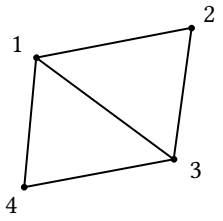
[CS1961: Lecture 15] Reversible Markov Chain, Expansion

Instructor: Chihao Zhang;
Scribed by Yuchen He

1 Markov Chain

1.1 Random Walk on Undirected Graph

Consider a random walk on the following undirected graph. We start at $X_0 = 1$ and move to a neighbor of the current vertex u.a.r. at each step. The distribution of the next position X_{t+1} is determined only by the current state. This random walk is a simple *Markov chain.*



Definition 1 (Markov Chain) *A sequence of random variables $X_0, X_1, \ldots, X_t, X_{t+1}, \ldots$ is a Markov chain if for any $t \in \mathbb{N}$ and any states $j_0, j_1, \ldots, j_t, j$,*

$$\mathbf{Pr}\left[X_{t+1} = j \mid X_t = j_t, X_{t-1} = j_{t-1}, \ldots, X_0 = j_0\right] = \mathbf{Pr}\left[X_{t+1} = j \mid X_t = j_t\right].$$

In this lecture, we only consider the time-homogeneous Markov chains with finite state space Ω. Such a Markov chain can be characterized by a matrix $P = (p_{ij})_{i,j\in\Omega} \in [0,1]^{\Omega\times\Omega}$ where $p_{ij} = \mathbf{Pr}\left[X_{t+1} = j \mid X_t = i\right]$. The transition matrix P is a stochastic matrix since $\sum_{j\in\Omega} p_{ij} = 1$ for all $i \in \Omega$. For example, in the above random walk, we have $\Omega = [4]$ and

$$p_{ij} = \begin{cases} 0, & \text{if } i \not\sim j \\ \frac{1}{\deg(i)}, & \text{if } i \sim j \end{cases}.$$

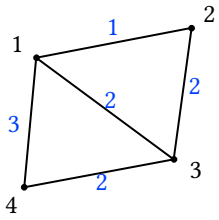
Sometimes we will simply denote the transition matrix P as the Markov chain for convenience.

Furthermore, we assume the Markov chains is *reversible*[1]. Then we can equate a Markov chain with a random walk on undirected weighted graphs. For a weighted graph $G = (V, E)$ with $w_{ji} = w_{ij} \geqslant 0$ for every $i, j \in V$,[2] let the transition probability p_{ij} be proportional to $w_{i,j}$. That is, $p_{ij} = \frac{w_{ij}}{d(i)}$ where $d(i) = \sum_{j\sim i} w_{ij}$. For example, in the following graph, we have $p_{12} = \frac{1}{1+2+3} = \frac{1}{6}$, $p_{13} = \frac{1}{3}$ and $p_{14} = \frac{1}{2}$.

For a transition matrix P, we can regard P as a directed graph G. When the Markov chain is reversible, G can actually be undirected by letting $w_{ij} = p_{ij} \cdot \pi(i)$ for each i, j.

[1] A Markov chain is reversible if there exists a distribution π on Ω such that $\pi(i)p_{ij} = \pi(j)p_{ji}$ for all $i, j \in \Omega$.

[2] $w_{ij} = 0$ iff $(i, j) \notin E$.



Note that given $X_t \sim \mu_t$, we can calculate $X_{t+1} \sim \mu_{t+1}$ since

$$\mu_{t+1}(j) = \sum_{i \in \Omega} \mathbf{Pr}\left[X_{t+1} = j \wedge X_t = i\right] = \sum_{i \in \Omega} \mu_t(i) p_{ij}.$$

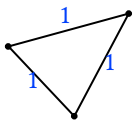
In brief, $\mu_{t+1}^T = \mu_t^T P = \mu_0^T P^{t+1}$. We will consider the following question about Markov chains: Will μ_t converge to some fixed distribution when t is sufficiently large?

