Lecture 12: Functional Analytic Tools for MCMC

October 19, 2023

In the next few lectures, we return to the study of Markov chain Monte Carlo methods. Our goal will be to *unify* all of the methods we have seen thus far in the following sense: Exponential decay of correlations or the absence of zeros of the partition function both imply local Markov chains like Glauber dynamics mix in *nearly-linear* time. Our aim in this lecture is to build some of the foundational tools required for this endeavor.

1 Mixing Time Bounds via Functional Analysis

Let μ be a probability distribution on some state space Ω, and let P be a Markov chain on Ω which is reversible w.r.t. μ. Our goal is to study the total variation mixing time of P, which controls how efficient it is to use P as a sampler. Previously, we saw two types of tools for bounding $\mathrm{T}_{\mathsf{mix}}(\epsilon)$: (path) coupling, and Poincaré Inequalities/spectral gap. Both of these are instantiations of a much more general strategy.

Theme 1.1. *Show that some other measure of "distance" between probability measures $\mathscr{D}(\cdot \,\|\, \cdot)$ contracts under every application of P. In other words, for some $0 < \alpha < 1$ which is not too small, we have*

$$\mathscr{D}(\nu\mathsf{P} \,\|\, \mu) \leq (1-\alpha) \cdot \mathscr{D}(\nu \,\|\, \mu), \qquad \forall \text{ probability measures } \nu \text{ on } \Omega. \tag{1}$$

Fact 1.2. *If Eq. (1) holds for some $\mathscr{D}(\cdot \,\|\, \cdot)$ such that $\mathscr{D}(\nu \,\|\, \mu) \leq \epsilon$ implies $\|\mu - \nu\|_{\mathsf{TV}} \leq O(\epsilon^c)$ for some $c > 0$, then*

$$\mathrm{T}_{\mathsf{mix}}(\epsilon) \leq O\left(\frac{1}{\alpha} \log\left(\frac{\max_{x\in\Omega} \mathscr{D}(\delta_x \,\|\, \mu)}{\epsilon}\right)\right)$$

We do not require $\mathscr{D}(\cdot \,\|\, \cdot)$ to be symmetric (e.g. KL-divergence) so it is not a metric in a formal sense. For the most part, we just need nonnegativity and that $\mathscr{D}(\nu \,\|\, \mu) = 0$ implies $\mu = \nu$. Depending on what $\mathscr{D}(\cdot \,\|\, \cdot)$ is, Eq. (1) is often called a *Strong Data Processing Inequality* in information theory contexts, since for many natural notions of "distance" $\mathscr{D}(\cdot \,\|\, \cdot)$, we have the standard Data Processing Inequality $\mathscr{D}(\nu\mathsf{P} \,\|\, \mu) \leq \mathscr{D}(\nu \,\|\, \mu)$. Of course, if $\mathscr{D}(\cdot \,\|\, \cdot)$ is total variation distance itself, then we immediately get rapid mixing. Notably, the total variation distance between two distributions is always at most 1 so $\max_{x\in\Omega} \mathscr{D}(\delta_x \,\|\, \mu) \leq 1$ in this case. However, TV-distance isn't very easy to work with in general, and it very well could decay in a highly irregular manner. So, we typically pick a nicer "smoother" $\mathscr{D}(\cdot \,\|\, \cdot)$ such that Eq. (1) holds and we make quantifiable progress in every single step. Here are two examples which we have already seen.

Example 1 (Bubley–Dyer Path Coupling). If we endow Ω with the structure of an undirected graph (Ω, E), then we can take $\mathscr{D}(\nu \,\|\, \mu)$ to be the *Wasserstein distance* (or *transportation distance*)

$$\mathscr{W}_1(\mu, \nu) \overset{\mathsf{def}}{=} \inf_{\xi} \mathbb{E}_{(x,y)\sim\xi}[\mathrm{dist}(x, y)], \tag{2}$$

w.r.t. the shortest path metric $\mathrm{dist}(x, y)$ in (Ω, E), where the infimum is over all couplings ξ of μ, ν. By composing couplings along shortest paths, to show $\mathscr{W}_1(\mu\mathsf{P}, \nu\mathsf{P}) \leq (1-\alpha) \cdot \mathscr{W}_1(\mu, \nu)$, it suffices to prove that for every pair of *neighboring* states $(x, y) \in E \subseteq \binom{\Omega}{2}$, we have $\mathscr{W}_1(\delta_x\mathsf{P}, \delta_y\mathsf{P}) \leq 1-\alpha$. This dramatically simplifies the task of proving mixing time upper bounds. In this context, the number α is sometimes called the *coarse Ricci curvature* (or *Ollivier–Ricci curvature*) of the measure metric space (Ω, E, P) [Oll09]. We also have the trivial bound $\max_{x\in\Omega} \mathscr{W}(\delta_x \,\|\, \mu) \leq \mathrm{diam}(\Omega, E)$. Note that since $\mathrm{dist}(\cdot, \cdot)$ takes values in $\mathbb{N}$, we always have $\mathscr{W}_1(\mu, \nu) \geq \|\mu - \nu\|_{\mathsf{TV}}$. Furthermore, if we took $E = \binom{\Omega}{2}$ so that $\mathrm{dist}(\cdot, \cdot)$ becomes the discrete metric, then $\mathscr{W}_1$ is exactly TV-distance. However, if Ω has product structure for instance, we can do much better by using Hamming distance.

