CS294 Partition Functions **Fall 2023**

Lecture Note 10

Instructor: Alistair Sinclair

Disclaimer: *These notes have not been subjected to the usual scrutiny accorded to formal publications. They may be distributed outside this class only with the permission of the Instructor.*

10.1 Functional analysis for Markov chains

Continuing our study of MCMC as a method for designing approximate counting algorithms, we turn now to a completely different approach to mixing times based on functional analysis, i.e., viewing the Markov chain as an operator on real-valued functions on its state space. Our starting point is the so-called *Poincaré constant* that quantifies the rate of convergence in terms of the decrease in the variance of any function in one step of the Markov chain; this will turn out to be a generalization of the classical spectral theory for reversible chains. We will then go on to develop combinatorial tools based on multicommodity flows to bound the Poincaré constant, and hence the mixing time. These tools will lead to several significant applications, notably approximation algorithms for the permanent (bipartite perfect matchings) and the ferromagnetic Ising model. We'll finish up with a discussion of log-Sobolev inequalities, in which convergence is measured using relative entropy rather than variance. These tools are harder to apply but sometimes give stronger results; we will see a recent important application to counting matroid bases.

10.2 A Poincaré inequality

We begin with some definitions. Let P be the transition matrix of an ergodic Markov chain with finite state space Ω, and π its stationary distribution.

Definition 10.1. *For any real-valued function $\varphi\colon \Omega \to \mathbb{R}$, define the* expectation $\mathrm{E}_\pi[\varphi] := \sum_{x\in\Omega} \pi(x)\varphi(x)$ *and the* variance $\mathrm{Var}_\pi[\varphi] := \sum_{x\in\Omega} \pi(x) \cdot (\varphi(x) - \mathrm{E}_\pi[\varphi])^2$.

We will view the transition matrix P as an operator on functions $\varphi\colon \Omega \to \mathbb{R}$, viz.

$$P\varphi(x) := \sum_{y\in\Omega} P(x,y)\varphi(y).$$

Note that $P\varphi(x)$ is the expectation of φ under the one-step distribution of the chain starting at x (the "one-step averaging" of φ). Similarly, $P^t\varphi(x) = \sum_y P^t(x,y)\varphi(y)$ is the t-step averaging of φ. Since π is stationary, $\mathrm{E}_\pi[P^t\varphi] = \mathrm{E}_\pi[\varphi]$ for all t. Moreover, by ergodicity of P, as $t \to \infty$, $P^t\varphi$ converges to the constant function $\mathrm{E}_\pi[\varphi]$. Thus $\mathrm{Var}_\pi[P^t\varphi] \to 0$ as $t \to \infty$.

We will measure the rate of convergence of the Markov chain by quantifying the rate of convergence of $\mathrm{Var}_\pi[P^t\varphi]$ to zero. The key quantity that will arise in our bound is the so-called "Dirichlet form" $\mathcal{E}_P(\varphi,\varphi)$. We introduce the following more general definition:

Definition 10.2. *For functions $\varphi, \psi\colon \Omega \to \mathbb{R}$, the* Dirichlet form *of P is*

$$\mathcal{E}_P(\varphi,\psi) := \langle \varphi, L\psi\rangle_\pi, \tag{10.1}$$

where $L := I - P$ is the Laplacian *of P and the inner product $\langle\cdot,\cdot\rangle_\pi$ is defined as $\langle f,g\rangle_\pi := \sum_x \pi(x)f(x)g(x)$. (Note that $\langle f,g\rangle_\pi$ can also be written as $\mathrm{E}_\pi[fg]$.)*

