MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.265/15.070J Lecture — Mar 22, SP17
Lecturer: Yury Polyanskiy — Scribe notes by Julia Romanski

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

Mixing Time and Sampling

Content.

1. Summary from last time
2. Lower bounds on mixing time: Cheeger constant
3. Sampling
4. Metropolis Algorithm
5. Glauber Dynamics

1 Summary from last time

There are several distances that we consider.

Definition 1.

$$d(t) = \sup_{x \in \mathcal{X}} d_{TV}\left(P_t(x, \cdot), \pi\right)$$

Definition 2.

$$\overline{d}(t) = \sup x, y \in \mathcal{X} = d_{TV}\left(P_t(x, \cdot), P_t(y, \cdot)\right)$$

Definition 3.

$$d^{(2)}(t) = \sup_{x \in \mathcal{X}} \|q_t(x, \cdot) - 1\|_2$$

where

$$q_t(x, y) = \frac{P_t(x, y)}{\pi(y)}$$

is the relative density.

Aside: The distance $d(t)$ typically falls sharply at a cutoff point while $d^{(2)}(t)$ falls exponentially.

Also recall the definition for mixing time:

Definition 4.

$$t_{mix} = \inf\left\{t : d(t) \leq \frac{1}{4}\right\}$$

Review of some proofs from last time:

1. For an irreducible, aperiodic Markov chain,

$$d(t) \searrow 0.$$

   The proof was through a coupling induction via Dobrushin extension.

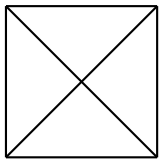
2. Mixing time for lazy random walks on the circle and hypercube.

   a) On the circle C_n, the $t_{mix} = \Theta(n^2)$. The coupling was done by only moving one chain at a time, which reduces to a simple random walk on $[0, \frac{n}{2}]$. Additionally, we learned from homework that the cover time for the simple random walk on C_n is $\Theta(n^2)$.

   b) For the lazy random walk on the hypercube H_n, $t_{mix} = \Theta(n \log n)$. At each time step, one coordinate of the chains would be chosen at random and a new sample would be generated. This coupling reduces to the Coupon Collector Problem.

Remark: Coupling two chains by moving them in the same direction is not always a good idea (think of symmetric random walk on $[-n, n]$). In some situations, like for diffusion processes, reflection coupling gives faster mixing.

Example: Symmetric random walk on K_n (clique of size n). For example, this is K_4:



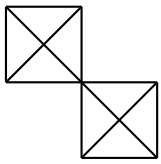
- The symmetric random walk on K_n is irreducible, and aperiodic for $n \geq 3$.

- The stationary distribution is uniform over the n states. Wherever the chain starts, $P_1(x,\cdot)$ will be a permutation of $\left(0, \frac{1}{n-1}, \frac{1}{n-1}, ..., \frac{1}{n-1}\right)$, which means $d(1) = \frac{1}{n}$, and so $t_{mix} = 1$.
- Similarly, $P_1(y,\cdot)$ will be a permutation of $\left(0, \frac{1}{n-1}, \frac{1}{n-1}, ..., \frac{1}{n-1}\right)$, so

$$\overline{d}(1) = \frac{1}{n-1}$$

.

Now let's consider $K_n \vee K_n$, which is two cliques of size n that share one point. For example, this is $K_4 \vee K_4$:



Let's work on finding an upper bound for t_{mix}. Let X_t and Y_t be two symmetric random walks on $K_n \vee K_n$ starting at x_0 and y_0 respectively, where $x_0 \neq y_0$. Let's assume x_0 and y_0 are in opposite cliques and not equal to the joining vertex, which we call ρ. For the coupling, let X_t and Y_t advance to X_{t+1} and Y_{t+1} independently if they are in opposite cliques. Otherwise use the optimal coupling that we had for K_n, which gives

$$P(X_{t+1} \neq Y_{t+1} | \mathcal{F}_t) \leq \frac{1}{n-1}.$$

- Now let

$$\tau_i = \inf\left\{t \geq 0 : X_t \neq \rho, Y_t \neq \rho, X_t \text{ and } Y_t \text{are in the same clique}\right\}.$$

  Then we have $\mathbb{P}(X_{\tau_i+1} \neq Y_{\tau_i+1}) \leq \frac{1}{n-1}$.
- If we prove that $\mathbb{E}[\tau_i] \leq 4n$ then $\mathbb{P}(X_{t+1} \neq Y_{t+1}) \leq \frac{4n}{t} + \frac{1}{n-1}$ by Chebyshev.
- Let $\tau_\rho = \inf\{t > 0 : X_t = \rho\}$. At time $\tau_\rho + 1$, with probability $\frac{1}{2}\left(1 - \frac{1}{n-1}\right) \geq \frac{1}{4}$, X and Y will be in the same clique. Then we have $\mathbb{E}[\tau_i] \leq \mathbb{E}[\tau_\rho + 1] \cdot 4$ by the property of geometric trials. Now τ_ρ is a geometric random variable with parameter $\frac{1}{n-1}$, which means $\mathbb{E}[\tau_\rho + 1] \cdot 4 = 4n$, and $t_{mix} \leq 16n$.

