Basics of Random walks and Markov chains

-Summary-

Reference : Foundations of Data Science by A.Blum et al.

Terminologies

- Correspondence between terminologies of Random walks and Markov chains

Random walks	Markov chain
Graph	Stochastic process
Vertex	State
Strongly connected	Connected
Aperiodic	Aperiodic
Strongly connected and Aperiodic	Ergodic
Edge-weighted undirected graph	Time reversible

- Strongly connected : for any pairs of vertices x, y, the graph contains a path of directed edges from x to y.

Terminologies of Markov chains

- Transition probability p_{xy} : probability going from state x to state y (s.t $\sum_y p_{xy} = 1$)
- Transition probability matrix P : a matrix whose (x, y) entry is p_{xy}

- Connected : **Markov chain is connected** if the underlying directed graph is strongly connected.

- Persistent (state) : **A state of a Markov chain** is persistent if $p_{xx} = 1$
- Aperiodic : A connected Markov chain is aperiodic if the gcd of the lengths of directed cycles is one

Terminologies of Markov chains

- Procedure of RW:
  - Start a RW at a vertex x with a starting probability distribution p (where p is a row vector with non-negative values summing to 1 and with p_x being the probability of starting at vertex x)

  - Let $p(t)$ be a row vector with a component for each vertex specifying the probability mass of the vertex at time t.

  - Now $p(t)P = p(t+1)$ holds, where $[P]_{ij} = p_{ij}$ (transition probability from state i to j)

  - If the $\lim_{t\to\infty} p(t)$ exists, then the random walk is "stationary" and $p(t)$ is called "stationary (probability) distribution"

Theorems of Markov chains

- Let $p(t)$ be the probability distribution at step t of RW.
- Define “long-term average probability distribution”:

$$a(t) = \frac{1}{t}\big(p(0) + p(1) + \cdots p(t-1)\big)$$

Theorem 1 (Fundamental Theorem of Markov Chains)

For a connected Markov chain, there is a unique probability vector π satisfying:

$$\pi P = \pi$$

Moreover, for any starting distribution, $\lim_{t\to\infty} a(t)$ exists and equals to π. (stationary distribution)

Theorems of Markov chains

Lemma 1

For a random walk on a strongly connected graph with probabilities on the edges, ($\Leftrightarrow$ Connected MC), if the non-negative vector $\pi = (\pi_i)$ satisfies:

$$\pi_i p_{ij} = \pi_j p_{ji} \text{(Detailed balanced equation)}$$

for all i and j and $\sum_i \pi_i = 1$, then π is the stationary distribution of the walk.

- If the Markov chain satisfies "Detailed balanced equation", then the chain is "time-reversible chain".

- We can easily show that MH step is necessary for MH algorithm to satisfy the detailed balanced equation, also Gibbs sampling does not require MH step.

Convergence of Random Walks on Undirected Graphs

- Is there a natural way to define the convergence speed of random works to its stationary distribution π? $\rightarrow$ ϵ-mixing

Definition 1 [ϵ-mixing]

Fix $\epsilon > 0$, The ϵ-mixing time of a Markov chain is the minimum integer t such that for any starting distribution,

$$\|a(t) - \pi\|_1 = \sum_i |a(t)_i - \pi_i| < \epsilon$$

- Can we bound the ϵ-mixing time using some information on random walk on an undirected graph? $\rightarrow$ yes, using the concept of normalized conductance.

Convergence of Random Walks on Undirected Graphs

Definition 2 [Normalized conductance]

For a subset S of vertices, let $\pi(S) \coloneqq \sum_{x \in S} \pi_x$. The normalized conductance $\Phi(S)$ of S is

$$\Phi(S) = \frac{1}{\min\big(\pi(S), \pi(S^C)\big)} \cdot \sum_{(x,y) \in S \times S^C} \pi_x p_{xy}$$

The normalized conductance of the Markov chain, denoted Φ is defined by

$$\Phi = \min_{\emptyset \neq S \subset V} \Phi(S)$$

Theorem 3

The ϵ-mixing time of a random walk on an undirected graph is

$$O\left(\frac{\log(1/\pi_{min})}{\Phi^2 \epsilon^3}\right)$$

where π_{min} is the minimum stationary probability of any state.

Convergence of Random Walks on Undirected Graphs

- From Theorem3, we require to maximize the Φ (normalized conductance of the chain) to reduce the mixing time (or the time for our chain to approximately close to target distribution)

- Note that $\Phi(S) = \frac{1}{\pi(S)} \sum_{(x,y) \in S \times S^c} \pi_x p_{xy} = \sum_{x \in S} \frac{\pi_x}{\pi(S)} \sum_{y \in S^c} p_{xy} = \mathbb{P}(Y \in S^c | X \in S)$.
  (By assuming $\pi(S) < \frac{1}{2}$ without loss of generality)

- Hence, the theorem 3 implies that if the probability of transition from S to S^c tends to be low, then, the mixing time becomes larger which implies bad structure of MCMC sampling. (This may explain the reason why we try to improve the mixing in MCMC sampling)