Random Walk and Markov Chains

Guo Yuanxin

CUHK-Shenzhen

February 5, 2020

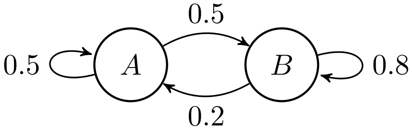
Table of Contents

Table of Contents

Toy Example: A Two-state Random Walk

Example

Consider the two-state random walk below:



Notations

- $\mathbf{p}$: **probability vector**. A row vector with nonnnegative components that sum up to one. Each component specifies the probability mass of a vertex.
- $\mathbf{p}_t$: **probability vector at time** t, specifying the probability masses of vertices at time t.
- $P = (p_{ij})$: **transition matrix**. Entry p_{ij} is the probability of the walk at vertex i selecting the edge to vertex j.
- The defining relationship of a random walk is

$$\mathbf{p}_t P = \mathbf{p}_{t+1}$$

Toy Example(Cont'd): Observations

- In our two-state random walk example:

$$P = \begin{pmatrix} 0.5 & 0.5 \\ 0.2 & 0.8 \end{pmatrix}$$

- We observe that, given initial distribution $\mathbf{p}_1$, we can compute $\mathbf{p}_t$ by the recursive formula:

$$\mathbf{p}_t = \mathbf{p}_{t-1} P = \cdots = \mathbf{p}_1 P^{t-1}$$

- We call P^k the k-step transition matrix.

Long Term Behavior: A Computational View

- We are interested in the asymptotic behavior of $\mathbf{p}_t$, namely when $t \to \infty$.
- We observe that P can be diagonalized as

$$P = Q^{-1}\Lambda Q = \begin{pmatrix} 2/7 & 5/7 \\ 1 & -1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 0 \\ 0 & 0.3 \end{pmatrix} \begin{pmatrix} 2/7 & 5/7 \\ 1 & -1 \end{pmatrix}$$

- As $t \to \infty$,

$$\lim_{t\to\infty} P^{t-1} = \begin{pmatrix} 2/7 & 5/7 \\ 2/7 & 5/7 \end{pmatrix}.$$

- We can verify that, given arbitrary initial probability vector $\mathbf{p}_1$, $\mathbf{p}_t$ will converge to $(2/7, 5/7)$ after sufficiently long time.

Natural Questions Arise...

- We have seen an example of a random walk whose probability vector converges to equilibrium despite the initial probability vector.
- It is natural for us to ask whether every random walk has this property.
- Also, can two different initial distributions converge to different limits?
- Both answers are NO.
- However, we will make certain assumptions of the random walk and instead focus on an another distribution other than $\mathbf{p}_t$.

(Discrete time) Markov Chains

- In statistical literature, a concept of **Markov chains** is usually regarded as the synonym of random walks.

Definition (Markov Chain)

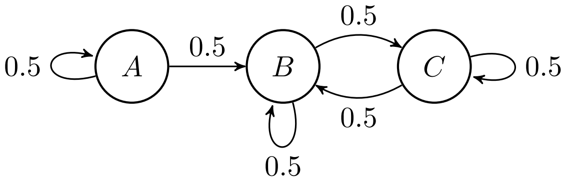
A *Markov chain* is a *stochastic process* in which future states are independent of past states given the present state.

- Consider a sequence of random variables $X_1, X_2, \ldots, X_t$, where X_i is the state at time i. If the random variables form a Markov chain, the state at time $t+1$ only depends on the state at time t, not on any of the past states.
- This is the **Markov property**:

$$\mathbb{P}(X_{t+1}|X_1, X_2, \ldots, X_t) = \mathbb{P}(X_{t+1}|X_t)$$

Basic Assumption: Connected/Irreducible

- We say a Markov chain is **connected/irreducible** if the underlying graph is strongly connected.
- In other words, there exists a directed path from every vertex to every other vertex.
- Here is an example of a not connected Markov chain/random walk:

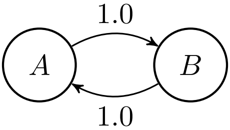


- State B cannot reach state A, thus it is not connected.