1.2 Stationary Distribution

Definition 2 (Stationary Distribution) *A distribution π is a stationary distribution of P if it remains unchanged in the Markov chain as time progresses, i.e.,*

$$\pi^T P = \pi^T.$$

For example, it is easy to varify that $\pi^T = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ is a stationary distribution of the Markov chain corresponding to the following random walk.



Proposition 3 *For a weighted graph G, the distribution π that $\pi(i) \sim d(i)$ [3] is a stationary distribution of the corresponding Markov chain P.*

[3] Here $\pi(i) \sim d(i)$ means that each $\pi(i)$ is proportional to $d(i)$.

Proof. We have

$$w_{ij} = p_{ij} d(i) = w_{ji} = p_{ji} d(j)$$

by definition. Let π be a distribution on Ω that $\pi(i) = \frac{d(i)}{\sum_{j \in \Omega} d(j)}$. Then π satisifies the *detailed balance condition*

$$\forall i, j \in \Omega, \quad \pi(i) p_{ij} = \pi(j) p_{ji}.$$

Therefore,

$$\pi^T P(j) = \sum_{i \in \Omega} \pi(i) p_{ij} = \sum_{i \in \Omega} \pi(j) p_{ji} = \pi(j),$$

which indicates that π is a stationary distribution w.r.t. P. □

One of the major algorithmic applications of Markov chains is the *Markov chain Monte Carlo (MCMC)* method. It is a general method for designing an algorithm to sample from a certain distribution π. The idea of MCMC is

- First design a Markov Chain of which the stationary distribution is the desired π;
- Simulate the chain from a certain initial distribution for several steps and output the state.

Therefore, the following three basic questions regarding stationary distributions are important.

- If a Markov chain has a stationary distribution, is it unique?
- If the chain has a unique stationary distribution, does μ_t always converge to it from any μ_0?
- If μ_t always converges to the stationary distribution, what is the rate of convergence?

Card shuffling is a typical example of MCMC. The state space Ω is the set of $n!$ permutations of the n cards. We want the shuffling to output a permutation u.a.r.. In other words, we want the stationary distribution to be uniform distribution. How fast can we do this? Does this rely on the initial state?

2 The Spectrum of Markov Chains

2.1 Spectral Decomposition

Another advantage to use reversible chains is that their transition matrices are symmetric in some sense. Suppose P is reversible with respect to π. Let $\Pi = \operatorname{diag}(\pi)$ be the diagonal matrix with $\Pi(i,i) = \pi(i)$. Define $Q = \Pi^{\frac{1}{2}} P \Pi^{-\frac{1}{2}}$, then we can verify that Q is symmetric:

$$Q(i,j) = \pi(i)^{\frac{1}{2}} P(i,j) \pi(j)^{-\frac{1}{2}} = \pi(j)^{\frac{1}{2}} P(j,i) \pi(i)^{-\frac{1}{2}} = Q(j,i).$$

So we can apply the spectral decomposition theorem for Q, which yields

$$Q = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^T,$$

where $\lambda_1 \geqslant \cdots \geqslant \lambda_n$ are eigenvalues of Q with corresponding orthonormal eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$. If we let $\mathbf{v}_i \triangleq \Pi^{-\frac{1}{2}} \mathbf{u}_i$, then the above is equivalent to

$$P = \sum_{i=1}^{n} \lambda_i \Pi^{-\frac{1}{2}} \mathbf{u}_i \mathbf{u}_i^T \Pi^{\frac{1}{2}} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T \Pi.$$

We claim that $\lambda_1, \ldots, \lambda_n$ are eigenvalues of P with corresponding eigen-

vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. To see this, we have for any $j \in [n]$:

$$\begin{aligned} P\mathbf{v}_j &= \sum_{i=1}^{n} \lambda_i \Pi^{-\frac{1}{2}} \mathbf{u}_i \mathbf{u}_i^T \Pi^{\frac{1}{2}} \mathbf{v}_j \\ &= \sum_{i=1}^{n} \lambda_i \Pi^{-\frac{1}{2}} \mathbf{u}_i \mathbf{u}_i^T \Pi^{\frac{1}{2}} \Pi^{-\frac{1}{2}} \mathbf{u}_j \\ &= \lambda_j \mathbf{v}_j. \end{aligned}$$

