6.S891 Lecture 4: Spectral Methods, Conductance, and Canonical Paths

Kuikui Liu

September 19, 2023

Our second class of methods is based on the spectrum of the transition probability matrix P. This an extremely powerful method which gives a beautiful connection between probability and linear algebra. We show combinatorial methods for controlling the spectral gap of P based on cuts and flows. The main application we will eventually build up to is designing an FPRAS for the famous ferromagnetic Ising model. Later on in the course, we'll see even more powerful functional analytic tools for studying mixing times.

1 The Poincaré Inequality

Let μ be a probability distribution over a finite state space Ω. This distribution induces an inner product on the space of functions $\{f : \Omega \to \mathbb{R}\} \cong \mathbb{R}^{\Omega}$ given by

$$\langle f, g\rangle_{\mu} \stackrel{\mathsf{def}}{=} \mathbb{E}_{\mu}[fg] = \sum_{x\in\Omega} \mu(x)f(x)g(x).$$

Fact 1.1. *A Markov chain P is reversible w.r.t. μ if and only if P is self-adjoint w.r.t. $\langle\cdot,\cdot\rangle_{\mu}$, i.e. $\langle f, \mathsf{P}g\rangle_{\mu} = \langle \mathsf{P}f, g\rangle_{\mu}$ for all $f, g : \Omega \to \mathbb{R}$.*

A proof is given in Appendix B. This tells us that reversible Markov chains P must have real eigenvalues, and we can order them as $-1 \le \lambda_{|\Omega|} \le \cdots \le \lambda_2 \le \lambda_1 = 1$. We write $\lambda_* \stackrel{\mathsf{def}}{=} \max_{i>1} |\lambda_i| = \max\{\lambda_2, |\lambda_{|\Omega|}|\}$ for the second largest eigenvalue in absolute value. Note that irreversibility translates to $\lambda_2 < 1$, and aperiodicity translates to $\lambda_{|\Omega|} > -1$. In particular, ergodicity yields $\lambda_* < 1$. We also write $\gamma = 1 - \lambda_2$ for the *spectral gap*, and $\gamma_* = 1 - \lambda_*$ for the *absolute spectral gap*. Note that if P has been lazified, all eigenvalues of P are nonnegative, in which case $\gamma = \gamma_*$.

Theorem 1.2 (Spectral Gap Implies Rapid Mixing; see e.g. [LPW17])**.** *Let P be an ergodic Markov chain on Ω which is reversible w.r.t. μ. Then for every $x \in \Omega$ and every $\epsilon > 0$,*

$$\mathrm{T}_{\mathsf{mix}}(\epsilon; \delta_x, \mathsf{P}) \le \frac{1}{\gamma_*} \log\left(\frac{1}{2\epsilon \cdot \sqrt{\mu(x)}}\right).$$

In particular,

$$\mathrm{T}_{\mathsf{mix}}(\epsilon) \le \frac{1}{\gamma_*}\left(\frac{1}{2}\log\frac{1}{\mu_{\min}} + \log\frac{1}{2\epsilon}\right),$$

where $\mu_{\min} \stackrel{\mathsf{def}}{=} \min_{x\in\mathrm{supp}(\mu)} \mu(x)$ is the minimum nonzero probability under μ.

Note that having a spectral gap is *necessary* for rapid mixing. We prove this converse in Appendix A.

Lemma 1.3 (Rapid Mixing Implies Spectral Gap; see e.g. [LPW17])**.** *Let P be an ergodic Markov chain on Ω which is reversible w.r.t. μ. Then*

$$\mathrm{T}_{\mathsf{mix}}(\epsilon) \ge \left(\frac{1}{\gamma_*} - 1\right)\log\frac{1}{2\epsilon}.$$

