Lecture 9: Random Walks and Markov Chain (Chapter 4 of Textbook B)

Jinwoo Shin

AI503: Mathematics for AI

Roadmap

Roadmap

Introduction - Random walk

KAIST AI
Graduate School of AI

- A *random walk* is defined under a (directed) graph $G = (V, E)$ where V, E are sets of vertices and edges, respectively.
- Start with a probability vector $\mathbf{p(0)} \in [0,1]^V$, where $p_x(0)$ is the probability of starting at vertex $x \in G$.
- The probability vector at time $t+1$ is defined by the probability vector at time t, namely $\mathbf{p(t{+}1)} = \mathbf{p(t)}P$.
  - P_{ij} is the probability of the walk at vertex i selecting the edge to vertex j.
- The long-term average probability of being at a particular vertex is independent of the choice of $\mathbf{p(0)}$ if G is strongly connected.
  - The limiting probabilities are called *stationary probabilities*.
- This property allows to design an efficient sampling algorithm from a desired probability distribution, called "Markov Chain Monte Carlo (MCMC)".

Graph	Stochastic process
vertex	state
strongly connected	persistent
aperiodic	aperiodic
strongly connected and aperiodic	ergodic
undirected graph	time reversible

Table 1: Correspondence between terminology of random walks and Markov chains

- A Markov chain has a finite set of *states*.
- For each pair of states x and y, there is a *transition probability* p_{xy} of going from x to y where $\sum_y p_{xy} = 1$ for each x.
- The terms "random walk" and "Markov chain" are used interchangeably.

Roadmap

Stationary Distribution

Definition 1. Long-term average probability distribution

Let $\mathbf{p(t)}$ be the probability distribution after t steps of random walks. Then, *long-term average probability distribution* is defined by:

$$\mathbf{a(t)} = \frac{1}{t}(\mathbf{p(0)} + \mathbf{p(1)} + ... + \mathbf{p(t\text{-}1)}).$$

Theorem 1. Fundamental Theorem of Markov Chains

For a connected Markov chain, there is a unique probability vector $\boldsymbol{\pi} P = \boldsymbol{\pi}$. Moreover, for any starting distribution, $\lim_{t\to\infty} \mathbf{a(t)}$ exists and equals to $\boldsymbol{\pi}$.

Proof. See Page 80-81 of Textbook B.

- Theorem 1 will be used to prove the convergence of Markov Chain Monte Carlo (MCMC) algorithm.

Roadmap

KAIST AI Graduate School of AI

Markov Chain Monte Carlo (MCMC)

KAIST AI
Graduate School of AI

- Let $f : \mathbb{R}^d \to \mathbb{R}$ and a probability distribution $p(\mathbf{x})$ where $\mathbf{x} = (x_1, x_2, ..., x_d)$.
- We want to calculate

$$E(f) = \sum_{\mathbf{x}} f(\mathbf{x})p(\mathbf{x})$$

- The sample space grows exponentially on d.
- Therefore, explicit summation requires exponential time on d, which is not desirable.

Markov Chain Monte Carlo (MCMC)

- Let $f : \mathbb{R}^d \to \mathbb{R}$ and a probability distribution $p(\mathbf{x})$ where $\mathbf{x} = (x_1, x_2, ..., x_d)$.
- We want to calculate

$$E(f) = \sum_{\mathbf{x}} f(\mathbf{x})p(\mathbf{x})$$

- Markov Chain Monte Carlo (MCMC) method approximates the summation by a summation of a set of samples, where each sample $\mathbf{x}$ is selected with probability $p(\mathbf{x})$.
  - Metropolis-Hastings algorithm
  - Gibbs sampling
- We construct a Markov chain that has the desired distribution as its stationary distribution.
- By Theorem 1, the average of the function f over states seen in a sufficiently long run is a good estimate of $E(f)$.