Of particular importance to us is the symmetric Dirichlet form $\mathcal{E}_P(\varphi,\varphi)$, which can be rewritten as follows:

$$\begin{aligned}
\mathcal{E}_P(\varphi,\varphi) &= \langle\varphi, L\varphi\rangle_\pi \\
&= \sum_{xy}\pi(x)\varphi(x)(I(x,y)-P(x,y))\varphi(y) \\
&= \sum_{x}\pi(x)\varphi(x)^2 - \sum_{xy}\pi(x)P(x,y)\varphi(x)\varphi(y) \\
&= \frac{1}{2}\sum_{xy}\pi(x)P(x,y)(\varphi(x)^2+\varphi(y)^2) - \sum_{xy}\pi(x)P(x,y)\varphi(x)\varphi(y) \\
&= \frac{1}{2}\sum_{xy}\pi(x)P(x,y)(\varphi(x)-\varphi(y))^2. \qquad (10.2)
\end{aligned}$$

To compare this to the variance itself, note that

$$\begin{aligned}
\mathrm{Var}_\pi[\varphi] &:= \sum_x \pi(x)\cdot(\varphi(x)-\mathrm{E}_\pi[\varphi])^2 \\
&= \sum_x \pi(x)\varphi(x)^2 - \mathrm{E}_\pi[\varphi]^2 \\
&= \frac{1}{2}\sum_{xy}\pi(x)\pi(y)(\varphi(x)^2+\varphi(y)^2) - \sum_x\pi(x)\varphi(x)\sum_y\pi(y)\varphi(y) \\
&= \frac{1}{2}\sum_{xy}\pi(x)\pi(y)\left(\varphi(x)-\varphi(y)\right)^2 \ . \qquad (10.3)
\end{aligned}$$

Now we can see that expressions (10.2) and (10.3) take the same form, except that in (10.2) the sum is taken only over neighbors in the Markov chain. Thus we can think of $\mathcal{E}_P$ as a "local variance" along the transitions of the chain.

In our context, a *Poincaré inequality* bounds the ratio of the local to the global variance for any non-constant function φ. This leads to the following definition.

Definition 10.3. *The* Poincaré constant *of P is defined by*

$$\alpha := \inf_{\varphi \text{ non-constant}} \frac{\mathcal{E}_P(\varphi,\varphi)}{\mathrm{Var}_\pi[\varphi]} \ .$$

The following theorem says that the Poincaré constant provides an upper bound on the mixing time; the intuition for this is that the ratio of local to global variance measures the rate at which the function φ is "averaged" at each step of the Markov chain. Recall that we call P *lazy* if it has a self-loop probability of at least 1/2 at every state.

Theorem 10.4. *For any lazy ergodic P and any initial state $x \in \Omega$,*

$$\tau_x(\varepsilon) \le \frac{1}{\alpha}\left(2\ln\varepsilon^{-1} + \ln(4\pi(x))^{-1}\right) \ .$$

Proof. This is a version of a classical result; we follow the proofs of Jerrum [Jer03] and Mihail [Mih89].

The following lemma is the main content of the proof.

Lemma 10.5. *For any* $\varphi : \Omega \to \mathbb{R}$

$$\mathrm{Var}_\pi[P\varphi] \le \mathrm{Var}_\pi[\varphi] - \mathcal{E}_P(\varphi,\varphi)\ .$$

We defer the proof of the lemma for a moment. For now, we state an immediate corollary that establishes a contraction $1-\alpha$ for the variance in each step of the chain.

Corollary 10.6. *For any non-constant* $\varphi : \Omega \to \mathbb{R}$*, we have* $\mathrm{Var}_\pi[P^t\varphi] \le (1-\alpha)^t\,\mathrm{Var}_\pi[\varphi]$.

