CSE 521: Design and Analysis of Algorithms I **Spring 2016**

Lecture 14: Random Walks

Lecturer: Shayan Oveis Gharan *May 11th* *Scribe: ??*

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

A Markov Chain is a random process on a set of states Ω. There are weighted (directed) links between the states of the chain. For any state $a \in \Omega$ the chain moves to random neighbor of a independent of the past. This transition is proportional to the weight of the edge. In particular, say an state a is connected to b, c, d and $w_{a,b} = 0.1, w_{a,c} = 0.4, w_{a,d} = 0.2$. Then,

$$\mathbb{P}[a \to b] = \frac{0.1}{0.1 + 0.2 + 0.4} = \frac{1}{7}, \ \mathbb{P}[a \to c] = \frac{4}{7}, \text{ and } \mathbb{P}[a \to d] = \frac{2}{7}.$$

The latter is known as the *Markovian* property. That is, let $X_0, X_1, \ldots, X_t$ be the sequence of states that we took at times $0, 1, \ldots, t$. Then,

$$\mathbb{P}[X_{t+1} = a | X_0, \ldots, X_t] = \mathbb{P}[X_{t+1} = a | X_t].$$

Markov Chains have many applications in different areas of science including computer science, mathematics, finance, economics, etc. Let us describe an application in speech recognition. One of the important tasks in natural language processing is to predict the next word of a sentence given the past words. One way to model this problem is by a Markov chain. Say we have a chain where each state represents one of the words in English. There is an edge from state a to b if there is a chance that word b appears after a. The weight of the edge connecting a, b is the probability that b comes after a, i.e.,

$$\mathbb{P}[b|a] = \frac{\mathbb{P}[ab]}{\mathbb{P}[a]}.$$

One can use a large text corpus to empirically estimate the probabilities in the RHS, so construct the chain. Given the chain and the current last word of the sentence we can stochastically predict the next work.

In the rest of this lecture we study *reversible* Markov chains. We will describe the formal definition later. This family of Markov chains correspond to random walks on (weighted) undirected graphs. So, from now on, let $G = (V, E)$ be a weighted undirected graph corresponding to our Markov chain. The *transition probability* of the chain is the matrix P where for each pair of vertices i, j let $P_{i,j}$ is the probability that the next state is j conditioned on the current state being i. We have

$$P_{i,j} = \frac{w_{i,j}}{d_w(i)},$$

where as usual $w_{i,j}$ is the weight of the edge connecting i to j and $d_w(i)$ is the weighted degree of i. We can also define P as follows:

$$P = D^{-1}A. \tag{14.1}$$

We say x is a distribution over V if $\sum_i x_i = 1$ and $x_i \geq 0$ for all i. Suppose at time 0 we start the chain at state 1. Then, at time 1 we will be at a state chosen random according to the distribution $\mathbf{1}^1 P$, where $\mathbf{1}^1$ is the indicator vector of 1. In general if we start the chain at a state chosen according to a distribution x, in the next step we will be at $x^T P$.

Definition 14.1 (Stationary distribution)**.** *We say a vector $\pi \in \mathbb{R}^n$ is a stationary distribution of the chain if for any state i,*

$$\sum_j P_{j,i}\pi_j = \pi_i,$$

or in other words,

$$\pi P = \pi.$$

This says that if we start the chain according to π, the distribution of the next step is the same as what we started with.

.

The following is a fundamental theorem of Markov Chains:

Theorem 14.2. *For any graph G, if G is connected and non-bipartite, then it has a unique stationary distribution. Furthermore, for any starting distribution x,*

$$\lim_{t\to\infty} xP^t = \pi.$$

Above theorem naturally extends to non-reversible chain. Let us give a few example to show that the assumptions in the above theorem are necessary. First, let G be a single edge $(1, 2)$ and we start at 1. It follows that at odd times we will be at 2 and at even times we will be at 1. So, $\mathbf{1}^1 P^t$ never converges.

Now, suppose G is a non-bipartite graph but it has two connected components. In this case the chain has multiple stationary distributions and depending on the state that we start the chain we may converge to either of them.