Markov Chain Monte Carlo (MCMC)

KAIST AI
Graduate School of AI

- Recall the definition of long-term average probability distribution.

$$\mathbf{a(t)} = \frac{1}{t}[\mathbf{p(0)} + \mathbf{p(1)} + \ldots + \mathbf{p(t\text{-}1)}]$$

- Let γ be the average value of f at the states seen in a t step walk, which is an estimate of $E(f) = \sum_i f_i p_i$.
- The expected value of γ is calculated by:

$$E(\gamma) = \sum_i f_i (\frac{1}{t} \sum_{j=1}^{t} \text{Prob(walk is in state } i \text{ at time } j)) = \sum_i f_i a_i(t).$$

- Letting $f_{max} := max_x |f(x)|$, we have:

$$|E(f) - E(\gamma)| \leq f_{max} \sum_i |p_i - a_i(t)| = f_{max} ||\mathbf{p} - \mathbf{a(t)}||_1.$$

- If $\mathbf{p}$ is the stationary distribution, $E(\gamma)$ converges to $E(f)$ by the rate of convergence of the Markov chain to its steady state.

Markov Chain Monte Carlo (MCMC) - Metropolis-Hasting Algorithm

KAIST AI
Graduate School of AI

Algorithm 1. Metropolis-Hasting Algorithm

Let r be the maximum degree of any vertex in a connected undirected graph G. We define the transition matrix as follows:

$$p_{ij} = \frac{1}{r}\min(1, \frac{p_j}{p_i}) \quad \text{if} \quad i \neq j,$$

$$p_{ii} = 1 - \sum_{j \neq i} p_{ij}.$$

- At state i, we select neighbor j with probability $\frac{1}{r}$.
- For the selected j, if $p_i \leq p_j$, go to j. If $p_i > p_j$, go to j with probability p_j/p_i.
- Otherwise, stay at i.

- The Metropolis-Hasting Algorithm is a method to design a Markov chain whose stationary distribution is a given target distribution $\mathbf{p}$.

Lemma 1.

For a random walk on a strongly connected graph G with probabilities on the edges, if the vector $\boldsymbol{\pi}$ satisfies $\pi_x p_{xy} = \pi_y p_{yx}$ for all $x, y \in G$ and $\sum_x \pi_x = 1$, then $\boldsymbol{\pi}$ is the stationary distribution of the walk.

Proof. See Page 81 of Textbook B.

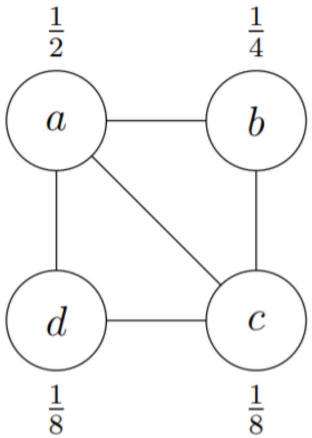
Theorem 2. Convergence of Metropolis-Hasting Algorithm

In algorithm 1 (Metropolis-Hasting Algorithm), the stationary probabilities are indeed p_i's.

Proof. The proof follows from Lemma 1.
As $p_i p_{ij} = \frac{1}{r}\texttt{min}(1, \frac{p_j}{p_i}) = \frac{1}{r}\texttt{min}(p_i, p_j) = \frac{p_j}{r}\texttt{min}(1, \frac{p_i}{p_j}) = p_j p_{ji}$, by Lemma 1., the theorem follows.

Markov Chain Monte Carlo (MCMC) - Metropolis-Hasting Algorithm

$p(a) = \frac{1}{2}$
$p(b) = \frac{1}{4}$
$p(c) = \frac{1}{8}$
$p(d) = \frac{1}{8}$



$a \to b \quad \frac{1}{3}\ \frac{1}{4}\ \frac{2}{1} = \frac{1}{6}$
$a \to c \quad \frac{1}{3}\ \frac{1}{8}\ \frac{2}{1} = \frac{1}{12}$
$a \to d \quad \frac{1}{3}\ \frac{1}{8}\ \frac{2}{1} = \frac{1}{12}$
$a \to a \quad 1 - \frac{1}{6} - \frac{1}{12} - \frac{1}{12} = \frac{2}{3}$

$b \to a \quad \frac{1}{3}$
$b \to c \quad \frac{1}{3}\ \frac{1}{8}\ \frac{4}{1} = \frac{1}{6}$
$b \to b \quad 1 - \frac{1}{3} - \frac{1}{6} = \frac{1}{2}$

