and the spectral bound when $\mathrm{Var}(h_t) < (DC/T)^D - 1$. The Nash portion of the integral is then

$$
\begin{aligned}
&\int_{\mathrm{Var}(h_0)}^{(DC/T)^D-1} \frac{dr}{-2\,(1+I)\left(\frac{(1+I)^{1/D}}{C} - \frac{1}{T}\right)} \\
&= -\frac{DT}{2}\log\left(1 - \frac{C/T}{(1+I)^{1/D}}\right)\Bigg|_{\mathrm{Var}(h_0)}^{(DC/T)^D-1} \\
&\leq -\frac{DT}{2}\log(1-1/D) \leq \frac{DT}{2(D-1)} \leq T
\end{aligned}
$$

The second inequality is because $\log(1-1/x) \geq -\frac{1}{x-1}$.

For the final mixing bound, use the Nash inequality when $\mathrm{Var}(h_t) \geq (DC/T)^D - 1$, the log-Sobolev bound for $(DC/T)^D - 1 > \mathrm{Var}(h_t) \geq 4$ and the spectral bound when $\mathrm{Var}(h_t) < 4$.

For the discrete time case proceed similarly, but with equation (2.2) instead of Equation (2.1). □

The continuous time log-Sobolev bound is comparable to a result of Diaconis and Saloff-Coste [25], while the discrete time log-Sobolev bound is comparable to a bound of Miclo [54].

Hypercontractivity ideas can be used to improve these results slightly for a reversible, continuous time chain. For such chains, given $t_0 \in \mathsf{R}$ then

$$\frac{d}{dt}\,\|h_t\|_{1+e^{4\rho(t-t_0)}} \leq 0\,. \tag{2.3}$$

This follows from a tedious differentiation and a few approximations (see around Equation (3.2) of [25]).

Now, let $t_0 = k+\frac{1}{4\rho}\log\left[\log(1+\mathrm{Var}(h_k))-1\right]$ for some fixed $k \in \mathsf{R}_+$. Since $t_0 \geq k$ then by Equation 2.3,

$$\|h_{t_0}\|_2 \leq \|h_k\|_{1+e^{-4\rho t_0}} \leq \|h_k\|_1^{1-\frac{2}{\log(1+\mathrm{Var}(h_k))}}\;\|h_k\|_2^{\frac{2}{\log(1+\mathrm{Var}(h_k))}} = e\,.$$

The second inequality was the relation $\|f\|_{q^*} \leq \|f\|_1^{1-2/q}\|f\|_2^{2/q}$ when $q \geq 2$ and $1/q+1/q^* = 1$; see Chapter 8, Lemma 41 of [1]. The equality is because $\|h_k\|_1 = 1$ and $\|h_k\|_2^2 = 1+\mathrm{Var}(h_k)$. It follows that for any $k \geq 0$ that $\mathrm{Var}(h_{t_0}) = \|h_{t_0}\|_2^2 - 1 \leq e^2-1$.

Combining this bound on $\mathrm{Var}(h_{t_0})$ with our earlier Nash and spectral work we obtain

$$\begin{aligned}
\tau_2(\epsilon) \;\; &\leq \;\; \int_{\mathrm{Var}(h_0)}^{(DC/T)^D-1} \frac{dr}{-2\,(1+I)\left(\frac{(1+I)^{1/D}}{C}-\frac{1}{T}\right)} \\
&\qquad + \frac{1}{4\rho}\log\left[\log(DC/T)^D-1\right] + \int_{e^2-1}^{\epsilon^2}\frac{dI}{-2\lambda\,I} \\
&< \;\; T+\frac{1}{4\rho}\log\log\left(\frac{DC}{T}\right)^D + \frac{1}{\lambda}\left(1+\log\frac{1}{\epsilon}\right)\,.
\end{aligned}$$

The factor of two difference between reversible and non-reversible cases is common, as in Corollary 1.14 and Remark 1.11. This seems to be a consequence of the Dirichlet property $\mathcal{E}_{\mathsf{P}}(f,f) = \mathcal{E}_{\mathsf{P}^*}(f,f) = \mathcal{E}_{\frac{\mathsf{P}+\mathsf{P}^*}{2}}(f,f)$, from which it follows that ρ and λ are the same for the non-reversible chain P and the reversible chain $\frac{\mathsf{P}+\mathsf{P}^*}{2}$.

2.2 Spectral profile

In the previous section it was found that log-Sobolev bounds on mixing time can improve on spectral gap results, by replacing the $\log(1/\pi_*)$ term with $\log\log(1/\pi_*)$. However, the log-Sobolev constant is much more difficult to bound than the spectral gap and, to date, bounds on it are known for only a handful of problems. Moreover, sometimes even log-Sobolev is not strong enough. In this section we develop the method of Spectral Profile, the idea of which is to improve on the elementary relation $\mathcal{E}(f,f) \geq \lambda \mathrm{Var}(f)$, i.e. $G(r) \geq \lambda r$, and instead make a relation depending on the size of the support of f. This improves on log-Sobolev bounds, while generalizing both spectral gap and conductance methods (defined in the next chapter).

Example 2.5. Consider a lazy simple random walk on the cycle $\mathbb{Z}/m\mathbb{Z}$, with $\mathsf{P}(i,i+1) = \mathsf{P}(i,i-1) = 1/4$ and $\mathsf{P}(i,i) = 1/2$. Then $\lambda, \rho = O(1/m^2)$, by taking $f = 1_{\{0\ldots m/2\}}$ in the definitions of λ and ρ. Then, Corollary 1.15 and 2.4 show at best

$$\tau_2(1/2e) = O(m^2 \log m) \quad \text{and} \quad \tau_2(1/2e) = O(m^2 \log\log m)$$

respectively. The correct bound is $\tau_2(1/2e) = \Theta(m^2)$, as will be established in Example 2.11. The spectral and log-Sobolev bounds were incorrect because during it's early phases the walk rapidly reaches new vertices, but fewer and fewer new ones as it continues – order $\sqrt{n}$ vertices have been reached after n steps. Hence, a good mixing time bound should distinguish between early in the walk, when the probability density is highly concentrated at a few vertices, and later when it is more spread around the space.

Example 2.6. The Thorp shuffle is a model for card shuffling for which mixing bounds have been very hard to come by. Morris [64] recently used the method of Evolving Sets (see next chapter) to give the first polynomial bound in d on the mixing time of this shuffle for a 2^d-card deck. In Section 5.4.2 we discover that his method can be used to lower bound $\frac{\mathcal{E}(f,f)}{\mathrm{Var}(f)}$ in terms of the size of the support of f, leading to an improved mixing result with Theorem 2.10.

Example 2.7. Consider an N vertex graph G, with edge set E, and constant R such that $1 \leq R \leq N/2$. Place R particles on the vertices of the graph, with no two particles on the same vertex. Then, a step of the R-particle Bernoulli-Laplace walk consists of choosing a particle and an unoccupied location, both uniformly at random, and then moving the particle to the new location. As far back as 1987 Diaconis and Shahshahani [16] showed that $\lambda = \frac{N}{R(N-R)}$. It is only in 1998 that Lee and Yau [49] determined the log-Sobolev constant, by showing:

$$\rho \geq \frac{\log 2}{2 \log \frac{N^2}{R(N-R)}} \,.$$

Perhaps a spectral gap type argument depending on the size of the support of f would have avoided the need to compute the log-Sobolev constant? More importantly, even the log-Sobolev constant gives too pessimistic a lower bound on $\mathcal{E}(f,f)$ for functions of small support (see Example 2.13), and so a more general method may make it possible to prove sharper bounds.

We now proceed to more concrete details. Faber-Krahn inequalities were developed by Grigor'yan, Coulhon and Pittet [18] (see also [35] and [19]) to study the rate of decay of the heat kernel, and in the finite Markov setting by Goel, Montenegro and Tetali [34]. As mentioned earlier, the argument is based on improving on the elementary relation $\mathcal{E}(f,f) \geq \lambda \text{Var}(f)$, and instead making a relation depending on the size of the support of f.

Definition 2.8. For a non-empty subset $S \subset \Omega$ the *first Dirichlet eigenvalue* on S is given by

$$\lambda_1(S) = \inf_{f \in c_0^+(S)} \frac{\mathcal{E}(f,f)}{\text{Var}(f)}$$

where $c_0^+(S) = \{f \geq 0 : \ \mathit{supp}(f) \subset S\}$ is the set of non-negative functions supported on S. The *spectral profile* $\Lambda : \ [\pi_*, \infty) \to \mathsf{R}$ is given by $\Lambda(r) = \inf_{\pi_* \leq \pi(S) \leq r} \lambda_1(S)$.

The spectral profile is a natural extension of spectral gap λ, and we will now see that it can be used to improve on the basic bound $\mathcal{E}(f,f) \geq \lambda \text{Var}(f)$ used earlier.

Lemma 2.9. For every non-constant function $f : \Omega \to \mathsf{R}_+$,

$$\mathcal{E}(f,f) \geq \frac{1}{2}\,\Lambda\left(\frac{4(\mathbb{E}f)^2}{\operatorname{Var} f}\right)\operatorname{Var}(f)\,.$$

Proof. Given $a \in \mathsf{R}$ use the notation $a_+ = \max\{a,0\}$ to denote the positive part. For c constant, $\mathcal{E}(f,f) = \mathcal{E}(f-c, f-c)$. Also, $\mathcal{E}(f-c, f-c) \geq \mathcal{E}((f-c)_+, (f-c)_+)$ because $\forall a,b \in \mathsf{R} :\ (a-b)^2 \geq (a_+ - b_+)^2$. It follows that when $0 \leq c < \max f$ then

$$\begin{aligned}
\mathcal{E}(f,f) &\geq \mathcal{E}((f-c)_+, (f-c)_+) \\
&\geq \operatorname{Var}((f-c)_+) \inf_{u \in c_0^+(f>c)} \frac{\mathcal{E}(u,u)}{\operatorname{Var}(u)} \\
&\geq \operatorname{Var}((f-c)_+)\,\Lambda(\pi(f>c))\,.
\end{aligned}$$

The inequalities $\forall a, b \geq 0 :\ (a-b)_+^2 \geq a^2 - 2b\,a$ and $(a-b)_+ \leq a$ show that

$$\operatorname{Var}((f-c)_+) = \mathbb{E}(f-c)_+^2 - (\mathbb{E}(f-c)_+)^2 \geq \mathbb{E}f^2 - 2c\,\mathbb{E}f - (\mathbb{E}f)^2.$$

Let $c = \operatorname{Var}(f)/4\mathbb{E}f$ and apply Markov's inequality $\pi(f>c) < (\mathbb{E}f)/c$,

$$\mathcal{E}(f,f) \geq (\operatorname{Var}(f) - 2c\,\mathbb{E}f)\,\Lambda(\mathbb{E}f/c) = \frac{1}{2}\operatorname{Var}(f)\,\Lambda\left(\frac{4(\mathbb{E}f)^2}{\operatorname{Var} f}\right)$$

□

A mixing time theorem then follows easily.

Theorem 2.10. Consider a Markov chain P, initial distribution σ and holding probability α. In continuous time,

$$\tau_2(\epsilon) \leq \int_{4/\operatorname{Var}(\sigma/\pi)}^{1/2} \frac{dr}{r\,\Lambda(r)} + \frac{1}{\lambda}\log\frac{2\sqrt{2}}{\epsilon}\,.$$

In discrete time,

$$\begin{aligned}
\tau_2(\epsilon) &\leq \left\lceil \int_{4/\operatorname{Var}(\sigma/\pi)}^{1/2} \frac{2\,dr}{r\,\Lambda_{\mathsf{PP}^*}(r)} + \frac{2}{\lambda_{\mathsf{PP}^*}}\log\frac{2\sqrt{2}}{\epsilon} \right\rceil \\
&\leq \left\lceil \int_{4/\operatorname{Var}(\sigma/\pi)}^{1/2} \frac{dr}{\alpha\,r\,\Lambda(r)} + \frac{1}{\alpha\lambda}\log\frac{2\sqrt{2}}{\epsilon} \right\rceil\,.
\end{aligned}$$

Moreover, $\operatorname{Var}(\sigma/\pi) \leq \frac{1}{\pi_*} - 1 < \frac{1}{\pi_*}$.

Proof. In continuous time let $h_0(x) = \frac{\sigma(x)}{\pi(x)}$, and apply Equation (2.1) with the spectral profile bound of Lemma 2.9 to obtain

$$\tau_2(\epsilon) \leq \int_{\mathrm{Var}(h_0)}^{8} \frac{dI}{-I\,\Lambda(4/I)} + \int_8^{\epsilon^2} \frac{dI}{-2\lambda\, I}\,.$$

A change of variables to $r = 4/I(t)$ gives the mixing time bound.

In discrete time use equation (2.2) instead of (2.1). Then simplify with the relation $\Lambda_{\mathsf{PP}^*}(r) \geq 2\alpha\Lambda(r)$, an immediate consequence of Equation 1.12. □

The theorem, with the trivial bound $\Lambda(r) \geq \lambda$, produces bounds about a factor of two worse than those of Corollaries 1.6, 1.14 and 1.15. However, there can be a significant improvement if $\Lambda(r) \gg \lambda$ for small values of r.

As an elementary example, let us give our first proof of both upper and lower bounds on a mixing time.

Example 2.11. It was established at the beginning of this section that spectral gap and the log-Sobolev constant cannot be used to establish a sharp mixing time bound for a walk on the cycle $\mathbb{Z}/m\mathbb{Z}$. Instead we use the spectral profile. The conductance profile is a geometric analog of spectral profile, and by equation (3.1) it can be used to lower bound spectral profile. Given $r \in [0,1]$ the conductance profile is minimized at the set $A = \{0, 1, \ldots, \lfloor rm \rfloor - 1\}$, with $\Phi(A) = \frac{p+q}{m\pi(A)} \geq \frac{p+q}{2mr}$. Then

$$\Lambda(r) \geq \frac{\Phi^2(r)}{2(1-\alpha)} \geq \frac{p+q}{8m^2r^2}\,.$$

The discrete time mixing time is $\tau_2(1/2e) = O\left(\frac{m^2}{(1-p-q)(p+q)}\right)$ by Theorem 2.10.

We now show a matching lower bound. Let x_k denote the direction of the kth step, so $x \in \{-1, 0, +1\}$. By the Central Limit Theorem there is a 99.7% chance that the long range average of x_k is within 3 standard deviations of the expected value. The standard deviation of the average is $\sigma = \sqrt{\frac{p+q-(p-q)^2}{N}}$, and so in particular

$$Prob\left(-3\sigma < \frac{1}{N}\left(\sum_{i=1}^{N} x_i - \mathbb{E}\sum_{i=1}^{N} x_i\right) < 3\sigma\right) > 99.7\%\,.$$

Hence, with high probability, if $N < \frac{m^2}{3^2 * 4^2 * (p+q-(p-q)^2)}$ then the walk is still in the same half of the cycle as it's expected location, and so for constants C_1, C_2,

$$C_1 \frac{m^2}{p+q-(p-q)^2} \leq \tau(1/2e) \leq \frac{1}{2}\tau_2(1/2e) \leq C_2 \frac{m^2}{p+q-(p+q)^2} \,. \tag{2.4}$$

The order of the upper and lower bounds cannot be made to match more closely, because if $p = q = 1/2$ then $\tau_2(1/2e) = O(m^2)$ when there are an odd number of vertices, but the walk is periodic when there are an even number of vertices.

Given that spectral profile is a fairly new tool, it has not been widely studied yet. However, mixing time methodologies that have been developed separately can sometimes be used to lower bound spectral profile, and still obtain the same mixing results. Hence this method subsumes many other results. For instance, in [34] the authors show that the log-Sobolev constant and a Nash inequality induce the following lower bounds:

$$\Lambda(r) \geq \rho \frac{\log(1/r)}{1-r} \quad \text{and} \quad \Lambda(r) \geq \frac{1}{C\, r^{1/2D}} - \frac{1}{T} \,.$$

By applying the Nash bound on $\Lambda(r)$ for $r \leq (T/2DC)^{2D}$ and the log-Sobolev bound when $(T/2DC)^{2D} \leq r \leq 1/2$, then integration in Theorem 2.10 establishes the bound

$$\tau_2(\epsilon) \leq 2T + \frac{1}{\rho} \log\log\left(\frac{2DC}{T}\right)^{2D} + \frac{1}{\lambda} \log\frac{2\sqrt{2}}{\epsilon} \tag{2.5}$$

This is only a factor two weaker than that found with our more direct approach earlier.

Another example of the utility of spectral profile, briefly discussed at the top of this section, are bounds on the mixing time of the Thorp shuffle. The proof of Theorem 5.13 proceeds by first upper bounding $\mathrm{Var}(k_N)$ for some N, then lower bounding $\Lambda(r)$, which when substituted into Theorem 2.10 with $\sigma(x) = \pi(x)k_N(x)$ produces a mixing time bound.

2.3 Comparison methods

It sometimes happens that a Markov chain is difficult to study, but a related chain is more manageable. In this situation the comparison method has been widely used to bound spectral gap, log-Sobolev constant and Nash inequalities (see [70, 24, 25, 26]). The argument applies to the quantities considered in this chapter as well. In order to motivate the subject, we first consider a few examples.

Example 2.12. Goel [33] solves a card-shuffling problem by comparison methods. He considers a slow card shuffle where either the top card in the deck is put in one of the bottom k positions, or one of the bottom k cards is put at the top of the deck. Mixing time upper and lower bounds are shown by comparison to the relevant quantities for the well studied random transposition shuffle, in which a pair of cards is chosen uniformly and the positions of the two cards are then swapped.

The following example illustrates how comparison of spectral profile might make it possible to simplify some difficult results.

Example 2.13. Recall the Bernoulli-Laplace random walk of Example 2.7. The R-particle Exclusion process is a similar random walk on an *arbitrary graph*; here a step consists of choosing a particle with probability proportional to the degree of the vertex that the particle is currently occupying, choosing a neighboring vertex uniformly, and then moving the particle if the neighboring position is vacant.

The Bernoulli-Laplace walk is then, within a factor of $\frac{N-R}{N}$, just the R-particle exclusion process on the complete graph K_N. Diaconis and Saloff-Coste [24] use a comparison method to lower bound the spectral gap and log-Sobolev constant of the Exclusion process in terms of those of the much simpler Bernoulli-Laplace walk, and thus show bounds on mixing of the Exclusion process.

Morris [63], by a clever, but fairly involved argument, shows that the exclusion process on $G = \mathbb{Z}^d/L\mathbb{Z}^d$, the d-dimensional grid of side length L, mixes faster than predicted by log-Sobolev. An alternate approach would be to consider the Bernoulli-Laplace model, apply the relation $\Lambda(r) \geq \rho \log(1/r)$, and then use an alternate method to bound $\Lambda(r)$

when r is extremely small (such as $r < 2^{-d}$). Comparing the Exclusion process to this may then match Morris' bound.

Before deriving comparison results for the quantities in this chapter, a preliminary result is needed.

Theorem 2.14. Consider two Markov chains P and $\hat{\mathsf{P}}$ on the same state space Ω, and for every $x \neq y \in \Omega$ with $\hat{\mathsf{P}}(x,y) > 0$ define a directed path γ_{xy} from x to y along edges in P. Let Γ denote the set of all such paths. Then

$$\mathcal{E}_{\mathsf{P}}(f,f) \geq \frac{1}{A}\,\mathcal{E}_{\hat{\mathsf{P}}}(f,f)\,,$$

$$\mathrm{Var}_{\pi}(f) \leq M\,\mathrm{Var}_{\hat{\pi}}(f)\,, \quad \mathrm{Ent}_{\pi}(f^2) \leq M\,\mathrm{Ent}_{\hat{\pi}}(f^2)\,,$$

where $M = \max_x \frac{\pi(x)}{\hat{\pi}(x)}$ and

$$A = A(\Gamma) = \max_{a\neq b:\mathsf{P}(a,b)\neq 0} \frac{1}{\pi(a)\mathsf{P}(a,b)} \sum_{x\neq y:(a,b)\in\gamma_{xy}} \hat{\pi}(x)\hat{\mathsf{P}}(x,y)|\gamma_{xy}|\,.$$

Proof. Without loss, assume that each path γ_{xy} does not cross the same edge more than once.

First, consider the Dirichlet forms:

$$\begin{aligned}
\mathcal{E}_{\hat{\mathsf{P}}}(f,f) &= \frac{1}{2}\sum_{x\neq y}(f(x)-f(y))^2\hat{\pi}(x)\hat{\mathsf{P}}(x,y) \\
&= \frac{1}{2}\sum_{x\neq y}\Big(\sum_{a\neq b:(a,b)\in\gamma_{xy}}(f(a)-f(b))\Big)^2\hat{\pi}(x)\hat{\mathsf{P}}(x,y) \\
&\leq \frac{1}{2}\sum_{x\neq y}\sum_{a\neq b:(a,b)\in\gamma_{xy}}(f(a)-f(b))^2|\gamma_{xy}|\hat{\pi}(x)\hat{\mathsf{P}}(x,y) \\
&= \frac{1}{2}\sum_{a\neq b}(f(a)-f(b))^2\pi(a)\mathsf{P}(a,b)\ \times \\
&\qquad\qquad \frac{1}{\pi(a)\mathsf{P}(a,b)}\sum_{x\neq y:(a,b)\in\gamma_{xy}}\hat{\pi}(x)\hat{\mathsf{P}}(x,y)|\gamma_{xy}| \\
&\leq \mathcal{E}_{\mathsf{P}}(f,f)\,A.
\end{aligned}$$

For variance we have

$$\mathrm{Var}_\pi(f) = \inf_{c\in\mathsf{R}} \mathbb{E}_\pi(f(x)-c)^2 \le \inf_{c\in\mathsf{R}} M\,\mathbb{E}_{\hat\pi}(f(x)-c)^2 = M\,\mathrm{Var}_{\hat\pi}(f).$$

For entropy, observe that

$$\begin{aligned} \mathrm{Ent}_\pi(f^2) &= \sum_{x\in\Omega} \pi(x)\left(f^2(x)\log\frac{f^2(x)}{\mathbb{E}_\pi f^2} - f^2(x) + \mathbb{E}_\pi f^2\right) \\ &= \inf_{c>0}\sum_{x\in\Omega}\pi(x)\left(f^2(x)\log\frac{f^2(x)}{c} - f^2(x)+c\right) \\ &\le \inf_{c>0} M\sum_{x\in\Omega}\hat\pi(x)\left(f^2(x)\log\frac{f^2(x)}{c} - f^2(x)+c\right) \\ &= M\,\mathrm{Ent}_{\hat\pi}(f^2)\,. \end{aligned}$$

The second equality follows from differentiating with respect to c to see that the minimum occurs at $c=\mathbb{E}_\pi f^2$, while the inequality required the fact that $a\,\log\frac{a}{b}-a+b \ge a\,(1-\frac{b}{a})-a+b = 0$ and so $f^2\log\frac{f^2}{c}-f^2+c\ge 0$. □

An easy consequence of this is that spectral gap, log-Sobolev and spectral profile bounds can be compared.

Corollary 2.15.

$$\lambda_{\mathsf{P}} \ge \frac{1}{M\,A}\lambda_{\hat{\mathsf{P}}}\,, \quad \rho_{\mathsf{P}} \ge \frac{1}{M\,A}\rho_{\hat{\mathsf{P}}}\,, \quad \Lambda_{\mathsf{P}}(r) \ge \frac{1}{M\,A}\Lambda_{\hat{\mathsf{P}}}(r).$$

The log-Sobolev and spectral profile mixing time bounds of P are thus at worst a factor MA times larger than those of $\hat{\mathsf{P}}$.

If the distribution $\pi=\hat\pi$ then a Nash inequality for $\hat{\mathsf{P}}$, along with the relation $\mathcal{E}_{\mathsf{P}}(f,f)\ge\frac{1}{A}\mathcal{E}_{\hat{\mathsf{P}}}(f,f)$, immediately yields a Nash inequality for P. It is not immediately clear how to compare Nash inequality bounds if $\pi\ne\hat\pi$. However, one can compare the spectral profile bounds used to show Equation (2.5), and so the mixing time of P is at most $M\,A$ times the bound Equation (2.5) gives for $\hat{\mathsf{P}}$. Alternatively, one can compare $\mathcal{E}_{\mathsf{P}}(f,f)$ to $\mathcal{E}_{\hat{\mathsf{P}}}(f,f)$ and $\mathrm{Var}_\pi(f)$ to $\mathrm{Var}_{\hat\pi}(f)$ in the original proofs of the mixing times.

In the case of a reversible chain Diaconis and Saloff-Coste [24] observe that it is also possible to compare λ_{n-1} if the paths are of odd length. First, a preliminary result.

Theorem 2.16. Consider two Markov chains P and $\hat{\mathsf{P}}$ on the same state space Ω, and for every $x, y \in \Omega$ with $\hat{\mathsf{P}}(x,y) > 0$ (including possibly $y = x$) define a directed path γ_{xy} of odd length $|\gamma_{xy}|$ from x to y along edges in P. Let Γ^* denote the set of all such paths. Then

$$\mathcal{F}_{\mathsf{P}}(f,f) \geq \frac{1}{A^*}\,\mathcal{F}_{\hat{\mathsf{P}}}(f,f)\,,$$

where $M = \max_x \frac{\pi(x)}{\hat{\pi}(x)}$ and

$$\begin{aligned} A^* &= A^*(\Gamma^*) \\ &= \max_{a,b:\mathsf{P}(a,b)\neq 0} \frac{1}{\pi(a)\mathsf{P}(a,b)} \sum_{x,y:(a,b)\in\gamma_{xy}} \hat{\pi}(x)\hat{\mathsf{P}}(x,y)|\gamma_{xy}|\,r_{xy}(a,b)\,, \end{aligned}$$

where $r_{xy}(a,b)$ is the number of times the edge (a,b) appears in path γ_{xy}.

Proof. The proof is nearly identical to that for comparison of $\mathcal{E}(f,f)$, except that due to the odd path length criterion a path may cross an edge twice. Also, if the path γ_{xy} is given by $x = x_0, x_1, x_2 \ldots, x_m = y$ for m odd then $f(x) + f(y)$ is rewritten as

$$\begin{aligned} f(x) + f(y) &= (f(x) + f(x_1)) - (f(x_1) + f(x_2)) + \cdots \\ &\quad -(f(x_{m-2}) + f(x_{m-1})) + (f(x_{m-1}) + f(y))\,. \end{aligned}$$

□

In particular, if P and $\hat{\mathsf{P}}$ are reversible then

$$1 - \lambda_{max}(\mathsf{P}) \geq \frac{1}{MA^*}\,(1 - \lambda_{max}(\hat{\mathsf{P}}))\,,$$

where we recall that $\lambda_{max}(\mathsf{K}) = \max\{\lambda_1, |\lambda_{n-1}|\}$ denotes the size of the second largest magnitude eigenvalue of Markov kernel K.

The most widely used example of these comparison results is the "canonical path theorem" (see [75, 27] for numerous examples).

