CSE 525: Randomized Algorithms **Spring 2023**

Lecture 18: Path Technology

Lecturer: Shayan Oveis Gharan *May 25th*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

18.1 Dirichlet Form

Consider a reversible Markov chain with Kernel K on state space Ω and stationary π. For two functions $f, g \in \Omega \to \mathbb{R}$ define

$$\langle f, g\rangle_\pi = \sum_x f(x)g(x)\pi(x).$$

So, far we have discussed that for a probability distribution p on Ω, $p^T K$ is the distribution of the chain after one step, i.e., if $X_0 \sim p$, then $X_1 \sim p^T K$.

What happens if we multiply a vector with K on the right? In that case, K acts as an averaging operator: For any $y \in \Omega$,

$$Kf(x) = \sum_y K(x,y)f(y),$$

in other words, we sample a state $y \sim x$ according to $K(x, .)$ and then we average f.

Fact 18.1. *for any pair of functions $f, g \in \Omega \to \mathbb{R}$,*

$$\langle Kf, g\rangle_\pi = \langle f, Kg\rangle_\pi.$$

Proof. We write

$$\begin{aligned}\langle Kf, g\rangle_\pi &= \sum_x \pi(x)g(x)\sum_y K(x,y)f(y) \underset{\text{reversibility}}{=} \sum_x g(x)\sum_y K(y,x)\pi(y)f(y)\\ &= \sum_y f(y)\pi(y)\sum_x P(y,x)g(x) = RHS.\end{aligned}$$

□

So the Markov kernel K is self-adjoint w.r.t. the inner product $\langle ., .\rangle_\pi$. It follows that for any self-adjoint operator w.r.t. an inner product one can apply the spectral theorem and obtain real eigenvalues $\lambda_1, \dots, \lambda_n$ with corresponding eigenvectors $v_1, \dots, v_n$ such that $\langle v_i, v_j\rangle_\pi = 0$ and $\|v_i\|_\pi^2 = 1$ for all i, j.

18.2 Mixing Time via Spectral Gap

Definition 18.2 (Variance)**.** *For a function $f : \Omega \to \mathbb{R}$ define*

$$\mathrm{Var}(f) = \langle f - \mathbb{E}f, f - \mathbb{E}f\rangle_\pi = \langle f, f\rangle_\pi - 2\mathbb{E}f\langle f, \mathbf{1}\rangle_\pi + (\mathbb{E}f)^2 = \langle f, f\rangle_\pi - (\mathbb{E}f)^2$$

where as usual we used $\|\mathbf{1}\| = 1$ *and* $\mathbb{E}f = \langle f, \mathbf{1}\rangle_\pi$. *Note that by definition* $\mathrm{Var}(f) = \mathrm{Var}(f + \alpha 1)$ *for any* $\alpha \in \mathbb{R}$. *Furthermore,*

$$\mathrm{Var}(K(f + \alpha 1)) = \mathrm{Var}(Kf + \alpha 1) = \mathrm{Var}(Kf).$$

Lemma 18.3. *For any function* $f : V \to \mathbb{R}$, $\mathrm{Var}(f) - \mathrm{Var}(Kf) = \langle (I - K^2)f, f\rangle_\pi$.

Proof. First, by definition of variance we can shift f and assume without loss of generality that $\mathbb{E}f = \langle f, \mathbf{1}\rangle_\pi = 0$. Also note that $\mathbb{E}Kf = \langle \mathbf{1}, Kf\rangle_\pi = 0$. Therefore,

$$\mathrm{Var}(f) - \mathrm{Var}(Kf) = \langle f, f\rangle_\pi - \langle Kf, Kf\rangle_\pi = \langle (I - K^2)f, f\rangle_\pi$$

The RHS is also known as the Dirichlet form of f with respect to K^2. The operator $I - K^2$ is called the normalized Laplacian of K^2. □

Equivalently, the Dirichlet form can be written as

$$\langle (I - K)f, f\rangle_\pi = \sum_{x,y} \pi(x)K(x, y)(f(x) - f(y))^2 \geq 0$$

The above lemma in particular implies that $\mathrm{Var}(Kf) \leq \mathrm{Var}(f)$ for any function $f : \Omega \to \mathbb{R}$.

Let

$$\lambda_2(I - K^2) = \min_f \frac{\langle (I - K^2)f, f\rangle_\pi}{\mathrm{Var}(f)}$$

The minimum is over all functions f that are orthogonal to the all ones vector, $\langle f, \mathbf{1}\rangle_\pi = 0$.