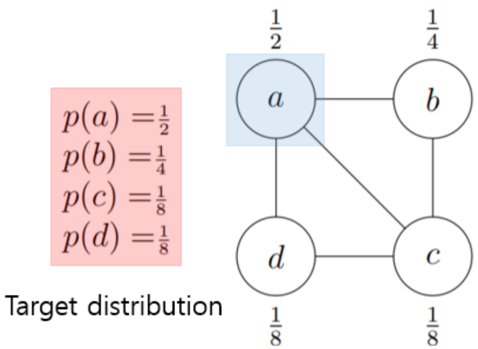
$c \to a \quad \frac{1}{3}$
$c \to b \quad \frac{1}{3}$
$c \to d \quad \frac{1}{3}$
$c \to c \quad 1 - \frac{1}{3} - \frac{1}{3} - \frac{1}{3} = 0$

$d \to a \quad \frac{1}{3}$
$d \to c \quad \frac{1}{3}$
$d \to d \quad 1 - \frac{1}{3} - \frac{1}{3} = \frac{1}{3}$

Figure 1: A connected undirected graph with vertices $\{a, b, c, d\}$

- Let's follow the Metropolis-Hasting Algorithm to construct the transition matrix P on the graph in Figure 1.

Markov Chain Monte Carlo (MCMC) - Metropolis-Hasting Algorithm



$a \to b \quad \frac{1}{3}\,\frac{1}{4}\,\frac{2}{1} = \frac{1}{6}$

$a \to c \quad \frac{1}{3}\,\frac{1}{8}\,\frac{2}{1} = \frac{1}{12}$

$a \to d \quad \frac{1}{3}\,\frac{1}{8}\,\frac{2}{1} = \frac{1}{12}$

$a \to a \quad 1-\frac{1}{6}-\frac{1}{12}-\frac{1}{12}=\frac{2}{3}$

$b \to a \quad \frac{1}{3}$

$b \to c \quad \frac{1}{3}\,\frac{1}{8}\,\frac{4}{1} = \frac{1}{6}$

$b \to b \quad 1-\frac{1}{3}-\frac{1}{6} = \frac{1}{2}$

$c \to a \quad \frac{1}{3}$

$c \to b \quad \frac{1}{3}$

$c \to d \quad \frac{1}{3}$

$c \to c \quad 1-\frac{1}{3}-\frac{1}{3}-\frac{1}{3} = 0$

$d \to a \quad \frac{1}{3}$

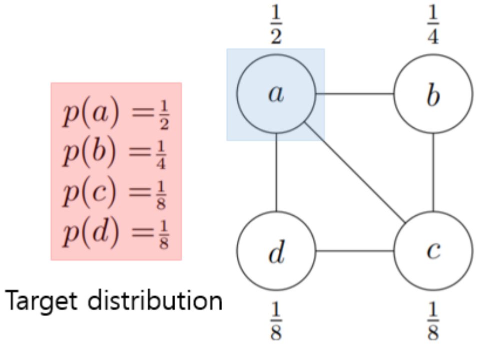
$d \to c \quad \frac{1}{3}$

$d \to d \quad 1-\frac{1}{3}-\frac{1}{3} = \frac{1}{3}$

Figure 2: A connected undirected graph with vertices $\{a, b, c, d\}$

- We want to make the stationary distribution $\pi = [\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}]$.
- For the vertex a,
  - $p_{ab} = \frac{1}{r}\mathtt{min}(1, \frac{p_b}{p_a}) = \frac{1}{3}\mathtt{min}(1, \frac{1}{2}) = \frac{1}{6}$
  - $p_{ac} = \frac{1}{r}\mathtt{min}(1, \frac{p_c}{p_a}) = \frac{1}{3}\mathtt{min}(1, \frac{1}{4}) = \frac{1}{12}$
  - $p_{ad} = \frac{1}{r}\mathtt{min}(1, \frac{p_d}{p_a}) = \frac{1}{3}\mathtt{min}(1, \frac{1}{4}) = \frac{1}{12}$
  - $p_{aa} = 1 - \sum_{j \neq a} p_{aj} = 1 - \frac{1}{6} - \frac{1}{12} - \frac{1}{12} = \frac{2}{3}$