Example 2 (Poincaré Inequality). If we take $\mathscr{D}(\nu \,\|\, \mu)$ to be $\chi^2(\nu \,\|\, \mu) = \mathrm{Var}_\mu\left(\frac{d\nu}{d\mu}\right)$, then contraction Eq. (1) follows from the Poincaré Inequality

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

where recall

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

is the Dirichlet form of P. The best choice of γ is the (absolute) spectral gap of P. Combining this with the comparison $\|\mu - \nu\|_{\mathsf{TV}}^2 \le \frac{1}{4}\chi^2(\nu \,\|\, \mu)$ implies rapid mixing assuming a Poincaré Inequality holds with a good γ. We saw earlier how to bound the Poincaré constant γ using the conductance method and canonical paths/flows.

1.1 (Modified) Logarithmic Sobolev Inequalities

Path coupling (see Example 1) is very useful in practice, and in many settings (e.g. graph colorings and the ferromagnetic Ising model) gives optimal nearly-linear mixing time. However, there are natural rapidly mixing Markov chains on non-contrived state spaces for which no such argument can certify this rapid mixing [KR01]. The spectral gap (see Example 2) in a concrete sense fully "characterizes" the mixing time up to polynomial factors. However, in many settings of interest (e.g. Gibbs distributions), even if one were to obtain the best possible bound on γ, it would still give an extraneous factor of n due to the initial distance $\max_{x\in\Omega} \chi^2(\delta_x \,\|\, \mu) = \frac{1}{\mu_{\min}}$, which is often exponentially large in n. So at the very best, we'd get a suboptimal $O(n^2)$-mixing, without even accounting for possible additional losses in bounding γ.

To remedy this situation, we turn to the KL-*divergence* (or *relative entropy*).

$$\mathscr{D}_{\mathrm{KL}}(\nu \,\|\, \mu) \overset{\mathsf{def}}{=} \sum_{x \in \Omega} \nu(x) \log \frac{\nu(x)}{\mu(x)}. \tag{3}$$

More generally, for a nonnegative function $f : \Omega \to \mathbb{R}_{\ge 0}$, define

$$\mathrm{Ent}_\mu(f) \overset{\mathsf{def}}{=} \mathbb{E}_\mu[f \log f] - \mathbb{E}_\mu[f] \log \mathbb{E}_\mu[f]. \tag{4}$$

One could also consider the Φ-*divergences*/Φ-*entropies* given by $\mathrm{Ent}_\mu^\Phi(f) \overset{\mathsf{def}}{=} \mathbb{E}_{x\sim\mu}[\Phi(f(x))] - \Phi\left(\mathbb{E}_{x\sim\mu}[f(x)]\right)$ for convex Φ; see [Cha04] and references therein. We will not do so here, although we mention that many of the techniques we will see later on also apply to these types of "distances".

Note that $\mathscr{D}_{\mathrm{KL}}(\nu \,\|\, \mu) = \mathrm{Ent}_\mu\left(\frac{d\nu}{d\mu}\right)$. Moreover, $\max_{x\in\Omega} \mathscr{D}_{\mathrm{KL}}(\delta_x \,\|\, \mu) = \log \frac{1}{\mu_{\min}}$, which is an exponential improvement over what we get using $\chi^2(\nu \,\|\, \mu)$. We now study the decay of $\mathscr{D}_{\mathrm{KL}}(\nu \,\|\, \mu)$ w.r.t. P, which is captured by the following functional analytic quantities.

Definition 1 (Standard/Modified Log-Sobolev Inequalities). *Let P be a Markov chain which is reversible w.r.t. a distribution μ on a domain Ω. We say P satisfies a* (standard) log-Sobolev inequality *with constant κ if*

$$\kappa \cdot \mathrm{Ent}_\mu(f) \le \mathcal{E}_{\mathsf{P}}\left(\sqrt{f}, \sqrt{f}\right), \qquad \forall f : \Omega \to \mathbb{R}_{\ge 0}. \tag{5}$$

We say P satisfies a modified log-Sobolev inequality *with constant ϱ if*

$$\varrho \cdot \mathrm{Ent}_\mu(f) \le \mathcal{E}_{\mathsf{P}}(f, \log f), \qquad \forall f : \Omega \to \mathbb{R}_{\ge 0}. \tag{6}$$

We define the standard/modified log-Sobolev constants *$\kappa(\mathsf{P}), \varrho(\mathsf{P})$ of P to be the best possible constants in Eqs. (5) and (6), respectively.*

Remark 1. Unlike its modified counterpart, it was previously observed e.g. in [HS20] that $\kappa(\mathsf{P})$ is sensitive to $\mu_{\min}$. In particular,

$$\kappa(\mathsf{P}) \le \min_{x \in \mathrm{supp}(\mu)} \frac{\mathcal{E}_{\mathsf{P}}\left(\sqrt{\mathbb{I}_x}, \sqrt{\mathbb{I}_x}\right)}{\mathrm{Ent}_\mu\left(\mathbb{I}_x\right)} = \min_{x \in \mathrm{supp}(\mu)} \frac{\mu(x) \cdot (1 - \mathsf{P}(x \to x))}{\mu(x) \log \frac{1}{\mu(x)}} \le \frac{1}{\log \frac{1}{\mu_{\min}}}.$$

While this isn't necessarily an issue for spin systems on bounded-degree graphs, there are many other applications (e.g. determinantal point processes) where we can have $\kappa(\mathsf{P}) \ll \varrho(\mathsf{P})$. We mention here a beautiful recent result of Salez–Tikhomirov–Youssef [STY23] on reverse inequalities between $\varrho(\mathsf{P})$ and $\kappa(\mathsf{P})$.