Corollary 2.17. Given a Markov chain P on state space Ω, and directed paths γ_{xy} between every pair of vertices $x \neq y \in \Omega$, then

$$\lambda \geq \left(\max_{a\neq b:\mathsf{P}(a,b)\neq 0} \frac{1}{\pi(a)\mathsf{P}(a,b)} \sum_{x\neq y:(a,b)\in\gamma_{xy}} \pi(x)\pi(y)|\gamma_{xy}| \right)^{-1}.$$

Proof. Let $\hat{\mathsf{P}}(x,y) = \pi(y)$, $\hat{\pi} = \pi$ and $M = 1$ in Theorem 2.14. Given $f : \Omega \to \mathsf{R}$ then

$$\begin{aligned} \mathcal{E}_{\hat{\mathsf{P}}}(f,f) &= \frac{1}{2} \sum_{x,y\in\Omega} (f(x) - f(y))^2 \pi(x)\hat{\mathsf{P}}(x,y) \\ &= \frac{1}{2} \sum_{x,y\in\Omega} (f(x) - f(y))^2 \pi(x)\pi(y) = \mathrm{Var}_\pi(f)\,. \end{aligned}$$

It follows that $\mathcal{E}_{\mathsf{P}}(f,f) \geq \frac{1}{A}\mathcal{E}_{\hat{\mathsf{P}}}(f,f) \geq \frac{1}{A}\mathrm{Var}_\pi(f)$, and the result follows by definition of λ and A. □

There is a related bound on the smallest eigenvalue of a reversible chain. This is useful in Corollary 1.15 when studying a random walk with no holding probability, such as the card shufflings considered by Goel (Lemma 4.1 of [33]).

Corollary 2.18. Consider a reversible Markov chain P on state space Ω, and a set of cycles γ_x of odd length from each vertex $x \in \Omega$ to itself. Then the smallest eigenvalue λ_{n-1} of P satisfies the relation

$$1 + \lambda_{n-1} \geq 2\left(\max_{a,b:\,\mathsf{P}(a,b)\neq 0} \frac{1}{\pi(a)\mathsf{P}(a,b)} \sum_{x:\,(a,b)\in\gamma_x} \pi(x)|\gamma_{xy}| r_x(a,b) \right)^{-1},$$

where $r_x(a,b)$ is the number of times edge (a,b) appears in path γ_x.

Proof. Let $\hat{\mathsf{P}}(x,y) = \delta_{x=y}$, $\hat{\pi} = \pi$ and $M = 1$ in Theorem 2.16. Then $\mathcal{F}_{\hat{\mathsf{P}}}(f,f) = \frac{1}{2}\sum_{x\in\Omega}(2f(x))^2\pi(x) = 2\mathbb{E}_\pi f^2$. It follows from Lemma 1.21 that for the walk P,

$$\begin{aligned} 1 + \lambda_{n-1} &= \inf_{\mathrm{Var}_\pi(f)\neq 0} \frac{\mathcal{F}(f,f)}{\mathrm{Var}_\pi(f)} \\ &\geq \frac{1}{A^*} \inf_{\mathrm{Var}_\pi(f)\neq 0} \frac{2\mathbb{E}_\pi(f^2)}{\mathbb{E}_\pi(f^2) - (\mathbb{E}_\pi f)^2} \geq \frac{2}{A^*} \end{aligned}$$

□

3

Evolving Set Methods

In many mixing time results the authors first estimate set expansion and then relate it to mixing time bounds. An early breakthrough in the study of mixing times was the conductance bound

$$\lambda \geq \Phi^2/2 \quad \text{where} \quad \Phi = \min_{A \subset \Omega} \frac{\mathsf{Q}(A, A^c)}{\min\{\pi(A), \pi(A^c)\}}$$

(see [39, 48]). Essentially the same proof can be used (see [34]) to show a conductance profile bound, that

$$\Lambda(r) \geq \Phi^2(r)/2 \quad \text{where} \quad \Phi(r) = \min_{\pi(A) \leq r} \frac{\mathsf{Q}(A, A^c)}{\min\{\pi(A), \pi(A^c)\}} . \tag{3.1}$$

Given holding probability α, this can be boosted to

$$\begin{aligned} \Lambda(r) &= (1-\alpha)\Lambda_{\frac{\mathsf{P}-\alpha I}{1-\alpha}}(r) \geq \frac{1-\alpha}{2}\,\Phi^2_{\frac{\mathsf{P}-\alpha I}{1-\alpha}}(r) \\ &= \frac{1-\alpha}{2}\left(\frac{\Phi(r)}{1-\alpha}\right)^2 = \frac{\Phi^2(r)}{2(1-\alpha)} . \end{aligned}$$

In the common setting of a reversible, lazy (i.e. $\alpha \geq 1/2$) chain Corollary 1.15 then implies the bound

$$\tau_2(\epsilon) \leq \left\lceil \frac{1}{\Phi^2} \log \frac{1}{\epsilon\sqrt{\pi_*}} \right\rceil . \tag{3.2}$$

More generally, by Corollary 1.14 and Theorem 2.10 discrete time mixing satisfies

$$\begin{aligned}
\tau_2(\epsilon) &\leq \left\lceil \frac{2}{\frac{\alpha}{1-\alpha}\,\Phi^2} \log \frac{1}{\epsilon\sqrt{\pi_*}} \right\rceil, \\
\tau_2(\epsilon) &\leq \left\lceil \int_{4\pi_*}^{4/\epsilon^2} \frac{2\,dr}{\frac{\alpha}{1-\alpha}\,r\Phi^2(r)} \right\rceil.
\end{aligned} \tag{3.3}$$

In this chapter we develop a more direct method of proof. This can give stronger set bounds, bounds for distances other than L^2-distance, and also leads to an extension of conductance which applies even with no holding probability. In particular, it is one of the few methods for studying relative entropy mixing $\tau_D(\epsilon)$ of a discrete time chain. Work will be done in discrete time, but carries over easily to continuous time, as discussed at the end of the chapter. The results and their proofs are based on the work of Morris and Peres [65] and Montenegro [61]. We also briefly consider Blocking Conductance, an alternate method which shows better mixing results when set expansion is poor on a small set, but high for larger sets.

3.1 Bounding Distances by Evolving Sets

In order to relate a property of sets (conductance) to a property of the original walk (mixing time) we construct a walk on sets that is a dual to the original Markov chain. Given a Markov chain on Ω with transition matrix P, a *dual process* consists of a walk P_D on some state space V and a *link*, or transition matrix, Λ from V to Ω such that

$$\mathsf{P}\Lambda = \Lambda\mathsf{P}_D\,.$$

In particular, $\mathsf{P}^n\Lambda = \Lambda\mathsf{P}_D^n$ and so the evolution of P^n and P_D^n will be closely related. This relation is given visually by Figure 3.1.

Diaconis and Fill [22] studied the use of dual Markov chains in bounding separation distance. Independently, Morris and Peres [65] proposed the same walk on sets and used it to bound L^2 distance. Montenegro [61] sharpened this technique and extended it to other distances.

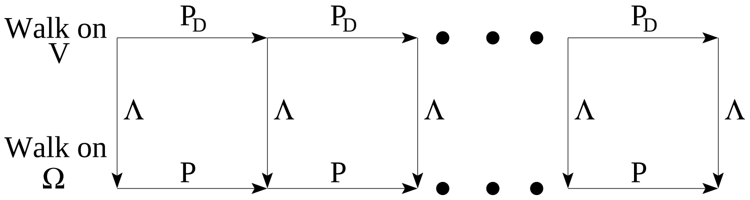

Fig. 3.1 The dual walk P_D projects onto the original chain P.

A natural candidate to link a walk on sets to a walk on states is the projection $\Lambda(S,y) = \frac{\pi(y)}{\pi(S)}\,1_S(y)$. Diaconis and Fill [22] have shown that for certain classes of Markov chains that the walk $\hat{\mathsf{K}}$ below is the unique dual process with link Λ, so this is the walk on sets that should be considered. We use notation of Morris and Peres [65].

Definition 3.1. Given set $A \subset \Omega$ a step of the *evolving set process* is given by choosing $u \in [0,1]$ uniformly at random, and transitioning to the set

$$A_u = \{y \in \Omega : \mathsf{Q}(A,y) \geq u\,\pi(y)\} = \{y \in \Omega : \mathsf{P}^*(y,A) \geq u\}$$

The walk is denoted by S_0, S_1, S_2, …, S_n, with transition kernel $\mathsf{K}^n(A,S) = Prob(S_n = S | S_0 = A)$.

The Doob transform of this process is the Markov chain on sets given by $\hat{\mathsf{K}}(S,S') = \frac{\pi(S')}{\pi(S)}\,\mathsf{K}(S,S')$, with n-step transition probabilities $\hat{\mathsf{K}}^n(S,S') = \frac{\pi(S')}{\pi(S)}\,\mathsf{K}^n(S,S')$.

Heuristically, a step of the evolving set process consists of choosing a uniform value of u, and then A_u is the set of vertices y that get at least a u-fraction of their size $\pi(y)$ from the set A.

The Doob transform produces another Markov chain because of a Martingale property.

Lemma 3.2. If $A \subset \Omega$ then

$$\sum_{A' \subset \Omega} \pi(A')\mathsf{K}(A,A') = \int_0^1 \pi(A_u)\,du = \pi(A)$$

Proof.

$$\int_0^1 \pi(A_u)du = \sum_{y\in\Omega} \pi(y)Prob(y \in A_u) = \sum_{y\in\Omega} \pi(y)\frac{\mathsf{Q}(A,y)}{\pi(y)} = \pi(A)$$

□

The walk $\hat{\mathsf{K}}$ is a dual process of P.

Lemma 3.3. If $S \subset \Omega$, $y \in \Omega$ and $\Lambda(S,y) = \frac{\pi(y)}{\pi(S)}1_S(y)$ is the projection linkage, then

$$\mathsf{P}\Lambda(S,y) = \Lambda\hat{\mathsf{K}}(S,y)\,.$$

Proof.

$$\begin{aligned}\mathsf{P}\Lambda(S,y) &= \sum_{z\in S} \frac{\pi(z)}{\pi(S)}\mathsf{P}(z,y) = \frac{\mathsf{Q}(S,y)}{\pi(S)} \\ \Lambda\hat{\mathsf{K}}(S,y) &= \sum_{S'\ni y} \hat{\mathsf{K}}(S,S')\,\frac{\pi(y)}{\pi(S')} = \frac{\pi(y)}{\pi(S)}\sum_{S'\ni y} \mathsf{K}(S,S') = \frac{\mathsf{Q}(S,y)}{\pi(S)}\end{aligned}$$

The final equality is because $\sum_{S'\ni y} \mathsf{K}(S,S') = Prob(y \in S') = \mathsf{Q}(S,y)/\pi(y)$. □

With duality it becomes easy to write the n step transitions in terms of the walk $\hat{\mathsf{K}}$.

Lemma 3.4. Let $\hat{\mathbb{E}}_n$ denote expectation under $\hat{\mathsf{K}}^n$. If $x \in \Omega$ and $S_0 = \{x\}$ then

$$\mathsf{P}^n(x,y) = \hat{\mathbb{E}}_n \pi_{S_n}(y)\,,$$

where $\pi_S(y) = \frac{1_S(y)\pi(y)}{\pi(S)}$ denotes the probability distribution induced on set S by π.

Proof.

$$\mathsf{P}^n(x,y) = (\mathsf{P}^n\Lambda)(\{x\},y) = (\Lambda\hat{\mathsf{K}}^n)(\{x\},y) = \hat{\mathbb{E}}_n \pi_{S_n}(y)$$

The final equality is because $\Lambda(S,y) = \pi_S(y)$. □

Recall from Equation (1.1) that if a distance $dist(\mu, \pi)$ is convex in μ then the worst initial distribution is a point mass. Given the preceding lemma it is easy to show a distance bound for all such convex distances.

Theorem 3.5. Consider a finite Markov chain with stationary distribution π. Any distance $dist(\mu, \pi)$ which is convex in μ satisfies

$$dist(\mathsf{P}^n(x, \cdot), \pi) \leq \hat{\mathbb{E}}_n dist(\pi_{S_n}, \pi)$$

whenever $x \in \Omega$ and $S_0 = \{x\}$.

Proof. By Lemma 3.4 and convexity,

$$dist(\mathsf{P}^n(x, \cdot), \pi) = dist(\hat{\mathbb{E}}_n \pi_{S_n}, \pi) \leq \hat{\mathbb{E}}_n dist(\pi_{S_n}, \pi)\,.$$

□

In particular, if $dist(\mu, \pi) = \mathcal{L}_\pi\left(\frac{\mu}{\pi}\right)$ for a convex functional $\mathcal{L}_\pi : (\mathsf{R}_+)^\Omega \to \mathsf{R}$ then the distance is convex and the conditions of the theorem are satisfied. The total variation distance satisfies this condition with $\mathcal{L}_\pi(f) = \frac{1}{2}\|f-1\|_{1,\pi}$, relative entropy with $\mathcal{L}_\pi(f) = \mathbb{E}_\pi f \log f$, and L^2 distance with $\mathcal{L}_\pi(f) = \|f-1\|_{2,\pi}$, and so the following bounds are immediate:

Theorem 3.6. If $x \in \Omega$ and $S_0 = \{x\}$ then in discrete time

$$\begin{aligned}
\|\mathsf{P}^n(x,\cdot) - \pi\|_{\mathrm{TV}} &\leq \hat{\mathbb{E}}_n (1 - \pi(S_n)), \\
\mathsf{D}(\mathsf{P}^n(x,\cdot)\|\pi) &\leq \hat{\mathbb{E}}_n \log \frac{1}{\pi(S_n)}, \\
\|\mathsf{P}^n(x,\cdot) - \pi\|_2 &\leq \hat{\mathbb{E}}_n \sqrt{\frac{1-\pi(S_n)}{\pi(S_n)}}\,.
\end{aligned}$$

3.2 Mixing Times

Mixing time bounds can be shown via an argument similar to that used for spectral profile bounds. One result that follows from this is what appears to be the only general method of bounding discrete-time convergence in relative entropy (recall there was no discrete-time analog to ρ_0); by Corollary 3.9 and Theorem 3.6 this distance decreases at a rate of $\mathcal{C}_{z\log(1/z)}$ each step, as is the case with $e^{-\rho_0}$ in continuous time. Also, when examining the Thorp shuffle in Section 5.4.3, the L^2 mixing time bound of Equation (3.5) will be used to give an alternate approach to the spectral profile bounds.

We restrict attention to distances satisfying

$$dist(\mathsf{P}^n(x,\cdot),\pi) \leq \hat{\mathbb{E}}_n f(\pi(S_n))$$

for a decreasing function $f: [0,1] \to \mathsf{R}_+$ (such as those in Theorem 3.6), and define $\tau(\epsilon) = \min\{n : \hat{\mathbb{E}}_n f(\pi(S_n)) \leq \epsilon\}$ to be an upper bound on the mixing time of our distance.

The analog of spectral profile $\Lambda_{\mathsf{PP}^*}(r)$ will be the f-congestion:

Definition 3.7. Given a function $f: [0,1] \to \mathsf{R}_+$ the *f-congestion profile* is

$$\mathcal{C}_f(r) = \max_{\pi(A)\leq r} \mathcal{C}_f(A) \quad \text{where} \quad \mathcal{C}_f(A) = \int_0^1 \frac{f(\pi(A_u))}{f(\pi(A))}\,du\,.$$

The *f-congestion* is $\mathcal{C}_f = \max_{A\subset\Omega} \mathcal{C}_f(A)$.

The analog of Lemma 1.13 will be the following:

Lemma 3.8.

$$\begin{aligned}\hat{\mathbb{E}}_{n+1} f(\pi(S_{n+1})) - \hat{\mathbb{E}}_n f(\pi(S_n)) &= -\hat{\mathbb{E}}_n f(\pi(S_n))\,(1-\mathcal{C}_{zf(z)}(S_n)) \\ &\leq -(1-\mathcal{C}_{zf(z)})\,\hat{\mathbb{E}}_n f(\pi(S_n))\end{aligned}$$

Proof. The inequality is because $1-\mathcal{C}_{zf(z)} \leq 1-\mathcal{C}_{zf(z)}(S)$ for all $S\subset\Omega$. For the equality,

$$\begin{aligned}\hat{\mathbb{E}}_{n+1} f(\pi(S_{n+1})) &= \hat{\mathbb{E}}_n \sum_S \hat{\mathsf{K}}(S_n,S) f(\pi(S)) \\ &= \hat{\mathbb{E}}_n f(\pi(S_n)) \frac{\sum_S \mathsf{K}(S_n,S)\pi(S)f(\pi(S))}{\pi(S_n)f(\pi(S_n))} \\ &= \hat{\mathbb{E}}_n f(\pi(S_n))\mathcal{C}_{zf(z)}(S_n)\end{aligned}$$

□

The analog of Corollary 1.14 is the following:

Corollary 3.9. In discrete time

$$\begin{aligned} dist(\mathsf{P}^n(x,\cdot),\pi) &\leq \mathcal{C}^n_{zf(z)}\, f(\pi(x)) \\ \text{and} \qquad \tau(\epsilon) &\leq \left\lceil \frac{1}{1-\mathcal{C}_{zf(z)}}\,\log\frac{f(\pi_*)}{\epsilon}\right\rceil . \end{aligned}$$

Proof. By Lemma 3.8, $\hat{\mathbb{E}}_{n+1} f(\pi(S_{n+1})) \leq \mathcal{C}_{zf(z)}\, \hat{\mathbb{E}}_n f(\pi(S_n))$, and by induction $\hat{\mathbb{E}}_n f(\pi(S_n)) \leq \mathcal{C}^n_{zf(z)}\, f(\pi(S_0))$. Solving for when this drops to ϵ and using the approximation $\log \mathcal{C}_{zf(z)} \leq -(1-\mathcal{C}_{zf(z)})$, gives the corollary. □

Note that u-almost everywhere $(A_u)^c = (A^c)_{1-u}$, and so if $zf(z) = zf(1-z)$ then a simple calculation shows that $\mathcal{C}_{zf(z)}(A) = \mathcal{C}_{zf(z)}(A^c)$. In particular, when $zf(z) = zf(1-z)$ then we may let $\mathcal{C}_{zf(z)} = \mathcal{C}_{zf(z)}(1/2)$ in the corollary, and more generally when $r \geq 1/2$ then we may take $\mathcal{C}_{zf(z)}(r) = \mathcal{C}_{zf(z)}$.

Theorem 2.10 will have two analogs: a bound under a weak convexity condition, with about a factor of two lost in the general case.

Theorem 3.10. In discrete time, if f is differentiable then

$$\tau(\epsilon) \leq \left\lceil \int_{\pi_*}^{f^{-1}(\epsilon)} \frac{-f'(r)\,dr}{f(r)(1-\mathcal{C}_{zf(z)}(r))}\right\rceil$$

if $r\left(1-\mathcal{C}_{zf(z)}(f^{-1}(r))\right)$ is convex, while in general

$$\tau(\epsilon) \leq \left\lceil \int_{f^{-1}(f(\pi_*)/2)}^{f^{-1}(\epsilon/2)} \frac{-2f'(r)\,dr}{f(r)(1-\mathcal{C}_{zf(z)}(r))}\right\rceil$$

Proof. First consider the convex case.

By Lemma 3.8 and Jensen's inequality for the convex function $x\left(1-\mathcal{C}_{zf(z)}(f^{-1}(x))\right)$,

$$\begin{aligned} &\hat{\mathbb{E}}_{n+1} f(\pi(S_{n+1})) - \hat{\mathbb{E}}_n f(\pi(S_n)) \\ &\quad = -\hat{\mathbb{E}}_n f(\pi(S_n))\,(1-\mathcal{C}_{zf(z)}(S_n)) \\ &\quad \leq -\hat{\mathbb{E}}_n f(\pi(S_n))\,\left[1-\mathcal{C}_{zf(z)}\left(f^{-1}\circ f(\pi(S_n))\right)\right] \\ &\quad \leq -\left[\hat{\mathbb{E}}_n f(\pi(S_n))\right]\left[1-\mathcal{C}_{zf(z)}\left(f^{-1}(\hat{\mathbb{E}}_n f(\pi(S_n)))\right)\right] . \qquad (3.4) \end{aligned}$$

Since $I(n) = \hat{\mathbb{E}}_n f(\pi(S_n))$ and $1 - \mathcal{C}_{zf(z)}(f^{-1}(x))$ are non-increasing, the piecewise linear extension of $I(n)$ to $t \in \mathsf{R}_+$ satisfies

$$\frac{dI}{dt} \leq -I(t)\ \left[1 - \mathcal{C}_{zf(z)}(f^{-1}(I(t)))\right]$$

At integer t the derivative can be taken from either right or left. Make the change of variables $r = f^{-1}(I(t))$, and integrate,

$$\begin{aligned}\tau(\epsilon) &= \int_0^{\tau_2(\epsilon)} 1\, dt \\ &\leq -\int_{I(0)}^{I(\tau_2(\epsilon))} \frac{dI}{I\left(1 - \mathcal{C}_{zf(z)}(f^{-1}(I(t)))\right)} \\ &= -\int_{r(0)}^{r(\tau_2(\epsilon))} \frac{f'(r)dr}{f(r)\left(1 - \mathcal{C}_{zf(z)}(r)\right)},\end{aligned}$$

as in Equation 2.2 and the proof of Theorem 2.10.

For the general case, use Lemma 3.11 instead of convexity at (3.4). □

Lemma 3.11. If $Z \geq 0$ is a nonnegative random variable and g is a nonnegative increasing function, then

$$E\left(Z\, g(Z)\right) \geq \frac{EZ}{2}\, g(EZ/2)\,.$$

Proof. (from [65]) Let A be the event $\{Z \geq EZ/2\}$. Then $E(Z\, 1_{A^c}) \leq EZ/2$, so $E(Z1_A) \geq EZ/2$. Therefore,

$$E\left(Z\, g(2Z)\right) \geq E\left(Z1_A\, g(EZ)\right) \geq \frac{EZ}{2}\, g(EZ)\,.$$

Let $U = 2Z$ to get the result. □

It is fairly easy to translate these to mixing time bounds. For instance, by Theorem 3.6 it is appropriate to let $f(z) = \sqrt{\frac{1-z}{z}}$ for L^2

bounds. Then the bounds from Corollary 3.9 and Theorem 3.10 imply:

$$\tau_2(\epsilon) \leq \begin{cases} \left\lceil \dfrac{1}{1-\mathcal{C}_{\sqrt{z(1-z)}}} \log \dfrac{1}{\epsilon\sqrt{\pi_*}} \right\rceil \\ \left\lceil \displaystyle\int_{\pi_*}^{\frac{1}{1+\epsilon^2}} \frac{dx}{2x(1-x)(1-\mathcal{C}_{\sqrt{z(1-z)}}(x))} \right\rceil \\ \left\lceil \displaystyle\int_{\frac{4\pi_*}{1+3\pi_*}}^{\frac{1}{1+\epsilon^2/4}} \frac{dx}{x(1-x)(1-\mathcal{C}_{\sqrt{z(1-z)}}(x))} \right\rceil \end{cases}, \tag{3.5}$$

with the first integral requiring $x\left(1-\mathcal{C}_{\sqrt{z(1-z)}}\left(\frac{1}{1+x^2}\right)\right)$ to be convex. By making the change of variables $x = \frac{r}{1+r}$ and applying a few pessimistic approximations one obtains a result more strongly resembling spectral profile bounds:

$$\tau_2(\epsilon) \leq \begin{cases} \left\lceil \dfrac{1}{1-\mathcal{C}_{\sqrt{z(1-z)}}} \log \dfrac{1}{\epsilon\sqrt{\pi_*}} \right\rceil \\ \left\lceil \displaystyle\int_{\pi_*}^{1/\epsilon^2} \frac{dr}{2r(1-\mathcal{C}_{\sqrt{z(1-z)}}(r))} \right\rceil \\ \left\lceil \displaystyle\int_{4\pi_*}^{4/\epsilon^2} \frac{dr}{r(1-\mathcal{C}_{\sqrt{z(1-z)}}(r))} \right\rceil \end{cases}$$

For total variation distance related results are in terms of $\mathcal{C}_{z(1-z)}(r)$, and $\mathcal{C}_{z\log(1/z)}(r)$ for relative entropy. The mixing time bounds are left to the interested reader.

3.3 Conductance

The most common geometric tool for studying mixing time is the *conductance* Φ, a measure of the chance of leaving a set after a single step. The conductance profile can also be used to lower bound the various f-congestion quantities $\mathcal{C}_f$ when the Markov chain is lazy. In Section 5.4.3 conductance profile, and in particular Lemma 3.13, is used to bound $\mathcal{C}_{\sqrt{a(1-a)}}(r)$ for the Thorp shuffle.

The argument is fairly simple (see also [65]).

Theorem 3.12. Given a lazy Markov chain, and $f : \mathsf{R}_+ \to \mathsf{R}_+$ concave, then

$$\mathcal{C}_f(A) \le \frac{f(\pi(A) + 2\mathsf{Q}(A,A^c)) + f(\pi(A) - 2\mathsf{Q}(A,A^c))}{2f(\pi(A))} .$$

Proof. For a lazy chain, if $u > 1/2$ then $A_u \subset A$, and so

$$\begin{aligned}\int_{1/2}^1 \pi(A_u)\,du &= \sum_{y\in A}\left(\frac{\mathsf{Q}(A,y)}{\pi(y)} - \frac{1}{2}\right)\pi(y)\\ &= \mathsf{Q}(A,A) - \frac{\pi(A)}{2} = \frac{\pi(A)}{2} - \mathsf{Q}(A,A^c) .\end{aligned}$$

By Lemma 3.2 $\int_0^1 \pi(A_u)du = \pi(A)$, from which it follows that

$$\int_0^{1/2} \pi(A_u)\,du = \pi(A) - \int_{1/2}^1 \pi(A_u)du = \frac{\pi(A)}{2} + \mathsf{Q}(A,A^c) . \qquad (3.6)$$

Jensen's inequality shows that if $f, g : \mathsf{R}_+ \to \mathsf{R}_+$ with f concave then $\int_0^1 f(g(x))\,dx \le f\left(\int_0^1 g(x)dx\right)$. In particular,

$$\begin{aligned}\mathcal{C}_f(A) &= \frac{\int_0^{1/2} f(\pi(A_u))\,\frac{du}{1/2} + \int_{1/2}^1 f(\pi(A_u))\,\frac{du}{1/2}}{2f(\pi(A))}\\ &\le \frac{f\left(\int_0^{1/2}\pi(A_u)\,\frac{du}{1/2}\right) + f\left(\int_{1/2}^1 \pi(A_u)\frac{du}{1/2}\right)}{2f(\pi(A))}\end{aligned}$$

□

For each choice of f a bit of simplification leads to bounds on $\mathcal{C}_f$. For instance, a lazy Markov chain will have

$$\mathcal{C}_{\sqrt{z(1-z)}}(A) \le \sqrt{1 - \tilde{\Phi}(A)^2} \quad \text{and} \quad \tau_2(\epsilon) \le \left\lceil \frac{2}{\tilde{\Phi}^2}\log\frac{1}{\epsilon\sqrt{\pi_*}}\right\rceil ,$$

where conductance

$$\tilde{\Phi}(A) = \frac{\mathsf{Q}(A,A^c)}{\pi(A)\pi(A^c)}, \quad \tilde{\Phi}(r) = \min_{\pi(A)\le r}\tilde{\Phi}(A), \quad \tilde{\Phi} = \min_{A\subset\Omega}\tilde{\Phi}(A) .$$

Note that $\tilde{\Phi}(A) > \Phi(A)$, and so this is an improvement on (3.3) for lazy Markov chains. See Example 3.15 for an example of how to use conductance.

Conductance is inappropriate for non-lazy chains because it cannot distinguish a periodic chain from an aperiodic one. In Section 1.3 it was found that for a discrete time chain the L^2 mixing time is closely related to λ_{PP^*}, and via Cheeger's inequality it is thus related to Φ_{PP^*}. In fact, the same holds for the evolving set bound on L^2 mixing.