Everything looks nice if we equip $\mathbb{R}^n$ with the inner product $\langle \cdot, \cdot \rangle_\Pi$ defined as $\langle \mathbf{x}, \mathbf{y} \rangle_\Pi = \mathbf{x}^T \Pi \mathbf{y} = \sum_{i=1}^{n} \pi(i) \mathbf{x}(i) \mathbf{y}(i)$. It is clear that $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are orthonormal with respect to the inner product:

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle_\Pi = \begin{cases} 0, & \text{if } i \neq j; \\ 1, & \text{if } i = j. \end{cases}$$

2.2 Fundamental Theorem for Reversible Markov Chains

We want that $\mu_0^T P^t \to \pi^T$ when $t \to \infty$. That is, for any initial distribution μ_0, we want $\mu_0 R = \pi^T$ if $P^t \to R$. Note that $\mu_0 R$ is the weighted average of the rows of R. Therefore, the only solution for R is that $R^T = [\pi\, \pi\, \cdots\, \pi]$.

We know that $\mathbf{1}^T P = \mathbf{1}^T$ since P is a stochastic matrix. Recall that λ_1 is no larger than the max absolute row sum. Therefore, we have $\lambda_1 = 1$ and $\mathbf{v}_1 = \mathbf{1}$. Since

$$P = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T \Pi,$$

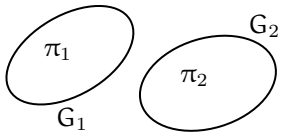
by direct calculation, we have

$$P^t = \sum_{i=1}^{n} \lambda_i^t \mathbf{v}_i \mathbf{v}_i^T \Pi = \mathbf{1} \cdot \mathbf{1}^T \pi^T + \sum_{i=2}^{n} \lambda_i^t \mathbf{v}_i \mathbf{v}_i^T \Pi.$$

Note that $\mathbf{1} \cdot \mathbf{1}^T \pi^T = [\pi\, \pi\, \cdots\, \pi]^T$. Therefore, $P^t \to R$ is equivalent with $\sum_{i=2}^{n} \lambda_i^t \nu_i \nu_i^T \Pi \to 0$. Since P is stochastic, we know that $|\lambda_i| \leqslant 1$ for all eigenvalues λ_i of P. Therefore, there are two ways to prohibit $\lim_{t\to\infty} \mu^T \left(\sum_{i=2}^{n} \lambda_i^t \nu_i \nu_i^T \Pi\right) = 0$, $\lambda_2 = 1$ or $\lambda_n = -1$. Recall that for a d-regular graph G, $\lambda_2 < 1$ and $\lambda_n > -1$ means that G is connected and is not bipartite. These conditions can be generalized to other weighted graphs.

Theorem 4 (Fundamental Theorem for Reversible Markov Chains) *For a finite Markov chain P which is reversible w.r.t. π, if the corresponding graph is connected and is not bipartite, then for any initial distribution μ_0, $\mu_0^T P^t \to \pi$ as $t \to \infty$.*

For example, when the graph is not connected, it can have more than one stationary distributions. When the graph is bipartite, although the stationary distribution might be unique, the Markov process will oscillate and never converge.



$\lambda_2 = 1$

$\lambda_n = -1$

2.3 Metropolis Algorithm

Given a distribution π over a state space Ω, how can we design a Markov chain P so that π is the stationary distribution of P? The *Metropolis algorithm* provides a way to achieve the goal as long as the transition graph G is connected and undirected.

Let Δ be the maximum degree of the transition graph except selfloop (that is $\Delta \triangleq \max_{u \in [n]} \sum_{v \neq u \in [n]} \mathbf{1}[(u,v) \in E]$). We describe the following process to construct a transition matrix P: Choose $k \in [\Delta + 1]$ uniformly at random. For any $i \in [n]$, let $\{j_1, j_2, \ldots, j_d\}$ be the d neighbours of i. We consider the transition at state i:

- If $d + 1 \leqslant k \leqslant \Delta + 1$, do nothing.
- If $k \leqslant d$,
  - propose to move from i to j_k.
  - accept the proposal with probability $\min\left\{\frac{\pi(j_k)}{\pi(i)}, 1\right\}$.