The standard version was first proposed by Gross [Gro75] in the continuous space, where the two versions are equivalent as observed by [ELL17]; see [Led99; GZ03; MT06] for more comprehensive material on these constants and inequalities. The term "modified" is a bit overloaded, especially in continuous settings, but we use it following Bobkov–Tetali [BT03]. It is well-known that the standard log-Sobolev inequality is equivalent to *hypercontractivity* of the associated *heat semigroup* Eq. (8) [DS96], which is a fundamental tool e.g. in the Fourier analysis of Boolean functions [ODo14]. On the other hand, the modified version tends to be more useful in mixing time applications because $\kappa(\mathsf{P})$ is sensitive to $\mu_{\min}$; see Remark 1. Like the spectral gap, lower bounds on these constants yield upper bounds on the mixing time.

Theorem 1.3 ((Modified) Log-Sobolev Implies Rapid Mixing)**.** *Let P be a reversible ergodic Markov chain with stationary distribution μ on a domain Ω. Then for every $\epsilon > 0$,*

$$\mathrm{T}_{\mathsf{mix}}(\epsilon) \leq \frac{1}{\varrho(\mathsf{P})}\left(\log\log\frac{1}{\mu_{\min}} + \log\frac{1}{2\epsilon^2}\right) \qquad \text{[BT03]}$$

$$\mathrm{T}_{\mathsf{mix}}(\epsilon) \leq \frac{1}{4\kappa(\mathsf{P})}\left(\log\log\frac{1}{\mu_{\min}} + \log\frac{1}{2\epsilon^2}\right) \qquad \text{[DS96]}$$

where recall that $\mu_{\min} = \min_{x\in\Omega:\mu(x)>0}\mu(x)$.

Besides mixing, these constants turn out to also have incredibly useful consequences for concentration of measure phenomena.

Theorem 1.4 ((Modified) Log-Sobolev Implies Concentration; see e.g. [Goe04; Sam05; BLM16])**.** *Let P be a reversible ergodic Markov chain with stationary distribution μ on a domain Ω. Fix an arbitrary function $f : \Omega \to \mathbb{R}$, and define the* maximum one-step variance of f *by*

$$v(f) \overset{\mathsf{def}}{=} \max_{x\in\Omega}\left\{\sum_{y\in\Omega}\mathsf{P}(x\to y)\cdot(f(x)-f(y))^2\right\}. \tag{7}$$

Then for every $t \geq 0$, we have the following sub-Gaussian concentration inequalities

$$\Pr_{x\sim\mu}[f(x) \geq \mathbb{E}_\mu[f] + \epsilon] \leq \exp\left(-\frac{\varrho(\mathsf{P})\epsilon^2}{2v(f)}\right)$$

$$\Pr_{x\sim\mu}[f(x) \geq \mathbb{E}_\mu[f] + \epsilon] \leq \exp\left(-\frac{2\kappa(\mathsf{P})\epsilon^2}{v(f)}\right).$$

Here, $v(f)$ quantifies how Lipschitz f is w.r.t. the underlying graph induced by P on Ω. In particular, if f is 1-Lipschitz in the sense that $|f(x) - f(y)| \leq 1$ for all x, y such that $\mathsf{P}(x \to y) > 0$, then $v(f) \leq 1$.

Finally, we have the following comparison inequalities between $\gamma(\mathsf{P}), \varrho(\mathsf{P}), \kappa(\mathsf{P})$, which says that lower bounding the spectral gap is easier than lower bounding the standard/modified log-Sobolev constants.

Proposition 1.5 ([BT03])**.** *For every reversible Markov chain P, $4\kappa(\mathsf{P}) \leq \varrho(\mathsf{P}) \leq 2\gamma(\mathsf{P})$.*

Remark 2. These constants really can behave very differently (see e.g. Remark 1) even for very simple and natural Markov chains, so it isn't obvious at all that working with $\varrho(\mathsf{P}), \kappa(\mathsf{P})$ would actually result in better mixing times compared to using $\gamma(\mathsf{P})$. However, we will see that in the context of spin systems on bounded-degree graphs, they are often all of the same order (at least in the regime where polynomial-time algorithms exist).

Historically, the standard and modified log-Sobolev constants are notoriously difficult to lower bound, especially in the absence of product structure or special symmetries [DS81; DS87; DS96; Sca97; LY98; DH02; ST10; FOW22]. In the next few lectures, we'll see new techniques for bounding these quantities based on quantitative correlation inequalities, which can then be established via techniques we've already seen. Proofs of Theorem 1.4 and Proposition 1.5 are provided in Section 2 and Appendix A, respectively. We now turn to a proof (sketch) of Theorem 1.3, which will also explain where the Dirichlet forms $\mathcal{E}_{\mathsf{P}}\left(\sqrt{f}, \sqrt{f}\right)$ and $\mathcal{E}_{\mathsf{P}}(f, \log f)$ come from.

1.2 The Heat Semigroup

It will be convenient to evolve P in *continuous* time. This is a standard tool which allows us to do differential calculus. For this part, we follow the presentation in [LPW17]. Let $\{t_k\}_{k=1}^{\infty} \subseteq \mathbb{R}_{\geq 0}$ be a sequence of i.i.d. mean 1 exponential random variables, i.e. $\Pr[t_k \geq t] = \exp(-t)$ for all $t \in \mathbb{R}_{\geq 0}$ and all $k \in \mathbb{N}$. Think of these as time *increments*. We now define a continuous-time stochastic process $t \mapsto Y_t \in \Omega$, which depends on $\{t_k\}_{k=1}^{\infty}$, as follows:

- We sample Y_0 according to some initial distribution ν.
- At *transition time* $T_k = \sum_{i=1}^{k} t_i$, we take a single discrete step according to P. At all other times, Y_t stays constant.