Now, let us study the stationary distribution of reversible chains. We need to find a vector π such that for all i

$$\sum_j P_{j,i}\pi_j = \pi_i.$$

By the definition of P it is enough to have π such that

$$\sum_j \frac{w_{i,j}}{d_w(j)}\pi_j = \pi_i.$$

So, it is enough to let $\pi_j \propto d_w(j)$. More precisely set

$$\pi_j = \frac{d_w(j)}{\sum_k d_w(k)}.$$

In this case,

$$\sum_j P_{j,i}\pi_j = \sum_j \frac{w_{i,j}}{d_w(j)} \cdot \frac{d_w(j)}{\sum_k d_w(k)} = \frac{d_w(i)}{\sum_k d_w(k)} = \pi_i.$$

14.1 Mixing Time

Mixing time is essentially the time it takes for the chain to reach or get close to the stationary distribution.

Definition 14.3 (Mixing Time)**.** *The mixing time of the chain corresponding to a graph G is the smallest time t such that for any starting distribution x,*

$$\left\|xP^t - \pi\right\|_1 \le 1/4.$$

As usual, $\|x - y\|_1 = \sum_i |x_i - y_i|$.

There is nothing special about the number $1/4$ in the above definition. As we will see, if the mixing time of a chain is t, then for any ϵ, and any starting state x,

$$\left\| xP^{t \log \frac{1}{\epsilon}} - \pi \right\|_1 \leq \epsilon.$$

The above definition is very strong. In particular, if the ℓ_1 distance of two probability distributions x, y is ϵ, then for any event $\mathcal{E} \subseteq V$,

$$|\mathbb{P}_x[\mathcal{E}] - \mathbb{P}_y[\mathcal{E}]| \leq \epsilon.$$

In the rest of this section we prove a strong bound on the mixing time using the second smallest eigenvalue of the normalized Laplacian of G.

Firstly, recall normalized adjacency matrix $\tilde{A} = D^{-1/2}AD^{-1/2}$ and the normalized Laplacian matrix $\tilde{L} = D^{-1/2}LD^{-1/2}$. In the previous lectures we showed that any eigenvalue λ of $\tilde{L}$ corresponds to an eigenvalue $1 - \lambda$ of $\tilde{A}$. Now, we see that any eigenvalue of $\tilde{A}$ is also an eigenvalue of P. In particular, assume λ is an eigenvalue of $\tilde{A}$ with eigenvector v. Then,

$$v\tilde{A} = vD^{-1/2}AD^{-1/2} = \lambda v$$

Multiply both sides by $D^{1/2}$ from the right.

$$\lambda D^{1/2}v = vD^{-1/2}A = vD^{-1/2}D^{-1}A = vD^{-1/2}P.$$

So, the vector $u = D^{-1/2}v$ is an eigenvector of P with eigenvalue λ. Let us summarize the above discussion. Since we said that the eigenvalues of $\tilde{A}$ are always in the interval $[-1, 1]$, we obtain the same holds for P. In particular,

$$1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq -1.$$

Furthermore, the stationary distribution, π, is an eigenvector of the eigenvalue 1.

Theorem 14.4. *Let $1 = \lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of P. Then, the mixing time is at most*

$$O\left(\frac{\log \min_i \frac{1}{\pi_i}}{1 - \max\{\lambda_2, |\lambda_n|\}}\right).$$

In particular, if G is d-regular, then the mixing time is

$$O\left(\frac{\log n}{1 - \max\{\lambda_2, |\lambda_n|\}}\right).$$

Here is a high-level intuition of the above bounds. λ_2 corresponds to how far G is from being disconnected. In particular, if G is disconnected $\lambda_2 = 1$ and the bound becomes infinity. λ_n measures how far G is from being a bipartite graph. If G is biparitte, $\lambda_n = -1$ and the chain never mixes.

There is a simple trick to get around the bipartiteness and λ_n in the statement of the above theorem. The idea is to make the chain *lazy*. For each vertex i we add loop with weight $d_w(i)$. It is not hard to see that this change preserves the stationary distribution of the chain and makes $\lambda_n \geq 0$. In the language of Markov chains, this means that at any state i with probability $1/2$ we wait and do nothing and with the remaining probability we follow the chain. It is not hard to see that the mixing time of the lazy chain is no more than twice the mixing time of the original chain.