The goal of this section is to prove Theorem 1.2; an "alternative" proof based on a full spectral decomposition of P is provided in Appendix A. The high-level idea is to replace $\|\mu - \nu\|_{\mathsf{TV}}$ with χ^2-divergence

$$\chi^2(\nu \,\|\, \mu) \stackrel{\text{def}}{=} \mathrm{Var}_\mu\left(\frac{d\nu}{d\mu}\right),$$

as our measure of distance to stationarity.[1] This distance turns out to be easier to work with than total variation since it looks more like Euclidean distance. In particular, for every $f : \Omega \to \mathbb{R}$, its variance admits the following linear algebraic interpretation

$$\mathrm{Var}_\mu(f) = \langle f, f\rangle_\mu - \langle f, \mathbf{1}\rangle_\mu^2.$$

We need the following comparison lemma between χ^2-divergence and TV-distance.

Lemma 1.4. *For every pair of distributions μ, ν on Ω, writing $\frac{d\nu}{d\mu}(x) = \frac{\nu(x)}{\mu(x)}$ for the* density *of ν w.r.t. μ, we have the inequality*

$$\|\mu - \nu\|_{\mathsf{TV}}^2 \le \frac{1}{4}\,\mathrm{Var}_\mu\left(\frac{d\nu}{d\mu}\right).$$

We prove this in Appendix B. Before we give the full proof of Theorem 1.2, we first present another viewpoint on the spectral gap γ which will be useful. Define the *Dirichlet form* of a reversible Markov chain P by

$$\mathcal{E}_{\mathsf{P}}(f, g) \stackrel{\text{def}}{=} \langle f, (\mathsf{Id} - \mathsf{P})g\rangle_\mu = \frac{1}{2}\sum_{x,y\in\Omega} \mu(x)\mathsf{P}(x \to y)\cdot(f(x) - f(y))\cdot(g(x) - g(y)).$$

If $f = g$, this can be interpreted as a measure of the "local variance" of the function f w.r.t. the transition of P. The matrix $\mathsf{Id} - \mathsf{P}$ is sometimes called the *Laplacian*. On the other hand, a simple calculation reveals that the global variance of a function f can be rewritten as

$$\mathrm{Var}_\mu(f) = \frac{1}{2}\sum_{x,y\in\Omega} \mu(x)\mu(y)\cdot(f(x) - f(y))^2.$$

We say P satisfies a *Poincaré Inequality* with constant $C > 0$ if

$$C \cdot \mathrm{Var}_\mu(f) \le \mathcal{E}_{\mathsf{P}}(f, f), \qquad \forall f : \Omega \to \mathbb{R}.$$

Fact 1.5. *Let P be a reversible Markov chain. Then we have the identity*

$$\gamma = \inf_f \frac{\mathcal{E}_{\mathsf{P}}(f, f)}{\mathrm{Var}_\mu(f)},$$

where the infimum is over all functions with nonzero global variance.

This is just the variational characterization of eigenvalues. The functions with nonzero global variance are precisely those which are not constant; in particular, they have a nonzero component orthogonal to the top (right) eigenfunction $\mathbf{1}$. Because of Fact 1.5, the spectral gap γ is also sometimes called the *Poincaré constant* of P. Later on, we'll see how to define other useful functional analytic constants which better capture mixing times.

Proof of Theorem 1.2. Let $x \in \Omega$ be arbitrary. By Lemma 1.4, it suffices to show that for all $t \in \mathbb{N}$,

$$\mathrm{Var}_\mu\left(\frac{d(\delta_x\mathsf{P}^t)}{d\mu}\right) \le \lambda_*^{2t}\cdot\mathrm{Var}_\mu\left(\frac{d\delta_x}{d\mu}\right) = \lambda_*^{2t}\cdot\frac{1-\mu(x)}{\mu(x)}.$$

[1] For $p \ge 1$, one can also define L_p-norms via $\|f\|_{\mu,p} \stackrel{\text{def}}{=} \mathbb{E}_\mu\left[|f|^p\right]^{1/p}$, and L_p-divergences via between probability measures μ, ν via $\left\|\frac{d\mu}{d\nu} - \mathbf{1}\right\|_{\mu,p}$. The case $p = 1$ recovers total variation (up to a factor of $1/2$), while the case $p = 2$ yields χ^2-divergence (after squaring).