This quantity is the second smallest eigenvalue of $I - K^2$ which can also be seen as difference of 1 and the square of the second largest eigenvalue of K in absolute value. It is also called the Poincare constant. It then follows that for any function $f : \Omega \to \mathbb{R}$ that is not the all-ones function,

$$\frac{\mathrm{Var}(Kf)}{\mathrm{Var}(f)} = 1 - \frac{\langle (I - K^2)f, f}{\mathrm{Var}(f)} = \frac{\langle K^2 f, f\rangle}{\mathrm{Var}(f)} \leq \lambda_2(K^2).$$

Applying this repeatedly, we get

Corollary 18.4. *For any function* $f : V \to \mathbb{R}$,

$$\mathrm{Var}(K^t f) \leq \lambda_2(K^2)^t \,\mathrm{Var}(f).$$

Lemma 18.5. *For any* $x \in \Omega$, *and any* $\epsilon > 0$ *the walk started at* x *satisfies:*

$$\sum_y |K^t(x, y) - \pi(y)| \leq \epsilon$$

as long as $t \geq \frac{\log \epsilon^{-1} \ln \pi(x)^{-1}}{1 - \lambda_2(K^2)}$. *In fact the same proof also bounds the L-2 mixing.*

Proof. Let $f = \mathbf{1}_x/\pi(x)$, i.e., $f(y) = 0$ for $y \neq x$ and $1/\pi(x)$ otherwise. Note that $\mathbb{E}f = 1$.

For some t that we choose later we write

$$\begin{aligned}
\mathrm{Var}(K^{2t} f) &= \langle K^t f, K^t f\rangle_\pi - (\mathbb{E}K^t f)^2 \\
&= \mathbb{E}_{y\sim\pi} K^t f(y) \cdot K^t f(y) - \langle K^t f, \mathbf{1}\rangle^2 \\
&\underset{\text{using } f = \mathbf{1}_x/\pi(x)}{=} \mathbb{E}_{y\sim\pi} \left(\frac{K^t(y, x)}{\pi(x)}\right)^2 - (\mathbb{E}f)^2 \\
&\underset{\text{using reversibility}}{=} \mathbb{E}_{y\sim\pi} \left(\frac{K^t(x, y)}{\pi(y)}\right)^2 - 1 = \mathbb{E}_{y\sim\pi} \left(\frac{K^t(x, y)}{\pi(x)} - 1\right)^2
\end{aligned}$$

The RHS is called the L-2 distance for the random walk started at x.

The L-2 distance is always greater than or equal to the total variation distance. That can be seen by an application of the Cauchy-Schwarz inequality. In particular,

$$\sum_y |K^t(x,y) - \pi(y)| = \mathbb{E}_{y\sim\pi}\left|\frac{K^t(x,y)}{\pi(y)} - 1\right| \leq \mathbb{E}_{y\sim\pi}\left(\frac{K^t(x,y)}{\pi(y)} - 1\right)^2$$
$$= \operatorname{Var}(K^{2t}f) \underset{Corollary\ 18.4}{\leq} \lambda_2(K^2)^t \operatorname{Var}(f) \leq \frac{\lambda_2(K^2)^t}{\pi(x)} \leq \epsilon$$

Where to get the last inequality it is enough to let $t = \frac{\ln \epsilon^{-1} \ln \pi(i)^{-1}}{1-\lambda_2(K^2)}$. □

Definition 18.6 (L-p mixing time)**.** *For a function $f : \Omega \to \mathbb{R}$ define,*

$$\|f\|_{p,\pi} := \left(\sum_{x\in\Omega} \pi(x) f(x)^p\right)^{1/p}.$$

For a state x, the ℓ_p mixing time of the walk started at x is defined as

$$\tau_{x,p}(\epsilon) := \min\left\{t : \left\|\frac{K^t(x,.)}{\pi} - 1\right\|_{p,\pi} \leq \epsilon\right\}$$

In particular, for $p = \infty$, it is:

$$\tau_{x,\infty}(\epsilon) := \min\left\{t : |\frac{K^t(x,y)}{\pi(y)} - 1| \leq \epsilon, \forall y\right\}$$

Lemma 18.7. *For any reversible chain,*

$$\max_x \tau_{x,\infty}(\epsilon) \leq 2 \max_x \tau_{x,2}(\sqrt{\epsilon})$$

Proof. For any x, y,

$$\frac{K^{2t}(x,y)}{\pi(y)} = \sum_z \frac{K^t(x,z)}{\pi(z)} \frac{K^t(z,y)}{\pi(y)} \pi(z) \underset{\text{reversibility}}{=} \sum_z \frac{K^t(x,z)}{\pi(z)} \frac{K^T(y,z)}{\pi(z)} \pi(z)$$

Therefore,

$$\left\langle \frac{K^t(x,.)}{\pi} - 1, \frac{K^t(y,.)}{\pi} - 1 \right\rangle = \frac{K^{2t}(x,y)}{\pi(y)} - 1$$

It thus follows that,

$$\max_{x,y} \left|\frac{K^{2t}(x,y)}{\pi(y)} - 1\right| \leq \max_{x,y} \left\langle \frac{K^t(x,.)}{\pi} - 1, \frac{K^t(y,.)}{\pi} - 1 \right\rangle \leq \max_x \|\frac{K^t(x,.)}{\pi} - 1\|_{2,\pi}^2$$

The conclusion follows. □

18.3 Path Technology

Next we discuss a new method to bound the Poincaré constant.