Markov Chain Monte Carlo (MCMC) - Metropolis-Hasting Algorithm



$a \to b \quad \frac{1}{3}\,\frac{1}{4}\,\frac{2}{1} = \frac{1}{6}$

$a \to c \quad \frac{1}{3}\,\frac{1}{8}\,\frac{2}{1} = \frac{1}{12}$

$a \to d \quad \frac{1}{3}\,\frac{1}{8}\,\frac{2}{1} = \frac{1}{12}$

$a \to a \quad 1-\frac{1}{6}-\frac{1}{12}-\frac{1}{12}=\frac{2}{3}$

$b \to a \quad \frac{1}{3}$

$b \to c \quad \frac{1}{3}\,\frac{1}{8}\,\frac{4}{1} = \frac{1}{6}$

$b \to b \quad 1-\frac{1}{3}-\frac{1}{6} = \frac{1}{2}$

$c \to a \quad \frac{1}{3}$

$c \to b \quad \frac{1}{3}$

$c \to d \quad \frac{1}{3}$

$c \to c \quad 1-\frac{1}{3}-\frac{1}{3}-\frac{1}{3} = 0$

$d \to a \quad \frac{1}{3}$

$d \to c \quad \frac{1}{3}$

$d \to d \quad 1-\frac{1}{3}-\frac{1}{3} = \frac{1}{3}$

Figure 3: A connected undirected graph with vertices $\{a, b, c, d\}$

- We want to make the stationary distribution $\boldsymbol{\pi} = [\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}]$.
- Applying Metropolis-Hasting Algorithm for other vertices, we get:

$$P = \begin{bmatrix} 2/3 & 1/6 & 1/12 & 1/12 \\ 1/3 & 1/2 & 1/6 & 0 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 0 & 1/3 & 1/3 \end{bmatrix}$$

- It is easy to check $\boldsymbol{\pi} P = \boldsymbol{\pi}$.

Markov Chain Monte Carlo (MCMC) - Gibbs Sampling

Algorithm 2. Gibbs Sampling

Let G be an undirected graph whose vertices corresponds to the values $\mathbf{x} = (x_1, ..., x_d)$. Also, assume that there is an edge from $\mathbf{x}$ to $\mathbf{y}$ if $\mathbf{x}$ and $\mathbf{y}$ differ in only one coordinate (W.L.O.G. assume it the first coordinate). We define the transition matrix as follows:

$$p_{\mathbf{xy}} = \frac{1}{d} p(y_1 | x_2, x_3, ..., x_d),$$

$$\text{where} \quad d = \sum_i \sum_{y_i} p(y_i | x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_d).$$

- Choose one coordinate (randomly or sequentially).
- The transition probability depends on the conditional probability of chosen coordinate while fixing other coordinates.

Markov Chain Monte Carlo (MCMC) - Gibbs Sampling

KAIST AI
Graduate School of AI

- Gibbs sampling is a method to design a Markov chain whose stationary distribution is a given target distribution **p**.

Theorem 3. Convergence of Gibbs Sampling

In algorithm 2 (Gibbs Sampling), the stationary probability is indeed p.

Proof. The proof follows again from Lemma 1.

$$p_{\mathbf{xy}} = \frac{1}{d} \frac{p(\mathbf{y})}{p(x_2, x_3, ..., x_d)}$$

$$p_{\mathbf{yx}} = \frac{1}{d} \frac{p(\mathbf{x})}{p(x_2, x_3, ..., x_d)}$$

Therefore, $p(\mathbf{x})p_{xy} = p(\mathbf{y})p_{yx}$.

