MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.265/15.070J Lecture 15 — April 10, 2017
Lecturer: Yury Polyanskiy — Scribe notes by Paxton Turner

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They are posted to serve class purposes.*

The Dirichlet form and monotone chains

Content.

1. Review of eigenfunctions and spectral gap
2. The Dirichlet form and applications
3. Monotone chains

1 Review of eigenfunctions and spectral gap

We recall some of the results from last time and derive some new ones. Let $\mathcal{X}$ be a finite state space for an irreducible reversible Markov chain X_t with stationary distribution π.

- Recall the *relaxation time* $t_{rel} := 1/\gamma_a$ where $\gamma_a := sup_{\lambda \neq 1}(1 - |\lambda|)$ is the *absolute* spectral gap. Then

$$t_{rel} \leq t_{mix} \leq \frac{1}{t_{rel}} \log\left(\frac{1}{\pi_{min}}\right)$$

  where π_{min} is the minimal atom of π.

- If $\mathcal{X}$ is a transitive chain (i.e., the underlying graph of $\mathcal{X}$ is vertex transitive), then

$$d_2(t) = \left(\sum \lambda_j^{2t}\right)^{1/2}.$$

- From problem 3 in homework 7, we have for any function $g : \mathcal{X} \to \mathbb{R}$,

$$Var[P_t g] \leq (1 - \gamma_a)^{2t} Var[g].$$

- The following is an analogous result for covariances.

$$\begin{aligned} Cov(f(X_t), g(X_0)) = (P_t f, g) &= \sum_{j \neq 1} \lambda_j^t a_j b_j \\ &\leq (1-\gamma_a)^t \sum_{j=1} |a_j b_j| \leq (1-\gamma_a)^t \sqrt{Var(f) Var(g)} \end{aligned}$$

- Now we can dope out a more concrete interpretation of the relaxation time by considering fluctuations of the empirical mean of $f(X_t)$:

$$\begin{aligned} Var\left(\frac{1}{T}\sum_{t=0}^{T-1} f(X_t)\right) &\leq \frac{1}{T^2}\sum_{s=0}^{T-1} 2TCov(f(X_t, f(X_0)) \\ &\leq \frac{1}{T}\sum_{t=0}^{\infty}(1-\gamma_a)^t Var[f] = \frac{2Var[f]}{T/t_{rel}}. \end{aligned}$$

The last result tells us that to generate multiple samples va MCMC, it suffices for the sake of an approximation to run the MCMC once and sample from it every t_{rel} steps. This is useful computationally speaking because the mixing time is often much larger than the relaxation time.

Consider the following interpretation of an eigenfunction f on $\mathcal{X}$. If we know the value of $f(X_t)$, then a natural estimator is

$$\hat{f}_{t+1} = \lambda f(X_t).$$

Observe the following:

- $\lambda \approx 1 \Rightarrow f$ is almost constant
- $\lambda \approx 0 \Rightarrow f(X_t)$ is independent of $f(X_{t-1})$.
- $\lambda \approx -1 \Rightarrow f$ is oscillatory.

Moreover, if

$$\lambda_2 = \max\left(\mathbb{E}f(X_{t+1})f(X_t) : \mathbb{E}f = 0, \mathbb{E}f^2 = 1\right)$$

is large, then it is impossible to partition $\mathcal{X}$.

2 The Dirichlet form and applications

The Dirichlet form allows us to develop a notion of spectral gap for Markov chains that are not reversible.

Definition 1 (Dirichlet form)**.**

$$\mathcal{E}(f,f) \triangleq ((I-P)f,f).$$

Definition 2 (Poincaré constant/spectral gap)**.**

$$\gamma \triangleq \inf_{f\neq 0} \frac{\mathcal{E}(f,f)}{Var[f]}$$

By definition, we have the so-called *Poincaré inequality*:

$$Var[f] \leq \frac{1}{\gamma}\mathcal{E}(f,f).$$

Proposition 1. *Let P be a reversible MC.*

1. $\mathcal{E}(f,f) = \frac{1}{2}\mathbb{E}[f(X_1)-f(X_0)]$.
2. $\gamma_a = 1-\lambda_2 \geq \gamma$
3. *If P is a simple random walk on a d-regular graph G,*

$$\mathcal{E}(f,f) = \frac{1}{2d}\mathbb{E}\left\|\nabla f\right\|^2$$

where $\nabla f(x) = \sum_{y\sim x} f(x)-f(y)$.

We think of *1.* in the above proposition as the *local variance* and *3.* as an interpretation of how the stationary distribution π plays against the ambient metric.