For this, since $\mathrm{Var}_\mu(f) = \langle f, f\rangle_\mu - \langle f, \mathbf{1}\rangle_\mu^2$ and $\frac{d(\nu\mathsf{P})}{d\mu} = \mathsf{P}\cdot\frac{d\nu}{d\mu}$ for all distributions ν,[2]

$$\begin{aligned}
\mathrm{Var}_\mu\left(\frac{d\nu}{d\mu}\right) - \mathrm{Var}_\mu\left(\frac{d(\nu\mathsf{P})}{d\mu}\right) &= \left\langle \frac{d\nu}{d\mu}, \frac{d\nu}{d\mu}\right\rangle_\mu - \left\langle \frac{d\nu}{d\mu}, \mathbf{1}\right\rangle_\mu^2 - \left\langle \mathsf{P}\frac{d\nu}{d\mu}, \mathsf{P}\frac{d\nu}{d\mu}\right\rangle_\mu + \left\langle \mathsf{P}\frac{d\nu}{d\mu}, \mathbf{1}\right\rangle_\mu^2 \\
&= \left\langle \frac{d\nu}{d\mu}, (I - \mathsf{P}^2)\frac{d\nu}{d\mu}\right\rangle_\mu \qquad \text{(Using } \mathsf{P}\mathbf{1} = \mathbf{1} \text{ and self-adjointness)} \\
&= \mathcal{E}_{\mathsf{P}^2}\left(\frac{d\nu}{d\mu}, \frac{d\nu}{d\mu}\right).
\end{aligned}$$

Using Fact 1.5 and the fact that the eigenvalues of P^2 are squares of the eigenvalues of P, we see that P^2 satisfies a Poincaré Inequality with constant $1 - \lambda_*^2$. It follows that

$$\mathrm{Var}_\mu\left(\frac{d\nu}{d\mu}\right) - \mathrm{Var}_\mu\left(\frac{d(\nu\mathsf{P})}{d\mu}\right) \geq (1 - \lambda_*^2)\cdot \mathrm{Var}_\mu\left(\frac{d\nu}{d\mu}\right),$$

which upon rearranging, yields

$$\mathrm{Var}_\mu\left(\frac{d(\nu\mathsf{P})}{d\mu}\right) \leq \lambda_*^2 \cdot \mathrm{Var}_\mu\left(\frac{d\nu}{d\mu}\right), \qquad \forall \nu.$$

Applying this iteratively to the distributions $\nu = \delta_x \mathsf{P}^j$ for $j = 0, \ldots, t-1$ finishes the proof. □

Spectral Methods vs. Markovian Couplings In the previous lecture, we saw simple probabilistic methods for proving fast mixing based on Markovian couplings. It turns out that there are natural Markov chains for natural sampling problems where

- one can certify fast mixing by bounding its spectral gap, and
- provably, *every* Markovian coupling necessarily requires exponential time to coalesce.

This was shown for a simple swap-based Markov chain on the collection of perfect and "near-perfect" matchings in a bipartite graph. Fast mixing was proved by Jerrum–Sinclair using the methods we discuss in this lecture [JS89]. The fact that no Markovian coupling can certify fast mixing is a beautiful result due to Kumar–Ramesh [KR01].

2 The Conductance Method

We now describe combinatorial ways of getting a handle on the Poincaré constant γ. The first is to study the sparsity of cuts in the state space Ω.