Lemma 3.13.

$$1 - \mathcal{C}_{\sqrt{z(1-z)}}(A) \geq 1 - \sqrt[4]{1 - \Phi^2_{\mathsf{PP}^*}(A)} \geq \frac{1}{4}\,\Phi^2_{\mathsf{PP}^*}(A)$$

Proof. Given $A, B \subset \Omega$, and $u, w \in [0,1]$ chosen uniformly at random, then

$$\begin{aligned}
\mathsf{Q}_{\mathsf{PP}^*}(A,B) &= \sum_x \mathsf{Q}_{\mathsf{P}}(A,x)\mathsf{P}^*(x,B) \\
&= \sum_x \pi(x)\,Prob(x \in A_u)\,Prob(x \in B_w) \\
&= \mathbb{E}\pi(A_u \cap B_w)\,.
\end{aligned}$$

In particular,

$$\mathbb{E}\pi(A_u \cap A_w) = \mathsf{Q}_{\mathsf{PP}^*}(A,A) = \pi(A) - \mathsf{Q}_{\mathsf{PP}^*}(A,A^c)\,.$$

This suggests that we should rewrite $\mathcal{C}_{\sqrt{z(1-z)}}(A)$ in terms of $\pi(A_u \cap A_w)$. Let $X = \pi(A_u \cap A_w) = \min\{\pi(A_u),\, \pi(A_w)\}$ and $Y = \pi(A_u \cup A_w) = \max\{\pi(A_u),\, \pi(A_w)\}$. Then

$$\begin{aligned}
&\mathbb{E}\sqrt{\pi(A_u)(1-\pi(A_u))} \\
&\quad= \sqrt{\mathbb{E}\sqrt{\pi(A_u)(1-\pi(A_u))\pi(A_w)(1-\pi(A_w))}} \\
&\quad= \sqrt{\mathbb{E}\sqrt{X(1-X)Y(1-Y)}} \\
&\quad\leq \sqrt[4]{\mathbb{E}X(1-X)\,\mathbb{E}Y(1-Y)} \\
&\quad\leq \sqrt[4]{\mathbb{E}X(1-\mathbb{E}X)\,\mathbb{E}Y(1-\mathbb{E}Y)} \\
&\quad= \sqrt[4]{\mathbb{E}X(1-\mathbb{E}X)\,(2\pi(A)-\mathbb{E}X)(1-2\pi(A)+\mathbb{E}X)}
\end{aligned}$$

To complete the lemma, substitute in the identity $\mathbb{E}X = \pi(A) - \mathsf{Q}_{\mathsf{PP}^*}(A, A^c)$, then divide by $\sqrt{\pi(A)\pi(A^c)}$ to obtain

$$\mathcal{C}_{\sqrt{z(1-z)}}(A) \leq \sqrt[4]{\left(1 - \Phi^2_{\mathsf{PP}^*}(A)\right)\left(1 - \Phi^2_{\mathsf{PP}^*}(A)\frac{\pi(A)^2}{\pi(A^c)^2}\right)}$$

and then pessimistically assume that $\pi(A) = 0$. □

This is comparable to the spectral profile result. For instance, it follows that

$$\tau_2(\epsilon) \leq \left\lceil \int_{4\pi_*}^{4/\epsilon^2} \frac{4\,dr}{r\,\Phi^2_{\mathsf{PP}^*}(r)} \right\rceil$$

which matches the spectral profile bound but can be improved by a factor of two if $x\Phi^2_{\mathsf{PP}^*}\left(\frac{1}{1+x^2}\right)$ is convex.

3.4 Modified Conductance

In the previous section we discussed use of conductance of PP^* to study mixing time of a non-lazy chain. We now consider an alternative approach which does not require use of the chain PP^*. Recall that conductance cannot be used for non-lazy chains because, for example, in a periodic chain the conductance may be high but the walk alternates between two sets of equal sizes (the partitions) and never reaches more than half the space at once, and hence never mixes. It seems more appropriate, therefore, to consider the chance of stepping from a set A into a strictly larger set, that is, the worst flow into a set of size $\pi(A^c)$. With this motivation, consider

$$\Psi(A) = \min_{\substack{B\subset\Omega,\, v\in\Omega \\ \pi(B)\leq\pi(A^c)<\pi(B\cup v)}} \mathsf{Q}(A, B) + \frac{\pi(A^c) - \pi(B)}{\pi(v)}\,\mathsf{Q}(A, v)\,.$$

The conductance-like quantity to be considered in this section is the following:

Definition 3.14. The *modified conductance* $\tilde{\phi}$ and its profile $\tilde{\phi}(r)$ are given by defining

$$\tilde{\phi}(A) = \frac{\Psi(A)}{\pi(A)\pi(A^c)}, \quad \tilde{\phi}(r) = \min_{\pi(A)\leq r} \tilde{\phi}(A), \quad \tilde{\phi} = \min_{A\subset\Omega} \tilde{\phi}(A)\,.$$

Define $\phi(A)$ similarly but without $\pi(A^c)$ in the denominator.

Observe that for a lazy chain $\Psi(A) = \mathsf{Q}(A, A^c)$, and so $\tilde{\phi}(A) = \tilde{\Phi}(A) \geq \Phi(A)$, showing that modified conductance bounds extend conductance to the non-lazy case.

Example 3.15. Consider a random walk on a cycle of even length, $\mathbb{Z}/m\mathbb{Z}$.

The lazy random walk with $\mathsf{P}(i,i) = 1/2$ and $\mathsf{P}(i, i \pm 1) = 1/4$ has $\tilde{\Phi} = 2/m$, as demonstrated in Figure 3.2, and so we can conclude that

$$\begin{aligned} 1 - \mathcal{C}_{\sqrt{z(1-z)}} &\geq 1 - \sqrt{1 - 4/m^2} \geq 2/m^2 \\ \text{and} \qquad \tau_2(\epsilon) &\leq \left\lceil \frac{m^2}{2} \log \frac{\sqrt{m}}{\epsilon} \right\rceil . \end{aligned}$$

Meanwhile, for the non-lazy random walk with $\mathsf{P}(i, i \pm 1) = 1/2$ Figure 3.2 shows the worst case of A and B, with $\Psi(A) = \mathsf{Q}(A, B) = 0$, and so $\tilde{\phi} = 0$.

The conductance of the lazy version of this chain was good, and we correctly found that the chain converged, while the non-lazy version is periodic and has zero modified conductance, so modified conductance captured the key differences between these chains.

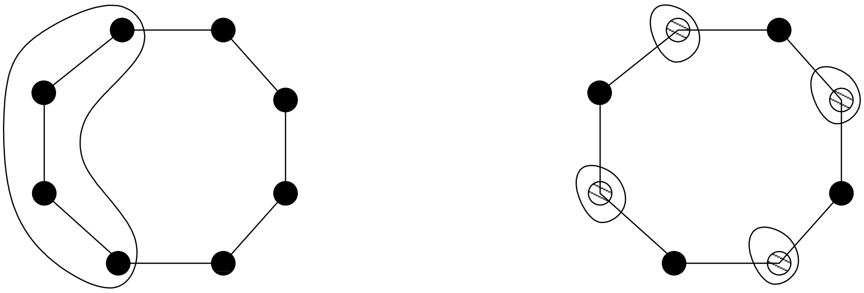

Fig. 3.2 Circled region on left gives $\tilde{\Phi}$. For $\tilde{\phi}$, let A be white vertices and B circled vertices.

The main result of this section is the following:

Theorem 3.16. Given a subset $A \subset \Omega$ then

$$\begin{array}{rclcl} \tilde{\phi}(A) & \geq & 1-\mathcal{C}_{\sqrt{z(1-z)}}(A) & \geq & 1-\sqrt{1-\tilde{\phi}(A)^2} \geq \tilde{\phi}(A)^2/2 \\ \tilde{\phi}(A) & \geq & 1-\mathcal{C}_{z\log(1/z)}(A) & \geq & \dfrac{2\phi(A)^2}{\log(1/\pi(A))} \\ \tilde{\phi}(A) & \geq & 1-\mathcal{C}_{z(1-z)}(A) & \geq & 4\tilde{\phi}(A)^2\pi(A)(1-\pi(A)) \end{array}$$

In order to prove this it is necessary to extend (3.6) to a result for writing $\Psi(A)$ in terms of evolving sets.

Lemma 3.17. Given $A \subset \Omega$ and $\wp_A \in [0,1]$ satisfying

$$\inf\{y:\, \pi(A_y) \leq \pi(A)\} \leq \wp_A \leq \sup\{y:\, \pi(A_y) \geq \pi(A)\}$$

then

$$\Psi(A) = \int_0^{\wp_A} (\pi(A_u) - \pi(A))\, du = \int_{\wp_A}^1 (\pi(A) - \pi(A_u))\, du\,.$$

For a lazy chain one may let $\wp_A = 1/2$, and we get (3.6) again.

Proof. The second equality is from the Martingale property Lemma 3.2. The final equality follows from the second equality and the definition of $\wp_A$.

To prove the first equality, observe that $\forall x \in \Omega$: $\mathsf{Q}(A,x) = \int_0^1 \pi(A_u \cap x)\, du$ and so if $w \in [0,1]$ then

$$\int_0^w (\pi(A_u) - \pi(A_w))\; du = \mathsf{Q}(A, \Omega \setminus A_w)\,.$$

In particular, if $\pi(A_{\wp_A}) = \pi(A)$, then $B = \Omega \setminus A_{\wp_A}$ in the definition of $\Psi(A)$, and the first equality follows.

More generally, if $\pi(A_{\wp_A}) > \pi(A)$ then $B \supset \Omega \setminus A_{\wp_A}$ and the points $x \in (B \cup v) \setminus (\Omega \setminus A_{\wp_A})$ satisfy $\mathsf{Q}(A,x) = \wp_A \pi(x)$. But then

$$\begin{array}{rcl} \Psi(A) & = & \mathsf{Q}(A, \Omega \setminus A_{\wp_A}) + (\pi(A^c) - \pi(\Omega \setminus A_{\wp_A}))\, \wp_A \\ & = & \displaystyle\int_0^{\wp_A} (\pi(A_u) - \pi(A))\, du\,, \end{array}$$

which completes the general case. □

Theorem 3.16 can be shown via Jensen's inequality and this lemma, although the upper bounds require a careful setup. However, we will follow an alternative approach in which the extreme cases are constructed explicitly. The following analytic fact will be needed.

Lemma 3.18. Given two non-increasing functions $g,\, \hat{g} :\ [0,1] \to [0,1]$ such that $\int_0^1 g(u)\,du = \int_0^1 \hat{g}(u)\,du$ and $\forall t \in [0,1] :\ \int_0^t g(u)\,du \geq \int_0^t \hat{g}(u)\,du$, then

$$\int_0^1 f \circ g(u)\,du \leq \int_0^1 f \circ \hat{g}(u)\,du,$$

for every concave function $f : [0,1] \to \mathsf{R}$.

Proof. The concavity of $f(x)$ implies that

$$\forall x \geq y,\, \delta \geq 0 :\ f(x) + f(y) \geq f(x+\delta) + f(y-\delta)\,. \tag{3.7}$$

This follows because $y = \lambda\,(y-\delta)+(1-\lambda)\,(x+\delta)$ with $\lambda = 1-\frac{\delta}{x-y+2\delta} \in [0,1]$ and so by concavity $f(y) \geq \lambda\,f(y-\delta)+(1-\lambda)\,f(x+\delta)$. Likewise, $x = (1-\lambda)\,(y-\delta)+\lambda\,(x+\delta)$ and $f(x) \geq (1-\lambda)\,f(y-\delta)+\lambda\,f(x+\delta)$. Adding these two inequalities gives (3.7).

The inequality (3.7) shows that if a bigger value (x) is increased by some amount, while a smaller value (y) is decreased by the same amount, then the sum $f(x)+f(y)$ decreases. In our setting, the condition that $\forall t \in [0,1] :\ \int_0^t g(u)\,du \geq \int_0^t \hat{g}(u)\,du$ shows that changing from $\hat{g}$ to g increased the already large values of $\hat{g}(u)$, while the equality $\int_0^1 g(u)\,du = \int_0^1 \hat{g}(u)\,du$ assures that this is canceled out by an equal decrease in the already small values. The lemma then follows from (3.7). □

Proof. [Proof of Theorem 3.16] Observe that $\pi(A_u) \in [0,1]$ is non-increasing and Lemma 3.17 shows $\Psi(A)$ is the area below $\pi(A_u)$ and above $\pi(A)$, and also below $\pi(A)$ and above $\pi(A_u)$. It is easily seen that, subject to these conditions, for all $t \in [0,1]$ the integral $\int_0^t \pi(A_u)\,du$ is upper (or lower) bounded by the case when $\pi(A_u)$ has the shapes given in Figure 3.3. By Lemma 3.18 these are the extreme cases minimizing (or maximizing) $\mathcal{C}_f(A)$ for *every* concave function $f(x)$!

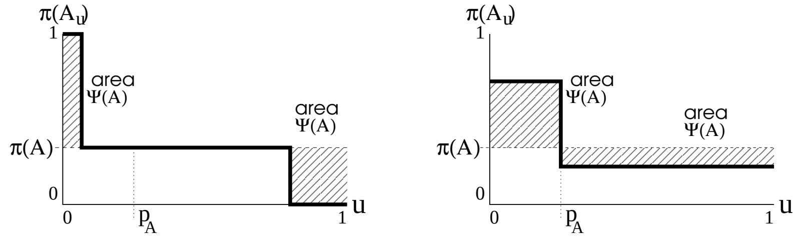

Fig. 3.3 Maximizing $\int_0^t \pi(A_u)\,du$ and minimizing $\int_0^t \pi(A_u)\,du$ given $\Psi(A)$ and $\wp_{AA}$.

First consider the upper bound. If we let $M(u)$ denote the maximizing case in the figure, then $\forall t \in [0,1]: \ \int_0^t \pi(A_u)\,du \leq \int_0^t M(u)\,du$ and $\int_0^1 \pi(A_u)\,du = \pi(A) = \int_0^1 M(u)\,du$, where

$$M(u) = \begin{cases} 1 & \text{if } u \leq \frac{\Psi(A)}{1-\pi(A)} \\ \pi(A) & \text{if } u \in \left(\frac{\Psi(A)}{1-\pi(A)},\, 1 - \frac{\Psi(A)}{\pi(A)}\right] \\ 0 & \text{if } u > 1 - \frac{\Psi(A)}{\pi(A)} \end{cases}$$

By Lemma 3.18 any choice of $f(z)$ which is concave and non-negative will therefore satisfy

$$\begin{aligned} \mathcal{C}_f(A) &\geq \frac{\int_0^1 f \circ M(u)\,du}{f(\pi(A))} \\ &= \frac{\Psi(A)}{1-\pi(A)}\,\frac{f(1)}{f(\pi(A))} + \left(1 - \frac{\Psi(A)}{\pi(A)\pi(A^c)}\right)\frac{f(\pi(A))}{f(\pi(A))} \\ &\quad + \frac{\Psi(A)}{\pi(A)}\,\frac{f(0)}{f(\pi(A))} \\ &\geq 1 - \tilde{\phi}(A) \end{aligned}$$

This shows all of the upper bounds.

For the lower bound, this time the figure shows that $\int_0^t \pi(A_u)\,du \geq \int_0^t m(u)\,du$ when

$$m(u) = \begin{cases} \pi(A) + \frac{\Psi(A)}{\wp_A} & \text{if } u < \wp_A \\ \pi(A) - \frac{\Psi(A)}{1-\wp_A} & \text{if } u > \wp_A \end{cases}$$

By Lemma 3.18, if $f(z) = z(1-z)$ then

$$\begin{aligned}
\mathcal{C}_{z(1-z)}(A) &\leq \frac{\int_0^1 f \circ m(u)\, du}{f(\pi(A))} \\
&= \wp_A \, \frac{\pi(A) + \frac{\Psi(A)}{\wp_A}}{\pi(A)} \, \frac{1 - \pi(A) - \frac{\Psi(A)}{\wp_A}}{1-\pi(A)} \\
&\quad + (1-\wp_A) \, \frac{\pi(A) - \frac{\Psi(A)}{1-\wp_A}}{\pi(A)} \, \frac{1 - \pi(A) + \frac{\Psi(A)}{1-\wp_A}}{1-\pi(A)} \\
&= 1 - \frac{\tilde{\phi}(A)^2\, \pi(A)\pi(A^c)}{\wp_A(1-\wp_A)} \leq 1 - 4\, \tilde{\phi}(A)^2\, \pi(A)\pi(A^c)
\end{aligned}$$

The other cases are similar, but with harder inequalities to eliminate the variable $\wp_A$. See [61] for details. □

It follows that, for instance,

$$\tau_2(\epsilon) \leq \left\lceil \frac{2}{\tilde{\phi}^2} \log \frac{1}{\epsilon\sqrt{\pi_*}} \right\rceil \quad \text{and} \quad \tau_2(\epsilon) \leq \left\lceil \int_{4\pi_*}^{4/\epsilon^2} \frac{2\, dr}{r\tilde{\phi}^2(r)} \right\rceil . \qquad (3.8)$$

In a lazy Markov chain, then $\tilde{\phi}(r) = \tilde{\Phi}(r) > \Phi(r)$ and this is a strict improvement on the spectral bound (3.3), with a further improvement if $x\tilde{\Phi}^2\left(\frac{1}{1+x^2}\right)$ is convex.

In this section we used the modified conductance $\tilde{\phi}(A)$ as a direct measure of congestion of P, rather than use congestion of PP^* via $\Phi_{\mathsf{PP}^*}(A)$. In fact, the two are related.

Lemma 3.19.

$$\tilde{\phi}(A) \geq \frac{1}{2}\, \tilde{\Phi}_{\mathsf{PP}^*}(A)$$

Proof. Let B be the set of size $\pi(A^c)$ such that $\Psi(A) = \mathsf{Q}(A,B)$. Observe that

$$\begin{aligned}
\mathsf{Q}_{\mathsf{P}^*}(B^c, A^c) &= \mathsf{Q}(A^c, B^c) = \pi(A^c) - \mathsf{Q}(A^c, B) \\
&= \pi(A^c) - (\pi(B) - \mathsf{Q}(A,B)) = \mathsf{Q}(A,B)
\end{aligned}$$

All ergodic flow from A to A^c in PP^* must involve a transition (via P) from A to B or B^c, and then (via P^*) to A^c, and hence

$$\mathsf{Q}_{\mathsf{PP}^*}(A, A^c) \leq \mathsf{Q}_{\mathsf{P}}(A,B) + \mathsf{Q}_{\mathsf{P}*}(B^c, A^c) = 2\Psi(A)$$

□

It follows that

$$\begin{aligned} 1 - \mathcal{C}_{\sqrt{z(1-z)}}(r) &\geq 1 - \sqrt{1 - \tilde{\phi}^2(r)} \\ &\geq \ 1 - \sqrt{1 - \tilde{\Phi}^2_{\mathsf{PP}^*}(r)/4} \geq \frac{1}{8}\,\tilde{\Phi}^2_{\mathsf{PP}^*}(r) \end{aligned} \tag{3.9}$$

and a corresponding mixing time bound follows immediately as well. In particular, modified conductance will not be more than a factor two weaker than working with conductance of PP^*, via Lemma 3.13, but may potentially be much better.

3.5 Continuous Time

Not much need be changed for continuous time. It is easily verified that if $\hat{\mathsf{K}}_t = e^{-t(\mathsf{I}-\hat{\mathsf{K}})}$ then

$$H_t(x,y) = \hat{\mathbb{E}}_t \pi_{S_t}(y)$$

where $S_0 = \{x\}$ and $\hat{\mathbb{E}}_t$ is the expectation under the walk $\hat{\mathsf{K}}_t$. Bounds in terms of k_n^x then translate directly into bounds in terms of h_t^x. Once Lemma 3.8 is replaced by

$$\frac{d}{dt}\hat{\mathbb{E}}_t f(\pi(S_t)) = -\hat{\mathbb{E}}_t f(\pi(S_t))(1 - \mathcal{C}_{zf(z)}(S_t))$$

then mixing time bounds also carry over to the continuous-time case, although it is no longer necessary to approximate by a derivative in the proofs nor necessary to take the ceiling of the bounds. One advantage of working in continuous time is that the chain P mixes in exactly half the time of the chain $\frac{\mathsf{I+P}}{2}$, and so conductance applies even if a chain is non-lazy:

$$\begin{aligned} \|h_t - 1\|_2 \ &\leq \ e^{-\left(1-\mathcal{C}^{(\mathsf{I+P})/2}_{\sqrt{z(1-z)}}\right)2t}\sqrt{\frac{1-\pi_*}{\pi_*}} \\ &\leq \ e^{-2t(1-\sqrt{1-\tilde{\Phi}^2/4})}\sqrt{\frac{1-\pi_*}{\pi_*}} \end{aligned} \tag{3.10}$$

and

$$\tau_2(\epsilon) \leq \frac{1}{2}\int_{4\pi_*}^{4/\epsilon^2} \frac{2dr}{r\,\tilde{\Phi}^2_{(\mathsf{I+P})/2}(r)} = \int_{4\pi_*}^{4/\epsilon^2} \frac{4dr}{r\,\tilde{\Phi}^2(r)}\,. \tag{3.11}$$

3.6 Blocking Conductance

We finish this chapter by discussing an alternative method for geometric bounds on mixing times. Generally, evolving set methods give excellent L^2 mixing bounds. However, L^2 mixing can be slower than total variation if there is a bottleneck at a very small set. For instance, take a complete graph K_m, attach a vertex v to it by a single edge, and consider the random walk which chooses a neighbor uniformly with probability $1/2m$ each, otherwise does nothing. Then $\tau(1/e) = \Theta(m)$ but $\tau_2(1/e) = \Theta(m \log m)$. More interesting examples include the lamplighter walks [68] and a non-reversible walk of [23].

In each of these cases the walk stays in a small set of vertices for a long time, and so even as variation distance falls the L^2 distance stays large. Evolving set arguments do not work well with these examples, because evolving set bounds involve the increasing function $\mathcal{C}_f(r) = \max_{\pi(A)\leq r} \mathcal{C}_f(A)$, and hence a bottleneck at a very small set A will effect all values of $r \geq \pi(A)$. One remedy to this is to work with Blocking Conductance, an alternate approach to bounding total variation distance on lazy reversible Markov chains developed by Kannan, Lovász and Montenegro [46]. A specific case of their result is the following:

Theorem 3.20. The total variation mixing time of a reversible, lazy Markov chain satisfies

$$\tau(\epsilon) \leq C \left(\int_{\pi_*}^{1/2} \frac{dr}{r\psi^+(r)} + \frac{1}{\psi^+(1/2)} \right) \log(1/\epsilon)\,,$$

where C is a universal constant, $\psi^+(r) = \min_{\frac{r}{2}\leq\pi(A)\leq r} \psi^+(A)$ and if $A \subset \Omega$ then

$$\begin{aligned}
\psi^+(A) &= \frac{1}{2}\int_{1/2}^{1} \left(1 - \frac{\pi(A_u)}{\pi(A)}\right)^2 du \\
&\geq \sup_{\lambda\leq\pi(A)} \min_{\substack{S\subset A \\ \pi(S)<\lambda}} \frac{\lambda \mathsf{Q}(A\setminus S, A^c)}{\pi(A)^2} \\
&\geq \frac{1}{2}\Phi^2(A)\,.
\end{aligned}$$

The key difference from the evolving set bounds is that if $\psi^+(A)$ is small then this only effects the integral with $r \in [\pi(A), 2\pi(A)]$. It is known [60] that $1 - \mathcal{C}_{\sqrt{z(1-z)}}(A) \geq \frac{1}{4}\psi^+(A)$, and so the sole improvement here over using evolving sets is in the treatment of bottlenecks.

The second characterization of $\psi^+(A)$ above can be interpreted as follows. Let λ denote the maximal size of a "blocking set", such that if any set S smaller than this is blocked from transitioning then it does not block too much of the ergodic flow $\mathsf{Q}(A, A^c)$. In particular, if $S \subset A$ then $\mathsf{Q}(A \setminus S, A^c) = \mathsf{Q}(A, A^c) - \mathsf{Q}(S, A^c) \geq \mathsf{Q}(A, A^c) - \pi(S)/2$, and so by setting $\lambda = \mathsf{Q}(A, A^c)$ then the first lower bound on $\psi^+(A)$ implies the second.

For instance, in the example of a complete graph with a single vertex v attached, let $\lambda = \pi(\{v\}) = \frac{1}{m+1}$. The only set $\pi(S) < \lambda$ is $S = \emptyset$, and so $\psi^+(\{v\}) \geq \frac{\mathsf{Q}(\{v\}, K_m)}{\pi(\{v\})} = \frac{1}{2m}$. All sets $A \neq \{v\}$ satisfy $\Phi(A) \geq 1/8$, and so $\psi^+(A) \geq \frac{1}{128}$. Hence $\psi^+(r) = \frac{1}{m+1}$ if $\frac{1}{m+1} \leq r \leq \frac{2}{m+1}$, and $\psi^+(r) \geq \frac{1}{128}$ otherwise. The above integral then evaluates to

$$\tau(\epsilon) = O(m \log(1/\epsilon)),$$

which is correct. Moreover, Equation (3.5) gives the correct $\tau_2(\epsilon) = O(m \log(m/\epsilon))$ bound because $1 - \mathcal{C}_{\sqrt{z(1-z)}}(A) \geq \frac{1}{4}\psi^+(A)$, as remarked above.

A more interesting example is given by Fountoulakis and Reed [31]. They have studied mixing time for the largest component of the random graph $G_{n,p}$. This is the graph given by taking n vertices, and connecting each pair of vertices by an edge with probability p. They report that the conductance is poor at very small sets, and increases with set size, leading to the following result:

Theorem 3.21. The random graph $G_{n,p}$, with $p = p(n)$ such that $1 + \Theta(1) < np$, satisfies

$$\tau(1/e) = \Theta\left(\max\left\{\left(\frac{\ln n}{np}\right)^2, \frac{\ln n}{\ln np}\right\}\right)$$

with probability $1 - o(1)$.

Remark 3.22. The proofs of the Blocking Conductance and Evolving set results are very different, and yet similar in some ways. In the former case, after n-steps of the walk, the vertices are ordered from those where the n-step average density $\rho_n(y) = \frac{1}{n+1}\sum_{i=0}^n k_i(y)$ is largest, down to those where $\rho_n(x)$ is smallest, as v_1, v_2, $\ldots, v_k$. Then, for each $\{v_1, v_2, \ldots, v_i\}$ they consider how much the set will grow in another step of the walk. Similarly, with evolving sets, $\rho_n(y) = \frac{1}{n+1}\sum_{i=0}^n \frac{Prob(y \in S_i)}{\pi(x)}$, and so the points with the highest average chance of being in an S_i are exactly those appearing at the beginning of the list v_1, v_2, $\ldots$, v_k. Likewise, given set A, the set A_u is a measure of how quickly the walk expands in a single step.

Remark 3.23. In [46] they prove a very general theorem bounding total variation mixing time, then write three corollaries of it, one of which is Theorem 3.20 given above. Surprisingly, although the methods of proof used to show Blocking conductance and Evolving set results were entirely different, it turns out that these three corollaries are almost exactly the total variation, relative entropy, and L^2 mixing time bounds of Theorem 3.6 and Corollary 3.9, but with better treatment of bottlenecks [61].

4

Lower Bounds on Mixing Times and their Consequences

In previous chapters we have considered the problem of bounding various mixing times from above. Once one has shown an upper bound on mixing time, it is natural to hope for a matching lower bound. In this chapter such lower bounds will be considered. Just as we found spectral and geometric (conductance) methods for upper bounds, there are also spectral and geometric arguments for lower bounding mixing times. A log-Sobolev lower bound will also be considered. We finish the chapter with discussion of a consequence of the lower bounds: a method of comparing mixing times.

