Theoretician's Toolkit **Spring 2019**

Lecture 4: Spectral Graph Theory 2

Scribes: Robert Wang and Alex Yu March 11

In the previous week, we learned about the variational characterization of eigenvalues and some applications of the graph adjacency and Laplacian matrices. We will now use this knowledge in the analysis of Markov chains and in particular mixing times. The bounds on mixing times we derive will serve as a theoretical foundation and justification for the popular Markov Chain Monte Carlo (MCMC) technique.

4.1 Markov Chain Review

Definition 4.1. *(Markov Chain) A Markov chain is a sequence of random variables $X_0, X_1, \ldots$ with the Markov property:*

$$P_{ij} \triangleq \Pr[X_t = j | X_{t-1} = i, \ldots, X_0] = \Pr[X_t = j | X_{t-1} = i] \qquad \text{For } i, j \in [n]$$

where n is the number of states.

A finite Markov chain may be completely specified by its transition matrix, $\mathbf{M} \triangleq (P_{ij})$. Letting $\pi^{(t)}$ be the distribution of X_t at time t, we have $\pi^{(t+1)} = \pi^{(t)}\mathbf{M}$. Note that following convention, we define $\pi^{(t)}$ as a row vector.

Finite Markov chains are naturally representable as graphs with states as vertices and edges labelled with transition probabilities. As such, they may be studied using spectral graph theory.

4.1.1 Stationary Distributions

Definition 4.2. *(Stationary Distribution) A distribution π for the Markov Chain M is stationary if $\pi\mathbf{M} = \pi$.*

It turns out that not all Markov chains has a stationary distribution. But we can say:

Theorem 4.3. *(Fundamental Theorem of Markov Chains) Any irreducible, aperiodic Markov chain converges to a unique stationary distribution, π. Formally, if M is*

1. *irreducible: $\forall i, j,\ \exists t,\ \mathbf{M}^t_{i,j} > 0$, and*
2. *aperiodic: $\forall i,\ \exists j,\ gcd\{t, \mathbf{M}^t_{i,j} > 0\} = 1$*

then for every initial distribution $\pi^{(0)}$, $\lim_{t\to\infty} \pi^{(0)}\mathbf{M}^t = \pi$.

In graph theoretic terms, the graph associated with M must be strongly connected and non-bipartite.

The aperiodic (non-bipartite) condition may seem difficult to ensure, but in fact we can make any Markov aperiodic using the following method:

Lemma 4.4. *Given any Markov chain with transition matrix* $\mathbf{M}$*, we can construct an aperiodic variant* $\mathbf{M'}$ *by adding a self loop at every vertex with transition probability* 0.5*, and also scaling all existing probabilities by* 0.5*. In other words,* $\mathbf{M'} = \frac{1}{2}(\mathbf{M} + \mathbf{I})$.

Proof. Following the construction, at every time step t after the first visit t'_i to vertex i, $\forall i, \exists j \mathbf{M}^t(i,j) > 0$, since there is a self loop back to i. Naturally, $\forall i,\ \exists j,\ gcd\{t, \mathbf{M}^t_{i,j} > 0\} = 1$ □

4.2 Mixing Times

Definition 4.5. *(Mixing Time) The mixing time of a Markov chain* M *with a unique stationary distribution* π *is the first time* t *such that* $\forall \pi^{(0)},\ |\pi^{(0)}\mathbf{M}^t - \pi|_1 \leq \frac{1}{4}$*. (The choice* $1/4$ *is arbitrary and only for algebraic convenience)*

Above, $|\cdot|_1$ indicates the ℓ_1-norm: $|\pi|_1 \triangleq \sum_{i=1}^n \pi_i$. The ℓ_1-norm, also known as the *variation distance*, is used because of the following property:

Lemma 4.6. *For any distributions* π, π' *on* $[n]$*, let* $\pi(S) \triangleq \sum_{i \in S} \pi_i$*. Then:*

$$|\pi - \pi'|_1 = 2 \max_{S \subseteq [n]} |\pi(S) - \pi'(S)|$$

Note that intuitively, the right hand side is the maximum difference between the probabilities assigned by π, π' to any event.