Definition 1 (Conductance)**.** *Let P be a Markov chain which is reversible w.r.t. a distribution μ on Ω. For every $S \subseteq \Omega$, define the* conductance *of S w.r.t. P to be the ratio*

$$\Phi(S) \stackrel{\text{def}}{=} \frac{\sum_{x\in S, y\in\Omega\setminus S} \mu(x)\mathsf{P}(x\to y)}{\sum_{x\in S}\mu(x)}.$$

Further define the conductance[3] *of P to be*

$$\Phi(\mathsf{P}) \stackrel{\text{def}}{=} \inf_{S\subseteq\Omega:\mu(S)\leq 1/2} \Phi(S).$$

Roughly speaking, the conductance of a set $S \subseteq \Omega$ measures how likely the Markov chain started within S is to leave S in one step. If $\Phi(S)$ is small, then the Markov chain is likely to stay trapped within S, preventing mixing. In the other direction, a strong lower bound on $\Phi(\mathsf{P})$ says that there are no such "bottlenecks" in the state space Ω. In this case, we expect P to mix rapidly. The following theorem is one way to formalize this; we can quantitatively connect the conductance $\Phi(\mathsf{P})$, a purely combinatorial quantity, with the spectral gap γ of P.

[2] This follows from reversibility, since for all functions $g : \Omega \to \mathbb{R}$, $\left\langle \frac{d(\nu\mathsf{P})}{d\mu}, g\right\rangle_\mu = \nu^\top \mathsf{P} g = \left\langle g, \mathsf{P}\frac{d\nu}{d\mu}\right\rangle_\mu$.

[3] The number $\Phi(\mathsf{P})$ is also sometimes called the *isoperimetric constant* or the *Cheeger constant*.

Theorem 2.1 (Cheeger's Inequality; [SJ89; LS88])**.** *For every reversible Markov chain* P*,*

$$\frac{1}{2}\Phi(\mathsf{P})^2 \le \gamma \le 2\Phi(\mathsf{P}).$$

The upper bound is easy; one can just plug in the indicator function $\mathbf{1}_S$ into the Poincaré Inequality. The lower bound is more nontrivial; for a proof, see e.g. [LPW17].

Strategy for Mixing Lower Bounds Suppose one wanted to show that a reversible Markov chain mixes *slowly*. Combining Theorem 2.1 with Lemma 1.3, it suffices to find *any* $S \subseteq \Omega$ such that $\Phi(S)$ is small.

2.1 Flows and Canonical Paths

Theorem 2.1 gives a connection between the Poincaré Inequality and cuts in Ω viewed as a graph with edges given by P. A dual viewpoint is to look at path/flows in Ω. Imagine we set up a multicommodity flow problem on Ω as follows: We wish to send $\mu(x)\mu(y)$ "units" of flow from state x to y (and vice versa), routed using the transitions of P. In particular, we choose a collection of paths $\mathcal{P} = \{\mathcal{P}_{x\to y}\}_{x,y\in\Omega}$, one for each pair $x, y \in \Omega$, where $\mathcal{P}_{x\to y}$ is a sequence of transitions $x = z_0 \to \cdots \to z_\ell = y$ using P. However, each possible transition $a \to b$ has a "cost" for usage (or "capacity"), given by $\mu(a)\mathsf{P}(a \to b)$. We want to route the flow efficiently, as quantified by the length of the paths $|\mathcal{P}_{x\to y}|$ and the *congestion* of a transition $a \to b$, defined as

$$C_{\mathcal{P}}(a \to b) \stackrel{\text{def}}{=} \frac{1}{\mu(a)\mathsf{P}(a \to b)} \sum_{x,y\in\Omega:(a,b)\in\mathcal{P}_{x\to y}} \mu(x)\mu(y).$$

Theorem 2.2 (Jerrum–Sinclair; [JS89])**.** *For every collection of paths* $\{\mathcal{P}_{x\to y}\}_{x,y\in\Omega}$*,*

$$\frac{1}{\gamma} \le \max_{x,y\in\Omega} |\mathcal{P}_{x\to y}| \cdot \max_{a\to b} C_{\mathcal{P}}(a \to b).$$