To make this formal, let $(X_k)_{k=0}^{\infty}$ be the discrete-time Markov chain described by P, where $X_0 \sim \nu$, and let $\{t_k\}_{k=1}^{\infty} \overset{\text{i.i.d.}}{\sim} \mathsf{Exp}(1)$ be completely independent of $(X_k)_{k=0}^{\infty}$. Then for every $k \in \mathbb{N}$, let $Y_t = X_k$ for all $t \in [T_k, T_{k+1})$.

Let us now study the distribution of Y_t. If we define the random variable $N_t = \max\{k \in \mathbb{N} : T_k \leq N_t\}$ which counts the number of transitions up to time t for every $t \in \mathbb{R}_{\geq 0}$, then

$$Y_t \sim \nu \sum_{k=0}^{\infty} \Pr[N_t = k] \cdot \mathsf{P}^k.$$

Lemma 1.6. *For every $t \in \mathbb{R}_{\geq 0}$, N_t distributed as a Poisson random variable with mean t.*

A proof is provided in Appendix B. This tells us that

$$\begin{aligned}\sum_{k=0}^{\infty} \Pr[N_t = k] \cdot \mathsf{P}^k &= \sum_{k=0}^{\infty} \frac{e^{-t} t^k}{k!} \cdot \mathsf{P}^k \\ &= e^{-t} \sum_{k=0}^{\infty} \frac{(t\mathsf{P})^k}{k!} \\ &= \exp\left(-t \cdot (\mathsf{Id} - \mathsf{P})\right) \overset{\mathsf{def}}{=} H_t.\end{aligned} \tag{8}$$

This is the *heat semigroup* (or *heat kernel*) of P. It itself is a reversible Markov chain with stationary measure μ. Instead of proving Theorem 1.3, we will prove the following which is sufficient for our purposes.

Theorem 1.7 ([DS96; BT03])**.** *For every $t \in \mathbb{R}_{\geq 0}$ and every initial distribution ν,*

$$\mathscr{D}_{\mathrm{KL}}(\nu H_t \,\|\, \mu) \leq e^{-\varrho(\mathsf{P}) \cdot t} \cdot \mathscr{D}_{\mathrm{KL}}(\nu \,\|\, \mu).$$

The same inequality holds if we replace $\varrho(\mathsf{P})$ with $4\kappa(\mathsf{P})$.

Theorem 1.7 essentially implies Theorem 1.3 except one has to convert continuous-time mixing to discrete-time mixing. The intuition here is that because N_t is Poisson with mean t, we expect via concentration for Poisson random variables that $H_t \approx \mathsf{P}^t$. In particular, $\mathsf{P}^{C \cdot t}$ for a large enough constant $C > 1$ "should mix better" than H_t. One slight subtlety here is that H_t is automatically aperiodic and in fact, all of its eigenvalues are nonnegative, even if P has nontrivial periodicity. So, this approximation $H_t \approx \mathsf{P}^t$ can't actually hold for arbitrary reversible chains P. For more details on the translation between continuous-time and discrete-time mixing, see [LPW17].

Proof of Theorem 1.7. For convenience, define $f_t = \frac{d(\nu H_t)}{d\mu}$ so that $\mathscr{D}_{\mathrm{KL}}(\nu H_t \,\|\, \mu) = \mathrm{Ent}_{\mu}(f_t)$. Note that since νH_t is a probability distribution, $\mathbb{E}_{\mu}[f_t] = 1$ for all $t \in \mathbb{R}_{\geq 0}$. Differentiating w.r.t. time, we see that

$$\begin{aligned}
\frac{d}{dt}\operatorname{Ent}_\mu(f_t) &= \sum_{x\in\Omega}\mu(x)\cdot\frac{d}{dt}(H_tf_0)(x)\log(H_tf_0)(x) && \left(\text{Using } \tfrac{d(\nu H_t)}{d\mu} = H_t\tfrac{d\nu}{d\mu} \text{ and } \mathbb{E}_\mu[f_t]=1\right)\\
&= -\sum_{x\in\Omega}\mu(x)\cdot\left((\mathsf{Id}-\mathsf{P})H_tf_0\right)(x)\cdot\log\left(H_tf_0\right)(x) - \underbrace{\sum_{x\in\Omega}\mu(x)\cdot\left((\mathsf{Id}-\mathsf{P})f_0\right)(x)}_{=\mathbb{E}_\mu[f_0]-\mathbb{E}_\mu[\mathsf{P}f_0]=1-1=0}\\
&= -\sum_{x\in\Omega}\mu(x)\cdot\left((\mathsf{Id}-\mathsf{P})\,f_t\right)(x)\cdot\log f_t(x) && (\text{Using } f_t = H_tf_0)\\
&= -\left\langle(\mathsf{Id}-\mathsf{P})\,f_t,\log f_t\right\rangle_\mu\\
&= -\mathcal{E}_{\mathsf{P}}(f_t,\log f_t). && (\text{Definition of Dirichlet form})
\end{aligned}$$