Proof.

$$\begin{aligned}
|\pi - \pi'|_1 &= \sum_{i=1}^n |\pi_i - \pi'_i| \\
&= \sum_{i:\pi_i \geq \pi'_i} (\pi_i - \pi'_i) + \sum_{i:\pi_i < \pi'_i} (\pi'_i - \pi_i) \\
&= \sum_{i:\pi_i \geq \pi'_i} (\pi_i - \pi'_i) + \left(1 - \sum_{i:\pi_i \geq \pi'_i} \pi'_i\right) - \left(1 - \sum_{i:\pi_i \geq \pi'_i} \pi_i\right) \\
&= \sum_{i:\pi_i \geq \pi'_i} (2\pi_i - 2\pi'_i) = 2 \sum_{i:\pi_i \geq \pi'_i} (\pi_i - \pi'_i) \\
&= 2 \max_{S \subseteq [n]} \left| \sum_{i:i \in S} (\pi_i - \pi'_i) \right| \\
&= 2 \max_{S \subseteq [n]} |\pi(S) - \pi'(S)|
\end{aligned}$$

The second-to-last step follows from the fact that adding any index i such that $\pi_i > \pi'_i$ will increase the sum of distances, while adding one such that $\pi_i < \pi'_i$ will decrease the sum, meaning $S = \{i \in [n] \,|\, \pi_i \geq \pi'_i\}$ results in the maximum. □

4.2.1 Bounding Mixing Times

For simplicity, we will limit our analysis to Markov chains defined by d-regular undirected graphs as in the previous lecture. Note that this means $P[X_t = j|X_{t-1} = i] = P[X_t = i|X_{t-1} = j]$. Since M is symmetric, it has eigenvalues $\lambda_1 \geq \lambda_2 \geq ...\lambda_n$, and orthogonal eigenvectors $\mathbf{e_1}, ...\mathbf{e_n}$ such that $v_i\mathbf{M} = \lambda_i\mathbf{e_i}$. In general, any eigenvalue of a markov chain must have absolute value less than or equal to 1.

Proof. Suppose for sake of contradiction that $\mathbf{M}$ has an eigenvalue λ with $|\lambda| > 1$. Then λ is also an eigenvalue of $\mathbf{M}^\top$. Let v be a corresponding eigenvector of $\mathbf{M}^\top$ and let i be the index of the item in v with the largest absolute value, i.e. $i = \text{argmax}_i|v_i|$. Then observe v_i must be nonzero, and $|(v\mathbf{M}^\top)_i| = |v \cdot \mathbf{M}_i| = |\lambda v_i| > |v_i|$, where $\mathbf{M}_i$ is the ith row of $\mathbf{M}$. But since the rows of $\mathbf{M}$ sum to 1 and are non-negative, $v \cdot \mathbf{M}_i$ is a just a linear combination of the entries of v, and since v_i has the largest absolute value by definition, $|v \cdot M_i| \leq |v_i|$. $\rightarrow\leftarrow$ □

Theorem 4.7. *A Markov Chain on a d-regular graph with a unique stationary distribution and transition maxtrix* $\mathbf{M}$ *with eigenvalues* $1 \geq \lambda_1 \geq ...\lambda_n \geq -1$ *and has mixing time at most* $O\left(\frac{\log n}{1-\lambda_{max}}\right)$*, where* $\lambda_{max} \triangleq \max\{\lambda_2, |\lambda_n|\}$ *is the eigenvalue with the largest absolute value among all eigenvalues of* $\mathbf{M}$ *other than* λ_1*.*

To give some intuition, consider the weighted Laplacian, $\mathcal{L} \triangleq \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$ of the graph ($\mathbf{L}, \mathbf{D}$ are the Laplacian and degree matrices respectively):

$$\mathcal{L} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{D}^{-1/2}\left(\mathbf{D} - \mathbf{A}\right)\mathbf{D}^{-1/2} = \mathbf{I} - \frac{1}{d}\mathbf{A} = \mathbf{I} - \mathbf{M}$$