Corollary 14.5. *Let $1 = \lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of a lazy random walk on G. The mixing time is at most*

$$O\left(\frac{\log \min_i \frac{1}{\pi_i}}{1 - \lambda_2}\right)$$

As we mentioned before $1-\lambda_2$ is equal to the second smallest eigenvalue of $\tilde{L}$. It is also known as the *spectral gap* of the chain. Now, we can use Cheeger's inequality to lower bound $1-\lambda_2$. By Cheeger's inequality,

$$\phi(G) \leq \sqrt{2(1-\lambda_2)} \Rightarrow (1-\lambda_2) \geq \phi(G)^2/2.$$

Corollary 14.6. *The mixing time of the lazy random walk on any graph G is at most*

$$O\left(\frac{\log \min_i \frac{1}{\pi_i}}{\phi(G)^2}\right).$$

In particular, if G is d-regular, then the mixing time is $O(\log n/\phi(G)^2)$.

For example, using the above theorem we have:

i) The mixing time of a cycle of length n is $O(n^2 \log n)$.

ii) The mixing time of a $\sqrt{n} \times \sqrt{n}$ grid is $O(n \log n)$.

iii) The mixing time of the complete graph K_n is $O(\log n)$.

iv) The mixing time of the hypercube $\{0,1\}^{\log n}$ is $O(\log^3 n)$.

In the rest of this section we prove Theorem 14.4. For the simplicity of the notation, we only prove the theorem for regular graph. In this case the stationary distribution π is the uniform distribution. Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of P with corresponding orthonormal eigenvectors $v^1, \ldots, v^n$. Instead of directly bounding the ℓ_1 norm we upper bound the ℓ_2 norm, i.e., we show

$$\left\|xP^t - \pi\right\|_2 \leq \frac{1}{4\sqrt{n}}. \tag{14.2}$$

Then, the theorem follows from the fact that

$$\left\|xP^t - \pi\right\|_1 \leq \sqrt{n}\left\|xP^t - \pi\right\|_2.$$

The proof is very similar to the proof of the power method.

Let x be probability distribution vector on V. We can write,

$$x = \sum_i \langle x, v_i\rangle v^i = \sum_i a_i v^i,$$

for $a_i = \langle x, v^i\rangle$. Therefore, we can write

$$x^T P^t = \sum_i a_i \lambda_i^t v^i$$

Now, we show $a_1\lambda_1^t v_1 = \pi_1$. Firstly, $\lambda_1^t = \lambda_1 = 1$. Secondly, $v_i^1 = 1/\sqrt{n}$ for all i. So,

$$a_1\langle x, v^1\rangle = \sum_i x_i \cdot \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{n}} \cdot \|x\|_1 = \frac{1}{\sqrt{n}},$$

where the last equality uses that x is a probability distribution. Therefore,

$$a_1\lambda_1^t v_1 = \frac{1}{\sqrt{n}} \cdot v_1 = \pi.$$

So, we can write,

$$x^T P^t - \pi = \sum_{i=1}^{n} a_i \lambda_i^t v^i - \pi = \sum_{i=2}^{n} a_i \lambda_i^t v^i.$$

Let $\lambda^* = \max\{\lambda_2, |\lambda_n|\}$. By the orthonormality of v^i's we have

$$\begin{aligned} \left\|x^T P^t - \pi\right\|_2^2 &\leq \left\|\sum_{i=2}^{n} a_i \lambda_i^t v^i\right\|_2^2 \\ &= \sum_{i=2}^{n} a_i^2 \lambda_i^{2t} \\ &\leq \sum_{i=2}^{n} a_i^2 \lambda^{*2t} \\ &\leq \|x\|^2 \lambda^{*2t} \leq \lambda^{*2t}. \end{aligned}$$

So, for $t = O(\frac{\log n}{1-\lambda^*})$, we have

$$\left\|x^T P^t - \pi\right\|_2^2 \leq 1/n,$$

which proves (14.2). This completes the proof of Theorem 14.4.