CPSC 661: Sampling Algorithms in Machine Learning February 10, 2021

Lecture 4

Lecturer: Andre Wibisono *Scribe: Daniel Edelberg*

1 Outline

Today's lecture covers three main parts:

- Mixing Time in χ^2-divergence
- Conductance
- Cheeger's Inequality

2 Mixing Time in χ^2-Divergence

We will revisit the theorem on mixing time in χ^2-divergence from last lecture in order to present the proof on the mixing time bound.

Theorem 1 (Mixing Time in χ^2-Divergence)**.** *Let P be a reversible Markov chain with respect to ν and with spectral gap γ. Along the Markov chain $X_k \sim \rho_k$:*

$$\chi^2_\nu(\rho_k) \le (1-\gamma)^{2k}\, \chi^2_\nu(\rho_0)$$

Proof. Let $h_k = \frac{d\rho_k}{d\nu}$ be the density of ρ_k with respect to ν. By assumption, P is reversible, so that $h_{k+1} = Ph_k$ (left as an exercise). This implies that $h_k = P^k h_0$. By previous Lemma, P is self-adjoint, so P has real eigenvalues

$$1 = \lambda_1 > \lambda_2 \ge \cdots \ge \lambda_n > -1$$

with orthonormal basis of eigenfunctions $\mathbb{1} = \phi_1, \phi_2, \ldots, \phi_n \in L^2(\nu)$ (where n may be ∞). We obtain the strict inequalities $\lambda_1 > \lambda_2$ and $\lambda_n > -1$ by the assumptions that P is irreducible and that P is not periodic, respectively. The spectral gap γ implies that

$$|\lambda_i| \le 1 - \gamma$$

for $i = 2, \ldots, n$. We may write $h_0 = \frac{d\rho_0}{d\nu}$ as a linear combination of the eigenfunctions as they form a basis:

$$h_0 = \mathbb{1} + c_2\phi_2 + \cdots + c_n\phi_n$$

where we may find the constants c_i via inner products:

$$c_i = \langle h_0, \phi_i \rangle_\nu.$$

Note that $c_1 = \langle h_0, \phi_1 \rangle_\nu = \langle h_0, \mathbb{1} \rangle_\nu = 1$, so it is omitted from the basis expansion of h_0. Now apply P to the equation:

$$Ph_0 = P\mathbb{1} + c_2 P\phi_2 + \cdots + c_n P\phi_n$$

and use the eigenvalue relation to get:

$$h_1 = Ph_0 = \mathbb{1} + c_2\lambda_2\phi_2 + \cdots + c_n\lambda_n\phi_n.$$

Applying this k times:

$$h_k = P^k h_0 = \mathbb{1} + c_2\lambda_2^k\phi_2 + \cdots + c_n\lambda_n^k\phi_n.$$

Recall from the definition of χ^2-divergence:

$$\chi^2_\nu(\rho_k) = \mathrm{Var}_\nu(h_k) = \|h_k - \mathbb{1}\|^2_\nu = \|c_2\lambda_2^k\phi_2 + \cdots + c_n\lambda_n^k\phi_n\|^2_\nu$$

Since the eigenfunctions are orthonormal, i.e.

$$\langle \phi_i, \phi_j \rangle_\nu = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

then

$$\chi^2_\nu(\rho_k) = c_2^2\lambda_2^{2k} + \cdots + c_n^2\lambda_n^{2k}$$

By the spectral gap definition: $|\lambda_i| \leq 1 - \gamma$, so

$$\begin{aligned} \chi^2_\nu(\rho_k) &\leq (1-\gamma)^{2k}\,(c_2^2 + \cdots + c_n^2) \\ &= (1-\gamma)^{2k}\,\|c_2\phi_2 + \cdots + c_n\phi_n\|^2_\nu \\ &\leq (1-\gamma)^{2k}\,\|h_0 - \mathbb{1}\|^2_\nu \\ &\leq (1-\gamma)^{2k}\,\chi^2_\nu(\rho_0). \end{aligned}$$

□

2.1 Examples

Random Walk on Graph:

1. Cycle Graph on n vertices

   - n even (bipartite): $\lambda_n = -1$, $\gamma = 0$. (The random walk does not converge because there is a parity constraint that prevents mixing.)