(One easy way to see the second step is by differentiating $\left(H_tf_0\right)(x) = e^{-t}\sum_{k=0}^{\infty}\frac{t^k}{k!}\cdot\left(\mathsf{P}^kf_0\right)(x)$ term-by-term.) Hence, if we have a modified log-Sobolev inequality with constant $\varrho = \varrho(\mathsf{P})$, then

$$\frac{d}{dt}\operatorname{Ent}_\mu(f_t) \le -\varrho\cdot\operatorname{Ent}_\mu(f_t).$$

This is great for us because rearranging yields a constant bound on the logarithmic derivative

$$\frac{d}{dt}\log\operatorname{Ent}_\mu(f_t) = \frac{\frac{d}{dt}\operatorname{Ent}_\mu(f_t)}{\operatorname{Ent}_\mu(f_t)} \le -\varrho.$$

Integrating from 0 to t, we obtain

$$\log\operatorname{Ent}_\mu(f_t) - \log\operatorname{Ent}_\mu(f_0) \le -\varrho\cdot t.$$

Rearranging again yields the desired inequality. We can also replace $\varrho(\mathsf{P})$ with $4\kappa(\mathsf{P})$ by Proposition 1.5. □

Remark 3. A similar calculation reveals that

$$\frac{d}{dt}\operatorname{Var}_\mu\left(f_t\right) = -2\cdot\mathcal{E}_{\mathsf{P}}(f_t,f_t),$$

which implies that $\chi^2(\nu H_t \,\|\, \mu) \le e^{-\gamma(\mathsf{P})\cdot t}\cdot\chi^2(\nu\,\|\,\mu)$. This is essentially the statement that having a Poincaré Inequality/spectral gap implies fast mixing, except from the continuous-time lens.

2 Concentration via Functional Inequalities

As in most proofs of Chernoff-type concentration inequalities, to prove Theorem 1.4, we need a strong bound on the *moment generating function* $\mathbb{E}_\mu\left[e^{tf}\right]$.

Proposition 2.1. *Let* P *be a reversible Markov chain on* Ω *with stationary measure* μ*. Then for every* $t\ge 0$*, we have the differential inequality*

$$\frac{d}{dt}\left[\frac{\log\mathbb{E}_\mu\left[e^{tf}\right]}{t}\right] \le \frac{v(f)}{2\varrho(\mathsf{P})}, \tag{9}$$

which in particular, implies the bound

$$\mathbb{E}_\mu\left[e^{tf}\right] \le \exp\left(t\cdot\mathbb{E}_\mu[f] + t^2\cdot\frac{v(f)}{2\varrho(\mathsf{P})}\right). \tag{10}$$

That Proposition 2.1 (more specifically, Eq. (10)) implies Theorem 1.4 is standard, and a proof is provided in Appendix B. We prove Proposition 2.1 via the famous *Herbst argument*. Really, the key inequality here is Eq. (9), which is why we decided to include it in the statement of Proposition 2.1.

Proof of Proposition 2.1. Let us first argue that Eq. (9) indeed implies Eq. (10). To see this, observe that by integrating Eq. (9) from 0 to t, we obtain

$$\frac{\log \mathbb{E}_\mu\left[e^{tf}\right]}{t} - \lim_{s\to 0}\left\{\frac{\log \mathbb{E}_\mu\left[e^{sf}\right]}{s}\right\} \leq \frac{v(f)}{2\varrho(\mathsf{P})}\cdot t.$$

Noting that the limit in the left-hand side evaluates to $\mathbb{E}_\mu[f]$ by L'Hôpital's Rule, and so rearranging immediately gives Eq. (10). Hence, all that remains is to establish Eq. (9) by leveraging the modified log-Sobolev inequality. By explicit calculation, we have

$$\begin{aligned}\frac{d}{dt}\left[\frac{\log \mathbb{E}_\mu\left[e^{tf}\right]}{t}\right] &= \frac{\mathbb{E}_\mu\left[f\cdot e^{tf}\right]}{t\cdot \mathbb{E}_\mu\left[e^{tf}\right]} - \frac{\log \mathbb{E}_\mu\left[e^{tf}\right]}{t^2} \\ &= \frac{\mathrm{Ent}_\mu\left(e^{tf}\right)}{t^2\cdot \mathbb{E}_\mu\left[e^{tf}\right]} \\ &\leq \frac{\mathcal{E}_{\mathsf{P}}\left(tf, e^{tf}\right)}{\varrho(\mathsf{P})\cdot t^2\cdot \mathbb{E}_\mu\left[e^{tf}\right]}. && \text{(Definition of } \varrho(\mathsf{P}))\end{aligned}$$

Hence, Eq. (9) is equivalent to

$$\frac{2}{t}\mathcal{E}_{\mathsf{P}}\left(tf, e^{tf}\right) \leq t\cdot v(f)\cdot \mathbb{E}_\mu\left[e^{tf}\right].$$

