CS294 Partition Functions **Fall 2023**

Lecture Note 14

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.*

14.1 Entropy and the log-Sobolev constant

So far we've focused on variance reduction, as embodied in the Poincaré constant (Theorem 10.4), to bound the mixing time. Potentially sharper results can be obtained by considering instead the rate of decrease of relative entropy (or Kullback-Leibler divergence), though the resulting technology is usually more difficult (if not impossible) to apply. In this lecture we will examine entropy-based estimates of the mixing time.

To develop the necessary technology, it is convenient to work with a continuous-time version of the Markov chain. In this model, rather than making one transition per time step, the chain is equipped with a mean-1 Poisson clock; each time the clock rings, a transition is made according to the usual transition matrix P. More precisely, the times between successive transitions are iid exponential with mean 1, so that the number of transitions per unit time follows a Poisson Po(1) distribution.

We will use H_t (for "heat kernel") to denote the probability distribution of the continuous time Markov chain after t steps. Thus $H_t(x,y) := \Pr[X_t = y | X_0 = x]$. From the above description, we have

$$H_t(x,y) = \sum_{k=0}^{\infty} \frac{e^{-t}t^k}{k!} P^k(x,y), \tag{14.1}$$

which has the more compact matrix representation

$$H_t = \exp(t(P-I)) = \exp(-tL), \tag{14.2}$$

where $L = I - P$ is the Laplacian of P. (The notation $\exp(A)$ for a matrix A should be interpreted as the convergent sum $I + A + \frac{1}{2!}A^2 + \cdots$.) **Exercise:** Check that the matrix power series in (14.1) and (14.2) are equal.

Switching between discrete and continuous time is largely a bookkeeping exercise, but continuous time chains have two advantages: first, periodicity is not an issue for convergence; and second, difference equations are replaced by derivatives, which are sometimes simpler (though less intuitive). If we wish to simulate a continuous time chain, we specify the transition times by sampling a sequence of independent exponential random variables and make each transition as usual according to P.

The following proposition confirms that the mixing times of the discrete and continuous time versions of the Markov chain are essentially equivalent.

Proposition 14.1. *Let P be an irreducible (not necessarily aperiodic) Markov chain matrix with stationary distribution π, and let $\tilde{P} = \frac{1}{2}(I+P)$ be the lazy version of P. For a given initial state x, write $\tilde{p}_x^{(k)}$ for the k-step distribution of $\tilde{P}$ and $h_x^{(t)}$ for the time-t distribution of the heat kernel of P. Then for any fixed $\varepsilon > 0$ we have:*

(i) For sufficiently large k, if $\|\tilde{p}_x^{(k)} - \pi\|_{\mathrm{TV}} \le \varepsilon$ then $\|h_x^{(k)} - \pi\|_{\mathrm{TV}} \le 2\varepsilon$.

(ii) For sufficiently large t, if $\|h_x^{(t)} - \pi\|_{\mathrm{TV}} \le \varepsilon$ then $\|\tilde{p}_x^{(\lceil 4t\rceil)} - \pi\|_{\mathrm{TV}} \le 2\varepsilon$.

Proof (rough sketch). We briefly sketch the ideas; you should be able to fill in the details if you are interested, else see [LPW09, Chapter 20] for a full proof. To see (i), note that if $\tilde{P}$ is ε-close to stationarity after k steps then the corresponding heat kernel (of the lazy chain) $\tilde{H}$ will be 2ε-close to stationarity after $2k$ steps since the probability that $\tilde{H}$ performs less than k transitions is $\Pr[\mathrm{Po}(2k) < k] \to 0$ as $k \to \infty$. Then (i) follows from the easy fact that $H_t = \tilde{H}_{2t}$. To see (ii), suppose H is ε-close to stationarity after t steps; then the same is true if we add on a further t steps of P. The number of transitions performed is then $\mathrm{Po}(t) + t$. On the other hand, the number of transitions (of P) performed in $4t$ steps of the lazy chain $\tilde{P}$ is $\mathrm{Bin}(4t, \frac{1}{2})$. Now one can check that $\|\mathrm{Po}(t) + t - \mathrm{Bin}(4t, \frac{1}{2})\|_{\mathrm{TV}} \to 0$ as $t \to \infty$, so by the coupling lemma we can couple these two processes to achieve an overall variation distance from π of at most 2ε. □

Note that continuous time dispenses with the need for an aperiodicity (or laziness) assumption on P.