4.1 A geometric lower bound

We first consider the geometric argument for lower bounding mixing times. While this can be used to show lower bounds for a rapidly mixing Markov chain, it is more commonly used to show that a Markov chain mixes slowly, so-called "torpid mixing." For instance, Borgs, et. al. [8] studied the Swendsen-Wang and the Potts model near their phase transitions, and by constructing a set A where the conductance $\Phi(A)$ is exponentially small they were able to establish that both Markov chains take exponentially long to converge. This is of practical significance as both Markov chains have been used in statistical physics and were assumed to be rapidly mixing, even in the case considered by the above authors.

Now, the geometric argument. Recall that

$$d(n) = \max_x \|\mathsf{P}^n(x,\cdot) - \pi(\cdot)\|_{\mathrm{TV}}$$

denotes the worst variation distance after n steps. Also, let $d(t)$ denote the corresponding worst case in continuous time. As before, the conductance $\tilde{\Phi}$ is given by

$$\tilde{\Phi} = \max_{A\subset\Omega} \tilde{\Phi}(A) \quad \text{where} \quad \tilde{\Phi}(A) = \frac{\mathsf{Q}(A,A^c)}{\pi(A)\pi(A^c)}$$

and when $A, B \subset \Omega$ then $\mathsf{Q}(A,B) = \sum_{x\in A, y\in B} \pi(x)\mathsf{P}(x,y)$.

Theorem 4.1. In discrete time and continuous time, respectively,

$$d(n) \geq \frac{1}{2}\,(1 - n\,\tilde{\Phi}) \quad \text{and} \quad d(t) \geq \frac{1}{2}\,(1 - t\tilde{\Phi})$$

Proof. We follow an argument of Borgs [7]. First, consider the discrete-time case.

Let $\tilde{\Phi}_n(A)$ denote the conductance of set A for the n-step Markov chain P^n. It suffices to show that for every set $A \subset \Omega$ that

$$d(n) \geq \frac{1}{2}\,|1 - \tilde{\Phi}_n(A)|$$

and

$$\tilde{\Phi}_{k+\ell}(A) \leq \tilde{\Phi}_k(A) + \tilde{\Phi}_\ell(A) \tag{4.1}$$

since the latter inequality implies that $\tilde{\Phi}_n(A) \leq n\tilde{\Phi}(A)$.

For the first inequality, let $\pi_A(x) = \frac{\pi(x)}{\pi(A)} 1_A(x)$ be the distribution induced on set A by π. Then

$$\begin{aligned} d(n) &\geq \|\pi - \pi_A\mathsf{P}^n\|_{TV} = \sup_{S\subset\Omega} |\pi(S) - (\pi_A\mathsf{P}^n)(S)| \qquad (4.2) \\ &\geq |\pi(A^c) - (\pi_A\mathsf{P}^n)(A^c)| = \pi(A^c)\left|1 - \tilde{\Phi}_n(A)\right| . \end{aligned}$$

The first equality is because $\|\sigma - \pi\|_{TV} = \max_{S\subset\Omega} |\sigma(S) - \pi(S)|$, with equality when $S = \{x \in \Omega : \sigma(x) \geq \pi(x)\}$ or $S = \{x \in \Omega : \sigma(x) <$

$\pi(x)\}$. Also, since $\tilde{\Phi}_n(A) = \tilde{\Phi}_n(A^c)$ then without loss assume that $\pi(A) \leq 1/2$.

For the second inequality, choose X_0 from distribution π, and let X_n denote the location of the walk at time n. Then

$$\tilde{\Phi}_n(A) = \frac{\mathsf{P}((X_n \in A^c) \cap (X_0 \in A))}{\pi(A)\pi(A^c)},$$

and so

$$\begin{aligned}
&\mathsf{P}((X_{k+\ell} \in A^c) \cap (X_0 \in A)) \\
&\quad = \mathsf{P}((X_{k+\ell} \in A^c) \cap (X_\ell \in A) \cap (X_0 \in A)) \\
&\qquad + \mathsf{P}((X_{k+\ell} \in A^c) \cap (X_\ell \in A^c) \cap (X_0 \in A)) \\
&\quad \leq \mathsf{P}((X_{k+\ell} \in A^c) \cap (X_\ell \in A)) + \mathsf{P}((X_\ell \in A^c) \cap (X_0 \in A)) .
\end{aligned}$$

Dividing both sides by $\pi(A)\pi(A^c)$ gives the result, as X_ℓ is drawn from distribution π.

The only place the discrete time assumption was used was to show that $\tilde{\Phi}_n(A) \leq n\tilde{\Phi}(A)$. However, $\tilde{\Phi}_{s+dt}(A) - \tilde{\Phi}_s(A) \leq \tilde{\Phi}_{dt}(A) = \tilde{\Phi}(A)dt$ by Equation (4.1), showing $\frac{d}{dt}\tilde{\Phi}_t(A) \leq \tilde{\Phi}(A)$. Integration implies that $\tilde{\Phi}_t(A) \leq t\,\tilde{\Phi}(A)$. □

Example 4.2. Consider the barbell given by connecting two copies of the complete graph K_m by a single edge, transitioning to a neighbor with probability $\mathsf{P}(x, y) = 1/2m$ if x and y are adjacent, and otherwise staying at state x. The conductance is $\tilde{\Phi} = 1/4m^2$, which gives lower bound $d(n) \geq 1 - \frac{n}{4m^2}$. This correctly shows that $\tau(1/2e) = \Omega(m^2)$.

Remark 4.3. A lower bound in terms of modified conductance $\tilde{\phi}$ would be even more useful than one in terms of conductance, as $\tilde{\phi} \leq \tilde{\Phi}$. Such a bound in terms of n-step modified conductance $\tilde{\phi}_n$ does in fact hold, and since $\tilde{\phi}_n(A) \leq \tilde{\Phi}_n(A) \leq n\tilde{\Phi}(A)$ then this is a stronger bound. For instance, when P is the simple random walk $\mathsf{P}(i, i \pm 1) = 1/2$ on a cycle $\mathbb{Z}/m\mathbb{Z}$ then $\tilde{\phi}_n = 0$ when m is even, while $\tilde{\phi}_n > 0$ when m is odd, correctly distinguishing between the non-convergent and the convergent cases. Unfortunately, it is not true that $\tilde{\phi}_n(A) \leq n\tilde{\phi}(A)$ (for instance, $\tilde{\phi} = 0$ but $\tilde{\phi}_2 = 1/2$ in the walk of Example 5.2), so this result is of questionable utility.

Remark 4.4. The same argument shows a bound in terms of $\bar{d}(n)$, defined in the next section:

$$\bar{d}(n) \geq |1 - \tilde{\Phi}_n| \geq 1 - n\tilde{\Phi}\,.$$

4.2 A spectral lower bound

Next we turn our attention to a spectral lower bound. Recall that $\tilde{\Phi} \geq \lambda \geq \Phi^2/2$, and so a lower bound in terms of the spectral gap λ can potentially offer a significant improvement over the geometric bound of the previous section. However, in Example 5.5 we give a non-reversible walk where the lower bound we might expect in terms of λ, in particular $d(n) \geq \frac{1}{2}(1 - n\lambda)$, cannot hold. Instead, we show a lower bound in terms of the eigenvalues of P; in the reversible case this gives a bound in terms of λ, and hence offers the hoped for improvement.

Example 4.5. Consider the simple random walk on the cycle $\mathbb{Z}/m\mathbb{Z}$ with $\mathsf{P}(i, i+1) = \mathsf{P}(i, i-1) = p$ and $\mathsf{P}(i,i) = 1 - 2p$. In Example 2.11 the lower bound $\tau(1/2e) = \Omega(m^2/p)$ was given by use of the Central Limit Theorem. However, when $p = 1/2$ this bound cannot distinguish between the $\Theta(m^2)$ mixing time when m is odd, and the non-convergent case when m is even. Neither can the geometric bound of Theorem 4.1 distinguish between these two cases (despite Remark 4.3, establishing a concrete lower bound on $\tilde{\phi}_n$ when $m \gg 1$ is not likely to be feasible). However, the eigenvalues of this walk are $\lambda_k = \cos(2\pi k/m)$ for $k = 0 \ldots (m-1)$, and so in particular when m is odd then $|\lambda_k| \leq \cos(\pi/m) \approx 1 - \frac{\pi^2}{2m^2}$ for all k, but when m is even then $|\lambda_{m/2}| = |-1| = 1$. Theorem 4.9 below states that $d(n) \geq \frac{1}{2}|\lambda_k|^n$, establishing a lower bound of the correct order for this walk both when m is odd and when it is even.

The proof of a spectral result will involve a new notion of closeness to stationarity.

$$\bar{d}(n) = \max_{x,y} \|\mathsf{P}^n(x,\cdot) - \mathsf{P}^n(y,\cdot)\|_{\mathrm{TV}}\,.$$

Also let $\bar{d}(t)$ denote the corresponding worst case in continuous time. The quantity $\bar{d}(n)$ measures the slowest rate at which a pair of initial distributions will converge towards each other, whereas $d(n)$ measures the slowest rate at which an initial distribution will converge to π. The following lemma shows that $\bar{d}(n)$ is closely related to the usual variation distance $d(n)$.

Lemma 4.6.

$$d(n) \leq \bar{d}(n) \leq 2d(n).$$

Proof. The proof is basically two applications of the triangle inequality.

Given probability distributions μ_1 and μ_2 then

$$\begin{aligned} \|\mu_1 \mathsf{P}^n - \mu_2 \mathsf{P}^n\|_{\mathrm{TV}} &= \left\| \sum_{x,y\in\Omega} \mu_1(x)\mu_2(y)(\mathsf{P}^n(x,\cdot) - \mathsf{P}^n(y,\cdot)) \right\|_{\mathrm{TV}} \\ &\leq \sum_{x,y\in\Omega} \mu_1(x)\mu_2(y)\|\mathsf{P}^n(x,\cdot) - \mathsf{P}^n(y,\cdot)\|_{\mathrm{TV}} \leq \bar{d}(n). \end{aligned} \tag{4.3}$$

The inequality $d(n) \leq \bar{d}(n)$ follows when $\mu_1 = 1_{\{x\}}$ and $\mu_2 = \pi$.
Given $x, y \in \Omega$ then

$$\|\mathsf{P}^n(x,\cdot) - \mathsf{P}^n(y,\cdot)\|_{\mathrm{TV}} \leq \|\mathsf{P}^n(x,\cdot) - \pi\|_{\mathrm{TV}} + \|\pi - \mathsf{P}^n(y,\cdot)\|_{\mathrm{TV}} \leq 2d(n).$$

□

Given (real or complex valued) vector v, let $\|v\|_{\mathrm{TV}} = \frac{1}{2}\sum_{x\in\Omega} |v(x)|$. The distance $\bar{d}(n)$ can be written in a form akin to an operator norm.

Lemma 4.7. If $|\Omega| = N$ then

$$\bar{d}(n) = \sup_{v\in\mathsf{R}^N,\, v\cdot 1=0} \frac{\|v\mathsf{P}^n\|_{\mathrm{TV}}}{\|v\|_{TV}} \geq \frac{1}{\sqrt{2}} \sup_{v\in\mathbb{C}^N,\, v\cdot 1=0} \frac{\|v\mathsf{P}^n\|_{\mathrm{TV}}}{\|v\|_{TV}}.$$

Proof. One direction of the real case is easy:

$$\bar{d}(n) = \sup_{x,y} \|(\delta_x - \delta_y)\mathsf{P}^n\|_{TV} \leq \sup_{v\in\mathsf{R}^N,\, v\cdot 1=0} \frac{\|v\mathsf{P}^n\|_{TV}}{\|v\|_{TV}}.$$

For the converse, without loss assume $\|v\|_{TV} = 1$. Let $v_+ = \max\{v, 0\}$ and $v_- = \max\{-v, 0\}$. If $v \cdot 1 = 0$ then $\sum_{x\in\Omega} v_+(x) =$

$\sum_{x\in\Omega} v_-(x) = \|v\|_{\mathrm{TV}}$, and so $v = v_+ - v_-$ is a difference of probability distributions. Equation (4.3) then shows that

$$\|v\mathsf{P}^n\|_{TV} = \|v_+\mathsf{P}^n - v_-\mathsf{P}^n\|_{TV} \leq \bar{d}(n)\,.$$

When $v \in \mathbb{C}^N$ with $v \cdot 1 = 0$, then write $v = (\mathrm{Re}\, v) + i(\mathrm{Im}\, v)$ with $\mathrm{Re}\, v$, $\mathrm{Im}\, v \in \mathsf{R}^N$. Then

$$\begin{aligned}\|v\mathsf{P}^n\|_{\mathrm{TV}} &\leq \|(\mathrm{Re}\, v)\mathsf{P}^n\|_{\mathrm{TV}} + \|(\mathrm{Im}\, v)\mathsf{P}^n\|_{\mathrm{TV}} \\ &\leq \bar{d}(n)\left(\|\mathrm{Re}\, v\|_{\mathrm{TV}} + \|\mathrm{Im}\, v\|_{\mathrm{TV}}\right) \\ &\leq \sqrt{2}\bar{d}(n)\|v\|_{\mathrm{TV}}\,.\end{aligned}$$

The first and last inequalities are due to the relation $|a+bi| \leq |a|+|b| \leq \sqrt{2}|a+bi|$ when $a, b \in \mathsf{R}$. The second inequality is from the real case, since $(\mathrm{Re}\, v) \cdot 1 = (\mathrm{Im}\, v) \cdot 1 = 0$. □

One consequence of this is the following lemma, which is traditionally shown via a coupling argument.

Lemma 4.8. Given $m, n \geq 0$ then

$$\begin{aligned}d(n+m) &\leq d(n)\bar{d}(m)\,, \\ \bar{d}(n+m) &\leq \bar{d}(n)\bar{d}(m)\,,\end{aligned}$$

and in particular,

$$\tau(\epsilon) \leq \tau(1/2e)\left\lceil \log\frac{1}{2\epsilon}\right\rceil\,.$$

Proof. The first bound follows easily from Lemma 4.7:

$$\begin{aligned}\|\delta_x\mathsf{P}^{n+m} - \pi\|_{TV} &= \|(\delta_x\mathsf{P}^n - \pi)\mathsf{P}^m\|_{TV} \\ &\leq \|\delta_x\mathsf{P}^n - \pi\|_{TV}\bar{d}(m)\,.\end{aligned}$$

The second bound follows similarly:

$$\begin{aligned}\|\delta_x\mathsf{P}^{n+m} - \delta_y\mathsf{P}^{n+m}\|_{TV} &= \|(\delta_x\mathsf{P}^n - \delta_y\mathsf{P}^n)\mathsf{P}^m\|_{TV} \\ &\leq \|\delta_x\mathsf{P}^n - \delta_y\mathsf{P}^n\|_{TV}\bar{d}(m)\,.\end{aligned}$$

By these relations and Lemma 4.6 we have

$$d(k\tau(1/2e)) \leq d(\tau(1/2e))\bar{d}(\tau(1/2e))^{k-1} \leq 2^{k-1}d(\tau(1/2e))^k \leq 1/2e^k\,.$$

□

With these lemmas it is fairly easy to show a lower bound on variation distance. See Seneta [74] for a similar lower bound in terms of the real eigenvalues.

Theorem 4.9. Let λ_{max} denote the second largest magnitude (complex valued) eigenvalue of P, and λ' be the non-trivial eigenvalue with largest real part. In discrete and continuous time respectively

$$d(n) \geq \frac{1}{2}\,|\lambda_{max}|^n \quad \text{while} \quad d(t) \geq \frac{1}{2}\,e^{-(1-\mathrm{Re}\lambda')\,t}\,,$$

that is, the discrete and continuous time mixing times are lower bounded respectively by

$$\tau(\epsilon) \geq \frac{\log(1/2\epsilon)}{\log(1/|\lambda_{max}|)} \geq \frac{|\lambda_{max}|}{1-|\lambda_{max}|}\log(1/2\epsilon) \quad \text{and} \quad \tau(\epsilon) \geq \frac{\log(1/2\epsilon)}{1-\mathrm{Re}\lambda'}\,.$$

Proof. If $v_i \in \mathbb{C}^N$ is a left eigenvector corresponding to eigenvalue $\lambda_i \neq 1$ of P then $v_i 1 = 0$, as shown in Equation (1.13). From Lemma 4.7 it then follows that

$$\bar{d}(n) \geq \frac{1}{\sqrt{2}}\,\frac{\|v_i\,\mathsf{P}^n\|_{\mathrm{TV}}}{\|v_i\|_{TV}} = \frac{1}{\sqrt{2}}\,\frac{\|\lambda_i^n\,v_i\|_{\mathrm{TV}}}{\|v_i\|_{TV}} = \frac{1}{\sqrt{2}}|\lambda_i|^n\,.$$

This can be sharpened a bit. By Lemma 4.8, if $k \in \mathbb{N}$ then

$$\bar{d}(n) = \sqrt[k]{\bar{d}(n)^k} \geq \sqrt[k]{\bar{d}(nk)} \geq \sqrt[k]{\frac{1}{\sqrt{2}}|\lambda_i|^{nk}} = \frac{1}{2^{1/2k}}\,|\lambda_i|^n\,,$$

and taking $k \to \infty$ it follows that

$$\bar{d}(n) \geq |\lambda_i|^n\,.$$

In continuous time the proof of the lemmas and theorem are similar. However, the eigenvalues of H_t are of form $e^{-(1-\lambda_i)\,t}$, and so their magnitudes are $\left|e^{-(1-\lambda_i)\,t}\right| = e^{-(1-\mathrm{Re}\lambda_i)\,t}$. □

Remark 4.10. For a reversible chain $1 - \lambda_1 = \lambda \leq \tilde{\Phi}$, as seen by setting $f = 1_A$ to be the indicator of a set in the definition of λ. It follows that if $\tilde{\Phi} \leq 1$ then

$$d(n) \geq \frac{1}{2}\,|\lambda_1|^n \geq \frac{1}{2}\,(1-\tilde{\Phi})^n\,,$$

which is a slight improvement over the bound in the previous section. It is unclear if the bounds are directly comparable for non-reversible chains.

Remark 4.11. The argument in this section is taken from [62]. When the eigenvector corresponding to an eigenvalue is known explicitly then stronger lower bounds are possible. Wilson [78] gives such a lower bound for real valued eigenvalues (e.g. a reversible walk), Saloff-Coste [72] extends this to complex eigenvalues, and Wilson [79] shows a related result in terms of eigenvalue / eigenvector pairs of a "lifted" chain. For instance, Wilson uses this final bound to show a lower bound of order $m^3 \log m$ on the mixing of the m-card Rudvalis shuffle, matching an upper bound of the same order, with the $\log m$ term arising from his stronger lower bound.

4.3 A log-Sobolev lower bound

Another lower bound on mixing time involves the log-Sobolev constant ρ, discussed earlier in Definition 1.9. The advantage of a bound in terms of this constant over that in terms of λ is that $\rho \leq \lambda/2$ (see Remark 1.11), and so a lower bound in terms of this may be significantly larger than the spectral bound in terms of $\lambda_1 = 1 - \lambda$.

Example 4.12. Recall that the random transposition shuffle on an m card deck is one in which a pair of cards are chosen uniformly and their positions are then swapped. This shuffle is known to have $\tau_2(1/e) = \Theta(m \log m)$. The spectral gap is $\lambda = \Theta(1/m)$ and so the spectral lower bound in the previous section only establishes the bound $\tau_2(1/e) = \Omega(m)$. However, $\rho = \Theta(1/m \log m)$ is smaller than λ, and it will be seen below that the continuous time version of this shuffle has $\tau_2(1/e) \geq \frac{1}{2\rho} = \Omega(m \log m)$, the correct bound.

The lower bound we give is taken from [25], but we discuss it again here as the original write-up omits many details.

Theorem 4.13. A reversible, continuous time Markov chain will satisfy

$$\frac{1}{2\rho} \leq \tau_2(1/e) \leq \frac{1}{4\rho}\left(4 + \log\log\frac{1}{\pi_*}\right).$$

Remark 4.14. Such a bound is not possible in discrete time. Consider the walk on the complete graph K_m with transitions $\forall x, y : \mathsf{P}(x,y) = 1/m$. This reaches stationary in a single step, so $\tau_2(1/e) = 1$. However, $\rho < 1/\log m$ because if $x \in \Omega$ then $\mathcal{E}(\delta_x, \delta_x) = \frac{m-1}{m^2} < \frac{1}{m}$ and $\operatorname{Ent}(\delta_x^2) = \frac{1}{m}\log m$.

Likewise, the bound does not hold for non-reversible chains either, even with a laziness assumption. See Example 5.5 for an example of a lazy walk on a pair of cycles $\mathbb{Z}/m\mathbb{Z}$, where $\rho = \Theta(1/m^2)$ but $\tau_2(1/e) = \Theta(m)$.

One consequence of this is a lower bound on the log-Sobolev constant.

Corollary 4.15. The log-Sobolev constant ρ and spectral gap λ of a (non-reversible) Markov kernel P satisfy

$$\frac{\lambda}{2} \geq \rho \geq \frac{\lambda}{2 + \log\frac{1-\pi_*}{\pi_*}}.$$

Diaconis and Saloff-Coste improve the lower bound slightly to $\rho \geq (1-2\pi_*)\frac{\lambda}{\log\frac{1-\pi_*}{\pi_*}}$, which is an equality for a walk on the complete graph K_m. Even our weaker corollary is still sharp, since for any walk on the two-point space $\{0,1\}$ with $\pi(0) = \pi(1) = 1/2$ the upper and lower bounds are the same, and so $\rho = \lambda/2$ in this case.

Proof. Suppose P is reversible. Then by Theorem 4.13 and Corollary 1.6 the continuous-time Markov chain associated with P will satisfy

$$\frac{1}{2\rho} \leq \tau_2(1/e) \leq \frac{1}{\lambda}\left(\frac{1}{2}\log\frac{1-\pi_*}{\pi_*} + 1\right).$$

Re-arranging terms gives the lower bound in the corollary. The upper bound follows from Remark 1.11.

In the general case, recall that $\mathcal{E}_{\mathsf{P}}(f,f) = \mathcal{E}_{\frac{\mathsf{P}+\mathsf{P}^*}{2}}(f,f)$, and so $\rho = \rho_{\frac{\mathsf{P}+\mathsf{P}^*}{2}}$ and $\lambda = \lambda_{\frac{\mathsf{P}+\mathsf{P}^*}{2}}$. The corollary, in the case of the reversible chain $\frac{\mathsf{P}+\mathsf{P}^*}{2}$, then implies the non-reversible case as well. □

In order to motivate the proof of the theorem, observe that (by a tedious but elementary calculation, see around Equation (3.2) of [25])) if $p : \mathsf{R}_+ \to [2,\infty)$ is differentiable, $p(0) = 2$, and $f : \Omega \to \mathsf{R}$, then

$$\frac{d}{dt}\bigg|_{t=0} \|H_t f\|_{p(t)} = \|f\|_2^{-1}\left(\frac{p'(0)}{4}\mathrm{Ent}(f^2) - \mathcal{E}(f,f)\right). \tag{4.4}$$

In this relation, if $p(t) = \frac{2}{1-t/\tau_2(\epsilon)}$ then $p'(0) = 2/\tau_2(\epsilon)$ and this derivative contains entropy, a Dirichlet form, and mixing time. This suggests that the derivative of an appropriate matrix norm type quantity might allow us to relate log-Sobolev and mixing time.

The key to studying the derivative of a matrix norm will be Stein's Interpolation Theorem. Recall that a function $h : \Omega \to \mathsf{R}$ is log-convex if its logarithm is convex, or equivalently if $h((1-s)x + s\,y) \leq h(x)^{1-s}h(y)^s$ when $s \in [0,1]$ and $x, y \in \Omega$.

Theorem 4.16 (Stein Interpolation, see [71, 76]). Suppose that $T(z) : \mathbb{C}^\Omega \to \mathbb{C}^\Omega$ is an operator on complex-valued functions which is defined for $z \in \mathbb{C}$. If $T(\cdot)$ is continuous on the strip $\{z : 0 \leq \mathrm{Re}(z) \leq 1\}$, uniformly bounded, and analytic in the interior of the strip, then

$$(\alpha, \beta, \gamma) \to \max_{y\in\mathsf{R}} \|T(\gamma + iy)\|_{1/\alpha\to 1/\beta}$$

is log-convex on $[0,1] \times [0,1] \times [0,1]$, where $\|M\|_{p\to q}$ is the smallest value satisfying the relation

$$\|M\,f\|_{q,\pi} \leq \|M\|_{p\to q}\,\|f\|_{p,\pi} \quad \text{for all} \quad f : \Omega \to \mathbb{C}.$$

Proof. [Proof of Theorem 4.13] The upper bound is from the reversible, continuous time case considered after Corollary 2.4.

For the lower bound, let E denote the expectation operator, which is just the square matrix in which every row is just π, and define $T(z) = H_{\tau z} - E = e^{-(\mathsf{I}-\mathsf{P})\tau z} - E$ with $\tau = \tau_2(\epsilon)$. This is continuous, analytic and uniformly bounded on $0 \leq \mathrm{Re} z \leq 1$, because e^{Az} has these properties for any choice of A. We will later show that $\|H_{iy} - E\|_{2\to 2} \leq 1$ and $\|H_{\tau+iy} - E\|_{2\to\infty} \leq \epsilon$. Then interpolation along the line $\ell(s) = (1-s)\left(\frac{1}{2},\frac{1}{2},0\right) + s\left(\frac{1}{2},0,1\right)$ for $s \in [0,1]$ yields the relation

$$\|H_{s\tau} - E\|_{2\to\frac{2}{1-s}} \leq \max_{a,b\in\mathsf{R}} \|H_{0+ia} - E\|_{2\to 2}^{1-s}\|H_{\tau+ib} - E\|_{2\to\infty}^{s} \leq \epsilon^s.$$

It follows that if $f : \Omega \to \mathsf{R}$ then

$$\epsilon^{-s}\|(H_{s\tau} - E)f\|_{2/(1-s)} \leq \|f\|_2 \,.$$

This is an equality at $s = 0$, and as the right side is constant in s then the derivative of the left side at $s = 0$ is non-positive. Equation (4.4) and the product rule show that

$$\begin{aligned} \left.\frac{d}{ds}\right|_{s=0} \epsilon^{-s}\, \|(H_{s\tau} - E)f\|_{\frac{2}{1-s}} &= \left.\frac{d}{ds}\right|_{s=0} \epsilon^{-s}\, \|H_{s\tau}(f - Ef)\|_{\frac{2}{1-s}} \\ = &\ \frac{1}{2\|f - Ef\|_2}\big(-2\|f - Ef\|_2^2 \log\epsilon + \text{Ent}((f - Ef)^2) - 2\tau\mathcal{E}(f, f)\big) \\ \leq &\ 0\,. \end{aligned}$$

Re-arranging terms, this becomes

$$\text{Ent}((f - Ef)^2) \leq 2\tau\mathcal{E}(f, f) + 2\text{Var}(f)\log\epsilon\,.$$

To go from this to a relation involving $\text{Ent}(f^2)$, we require an inequality of Rothaus [21].

$$\text{Ent}(f^2) \leq \text{Ent}((f - Ef)^2) + 2\text{Var}(f)\,.$$

Combining these two previous inequalities, and setting $\epsilon = 1/e$ and $\tau = \tau_2(1/e)$ then this implies that

$$\tau \geq \sup \frac{\text{Ent}(f^2)}{2\mathcal{E}(f, f)} = \frac{1}{2\rho}\,.$$

It remains only to show the assumptions about $T(z)$. The following three points are key to the argument, where $z = x + iy \in \mathbb{C}$ and $f : \Omega \to \mathbb{C}$:

(1) $H_z = H_{x+iy} = H_x\, H_{iy} = H_{iy}\, H_x$.
(2) H_{iy} is unitary and so $\|H_{iy}f\|_2 = \|f\|_2$.
(3) $H_z\, Ef = E\, H_z f = Ef$.