This means that the eigenvalues of $\mathcal{L}$ are $\mu_1 \leq \mu_2, ... \leq \mu_n$ (note the ordering), where $\mu_i = 1 - \lambda_i$:

Proof. Suppose $\det(\mathbf{M} - \lambda_i\mathbf{I}) = 0$. Then $\det(\mathcal{L} - \mu_i\mathbf{I}) = \det(\mathbf{I} - \mathbf{M} - (1 - \lambda_i)\mathbf{I}) = \det(-(\mathbf{M} - \lambda_i\mathbf{I})) = 0$ □

This means that $\lambda_{max} = \max\left(|1 - \mu_2|, |1 - \mu_n|\right)$, and $1 - \lambda_{max} = \min\left(\mu_2, 2 - \mu_n\right)$. Thus the mixing time, t, is $O\left(\frac{\log n}{\min(\mu_2, 2-\mu_n)}\right)$. Recall that μ_2, the second eigenvalue of $\mathbf{L}$, is a measure of how well connected a graph is, while it turns out μ_n is a measure of bipartiteness, where $\mu_n = 2$ when the graph is bipartite. Our bound tells us that the more well connected and less bipartite a graph is, the faster the mixing time is.

Now, we proceed to prove Theorem 4.7.

Proof. Let M be a Markov chain on a d-regular symmetric directed graph. Let $\mathbf{A}$ be G's adjacency matrix. Then observe that the transition matrix

$$\mathbf{M} = \frac{1}{d}\mathbf{A}$$

Since G is d-regular, the uniform distribution $\frac{1}{n}\mathbf{1}$ is a stationary distribution, so it must be the unique stationary distribution π.

Further, since $\mathbf{M}$ is symmetric, by the spectral theorem, its eigenvectors form an orthogonal basis. Thus, any initial distribution $\pi^{(0)}$ can be written as the sum of projections onto the eigenvectors:

$$\pi_0 = \sum_{i=1}^{n} \frac{\pi^0 \cdot \mathbf{e_i}}{\mathbf{e_i} \cdot \mathbf{e_i}} \mathbf{e_i}$$

Let $\alpha_i \triangleq \frac{\pi^{(0)} \cdot \mathbf{e_i}}{\mathbf{e_i} \cdot \mathbf{e_i}}$. We know that $\mathbf{e_1} = \mathbf{1}$, and since $\pi^{(0)}$ is a probability vector, $\mathbf{1} \cdot \pi^{(0)} = 1$. Thus, we have

$$\pi^{(0)} = \frac{1}{n}\mathbf{1} + \sum_{i=2}^{n} \alpha_i \mathbf{e}_i$$

Thus, π^t, the probability at time t, can be expressed as:

$$\begin{aligned}
\pi^{(0)}\mathbf{M}^t &= \left(\frac{1}{n}\mathbf{1} + \sum_{i=2}^{n} \alpha_i \mathbf{e}_i\right)\mathbf{M}^t \\
&= \frac{1}{n}\mathbf{1} + \sum_{i=2}^{n} \alpha_i \lambda_i^t \mathbf{e}_i \\
\left|\pi^{(0)}\mathbf{M}^t - \frac{1}{n}\mathbf{1}\right|_2 &= \left|\sum_{i=2}^{n} \alpha_i \lambda_i^t \mathbf{e}_i\right|_2 \\
&\leq \lambda_{max}^t \left|\sum_{i=2}^{n} \alpha_i \mathbf{e}_i\right|_2 \\
&\leq \lambda_{max}^t \left|\pi^{(0)}\right|_2 \\
&\leq \lambda_{max}^t \left|\pi^{(0)}\right|_1 = \lambda_2^t \\
\left|\pi^{(0)}\mathbf{M}^t - \frac{1}{n}\mathbf{1}\right|_1 &\leq \sqrt{n}\left|\pi^{(0)}\mathbf{M}^t - \frac{1}{n}\mathbf{1}\right|_2 \\
&\leq \sqrt{n}\lambda_{max}^t
\end{aligned}$$