To get a feel for continous time, let's first prove a continuous time version of our variance reduction estimate, Lemma 10.5 from Lecture 10, which says that $\mathrm{Var}_\pi[P\varphi] - \mathrm{Var}_\pi[\varphi] \le -\mathcal{E}_P(\varphi,\varphi)$ for any function $\varphi : \Omega \to \mathbb{R}$.

Lemma 14.2. *Let P be an irreducible Markov chain with stationary distribution π, and let H_t be the corresponding heat kernel. For any function $\varphi : \Omega \to \mathbb{R}$, we have $\frac{d}{dt}\mathrm{Var}_\pi[H_t\varphi] = -2\mathcal{E}_P(H_t\varphi, H_t\varphi)$.*

Proof. Write $\varphi_t := H_t\varphi$. Note first that, from the definition $H_t = \exp(-tL)$, we have $\frac{d}{dt}\varphi_t = -L\varphi_t$. Hence

$$\begin{aligned}
\frac{d}{dt}\mathrm{Var}_\pi[\varphi_t] &= \frac{d}{dt}\big(\mathrm{E}_\pi[\varphi_t^2] - \mathrm{E}_\pi[\varphi_t]^2\big) \\
&= \frac{d}{dt}\Big(\sum_x \pi(x)\varphi_t(x)^2\Big) \\
&= 2\sum_x \pi(x)\varphi_t(x)\frac{d}{dt}\varphi_t(x) \\
&= -2\sum_x \pi(x)\varphi_t(x)[L\varphi_t](x) \\
&= -2\mathcal{E}_P(\varphi_t,\varphi_t),
\end{aligned}$$

where in the last step we used the definition of the Dirichlet form. In the second line, we used the fact that the expectation $\mathrm{E}_\pi[\varphi_t] = \mathrm{E}_\pi[\varphi]$ is constant. □

From the definition of the Poincaré constant, this immediately gives us the following analog of Corollary 10.6:

Corollary 14.3. *For any non-constant $\varphi : \Omega \to \mathbb{R}$, we have $\frac{d}{dt}\mathrm{Var}_\pi[H_t\varphi] \le -2\alpha\mathrm{Var}_\pi[H_t\varphi]$. Hence $\mathrm{Var}_\pi[H_t\varphi] \le \exp(-2\alpha t)\mathrm{Var}_\pi[\varphi]$.*

By exactly the same argument as earlier, we therefore get a continuous time version of Theorem 10.4 (with an improvement of a factor of 2):

Theorem 14.4. *For any irreducible P and any initial state $x \in \Omega$, the mixing time of the associated continuous time Markov chain satisfies*

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

In place of variance, let's now investigate the rate of convergence under a different notion of distance. We'll use the following entropy-like quantity:

$$\mathrm{Ent}_\pi(\varphi) := \sum_x \pi(x)\varphi(x)\log\varphi(x) - \mathrm{E}_\pi[\varphi]\log\mathrm{E}_\pi[\varphi] = \sum_x \pi(x)\varphi(x)\log\left(\frac{\varphi(x)}{\mathrm{E}_\pi[\varphi]}\right). \tag{14.3}$$

If we set $\varphi = \frac{\mu}{\pi}$, where μ is a probability distribution over Ω, then $\mathrm{E}_\pi[\varphi] = 1$ and

$$\mathrm{Ent}_\pi(\varphi) = \sum_x \mu(x)\log\frac{\mu(x)}{\pi(x)} = D(\mu\|\pi),$$

where $D(\mu\|\pi)$ is the *Kullback-Leibler divergence* or *relative entropy*, a widely used notion of distance between probability distributions. (Note, however, that $D(\mu\|\pi)$ is not a metric; e.g., it is not symmetric.)

Our goal is to bound the rate of convergence of $D(p_x^{(t)}\|\pi) = \mathrm{Ent}_\pi[\frac{p_x^{(t)}}{\pi}]$ to zero, just as we earlier bounded the rate of convergence of $\mathrm{Var}_\pi[\frac{p_x^{(t)}}{\pi}]$. It turns out that this rate is also bounded by a Dirichlet form, namely the asymmetric form $\mathcal{E}_P(\varphi, \log\varphi)$. Accordingly, following [BT06], we make the following definition.