Proof. *1.*

$$((I-P)f,f) = \mathbb{E}f^2 - (Pf,f) = \mathbb{E}f^2 - \mathbb{E}f(X_1)f(X_0) = \frac{1}{2}\mathbb{E}(f(X_0)-f(X_1))^2.$$

2. Let $f = \sum_{j\neq 1} a_j f_j$, then

$$\mathcal{E}(f,f) = \sum_j a_j^2(1-\lambda_j) \leq (1-\lambda_2)Var[f].$$

3.

$$\mathcal{E}(f,f) = \frac{1}{2n}\sum_{x\sim y}\frac{1}{d}(f(x)-f(y))^2 = \frac{1}{2d}\mathbb{E}\left\|\nabla f\right\|^2.$$

□

As an application, we get an interesting proof of Efron-Stein.

Proof of Efron-Stein. Define $P(x,y) = \pi(y)$ to be the Markov chain that mixes in one step. Note that $\gamma(P) = 1$. Next,

$$\mathcal{E}(f,f) = ((I-P)f,f) = ((I-\mathbb{E})f,f) = ((I-\mathbb{E})f,(I-\mathbb{E})f) = Var[f]$$

if $\mathbb{E}f = 0$. Define the product MC

$$P^{(n)} = \frac{1}{n}\sum_i I\otimes I\otimes\cdots\otimes P\otimes\cdots\otimes I,$$

and observe that $\gamma(P^{(n)}) = \frac{1}{n}\gamma(P) = \frac{1}{n}$. Let $\{X_i\}\sim\pi$ be iid random variables, and set $X = (X_1,\ldots,X_n)$. Note that

$$\mathcal{E}(f,f) = \frac{1}{2n}\sum_{i=1}\mathbb{E}(f(X)-f(X^{(i)})^2$$

We see that

$$Var(f(X)) \le n\frac{1}{2n}\sum_i\mathbb{E}(f(X)-f(X^{(i)})^2 = \frac{1}{2}\sum_i\mathbb{E}(f(X)-f(X^{(i)})^2.$$

□

Our second application is the Cheeger inequality.

Theorem 2 (Cheeger Inequality)**.** *Let P be a reversible MC. Recall the following:*

- *The capacitance of the edge from x to y is $c(x,y) \triangleq \pi(x)P(x,y)$;*
- *more generally, we define the capacitance of a subset S of the state space:*

$$c(S,S^c) \triangleq \sum_{x\in S, y\in S^c} c(x,y);$$

and

- *the Cheeger constant is defined to be*

$$\varphi^* = \min_{\pi(S)\leq \frac{1}{2}} \frac{c(S, S^c)}{\pi(s)}.$$

Then we have

$$\frac{\gamma}{2} \leq \varphi^* \leq \sqrt{2\gamma}.$$

Proof. First, we prove the lower bound. Take some $S \subset \mathcal{X}$. Then

$$\mathcal{E}(\mathbb{1}, \mathbb{1}) = \frac{1}{2}\mathbb{E}[\mathbb{1}_S(X_1) - \mathbb{1}_S(X_0)]^2$$
$$= \frac{1}{2}\left(\mathbb{P}[X_0 \in S, X_1 \notin S] + \mathbb{P}[X_0 \notin S, X_1 \in S]\right) = \mathbb{P}[X_0 \in S, X_1 \notin S] = C(S, S^c).$$

Moreover, $Var[\mathbb{1}_S] = \pi(S)\pi(S^c)$. Therefore,

$$\gamma \leq \frac{\mathcal{E}(f, f)}{Var[f]} = \frac{C(S, S^c)}{\pi(S)(1 - \pi(S))} \leq 2\frac{C(S, S^c)}{\pi(S)} = 2\varphi^*.$$

For the opposite direction, take an arbitrary $g \geq 0$ such that $\pi(g = 0) \geq |\frac{1}{2}|$. Observe that

$$\mathbb{P}[g(X_1) \leq \theta < g(X_0)] = \mathbb{P}[X_0 \in \{x : g(x) > \theta\}, X_1 \in S^c]$$
$$\geq \varphi^* \mathbb{P}[g(X_0) > \theta] = \varphi^* \mathbb{E}g.$$

Now integrate both sides with respect to θ:

$$\mathbb{E}\int d\theta\, \mathbb{1}\{g(X_1) \leq \theta < g(X_0)\} = \frac{1}{2}\mathbb{E}[g(X_1) - g(X_0)]$$
$$= \mathbb{E}(g(X_0) - g(X_1))_+ \geq \varphi^* \mathbb{E}g.$$

The third equality follows from reversibility of P. Take g such that $\pi(g = 0) \geq \frac{1}{2}$. Let $h = g^2$. By the previous inequality and Cauchy-Schwarz,

$$h \leq \frac{1}{2}\mathbb{E}[g^2(X_0) - g^2(X_1)]$$
$$\leq \frac{1}{2}\sqrt{4\mathbb{E}g^2 - \mathbb{E}[(g(X_0) - g(X_1))^2]} = \sqrt{\mathbb{E}g^2}\sqrt{2\mathcal{E}(g, g)}.$$