Proof. We verify the Poincaré Inequality. Let $f : \Omega \to \mathbb{R}$ be an arbitrary function. Then

$$\begin{aligned}
\mathrm{Var}_\mu(f) &= \frac{1}{2}\sum_{x,y\in\Omega} \mu(x)\mu(y)\cdot(f(x)-f(y))^2 \\
&= \frac{1}{2}\sum_{x,y\in\Omega} \mu(x)\mu(y)\cdot\left(\sum_{(a,b)\in\mathcal{P}_{x\to y}} (f(a)-f(b))\right)^2 && \text{(Telescoping)} \\
&\le \max_{x,y\in\Omega}|\mathcal{P}_{x\to y}|\cdot\frac{1}{2}\sum_{x,y\in\Omega}\mu(x)\mu(y)\sum_{(a,b)\in\mathcal{P}_{x\to y}}(f(a)-f(b))^2 && \text{(Cauchy–Schwarz)} \\
&= \max_{x,y\in\Omega}|\mathcal{P}_{x\to y}|\cdot\sum_{\text{transition }(a,b)}(f(a)-f(b))^2\sum_{x,y\in\Omega:(a,b)\in\mathcal{P}_{x\to y}}\mu(x)\mu(y) \\
& && \text{(Exchange summation order)} \\
&\le \max_{x,y\in\Omega}|\mathcal{P}_{x\to y}|\cdot\max_{a\to b}C_{\mathcal{P}}(a\to b)\cdot\sum_{a,b\in\Omega}\mu(a)\mathsf{P}(a\to b)\cdot(f(a)-f(b))^2 \\
&= \max_{x,y\in\Omega}|\mathcal{P}_{x\to y}|\cdot\max_{a\to b}C_{\mathcal{P}}(a\to b)\cdot\mathcal{E}_{\mathsf{P}}(f,f).
\end{aligned}$$

Rearranging and applying Fact 1.5 completes the proof. □

There is a precise duality with conductance related to linear programming, which we won't have time to discuss.[4]

[4]Technically, one needs to look at a slightly different notion of conductance given by $\tilde{\Phi}(S) = \frac{\sum_{x\in S,y\in\Omega\setminus S}\mu(x)\mathsf{P}(x\to y)}{\sum_{x\in\Omega,y\in\Omega\setminus S}\mu(x)\mu(y)}$; note that $\Phi(S) \le \tilde{\Phi}(S) \le 2\Phi(S)$ if $\mu(S) \le 1/2$.

Strategy for Mixing Upper Bounds Given Theorem 2.2, one strategy for upper bounding mixing times (or lower bounding spectral gaps) is to construct *any* set of short paths $\mathcal{P} = \{\mathcal{P}_{x\to y}\}$ with small congestion $\max_{a\to b} C_{\mathcal{P}}(a \to b)$. This sometimes called the *canonical paths* method. Typically, bounding the lengths of the paths isn't an issue. The main challenge is ensuring small congestion.

Example 1 (Hypercube). We illustrate a simple construction of canonical paths for Glauber dynamics on $\{\pm 1\}^n$. For two vectors $x, y \in \{\pm 1\}^n$, we define a canonical path $\mathcal{P}_{x\to y}$ by going in order of the coordinates $1, \ldots, n$, and flipping x_i if $y_i \neq x_i$. More precisely, $\mathcal{P}_{x\to y} = \{x = z_0 \to \cdots \to z_n = y\}$, where z_k agrees with y on the first k coordinates, and agrees with x on the remaining $n-k$ coordinates.

Clearly, $|\mathcal{P}_{x\to y}| \leq n$. Let us now bound the congestion. Note that

$$\mathsf{P}(a \to b) = \frac{1}{2n}.$$

Since μ is uniform over $\{\pm 1\}^n$,

$$C_{\mathcal{P}}(a \to b) = 2n \cdot 2^{-n} \cdot \#\{(x, y) : (a \to b) \in \mathcal{P}_{x\to y}\}.$$