Definition 14.5. *The* (modified) log-Sobolev constant *of P is defined by*

$$\rho := \inf_{\varphi\ge 0,\,\mathrm{Ent}_\pi[\varphi]\neq 0} \frac{\mathcal{E}_P(\varphi,\log\varphi)}{\mathrm{Ent}_\pi[\varphi]}.$$

The analog of Theorem 14.4 is the following:

Theorem 14.6. *For any irreducible P and any initial state $x \in \Omega$, the mixing time of the associated continuous time Markov chain satisfies*

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

This theorem looks very similar to Theorem 14.4, with the Poincaré constant α replaced by $\rho/2$. However, there is a crucial difference: here we have replaced $\ln\pi(x)^{-1}$ by $\ln\ln\pi(x)^{-1}$. Recall that in typical applications $\pi(x)$ is of the order $\exp(-cn)$, where n is the natural measure of input size and c is a constant. Thus the dependence of the mixing time on the initial state in Theorem 14.4 is typically $O(n)$, while in Theorem 14.6 it is $O(\log n)$, a substantial reduction. In cases where ρ and α are of similar order, this elimination of a factor of $O(n)$ sometimes allows one to obtain an optimal bound on the mixing time. (The overhead $\ln\pi(x)^{-1}$ in the Poincaré bound is almost always pessimistic.) However, it is often very difficult to obtain a good estimate of ρ.

In parallel with Lemma 14.2, the key to the proof of Theorem 14.6 is the following.

Lemma 14.7. *For any positive function $\varphi : \Omega \to \mathbb{R}^+$, define $\varphi_t := \frac{\varphi H_t}{\pi}$. Then $\frac{d}{dt}\mathrm{Ent}_\pi[\varphi_t] = -\mathcal{E}_P(\varphi_t, \log\varphi_t)$.*

Proof. To undo the normalization by π, note that $\varphi_t = H_t^*(\frac{\varphi}{\pi})$, where H_t^* is the heat kernel of the reversibilization of P. [**Exercise:** check this, as in the proof of Theorem 10.4. Recall that P^* is defined by $\pi(x)P(x,y) = \pi(y)P^*(y,x)\forall x,y.$] So, as in the proof of Lemma 14.2, we have $\frac{d}{dt}\varphi_t = \frac{d}{dt}H_t^*(\frac{\varphi}{\pi}) = -L^*\varphi_t$.

Hence

$$\begin{aligned}
\frac{d}{dt}\mathrm{Ent}_\pi[\varphi_t] &= \frac{d}{dt}\big(\mathrm{E}_\pi[\varphi_t \log \varphi_t - \mathrm{E}_\pi[\varphi_t]\log \mathrm{E}_\pi[\varphi_t]]\big)\\
&= \frac{d}{dt}\big(\sum_x \pi(x)\varphi_t(x)\log\varphi_t(x) - \mathrm{E}_\pi[\varphi_t]\log \mathrm{E}_\pi[\varphi_t]]\big)\\
&= \sum_x \pi(x)(1+\log\varphi_t(x))\frac{d}{dt}\varphi_t(x)\\
&= -\sum_x \pi(x)\log\varphi_t(x)(L^*\varphi_t)(x)\\
&= -\sum_x \pi(x)[L\log\varphi_t](x)\varphi_t(x)\\
&= -\mathcal{E}_P(\varphi_t, \log\varphi_t).
\end{aligned}$$

In the third line here we used the fact that $\mathrm{E}_\pi[\varphi_t]$ is constant; in the fourth line the fact that $\mathrm{E}_\pi[\frac{d}{dt}\varphi_t(x)] = 0$ (essentially conservation of mass); and in the fifth line the fact that $\langle\psi, L^*\eta\rangle_\pi = \langle\eta, L\psi\rangle_\pi$ for all ψ, η (i.e., L^* is the adjoint of L w.r.t. the inner product $\langle\cdot,\cdot\rangle_\pi$). [**Exercise:** Check these facts!] □

Corollary 14.8. *For any positive* $\varphi : \Omega \to \mathbb{R}^+$ *with* $\mathrm{Ent}_\pi[\varphi] \neq 0$*, we have* $\mathrm{Ent}_\pi[\frac{\varphi H_t}{\pi}] \le \exp(-\rho t)\mathrm{Ent}_\pi[\frac{\varphi}{\pi}]$.

Now it is just a short step to the proof of Theorem 14.6.