Then from the previous two inequalities,

$$\mathcal{E}(g, g) \geq \frac{\varphi^{*2}}{2}\mathbb{E}g^2. \tag{1}$$

Take $f = \max(f_2, 0)$, and WLOG $\pi(f = 0) \geq \frac{1}{2}$. From Jenson's inequality, $Pf \geq \max(Pf_2, 0) = \lambda_2 f$. Thus $(Pf, f) \geq \lambda_2(f, f)$ and $\mathcal{E}(f, f) \leq \gamma(f, f)$. Then apply the inequality (1) with $g = f$. $\square$

Note that for the n-cycle: $\gamma = c/n^2, \varphi* = 1/n$. And for the n-hypercube: $\gamma = 1/n, \varphi^* = c/n$.

3 Monotone chains

Consider the lazy random walk W on the integer points P_k of the interval $[1, k]$. This MC is reversible and irreducible, and last time we showed that

$$f_2 = \cos\left(\frac{\pi}{k+1}(x-1)\right);\ \lambda_2 = \frac{1}{2}\left(1 + \cos\frac{\pi}{k}\right).$$

Based on our heuristic for eigenfunctions, we interpret f_2 as the slowest varying non-constant eigenfunction. A suggestive way to see this is to interpret W as a projection of a lazy random walk on the n-cycle C. Then f_2 passes through the points of C closest to the x-axis, illustrating geometrically its slowly varying nature.

Recall that given a poset $\mathcal{X}$ with order relation $\leq$, Strassen's monotone coupling tells us if ν stochastically dominates μ, then there exists a coupling $X_1 \sim \mu, Y_1 \sim \nu$ such that

$$\mathbb{P}_{X_1,Y_1}[X_1 \leq Y_1] = 1.$$

Definition 3. *A transition kernel P on a poset $\mathcal{X}$ is* monotone *if $P(x, \cdot) \leq P(y, \cdot)$ for all $x \leq y$.*

Definition 4. *P has* monotone jumps *if*

$$P(x, y) > 0 \Rightarrow\ (x \geq y \text{ or } x \leq y).$$

Theorem 3. *Let P be monotone irreducible reversible MC, then*

1. *There exists f which is strictly increasing such that $Pf = \lambda_2 f$.*
2. *Suppose P has monotone jumps. Then if $Pf = \lambda f$ and f is strictly increasing, then $\lambda = \lambda_2$.*

Proof. We only give a proof of the first part. Let g_0 be any strictly increasing function. Let $g = g_0 - \mathbb{E}g_0$, which has 0 mean and is also strictly increasing. Set

$$V_2 = \text{span}\{\text{all } \lambda_2 - \text{eigenfunctions}\}.$$

WLOG g is not orthogonal to V_2. That is, if $g \perp V_2$, take $f_2 \in V_2$ and consider $g + \epsilon f_2$, which is increasing for $\epsilon << 1$. Observe that $(g + \epsilon f_2, f_2) > 0$. Set $g = h_2 + h_{\text{rest}}$, where $h_2 \in V_2$ and $h_{\text{rest}} \perp V_2$.

Next,

$$g_t := P_t g/\lambda_2^t \Rightarrow P_t g = \lambda_2^t h_2 + P_t h_{\text{rest}}$$

where $P_t h_{\text{rest}} \leq |\lambda_3^t| C$ for some constant $C > 0$ independent of t. Therefore,

$$P_t g/\lambda_2^t = h_2 \pm \left(\frac{\lambda_3}{\lambda^2}\right)^t C.$$

Therefore, $P_t g \to h_2 \in V_2$ pointwise and in any norm, since we are working in a finite-dimensional vector space.

As a sidenote: the above gives a numerical procedure for computing the second largest eigenvalue: choose some increasing function g on the data and then compute $P_t g/g$ for t large.

The key observation required is that g monotone implies P_g monotone for a monotone kernel P. Indeed, if $x > y$,

$$Pg(x) - Pg(y) = \mathbb{E}[g(X_1) - g(Y_1)] \geq 0,$$

which implies that h_2 is monotone. □

Proposition 4. *If P is irreducible, reversible, monotone, and has monotone jumps, then the stationary distribution π is also monotone.*

Proof. Note that

$$(f(X_1) - f(X_0))(g(X_1) - g(X_0)) \geq 0,$$

so that

$$\mathbb{E}fg \geq (P_t f, g) \to \mathbb{E}f\mathbb{E}g.$$

□