  Then the transition matrix is, for $i, j \in [n]$,

$$P(i,j) = \begin{cases} \frac{1}{\Delta+1} \min\left\{\frac{\pi(j)}{\pi(i)}, 1\right\}, & \text{if } i \neq j; \\ 1 - \sum_{k \neq i} P(i,k), & \text{if } i = j. \end{cases}$$

We can verify that P is reversible with respect to π:

$\forall i, j \in \Omega$:

$$\pi(i)P(i,j) = \pi(i) \cdot \frac{1}{\Delta+1} \min\left\{\frac{\pi(j)}{\pi(i)}, 1\right\} = \frac{\min\{\pi(i), \pi(j)\}}{\Delta+1} = \pi(j)P(j,i).$$

The advantage of the Metropolis algorithm is that we do not need to know π in order to implement the algorithm. We only need to know the quantity $\frac{\pi(j)}{\pi(i)}$, which is much easier to compute in many applications.

3 Relaxation Time and Expansion

3.1 Relaxation Time

By discussion above, if $\lambda_2 < 1$ and $\lambda_n > -1$, then

$$\lim_{t \to \infty} \mu^T \left(\sum_{i=2}^{n} \lambda_i^t v_i v_i^T \Pi \right) = 0.$$

The gap between 1 and the absolute value of these two eigenvalues also determines how fast P^t converges to $\mathbf{1} \cdot \mathbf{1}^T \pi^T$. Let $\lambda^* \triangleq \max\{|\lambda_2|, |\lambda_n|\}$, then

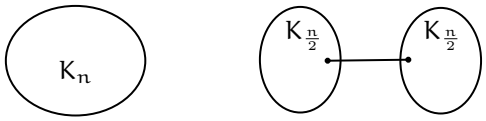
the *relaxation time* of P is defined to be

$$\tau_{\text{rel}} \triangleq \frac{1}{1-\lambda^*}.$$

For a reversible P, we can construct $P' = \frac{1}{2}(P + I)$. Note that the smallest eigenvalue of P', $\lambda_n(P')$, is non-negative. Therefore, we may only care about λ_2 when analyse the convergence rate.

3.2 Graph Expansion

Consider the following two graphs. Obviously, the Markov chain converges faster on K_n. In the second graph, if we start from certain vertex in the left $K_{\frac{n}{2}}$, it takes $\Theta(n^2)$ steps to go to the right part in expectation. The bottleneck significantly decrease the convergence rate.



Let P be a reversible chain. We let $G = (V, E)$ be its transition graph. Note that G is an undirected graph here. For any $S \subseteq V$,

$$Q(S, \overline{S}) \triangleq \sum_{i \in S, j \in V \setminus S} \pi(i) P(i, j).$$

Furthermore, we define the *expansion* of S as

$$\Phi(S) = \frac{Q(S, \overline{S})}{\pi(S)},$$

where $\pi(S) = \sum_{i \in S} \pi(i)$. Suppose $X_t \sim \pi$, then $\Phi(S) = \mathbf{Pr}[X_{t+1} \notin S \mid X_t \in S]$, which is the probability of escaping S. The expansion of P is the smallest $\Phi(S)$, i.e., $\Phi(P) = \min_{S \subseteq V : \pi(S) \leqslant \frac{1}{2}} \Phi(S)$. The expansion measures the difficulty of convergence. It can be bounded by the distance between 1 and λ_2.

Theorem 5 (Cheeger's Inequality) $\frac{1-\lambda_2}{2} \leqslant \Phi(P) \leqslant \sqrt{2(1-\lambda_2)}$.

Moreover, λ_2 also carries the information on how to partition G into the hardest $(S, \overline{S})$.