The first and third claims are easily verified. For the second, recall that a matrix T is unitary if $T^*\, T = I$. Then, if $f : \Omega \to \mathbb{C}$,

$$\|Tf\|_2^2 = \langle Tf, \overline{Tf}\rangle_\pi = \langle T^*Tf, \overline{f}\rangle_\pi = \langle f, \overline{f}\rangle_\pi = \|f\|_2^2$$

and so $\|T\|_{2\to 2}=1$. In the case at hand, P is reversible and so

$$H_{iy}^* = \left(e^{-iy(I-\mathsf{P})}\right)^* = \overline{e^{-iy(I-\mathsf{P}^*)}} = e^{iy(I-\mathsf{P}^*)} = e^{iy(I-\mathsf{P})}\,.$$

Checking the condition for a unitary matrix,

$$H_{iy}^*\,H_{iy} = e^{iy(I-\mathsf{P})}e^{-iy(I-\mathsf{P})} = I\,,$$

and so H_{iy} is unitary, and in particular $\|H_{iy}f\|_2=\|f\|_2$.

Now for the first desired bound, $\|H_{iy}-E\|_{2\to 2}\le 1$. If $f:\Omega\to\mathbb{C}$ then

$$\|(H_{iy}-E)f\|_2 = \|H_{iy}(f-Ef)\|_2 = \|f-Ef\|_2 \le \|f\|_2$$

where the inequality is because $\min_c \|f-c\|_2 = \|f-Ef\|_2$.

Now for the next bound, $\|H_{\tau_2(\epsilon)+iy}-E\|_{2\to\infty}\le\epsilon$. Let $f:\Omega\to\mathbb{C}$ and $\tau=\tau_2(\epsilon)$. Then

$$\begin{aligned}
\|(H_{\tau+iy}-E)f\|_\infty &= \|(H_\tau - E)H_{iy}f\|_\infty \\
&\le \|H_\tau - E\|_{2\to\infty}\|H_{iy}f\|_2 \\
&= \|H_\tau - E\|_{2\to\infty}^{\mathsf{R}}\|f\|_2 \\
&= \|H_\tau^* - E\|_{1\to 2}^{\mathsf{R}}\|f\|_2 \\
&\le \epsilon\,\|f\|_2
\end{aligned}$$

where $\|\cdot\|_{p\to q}^{\mathsf{R}}$ denotes the operator norm taken over real valued functions only.

The first equality and inequality in this bound are straightforward. Now, if $f:\Omega\to\mathbb{C}$ and T acts on functions, then

$$\begin{aligned}
\|Tf\|_\infty^2 &= \sup_x\left((T\mathrm{Re}f)(x)\right)^2 + \left((T\mathrm{Im}f)(x)\right)^2 \\
&\le \left(\|T\|_{2\to\infty}^{\mathsf{R}}\right)^2\left(\|\mathrm{Re}f\|_2^2+\|\mathrm{Im}f\|_2^2\right) = \left(\|T\|_{2\to\infty}^{\mathsf{R}}\right)^2\|f\|_2^2\,,
\end{aligned}$$

and so $\|T\|_{2\to\infty} = \|T\|_{2\to\infty}^{\mathsf{R}}$, which gives the second equality in the bound when $T=H_\tau - E$.

For the third equality, let $q\in[2,\infty]$ and $q^*\in[1,2]$ be conjugate exponents (i.e. $1/q+1/q^*=1$). If $f,g:\Omega\to\mathsf{R}$, and the operator norm is over real functions only, then

$$\|T\|_{2\to q} = \sup_{\|f\|_2=1}\|Tf\|_\infty = \sup_{\|f\|_2=1}\sup_{\|g\|_{q^*}=1}|\langle Tf,\,g\rangle_\pi|$$

$$\begin{aligned} &= \sup_{\|f\|_2=1} \sup_{\|g\|_{q^*}=1} |\langle f,\, T^* g\rangle_\pi| = \sup_{\|g\|_{q^*}=1} \sup_{\|f\|_2=1} |\langle T^* g,\, f\rangle_\pi| \\ &= \sup_{\|g\|_{q^*}=1} \|T^* g\|_2 = \|T^*\|_{q^*\to 2}\,. \end{aligned}$$

We used L^p duality via the relation $\|f\|_q = \sup_{\|g\|_{q^*}=1} |\langle f, g\rangle_\pi|$ in the argument above.

To bound the operator norm in the final inequality, without loss restrict attention to $g \geq 0$ and $\|g\|_{1,\pi} = 1$, i.e. g is a density. Denote this by $h_0 = g$ and $h_t = H_t^* g$. Then,

$$\|(H^*_{\tau_2(\epsilon)} - E)h_0\|_2 = \|h_{\tau_2(\epsilon)} - 1\|_2 \leq \epsilon$$

by definition of $\tau_2(\epsilon)$. □

4.4 Comparing Mixing Times

Recall from Section 2.3 that the spectral gap, log-Sobolev constant, Nash-inequality or spectral profile of a chain P can be bounded in terms of that of another chain $\hat{\mathsf{P}}$. However, suppose that none of these quantities are known for $\hat{\mathsf{P}}$, but it's mixing time is known (for instance by coupling). In this case it is still possible to use a comparison argument to say something about the mixing time of P.

Randall and Tetali [70] give a variety of problems where comparison of mixing times is the only method known for upper bounding convergence. One such problem is that of domino tilings. Suppose a rectangular region is tiled by dominoes (i.e. rectangles with dimensions 1×2). A natural Markov chain to sample domino tilings would choose a square 2×2 region, and if there are two parallel dominoes in it then rotate them by 90°, otherwise do nothing. The authors bound the mixing time of this walk by comparing it to a walk in which the steps involve modifying "towers" of dominoes at each step.

Now, recall that by Theorem 2.14 it follows that $\lambda_{\mathsf{P}} \geq \frac{1}{MA}\lambda_{\hat{\mathsf{P}}}$, where M and A are defined in the theorem. If (non-reversible) P has holding probability α then Corollary 1.14 shows that

$$\tau_{\mathsf{P}}(\epsilon) \leq \tau_{2,\mathsf{P}}(2\epsilon) \leq \left\lceil \frac{MA}{\alpha\,\lambda_{\hat{\mathsf{P}}}} \log \frac{1}{2\epsilon\sqrt{\pi_*}} \right\rceil . \tag{4.5}$$

More generally, if only a mixing time bound for some $\hat{\mathsf{P}}$ is known, then by Theorem 4.9 it follows that $d_{\hat{\mathsf{P}}}(n) \geq \frac{1}{2}\,|\lambda_1(\hat{\mathsf{P}})|^n$. If $\hat{\mathsf{P}}$ is reversible then

$$\lambda_{\hat{\mathsf{P}}} = 1 - \lambda_1(\hat{\mathsf{P}}) \geq 1 - \sqrt[n]{2d_{\hat{\mathsf{P}}}(n)}\,. \tag{4.6}$$

Equation (4.5) can now be applied, even if P is non-reversible.

If $\alpha \approx 0$ then the above result is insufficient. However, if $\hat{\mathsf{P}}$ is reversible then $d_{\hat{\mathsf{P}}}(n) \geq \frac{1}{2}\,\lambda_{max}(\hat{\mathsf{P}})^n$, and so $\lambda_{max}(\hat{\mathsf{P}}) \leq \sqrt[n]{2d_{\hat{\mathsf{P}}}(n)}$. By Theorem 2.16, if the path lengths are restricted to odd length then the smallest eigenvalue can be compared as well, and in particular it follows that

$$1 - \lambda_{max}(\mathsf{P}) \geq \frac{1 - \lambda_{max}(\hat{\mathsf{P}})}{MA^*} \geq \frac{1 - \sqrt[n]{2d_{\hat{\mathsf{P}}}(n)}}{MA^*}\,.$$

By the reversible case in Corollary 1.14,

$$\begin{aligned} \tau_{\mathsf{P}}(\epsilon) \leq \tau_{2,\mathsf{P}}(2\epsilon) &\leq \left\lceil \frac{1}{1-\lambda_{max}} \log \frac{1}{2\epsilon\sqrt{\pi_*}} \right\rceil \\ &\leq \left\lceil \frac{MA^*}{1 - \sqrt[n]{2d_{\hat{\mathsf{P}}}(n)}} \log \frac{1}{2\epsilon\sqrt{\pi_*}} \right\rceil. \end{aligned}$$

If $\hat{\mathsf{P}}$ is not reversible then Theorem 4.9 does not give useful information on spectral gap $\lambda_{\hat{\mathsf{P}}}$. Instead, recall from Theorem 4.1 that $d_{\hat{\mathsf{P}}}(n) \geq \frac{1}{2}\,(1 - n\tilde{\Phi}_{\hat{\mathsf{P}}})$. Re-arranging terms it follows that

$$\tilde{\Phi}_{\hat{\mathsf{P}}} \geq \max_{n>0} \frac{1 - 2d_{\hat{\mathsf{P}}}(n)}{n} \geq \frac{1 - 1/e}{\tau_{\hat{\mathsf{P}}}(1/2e)}\,.$$

By Theorem 5.7 we have

$$\lambda_{\hat{\mathsf{P}}} \geq \frac{\tilde{\Phi}^2_{\hat{\mathsf{P}}}}{4} \geq \frac{1}{11\tau_{\hat{\mathsf{P}}}(1/2e)^2}\,,$$

and Equation (4.5) then gives a mixing time bound for P.

What choice of n is good in the above bounds? One can let $n = \tau_{\hat{\mathsf{P}}}(1/2e)$ and use the approximation $1 - e^{-1/n} \geq 1/(n+1)$ for $n \geq 1$ to get the relation $\lambda_{\hat{\mathsf{P}}} \geq 1/(1 + \tau_{\hat{\mathsf{P}}}(1/2e))$ in Equation (4.6). However, generally bounds on distance are of the form $d(n) \leq B\,C^n$, while most mixing time bounds are of the form $\tau(\epsilon) \leq D + E\log(1/\epsilon)$, or equivalently $d(n) \leq e^{D/E}\, e^{-n/E}$. Taking $n \to \infty$ then we have the simpler expression $\lambda_{\hat{\mathsf{P}}} \geq 1 - C$ or $\lambda_{\hat{\mathsf{P}}} \geq 1 - e^{-1/E}$.

These various cases are summarized as follows:

Theorem 4.17. Suppose that P and $\hat{\mathsf{P}}$ are Markov chains, with holding probability α, relative density at most $M = \max_x \frac{\pi(x)}{\hat{\pi}(x)}$, and let A and A^* be as in Theorems 2.14 and 2.16. Then,

if $\hat{\mathsf{P}}$ *and* P *reversible:*

$$\tau_{\mathsf{P}}(\epsilon) \leq \tau_{2,\mathsf{P}}(2\epsilon) \quad \leq \left\lceil MA^*(1 + \tau_{\hat{\mathsf{P}}}(1/2e)) \log \frac{1}{2\epsilon\sqrt{\pi_*}} \right\rceil$$

if $\hat{\mathsf{P}}$ *reversible,* P *not:*

$$\tau_{\mathsf{P}}(\epsilon) \leq \tau_{2,\mathsf{P}}(2\epsilon) \quad \leq \left\lceil \frac{MA}{\alpha}(1 + \tau_{\hat{\mathsf{P}}}(1/2e)) \log \frac{1}{2\epsilon\sqrt{\pi_*}} \right\rceil$$

if $\hat{\mathsf{P}}$ *not reversible:*

$$\tau_{\mathsf{P}}(\epsilon) \leq \tau_{2,\mathsf{P}}(2\epsilon) \quad \leq \left\lceil \frac{11MA}{\alpha}\tau_{\hat{\mathsf{P}}}(1/2e)^2 \log \frac{1}{2\epsilon\sqrt{\pi_*}} \right\rceil$$

In short, little is lost if P and $\hat{\mathsf{P}}$ are reversible, while if P is non-reversible and $\alpha \approx 0$ then problems occur, and in the worst case if $\hat{\mathsf{P}}$ is non-reversible then its mixing time must be squared. See Example 5.4 for examples showing that the various cases just given above are necessary, and not just artifacts of the method of proof.

Related results hold for continuous time chains as well, with the advantage of there being no need for α. The details are similar and are left to the interested reader. See [29] for additional examples, for discussion on the continuous time case, and for an alternate comparison method involving vertex congestion instead of edge congestion $A(\Gamma)$.

5

Examples

We start with elementary examples illustrating sharpness of the various bounds derived in this survey. This is followed by discussion of time for Pollard's Rho algorithm for the discrete logarithm to have a collision, a case of studying mixing of a non-reversible chain with no holding probability. The next section involves turning the tables a bit, and instead of using Cheeger inequalities to study mixing, we use mixing results to generalize Cheeger inequalities to non-reversible Markov chains. The chapter finishes with a discussion of the Thorp shuffle, a problem in which the full generality of methods in this volume are required, that is, methods for non-reversible chains with no holding probability and a spectral or evolving set profile.

5.1 Sharpness of bounds

In this section we give examples demonstrating the strengths and weaknesses of the various results developed in this survey. These include mixing time upper bounds (Examples 5.1, 5.2 and 5.3), mixing time lower bounds (Examples 5.2, 5.3 and 5.5), reversibility and non-reversibility issues (Example 5.2, 5.3 and 5.5), comparison of mixing times (Example

5.4), and an example which illustrates both why squared terms occur in several bounds as well as why the log-Sobolev lower bound cannot be generalized to non-reversible walks (Example 5.5).

Perhaps the simplest test case is random walk on the complete graph K_m.

Example 5.1. Given $\alpha \in [-\frac{1}{m-1}, 1]$ consider the walk on K_m with $\mathsf{P}(x,y) = (1-\alpha)/m$ for all $y \neq x$ and $\mathsf{P}(x,x) = \alpha + (1-\alpha)/m$, that is, choose a point uniformly at random and move there with probability $1-\alpha$, otherwise do nothing.

The n step distribution is $\mathsf{P}^n(x,x) = \frac{1}{m} + \alpha^n\left(1 - \frac{1}{m}\right)$ and $\mathsf{P}^n(x,y) = \frac{1}{m} - \frac{\alpha^n}{m}$ for all $y \neq x$. Therefore, when $\alpha \in [0,1]$ then $\mathsf{D}(\mathsf{P}^n(x,\cdot)\|\pi) = (1+o_m(1))\alpha^n \log m$ as $m \to \infty$. When $\alpha \in \left[\frac{-1}{m-1},\, 1\right]$ then $\|\mathsf{P}^n(x,\cdot) - \pi\|_{\mathrm{TV}} = |\alpha|^n(1-1/m)$ and $\|\mathsf{P}^n(x,\cdot) - \pi\|_2 = |\alpha|^n\sqrt{m-1}$.

First, let us consider spectral methods. This walk has trivial eigenvalue 1 and $m-1$ copies of eigenvalue α. It follows that $\lambda_{max} = |\alpha|$ and hence by Corollary 1.14

$$\|\mathsf{P}^n(x,\cdot) - \pi\|_2 \leq |\alpha|^n\sqrt{m-1}\,,$$

the correct bound.

Now for evolving sets. If $\alpha \in [0,1]$ then

$$\pi(A_u) = \begin{cases} 0 & \text{if } u \in (\alpha + (1-\alpha)\pi(A), 1] \\ \pi(A) & \text{if } u \in ((1-\alpha)\pi(A), \alpha + (1-\alpha)\pi(A)] \\ 1 & \text{if } u \in [0, (1-\alpha)\pi(A)] \end{cases}$$

A quick calculation shows that $\mathcal{C}_{z(1-z)} = \mathcal{C}_{z\log(1/z)} = \mathcal{C}_{\sqrt{z(1-z)}} = \alpha$, and so Corollary 3.9 implies $\|\mathsf{P}^n(x,\cdot) - \pi\|_{\mathrm{TV}} \leq \alpha^n\,(1-1/m)$, $\mathsf{D}(\mathsf{P}^n(x,\cdot)\|\pi) \leq \alpha^n \log m$ and $\|\mathsf{P}^n(x,\cdot) - \pi\|_2 \leq \alpha^n\sqrt{m-1}$. Total variation and L^2 bounds are correct, while relative entropy is asymptotically correct.

When $\alpha \in \left[\frac{-1}{m-1},\, 0\right)$ then a similar calculation shows that $\mathcal{C}_{z(1-z)} = \mathcal{C}_{\sqrt{z(1-z)}} = -\alpha$ and so $\|\mathsf{P}^n(x,\cdot) - \pi\|_{\mathrm{TV}} \leq (-\alpha)^n\,(1-1/m)$ and $\|\mathsf{P}^n(x,\cdot) - \pi\|_2 \leq (-\alpha)^t\sqrt{m-1}$. Again, both are exact.

The lower bound of Theorem 4.9 is

$$\|\mathsf{P}^n(x,\cdot) - \pi\|_{\mathrm{TV}} \geq \frac{1}{2}|\alpha|^n$$

which is quite close to the correct value of $|\alpha|^n\,(1-1/m)$.

The conductance of the walk is $\tilde{\Phi} = 1-\alpha$ and so Theorem 4.1 implies $d(n) \geq \frac{1}{2}\,(1-n(1-\alpha))$ which is comparable to the spectral bound only when α is close to one.

In the non-reversible case, even when the discrete time upper bounds fail the lower bound may be useful.

Example 5.2. Consider a non-reversible walk on the triangle $\{0,1,2\}$ with $\mathsf{P}(0,1) = \mathsf{P}(1,2) = 1$ and $\mathsf{P}(2,2) = \mathsf{P}(2,0) = 1/2$. Then $\pi(0) = \pi(1) = 1/4$, $\pi(2) = 1/2$, the chain mixes quite rapidly, but $\lambda_{\mathsf{PP}^*} = 0$ and $\mathcal{C}_f = 1$ for any choice of $f(x)$, so all discrete time upper bounds fail. This failure occurs because these methods require that distance decrease at *every step*, but if the initial distribution is all at vertex $\{0\}$ then it will take two steps before distance decreases (once vertex $\{2\}$ is reached).

The continuous time bounds are better. Observe that $\tilde{\Phi} = 1$ and $\lambda \geq \tilde{\Phi}^2/4 = 1/4$, and so Corollary 1.6 and Equation 3.11 both show mixing in time $\tau_2(\epsilon) = O(\log(1/\epsilon))$, which is correct.

The eigenvalues of this chain are $\lambda_i = 1,\ \frac{-1\pm i\sqrt{7}}{4}$ and so $|\lambda_{max}| = 1/\sqrt{2}$ while $\mathrm{Re}\lambda' = -1/4$. The mixing time lower bounds are then

$$d(n) \geq \frac{1}{2^{1+n/2}} \quad \text{and} \quad d(t) \geq \frac{1}{2e^{5t/4}}\,.$$

The conductance is $\tilde{\Phi} = 1$ and the bound of Theorem 4.1 provides no information.

Generally, the discrete-time upper bounds in this paper tend to be poor for non-reversible chains which are strongly inclined to move to a specific neighbor, as in the previous case. We give now a more interesting example of this; a chain constructed by Diaconis, Holmes and Neal [23] specifically for the purpose of speeding mixing.

Example 5.3. Consider a random walk on the cycle $\mathbb{Z}/2m\mathbb{Z}$, labeled as

$$\Omega = \{-(m-1), -(m-2), \ldots, 0, \ldots, (m-1), m\}$$

with transitions $\mathsf{P}(i, i+1) = 1 - 1/m$ and $\mathsf{P}(i, -i) = 1/m$. Diaconis, Holmes and Neal [23] show that $\tau(\epsilon) = \Theta(m \log(1/\epsilon))$ and $\tau_2(\epsilon) = \Theta(m \log(m/\epsilon))$, both much faster than the time $\Theta(m^2 \log(1/\epsilon))$ required by a simple random walk on a cycle.

The discrete-time upper bound on mixing time given in Corollary 1.14 requires computation of λ_{PP^*}. Now, PP^* is given by $\mathsf{PP}^*(i, -i-1) = \frac{2}{m}(1 - \frac{1}{m})$ and $\mathsf{PP}^*(i, i) = (1 - 1/m)^2 + (1/m)^2 = 1 - \frac{2}{m}\left(1 - \frac{1}{m}\right)$. The space is then disjoint, such that no transitions can be made between set $A = \{-\lfloor m/2 \rfloor, \ldots, \lfloor m/2 \rfloor - 1\}$ and its complement. In particular, if $f = 1_A$ then $\mathcal{E}_{\mathsf{PP}^*}(f, f) = 0$ and so $\lambda_{\mathsf{PP}^*} = 0$. Again, the upper bound tells nothing. We note in passing that evolving set ideas do no better, because if $A = \{1, -2\}$ and $B = \Omega \setminus \{2, -1\}$ then $\pi(B) = \pi(A^c)$ and yet $\tilde{\phi}(A) \leq \frac{\mathsf{Q}(A,B)}{\pi(A)\pi(A^c)} = 0$; hence $0 = \tilde{\phi} \geq \mathcal{C}_f$ and so $\mathcal{C}_f = 0$ for all concave functions f.

In contrast, consider the continuous-time case. In [29] it is observed that $\lambda = O(1/m^2)$, as $\lambda \leq \mathcal{E}(|i|, |i|)/\mathrm{Var}_\pi(|i|)$. Of course, every set A will have at least one transition $i \to i+1$ from some $i \in A$ to $i + 1 \notin A$, and for instance the set $A = \{-\lfloor m/2 \rfloor, \ldots, \lfloor m/2 \rfloor\}$ has exactly one transition to its complement, so $\Phi = \Theta(1/m)$. Consequently, $\lambda \geq \Phi^2/2 = \Omega(1/m^2)$, and so $\lambda = \Theta(1/m^2)$. This time the conductance and spectral continuous-time upper bounds are both $\tau_2(\epsilon) = O(m^2 \log(m/\epsilon))$.

To lower bound mixing, recall that $\tilde{\Phi} = \Theta(1/m)$. Also, note that if $f(i) = (-1)^i$ for all $i \in \mathbb{Z}/2m\mathbb{Z}$, then $\mathsf{P}(f)(i) = -(1 - \frac{2}{m})(-1)^i$, and so $\lambda_k = -(1 - 2/m)$ is an eigenvalue of P. Then Theorems 4.1 and 4.9 both show that in both discrete and continuous time $\tau(1/2e) = \Omega(m)$, and so once again even though the upper bound on mixing is either weak or useless, the lower bound is of the correct order.

Many of the fast mixing non-reversible chains are designed in a similar fashion, moving with high probability to a specific neighbor, and while this can speed mixing, it worsens our bounds because PP^* becomes like the identity matrix. Another instance of similar behavior can be found in Example 5.5 later.

We now consider the various cases appearing in Theorem 4.17, the theorem on comparison of mixing times. Simple examples are given showing that each of the cases in the theorem are necessary. In each case the stationary distribution is uniform, so $M = 1$.

Example 5.4. First, suppose $\hat{\mathsf{P}}$ and P are the (reversible) lazy simple walks on the complete graph K_m with $\mathsf{K}(x,x) = 1/2$ and $\mathsf{K}(x,y) = 1/2(m-1)$ if $y \neq x$. Then $\tau_{\hat{\mathsf{P}}}(1/2e) = 3$, while $\tau_{2,\mathsf{P}}(1/2e) = \Theta(\log m)$ (see Example 5.1, with $\alpha = \frac{1}{2}(1 - \frac{1}{m-1}) \approx 1/2$).

For the remaining cases, we consider variants of walks on a cycle $\mathbb{Z}/m\mathbb{Z}$ of even length m. Please check Equation (2.4) to recall the mixing time of this walk.

Now, suppose $\hat{\mathsf{P}}$ is the (reversible) lazy simple walk (i.e. $\hat{\mathsf{P}}(i,i) = 1/2$ and $\hat{\mathsf{P}}(i, i \pm 1) = 1/4$), but P is the (reversible) periodic simple random walk (i.e. $\mathsf{P}(i, i \pm 1) = 1/2$). Then $\hat{\mathsf{P}}$ converges while P does not. This is reflected in A^*, because $\forall x \in \Omega :\ \hat{\mathsf{P}}(x,x) > 0$ and so an odd length path γ_{xx} must be given. However, this must then include a loop at some vertex j, with probability $\mathsf{P}(j,j) = 0$, and so $A^* = \infty$.

Next, if $\hat{\mathsf{P}}$ is still the (reversible) lazy simple walk, but P is the (non-reversible) walk with counterclockwise drift given by $\mathsf{P}(i, i-1) = 1 - e^{-m}$ and $\mathsf{P}(i,i) = e^{-m}$, then $\tau_{\hat{\mathsf{P}}}(1/2e) = \Theta(m^2)$ while $\tau_{2,\mathsf{P}}(1/2e) = \Theta(m^2\, e^m)$ and so the $\alpha = e^{-m}$ term is necessary in this case.

Last of all, we consider a variant of Example 5.3. Consider the walks in Figure 5.1 on two copies of the cycle $\mathbb{Z}/m\mathbb{Z}$, denoted as -1 (counterclockwise) and $+1$ (clockwise). Let $\hat{\mathsf{P}}$ be a lazy (non-reversible) walk which half the time does nothing, while with probability $\frac{1}{2}(1 - 1/100m)$ it moves according to the sign of the cycle it is on, and with probability $1/200m$ it changes to the other copy (but otherwise keeps the same position). Also, let P be a (reversible) walk defined the same, except that with equal probabilities $\frac{1}{4}(1 - 1/100m)$ it will move in direction $+1$ or -1, regardless of the cycle it is on. The walk $\hat{\mathsf{P}}$ circles the cycle roughly $100m$ times before changing cycles, and does this at a point essentially uniformly distributed, so $\tau_{\hat{\mathsf{P}}}(1/2e) = \Theta(m)$. Meanwhile, the position P takes on the cycles is basically the same as that of a lazy walk on a single cycle, and so $\tau_{2,\mathsf{P}}(1/2e) = \Theta(m^2)$. When $\hat{\mathsf{P}}(x,y) > 0$ then $\mathsf{P}(x,y) \geq \hat{\mathsf{P}}(x,y)/2$, so $A \leq 2$ is insignificant. This explains the need for squaring when $\hat{\mathsf{P}}$ is non-reversible.

The example above, of a walk on two copies of a cycle, illustrates the reason for which our reversible and non-reversible results often differ by square factors. Let us explore this more carefully.

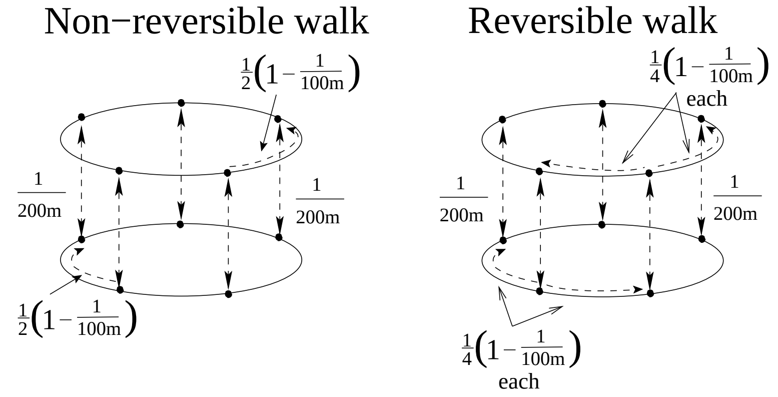
Non−reversible walk
Reversible walk
$\frac{1}{2}\left(1-\frac{1}{100m}\right)$
$\frac{1}{200m}$
$\frac{1}{200m}$
$\frac{1}{2}\left(1-\frac{1}{100m}\right)$
$\frac{1}{4}\left(1-\frac{1}{100m}\right)$
each
$\frac{1}{200m}$
$\frac{1}{200m}$
$\frac{1}{4}\left(1-\frac{1}{100m}\right)$
each

Fig. 5.1 Non-reversible and reversible walks on a pair of cycles.