Think of a multicommodity flow problem where we want to send $\pi(x)\pi(y)$ unit of flow between each pair of states x, y. Note that these are supposed to be disjoint commodities. This means that if a 1 unit $x \to y$ flow goes on edge (u, v) from u to v and 1 unit of $x' \to y'$ flow goes on the same edge from v to u they don not cancel out each other. The capacity of each edge $e = (u, v)$ is $Q(e) = \pi(u)K(u, v) = \pi(v)K(v, u)$. Recall that $Q(e)$ represents the flow of the probability mass along edge e at stationarity. Suppose we choose a path $P_{x,y}$ between each pair of vertices x, y (note that more generally this may be a distribution of paths). For an edge e let flow of e be defined as follows:

$$f(e) = \sum_{x,y:e\in P_{x,y}} \pi(x)\pi(y).$$

The congestion of e is defined as $\frac{f(e)}{Q(e)}$. In the following lemma we show that the inverse of the Poincaré is at most $\max_{x,y} |P_{x,y}| \cdot \max_e \frac{f(e)}{Q(e)}$.

Lemma 18.8. *For any reversible Markov chain, suppose for every pair of states $x, y \in \Omega$ we choose a path $P_{x,y}$.*

$$\frac{1}{\alpha} \leq \max_e \frac{f(e)}{Q(e)} \cdot \max_{x,y} |P_{x,y}|.$$

where α is the Poincaré constant.

Proof. Consider an arbitrary function f. Note that for every path $P_{x,y}$ we can choose an orientation of the edges of this say e^+ is the head and e^- is the tail such that $f(x) - f(y) = \sum_{e\in P_{x,y}} f(e^+) - f(e^-)$. We can write

$$\begin{aligned}
\mathrm{Var}(f) &= \frac{1}{2}\sum_{x,y}(f(x) - f(y))^2\pi(x)\pi(y) \\
&= \frac{1}{2}\sum_{x,y}\left(\sum_{e\in P_{x,y}} f(e^+) - f(e^-)\right)^2 \pi(x)\pi(y) \\
&\leq \frac{1}{2}\sum_{x,y}|P_{x,y}| \sum_{e\in P_{x,y}} (f(e^+) - f(e^-))^2\pi(x)\pi(y) \\
&= \frac{1}{2}\sum_{e=(x',y')} (f(x') - f(y'))^2 \frac{Q(e)}{Q(e)} \sum_{x,y:e\in P_{x,y}} |P_{x,y}|\pi(x)\pi(y) \leq max_{x,y}|P_{x,y}| \cdot \max_e \frac{f(e)}{Q(e)} \cdot \mathcal{E}(f, f).
\end{aligned}$$

where the inequality follows by Cauchy-Schwarz inequality. □

Note that we typically let $P_{x,y}$ be a shortest path from x to y. So, $P_{x,y}$ is no more than the diameter. Since in all Markov chains that we constant diameter is only a polynomial in the size of the input, usually, the most important parameter to bound the mixing time is the maximum congestion.

Let us use the above machinery to bound the mixing time of a simple lazy random walk on a path. Consider the a simple path of length n. The stationarity distribution is almost uniform. Observe that there is a unique path between each pair of vertices. So, the edge $(n/2, n/2+1)$ has the maximum congestion of about

$$\frac{(n/2)(n/2)\frac{1}{n}\frac{1}{n}}{\frac{1}{n}\frac{1}{4}} \approx n.$$

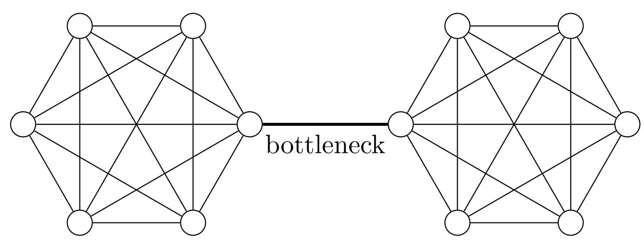
Since $\mathrm{Diam}(G) = n$, we get

$$\frac{1}{\alpha} \leq \frac{1}{n^2}.$$

It follows that the chain mixes in $O(n^2 \log(n))$. Note that this bound is $O(\log n)$ off from the bound we proved using strong stationarity time. The reason is that here we are just upper bound the second eigenvalue of K and we use a very crude bound on all other eigenvalues (we are upper bound each λ_i by λ_2). So, in many applications of the Path technology this $O(\log(n))$ loss in inherent, but usually it does not chain the mixing time significantly, because it is logarithmic in the size of the state space.

Dumbell Graph Note that if we have a Markov chain with a diameter that is polynomial in the size of the state space, then we should expect a very slow mixing. This is essentially what happens in the path example. So, one can ask if we have a Markov chain with a logarithmic size diameter, can it still have a mixing time polynomial in the size of the instance? The answer is yes, and the famous Dumbell graph is perhaps the worst example. In this graph the single edge which is connecting the two cliques will be a bottleneck because all of the flow between the two cliques must go over this edge.

Note that although we didn't prove, but the path technology gives a necessary and sufficient condition for bounding the Poincaré constant. Namely, if we can show that for any routing of the multicommodity flow there is an edge with large congestion, it would imply that the chain mixes very slowly.



References