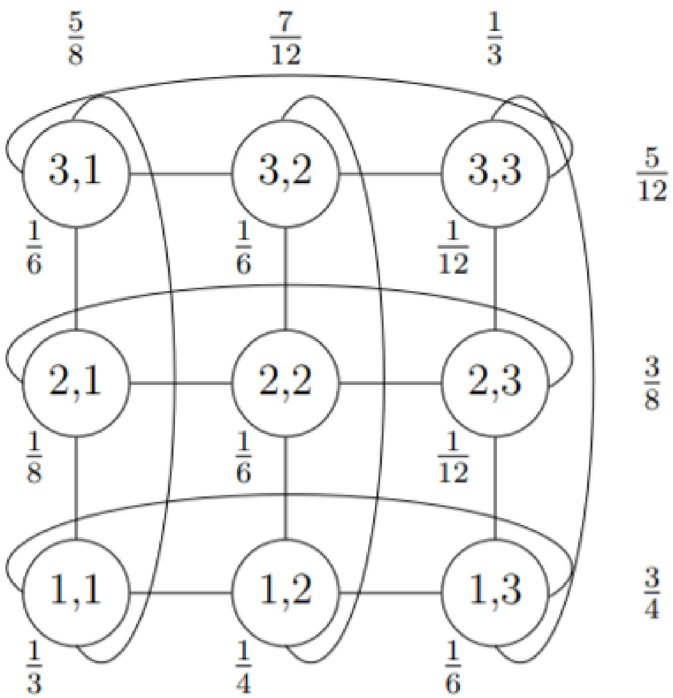
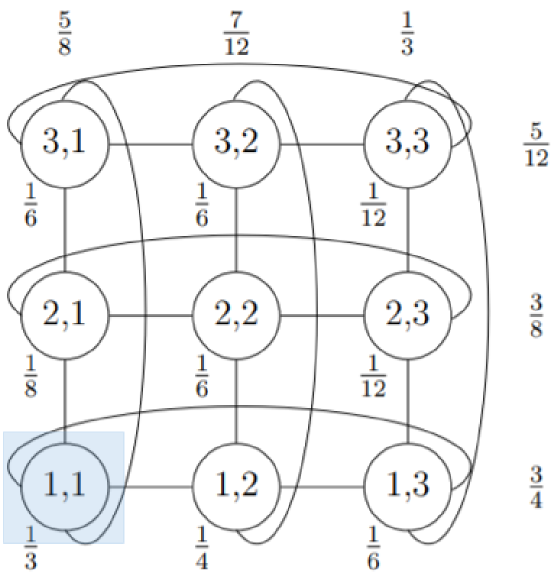


Figure 4: A connected undirected graph with vertices $\{(i,j) : i,j \in \{1,2,3\}\}$

- Let's follow the Gibbs sampling Algorithm to construct the transition matrix P on the graph in Figure 1.



Target distribution

$$p(1,1) = \tfrac{1}{3}$$
$$p(1,2) = \tfrac{1}{4}$$
$$p(1,3) = \tfrac{1}{6}$$
$$p(2,1) = \tfrac{1}{8}$$
$$p(2,2) = \tfrac{1}{6}$$
$$p(2,3) = \tfrac{1}{12}$$
$$p(3,1) = \tfrac{1}{6}$$
$$p(3,2) = \tfrac{1}{6}$$
$$p(3,3) = \tfrac{1}{12}$$

$$p_{(11)(12)} = \tfrac{1}{d} p_{12} / (p_{11} + p_{12} + p_{13}) = \tfrac{1}{2}\left(\tfrac{1}{4}\right) / \left(\tfrac{1}{3}\ \tfrac{1}{4}\ \tfrac{1}{6}\right) = \tfrac{1}{8} / \tfrac{9}{12} = \tfrac{1}{8}\ \tfrac{4}{3} = \tfrac{1}{6}$$

Calculation of edge probability $p_{(11)(12)}$

$p_{(11)(12)} = \frac{1}{2}\ \frac{1}{4}\ \frac{4}{3} = \frac{1}{6}$	$p_{(12)(11)} = \frac{1}{2}\ \frac{1}{3}\ \frac{4}{3} = \frac{2}{9}$	$p_{(13)(11)} = \frac{1}{2}\ \frac{1}{3}\ \frac{4}{3} = \frac{2}{9}$	$p_{(21)(22)} = \frac{1}{2}\frac{1}{6}\frac{8}{3} = \frac{2}{9}$
$p_{(11)(13)} = \frac{1}{2}\ \frac{1}{6}\ \frac{4}{3} = \frac{1}{9}$	$p_{(12)(13)} = \frac{1}{2}\ \frac{1}{6}\ \frac{4}{3} = \frac{1}{9}$	$p_{(13)(12)} = \frac{1}{2}\ \frac{1}{4}\ \frac{4}{3} = \frac{1}{6}$	$p_{(21)(23)} = \frac{1}{2}\frac{1}{12}\frac{8}{3} = \frac{1}{9}$
$p_{(11)(21)} = \frac{1}{2}\ \frac{1}{8}\ \frac{8}{5} = \frac{1}{10}$	$p_{(12)(22)} = \frac{1}{2}\ \frac{1}{6}\ \frac{12}{7} = \frac{1}{7}$	$p_{(13)(23)} = \frac{1}{2}\ \frac{1}{12}\ \frac{3}{1} = \frac{1}{8}$	$p_{(21)(11)} = \frac{1}{2}\frac{1}{3}\frac{8}{5} = \frac{4}{15}$
$p_{(11)(31)} = \frac{1}{2}\ \frac{1}{6}\ \frac{8}{5} = \frac{2}{15}$	$p_{(12)(32)} = \frac{1}{2}\ \frac{1}{6}\ \frac{12}{7} = \frac{1}{7}$	$p_{(13)(33)} = \frac{1}{2}\ \frac{1}{12}\ \frac{3}{1} = \frac{1}{8}$	$p_{(21)(31)} = \frac{1}{2}\frac{1}{6}\frac{8}{5} = \frac{2}{15}$