Limiting Distribution Does Not Exist

Example

We now consider a case where the probability vector does not necessarily converge. The transition diagram is given by:



- We consider $\mathbf{p}_1 = (1, 0)$, i.e., all the probability mass is at state A initially.
- It is straightforward to see $\mathbf{p}_{2k} = (0, 1)$, $\mathbf{p}_{2k+1} = (1, 0)$, for all $k \in \mathbb{N}$.
- This implies $\lim_{t\to\infty} \mathbf{p}_t$ does not exist.

Limit of the Long Term Avg. is Invariant

- However, if we consider the **long-term average probability distribution $\mathbf{a}_t$** given by

$$\mathbf{a}_t = \frac{1}{t}(\mathbf{p}_1 + \mathbf{p}_2 + \cdots + \mathbf{p}_t),$$

  We observe that this distribution converges:

$$\lim_{t\to\infty} \mathbf{a}_t = (0.5, 0.5).$$

- We also observe that:

$$\mathbf{a}_t P = \mathbf{a}_t,$$

  where P is the transition matrix

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Stationary Distribution

Theorem (FT of Markov Chains)

Let P be the transition probability matrix for a connected Markov chain, $\mathbf{p}_t$ be the probability distribution at time t, and $\mathbf{a}_t$ be the long term average probability distribution. Then there is a unique probability vector $\boldsymbol{\pi}$ satisfying $\boldsymbol{\pi} P = \boldsymbol{\pi}$. Moreover, for any starting distribution, $\lim_{t\to\infty} \mathbf{a}_t$ exists and equals $\boldsymbol{\pi}$.

- By $\boldsymbol{\pi} P = \boldsymbol{\pi}$, we have $\boldsymbol{\pi} P^k = \boldsymbol{\pi}$ for all $k \in \mathbb{N}$, which indicates running any number of steps of the Markov Chain starting with $\boldsymbol{\pi}$ leaves the distribution unchanged.
- For this reason, we call $\boldsymbol{\pi}$ the **stationary distribution**.

Proof

Proposition

For two probability distribution $\mathbf{p}$ and $\mathbf{q}$:

$$\|\mathbf{p}-\mathbf{q}\|_1 = 2\sum_i (p_i - q_i)^+ = 2\sum_i (q_i - p_i)^+,$$

where x^+ is defined to be $\max(x, 0)$.

Lemma

The $n \times (n+1)$ matrix $A = [P - I, \mathbf{1}_n]$ has rank n if P is the transition matrix for a connected Markov chain.

Detailed Balance

- We end this part with a sufficient condition for stationary distributions which will be of great use in the following part:

Lemma (Detailed Balance)

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

- **Proof**: Sum both sides of the detailed balance equation over y, we get $\pi_x = \sum_y \pi_y p_{yx}$. This is equivalent to $\boldsymbol{\pi} P = \boldsymbol{\pi}$, which indicates $\boldsymbol{\pi}$ is a stationary distribution.

Table of Contents

What is Monte Carlo?

- **Monte Carlo**: A fancy name of **simulation**.

Example: Monte Carlo Integration

We have a distribution $p(x)$ that we want to take quantities of interest from (e.g., mean, variance). To derive it analytically, we have to take integrals:

$$I = \int_{\mathbb{R}} g(x)p(x)dx$$

where $g(x)$ is some function of x (e.g. $g(x) = x$ for the mean). We can approximate the integrals via Monte Carlo, where each $x^{(i)}$ is simulated from $p(x)$:

$$\hat{I}_M = \frac{1}{M}\sum_{i=1}^{M} g(x^{(i)})$$

Why Monte Carlo works?

- Intuitive as Monte Carlo may seem, how can one justify that $\hat{I}_M = I$ as $M \to \infty$?

Theorem (Strong Law of Large Numbers)

Let $X_1, X_2, \ldots, X_M$ be a sequence of **independent and identically distributed (i.i.d.)** random variables, each having a finite mean $\mu = \mathbb{E}(X_i)$.
Then with probability 1,

$$\frac{X_1 + X_2 + \cdots + X_M}{M} \to \mu \text{ as } M \to \infty$$

- Recall our last example, every sample point $x^{(i)}$ is simulated independently.
- What if we can't generate **independent** draws?

Sampling with a Markov Chain

- In Bayesian framework, it is essential to sample from the posterior distribution as it allows Monte Carlo estimation of all posterior quantities of interest.
- Typically, it is not possible to sample directly from a posterior. For example, we may not know the normalizing constant.
- However, we can generate *slightly dependent* draws using a **Markov chain**.
- Under certain conditions, we can still find these quantities of interest from those draws.