The second to last inequality follows from Cauchy-Schwartz. Recalling that $\frac{1}{n}\mathbf{1}$ is the stationary distribution, by definition 4.5, the mixing time is the smallest t such that: $\left|\pi^{(0)}\mathbf{M}^t - \frac{1}{n}\mathbf{1}\right|_1 \leq \sqrt{n}\lambda_{max}^t$. If $t = \frac{c \log n}{1-\lambda_{max}}$, and $\lambda_{max} < 1$, then we have:

$$\begin{aligned}
\left|\pi^{(0)}\mathbf{M}^t - \frac{1}{n}\mathbf{1}\right|_1 &\leq \sqrt{n}(\lambda_{max})^{\frac{c \log n}{1-\lambda_{max}}} \\
&= n^{\frac{c \log \lambda_{max}}{1-\lambda_{max}} + \frac{1}{2}} \\
&= O\left(\frac{1}{n^c}\right)
\end{aligned}$$

Note, that if t goes up by a constant factor, the l_1 distance between the current distribution and the stationary distribution decreases exponentially. □

4.2.2 Mixing Time of Cycle Graphs

Next, we show how Theorem 4.7 may be applied to the analysis of the mixing time on an n-vertex cycle, which is 2-regular. To be more precise, we let G' be the undirected n-vertex cycle graph without added self-loops, and G be the same as G' but with self loops added as in Lemma 4.4.

Theorem 4.8. *The mixing time of the Markov chain on G, the n-vertex cycle graph with added self-loops, is:*

$$t = O(n^2 \log n)$$

Before we prove this theorem, recall the definitions of conductance and Cheeger's inequalities:

Definition 4.9. *(Conductance for regular graphs) Let $E(S, \overline{S})$ be the number of edges crossing the cut $(S, \overline{S})$. Then the conductance ϕ_G of a d-regular graph G is given by:*

$$\phi_G \triangleq \min_{\substack{S \subset V \\ |S| \leq \frac{|V|}{2}}} \frac{E(S, \overline{S})}{d|S|}$$

Theorem 4.10. *(Cheeger's Inequalities) Let μ_2 be the second-smallest eigenvalue of the normalized Laplacian matrix, $\mathcal{L}$, of G. Then:*

$$2\phi_G \geq \mu_2 \geq \frac{\phi_G^2}{2}$$

We begin by determining $\phi_{G'}$:

Claim 4.11.

$$\phi_{G'} = \frac{2}{n}$$

Proof. Let $v_1, \ldots, v_n$ be the vertices of G', where v_i is connected to v_{i+1} for all $1 \leq i < n$, and v_1 is connected to v_n. But clearly, setting $S = \{v_1, \ldots, v_{\lfloor n/2 \rfloor}\}$ minimizes $\frac{E(S,\overline{S})}{d|S|}$, since the numerator i.e. the number of edges crossing any cut must be at least 2, and the denominator i.e. the number of vertices in one set can be at most $\lfloor |S|/2 \rfloor$. Thus $\phi_{G'} = \frac{2}{2\frac{n}{2}} = \frac{2}{n}$ □

We now proceed to prove Theorem 4.8.

Proof. Let $\mathcal{L}, \mathcal{L}'$ be the normalized Laplacian matrices of G, G', with eigenvalues μ_i, μ_i' respectively. Let $\mathbf{M} = \mathbf{I} - \mathcal{L}, \mathbf{M}' = \mathbf{I} - \mathcal{L}'$ be the respective transition matrices with eigenvalues λ_i, λ_i'.