Example 5.5. We now consider more carefully the walks $\hat{\mathsf{P}}$ and P (non-reversible and reversible, respectively) on two copies of a cycle, as given at the end of the previous example. Observe that in this final example, $\frac{\mathsf{P}+\mathsf{P}^*}{2} = \frac{\hat{\mathsf{P}}+\hat{\mathsf{P}}^*}{2}$, and so even though the mixing times are substantially different the conductance and spectral gap will be the same (as $\mathcal{E}_{\hat{\mathsf{P}}}(f,f) = \mathcal{E}_{\frac{\hat{\mathsf{P}}+\hat{\mathsf{P}}^*}{2}}(f,f) = \mathcal{E}_{\mathsf{P}}(f,f)$). This illustrates the reason for which our results are often not sharp on non-reversible chains, as the reversible version of the same chain may mix much more slowly.

By the remark above, it suffices to calculate $\tilde{\Phi}$, λ, and ρ for P. Now, $\tilde{\Phi} = 1/100m$, with the extreme case $\tilde{\Phi}(A)$ where A is one of the cycles. Likewise, $\lambda = \Theta(1/m^2)$ because $\lambda \geq \tilde{\Phi}^2/4 = \Omega(1/m^2)$ by Cheeger's inequality, but $\lambda \leq \frac{1}{2}(1-\cos(2\pi/m)) \approx \frac{\pi^2}{m^2}$ as $\frac{1}{2}(1+\cos(2\pi/m))$ is an eigenvalue with eigenfunction $f(j) = \cos(2\pi j/m)$ on both cycles (where $j \in [0 \ldots m-1]$ indicates position on the cycle). Moreover, the log-Sobolev constant satisfies $\rho \leq \frac{\lambda}{2} = O(1/m^2)$ by Proposition 1.10, while in continuous time $\tau_2(1/e) = \Theta(m^2)$ and so $\rho \geq \frac{1}{2\tau_2(1/e)} = \Omega(1/m^2)$ by Theorem 4.13, which combine to show $\rho = \Theta(1/m^2)$.

Now let us consider the various mixing bounds. The upper bound $\tau_2(\epsilon) \leq \frac{1}{\lambda}\log(1/\epsilon\sqrt{\pi_*})$ is then asymptotically correct for the reversible walk P, but about the square of the correct value for the non-reversible walk $\hat{\mathsf{P}}$; the same holds for the bound $\tau_2(\epsilon) \leq \frac{1}{\tilde{\Phi}^2}\log(1/\epsilon\sqrt{\pi_*})$. Turning to lower bounds, the lower bound $d(n) \geq \frac{1}{2}(1-\tilde{\Phi})^n$ will be of the correct order for the non-reversible walk $\hat{\mathsf{P}}$, but far too pessimistic for the reversible walk P. The spectral lower bound for the reversible walk P will be $\frac{1}{2}(1-\lambda)^n \leq d(n)$, which is of the correct order, while for the non-reversible walk $\hat{\mathsf{P}}$ an eigenvalue satisfies $|\lambda_i| = 1 - \Theta(1/m)$, which gives a correct order lower bound. Finally, in continuous time $\tau_{2,\mathsf{P}}(1/2e) = \Theta(m^2)$ and $\tau_{2,\hat{\mathsf{P}}}(1/2e) = \Theta(m)$ due to the same reasons given earlier for the discrete-time walks. Since $\rho = \Theta(1/m^2)$, then the continuous time lower bound of Theorem 4.13 in terms of log-Sobolev constant ρ gives the correct bound for the reversible walk P. This also shows why Theorem 4.13 cannot apply to non-reversible walks, because it would imply that $\tau_{\hat{\mathsf{P}}} = \Omega(m^2)$, which is incorrect.

5.2 Discrete Logarithm

Certain encryption algorithms rely on the difficulty of finding a discrete logarithm, that is, given that $y = x^k$ for fixed x and y in some finite cyclic group G, determine the power k. The best algorithm for finding discrete logarithm over a general cyclic group seems to be the Pollard Rho algorithm, and it is widely conjectured to require $O(\sqrt{n})$ steps for the algorithm to work, where $n = |G|$. Miller and Venkatesan [56] recently gave a proof that the Pollard Rho algorithm for discrete logarithm runs in time $\sqrt{n}\log^3 n$, the first result within logarithmic factors of the conjectured bound. In this section we discuss their result.

Pollard's Rho algorithm works by taking a random walk with transitions $\mathsf{P}(z, zx) = \mathsf{P}(z, zy) = \mathsf{P}(z, z^2) = 1/3$ and stopping at the first collision time, that is, the first time the Markov chain returns to a previously visited state. Equivalently, consider the random walk on a cycle $\mathbb{Z}/p\mathbb{Z}$ with transitions given by $\mathsf{P}(i, i+1) = \mathsf{P}(i, i+k) = \mathsf{P}(i, 2i) = 1/3$. As we will see below, to find the first collision time it suffices to find L^∞ mixing time, but since the random walk is non-reversible and nonlazy then classical bounds in terms of λ could not be used. However, by Equation (7.4) in the Appendix and Proposition 1.12 it follows that if $a, b \in \mathbb{Z}/p\mathbb{Z}$ then

$$|k_n^a(b) - 1| \leq \|k_0^a - 1\|_2 \|k_n^{*b} - 1\|_2 \leq \|\mathsf{P} - E\|_{2\to 2}^n \, \frac{1-\pi_*}{\pi_*}\,,$$

where $\|\mathsf{P} - E\|_{2\to 2} = \sup_{f:\Omega\to\mathsf{R}} \frac{\|\mathsf{P}(f) - \mathbb{E}_\pi f\|_2}{\|f\|_2}$, as in Definition 1.18 and Remark 1.19. Consequently,

$$\tau_\infty(\epsilon) \leq \left\lceil \frac{1}{1 - \|\mathsf{P} - E\|_{2\to 2}} \log \frac{1-\pi_*}{\epsilon\pi_*} \right\rceil\,.$$

The authors' main results in [56] are a proof of a related (but equivalent) statement in the special case of a simple random walk on a regular degree directed graph, and a proof that for some $c \in \mathsf{R}_+$ that

$$\|\mathsf{P}f\|_2 \leq \left(1 - \frac{1}{c(\log n)^2}\right) \|f\|_2 \quad \text{for all } f : \Omega \to \mathbb{C},\, \mathbb{E}f = 0\,. \tag{5.1}$$

Recall from Remark 1.19 that this is equivalent to the statement that

$$\|\mathsf{P} - E\|_{2\to 2} \leq 1 - \frac{1}{c(\log n)^2}\,.$$

It follows that

$$\tau_\infty(\epsilon) \leq \left\lceil c\log^3 n + c\log^2 n \log\frac{1}{\epsilon}\right\rceil\,.$$

To check for a collision, let S consist of the states visited in the first $t = \lfloor\sqrt{n}\rfloor$ steps. If $|S| < t$ then a collision occurred, and we are done, so without loss assume $|S| = t$. Now, $\tau_\infty(1/2) \leq (c+1)\log^3 n$, and so every $(c+1)\log^3 n$ steps there is a $\geq \pi(S)/2 \approx 1/2\sqrt{n}$ chance of ending in set S, and having a collision. In particular, in $2\sqrt{n}\log(1/\epsilon)$ repetitions of a $(c+1)\log^3 n$-step process, the chance of never ending in S is at most

$$\mathit{Prob} \leq (1 - 1/2\sqrt{n})^{2\sqrt{n}\log(1/\epsilon)} \leq \epsilon.$$

More generally, this shows that if T denotes the first collision time of a Markov chain then

$$\mathit{Prob}(T > 2\sqrt{n}\tau_\infty(1/2)\log(1/\epsilon)) \leq \epsilon\,.$$

Remark 5.6. Miller and Venkatesan's proof of Equation 5.1 was based on considering characters of a transition matrix, and reducing the problem to one of bounding a quadratic form. We suggest here that a more elementary argument might be possible, as it is illustrative of how techniques in this survey can be extended further.

Observe that in $3\log_2 n$ steps the walk will make an $i \to 2i$ step about $\log_2 n$ times, for a total distance traveled of at least $2^{\log_2 n} = n$, i.e. at least one circuit of the cycle is completed. This suggests every pair of vertices $x, y \in \Omega$ can be connected by canonical paths of length $|\gamma_{xy}| = O(\log n)$. No particular edge appears to be a bottleneck in this graph, so each edge should have about

$$O\left(\frac{\binom{|\Omega|}{2} * \max|\gamma_{xy}|}{|\Omega| \min_{x\in\Omega} deg(x)}\right) = O(|\Omega| \log n)$$

paths passing through it, making for congestion $A = O(\log^2 n)$. Hence, perhaps canonical paths can be used to show $\lambda_{\mathsf{P}} = \Omega(1/\log^2 n)$.

Unfortunately, to upper bound mixing time with Corollary 1.14 requires consideration of λ_{PP^*}, which is only lower bounded by λ_{P} when the holding probability α is non-zero. This is sufficient to study a lazy version $\mathsf{P}' = \frac{1}{2}(I + \mathsf{P})$ of the Pollard Rho walk, as $\lambda_{\mathsf{P}'\mathsf{P}'^*} \geq \lambda_{\mathsf{P}'} = \frac{1}{2}\lambda_{\mathsf{P}}$, but this may slow the walk by a factor of two.

Instead, consider a slightly modified walk $\hat{\mathsf{P}}$ with $\hat{\mathsf{P}}(i, i{+}1) = \hat{\mathsf{P}}(i, i{+}2) = 1/6$ and $\mathsf{P}(i, i + k) = \mathsf{P}(i, 2i) = 1/3$; that is, split the $i \to i + 1$ step into two parts. This might potentially speed the walk, as the local $i \to i + 1$ move can now go twice as far with $i \to i + 2$.

The lazy version of this walk, $M = \frac{1}{2}(I + \hat{\mathsf{P}})$ will satisfy $\lambda_{MM^*} \geq \lambda_M = \frac{1}{2}\lambda_{\hat{\mathsf{P}}}$, and so it remains to relate λ_{MM^*} to $\lambda_{\hat{\mathsf{P}}\hat{\mathsf{P}}^*}$. Instead, recall from Remark 1.16 that since $\mathsf{K} = \hat{\mathsf{P}}\hat{\mathsf{P}}^*$ is reversible then

$$\lambda_{\hat{\mathsf{P}}\hat{\mathsf{P}}^*} = \lambda_{\mathsf{K}} \geq \frac{1}{2}\lambda_{\mathsf{KK}^*} = \frac{1}{2}\,\lambda_{(\hat{\mathsf{P}}\hat{\mathsf{P}}^*)^2}\,.$$

Furthermore, a direct computation verifies that

$$\forall x \neq y : \, (\hat{\mathsf{P}}\hat{\mathsf{P}}^*)^2(x, y) \geq \frac{1}{54}(MM^*)(x, y)$$

and so $\mathcal{E}_{(\hat{\mathsf{P}}\hat{\mathsf{P}}^*)^2}(f, f) \geq \frac{1}{54}\mathcal{E}_{MM^*}(f, f)$, and in particular, $\lambda_{(\hat{\mathsf{P}}\hat{\mathsf{P}}^*)^2} \geq \frac{1}{54}\lambda_{MM^*}$.

Putting all these statements together, we have that

$$\lambda_{\hat{\mathsf{P}}\hat{\mathsf{P}}^*} \geq \frac{1}{2}\lambda_{(\hat{\mathsf{P}}\hat{\mathsf{P}}^*)^2} \geq \frac{1}{108}\lambda_{MM^*} \geq \frac{1}{216}\lambda_{\hat{\mathsf{P}}}\,,$$

and so a canonical path result for $\lambda_{\hat{\mathsf{P}}}$ will suffice to show mixing of this walk, even though it has zero holding probability.

More generally, given any walk K, with lazy version $M = \frac{1}{2}(I+\mathsf{K})$, if $\forall x \neq y:\ (\mathsf{K}\mathsf{K}^*)^2(x,y) \geq c\,MM^*(x,y)$ for some $c > 0$ then $\lambda_{\mathsf{K}\mathsf{K}^*} \geq \frac{c}{4}\lambda_{\mathsf{K}}$, and in particular the mixing time can be bounded by constructing canonical paths.

5.3 New Cheeger Inequality

Early results in bounding mixing times relied heavily on Cheeger's inequality,

$$\tilde{\Phi} \geq \lambda \geq 1 - \sqrt{1 - \Phi^2} \geq \Phi^2/2\,.$$

Cheeger's inequality was originally shown via a direct lower bounding of the eigenvalue gap of a reversible chain, and then applied to bound mixing times. In this section we use mixing results to give an alternate method of proof for Cheeger-like inequalities. First, we consider a related inequality in terms of $\tilde{\Phi}$, instead of Φ, a bound which will be used to slightly improve bounds for the Thorp shuffle. Next, we generalize this and show that the f-congestion of $(\mathsf{P} + \mathsf{P}^*)/2$ provides a more general lower bound for $1 - \lambda$. Finally, given that Cheeger's inequality was originally shown in terms of eigenvalues of reversible chains, it is natural to wonder if a similar relation holds for eigenvalues of non-reversible chains. In an interesting turn of the tables, it is possible to use our mixing time results to answer this question.

The tools developed in the previous two chapters can be used to prove an alternate form of Cheeger's inequality, which is sometimes stronger:

Theorem 5.7.

$$\lambda \geq 2\left(1 - \sqrt{1 - \tilde{\Phi}^2/4}\right) \geq \tilde{\Phi}^2/4\,.$$

Proof. Suppose the Markov chain of interest is reversible. Then $\lambda = 1 - \lambda_1$. It follows from Theorem 4.9 that

$$d(t) \geq \frac{1}{2}\, e^{-(1-\max_{i\neq 0} \mathrm{Re}\lambda_i)t} = \frac{1}{2}\, e^{-\lambda\, t}.$$

However, from the evolving set bound (3.10) it is also known that

$$d(t) \leq \frac{1}{2}\, e^{-2t\left(1-\sqrt{1-\tilde{\Phi}^2/4}\right)} \sqrt{\frac{1-\pi_*}{\pi_*}}\,.$$

Combining these two bounds shows that

$$\exp\left[-t\left(\lambda - 2\left(1 - \sqrt{1-\tilde{\Phi}^2/4}\right)\right)\right] \leq \sqrt{\frac{1-\pi_*}{\pi_*}}\,.$$

Taking $t \to \infty$ implies that the exponent on the left is negative, which gives the theorem in the reversible case.

If the chain is not reversible then it is easily verified that $\lambda = \lambda_{(\mathsf{P}+\mathsf{P}^*)/2}$ and $\tilde{\Phi} = \tilde{\Phi}_{(\mathsf{P}+\mathsf{P}^*)/2}$. The general bound then follows immediately from the reversible case. □

Not only does the conductance $\tilde{\Phi}$ lower bound λ, but the f-congestion does as well.

Theorem 5.8. Given $f : [0,1] \to \mathsf{R}_+$, non-zero except possible at 0 and 1, then

$$1 - \max_{i>0} |\lambda_i| \geq 1 - \mathcal{C}_f\,.$$

To see this, observe that if $S_0 = \{x\}$ then

$$\|\mathsf{P}^n(x,\cdot) - \pi\|_{\mathrm{TV}} \leq \hat{\mathbb{E}}_n(1-\pi(S_n)) \leq c\,\hat{\mathbb{E}}_n \frac{f(\pi(S_n))}{\pi(S_n)}\,,$$

where $c = \max_{\pi(A)\neq 0,1} \frac{\pi(A)(1-\pi(A))}{f(\pi(A))} < \infty$. By Lemma 3.8,

$$d(n) \leq c\,\mathcal{C}_f^n \max_{x\in\Omega} \frac{f(\pi(\{x\}))}{\pi(\{x\})}.$$

The theorem then follows as before.

Example 5.9. Consider the lazy walk on a cycle of even length which steps to the left or right with probability 1/2 each. This has

$$\Phi = 1/n, \quad \tilde{\Phi} = 2/n \quad \text{and} \quad \lambda = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi}{n}\right)\right) \approx \frac{\pi^2}{n^2}.$$

The new Cheeger bound is a factor two better than the conventional one:

$$\frac{2}{n} = \tilde{\Phi} \geq \lambda \geq \tilde{\Phi}^2/4 = \frac{1}{n^2}.$$

Now, the quantity $\mathcal{C}_{\sin(\pi z)}$ is maximized at the set A consisting of a connected set of half the vertices. The values A_u are easily determined, and

$$\lambda = 1 - \lambda_1 \geq 1 - \mathcal{C}_{\sin(\pi z)} = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi}{n}\right)\right),$$

the correct value.

Note that the periodic walk $\mathsf{P}(i, i \pm 1) = 1/2$ on the even cycle has $\mathcal{C}_{\sin(\pi z)}$ maximized at the set A given by one of the bipartitions, with $\mathcal{C}_{\sin(\pi z)} = 1$, implying $1 - |\lambda_{max}| = 1 - \mathcal{C}_{\sin(\pi z)} = 0$, which is correct as there is an eigenvalue $\lambda_{n-1} = -1$.

If P is non-reversible then one can consider $\mathcal{C}_f$ for the chain $\frac{\mathsf{P}+\mathsf{P}^*}{2}$ to obtain a lower bound on $\lambda = \lambda_{\frac{\mathsf{P}+\mathsf{P}^*}{2}}$. Also, by upper bounding $\mathcal{C}_f$ for appropriate choices of f this can be used to show Cheeger-type bounds in terms of conductance profile, edge and vertex expansion, and other related quantities. See [62] for details.

The same method of proof can also be used to show a lower bound on eigenvalues of non-reversible chains, and not just on λ. Recall that on a reversible Markov chain $1-\lambda_1 = \lambda$, from which it also follows that $1 - \lambda_{\mathsf{PP}^*} = \lambda_1(\mathsf{PP}^*) = |\lambda_{max}|^2$, that is $1 - |\lambda_{max}| = 1 - \sqrt{1 - \lambda_{\mathsf{PP}^*}}$. We now generalize these relations on $1 - \lambda_1$ and $1 - |\lambda_{max}|$ to the case of eigenvalues of non-reversible chains.

To get the two cases we require both discrete and continuous time mixing relations.

$$\begin{aligned} \frac{1}{2} e^{-(1-\max_{i\neq 0} \mathrm{Re}\lambda_i)\, t} \leq \quad & d(t) \quad \leq \frac{1}{2} e^{-\lambda t} \sqrt{\frac{1-\pi_*}{\pi_*}} \\ \frac{1}{2} |\lambda_i|^n \leq \quad & d(n) \quad \leq \frac{1}{2} (1 - \lambda_{\mathsf{PP}^*})^{n/2} \sqrt{\frac{1-\pi_*}{\pi_*}} \end{aligned}$$

Then, arguing as in the proof of Theorem 5.7 implies the following:

Theorem 5.10. The non-trivial (complex-valued) eigenvalues of a finite, irreducible Markov chain P will satisfy

$$\begin{aligned} 1-\max_{i>0} \mathrm{Re}\lambda_i &\geq \lambda \\ 1-\max_{i>0} |\lambda_i| &\geq 1-\sqrt{1-\lambda_{\mathsf{PP}^*}} \geq \lambda_{\mathsf{PP}^*}/2 \end{aligned}$$

It follows immediately from the Cheeger inequality $\lambda \geq \Phi^2/2$ that

$$1-\max_{i>0} \mathrm{Re}\lambda_i \geq \Phi^2/2\,.$$

This gives an alternate proof of a result of [17], which argued as in a standard proof of Cheeger's inequality but with a specialized Laplacian for non-reversible chains.

The theorem shows that the real part of the eigenvalues is related to the eigenvalue gap of the additive reversibilization $\frac{\mathsf{P}+\mathsf{P}^*}{2}$ (since $\lambda = \lambda_{\mathsf{P}} = \lambda_{\frac{\mathsf{P}+\mathsf{P}^*}{2}}$), while the magnitude of the eigenvalues is related to the eigenvalue gap of the multiplicative reversibilization PP^*. This is not just an artifact of the method of proof, as the following shows:

Example 5.11. Consider the walk on a cycle of length n that steps in the counterclockwise direction with probability one. The eigenvalues are $\lambda_k = e^{2\pi ki/n}$. Then $\lambda_{\mathsf{PP}^*} = 0$ since $\mathsf{PP}^* = I$, and $\lambda = 1-\cos(2\pi/n)$ since $\lambda = \lambda\left(\frac{\mathsf{P}+\mathsf{P}^*}{2}\right)$ is twice that of the lazy chain considered earlier. The theorem then shows that

$$\begin{aligned} 1-\max_{i>0} \mathrm{Re}\lambda_i &\geq \lambda = 1-\cos(2\pi/n)\,, \\ 1-\max_{i>0} |\lambda_i| &\geq \lambda_{\mathsf{PP}^*} = 0\,, \end{aligned}$$

both of which are in fact equalities.

Not only can spectral gap distinguish between the two cases, but conductance bounds can as well. For instance, on the 3-cycle it is easily verified that $\Phi = 1$ and $\Phi_{\mathsf{PP}^*} = 0$. Then Cheeger's inequality shows that

$$\begin{aligned} 1-\max_{i>0} \mathrm{Re}\lambda_i &\geq 1-\sqrt{1-\Phi^2} = 1\,, \\ 1-\max_{i>0} |\lambda_i| &\geq 1-\sqrt{1-\Phi_{\mathsf{PP}^*}^2} = 0\,. \end{aligned}$$

5.4 The Thorp Shuffle

A classical problem in mixing times is to determine the number of shuffles required to make a deck of cards well mixed. Many methods of card shuffling have been proposed, the most famous being the riffle shuffle, and for most of these fairly rigorous mixing time bounds are known. One card shuffle which has proved difficult to analyze is the Thorp shuffle. Resolving a twenty or so year old conjecture, B. Morris [64] has shown in 2005 that the mixing time of the Thorp shuffle on 2^d cards is indeed polynomial in d.

The Thorp shuffle can be described as follows. Given a deck with an even number of cards m, split the deck exactly in half, putting the top $m/2$ cards in the left hand and the bottom $m/2$ in the right hand. Now, flip a coin and drop the bottom card from the left hand if the outcome is a heads or from the right hand if it is tails, then drop the card from the other hand. Next, flip the coin again to decide the order in which to drop the second pair of cards, and continue flipping and dropping until all cards have been dropped. This makes a single shuffle.

There are three issues to consider in analyzing the Thorp shuffle by conventional spectral or conductance methods. First, the shuffle is non-reversible, and in fact once a shuffle is performed it is impossible to get back to the original ordering in under $d-1$ shuffles. Also, the holding probability is only $\alpha = 1/2^{2^{d-1}}$, so spectral gap or conductance methods do not work as they require α to not be too small. Third, spectral gap and conductance bounds have a $\log(1/\pi_*)$ term, and for the Thorp shuffle $\pi_* = 1/(2^d)! < 1/2^{d2^{d/2-1}}$, and so even taking the log of this it is still exponential.

All of these problems can be remedied by spectral profile $\Lambda_{\mathsf{PP}^*}(r)$ or evolving set $\mathcal{C}_{\sqrt{z(1-z)}}(r)$ methods. We borrow heavily from an evolving set argument of Morris [64], but our more careful analysis sharpens the mixing time bound of Morris from $\tau_2(1/e) = O(d^{44})$ to $\tau_2(1/e) = O(d^{29})$.

5.4.1 Modeling the Thorp Shuffle

Before delving into the mixing time bound we consider why a 2^d card deck is more amenable to analysis than other deck sizes. This is best seen be considering various alternate models of the Thorp shuffle, which we borrow from the discussion of Morris.

First, counting from the bottom of the deck, denote the initial ordering of the cards as 0, 1, 2, ..., $2^d - 1$. A single Thorp shuffle is equivalent to taking the cards at positions i and $2^{d-1} + i$ (for each $i \in 1 \ldots 2^{d-1}$), and placing them in positions $2i$ and $2i + 1$, with order determined by the coin flips.

Equivalently, consider the bit sequence corresponding to the card label, for instance $1 = 001_2$ in an 8 card deck, and flip a coin to decide whether to change order of the cards differing only in the most significant bit, then shift the bits to the left, wrapping around the final bit (e.g. $001_2 \to 101_2 \to 011_2$ if cards 1 and 5 are interchanged). Note that d consecutive shuffles will put the bits back in their original positions.

For a final model, take a d-dimensional cube $\{0, 1\}^d$, label a vertex by the binary value of its coordinates, so that vertex $(0, 1, 1, 0) = 0110_2 = 6$, and place card i at vertex i. Then, to do d consecutive Thorp shuffles, first consider all edges in direction 0 (i.e. corresponding to the highest order bit) and decide whether to interchange the cards at opposite ends of each edge independently, then do the same in direction 1, 2, etc. until all d directions have been considered.

This last model has several nice features. For one, even though the Thorp shuffle is non-reversible, flipping in a fixed direction on the cube is reversible. Also, this makes an inductive argument possible, by considering a pseudo k-step shuffle in which only the first k coordinate directions will be flipped. These properties suggest that it will be beneficial to base a mixing time proof on the d-step Thorp shuffle, rather than on single shuffles.

5.4.2 Spectral Profile

We begin by considering the spectral profile argument, because there is a large audience familiar with spectral gap bounds on mixing, but far fewer familiar with evolving sets.

By Theorem 2.10 the holding probability is not an issue if bounds on $\Lambda_{\mathsf{PP}^*}(r)$ are known. A standard proof of Cheeger's inequality, restricted to functions supported on a set of size at most r, can be used (see [34]) to show that

$$\Lambda_{\mathsf{PP}^*}(r) \geq 1 - \sqrt{1 - \Phi^2_{\mathsf{PP}^*}(r)} \geq \Phi^2_{\mathsf{PP}^*}(r)/2. \tag{5.2}$$

This shows that it is natural to consider the conductance profile of the chain PP^*.

Theorem 5.12. Given a 2^d card deck, the Thorp shuffle satisfies

$$\Phi_{\mathsf{KK}^*}(A) \geq 1 - \pi(A)^{C/d^{14}} \quad \text{and} \quad \tilde{\Phi}_{\mathsf{KK}^*} \geq \frac{C}{d^{14}}\,,$$

where K indicates the transition kernel of the d-step shuffle, and C is a constant independent of d.