Ergodic Theorem

Theorem (Ergodic Theorem)

Let $x^{(1)}, x^{(2)}, \ldots, x^{(M)}$ be M values from a Markov chain that is aperiodic, irreducible, and positive recurrent (then the chain is ergodic), and $\mathbb{E}[g(x)] < \infty$. Then with probability 1,

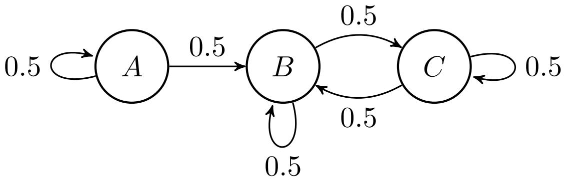
$$\frac{1}{M}\sum_{i=1}^{M} g[x^{(i)}] \to \sum_{\mathcal{X}} g(x)\pi(x) \text{ as } M \to \infty$$

where π is the stationary distribution.

- Note that by letting g be the indicator function $\mathbb{I}_{\{x=s\}}$, one can interpret the stationary distribution as the long-run fraction of time spent in each state.

Recap: Connected/Irreducible

- We say a Markov chain is **connected/irreducible** if the underlying graph is strongly connected.
- In other words, there exists a directed path from every vertex to every other vertex.
- Here is an example of a not connected Markov chain/random walk:



- State B cannot reach state A, thus it is not connected.

Technical Condition: Positive Recurrence

Definition (Recurrence)

A Markov chain is **recurrent** if for any given state i, if the chain starts at i, it will eventually return to i with probability 1.

Definition (Positive Recurrence)

A Markov chain is **positive recurrent** if the expected return time to state i is finite; otherwise it is **null recurrent**.

- The *simple symmetric random walk* on $\mathbb{Z}$ is null recurrent.

Identifying Positive Recurrence

Theorem (Positive Recurrence & Stationary Distribution)

Suppose $\{X_n\}$ is an irreducible Markov chain with transition matrix P. Then $\{X_n\}$ is positive recurrent if and only if there exists a (non-negative, summing to 1) solution π, to the set of linear equations $\pi = \pi P$.

Moreover, the stationary distribution π is given by:

$$\pi_i = \frac{1}{\mathbb{E}(T_{ii})} > 0,$$

where $\mathbb{E}(T_{ii})$ is the expected return time to state i.

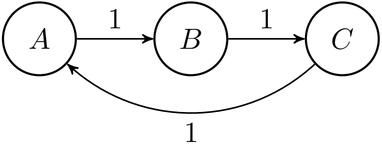
- Intuition: On average, the chain visits state i once every $\mathbb{E}(T_{ii})$ amount of time.

Technical Condition: Aperiodicity

Definition (Aperiodicity)

A Markov chain is **aperiodic** if for any set A, the number of steps required to return to A must not always be a multiple of some value k.

- Here is an example of a periodic Markov chain:



- It always takes $3k$ ($k \in \mathbb{N}$) steps for the chain to return to state A. Thus the chain is periodic.

Goal Revisited

- **Goal**: We want to generate slightly dependent samples from a known distribution using a Markov chain in order to use **Ergodic Theorem**.
- Now we will introduce two algorithms to do it.

Table of Contents

Metropolis-Hastings Algorithm: Overview

- Given a target distribution p over the states, the *Metropolis-Hastings algorithm* is as follows:
  1. Pick an initial state: $X^{(0)} = x$.
  2. At iteration t, suppose $X^{(t)} = y$, propose a move to z with probability $q(z|y)$.
  3. Compute the acceptance ratio:
  $$r(z|y) = \frac{p(z)q(y|z)}{p(y)q(z|y)}$$
  4. Accept the proposed move (i.e., $X^{(t+1)} = z$) with probability
  $$\alpha(z|y) = \min\{1, r(z|y)\}.$$
  Otherwise, $X^{(t+1)} = X^{(t)} = y$
  5. Repeat $2 \sim 4$.

Justification of M-H Algorithm

Theorem (Target distribution is stationary)

The Markov chain with transition probabilities arising from the *Metropolis-Hastings algorithm* has the distribution p as a stationary distribution.