Proceeding with the proof of Theorem 10.4, we need to relate $\mathrm{Var}_\pi[P^t\varphi]$ to the variation distance. A standard application of Cauchy-Schwarz **[exercise]** yields

$$\|p_x^{(t)} - \pi\|_{\mathrm{TV}}^2 \le \frac{1}{4}\mathrm{Var}_\pi\Big[\frac{p_x^{(t)}}{\pi}\Big]. \tag{10.4}$$

Now define $\varphi := \frac{p_x^{(0)}}{\pi}$, where $p_x^{(0)}$ is the initial distribution concentrated at state x. Unfortunately $P^t\varphi$ is not equal to the quantity $\frac{p_x^{(t)}}{\pi}$ on the rhs of (10.4). However, we can relate them by introducing the *time reversal* P^* of P, defined by $P^*(x,y) = \frac{\pi(y)}{\pi(x)}P(y,x)$.

Exercise: Verify the following simple properties: (i) P^* is also ergodic with the same stationary distribution π; (ii) $(P^*)^* = P$; (iii) $P^* = P$ iff P is reversible; (iv) $\mathcal{E}_{P^*}(\psi,\psi) = \mathcal{E}_P(\psi,\psi)$ for any ψ, so the Poincaré constants of P and P^* are the same.

Now note that $P^{*t}\varphi = P^{*t}\frac{p_x^{(0)}}{\pi} = \frac{p_x^{(0)}P^t}{\pi} = \frac{p_x^{(t)}}{\pi}$. [**Exercise:** verify this!] Hence

$$4\|p_x^{(t)} - \pi\|_{\mathrm{TV}}^2 \le \mathrm{Var}_\pi\Big[\frac{p_x^{(t)}}{\pi}\Big] = \mathrm{Var}_\pi[P^{*t}\varphi] \le (1-\alpha)^t\,\mathrm{Var}_\pi[\varphi],$$

where in the last step we used Corollary 10.6 and the fact that the Poincaré constant of P^* is also α. Finally, noting that $\mathrm{Var}_\pi[\varphi] = \frac{1}{\pi(x)} - 1 \le \frac{1}{\pi(x)}$, we see that setting $t = \frac{1}{\alpha}(\ln\varepsilon^{-1} + \frac{1}{2}\ln(4\pi(x))^{-1})$ ensures that $\|p_x^{(t)} - \pi\|_{\mathrm{TV}} \le \varepsilon$, as required. □

Proof of Lemma 10.5. Because P is lazy we have $P = \frac{1}{2}(I + \widehat{P})$, where $\widehat{P}$ is stochastic and has the same stationary distribution π as P. Now we may write

$$\begin{aligned}[P\varphi](x) &= \sum_y P(x,y)\varphi(y)\\ &= \frac{1}{2}\varphi(x) + \frac{1}{2}\sum_y \widehat{P}(x,y)\varphi(y)\\ &= \frac{1}{2}\sum_y \widehat{P}(x,y)\left(\varphi(x)+\varphi(y)\right)\ .\end{aligned}$$

Assume w.l.o.g. that $\mathrm{E}_\pi[\varphi] = 0$ (shifting φ by a constant value does not affect any of the quantities we are interested in, all of which are variances). Then

$$\begin{aligned}\mathrm{Var}_\pi[P\varphi] &= \sum_x \pi(x)\left([P\varphi](x)\right)^2\\ &= \frac{1}{4}\sum_x \pi(x)\left(\sum_y \widehat{P}(x,y)\left(\varphi(x)+\varphi(y)\right)\right)^2\\ &\le \frac{1}{4}\sum_{xy} \pi(x)\widehat{P}(x,y)\left(\varphi(x)+\varphi(y)\right)^2\ ,\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality (or the fact that the square of an expectation is bounded by the expectation of the square). Note that this is the only inequality in this proof. Moreover, we can rewrite the variance yet again as

$$\begin{aligned}\mathrm{Var}_\pi[\varphi] &= \frac{1}{2}\sum_x \pi(x)\varphi(x)^2 + \frac{1}{2}\sum_y \pi(y)\varphi(y)^2 \\ &= \frac{1}{2}\sum_{xy} \pi(x)\varphi(x)^2\widehat{P}(x,y) + \frac{1}{2}\sum_{xy} \pi(x)\widehat{P}(x,y)\varphi(y)^2 \\ &= \frac{1}{2}\sum_{xy} \pi(x)\widehat{P}(x,y)\left(\varphi(x)^2+\varphi(y)^2\right) \ .\end{aligned}$$