Proof. Given set $A \subset \Omega$ then

$$\mathsf{Q}_{\mathsf{KK}^*}(A,A) = \sum_{z\in\Omega} \mathsf{Q}(A,z)\mathsf{K}^*(z,A) = \sum_{z\in\Omega} \mathsf{K}^*(z,A)^2\,\pi(z) = \|\mathsf{K}^*\,1_A\|_2^2 \tag{5.3}$$

where the second equality is because $\mathsf{Q}_{\mathsf{K}}(A,z) = \mathsf{Q}_{\mathsf{K}^*}(z,A) = \pi(z)\mathsf{K}^*(z,A)$. Lemma 12 of Morris [64] states that if $f:\ \Omega \to [0,1]$ then

$$\|\mathsf{K}^*\,f\|_2^2 \leq \|f\|_1^{1+C/d^{14}}$$

for some constant $C \in (0,1)$. *(See Section 5.4.4 below for a discussion of Morris' proof of this result.)* Letting $f = 1_A$ then leads to the bound

$$\begin{aligned} \Phi_{\mathsf{KK}^*}(A) &= \frac{\pi(A) - \mathsf{Q}_{\mathsf{KK}^*}(A,A)}{\pi(A)} \\ &\geq \frac{\pi(A) - \pi(A)^{1+C/d^{14}}}{\pi(A)} = 1 - \pi(A)^{C/d^{14}}\,. \end{aligned}$$

The bound on $\tilde{\Phi}_{\mathsf{KK}^*}$ follows from this:

$$\begin{aligned} \tilde{\Phi}_{\mathsf{KK}^*} &= \min_{\pi(A)\leq 1/2} \frac{\Phi_{\mathsf{KK}^*}(A)}{\pi(A^c)} \\ &\geq \min_{x\in(0,1/2]} \frac{1 - x^{C/d^{14}}}{1-x} = 2\left(1 - 2^{-C/d^{14}}\right) \end{aligned}$$

followed by the approximation $2^{-\gamma} \leq 1 - \gamma/2$ when $\gamma \in [0,1]$. □

This shows that when r is extremely small then $\Lambda_{\mathsf{KK}^*}(r) \geq 1 - \sqrt{2}r^{1/2d^{14}} \approx 1$. However, even if $\Lambda_{\mathsf{KK}^*}(r) = 1$ then Theorem 2.10 shows at best $\tau_2(\epsilon) = O\left(\log \frac{1}{\pi_* \epsilon^2}\right)$, which is still exponential. Thus, our approach to this problem will be to show that variance drops exponentially fast from $\mathrm{Var}(k_0^x) = (2^d)! - 1 \geq 2^{d2^{d/2-1}}$ until it reaches a singly exponential size, after which Theorem 2.10 can be used to finish the proof.

Theorem 5.13. The mixing time of the Thorp shuffle is

$$\tau_2(\epsilon) \leq 8\, C^{-2}\, d^{29}\, (25 + \log(1/\epsilon))\,,$$

where C is the constant of Theorem 5.12.

Proof. To reduce clutter in the proof somewhat we define $D = d^{14}/C$.

Let $c = \delta \frac{\mathrm{Var}(f)}{2\mathbb{E}f}$ in the final step of the proof of Lemma 2.9. Then

$$\mathcal{E}_{\mathsf{KK}^*}(k_n, k_n) \geq \mathrm{Var}(k_n)(1-\delta)\Lambda_{\mathsf{KK}^*}(2/\delta \mathrm{Var}(k_n))\,. \tag{5.4}$$

Theorem 5.12 and the strengthened Cheeger inequality (5.2) imply that

$$\Lambda_{\mathsf{KK}^*}(r) \geq 1 - \sqrt{1 - \Phi^2_{\mathsf{KK}^*}(r)} \geq 1 - \sqrt{2}\, r^{1/2D}\,. \tag{5.5}$$

Letting $\delta = \frac{1}{2\mathrm{Var}(k_n)^{1/4D}}$ it follows from (5.4) and (5.5) that

$$\begin{aligned}
&\mathcal{E}_{\mathsf{KK}^*}(k_n, k_n) \\
&\geq \mathrm{Var}(k_n)\left(1 - \frac{1}{2\mathrm{Var}(k_n)^{1/4D}}\right)\left(1 - \sqrt{2}\left(\frac{4}{\mathrm{Var}(k_n)^{1-1/4D}}\right)^{1/2D}\right) \\
&\geq \mathrm{Var}(k_n)\left(1 - \frac{1}{\mathrm{Var}(k_n)^{1/4D}}\left(\frac{1}{2} + \left(\frac{2^{2+D}}{\mathrm{Var}(k_n)^{1/2-1/4D}}\right)^{1/2D}\right)\right)
\end{aligned}$$

If $\mathrm{Var}(k_n) \geq 2^{(8+12D)}$ then this simplifies to $\mathcal{E}_{\mathsf{KK}^*}(k_n, k_n) \geq \mathrm{Var}(k_n) - \mathrm{Var}(k_n)^{1-1/4D}$. Then by Lemma 1.13

$$\mathrm{Var}(k_{n+1}) = \mathrm{Var}(k_n) - \mathcal{E}_{\mathsf{KK}^*}(k_n, k_n) \leq \mathrm{Var}(k_n)^{1-1/4D}\,.$$

It follows by induction on n that if $N = 4Dd$ and $\mathrm{Var}(k_N) \geq 2^{(8+12D)}$ then

$$\mathrm{Var}(k_N) \leq \mathrm{Var}(k_0)^{\left((1-1/4D)^N\right)} \leq 2^{d\,2^d\,e^{-d}} \leq 2^d\,.$$

This gives a contradiction, and it follows that $\mathrm{Var}(k_N) < 2^{(8+12D)}$.

Now spectral profile will be useful. For large values of r, apply the inequality $\forall x, \alpha \in [0,1]: \ (1-x)^\alpha \leq 1 - \alpha\, x$ to Theorem 5.12 to obtain

$$\begin{aligned} \Phi_{\mathsf{KK}^*}(r) &\geq 1 - r^{1/D} = 1 - \left(\frac{1}{2}\right)^{\log_2(1/r)/D} \\ &\geq \begin{cases} 1/2 & \mathit{if}\ r \leq 1/2^D \\ \frac{\log_2(1/r)}{2D} & \mathit{if}\ r > 1/2^D\,. \end{cases} \end{aligned} \tag{5.6}$$

Recall that by Cheeger's Inequality $\Lambda_{\mathsf{KK}^*}(r) \geq \frac{1}{2}\,\Phi^2_{\mathsf{KK}^*}(r)$, and by Theorem 5.7, $\lambda_{\mathsf{KK}^*} \geq \frac{1}{4}\,\tilde{\Phi}^2_{\mathsf{KK}^*}$.

Let $\sigma/\pi = k_N$ in Theorem 2.10, to obtain

$$\begin{aligned} \tau_2(\epsilon) &\leq N + \left\lceil \int_{4/\mathrm{Var}(k_N)}^{1/2} \frac{2\,dr}{r\,\Lambda_{\mathsf{KK}^*}(r)} + \frac{2}{\lambda_{\mathsf{KK}^*}} \log\frac{2\sqrt{2}}{\epsilon} \right\rceil \\ &\leq N + \left\lceil \int_{4/\mathrm{Var}(k_N)}^{1/2} \frac{4\,dr}{r\,\Phi^2_{\mathsf{KK}^*}(r)} + \frac{8}{\tilde{\Phi}^2_{\mathsf{KK}^*}} \log\frac{2\sqrt{2}}{\epsilon} \right\rceil \\ &\leq N + \left\lceil \int_{4/2^{(8+12D)}}^{1/2^D} \frac{16\,dr}{r} + \int_{1/2^D}^{1/2} \frac{16D^2\,dr}{r\,(\log_2(1/r))^2} + 8D^2\,\log\frac{2\sqrt{2}}{\epsilon} \right\rceil \\ &= N + \left\lceil 4(\log 2)\,\left(7D^2 + 40D + 24\right) + 8D^2\log(1/\epsilon)\right)\right\rceil\,. \end{aligned}$$

Substitute back in $D = d^{14}/C$, then recall that each of these shuffles was in fact d steps of a "regular" Thorp shuffle, resulting in a mixing time bound d times larger, and hence the $O(d^{29})$ bound for the Thorp shuffle. □

5.4.3 Evolving Sets

The approach to this section will be roughly the same as that in the spectral case, but with $\hat{\mathbb{E}}_n\sqrt{\frac{1-\pi(S_n)}{\pi(S_n)}}$ and $1 - \mathcal{C}_{\sqrt{z(1-z)}}(r)$ in place of $\mathrm{Var}(k_n)$ and $\Lambda_{\mathsf{KK}^*}(r)$ respectively. The proof of a bound for the Thorp shuffle will be similar to that of Theorem 5.13. First, a careful analysis will be used to show that $\hat{\mathbb{E}}_n\sqrt{\frac{1-\pi(S_n)}{\pi(S_n)}}$ drops from doubly exponential to singly exponential after some N steps. Then the congestion profile bound on mixing time will be used for the remainder.

Theorem 5.14. The mixing time of the Thorp shuffle is

$$\tau_2(\epsilon) \leq 8\, C^{-2}\, d^{29}\, \left(2 + \log(1/\epsilon)\right) ,$$

where C is the constant of Theorem 5.12.

Proof. As before, we reduce clutter by defining $D = d^{14}/C$.

To show exponential contraction of distance an argument similar to the spectral case can be given, with Lemma 3.11 improved to $E(Zg(Z)) \geq (1-\delta)EZ\, g(\delta EZ)$, just as Lemma 2.9 was sharpened previously. However, a somewhat sharper bound can be obtained by using convexity. A few preliminaries are required.

Suppose $h(x) \geq x\mathcal{C}_{\sqrt{z(1-z)}}\left(\frac{1}{1+x^2}\right)$ and $f(x) = \sqrt{\frac{1-x}{x}}$. By Equation (3.4), if $x - h(x)$ is convex then $\hat{\mathbb{E}}_{n+1} f(\pi(S_{n+1})) \leq h(\hat{\mathbb{E}}_n f(\pi(S_n)))$. If h is increasing then by induction $\hat{\mathbb{E}}_n f(\pi(S_n)) \leq h^n(f(\pi(S_0)))$, and if, moreover, $g \geq h$ for some function g then

$$\hat{\mathbb{E}}_n f(\pi(S_n)) \leq h^n(f(\pi(S_0))) \leq g^n(f(\pi(S_0)))\,. \tag{5.7}$$

It remains to verify these conditions and determine $g^n(f(\pi(S_0)))$. By our earlier work

$$\mathcal{C}_{\sqrt{z(1-z)}}(r) \leq \sqrt[4]{1 - \Phi^2_{\mathsf{KK}^*}(r)} \leq \sqrt[4]{1 - (1 - r^{1/D})^2}\,.$$

Let $h(x) = x\sqrt[4]{1 - \left(1 - \frac{1}{(1+x^2)^{1/D}}\right)^2} \geq x\mathcal{C}_{\sqrt{z(1-z)}}\left(\frac{1}{1+x^2}\right)$ and $g(x) = \sqrt[4]{2}\, x^{1-1/2D}$. Then $x - h(x)$ is convex, h is increasing, and $g \geq h$, so Equation (5.7) applies. It follows that if $\hat{\mathbb{E}}_N f(\pi(S_N)) > 2^D$ at $N = 4Dd$ then for all $\forall k \leq N :\ g^k(f(\pi(S_0))) > \hat{\mathbb{E}}_N f(\pi(S_N)) > 2^D$. Since $f(\pi(S_0)) < 1/\sqrt{\pi(S_0)}$ and $g(x) \leq x^{1-1/4D}$ when $x \geq 2^D$ then

$$\begin{aligned} \hat{\mathbb{E}}_N f(\pi(S_N)) &\leq g^N(f(\pi(S_0))) < \left(\frac{1}{\sqrt{\pi(S_0)}}\right)^{(1-1/4D)^{4Dd}} \\ &\leq 2^{d2^d\, e^{-d}} < 2^d \leq 2^D\,, \end{aligned}$$

a contradiction. It follows that $\hat{\mathbb{E}}_N f(\pi(S_N)) \leq 2^D$ at $N = 4dD$.

The proof of Theorem 3.10 still holds when π_* is replaced by $f^{-1}(\hat{\mathbb{E}}_N f(\pi(S_N))) = \frac{1}{1+\left(\hat{\mathbb{E}}_N f(\pi(S_N))\right)^2}$. Hence,

$$\begin{aligned}
\tau_2(\epsilon) &\leq N + \left\lceil \int_{f^{-1}(\hat{\mathbb{E}}_N f(\pi(S_N)))}^{f^{-1}(\epsilon)} \frac{dr}{2r(1-r)(1-\mathcal{C}_{\sqrt{z(1-z)}}(r))} \right\rceil \\
&\leq N + \left\lceil \int_{\frac{1}{1+\left(\hat{\mathbb{E}} f(\pi(S_N))\right)^2}}^{1/2} \frac{2\,dr}{r(1-r)\Phi^2_{\mathsf{KK}^*}(r)} + \int_{1/2}^{\frac{1}{1+\epsilon^2}} \frac{4\,dr}{r(1-r)\tilde{\Phi}^2_{\mathsf{KK}^*}} \right\rceil \\
&\leq N \\
&\quad + \left\lceil \int_{\frac{1}{1+2^{2D}}}^{\frac{1}{2^D}} \frac{8\,dr}{r(1-r)} + \int_{\frac{1}{2^D}}^{1/2} \frac{8D^2\,dr}{r(1-r)(\log_2(1/r))^2} + \int_{1/2}^{\frac{1}{1+\epsilon^2}} \frac{4D^2\,dr}{r(1-r)} \right\rceil \\
&\leq N + \left\lceil 16(\log 2)D^2 + 4D^2 \log(1/\epsilon^2) \right\rceil ,
\end{aligned}$$

where the second inequality applied Equation (3.9), the third inequality applied the conductance bounds of Equation (5.6), and in the fourth inequality the second integral was evaluated by first approximating $1 - r \geq 1/2$.

Substitute back in $D = d^{14}/C$, then recall that each of these shuffles was in fact d steps of a “regular” Thorp shuffle, resulting in a mixing time bound d times larger, and hence the $O(d^{29})$ bound for the Thorp shuffle. □

5.4.4 Bounding the ℓ_2 norm of functions

The mixing time argument given above relied crucially on Lemma 12 of Morris [64]. The proof of this result and the preliminary tools leading to it are quite involved, and so it would be beyond the scope of this survey to prove the lemma here. Instead, we merely sketch the key principles behind this proof.

Recall the model of the d-step Thorp shuffle given by flipping along the d-coordinate directions of a d-dimensional cube. If only the first k coordinate directions are flipped this gives a pseudo k-step shuffle, and opens the possibility of an inductive proof on $k = 1 \ldots d$. Note that this is not the same as k Thorp shuffles on the original 2^d-card deck, but is rather k Thorp shuffles on 2^{d-k} independent 2^k-card subdecks.

We now give a very rough sketch of Morris' proof; all references are to version [64]. Recall that the goal is to show, for every set A of card orderings, that $\|\mathsf{K}^*1_A\|_2^2$ is extremely small, where K denotes d consecutive Thorp shuffles. As we saw, $\|\mathsf{K}^*1_A\|_2^2 = \mathsf{Q}_{\mathsf{KK}^*}(A, A)$ and so intuitively this will be similar to showing that, given a set A of card orderings, there is a very small chance of a KK^* shuffle taking an ordering in A to another ordering in A. When we refer to the "k-coordinate shuffle" we will mean a shuffle in which the first k-coordinates are shuffled, followed by a time-reversed k-coordinate shuffle (i.e. KK^*, called the "zigzag shuffle" by Morris).

- Assume the $k-1$ coordinate shuffle has been studied already. Adding a shuffle in the kth coordinate will merge pairs of 2^{k-1}-card groups into common 2^k-card groups. In each pair refer to the two 2^{k-1}-card groups as the "top" and "bottom". In the best case scenario the induction will require only that the $k-1$ coordinate shuffle sufficiently mixed the "top" and "bottom". When this doesn't work then it is necessary to understand the degree to which the kth step of the shuffle randomizes the choice of cards in the lower half space (i.e. how much it mixes cards from the "top" into locations in the "bottom").
- For this use Corollary 5, which shows that after $O(d(k-1)^4)$ k-coordinate shuffles the 2^{k-1} cards in the "bottom" of each of the 2^{d-k} sub-decks are roughly uniformly chosen from the 2^k cards available.
- To show this, let Λ' denote a set of positions in a deck. By Lemma 4, after $O(d^4)$ full d-coordinate shuffles any of the $C(2^d, |\Lambda'|)$ subsets of $|\Lambda'|$ cards are about equally likely to end at Λ'. In particular, a k-coordinate shuffle is a full shuffle for each 2^k-card sub-deck, so $O(k^4)$ such shuffles randomize the selection of cards ending in the "bottom".
- This requires equation (26), that if $O(d^4)$ full d-step shuffles sends unordered set S of cards to some unordered set S' of positions (with $|S| = |S'|$), then each $x \in S$ is about equally likely to end at any fixed $y \in S'$.

- For this use equation (17), which uses the "Chameleon process" to show that if $O(d^3)$ full d-step shuffles have distribute the first b cards (for any b) fairly widely in the deck of 2^d cards, then card $b+1$ is also well mixed up in the deck. Induction on b shows there to be a good chance the initial subsets are at random positions.

Heuristically speaking, the proof is building up from the most local property to the most global – starting by showing that if the first b cards have moved around the deck a lot then the next card probably has too, then showing that subsets of cards do in fact move around the space fairly well, and eventually showing a property of averages over the whole space with Lemma 12.

6

Miscellaneous

6.1 The Fastest Mixing Markov Process Problem

Throughout this section, we will assume that the Markov kernel P under consideration is reversible. Consider the general problem of designing a fast Markov chain with a prescribed stationary measure π on the finite state space M. Trivially one can always consider the kernel $P(x, y) = \pi(y)$, for each $x, y \in \Omega$ which reaches stationarity in one step! In situations where Ω is of large size, such an answer is not satisfactory, since the above mentioned P may not be practical (implementable). To make the above problem more interesting and nontrivial, following the work of [10, 11, 77, 66], let us now further impose that we are given an arbitrary graph $G = (V, E)$, with π a probability measure on V as before, and demand that any kernel P has to be zero on the pairs which are not adjacent in G (i.e., not connected by an edge from E) and such that π is the stationary measure for P. Now given this setting, the main issue is to come up with the fastest such P, in the sense that such a P has as large a *spectral gap* as possible. Let $\lambda^*(G, \pi)$ denote the spectral gap of such a fastest chain. When π is uniform over the state space, we denote this by $\lambda^*(G)$. Boyd et al.

also investigated [77] the closely related problem of designing a fastest Markov process (meaning, a continuous-time Markov chain) under the above stipulation. Since the rates of a continuous-time Markov process can be arbitrary (up to scaling), a particular normalization such as the *sum of all non-diagonal entries* being at most 1, was imposed by [77] to make the problem meaningful in continuous-time. Perhaps it would have been more common to have made the assumption that the sum in each row (barring the non-diagonal entry) be at most 1, however the above assumption seems rather interesting and powerful (see the remark at the end of this section).

One of the main conclusions of [10, 11, 77] is that the problem of determining the spectral gap of the fastest chain or process can be modeled as a convex optimization problem, and that it can be solved in polynomial time in the size ($|V(G)|$) of the chain. These papers also provide examples showing that some of the natural chains, such as the Metropolis walk, or the maximum degree based simple random walk, can be worse than the optimal gap ($\lambda^*(G)$) by an arbitrary multiplicative factor – in some instances off by a factor as high as $|V(G)|$.

For the problem of designing the fastest (continuous-time) Markov process, tight bounds on the optimal gap have been given in the recent work of Naor et al. [66]. In particular, using a dual characterization in [77], they provide tight bounds on the spectral gap $\lambda^*(G)$ in terms of a maximal variance quantity, introduced in [2], as the *spread constant* of G:

Let $\mathcal{L}(G)$ denote the set of Lipschitz functions (with respect to the graph distance) on G with constant 1. Equivalently,

$$\mathcal{L}(G) = \{f : V(G) \to \mathsf{R} : \{i, j\} \in E(G) \to |f(i) - f(j)| \le 1\}.$$

The spread constant of G with respect to π is defined to be

$$c_{var}(G, \pi) = \max_{f \in \mathcal{L}(G)} \mathrm{Var}_\pi f. \tag{6.1}$$

It is easily seen that the spread constant offers an upper bound on the inverse spectral gap of any P with support on the edges of G. Indeed

$$\begin{aligned}\lambda(P) &= \inf_{f:\mathrm{Var}_\pi f\neq 0} \frac{\mathcal{E}(f,f)}{\mathrm{Var}_\pi f} \leq \inf_{\substack{f:\mathrm{Var}_\pi f\neq 0\\ f:\in\mathcal{L}(G)}} \frac{\mathcal{E}(f,f)}{\mathrm{Var}_\pi f} \\ &\leq \inf_{\substack{f:\mathrm{Var}_\pi f\neq 0\\ f:\in\mathcal{L}(G)}} \frac{1}{2\mathrm{Var}_\pi f} = \frac{1}{2c_{var}(G,\pi)}. \end{aligned} \tag{6.2}$$

On the other hand, the work of [66] shows that for the *fastest* continuous-time P, the above inequality can be off by no more than a factor of $\log n$, where n is the number of vertices of G (i.e., the number of states of P). The precise formulation is as follows. Let $c_{var}(G)$ denote $c_{var}(G,\pi)$ when π is uniform over the vertex set.

Theorem 6.1. Given an undirected graph $G=(V,E)$ with $|V|=n$, let P be a chain with the largest spectral gap $\lambda^*(G)$ over all reversible continuous-time random walks with support contained in E, and with uniform stationary distribution. Then

$$2c_{var}(G) \leq \frac{1}{\lambda^*(G)} = O(c_{var}(G)\log n). \tag{6.3}$$

Moreover, there exist arbitrarily large n-vertex graphs G for which

$$\frac{1}{\lambda^*(G)c_{var}(G)} = \Omega\left(\frac{\log n}{\log\log n}\right). \tag{6.4}$$

We recall here the proof of (6.3), which follows fairly easily using some known results; we refer the reader to [66] for the proof of (6.4), for results on the (complexity of) approximation of spread constant, and further related material.

Proof. [Proof of Theorem 6.1] The lower bound is immediate from the observation made above, leading to (6.2). For the upper bound, let $\pi(i)=1/n$ for all $i\in V$, and given $X\in\mathsf{R}^n$ let $\|X\| = \left(\sum_{k=1}^n X(k)^2\right)^{1/2}$ denote the Euclidean length of vector X.

Sun et al. [77] show that finding λ^* is equivalent (by duality) to solving the following n-dimensional maximal variance problem (see eqn. (12) in Section 6 of [77]). In fact λ^* is equal to the inverse of the optimum value in the following problem.

$$\begin{array}{ll} \text{Maximize} & \sum_i \pi(i)\|X_i\|^2 \\ \text{subject to} & \|X_i - X_j\| \le 1, \quad \{i,j\} \in E, \text{ and } \sum_i \pi(i)X_i = \vec{0}. \end{array}$$

The above maximization is over $X_i \in \mathsf{R}^n$, for $i = 1, 2, \ldots, n$. Since $\sum_i \pi(i)X_i = 0$, we may rewrite the above problem:

$$\begin{array}{ll} \text{Maximize} & \frac{1}{2}\sum_{i,j} \|X_i - X_j\|^2\pi(i)\pi(j) \\ \text{subject to} & \|X_i - X_j\| \le 1, \quad \{i,j\} \in E, \text{ and } \sum_i \pi(i)X_i = \vec{0}. \end{array}$$

Note that when the above problem is formulated with $X_i \in \mathsf{R}$, it reduces to solving for the spread constant. On the other hand, the now-famous lemma of Johnson and Lindenstrauss [44] guarantees that there exists a map $f : \mathsf{R}^n \to \mathsf{R}^d$ with $d = O(\log n)$, and

$$\frac{1}{2}\|X_i - X_j\| \le \|f(X_i) - f(X_j)\| \le \|X_i - X_j\| \text{ for all } i, j \in V. \quad (6.5)$$

More over, such an f can be found in polynomial time in n using a randomized algorithm. Thus such a map, while maintaining the constraint that $\|f(X_i) - f(X_j)\| \le 1$, for $\{i,j\} \in E$, gives solution vectors in R^d, with no worse than a factor of 4 loss in the optimal variance, as

$$\frac{1}{4}\sum_{i,j} \|X_i - X_j\|^2\pi(i)\pi(j) \le \sum_{i,j} \|f(X_i) - f(X_j)\|^2\pi(i)\pi(j).$$

The other constraint about the mean of the configuration being zero should not matter, since $\sum_i \pi(i)X_i$ (or $\sum_i \pi(i)f(X_i)$) can be subtracted from any solution without affecting $\|X_i - X_j\|$. Finally, to complete the first part of the theorem, observe that

$$\begin{aligned} \frac{1}{4d}\frac{1}{\lambda^*} &= \frac{1}{8d}\sum_{i,j} \|X_i - X_j\|^2\pi(i)\pi(j) \\ &\le \frac{1}{2d}\sum_{i,j}\sum_{k=1}^{d} |f(X_i)(k) - f(X_j)(k)|^2\pi(i)\pi(j) \\ &\le \max_k \frac{1}{2}\sum_{i,j} |f(X_i)(k) - f(X_j)(k)|^2\pi(i)\pi(j) \le c_{var}(G), \end{aligned}$$

the last inequality following from the fact that f automatically satisfies the Lipschitz condition in each coordinate as $1 \geq \|f(X_i) - f(X_j)\| \geq |f(X_i)(k) - f(X_j)(k)|$ for all k. □

As noted in [66], the theorem generalizes to the case of general π and nonnegative real weights $d_{ij} \geq 0$ on edges $\{i, j\}$. (The crucial Johnson-Lindenstrauss lemma used above is completely general, for any n-point metric space.) For $S \subset V(G)$, and $i \in V(G)$, let $\bar{d}(i, S) = \min_{j \in S} d_G(i, j)$, where $d_G(\cdot)$ denotes the graph distance. Now let S be any set with $\pi(S) \geq 1/2$. Then Theorem 3.15 of [2] shows that the spread constant is at least the square of the (average) distance of a random point to such a set S. For every $S \subset V(G)$, with $\pi(S) \geq 1/2$, we have

$$(\mathbb{E}_\pi \bar{d}(S)) := \sum_{i \in V(G)} \bar{d}(i, S)\pi(i) \leq \sqrt{c_{var}(G)}\,.$$

On the other hand, an upper bound on the spread constant turns out to be $D^2(G)/4$ (see [2], [66]). Hence the following corollary.

Corollary 6.2. Let G and P be as in Theorem 6.1, and let $D(G)$ be the diameter of G. Then for $\bar{d}(S)$ defined as above, with $S : \pi(S) \geq 1/2$,

$$(\mathbb{E}_\pi \bar{d}(S))^2 \leq \frac{1}{2\lambda^*(G)} = O(D^2(G) \log n). \tag{6.6}$$

Remark 6.3. The above corollary suggests a way of showing the *existence* of rapidly mixing Markov processes on a given graph structure, without necessarily explicitly constructing them! Given a large set (of size exponential in n, say), once we construct a graph G with the given set as its vertex set, and *any* edge set such that the graph has a small diameter (of size polynomial in n, say), then $\lambda^*(G)$ of such a graph is automatically bounded from below by a polynomial in n; implying that there *exists* a fast Markov process which samples (uniformly, say) from the given large set. Of course, it might be challenging and in general impractical a task to actually find such a process explicitly! In light of several *slow mixing* results for the standard Glauber-type dynamics for various statistical physics models, the above result nevertheless seems puzzling. As noted earlier in the section, the key to the mystery might be that we were implicitly providing the continuous-time chain with more power – by not requiring the rates in each row to sum to 1, but only the rates in the entire matrix to sum to 1, barring the diagonal entries in each sum.

6.2 Alternative description of the spectral gap, and the entropy constant

While the following characterization of the spectral gap apparently dates back to Bochner [6] and Lichnérowicz [50], it does not seem to be well-known in the Markov chain mixing time community. For recent applications and further references, see [9].