Justification of M-H Algorithm

- **Proof:** The transition probability from state i to j of this chain constructed by M-H algorithm is given by $q(j|i)\alpha(j|i)$.
- Without loss of generality, assume $p(j)q(i|j) < p(i)q(j|i)$, then

$$\begin{aligned} p(i)q(j|i)\alpha(j|i) &= p(i)q(j|i) \cdot \frac{p(j)q(i|j)}{p(i)q(j|i)} \\ &= p(j)q(i|j) \cdot 1 \\ &= p(j)q(i|j)\alpha(i|j), \end{aligned}$$

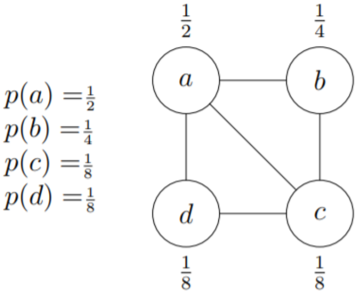
  which is the **detailed balance equation**.
- By the previous lemma, p is the stationary distribution indeed.

Random Walk Metropolis Sampling

- To simplify things, we can have a *symmetric* proposal distribution, i.e., $q(y|x) = q(x|y)$, the acceptance ratio is simply $r(y|x) = p(y)/p(x)$. We call this **random walk Metropolis sampling**.
- If $p > 0$, it is not difficult to establish the ergodicity of this chain.
- This chain favors "heavier" states (with higher p_x), since heavier states have relatively low acceptance rates. .

Example

- We consider the example given in the textbook. The target distribution p is given below.
- We further assume that choosing any edge at a vertex has equal probability.



Python Code Implementation

```
import numpy as np
import matplotlib.pyplot as plt

n = 200000
path = [0]
pr = [1/2,1/4,1/8,1/8]
proposal = [[0,1/3,1/3,1/3],[1/2,0,1/2,0],[1/3,1/3,0,1/3],[1/2,0,1/2,0]]
count = [1,0,0,0]

for i in range(n-1):
    now = path[i]
    new = np.random.choice([0,1,2,3], p=proposal[now])
    r = pr[new]*proposal[new][now]/(pr[now]*proposal[now][new])
    accept = min(1,r)
    gen = np.random.uniform()
    if accept > gen:
        path.append(new)
        count[new] += 1
    else:
        path.append(now)
        count[now] += 1

freq = [i/200000 for i in count]
print(freq)

plt.plot(path[100000::500], lw=1)
plt.show()
```

Figure: Python code for M-H algorithm

Simulation Results

```
[0.50136, 0.24929, 0.12413, 0.12522]
```

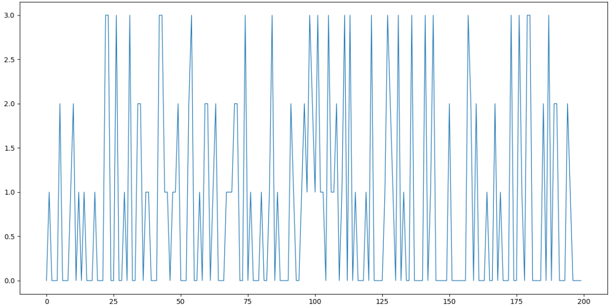


Figure: Simulated stationary distribution & Trace plot

Table of Contents

Gibbs Sampling: Idea

- *Gibbs sampling* is a technique to sample from multivariate distributions.
- The basic idea is to split the multidimensional vector into scalars.
- The beauty of this technique lies in that it simplifies a complex, high-dimensional problem by breaking it down into simple, low-dimensional problems.
- **Note**: We can only use Gibbs sampling if we know the **full conditional** distributions of the variables.

Full Conditional Distribution

- Suppose we have a joint distribution $p(x_1, x_2, \ldots, x_d)$.
- The full conditional distribution of variable x_j is: $p(x_j | x_{-j})$, where x_{-j} denotes all variables except x_j.
- When the joint distribution is known, it is not difficult to find the full conditionals.