Expanding the definition of the Dirichlet form,

$$\begin{aligned}\frac{2}{t}\mathcal{E}_{\mathsf{P}}\left(tf, e^{tf}\right) &= \sum_{x,y\in\Omega}\mu(x)\mathsf{P}(x\to y)\cdot(f(x)-f(y))\cdot\left(e^{tf(x)}-e^{tf(y)}\right) \\ &= \sum_{x,y\in\Omega}\mu(x)\mathsf{P}(x\to y)\cdot(f(x)-f(y))^2\cdot\left(\frac{e^{tf(x)}-e^{tf(y)}}{f(x)-f(y)}\right) \\ &= \sum_{x\in\Omega}\mu(x)\cdot\left(\sum_{y\in\Omega}\mathsf{P}(x\to y)\cdot(f(x)-f(y))^2\right)\cdot\max_{y\in\Omega}\left\{\frac{e^{tf(x)}-e^{tf(y)}}{f(x)-f(y)}\right\} \\ &\leq v(f)\cdot\sum_{x\in\Omega}\mu(x)e^{tf(x)}\cdot\max_{y\in\Omega}\left\{\frac{1-e^{-t(f(x)-f(y))}}{f(x)-f(y)}\right\} \\ &\leq v(f)\cdot\sup_{z\in\mathbb{R}}\left\{\frac{1-e^{-tz}}{z}\right\}\cdot\mathbb{E}_\mu\left[e^{tf}\right] \\ &\leq t\cdot v(f)\cdot\mathbb{E}_\mu\left[e^{tf}\right]. && \text{(Using } 1-x\leq e^{-x} \text{ for all } x\in\mathbb{R})\end{aligned}$$

□

We conclude this section with a conjectured connection between path coupling and the modified log-Sobolev inequality.

Conjecture 1 (Peres–Tetali; see e.g. [ELL17])**.** *Let* P *be a reversible Markov chain on a metric space* (Ω, d) *with stationary measure* μ*. If there exists* $\alpha > 0$ *such that Eq. (1) holds for* P *w.r.t. the transportation distance* $\mathscr{W}_1(\cdot,\cdot)$*, then* P *also satisfies* $\varrho(\mathsf{P}) \geq \Omega(\alpha)$*.*

It is known by the work of Eldan–Lee–Lehec [ELL17] that such a contraction w.r.t. $\mathscr{W}_1(\cdot,\cdot)$ implies a *transport-entropy inequality*, which is *equivalent* to sub-Gaussian concentration statements like Eq. (10), and weaker than the modified log-Sobolev inequality by Proposition 2.1. Transport-entropy inequalities will appear again in a future lecture, but for now, we refer interested readers to an excellent monograph of Gozlan–Léonard [GL10]. For further positive results in support of Conjecture 1, see [Mar19; Liu21; Bla+22].

References

[Bla+22] Antonio Blanca, Pietro Caputo, Zongchen Chen, Daniel Parisi, Daniel Štefankovič, and Eric Vigoda. "On Mixing of Markov Chains: Coupling, Spectral Independence, and Entropy Factorization". In: *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2022, pp. 3670–3692 (cit. on p. 6).

[BLM16] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. 1st ed. Oxford University Press, 2016 (cit. on p. 3).

[BT03] Sergey Bobkov and Prasad Tetali. "Modified Log-Sobolev Inequalities, Mixing and Hypercontractivity". In: *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing*. STOC '03. San Diego, CA, USA: Association for Computing Machinery, 2003, pp. 287–296 (cit. on pp. 3, 4).

[Cha04] Djalil Chafaï. "Entropies, convexity, and functional inequalities, On Φ-entropies and Φ-Sobolev inequalities". In: *Journal of Mathematics of Kyoto University* 44.2 (2004), pp. 325–363 (cit. on p. 2).

[DH02] Persi Diaconis and Susan Holmes. "Random Walks on Trees and Matchings". In: *Electronic Journal of Probability* 7.6 (2002), pp. 1–17 (cit. on p. 3).

[DS81] Persi Diaconis and Mehrdad Shahshahani. "Generating a random permutation with random transpositions". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 57 (1981), pp. 159–179 (cit. on p. 3).

[DS87] Persi Diaconis and Mehrdad Shahshahani. "Time to Reach Stationarity in the Bernoulli–Laplace Diffusion Model". In: *SIAM Journal on Mathematical Analysis* 18 (1 1987) (cit. on p. 3).

[DS96] Persi Diaconis and Laurent Saloff-Coste. "Logarithmic Sobolev inequalities for finite Markov chains". In: *The Annals of Applied Probability* 6.3 (Aug. 1996) (cit. on pp. 3, 4).

[ELL17] Ronen Eldan, James R. Lee, and Joseph Lehec. "Transport-Entropy Inequalities and Curvature in Discrete-Space Markov Chains". In: *A Journey Through Discrete Mathematics* (2017), pp. 391–406 (cit. on pp. 3, 6).

[FOW22] Yuval Filmus, Ryan O'Donnell, and Xinyu Wu. "Log-Sobolev inequality for the multislice, with applications". In: *Electronic Journal of Probability* 27 (2022), pp. 1–30 (cit. on p. 3).

[GL10] Nathael Gozlan and Christian Léonard. "Transport Inequalities. A Survey". In: *Markov Processes And Related Fields* 16 (4 2010), pp. 635–736 (cit. on p. 6).

[Goe04] Sharad Goel. "Modified logarithmic Sobolev inequalities for some models of random walk". In: *Stochastic Processes and their Applications* 114 (1 2004), pp. 51–79 (cit. on p. 3).

[Gro75] Leonard Gross. "Logarithmic Sobolev inequalities". In: *American Journal of Mathematics* 97.4 (1975), pp. 1061–1083 (cit. on p. 3).

[GZ03] A. Guionnet and B. Zegarlinksi. "Lectures on Logarithmic Sobolev Inequalities". In: *Séminaire de Probabilités, XXXVI, volume 1801 of Lecture Notes in Math.* Springer, Berlin, 2003, pp. 1–134 (cit. on p. 3).

[HS20] Jonathan Hermon and Justin Salez. "Modified log-Sobolev inequalities for strong-Rayleigh measures". In: *arXiv preprint arXiv:1902.02775* (2020) (cit. on p. 2).

[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. 2).