Proof of Theorem 14.6. Setting $\varphi = p_x^{(0)}$ (the initial distribution concentrated at state x), and using our earlier observation that $\mathrm{Ent}_\pi[\frac{p_x^{(t)}}{\pi}] = D(p_x^{(t)}\|\pi)$, Corollary 14.8 tells us that[1]

$$D(p_x^{(t)}\|\pi) \le \exp(-\rho t)D(p_x^{(0)}\|\pi) = \exp(-\rho t)\log\pi(x)^{-1}. \tag{14.4}$$

Note the initial value $\log\pi(x)^{-1}$ for relative entropy here, in contrast to the initial value $\pi(x)^{-1}$ for variance; this is the source of the improved dependence on the initial distribution.

To finish, we may appeal to Pinsker's inequality, which says that for any distribution p, $2\|p-\pi\|_{\mathrm{TV}}^2 \le D(p\|\pi)$. Combining this with (14.4) ensures that $\|p_x^{(t)} - \pi\|_{\mathrm{TV}} \le \varepsilon$ for all $t \ge \frac{1}{\rho}(\ln\frac{1}{2\varepsilon^2} + \ln\ln\pi(x)^{-1})$, as claimed. □

Notes:

1. By definition, the modified log-Sobolev constant ρ is the largest quantity that satisfies the inequality

   $$\mathcal{E}_P(\varphi, \ln\varphi) \ge \rho\,\mathrm{Ent}_\pi(\varphi)$$

   for all non-negative functions $\varphi : \Omega \to \mathbb{R}^+$. By contrast, the traditional log-Sobolev inequality takes the form

   $$\mathcal{E}_P(\sqrt{\varphi}, \sqrt{\varphi}) \ge \rho_0\,\mathrm{Ent}_\pi(\varphi) \tag{14.5}$$

   over the same class of functions; ρ_0 is then just called the *log-Sobolev constant.* This latter inequality was largely pioneered by Gross [Gro76], in the context of the theory of *hypercontractivity*, which deals with inequalities of the form $\|f(x)\|_p \le \beta\|x\|_q$ for some constant β, for all functions f. When suitably formulated, the logarithmic Sobolev inequality (14.5) turns out to be equivalent to hypercontractivity, which explains its importance. The theory of log-Sobolev inequalities in the context of finite Markov

[1] Strictly speaking this step needs further justification because $p_x^{(t)}$ is not strictly positive at $t = 0$. However, $p_x^{(t)}$ is strictly positive for all $t > 0$, and the step can be justified by continuity. We omit the details.

chains was worked out by Diaconis and Saloff-Coste [DSC96], including in particular the mixing time bound of Theorem 14.6 with ρ replaced by ρ_0. The modified log-Sobolev constant we are using here was implicit in [DSC96] and was explicitly studied by Bobkov and Tetali [BT06]. For the purposes of mixing time, a bound on ρ is essentially equivalent to a bound on ρ_0, so it is usually preferable to work with the simpler modified log-Sobolev constant in this context. More significantly, there are several applications where ρ can be effectively bounded while the log-Sobolev constant ρ_0 itself can be arbitrarily small. On the other hand, it should be noted that hypercontractivity (guaranteed by ρ_0) is a stronger property than convergence in variation distance (or KL divergence).

2. It can be shown that $\rho_0 \leq 4\rho \leq 2\alpha$ [BT06] (and for our purposes the constants don't matter). Note that relative entropy is bounded above by relative variance, so the generic bounds in Corollaries 14.8 and 14.3 are incomparable: the former controls a smaller distance (relative entropy) via an exponential decay at possibly lower rate. When we translate both to mixing time, as we've seen, in cases where $\rho \approx \alpha$ (and we can approximate ρ well), we do better with log-Sobolev due to the improved dependence on the initial distribution. However, in general α can be much larger than either ρ or ρ_0. Moreover, bounds on log-Sobolev constants are typically much harder to obtain than those on the Poincaré constant, and usually require some symmetry or recursive structure we can exploit. We will see some examples in the next lecture.

References

[BT06] S. Bobkov and P. Tetali. Modified logarithmic Sobolev inequalities in discrete settings. *Journal of Theoretical Probability*, 19:289–336, 2006.

[DSC96] P. Diaconis and L. Saloff-Coste. Logarithmic Sobolev inequalities for finite Markov chains. *Annals of Applied Probability*, 6:695–750, 1996.

[Gro76] L. Gross. Logarithmic Sobolev inequalities for finite Markov chains. *American Journal of Mathematics*, 97:1061–1083, 1976.

[LPW09] D. Levin, Y. Peres, and E. Wilmer. *Markov Chains and Mixing Times.* American Mathematical Society, Providence, RI, 2009.