Gibbs Sampling: Algorithm

- To generate samples of $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ given a target distribution $p(\mathbf{x})$, do the following steps:

1. Pick an initial state $\mathbf{x}^{(0)}$.
2. At iteration t, current state $\mathbf{x}^{(t)} = (x_1^{(t)}, x_2^{(t)}, \ldots, x_d^{(t)})$. Randomly choose a coordinate x_i to update, while leaving the rest to be unchanged. WLOG, let the coordinate be the first: x_1.
3. Draw $x_1^{(t+1)}$ from $p(x_1|x_2, \ldots, x_d)$
4. Then $\mathbf{x}^{(t+1)} = (x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_d^{(t+1)}) = (x_1^{(t+1)}, x_2^{(t)}, \ldots, x_d^{(t)})$
5. Repeat $2 \sim 4$.

Selecting the Coordinate

- Randomly picking a coordinate to update is not the only scheme to choose the coordinate to update. Another option is to sequentially scan the coordinates from x_1 to x_d.

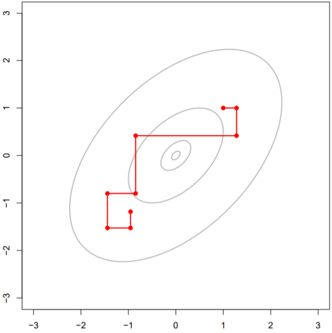


Figure: An illustration of the sequentially scanning scheme

Justification of Gibbs Sampling

- Let $\mathbf{x}$ and $\mathbf{y}$ be two states that differ in only one coordinate, say the first coordinate.
- Then the transition probability from $\mathbf{x}$ to $\mathbf{y}$ is given by:

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

- Note that the normalizing constant is $1/d$ because $\sum_{y_1} p(y_1|x_2, \ldots, x_d) = 1$, and there are totally d directions to move towards.

Justification of Gibbs Sampling

- Similarly,

$$\begin{aligned} p_{\mathbf{yx}} &= \frac{1}{d} p(x_1 | y_2, \ldots, y_d) \\ &= \frac{1}{d} p(x_1 | x_2, \ldots, x_d) \end{aligned}$$

since the algorithm changes only one coordinate at a time.

- By **Law of Total Probability**,

$$p_{\mathbf{xy}} = \frac{1}{d} \frac{p(\mathbf{x})}{p(x_2, \ldots, x_d)}, \; p_{\mathbf{yx}} = \frac{1}{d} \frac{p(\mathbf{y})}{p(x_2, \ldots, x_d)}$$

Justification of Gibbs Sampling

- This is just

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

  which is again the **detailed balance** equation, indicating that p is the stationary distribution.

Gibbs Sampling: Metropolis-Hastings in Disguise

- *Gibbs Sampling* is actually a special case of *Metropolis-Hastings algorithm*, although they look quite different.
- We can see $p_{\mathbf{xy}}$ as the proposal distribution in M-H algorithm.
- The acceptance ratio of any move is 1, i.e. all moves that are proposed are accepted.
- Recall acceptance ratio:

$$r(y|x) = \frac{p(y)q(x|y)}{p(x)q(y|x)}.$$

Summary

- **Metropolis-Hastings Algorithm**: Constructing a Markov chain with target distribution in an accept-reject manner.
- **Gibbs Sampling**: A special form of **Metropolis-Hastings Algorithm** that converts a high-dimensional problem into low-dimensional (usually 1) problems.
- Both algorithm works when the state space is a continuum, where $p(x)$ is changed to the density.

Table of Contents

Motivation

- We can regard both the M-H algorithm and Gibbs sampling as random walks.
- We have demonstrated that no matter what initial state is picked, the walk will eventually converge.
- However, it is also intuitive that the first few states will be highly dependent on initial state.
- A natural question will be how fast the walk starts to starts to reflect the stationary probability?

We will assume our Markov chain is connected in the following part.

Random Walks on Edge-weighted Undirected Graphs

- We exploit one nice property of the random walks involved in the M-H algorithm and Gibbs sampling: they are random walks on edge-weighted undirected graphs.
- These Markov chains are derived from electrical networks.

Conductance: A Notion from Electrical Networks

- Given a network of resistors, the *conductance* of edge (x, y) is denoted c_{xy} and the normalizing constant c_x equals $\sum_y c_{xy}$.
- The Markov chain has transition probabilities proportional to edge conductances, i.e.,

$$p_{xy} = \frac{c_{xy}}{c_x}$$

- Since $c_{xy} = c_{yx}$, we have

$$c_x p_{xy} = c_{xy} = c_{yx} = \frac{c_y}{c_{yx}} = c_y p_{yx},$$

we have from the detailed balance equation that the stationary distribution $\boldsymbol{\pi}$ is given by $\pi_i = c_i / \sum_x c_x$.