[Led99] Michel Ledoux. "Concentration of measure and logarithmic Sobolev inequalities". In: *Séminaire de Probabilités XXXIII (Strasbourg)* 33 (1999), pp. 120–216 (cit. on p. 3).

[Liu21] Kuikui Liu. "From Coupling to Spectral Independence and Blackbox Comparison with the Down-Up Walk". In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2021)*. Vol. 207. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, 32:1–32:21 (cit. on p. 6).

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

[LY98] Tzong-Yow Lee and Horng-Tzer Yau. "Logarithmic Sobolev inequality for some models of random walks". In: *The Annals of Probability* 26.4 (1998), pp. 1855–1873 (cit. on p. 3).

[Mar19] Katalin Marton. "Logarithmic Sobolev inequalities in discrete product spaces". In: *Combinatorics, Probability and Computing* 28.6 (2019), pp. 919–935. DOI: 10.1017/S0963548319000099 (cit. on p. 6).

[MT06] Ravi Montenegro and Prasad Tetali. "Mathematical Aspects of Mixing Times in Markov Chains". In: *Foundations and Trends in Theoretical Computer Science* 1.3 (2006), pp. 237–354 (cit. on p. 3).

[ODo14] Ryan O'Donnell. *Analysis of Boolean Functions*. USA: Cambridge University Press, 2014. ISBN: 1107038324 (cit. on p. 3).

[Oll09] Yann Ollivier. "Ricci curvature of Markov chains on metric spaces". In: *Journal of Functional Analysis* 256 (3 Feb. 2009), pp. 810–864 (cit. on p. 1).

[Sam05] Marcus D. Sammer. "Aspects of Mass Transportation in Discrete Concentration Inequalities". PhD thesis. Georgia Institute of Technology, Apr. 2005 (cit. on p. 3).

[Sca97] Fabio Scarabotti. "Time to Reach Stationarity in the Bernoulli–Laplace Diffusion Model with Many Urns". In: *Advances in Applied Mathematics* 18 (3 1997), pp. 351–371 (cit. on p. 3).

[ST10] Fabio Scarabotti and Filippo Tolli. In: *Forum Mathematicum* 22 (5 2010), pp. 879–911 (cit. on p. 3).

[STY23] Justin Salez, Konstantin Tikhomirov, and Pierre Youssef. "Log concavity and concentration of Lipschitz functions on the Boolean hypercube". In: *Journal of Functional Analysis* 285 (9 2023), p. 110076 (cit. on p. 3).

A Comparison Inequalities and Entropy vs. Variance

The goal of this section is to prove Proposition 1.5. We do this in a sequence of lemmas. Throughout, P is some fixed reversible Markov chain on Ω with stationary measure μ.

Lemma A.1. *For every nonnegative function $f : \Omega \to \mathbb{R}_{\geq 0}$, $\mathcal{E}_{\mathsf{P}}(f, \log f) \geq 4 \cdot \mathcal{E}_{\mathsf{P}}\left(\sqrt{f}, \sqrt{f}\right)$. In particular, $4\kappa(\mathsf{P}) \leq \varrho(\mathsf{P})$.*

Proof. Note that the second inequality follows immediately from the first. To prove the first claim, it suffices to show that for any $x, y \in \Omega$,

$$(f(x) - f(y)) \cdot (\log f(x) - \log f(y)) \geq 4 \cdot \left(\sqrt{f(x)} - \sqrt{f(y)}\right)^2.$$

Without loss of generality, we may assume $f(x) \geq f(y)$. Rearranging yields that this is equivalent to

$$\log \frac{f(x)}{f(y)} \geq 4 \cdot \frac{\sqrt{f(x)} - \sqrt{f(y)}}{\sqrt{f(x)} + \sqrt{f(y)}} = 4 \cdot \frac{\sqrt{\frac{f(x)}{f(y)}} - 1}{\sqrt{\frac{f(x)}{f(y)}} + 1}.$$

From here, it suffices to verify the simple one-dimension inequality $\log z \geq 2 \cdot \frac{z-1}{z+1}$ for $z \geq 1$. This holds for $z = 1$, and so it suffices to show that the derivative of the left-hand side is greater than the derivative of the right-hand side for all $z \geq 1$, i.e. $\frac{1}{z} \geq \frac{4}{(z+1)^2}$ for all $z \geq 1$. This holds by "completing the square". $\square$

Lemma A.2. *For every real-valued function $f : \Omega \to \mathbb{R}$, $\lim_{c \to \infty} \operatorname{Ent}_\mu\left((c + f)^2\right) = 2 \operatorname{Var}_\mu(f)$. In particular, $\kappa(\mathsf{P}) \leq \frac{1}{2}\gamma(\mathsf{P})$.*

Proof. The first statement is equivalent to

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \operatorname{Ent}_\mu\left((1 + \sqrt{\epsilon} f)^2\right) = 2 \operatorname{Var}_\mu(f).$$

We use the second-order Taylor series for $x \mapsto (1+x)\log(1+x)$, which is $x + \frac{1}{2}x^2 + O(x^3)$ and is valid for all $|x| < 1$. This trick is sometimes called *linearization*. Applying this, we get

$$\begin{aligned}
\operatorname{Ent}_\mu\left((1+\sqrt{\epsilon}f)^2\right) &= \mathbb{E}_\mu\left[\left(2\sqrt{\epsilon}f + \epsilon f^2\right) + \frac{1}{2}\left(2\sqrt{\epsilon}f + \epsilon f^2\right)^2\right] \\
&\quad - \left(2\sqrt{\epsilon}\mathbb{E}_\mu[f] + \epsilon\mathbb{E}_\mu[f^2]\right) - \frac{1}{2}\left(2\sqrt{\epsilon}\mathbb{E}_\mu[f] + \epsilon\mathbb{E}_\mu[f^2]\right)^2 \\
&\quad + O(\epsilon^{3/2}) \\
&= 2\operatorname{Var}_\mu(f) + O(\epsilon^{3/2}).
\end{aligned}$$