Taking differences,

$$\mathrm{Var}_\pi[\varphi] - \mathrm{Var}_\pi[P\varphi] \quad \geq \quad \frac{1}{4}\sum_{xy} \pi(x)\widehat{P}(x,y)\left(\varphi(x)-\varphi(y)\right)^2 \ .$$

Observe that all entries in $\widehat{P}$ are twice as large as the entries in its lazy version P, except for the diagonal elements. Diagonal elements can be ignored, however, as they contribute a value of 0 to the above sum. Hence, the right-hand side is equal to

$$\frac{1}{2}\sum_{xy} \pi(x)P(x,y)\left(\varphi(x)-\varphi(y)\right)^2 = \mathcal{E}_P(\varphi,\varphi) \ .$$

This completes the proof of Lemma 10.5 and the proof of Theorem 10.4. □

Remark: The above proof assumes that the Markov chain is lazy, which as we have seen is not in practice a restriction in algorithmic applications as we can always insert a self-loop probability. The role of this assumption is to ensure that there are no periodicity issues. As we shall see later, it can be avoided by passing to a *continuous time* version of the chain, in which periodicity can never arise.

10.3 Connection to eigenvalues

If the Markov chain P is *reversible* (as is often the case in applications), then we can take an alternative approach to proving Theorem 10.4 based on classical spectral graph theory. Recall that P is reversible iff it satisfies the detailed balance condition

$$\pi(x)P(x,y) = \pi(y)P(y,x) \quad \forall x,y \in \Omega,$$

and in that case π is also the unique stationary distribution of the chain.

First, recall from a previous lecture that if P is reversible w.r.t. π then the matrix $S = DPD^{-1}$, where $D = \mathrm{diag}(\sqrt{\pi(x)})$, is non-negative and symmetric (but not necessarily stochastic). This implies that the eigenvalues of P are the same as those of S and hence are all real, and that the eigenvectors of P form a basis for $\mathbb{R}^N$, where $N = |\Omega|$.

Thus by the Perron-Frobenius theorem the spectrum of P takes the form

$$1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_N > -1.$$

Now as we observed in our sketch proof of the Fundamental Theorem for reversible chains in an earlier lecture, we may express the initial distribution $p^{(0)}$ at time t as a linear combination of eigenvectors $e_1, e_2, \ldots, e_N$

of P, namely: $p^{(0)} = \sum_i \alpha_i e_i$, where $e_1 = \pi$ and $\alpha_1 = 1$. But then $p^{(t)} = p^{(0)} P^t = \sum_i \alpha_i \lambda_i^t e_i$. Since $|\lambda_i| < 1$ for all $i > 1$, this implies that $p^{(t)} \to \alpha_1 e_1 = \pi$, and the slowest rate of decay of the other components is determined by $\max_{i \geq 2} |1 - \lambda_i|$. If in addition P is lazy, then all eigenvectors are non-negative (i.e., $\lambda_N \geq 0$), so $\max_{i \geq 2} |1 - \lambda_i| = 1 - \lambda_2$. Thus the rate at which $x^{(t)}$ approaches π can be bounded in terms of the *spectral gap* $1 - \lambda_2$. In fact one can prove via standard linear algebra:

Claim 10.7. *For any ergodic, reversible, lazy Markov chain P, and any initial state x,*

$$\Delta_x(t) \leq \frac{\lambda_2^t}{2\sqrt{\pi(x)}},$$

and hence

$$\tau_x(\varepsilon) \leq \frac{1}{1 - \lambda_2} \left(\ln \varepsilon^{-1} + \tfrac{1}{2} \ln(4\pi(x))^{-1} \right). \tag{10.5}$$

Proof. Left as a (slighly involved) **exercise**. [Hint: Do the proof first for the case that P is symmetric, then generalize to reversible P.] □

Note that the overhead in approximating the overall rate of decay in variation distance just by the second eigenvalue is captured by the factor $\ln \pi(x)^{-1}$, as in Theorem 10.4.