Now let's work on finding a lower bound for t_{mix}. Let S^c be the set of vertices in one of the K_n cliques, including ρ. Then clearly $\pi(S^c) \geq \frac{1}{2}$. Let's put x_0 in S. We have

$$\begin{aligned}
\mathbb{P}(X_t \in S^c) &\leq \mathbb{P}(\tau_\rho \leq t) \\
&\leq \frac{t}{n-1} && \text{Union Bound} \\
&\leq \frac{1}{4} && \text{if } t < \frac{n}{4}
\end{aligned}$$

Therefore $t_{mix} > \frac{n}{4}$. Combining with the upper bound, $t_{mix} = \Theta(n)$.

2 Lower bounds on mixing time: Cheeger constant

We will begin by defining the Cheeger constant and related quantities for a Markov chain, working with the graph representation of the chain.

Definition 5 (Edge weights)**.** *For a directed edge $\vec{e}$ from x to y in the graph representation of the chain, let $C(\vec{e}) = \pi(x)P(x,y)$. This quantity is often called the capacity or conductance.*

Definition 6 (Cheeger constant for a set)**.** *For any set S of vertices, define*

$$\varphi(S) = \frac{C(S, S^C)}{\pi(S)} = \frac{\sum_{x \in S, y \in S^c} C(x,y)}{\pi(S)}.$$

Definition 7 (Cheeger constant of a Markov chain)**.**

$$\varphi_\star = \min_{S: 0 < \pi(s) \leq \frac{1}{2}} \varphi(S)$$

Proposition 1.

$$\varphi(S) = \mathbb{P}^\pi \left(X_1 \notin S | X_0 \in S \right).$$

In this way $\varphi(S)$ can be thought of an escape probability.

Proof. The proof follows from the definition. □

Proposition 2. *Let*

$$\tilde{\varphi}_\star = \min_S \frac{C(S, S^c)}{\pi(S)\pi(S^c)}.$$

Proof. Follows from $C(S, S^c) = C(S^c, S)$. □

Proposition 3. *Consider a simple random walk on a graph $\mathcal{G}$. Let $E(S, S^c)$ be the set of edges from S to S^c. Similarly, let $E(S, S)$ be the set of edges connecting vertices within S. Also, let $deg(x)$ be the degree of a vertex x and let $\delta_e(S)$ be the set of edges on the boundary of S. Then*

$$\varphi(S) = \frac{|E(S, S^c)|}{\sum_{x \in S} deg(x)} = \frac{|\delta_e(S)|}{|E(S, S)| + |E(S, S^c)|}.$$

For a lazy random walk on $\mathcal{G}$,

$$\varphi(S) = \frac{1}{2} \frac{|E(S, S^c)|}{\sum_x deg(x)}$$

Proof. (Sketch) Recall that $C(e) = \frac{1}{2|E|}$ for a simple random walk. □

Theorem 4 (Relating Cheeger constant to mixing)**.**

1. *For any Markov chain,*

$$t_{mix} \geq \frac{1}{4\varphi_\star}.$$

   Note that this also means

$$t_{mix} \geq \frac{1}{4\varphi(S)}$$

   for any set S.

2. *For a reversible Markov chain, $d(t) \geq (1 - \varphi_\star)^t - \frac{1}{2}$.*

Proof. Let

$$\nu(x) = \frac{\pi(x)}{\pi(S)} \mathbb{1}_{x \in S},$$

which is a valid probability distribution on $\mathcal{X}$. Take $X_0 \sim \nu$. Then

$$\mathbb{P}(X_1 \notin S) = \varphi(S).$$

Also

$$\begin{aligned} P(X_1 = a) &= \sum_{x \in S} \nu(x) P(X, a) \\ &= \frac{1}{\pi(S)} \sum_{x \in S} \pi(x) P(x, a) \\ &\leq \frac{1}{\pi(S)} \sum_{x \in \mathcal{X}} \pi(x) P(x, a) \\ &\leq \frac{\pi(a)}{\pi(S)} \end{aligned}$$

By induction, $\mathbb{P}(X_t = a) \leq \frac{\pi(a)}{\pi(S)}, \forall t$. Then we have

$$\begin{aligned}
\mathbb{P}(X_t \in S^c) &\leq \mathbb{P}(\exists v : 0 \leq v < t, X_v \in S, X_{v+1} \notin S) \\
&\leq \sum_v \mathbb{P}(X_v \in, X_{v+1} \notin S) && \text{Union Bound} \\
&= \sum_v \sum_{a \in S} \mathbb{P}(X_v = a) P(a, S^c) \\
&\leq \sum_v \sum_{a \in S} \frac{\pi(a)}{\pi(S)} P(a, S^c) \\
&= \sum_v \varphi(S) \\
&= t\varphi(S)
\end{aligned}$$

which means that $t_{mix} \geq \frac{1}{4\varphi_\star}$. Now for a reversible chain,

$$\mathbb{P}(\tau_{S^c} > t) \geq (1 - \varphi(S))^t,$$

from which it follows that

$$d(t) \geq (1 - \varphi_\star)^t - \frac{1}{2}.$$

□

3 Sampling

Given a distribution π on a state space $\mathcal{X}$, our goal is to generate a sample $X_0 \sim \pi$. There are several complications that can arise. For example, the state space could be very large. Another issue is that sometimes distributions are unnormalized, i.e. π is specified as $\pi(x) = \frac{1}{Z}\pi^\star(x)$ where Z is unknown. Another difficulty arises when π is a complicated function, e.g.

$$\pi(x) = \frac{1}{Z} \exp(-x^2 + x^3 - x^4 + x^5 - x^6)$$

because sampling from π is equivalent in difficulty to sampling a point under the graph of this function.