Applying Cheeger's inequalities to claim 4.11, we see $\mu_2' \geq \frac{\phi_{G'}^2}{2} = \frac{2}{n^2}$. But since $\mathbf{M}' = \mathbf{I} - \mathcal{L}'$, we have $\lambda_2' \leq 1 - \frac{2}{n^2}$. Furthermore, by lemma 4.4, $\mathbf{M} = \frac{\mathbf{M}'+\mathbf{I}}{2}$, so this means $\lambda_2 = \frac{\lambda_2'+1}{2} \leq 1 - \frac{1}{n^2}$.

Applying theorem 4.7 (the mixing time bound):

$$\begin{aligned} t &= O\left(\frac{\log n}{1-\lambda_{max}}\right) \\ &= O\left(\frac{\log n}{1-\lambda_2}\right) \\ &= O\left(\frac{\log n}{1-\left(1-\frac{1}{n^2}\right)}\right) \\ &= O\left(n^2 \log n\right) \end{aligned}$$

n.b. the second step follows from that since the eigenvalues of $\mathbf{M}'$ are in the interval $[-1, 1]$, those of $\mathbf{M} = \frac{\mathbf{M}'+\mathbf{I}}{2}$ must be in $[0, 1]$, so $\lambda_{max} = \lambda_2$. □

4.3 Markov Chain Monte Carlo

4.3.1 0-1 Knapsack Counting

The 0-1 knapsack problem is a well-known **NP**-complete problem. We will consider the problem of finding the number of solutions to $0-1$ knapsack, which turns out to be difficult:

Definition 4.12. *(0-1 Knapsack Counting Problem) Find the size of A, the set of vectors* $(x_1, \ldots x_n)$ *that satisfy* $\sum_{i=1}^{n} a_i x_i \leq b$*, for some given* $a_1 \ldots a_n$ *and* $b \in \mathbb{Z}^+$*,*

By a theorem of Valiant from the 1970s, the above is #**P**-complete ($\#P$ is the complexity class corresponding to finding the number of solutions to an **NP** problem), so a polynomial time algorithm would imply **P** = **NP**. However, below we will see that using sampling on Markov Chains, we can achieve a good approximation.

Definition 4.13. *(FPRAS) A fully polynomial random approximation scheme (FPRAS) is a randomized algorithm that finds an answer correct within* $(1+\epsilon)$ *with probability* $(1-\delta)$*, in* poly $\left(\frac{n}{\epsilon} \log \frac{1}{\delta}\right)$ *time.*

Definition 4.14. *(Almost-uniform) Let* π *be the uniform distribution, i.e.,* $\frac{1}{n}\mathbf{1}$*. We say a distribution* π' *is almost-uniform if:*

$$|\pi' - \pi|_1 \leq \exp(-n^2)$$

Theorem 4.15. *(Jerrum-Valiant-Vazirani) A FPRAS exists for the knapsack counting problem iff there exists a method to sample almost-uniformly from the 0-1 knapsack solution set, A, in polynomial time.*

For simplicity, we will limit our analysis to proving that counting and uniform sampling are equivalent in the exact, as opposed to approximate, case.

Proof. $\Longleftarrow$. Suppose there exists an algorithm for sampling uniformly from A. Let

$$A_i = \left\{ (x_{i+1}, \ldots, x_n) \middle| \sum_{j=i+1}^{n} a_j x_j \leq b \right\}$$

Then we can write $|A|$ as:

$$|A| = \frac{|A_0|}{|A_1|}\frac{|A_1|}{|A_2|}\cdots\frac{|A_{n-1}|}{|A_n|}|A_n|$$

Note that $|A_n| = 1$, since the only solution is the empty vector, and also $A_0 = A$.