Now to complete the connection, we observe that in the reversible case the spectral gap coincides exactly with the Poincaré constant (see Definition 10.3)!

Claim 10.8. *For P ergodic and reversible, the spectral gap is given by*

$$1 - \lambda_2 = \inf_{\varphi \text{ non-constant}} \frac{\mathcal{E}_P(\varphi, \varphi)}{\mathrm{Var}_\pi[\varphi]} \ .$$

Proof. Note that the Laplacian $L := I - P$ has eigenvalues $\mu_i = 1 - \lambda_i$, with $0 = \mu_1 < \mu_2 \leq \cdots \leq \mu_N < 2$, and the same eigenvectors as P. So the spectral gap of P is just μ_2.

Suppose first that P is symmetric (so that it is reversible w.r.t. the uniform distribution $\pi = \mathbf{1}$); then of course L is also symmetric. By the classical variational characterization of eigenvalues of symmetric matrices, the principal eigenvalue of L minimizes the *Rayleigh quotient*, i.e.,

$$\mu_1 = \inf_{\varphi \neq 0} \frac{\langle \varphi, L\varphi \rangle}{\langle \varphi, \varphi \rangle},$$

where $\langle \varphi, \psi \rangle := \sum_{x \in \Omega} \varphi(x)\psi(x)$. A function φ that achieves the infimum is the principal eigenvector $\mathbf{1}$. This can be extended to higher eigenvalues in the obvious way. In particular,

$$\mu_2 = \inf_{\varphi \perp \mathbf{1}} \frac{\langle \varphi, L\varphi \rangle}{\langle \varphi, \varphi \rangle} = \inf_{\varphi \not\parallel \mathbf{1}} \frac{\langle \varphi, L\varphi \rangle}{\langle \varphi, \varphi \rangle - \langle \varphi, \mathbf{1} \rangle^2}. \tag{10.6}$$

To extend this to the case where P is reversible, we apply (10.6) to the symmetric matrix $I - DPD^{-1}$ as above and obtain, after some algebra **[exercise!]**,

$$\mu_2 = \inf_{\varphi \perp \mathbf{1}} \frac{\langle \varphi, L\varphi \rangle_\pi}{\langle \varphi, \varphi \rangle_\pi} = \inf_{\varphi \not\parallel \mathbf{1}} \frac{\langle \varphi, L\varphi \rangle_\pi}{\langle \varphi, \varphi \rangle_\pi - \langle \varphi, \mathbf{1} \rangle_\pi^2}, \tag{10.7}$$

where as before $\langle \varphi, \psi \rangle_\pi := \sum_{x \in \Omega} \varphi(x)\psi(x)\pi(x)$. Now notice that the numerator in (10.7) is just $\mathcal{E}_P(\varphi, \varphi)$, while the denominator is $\mathrm{Var}_\pi[\varphi]$. This completes the proof. □

Plugging Claim 10.8 into equation (10.5), we recover (up to a factor of 2) exactly the same bound on the mixing time as in Theorem 10.4.

Remark: For an extension of the spectral approach to non-reversible chains, see [Fil91].

References

[Fil91] J.A. Fill. Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *Annals of Applied Probability*, 1:62–87, 1991.

[Jer03] M. Jerrum. *Counting, Sampling and Integrating: Algorithms and Complexity.* Birkhäuser Lectures in Mathematics, 2003.

[Mih89] M. Mihail. Conductance and convergence of Markov chains: A combinatorial treatment of expanders. *Proceedings of the 30th ACM STOC*, pages 526–531, 1989.