Figure 5: A connected undirected graph with vertices $\{(i,j) : i,j \in \{1,2,3\}\}$

Roadmap

KAIST AI
Graduate School of AI

(1) Introduction

(2) Stationary Distribution

(3) Markov Chain Monte Carlo (MCMC)

(4) Convergence of Random Walks on Undirected Graphs

(5) Random Walks on Undirected Graphs with Unit Edge Weights

Convergence of Random Walks on Undirected Graphs

- Given an edge-weighted undirected graph G, let w_{xy} denote the weight of the edge between nodes $x, y \in G$.
  - If no such edge exists, let $w_{xy} = 0$.
- Let $w_x = \sum_y w_{xy}$, $p_{xy} = w_{xy}/w_x$, and $w_{\texttt{total}} = \sum_{x'} w_{x'}$.
- By Lemma 1, $\pi_x := w_x/w_{\texttt{total}}$ are the stationary probabilities.
  - $w_x p_{xy} = w_x \frac{w_{xy}}{w_x} = w_{xy} = w_{yx} = w_y \frac{w_{yx}}{w_y} = w_y p_{yx}$.
  - Therefore, $(w_x/w_{\texttt{total}})p_{xy} = (w_y/w_{\texttt{total}})p_{yx}$.
- We are interested in the convergence rate of Metropolis-Hasting algorithm and Gibbs sampling on edge-weighted undirected graph.

Convergence of Random Walks on Undirected Graphs

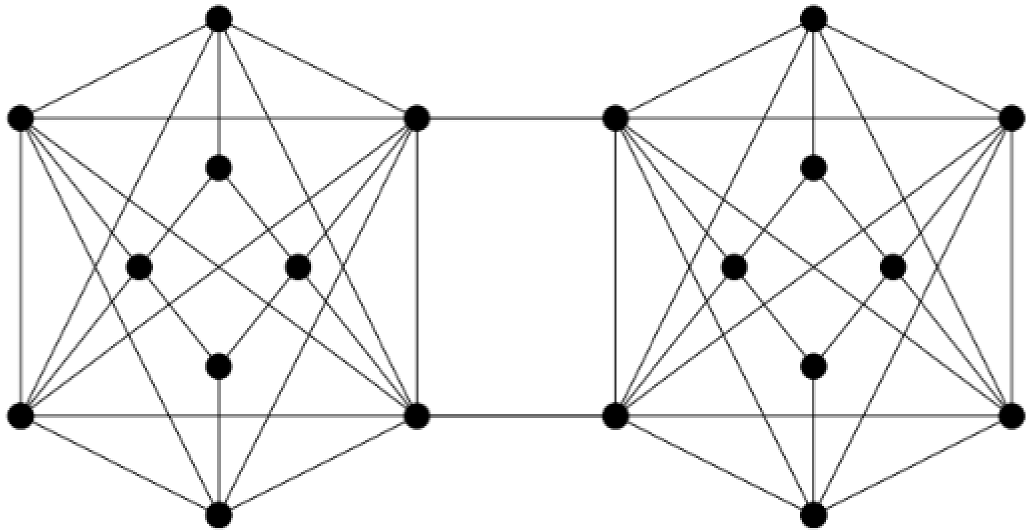


Figure 6: A connected graph. All edges have the same weight.

- In Figure 6, the random walk is unlikely to cross the narrow passage between the two halves.
- We will show that the time to converge is quantitatively related to the tightest constriction.

Definition 1. ε-mixing time

The ε-mixing time of a Markov chain is the minimum integer t such that for any starting distribution $\mathbf{p(0)}$, the 1-norm difference between the t-step *running average probability distribution* and the *stationary distribution* is at most ε.