The first equality follows immediately. For the second inequality, let f be any function attaining $\gamma(\mathsf{P})$, i.e. $\frac{\mathcal{E}_\mathsf{P}(f,f)}{\operatorname{Var}_\mu(f)} = \gamma(\mathsf{P})$. Then

$$\begin{aligned}
\kappa(\mathsf{P}) &\le \inf_{c\in\mathbb{R}_{\ge 0}} \frac{\mathcal{E}_\mathsf{P}(c+f, c+f)}{\operatorname{Ent}_\mu\left((c+f)^2\right)} \\
&= \inf_{c\in\mathbb{R}_{\ge 0}} \frac{\mathcal{E}_\mathsf{P}(f, f)}{\operatorname{Ent}_\mu\left((c+f)^2\right)} \\
&\le \frac{\mathcal{E}_\mathsf{P}(f, f)}{\lim_{c\to\infty}\operatorname{Ent}_\mu\left((c+f)^2\right)} \\
&= \frac{\mathcal{E}_\mathsf{P}(f, f)}{2\operatorname{Var}_\mu(f)} \\
&= \frac{1}{2}\gamma(\mathsf{P}).
\end{aligned}$$

□

Lemma A.3. *For every real-valued function $f : \Omega \to \mathbb{R}$,*

$$\begin{aligned}
\operatorname{Ent}_\mu\left(1 + \frac{f}{c}\right) &= \frac{1}{2c^2}\left(\operatorname{Var}_\mu(f) + o_c(1)\right) \\
\mathcal{E}_\mathsf{P}\left(1 + \frac{f}{c}, \log\left(1 + \frac{f}{c}\right)\right) &= \frac{1}{c^2}\left(\mathcal{E}_\mathsf{P}(f,f) + o_c(1)\right),
\end{aligned}$$

where $o_c(1)$ is a quantity tending to 0 as $c \to \infty$. In particular, $\varrho(\mathsf{P}) \le 2\gamma(\mathsf{P})$.

Proof. Again, using the second-order Taylor series for $x \mapsto (1+x)\log(1+x)$, which is $x+\frac{1}{2}x^2+O(x^3)$ and is valid for all $|x| < 1$, we get

$$\begin{aligned}
\operatorname{Ent}_\mu\left(1 + \frac{f}{c}\right) &= \mathbb{E}_\mu\left[\frac{f}{c} + \frac{1}{2}\cdot\frac{f^2}{c^2}\right] - \mathbb{E}_\mu\left[\frac{f}{c}\right] - \frac{1}{2}\mathbb{E}_\mu\left[\frac{f}{c}\right]^2 + O(1/c^3) \\
&= \frac{1}{2c^2}\left(\operatorname{Var}_\mu(f) + o_c(1)\right).
\end{aligned}$$

The rest of the proof is similar to the one for Lemma A.2. □

B Unfinished Proofs

Proof of Lemma 1.6. First, we claim that the probability density function T_k is given by

$$p_k(t) = \frac{t^{k-1}e^{-t}}{(k-1)!}.$$

This is a straightforward calculation obtained by inductively convolving k copies of the $x \mapsto e^{-x}$, the probability density function of $\mathsf{Exp}(1)$. It follows that

$$\begin{aligned}
\Pr[N_t = k] &= \Pr[T_k \le t \text{ and } t < T_{k+1}] \\
&= \int_0^t \frac{u^{k-1}e^{-u}}{(k-1)!} \underbrace{\int_{t-u}^{\infty} e^{-v}\, dv}_{=e^{-t}e^{u}}\, du \\
&= e^{-t} \int_0^t \frac{u^{k-1}}{(k-1)!}\, du \\
&= \frac{t^k e^{-t}}{k!}.
\end{aligned}$$

This is the probability mass function of the Poisson random variable with mean 1. $\square$

Proof of Proposition 2.1 $\implies$ Theorem 1.4. Observe that for a fixed parameter $t \ge 0$ to be determined later,

$$\begin{aligned}
\Pr_{x\sim\mu}\left[f(x) \ge \mathbb{E}_\mu[f] + \epsilon\right] &= \Pr_{x\sim\mu}\left[e^{tf(x)} \ge e^{t\cdot\mathbb{E}_\mu[f]+t\cdot\epsilon}\right] \\
&\le \frac{\mathbb{E}_\mu\left[e^{tf}\right]}{\exp\left(t\cdot\mathbb{E}_\mu[f] + t\cdot\epsilon\right)} && \text{(Markov's Inequality)} \\
&\le \exp\left(t^2 \cdot \frac{v(f)}{2\varrho(\mathsf{P})} - t\cdot\epsilon\right). && \text{(Proposition 2.1)}
\end{aligned}$$

The optimal choice for $t \ge 0$ is clearly $\frac{\varrho(\mathsf{P})\cdot\epsilon}{v(f)}$, which yields the first inequality. The second inequality follows by combining the first inequality with $4\kappa(\mathsf{P}) \le \varrho(\mathsf{P})$ (see Proposition 1.5). $\square$