We prove that this number of paths is at most $\frac{1}{2} \cdot 2^n$, from which it will follow that $C_{\mathcal{P}}(a \to b) \leq n$ for all transition $a \to b$. Fix a transition $a \to b$, where coordinate k is flipped for some $k \in [n]$. If $x, y \in \{\pm 1\}^n$ are such that $(a \to b) \in \mathcal{P}_{x\to y}$, then x must agree with a on coordinates $k, \ldots, n$ (since they haven't been processed yet by the time we reach a), and y must agree with b on coordinates $1, \ldots, k$ (since they have been processed already by the time we reach b). It follows that there are 2^{k-1} choices for x and 2^{n-k} choices for y. In particular,

$$\#\{(x, y) : (a \to b) \in \mathcal{P}_{x\to y}\} \leq 2^{k-1} \cdot 2^{n-k} = \frac{1}{2} \cdot 2^n.$$

It follows that $\max_{a\to b} C_{\mathcal{P}}(a \to b) \leq n$, and so $1/\gamma \leq n^2$ by Theorem 2.2. Further combined with Theorem 1.2, this implies that Glauber dynamics mixes in $O(n^3)$-steps.

We saw in the previous lecture that a simple coupling argument certifies $O(n \log n)$ mixing for Glauber dynamics with uniform stationary measure on $\{\pm 1\}^n$, which is sharp. Example 1 highlights a key downside behind spectral methods and canonical paths. Even if one obtains sharp bounds on the spectral gap, in many applications, the mixing time will often be off by a factor of n from the true mixing time. This is because we're only using the second largest eigenvalue of P, ignoring all other eigenvalues. One often loses an additional factor of n from using canonical paths to bound the spectral gap. Nevertheless, there are natural Markov chains for important sampling problems for which canonical paths remains the only known method certifying fast mixing.

References

[JS89] M. Jerrum and Alistair Sinclair. "Approximating the Permanent". In: *SIAM Journal on Computing* 18.6 (Dec. 1989), pp. 1149–1178 (cit. on pp. 3, 4).

[KR01] V.S. Anil Kumar and H. Ramesh. "Coupling vs. conductance for the Jerrum–Sinclair chain". In: *Random Structures & Algorithms* 18.1 (2001), pp. 1–17 (cit. on p. 3).

[LPW17] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. 2nd ed. American Mathematical Society, 2017 (cit. on pp. 1, 4).

[LS88] Gregory F. Lawler and Alan D. Sokal. "Bounds on the L^2 Spectrum for Markov Chains and Markov Processes: A Generalization of Cheeger's Inequality". In: *Transactions of the American Mathematical Society* 309.2 (1988), pp. 557–580 (cit. on p. 4).

[SJ89] Alistair Sinclair and Mark Jerrum. "Approximate counting, uniform generation and rapidly mixing Markov chains". In: *Information and Computation* 82.1 (1989), pp. 93–133. ISSN: 0890-5401 (cit. on p. 4).

A Spectral Theory

Recall that for reversible Markov chains P, the eigenvalues of P are real, and can be ordered as $-1 \leq \lambda_{|\Omega|} \leq \cdots \leq \lambda_2 \leq \lambda_1 = 1$.

Lemma A.1 (Spectral Decomposition of Reversible Markov Chains)**.** *Let* P *be reversible w.r.t.* μ*. Then* $\mathbb{R}^\Omega$ *admits a basis* $\{\psi_i\}_{i=1}^{|\Omega|}$ *satisfying the following:*

1. *Each* ψ_i *is a (right) eigenfunction of* P*, i.e.* $\mathsf{P}\psi_i = \lambda_i\psi_i$*. Furthermore,* $\psi_1 = \mathbf{1}$*.*
2. *The collection* $\{\psi_i\}_{i=1}^{|\Omega|}$ *is orthonormal w.r.t. the inner product* $\langle\cdot,\cdot\rangle_\mu$*.*
3. *For every* $x, y \in \Omega$ *and every* $t \in \mathbb{N}$*,*

$$\frac{\mathsf{P}^t(x \to y)}{\mu(y)} = \sum_{i=1}^{|\Omega|} \lambda_i^t \psi_i(x)\psi_i(y).$$