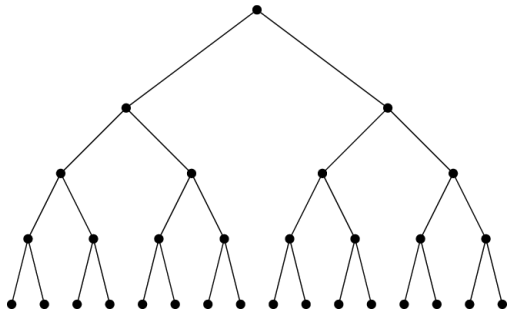

Figure 1: Complete Binary Tree with $h = 5$.

- n odd: $\gamma = \mathcal{O}(\frac{1}{n^2}) \implies \tau = \mathcal{O}(n^2)$

2. Complete Binary Tree on $n = 2^h - 1$ vertices (height h)
   - $\gamma = \mathcal{O}(\frac{1}{n}) \implies \tau = \mathcal{O}(n)$
3. Hypercube on $n = 2^h$ vertices
   - $\gamma = \mathcal{O}(\frac{1}{h}) = \mathcal{O}(\frac{1}{\log n}) \implies \tau = \mathcal{O}(h) = \mathcal{O}(\log n)$

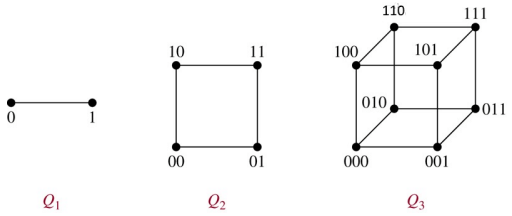

Figure 2: Hypercube with $h = 1, 2$ and 3.

For general P, recall we have the characterization:

$$\gamma = \inf_{f \in L^2(\nu)} \frac{\langle f, Lf \rangle_\nu}{\mathrm{Var}_\nu(f)} = \inf_{\substack{f \in L^2(\nu) \\ \mathbb{E}_\nu[f] = \langle f, \mathbb{1} \rangle_\nu = 0}} \frac{\langle f, Lf \rangle_\nu}{\|f\|_\nu^2}$$

So we may pick any f and plug into the above formula to get an upper bound on the spectral gap γ. For the cycle graphs, choosing a function that increases linearly will give the $\mathcal{O}(\frac{1}{n^2})$ upper bound. For a binary tree, we may choose a function that assigns 0 to nodes on the left and 1 to nodes on the right and again calculate the ratio to get an upper bound of $\mathcal{O}(\frac{1}{n})$. More clever choices of f will yield tighter upper bounds in many cases.

Exercise: Compute/bound the spectral gap of the Ornstein-Uhlenbeck process on $\mathcal{X} = \mathbb{R}^n$:

$$X_{k+1} = e^{-\eta} X_k + \sqrt{(1 - e^{-2\eta})} Z$$

where $Z \sim \mathcal{N}(0, I)$ and $\eta > 0$ is step size. Here the target distribution is $\nu = \mathcal{N}(0, I)$.

3 Conductance

A more intuitive understanding of mixing time for random walks begins by asking what prevents a random walk from mixing quickly. The short answer is a bottleneck: a region within the space $\mathcal{X}$ which is "narrow" and prevents a random walk from easily passing between two or more regions of the space. Being able to go from one part of the space to the other is integral to mixing, so a bottleneck that is hard to pass through causes the random walk to mix slowly. If the space has no bottleneck, then we expect a random walk to have fast mixing through the space. We will formalize this idea in this section.

3.1 Ergodic Flow

Let P be reversible with respect to ν on $\mathcal{X}$. Let subset $A \subseteq \mathcal{X}$ define a **partition** or a **cut**, such that $\mathcal{X} = A \cup A^c$, where $A^c = \mathcal{X} \setminus A$ is the complement.