Definition 2. Normalized conductance

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

$$\Phi(S) = \frac{\sum_{(x,y)\in(S,\bar{S})} \pi_x p_{xy}}{\min(\pi(S), \pi(\bar{S}))}$$

Definition 2. Normalized conductance

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

$$\Phi(S) = \frac{\sum_{(x,y) \in (S, \bar{S})} \pi_x p_{xy}}{\min(\pi(S), \pi(\bar{S}))}$$

- If $\pi(S) \leq \pi(\bar{S})$, we have $\Phi(S) = \sum_{x \in S} \frac{\pi_x}{\pi(S)} \sum_{y \in \bar{S}} p_{xy}$.
  - $\sum_{x \in S} \frac{\pi_x}{\pi(S)}$ is the probability of being in x if we were in the stationary distribution restricted to S.
  - $\sum_{y \in \bar{S}} p_{xy}$ is the probability of stepping from x to $\bar{S}$ in a single step.
- Then, $\Phi(S)$ is the probability of moving from S to $\bar{S}$ in one step if we were in the stationary distribution restricted to S.

Definition 3. Normalized conductance of the Markov chain

We define *the* normalized conductance of the Markov chain as follows:

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

Theorem 4. ε-mixing time of a random walk

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

$$O\left(\frac{ln(1/\pi_{\mathtt{min}})}{\Phi^2 \varepsilon^3}\right)$$

where $\pi_{\mathtt{min}}$ is the minimum stationary probability of any state.

Proof. See Page 90-92 of Textbook B.

Theorem 4. ε-mixing time of a random walk

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

$$O(\frac{ln(1/\pi_{\mathtt{min}})}{\Phi^2\varepsilon^3})$$

where $\pi_{\mathtt{min}}$ is the minimum stationary probability of any state.

- By theorem 4, one can show the convergence rate of the random walk of
  - 1-dimensional lattice with self-loops at both ends: $O(\frac{n^2 logn}{\varepsilon^3})$ where n denotes the number of vertices.
  - d-dimensional lattice with self-loops at each boundary point : $O(\frac{d^3 n^2 logn}{\varepsilon^3})$ where n denotes the number of vertices in each dimension.
  - a connected undirected graph : $O(\frac{n^4 logn}{\varepsilon^3})$ where n denotes the number of vertices.
- The convergence rate is polynomially bounded by n and d.

Theorem 4. ε-mixing time of a random walk

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

$$O\left(\frac{ln(1/\pi_{\mathtt{min}})}{\Phi^2\varepsilon^3}\right)$$

where $\pi_{\mathtt{min}}$ is the minimum stationary probability of any state.

- Proof sketch of 1-dimensional lattice with self-loops.
  With self-loops, $p_{xy} = \frac{1}{2}$ for all edges (x,y). Also, stationary distribution is $\frac{1}{n}$ for all vertices.
  The set S with $\frac{n}{2}$ vertices yields the minimum $\Phi(S)$ with $\Phi = \frac{1}{n}$.
  Therefore, by Theorem 4, the ε-mixing time is $O(\frac{n^2 logn}{\varepsilon^3})$.

Roadmap

Random Walks on Undirected Graphs with Unit Edge Weights

KAIST AI
Graduate School of AI

- For undirected graphs with unit edge weights, some concrete analysis for a random walk can be done.
  - Hitting time (h_{xy}): expected time of a random walk starting at vertex x to reach vertex y.
  - Commute time ($commute(x, y)$): expected time of a random walk starting at x reaching y and then returning to x. *i.e.*, $commute(x, y) = h_{xy} + h_{yx}$.
  - Cover time ($cover(G)$): expected time of any vertex to reach other vertices at least once.
- The proofs of this section is somewhat technical. You can find the proofs in section 4.6 of Textbook B.

Theorem 5.

If vertices x and y are connected by an edge, then $h_{xy} + h_{yx} \leq 2m$ where m is the number of edges in the graph.

Theorem 6.

For vertices x and y in an n vertex graph, $commute(x, y)$ is less than or equal to n^3.

Theorem 7.

Let G be a connected graph with n vertices and m edges. $Cover(G)$ is bounded above by $4m(n-1)$.

Questions?