Motivation for sampling

1. Sampling corresponds to counting or computing volumes

2. Sampling has applications in statistical physics. There are many situations in statistical physics where we wish to sample from $\pi(x) = \frac{1}{Z}e^{-f(x)}$.

3. In combinatorics, we may wish to sample uniformly from some set of objects. For example, π could be uniform over the set of 5-colorable graphs.

4. In inference, we may have conditional probabilities to sample from. For example, the cell phone signal reconstruction model is of the form

$$Y \sim \frac{1}{z}e^{-\beta d_H(y,x)}\pi(x),$$

and we wish to sample from $\mathbb{P}_{X|Y=y}$.

4 Metropolis Algorithm

The first approach to sampling that we will discuss is called the Metropolis Algorithm. The underlying idea is given some reversible Markov chain governed by $\tilde{P}(x, y)$ with stationary distribution $\tilde{\pi}(x)$, modify it so that the new Markov chain is governed by $P(x, y)$ and has $\pi(x)$ as its stationary distribution. We need to define a function $f(x, y)$ that defines the relationship between $P(x, y)$ and $\tilde{P}(x, y)$:

$$P(x, y) = \tilde{P}(x, y)f(x, y), \forall x \neq y$$

and

$$P(x, x) = 1 - \sum_{y \neq x} P(x, y).$$

The function must satisfy

1. $f > 0$
2. f is such that $P(x, x) \geq 0$.
3. f is such that $\pi(x)P(x, y) = \pi(y)P(y, x)$.

For example,

$$f(x, y) = c\sqrt{\frac{\pi(y)}{\tilde{\pi}(y)} \cdot \frac{\tilde{\pi}(x)}{\pi(x)}}$$

satisfies the requirements, for c small enough so that the second requirement holds.

The Metropolis algorithm uses

$$f(x,y) = \min\left\{1, \frac{\pi(y)}{\tilde{\pi}(y)} \cdot \frac{\tilde{\pi}(x)}{\pi(x)}\right\} (M).$$

This function is sometimes called the "Metropolis filter."

Theorem 5. *$f(x,y)$ as defined in (M) defines a reversible Markov chain P with reversing measure π.*

Remarks

1. The resulting chain is not necessarily irreducible.
2. The resulting chain is aperiodic if $\tilde{P}(x,x) > 0$.
3. The algorithm doesn't depend on Z, because it cancels in the expression for $f(x,y)$.

Aside: Suppose $\pi(x) = \frac{1}{Z}e^{-\beta f(x)}$. Then if β is large, a random sample from π is a global minimizer of f. This is related to simulated annealing.

5 Glauber Dynamics

We consider a special case where $\mathcal{X} = \mathcal{A}^n$, i.e. $\mathcal{X}$ is a product space. Given π_{X^n}, a distribution on $\mathcal{X}$, define

$$P^{(i)}(a^n, b^n) = \begin{cases} 0 \text{ if } a_{\sim i} \neq b_{\sim i} \\ \pi_{X_i|X_{\sim i}}(b_i|a_{\sim i}) \text{ otherwise} \end{cases}$$

Theorem 6.

$$Pn\sum_{i=1}^{n} P^{(i)}$$

is reversible and has π_{X^n} as its stationary (reversing) measure.

Proof. Proof omitted. □

Examples

1. As an example of Glauber dynamics, consider a graph with vertices labelled in $\{0,1\}$. The dynamics are: sample $i \in [n]$ uniformly at random, then resample the ith coordinate via $\pi_{X_i|X\ _i}$. Note that the normalizing constant Z cancels when sampling from the conditional distribution.

2. For sampling independent sets on a graph G, Glauber dynamics arrive at a sample as fast as the Metropolis algorithm would. The dynamics are to sample $v \in V(\mathcal{G})$ and include or exclude v from the independent set with probability $\frac{1}{2}$ each unless a neighbor of v already belongs to the independent set.

3. Graph coloring. Let

$$qCol(\mathcal{G}) = \{X_i : V(G) \to [q], X_i \neq X_j \forall i \leftrightarrow j\}.$$

For this example Glauber dynamics arrive at a sample faster than the Metropolis algorithm. The dynamics are to pick a vertex v uniformly at random and recolor it among the available colors. As q increases, the dependence between coordinates decreases.

Proposition 7. *For any fixed q and with $\mathcal{G}$ a star on n vertices,*

$$t_{mix} \geq 2^{C_q n}$$

where C_q is a constant that depends on q. The reason is that transitions between subsets of the state space where the central color is fixed are rare.