Definition 1 (Ergodic Flow)**.** *The* ***ergodic flow*** *of a partition A of $\mathcal{X}$, for Markov chain P is given by*

$$\Phi(A) = \int_A P_x(A^c)\, d\nu(x) = \Pr(X_0 \in A, X_1 \notin A)$$

where $X_0 \sim \nu$, $X_1 \mid X_0 \sim P_{X_0}$.

Since ν is stationary for P, one can verify that $\Phi(A) = \Phi(A^c)$, or equivalently,

$$\Pr(X_0 \in A, X_1 \notin A) = \Pr(X_0 \notin A, X_1 \in A)$$

This condition does not require reversibility of P. In the case of a "bottleneck" on $\mathcal{X}$ that would prevent fast mixing, there would be some subset A such that $\Phi(A)$ is small. However, we can also force $\Phi(A)$ to be small by making A itself small. Therefore, we should normalize and consider the ratio:

$$\frac{\Phi(A)}{\nu(A)}$$

so that the "bad case" of a bottleneck occurs when we have a large A and a small $\Phi(A)$. We also have that $\Phi(A^c) = \Phi(A)$, and $\nu(A^c) = 1 - \nu(A)$ since ν is a probability distribution, so we should normalize $\Phi(A)$ by the smaller of $\nu(A)$ and $1 - \nu(A)$. This brings us to the next definition.

Definition 2 (Cut-Ratio)**.** *The* ***cut-ratio*** *(or* ***expansion****) for a set $A \subseteq \mathcal{X}$ is given by*

$$C(A) = \frac{\Phi(A)}{\min\{\nu(A), 1 - \nu(A)\}}.$$

Suppose that $\nu(A) \leq \frac{1}{2}$. Then we have that

$$C(A) = \frac{\Phi(A)}{\nu(A)} = \frac{\Pr(X_0 \in A, X_1 \notin A)}{\Pr(X_0 \in A)}$$

which we recognize from an application of Bayes' Rule to be:

$$C(A) = \Pr(X_1 \notin A \mid X_0 \in A)$$

so that this cut-ratio represents a conditional probability of X_1 not in A given that X_0 is in A.

3.2 Conductance

Definition 3 (Conductance)**.** *The* ***conductance*** *of a Markov chain P is*

$$\phi = \inf_{\substack{A \subset \mathcal{X} \\ 0 < \nu(A) < 1}} \frac{\Phi(A)}{\min\{\nu(A), 1 - \nu(A)\}}.$$

Equivalently,

$$\phi = \inf_{0 < \nu(A) \leq \frac{1}{2}} \Pr(X_1 \notin A \mid X_0 \in A)$$

for $X_0 \sim \nu$, $X_1 \mid X_0 \sim P_{X_0}$.

A large conductance ϕ implies that every subset has a good expansion, and consequently that a random walk on the set $\mathcal{X}$ mixes quickly.

4 Cheeger's Inequality

Theorem 2 (Cheeger's Inequality)**.** *Let P be reversible with spectral gap γ and conductance ϕ. Then*

$$\frac{\phi^2}{2} \leq \gamma \leq 2\phi.$$

Proof. (Of the upper bound: $\gamma \leq 2\phi$) Recall by definition,

$$\gamma = \inf_{f \in L^2(\nu)} \frac{\langle f, Lf \rangle_\nu}{\mathrm{Var}_\nu(f)}$$

$$\phi = \inf_{A \subset \mathcal{X}} \frac{\Phi(A)}{\min\{\nu(A), 1 - \nu(A)\}}.$$

Given $A \subset \mathcal{X}$, let $f(x) = \mathbb{1}_A(x) \in L^2(\nu)$, i.e.,

$$f(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A. \end{cases}$$

Then

$$\langle f, Lf \rangle_\nu = \frac{1}{2}\mathbb{E}[f(x_1) - f(x_0))^2] = \frac{1}{2}\mathbb{E}[(\mathbb{1}_A(x_1) - \mathbb{1}_A(x_0))^2].$$