In particular, we may express the measure $\delta_x\mathsf{P}^t$ *in density form as*

$$\frac{d(\delta_x\mathsf{P}^t)}{d\mu} = \sum_{i=1}^{|\Omega|} \lambda_i^t \psi_i(x) \cdot \psi_i.$$

Proof. Since P is in general asymmetric, our strategy will be to reduce the Spectral Theorem for symmetric matrices. Let $D_\mu = \mathrm{diag}(\mu)$, and consider $A_\mu \stackrel{\mathsf{def}}{=} D_\mu^{1/2}\mathsf{P}D_\mu^{-1/2}$. (In the spectral graph theory literature, this is sometimes called the "normalized adjacency matrix" to distinguish it from the usual random walk matrix.) This matrix has entries $\frac{\sqrt{\mu(x)}}{\sqrt{\mu(y)}}\mathsf{P}(x \to y)$, and hence is symmetric by reversibility of P. Furthermore, it has the same set of eigenvalues as P, since if $\mathsf{P}\psi = \lambda\psi$, then $A_\mu D_\mu^{1/2}\psi = \lambda D_\mu^{1/2}\psi$ (and vice versa). Finally, a direct calculation reveals that $\sqrt{\mu}$ (taken entrywise) is a right eigenfunction of A_μ with eigenvalue 1. It follows from the Spectral Theorem that A_μ admits a decomposition

$$A_\mu = \sum_{i=1}^{|\Omega|} \lambda_i \varphi_i \varphi_i^\top,$$

where $\{\varphi_i\}_{i=1}^{|\Omega|}$ is a collection of eigenfunctions which are orthonormal w.r.t. the standard Euclidean inner product, and $\varphi_1 = \sqrt{\mu}$.

We set $\psi_i = D_\mu^{-1/2}\varphi_i$ for all $i = 1, \ldots, |\Omega|$; clearly, $\psi_1 = \mathbf{1}$. As we saw earlier when we argued that P and A_μ have the same eigenvalues, we have $\mathsf{P}\psi_i = \lambda_i\psi_i$ for all i. Furthermore, $\langle\psi_i, \psi_j\rangle_\mu = \langle\varphi_i, \varphi_j\rangle$ and so we get orthonormality w.r.t. $\langle\cdot,\cdot\rangle_\mu$. Only the third claim remains. Since

$$\begin{aligned}
\mathsf{P}^t &= (D_\mu^{-1/2}A_\mu D_\mu^{1/2})^t = D_\mu^{-1/2}A_\mu^t D_\mu^{1/2} \\
&= \sum_{i=1}^{|\Omega|} \lambda_i^t \cdot \left(D_\mu^{-1/2}\varphi_i\right) \cdot (D_\mu\varphi_i)^\top \\
&= \sum_{i=1}^{|\Omega|} \lambda_i^t \psi_i\psi_i^\top \cdot D_\mu,
\end{aligned}$$

we see that

$$\frac{d(\delta_x\mathsf{P}^t)}{d\mu} = \delta_x^\top \mathsf{P}^t D_\mu^{-1} = \sum_{i=1}^{|\Omega|} \lambda_i^t \langle\delta_x, \psi_i\rangle\psi_i^\top$$

as row vectors, and so we're done. □

"Alternative" Proof of Theorem 1.2. Applying Lemma 1.4 and the inequality $\lambda_*^{1/\gamma_*} \le 1/e$, it suffices to show that for every $t \in \mathbb{N}$,

$$\left\| \frac{d(\delta_x\mathsf{P}^t)}{d\mu} - \mathbf{1} \right\|_{\mu,2} \le \frac{\lambda_*^t}{\sqrt{\mu(x)}},$$

where $\|f\|_{\mu,2} \stackrel{\text{def}}{=} \sqrt{\langle f, f\rangle_\mu}$. Using Lemma A.1 and expanding, the left-hand side is precisely