Time-reversibility

- A Markov chain satisfying the detailed balance equation is said to be **time-reversible**.
- The name comes from the fact that for a particular path $(i_1, i_2, \ldots, i_k)$, the probability of observing the path is the same as observing its reversal:

$$\pi_{i_1} p_{i_1,i_2} p_{i_2,i_3} \ldots p_{i_{k-1},i_k} = \pi_{i_k} p_{i_k,i_{k-1}} p_{i_{k-1},i_{k-2}} \ldots p_{i_2,i_1}$$

- Given a sequence of states seen, one cannot tell whether the time runs forward or backward.

Slowly Mixing Random Walks

- In general, there are certain random walks that takes a long time to converge.

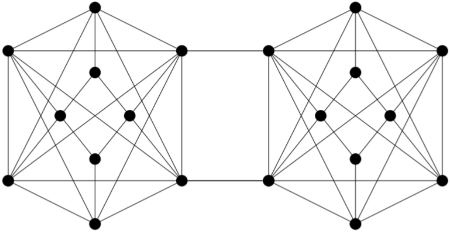


Figure: A network with a constriction

- The rapid mixing of a random walk on this graph is restricted by the narrow passage between two big components.

ϵ-mixing Time

Definition (ϵ-mixing Time)

Fix $\epsilon > 0$. 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 distance between the t-step running average probability distribution $\mathbf{a_t}$ and the stationary distribution $\boldsymbol{\pi}$ is at most ϵ.

Normalized Conductance

Definition (Normalized Conductance)

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

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

- Observe that conductance is symmetric, i.e., $\Phi(S) = \Phi(\bar{S})$

Interpreting the Normalized Conductance

- Suppose WLOG that $\pi(S) \leq \pi(\bar{S})$. Then we can write $\Phi(S)$ as:

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

- The red term is the probability that the walk is in state x given that the walk is in set S.
- The blue term is the probability of stepping from x to $\bar{S}$ in one step.
- Since the red terms sum to 1, it can be seen as a distribution. $\Phi(S)$ is thus the overall probability of stepping to $\bar{S}$ from S in one step.

Interpreting the Normalized Conductance

- Since the number of step needed to get into $\bar{S}$ follows $Geo(\Phi(S))$, the expected number of steps needed to get into $\bar{S}$ is $1/\Phi(S)$.
- Clearly, to be close to the stationary distribution, we must at least get to $\bar{S}$ once.
- Hence, $1/\Phi(S)$ is a lower bound of mixing time.
- And since we can choose any S to start with, mixing time is lower bounded by the minimum over all S of $\Phi(S)$.

Normalized Conductance of the Markov Chain

Definition (Normalized Conductance of the Markov Chain)

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

$$\Phi = \min_S \Phi(S).$$

Finding the ϵ-mixing time

Theorem (Mixing Time for Undirected Graph)

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

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

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

Proof

Lemma

Suppose $G_1, \ldots, G_r$ and $u_1, \ldots, u_{r+1}$ defined as before, then

$$\pi(G_1 \cup G_2 \cup \cdots \cup G_k)(u_k - u_{k+1}) \leq \frac{8}{t\Phi\epsilon} + \frac{2}{t}$$

Example: Mixing Time of 1-D Lattice

- Consider a random walk on an undirected graph consisting of an $2n$-vertex path with self-loops at the both ends. Conductance of each edge (self-loops included) is the same.
- The stationary distribution is thus uniform over all states.
- The set with minimum normalized conductance is the set with probability $\pi \leq 1/2$ and the maximum number of vertices with the minimum number of edges leaving it.
- This is just the set with the first n vertices. (Why?)

Example: Mixing Time of 1-D Lattice (Cont'd)

- The conductance from S to $\bar{S}$ is $\pi_n p_{n,n+1} = \frac{1}{4n} = O(\frac{1}{n})$.
- $\pi(S) = \pi(\bar{S}) = \frac{1}{2}$
- Hence, $\Phi = 2\pi_n p_{n,n+1} = O(\frac{1}{n})$
- The mixing time is thus $O(\frac{n^2 \ln n}{\epsilon})$.