Note that

$$(\mathbb{1}_A(x_1) - \mathbb{1}_A(x_0))^2 = \begin{cases} 1 & \text{if } (x_0 \in A, x_1 \notin A) \\ 1 & \text{if } (x_0 \notin A, x_1 \in A) \\ 0 & \text{otherwise.} \end{cases}$$

Then we can write

$$\begin{aligned} \langle f, Lf \rangle_\nu &= \frac{1}{2}\left(\Pr(x_0 \in A, x_1 \notin A) + \Pr(x_0 \notin A, x_1 \in A)\right) \\ &= \frac{1}{2}\left(\Phi(A) + \Phi(A^c)\right) \\ &= \Phi(A). \end{aligned}$$

We note that if $X \sim \nu$, then $f(X) = \mathbb{1}_A(X)$ is a Bernoulli random variable on the space $\mathcal{X}$ that takes value 1 with probability $\nu(A)$, and takes value 0 otherwise. Then its variance is:

$$\mathrm{Var}_\nu(f) = \nu(A)(1 - \nu(A)).$$

Since $0 < \nu(A) < 1$:

$$\mathrm{Var}_\nu(f) \geq \frac{1}{2}\min\{\nu(A), 1 - \nu(A)\}.$$

So:

$$\gamma = \inf_{f \in L^2(\nu)} \frac{\langle f, Lf \rangle_\nu}{\mathrm{Var}_\nu(f)} \leq \inf_{A \subset \mathcal{X}} \frac{\Phi(A)}{\frac{1}{2}\min\{\nu(A), 1 - \nu(A)\}} = 2\phi.$$

□

For proof of the lower bound, see [Lovász and Simonovits, 1993, Lemma 1.6].

4.1 Examples

In general, the upper and lower bounds on Cheeger's inequalities can be saturated. For example, for the random walk on graph:

1. Cycle on n vertices (n odd)
    - Choose cut such that the vertices are almost evenly split

- $\phi = \mathcal{O}(\frac{1}{n})$, $\gamma = \mathcal{O}(\frac{1}{n^2})$
- $\gamma \sim \phi^2$

2. Complete binary tree on n vertices
   - Choose cut that splits tree down the middle (adding the top node to either side)
   - $\phi = \mathcal{O}(\frac{1}{n})$, $\gamma = \mathcal{O}(\frac{1}{n})$
   - $\gamma \sim \phi$
3. Hypercube $\{0,1\}^h$ on $n = 2^h$ vertices
   - $\phi = \mathcal{O}(\frac{1}{h}) = \mathcal{O}(\frac{1}{\log n})$, $\gamma = \mathcal{O}(\frac{1}{\log n})$
   - $\gamma \sim \phi$

4.2 s-Conductance

Cheeger's inequality $\gamma \geq \frac{\phi^2}{2}$ implies that mixing time in χ^2-divergence is

$$\tau = \tilde{\mathcal{O}}\left(\frac{1}{\gamma}\right) = \tilde{\mathcal{O}}\left(\frac{1}{\phi^2}\right).$$

But sometimes conductance is too strict, e.g. a cone in $\mathbb{R}^n$. Starting at the "tip" of the cone with an extremely small subset can yield a very poor mixing time, but starting in other regions does not yield these kind of issues. Therefore, we want to devise a way to remove these particularly problematic regions of small measure in order to have a more consistent measurement of conductance that ignores certain small regions.

Definition 4 (s-Conductance)**.** *For $0 \leq s \leq \frac{1}{2}$, the **$s$-conductance** of a Markov chain P is*

$$\phi_s = \inf_{\substack{A \subset \mathcal{X} \\ s < \nu(A) < 1-s}} \frac{\Phi(A)}{\min\{\nu(A) - s, 1 - s - \nu(A)\}}.$$

The s value prevents using A that are too small or too large. We note that when $s = 0$, we get the usual definition of conductance. Furthermore, as we increase s, ϕ_s also increases. Using this, we can get a mixing time bound in Total Variation distance under warm-start conditions, which we will see next time.

References

[Lovász and Simonovits, 1993] Lovász, L. and Simonovits, M. (1993). Random walks in a convex body and an improved volume algorithm. *Random structures & algorithms*, 4(4):359–412.