$$\left\| \frac{d(\delta_x \mathsf{P}^t)}{d\mu} - \mathbf{1} \right\|_{\mu,2}^2 = \sum_{i=2}^{|\Omega|} \lambda_i^{2t} \psi_i(x)^2 \le \lambda_*^{2t} \sum_{i=2}^{|\Omega|} \psi_i(x)^2 = \frac{\lambda_*^{2t}}{\mu(x)^2} \sum_{i=2}^{|\Omega|} \langle \psi_i, \delta_x\rangle_\mu^2 \le \frac{\lambda_*^{2t}}{\mu(x)^2} \cdot \langle \delta_x, \delta_x\rangle_\mu.$$

This completes the proof. Note that in the last step, we used orthonormality of $\{\psi_i\}_{i=1}^{|\Omega|}$ w.r.t. the inner product $\langle \cdot, \cdot \rangle_\mu$ to write $\delta_x = \sum_{i=1}^{|\Omega|} \langle \psi_i, \delta_x\rangle_\mu \cdot \psi_i$. □

Proof of Lemma 1.3. Let ψ be a (right) eigenfunction of P with eigenvalue λ_*. Then $\langle \psi, \mathbf{1}\rangle_\mu = 0$, and so for every $t \in \mathbb{N}$ and every $x \in \Omega$,

$$\begin{aligned} \lambda_*^t \cdot |\psi(x)| &= \left|(\mathsf{P}^t \psi)(x)\right| \\ &= \left| \sum_{y \in \Omega} \left(\mathsf{P}^t(x \to y) - \mu(y)\right) \cdot \psi(y) \right| \\ &\le 2 \cdot \max_{y \in \Omega} |\psi(y)| \cdot \left\|\delta_x \mathsf{P}^t - \mu\right\|_{\mathsf{TV}}. \end{aligned}$$

Choosing x maximizing $|\psi(x)|$, it follows that

$$\lambda_*^t \le 2 \cdot \max_{x \in \Omega} \left\|\delta_x \mathsf{P}^t - \mu\right\|_{\mathsf{TV}}$$

Applying the definition of mixing time, we obtain $\lambda_*^{\mathrm{T}_{\mathsf{mix}}(\epsilon)} \le 2\epsilon$, from which it follows that

$$\mathrm{T}_{\mathsf{mix}}(\epsilon) \cdot \left(\frac{1}{\lambda_*} - 1\right) \ge \mathrm{T}_{\mathsf{mix}}(\epsilon) \cdot \log \frac{1}{\lambda_*} \ge \log \frac{1}{2\epsilon}.$$

Rearranging and using $\gamma_* = 1 - \lambda_*$ yields the claim. □

B Proofs of Technical Lemmas

Proof of Lemma 1.4.

$$\begin{aligned} \|\mu - \nu\|_{\mathsf{TV}} &= \frac{1}{2} \sum_{x \in \Omega} \mu(x) \cdot \left| \frac{\nu(x)}{\mu(x)} - 1 \right| \\ &\le \frac{1}{2} \sqrt{\sum_{x \in \Omega} \mu(x) \cdot \left| \frac{\nu(x)}{\mu(x)} - 1 \right|^2} \cdot \underbrace{\sqrt{\sum_{x \in \Omega} \mu(x)}}_{=1} && \text{(Cauchy–Schwarz)} \\ &= \frac{1}{2} \sqrt{\mathrm{Var}_\mu\left(\frac{d\nu}{d\mu}\right)}. \end{aligned}$$

□

Proof of Fact 1.1. Reversibility of P w.r.t. μ is exactly self-adjointness w.r.t. the standard basis functions $\{\delta_x : x \in \Omega\}$ of $\mathbb{R}^\Omega$, where δ_x places unit mass at x and is zero everywhere else. Hence, full self-adjointness implies reversibility. For the other direction, observe that self-adjointness w.r.t. a basis of $\mathbb{R}^\Omega$ implies self-adjointness for all functions just by linearity. □