Now observe that for each A_i,

$$\begin{aligned}
|A_i| &= \left|\left\{(x_{i+1},\ldots,x_n)\,\middle|\,\sum_{j=i+1}^{n} a_j x_j \le b\right\}\right| \\
&= \left|\left\{(x_i,\ldots,x_n)\,\middle|\,x_i = 0\,\middle|\,\sum_{j=i+1}^{n} a_j x_j \le b\right\}\right| \\
&= \left|\,\{(x_i,\ldots,x_n) \in A_{i-1}\,|\,x_i = 0\}\,\right|
\end{aligned}$$

Thus, to compute each fraction $\frac{|A_{i-1}|}{|A_i|} \triangleq \frac{1}{p_i}$, we can use the sampling algorithm to uniformly draw several samples from A_{i-1} to find $p_i = \frac{|A_i|}{|A_{i-1}|} = \frac{\left|\{(x_i,\ldots,x_n)\in A_{i-1}\,|\,x_i=0\}\right|}{|A_{i-1}|}$, the proportion of items in A_{i-1} with $x_i = 0$.

The only issue with this algorithm is the edge case where p_i is very (exponentially) small, such that it is 0 for all polynomial numbers of samples. To fix this problem, when p_i is small we can replace $|A_i|$ with $|A_i'|$, where $A_i' = \{(x_i,\ldots,x_n) \in A_{i-1}\,|\,x_i = 1\}$. Now all sampling for ratios with $j \ge i$ should be done with modified volume $b' \triangleq b - a_{i+1}$.

$\Longrightarrow$. For the converse, assume we have a counting algorithm. Then for each bit x_i starting at x_1, set knapsack volume $b' \triangleq b - \sum_{j=1}^{i-1} a_j x_j$ and let q_i be the number of solutions to the knapsack instance with items $(x_{i+1},\ldots,x_n)$ divided by that for the instance with items $(x_i,\ldots,x_n)$. Then we can just output 0 with probability q_i and 1 with probability $1 - q_i$. □

Note that Knapsack count is not an anomaly: in fact, counting and sampling are generally equivalent, so this method is widely applicable.

We will now demonstrate a Markov Chain Monte Carlo method for sampling from A almost uniformly, which using the above theorem can be used to create an approximation algorithm for 0-1 knapsack counting. Construct a Markov Chain, M, whose states are the binary vectors $(x_1,\ldots,x_n)$ of the solution space, as follows:

Start at $(x_1,\ldots,x_n) = (0,\ldots,0)$. Say the current state is $\mathbf{x} \triangleq (x_1,\ldots,x_n) \in A$. Then:

- With 0.5 probability remain at the same vertex.
- Else, pick an index $i \in [n]$ uniformly at random. Let $\mathbf{y} := (x_1,\ldots,x_{i-1},1-x_i,x_{i+1},x_n)$. If $\mathbf{y} \in A$, go to $\mathbf{y}$. If not, remain at the same vertex.

Lemma 4.16. *The uniform distribution over A is a stationary distribution of M.*

Proof. For any two vectors $\mathbf{v}_1, \mathbf{v}_2 \in A$, $\mathbf{v}_1$ has an edge to $\mathbf{v}_2$ iff the vectors differ by exactly one bit. Then there must also be an edge from $\mathbf{v}_2$ to $\mathbf{v}_1$, and since all edges have transition probability $\frac{1}{2n}$, for the uniform distribution the "flow" out and into each vertex is equivalent. n.b. there is no edge to any vertex outside of A by construction. Thus one time step the uniform distribution must remain uniform. □

Note that M is aperiodic because every vertex has a self-loop (Lemma 4.4), and moreover is irreducible because every vector can be transformed to 0 by flipping all set bits, which would only decrease the total volume of the selected items, so every vertex is connected to 0. Thus M has a unique stationary state π by the fundamental theorem (4.3). From Lemma 4.16 above, it follows that the uniform distribution must be this unique stationary state.

Thus if we simulate M for enough steps and output the resulting state, we can almost uniformly sample a solution from A. But is this truly practical, and how many steps are needed? This is where mixing time analysis becomes relevant.

Theorem 4.17. *(Morris-Sinclair) The mixing time for M as defined above is $O(n^8)$.*

Unfortunately, the proof (32 pages) is omitted due to its length. For more information, see their 1999 paper. However, it is quite intriguing that this method provides such a close approximation to a $\#P$-complete problem. In the next section, we will exhibit a bound for d-regular graphs.