Proposition 6.4.

$$\lambda = \inf_f \frac{\mathcal{E}(-\mathcal{L}f,f)}{\mathcal{E}(f,f)} \,. \tag{6.7}$$

Proof. Let

$$\mu := \inf_f \frac{\mathcal{E}(-\mathcal{L}f,f)}{\mathcal{E}(f,f)} \,. \tag{6.8}$$

The fact that $\lambda \le \mu$ follows by simply using Cauchy-Schwartz. Indeed, without loss assuming $\mathbb{E}_\pi f = 0$, we have

$$\begin{aligned} \mathcal{E}(f,f) = \mathbb{E}(f(-\mathcal{L}f))) &\le (\mathrm{Var}(f))^{1/2}(\mathbb{E}(-\mathcal{L}f)^2))^{1/2} \\ &\le \frac{1}{\sqrt{\lambda}}(\mathcal{E}(f,f))^{1/2}(\mathbb{E}(-\mathcal{L}f)^2))^{1/2} \\ &= \frac{1}{\sqrt{\lambda}}(\mathcal{E}(f,f))^{1/2}(\mathcal{E}(-\mathcal{L}f,f))^{1/2}, \end{aligned}$$

implying that $\lambda \le \mu$. To prove the other direction, observe that, as in the argument of Lemma 1.4,

$$\frac{d}{dt}\mathrm{Var}(H_t f) = -2\mathcal{E}(H_t f, H_t f),$$

$$\frac{d}{dt}\mathcal{E}(H_t f, H_t f) = -2\mathcal{E}(H_t f, -\mathcal{L}H_t f).$$

Also note that as $t \to \infty$, $\mathcal{E}(H_t f, H_t f) \to 0$, since $H_t f \to \mathbb{E}_\pi f$. Thus,

$$\mathcal{E}(f,f) \;=\; -\int_0^\infty \frac{d}{dt}\mathcal{E}(H_t f, H_t f)\;dt = 2\int_0^\infty \mathcal{E}(H_t f, -\mathcal{L}H_t f)\;dt$$

$$
\begin{aligned}
&= 2\int_0^\infty \mathcal{E}(-\mathcal{L}^* H_t f, H_t f)\ dt \geq 2\mu_{\mathsf{P}^*}\int_0^\infty \mathcal{E}(H_t f, H_t f)\ dt \\
&= -\mu_{\mathsf{P}^*}\int_0^\infty \frac{d}{dt}\mathrm{Var}(H_t f)\ dt = \mu_{\mathsf{P}^*}\ \mathrm{Var} f\,,
\end{aligned}
$$

implying that $\mu_{\mathsf{P}^*} \leq \lambda_{\mathsf{P}}$.

Combining these two cases, and the relation $\lambda_{\mathsf{P}} = \lambda_{\mathsf{P}^*}$, gives $\lambda_{\mathsf{P}} \leq \mu_{\mathsf{P}} \leq \lambda_{\mathsf{P}^*} = \lambda_{\mathsf{P}}$. □

While it is natural to ask for the entropy analog, it seems to extend only in one direction. (see Bakry-Emery [3]). We omit the analogous proof, which involves the second derivative of entropy:

$$
\rho_0 \geq \mathrm{e}_0 := \inf_{f>0}\frac{u(f)}{\mathcal{E}(f,\log f)}\,, \quad u(f) := \mathcal{E}(-\mathcal{L}f,\log f) + \mathcal{E}(f,(-\mathcal{L}f)/f)\,. \tag{6.9}
$$

In [14], the above formulation was used in estimating the entropy constant of the Bernoulli-Laplace model on $r \geq 1$ particles. Their proof showing $\mathrm{e}_0 \geq n$ (independently of r) for the B-L model, makes effective use of the commutators of the gradient operators, and yields the best known estimate – in terms of the absolute constant in front of n. It remains elusive, as to how to employ their technique to estimate the entropy constant of the closely-related random transposition model. However, such a computation (in fact achieving the optimal constant) for the *spectral gap* of the random transposition model has been carried out, inter alia, in [9].

6.3 Perturbation Bounds

One of the important results in the stability theory for Markov chains was the discovery of a connection between the stability of a chain and its speed of convergence to equilibrium (see [57] and references therein). Such a connection is quantitative, in the sense that we can always get a sensitivity bound for Markov chains, once we have a convergence bound (say, in total variation), and the sharpness of the convergence will in turn determine the accuracy of the sensitivity bound. As an illustration, we cite the following theorem of A.Yu. Mitrophanov.

Consider two continuous-time Markov chains, $X(t)$ and $\tilde{X}(t)$, $t \geq 0$, with finite state space $\Omega = \{1, 2, \ldots, N\}$, $N \geq 1$, and generators $\mathbf{Q} = (q_{ij})$ and $\tilde{\mathbf{Q}} = (\tilde{q}_{ij})$, respectively. Recall that a row stochastic matrix P may be used to get a continuous-time Markov chain, by letting the generator be $\mathbf{Q} = \mathsf{P} - \mathsf{I}$, where I is the identity matrix. Then $p_0^T H_t = p_0^T e^{\mathbf{Q}t}$ gives the distribution of the chain at time $t \geq 0$, when started in p_0 at time 0. (Here we are viewing p_0 as a column vector.)

In the following, for a vector p, let $\|p\|$ denote the l_1 norm, and for a matrix A, let $\|A\|$ denote the subordinate norm (namely, the maximum absolute row sum of A). Assume that $X(t)$ has a unique stationary distribution $\pi()$. Then the following theorem is proved in [57]. Let p_0 and $\tilde{p}_0$ denote the distribution of $X(0)$ and $\tilde{X}(0)$, respectively.

Theorem 6.5. If $b > 0, c > 2$ are constants such that for all x, y,

$$\|H_t(x, \cdot) - H_t(y, \cdot)\| \leq ce^{-bt}, \quad t \geq 0, \tag{6.10}$$

then for $\mathbf{z}(t) = p_0^T H_t - \tilde{p}_0^T \tilde{H}_t$ and $\mathbf{E} = \mathbf{Q} - \tilde{\mathbf{Q}}$,

$$\|\mathbf{z}(t)\| < \begin{cases} \|\mathbf{z}(0)\| + t\|\mathbf{E}\| & \text{if } 0 < t \leq \frac{\log(c/2)}{b}, \\ ce^{-bt}\|\mathbf{z}(0)\|/2 & \\ +\frac{1}{b}(\log(c/2) + 1 - ce^{-bt}/2)\|\mathbf{E}\| & \text{if } t \geq \frac{\log(c/2)}{b}. \end{cases}$$

Moreover, if $\tilde{\pi}$ is a stationary distribution of $\tilde{X}(t)$, then

$$\|\tilde{\pi} - \pi\| \leq b^{-1}(\log(c/2) + 1)\|\mathbf{E}\|.$$

An important feature of such bounds is the logarithmic dependence of the right hand side on c. A significant role in the proof of the above theorem is played by $\bar{d}(t)$, called the ergodicity coefficient of H_t, which was defined in Chapter 4. Further extensions of results of this type for discrete-time Markov chains on general state spaces and for hidden Markov models are discussed in [58] and [59], respectively.

7

Open Problems

We conclude with some open problems.

1. Is there a useful comparison argument for the entropy constant?

2. Provide lower bounds on the entropy constant in terms of inverse τ?

Based on several examples, we ask if the following could always be true? If so, it would strengthen the classical inverse spectral gap lower bound on the total variation mixing time. Is there a universal constant $c > 0$ such that for every irreducible Markov chain P, we have

$$\frac{c}{\rho_0} \leq \tau_{\mathrm{TV}}(1/e)\,?$$

3. A functional analog of (5.3) is that for all $f : \Omega \to \mathsf{R}$,

$$\mathcal{E}_{\mathsf{KK}^*}(f, f) = \|f\|_2^2 - \|\mathsf{K}^* f\|_2^2.$$

If this, in conjunction with Lemma 12 of [64] (or more likely, a strengthened version of that lemma), could be used to directly bound $\Lambda_{\mathsf{KK}^*}(r)$ from below, then the final bound on the mixing time of the Thorp shuffle could be substantially smaller – it should drop to d^{15} or so, since we would be avoiding the use of the generalized Cheeger inequality (5.2).

Bounding the entropy decay of the Thorp shuffle, using either the entropy constant or the log-Sobolev constant, is another way to improve upon the currently weak estimates on the mixing time of this shuffle.

4. Can the spectral profile be estimated for the lamplighter problem (see [68]) on a discrete cube or more generally on an expander graph? This would tighten the estimates on the L^2 mixing time for the lamplighter problem.

5. For simple random walks on n-vertex graphs, how small can the log-Sobolev and the entropy constants be? The spectral gap is lower bounded by $\Omega(1/n^3)$ by an old result of Landau and Odlyzko [47]. The bound is tight, since a barbell graph achieves such a bound.

6. For the Bernoulli Process of Example 2.7 and 2.13 can a better lower bound on the spectral profile $\Lambda(r)$ be achieved when r is small? As discussed, this might lead to an easier proof of the mixing time of the exclusion process on $\mathbb{Z}^d/L\mathbb{Z}^d$.

References

[1] D. J. Aldous and J. Fill, *Reversible Markov Chains and Random Walks on Graphs*, (book to appear); URL for draft at http://stat-www.berkeley.edu/users/aldous/RWG/book.html

[2] N. Alon, R. Boppana and J. Spencer, "An asymptotic isoperimetric inequality," *Geom. and Funct. Anal.*, vol. 8, pp. 411–436, 1998.

[3] D. Bakry and M. Emery, "Diffusions hypercontractives," *Séminaire de Probabilités XIX, Lecture Notes in Math. 1123*, Springer-Verlag, pp. 177–206, 1985.

[4] M. Barlow, Th. Coulhon and A. Grigor'yan, "Manifolds and graphs with slow heat kernel decay," *Invent. Math.* vol. 144, pp. 609–649, 2001.

[5] S. Bobkov and P. Tetali, "Modified Log-Sobolev Inequalities in Discrete Settings," *Proc. of the ACM Symp. on Theory of Computing*, pp. 287–296, 2003; journal version in *Jour. of Theor. Probab.*, to appear.

[6] S. Bochner, "Vector fields and Ricci Curvature," *Bull. Amer. Math. Soc.*, vol. 52, pp. 776–797, 1946.

[7] C. Borgs, *Statistical Physics Expansion Methods in Combinatorics and Computer Science*, CBMS-SIAM Lecture Notes, in preparation.

[8] C. Borgs, J. Chayes, A. Frieze, J.H. Kim, P. Tetali, E. Vigoda and V. Vu, "Torpid mixing of some MCMC algorithms in statistical physics," *Proceedings of the 40th IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 218–229, 1999.

[9] A-S. Boudou, P. Caputo, P. Dai Pra and G. Posta, "Spectral gap estimates for interacting particle systems via a Bochner type identity," *Journal of Functional Analysis*, vol. 232, pp. 222–258, 2005.

[10] S. Boyd, P. Diaconis and L. Xiao, "Fastest mixing Markov chain on a graph," *SIAM Review* vol. 46, no. 4, pp. 667–689, 2004.

[11] S. Boyd, P. Diaconis, J. Sun and L. Xiao, "Fastest mixing Markov chain on a path," *The American Mathematical Monthly*, vol. 113, no. 1, pp. 70–74, 2006.

[12] P. Caputo, "Spectral gap inequalities in product spaces with conservation laws," in: *Stochastic Analysis on Large Scale Systems*, T. Funaki and H. Osada eds. Advanced Studies in Pure Mathematics, Japan, 2004.

[13] P. Caputo and G. Posta, "Entropy dissipation estimates in a zero–range dynamics," Preprint (arXiv:math.PR/0405455), 2004.

[14] P. Caputo and P. Tetali, "Walks, Transpositions, and Exclusion," Preprint (July 2004).

[15] M. Cryan, M. Dyer and D. Randall, "Approximately Counting Integral Flows and Cell-Bounded Contingency Tables," Proc. of the 37th Annual ACM Symp. on Theory of Computing (STOC), pp. 413–422, 2005.

[16] P. Diaconis and M. Shahshahani, "Time to reach*SIAM J. Math. Anal.*, vol. 18, no. 1, pp. 208–218, 1987.

[17] F. Chung, "Laplacians and the Cheeger inequality for directed graphs," *Annals of Combinatorics*, vol. 9, no. 1, pp. 1–19, 2005.

[18] T. Coulhon, A. Grigor'yan and C. Pittet, "A geometric approach to on-diagonal heat kernel lower bound on groups," *Ann. Inst. Fourier*, vol. 51, pp. 1763–1827, 2001.

[19] T. Coulhon, "Ultracontractivity and Nash-type inequalities," *J. Funct. Anal.*, vol. 141, pp. 510–539, 1996.

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, 1991.

[21] J-D. Deuschel and D.W. Stroock, *Large Deviations*, pp. 246, Academic Press, 1989.

[22] P. Diaconis and J. Fill, "Strong stationary times via a new form of duality," *The Annals of Probability*, vol. 18, no. 4, pp. 1483–1522, 1990.

[23] P. Diaconis and S. Holmes and R. Neal, "Analysis of a Non-reversible Markov Chain Sampler," *Annals of Applied Probability*, vol. 10, no. 3, pp. 726–752, 2000.

[24] P. Diaconis and L. Saloff-Coste, "Comparison Theorems for Reversible Markov Chains," *The Annals of Applied Probability*, vol. 3, no. 3, pp. 696–730, 1993.

[25] P. Diaconis and L. Saloff-Coste, "Logarithmic Sobolev inequalities for finite Markov chains," *The Annals of Applied Probability*, vol. 6, no. 3, pp. 695–750, 1996.

[26] P. Diaconis and L. Saloff-Coste, "Nash Inequalities for Finite Markov Chains," *Journal of Theoretical Probability*, vol. 9, pp. 459–510, 1996.

[27] P. Diaconis and D. Stroock, "Geometric bounds for eigenvalues of Markov chains," *The Annals of Applied Probability*, vol. 1, pp. 36–61, 1991.

[28] M. Dyer, A. Frieze and M. Jerrum, "On counting independent sets in sparse graphs," *SIAM J. Computing*, vol. 33, pp. 1527–1541, 2002.

[29] M. Dyer, L. Goldberg, M. Jerrum and R. Martin, "Markov chain comparison," Preprint (arXiv:math.PR/0410331), 2005.

[30] J. Fill, "Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process," *The Annals of Applied Probability*, vol. 1, no. 1, pp. 62–87, 1991.

[31] N. Fountoulakis and B.A. Reed, “The evolution of the conductance of a random graph,” in preparation.

[32] A. Frieze and E. Vigoda, “Survey of Markov Chains for Randomly Sampling Colorings,” To appear in Festschrift for Dominic Welsh (2006).

[33] S. Goel, “Analysis of top to bottom-k shuffles,” *The Annals of Applied Probability*, vol. 16, no. 1, pp. 30–55, 2006.

[34] S. Goel, R. Montenegro and P. Tetali, “Mixing time bounds and the spectral profile,” *Electronic Journal of Probability*, vol. 11, pp. 1–26, 2006.

[35] A. Grigor’yan, ‘ Heat kernel upper bounds on a complete non-compact manifold,” *Revista Matemá,tica Iberoamericaná*, vol. 10, pp. 395–452, 1994.

[36] P. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[37] M. Jerrum, *Counting, Sampling and Integrating : Algorithms & Complexity*, Birkhäuser Verlag, Basel, 2003.

[38] M. Jerrum, “Mathematical Foundations of the Markov Chain Monte Carlo Method,” in *Probabilistic Methods for Algorithmic Discrete Mathematics* edited by Habib et al., pp. 116–165, Springer-Verlag, Germany, 1998.

[39] M. Jerrum and A. Sinclair, “Conductance and the rapid mixing property for Markov chains: the approximation of the permanent resolved,” *Proceedings of the 20th Annual ACM Symposium on Theory of Computing (STOC 1988)*, pp. 235–243, 1988.

[40] M. Jerrum and A. Sinclair, “The Markov chain Monte Carlo method: an approach to approximate counting and integration,” Chapter 12 of *Approximation Algorithms for NP-hard Problems* edited by Dorit Hochbaum, PWS Publishing, Boston, 1996.

[41] M. Jerrum, A. Sinclair and E. Vigoda, “A polynomial-time approximation algorithm for the permanent of a matrix with non-negative entries,” *Journal of the ACM*, vol. 51, pp. 671–697, 2004.

[42] M. Jerrum and J-B.Son, “Spectral Gap and log-Sobolev constant for balanced matroids,” *Proc. of the 43rd Annual IEEE Symposium on Foundations of Computer Science (FOCS 2002)*, pp. 721–729, 2002.

[43] M. Jerrum, L. Valiant and V. Vazirani, “Random generation of combinatorial structures from a uniform distribution,” *Theoretical Computer Science*, vol. 43, pp. 169–188, 1998.

[44] W. Johnson and J. Lindenstrauss, “Extensions of Lipschitz maps into a Hilbert space,” *Contemp. Math.*, vol. 26, pp. 189–206, 1984.

[45] R. Kannan, “Markov chains and polynomial time algorithms,” Plenary Talk at *Proc. of 35th Annual IEEE Symp. on the Foundations of Computer Science*, pp. 656–671, 1994.

[46] R. Kannan, L. Lovász and R. Montenegro, “Blocking conductance and mixing in random walks,” *Combinatorics, Probability and Computing*, to appear, 2006.

[47] H.J. Landau and A.M. Odlyzko, “Bounds for eigenvalues of certain stochastic matrices,” *Linear Algebra Appl.* vol. 38, pp. 5–15, 1981.

[48] G. Lawler and A. Sokal, “Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger’s inequality,” *Transactions of the American Mathematical Society*, vol. 309, pp. 557–580, 1988.

[49] T.-Y. Lee and H.-T. Yau, "Logarithmic Sobolev inequalities for some models of random walks," *The Annals of Probability*, vol. 26, no. 4, pp. 1855–1873, 1998.

[50] A. Lichnérowicz, *Géométrie des groupes de transformations*, Dunod, Paris, 1958.

[51] L. Lovász and R. Kannan, "Faster mixing via average conductance," *Proc. of the 31st Annual ACM Symp. on Theory of Computing*, pp. 282–287, 1999.

[52] L. Lovász and S. Vempala, "Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithm," Proc. of the 44th IEEE Found. of Computer Science, Boston, 2003. To appear in Jour. Comp. Sys. Sci.(FOCS '03 special issue).

[53] L. Lovász and S. Vempala, "Hit-and-run from a corner," Proc. of the 36th ACM Symp. on the Theory of Computing, Chicago, 2004. To appear in SIAM J. Computing (STOC '04 special issue).

[54] L. Miclo, "Remarques sur l'hypercontractivité et l'évolution de l'entropie pour des chaînes de Markov finies," *Séminaire de probabilités de Strasbourg*, vol. 31, pp. 136–167, 1997.

[55] M. Mihail, "Conductance and Convergence of Markov Chains-A Combinatorial Treatment of Expanders," *Proc. of the 30th Annual Symposium on Foundations of Computer Science*, pp. 526–531, 1989.

[56] S. Miller and R. Venkatesan, "Spectral Analysis of Pollard Rho Collisions," *Proc. of the 7th Algorithmic Number Theory Symposium (ANTS VII)* in series *Lecture Notes in Computer Science (LNCS)*, Springer, to appear, 2006.

[57] A.Yu. Mitrophanov, "Stability and exponential convergence of continuous-time Markov chains," *J. Appl. Probab.*, vol. 40, pp. 970–979, 2003.

[58] A.Yu. Mitrophanov, "Sensitivity and convergence of uniformly ergodic Markov chains," *J. Appl. Probab.*, vol. 42, pp. 1003–1014, 2005.

[59] A.Yu. Mitrophanov, A. Lomsadze and M. Borodovsky, "Sensitivity of hidden Markov models," *J. Appl. Probab.*, vol. 42, pp. 632-642, 2005.

[60] R. Montenegro, "Vertex and edge expansion properties for rapid mixing," *Random Structures & Algorithms*, vol. 26, no. 1–2, pp. 52–68, 2005.

[61] R. Montenegro, "Duality and evolving set bounds on mixing times," Preprint, 2006.

[62] R. Montenegro, "Eigenvalues of non-reversible Markov chains: their connection to mixing times, reversible Markov chains, and Cheeger inequalities," Preprint (arXiv:math.PR/0604362), 2006.

[63] B. Morris, "The mixing time for simple exclusion," *Annals of Applied Probability*, to appear, 2006.

[64] B. Morris, "The mixing time of the Thorp shuffle," *equation and lemma numbering taken from version arXiv:math.PR/0507307v1*, 2005.

[65] B. Morris and Y. Peres, "Evolving sets, mixing and heat kernel bounds," *Probability Theory and Related Fields*, vol. 133, no. 2, pp. 245–266, 2005.

[66] A. Naor, M. Sammer and P. Tetali, "The Fastest Mixing Markov Process and the Subgaussian Constant," *Preprint,* 2005.

[67] J.R. Norris, *Markov Chains,* Cambridge University Press, 1997.

[68] Y. Peres and D. Revelle, "Mixing times for random walks on finite lamplighter groups," *Electronic J. on Probab.*, vol. 9, pp. 825–845, 2004.

[69] D. Randall, "Rapidly mixing Markov chains with applications in computer science and physics," *Computing in Science & Engineering*, vol. 8, no. 2, pp. 30–41, March, 2006.

[70] D. Randall and P. Tetali, "Analyzing Glauber dynamics using comparison of Markov chains," *J. Math. Physics*, vol. 41, pp. 1598–1615, 2000.

[71] M. Reed and B. Simon, *Methods of Modern Mathematical Physics II: Fourier Analysis, Self-Adjointness*, Academic Press Inc, New York, 1975.

[72] L. Saloff-Coste, "Total variation lower bounds for finite Markov chains: Wilson's lemma," in *Random walks and geometry,* pp. 515–532, Walter de Gruyter GmbH & Co. KG, Berlin, 2004.

[73] M. Sammer and P. Tetali, "Concentration on the Discrete Torus using Transportation," submitted to *Combin. Probab. & Computing.*

[74] E. Seneta, "Coefficients of ergodicity: structure and applications," *Adv. Appl. Prob.*, vol. 11, pp. 576–590, 1979.

[75] A. Sinclair, "Improved bounds for mixing rates of Markov chains and multicommodity flow," *Combinatorics, Probability and Computing*, vol. 1, no. 4, pp. 351–370, 1992.

[76] E. Stein, "Interpolation of Linear Operators," *Trans. Amer. Math. Soc.*, vol. 83, pp. 482–492, 1956.

[77] J. Sun, S. Boyd, L. Xiao and P. Diaconis, "The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem," *SIAM Review,* to appear, 2006.

[78] D. Wilson, "Mixing times of lozenge tiling and card shuffling Markov chains," *The Annals of Applied Probability*, vol. 14, no. 1, pp. 274–325, 2004.

[79] D. Wilson, "Mixing Time of the Rudvalis Shuffle," *Electronic Communications in Probability*, vol. 8, no. 77–85, 2003.

[80] H. T. Yau, "Logarithmic Sobolev inequality for generalized simple exclusion process," *Probability Theory and Related Fields*, vol. 109, no. 4, pp. 507–538, 1997.

Appendix

Our main focus has been on bounds for L^2 distance. Our bounds on mixing times in L^2 and relative entropy also yield bounds on the total variation mixing time using the following well-known inequality relating probability measures ν and μ.

$$\|\nu - \mu\|_{\mathrm{TV}} = \frac{1}{2}\left\|\frac{\nu}{\mu} - 1\right\|_{1,\mu} \leq \frac{1}{2}\left\|\frac{\nu}{\mu} - 1\right\|_{2,\mu}. \tag{7.1}$$

Further assuming that ν is absolutely continuous with respect to μ, the so-called Pinsker inequality (see Lemma 12.6.1 in [20] for a proof), asserts that:

$$\|\nu - \mu\|_{\mathrm{TV}}^2 \leq \frac{1}{2}D(\nu\|\mu) \tag{7.2}$$

Finally the general inequality $(\mathbb{E}_\mu f)\mathrm{Ent}_\mu(f) \leq \mathrm{Var}_\mu(f)$, valid for all measurable functions on an arbitrary probability space (since $\log \frac{f}{\mathbb{E}_\mu f} \leq \frac{f}{\mathbb{E}_\mu f} - 1$), when applied to $f = \nu/\mu$ implies that,

$$D(\nu\|\mu) \leq \left\|\frac{\nu}{\mu} - 1\right\|_{2,\mu}^2. \tag{7.3}$$

In a sense the L^∞, or relative pointwise distance, is the strongest of all distances. Our L^2 bounds also induce L^∞ bounds. Observe that if

$t = t_1 + t_2$ then

$$\begin{aligned}\left|\frac{H_t(x,y)-\pi(y)}{\pi(y)}\right| &= \left|\frac{\sum_z \left(H_{t_1}(x,z)-\pi(z)\right)\left(H_{t_2}(z,y)-\pi(y)\right)}{\pi(y)}\right| \\ &= \left|\sum_z \pi(z)\left(\frac{H_{t_1}(x,z)}{\pi(z)}-1\right)\left(\frac{H^*_{t_2}(y,z)}{\pi(z)}-1\right)\right| \\ &\le \left\|h^x_{t_1}-1\right\|_2 \left\|h^{*y}_{t_2}-1\right\|_2 \end{aligned} \tag{7.4}$$

where the inequality follows from Cauchy-Schwartz. Several bounds on L_∞ mixing then follow immediately, including

$$\tau_2(\epsilon) \le \tau_\infty(\epsilon) \le \tau_2\left(\epsilon\sqrt{\frac{\pi_*}{1-\pi_*}}\right)$$

and

$$\tau_2(\epsilon) \le \tau_\infty(\epsilon) \le \tau_2^{\mathsf{P}}(\sqrt{\epsilon}) + \tau_2^{\mathsf{P}^*}(\sqrt{\epsilon})\,.$$

The first of these two is somewhat unappealing because the asymptotic portion of $\tau_2(\epsilon)$ is of the form $\lambda^{-1}\log(1/\epsilon)$, and so taking $\tau_2\left(\epsilon\sqrt{\frac{\pi_*}{1-\pi_*}}\right)$ adds an extra factor of $\lambda^{-1}\log(1/\pi_*)$ to the $\tau_2(\epsilon)$ bound, potentially large relative to spectral profile bounds. The second bound unfortunately requires study of both P and P^*. However, if P is reversible then this last bound becomes

$$\tau_2(\epsilon) \le \tau_\infty(\epsilon) \le 2\tau_2(\sqrt{\epsilon})\,. \tag{7.5}$$

More generally, most bounds in this paper were the same for P and P^*. For instance, (7.5) holds for the spectral profile bounds on L^2 mixing in terms of $\Lambda(r)$. In particular, the Dirichlet form satisfies

$$\mathcal{E}_{\mathsf{P}}(f,f) = \mathcal{E}_{\mathsf{P}^*}(f,f)$$

and so $\lambda(\mathsf{P}) = \lambda(\mathsf{P}^*)$, $\Lambda_{\mathsf{P}}(r) = \Lambda_{\mathsf{P}^*}(r)$, $\rho(\mathsf{P}) = \rho(\mathsf{P}^*)$ and $\Phi_{\mathsf{P}}(r) = \Phi_{\mathsf{P}^*}(r)$ (as $\mathsf{Q}(A,A^c) = \mathcal{E}(1_A,1_A)$).

It is not as clear how the f-congestion bounds behave for P^*. However, if π is uniform then

$$\begin{aligned}\Psi(A) &= \Psi(A^c) = \min_{\pi(B)=\pi(A)} \mathsf{Q}(A^c,B) \\ &= \min_{\pi(B)=\pi(A)} \mathsf{Q}_{\mathsf{P}^*}(B,A^c) \ge \min_{\pi(B)=\pi(A)} \Psi_{\mathsf{P}^*}(B)\end{aligned}$$

and so $\tilde\phi(r) \ge \tilde\phi_{\mathsf{P}^*}(r)$. The converse follows similarly, so $\tilde\phi_{\mathsf{P}}(r) = \tilde\phi_{\mathsf{P}^*}(r)$ when π is uniform, and (7.5) holds for modified-conductance bounds.