10-708 Recitation 3 - Monte Carlo Markov Chain

Fan Pu Zeng

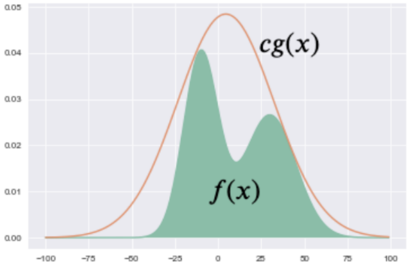
17 Feb 2023

Table of Contents

Rejection Sampling

Rejection Sampling (HW2 Q5)

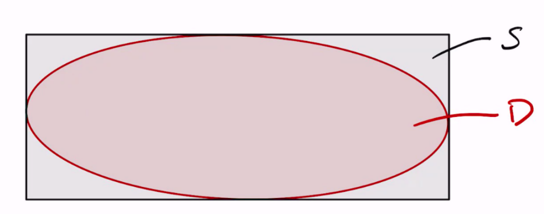
- Problem: you have some hard to sample distribution f (target distribution), and an easy to sample distribution g (proposal distribution)
- Rejection sampling algorithm: Choose c large enough such that $\forall x, f(x) \leq cg(x)$.
  - Generate sample x from g
  - Generate sample $u \sim \text{Unif}(0,1)$
  - Accept if $u \leq \frac{f(x)}{cg(x)}$



- Good when g is close to h and therefore c is small

Rejection Sampling (HW2 Q5)

Illustration for proposal distribution S and target distribution D in Q5:



$$D = \{(x, y) : (x/2)^2 + y^2 \leq 1\}$$
$$S = \{(x, y) : -2 \leq x \leq 2, -1 \leq y \leq 1\}$$

Markov Chains

Markov Chains

- Let T denote the transition matrix of a Markov chain.

Definition (Stationary Distribution)

A distribution $\boldsymbol{\pi}$ is stationary if $\boldsymbol{\pi}\mathbf{T} = \boldsymbol{\pi}$.

- When does a *unique* stationary distribution exist?

Markov Chains

- Let T denote the transition matrix of a Markov chain.

Definition (Stationary Distribution)

A distribution $\boldsymbol{\pi}$ is stationary if $\boldsymbol{\pi}\mathbf{T} = \boldsymbol{\pi}$.

- When does a *unique* stationary distribution exist?
- Sufficient conditions:
    1. Irreducibility: transition graph is connected, able to reach any state from any other state eventually
    2. Aperiodicity: random walk doesn't get trapped in cycles, i.e there exists some n where eventually there is positive probability of being in all states after n steps

Markov Chains

- Let T denote the transition matrix of a Markov chain.

Definition (Stationary Distribution)

A distribution $\boldsymbol{\pi}$ is stationary if $\boldsymbol{\pi}\mathbf{T} = \boldsymbol{\pi}$.

- When does a *unique* stationary distribution exist?
- Sufficient conditions:
  1. Irreducibility: transition graph is connected, able to reach any state from any other state eventually
  2. Aperiodicity: random walk doesn't get trapped in cycles, i.e there exists some n where eventually there is positive probability of being in all states after n steps
- When is $\boldsymbol{\pi}$ a stationary distribution? Sufficient condition: $\boldsymbol{\pi}$ satisfies detailed balance:

$$\boldsymbol{\pi}_i \mathbf{T}_{ij} = \boldsymbol{\pi}_j \mathbf{T}_{ji} \qquad \forall (i, j).$$

Monte Carlo Markov Chain

Monte Carlo Markov Chain

- Importance sampling may work in low dimensions, but becomes inefficient in high dimensions (ratio of volumes grow exponentially, always rejecting)
- Idea: construct a Markov Chain on the state space whose stationary distribution is the target distribution

Monte Carlo Markov Chain

- Importance sampling may work in low dimensions, but becomes inefficient in high dimensions (ratio of volumes grow exponentially, always rejecting)
- Idea: construct a Markov Chain on the state space whose stationary distribution is the target distribution

Example

For Ising models,the Markov Chain will move around state space $\{0, 1\}^n$.
After reaching stationary distribution, proportion of time spent in some state $\mathbf{x} \in \{0, 1\}^n$ proportionate to $p(\mathbf{x})$, so sampling from the Markov Chain is like sampling from p

- Question: How to determine $\mathbf{T}$?

Metropolis-Hastings (HW2 Programming)

Main idea:

- Suppose we have some easy to sample proposal distribution (also called transition kernel) $q(i,j)$, and we are in state j
- At each step, we sample a proposal i with probability $q(i,j)$
- Be clever about deciding the probability to accept the proposal
- The Markov Chain will eventually reach a stationary distribution

Algorithm:

$$\Pr(X_n = j \,|\, X_{n-1} = i) =$$

$$\begin{array}{ll} 1., & \text{from state } i \text{ go to state } j \text{ with prob. } q(i,j) \\ 2., & \begin{cases} \text{with prob } 1-\alpha(i,j) \text{ go back to state } i, \\ \text{with prob } \alpha(i,j) \text{ stay in state } j. \end{cases} \end{array}$$

where

$$\alpha(i,j) = \min\left(\frac{\pi_j q(j,i)}{\pi_i q(i,j)}, 1\right) = \min\left(\frac{b(j)q(j,i)}{b(i)q(i,j)}, 1\right).$$

Gibbs Sampling (HW2 Programming)

- Downside with Metropolis-Hastings: need to come up with a proposal distribution q, and acceptance rate may be low
- Gibbs sampling always accepts, and is a special case of MH

Algorithm:

Repeat:

Let current state be $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$

Pick $i \in [n]$ uniformly at random.

Sample x $\sim P(X_i = x | \boldsymbol{x}_{-i})$

Update state to $\boldsymbol{y} = (x_1, x_2, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)$

Why is $x \sim P(X_i = x \mid \mathbf{x}_{-i})$ easy to sample? You will show this in the HW.

HW2 Programming Hints

- The neighbors of node (i, j) are just its vertical and horizontal neighbors on the $n \times x$ grid
- In the setup of this problem, there is double counting of the edges. In general, whether there is double counting or not is a matter of convention and does not affect any of our results.

Linear Algebra Refresher

Linear Algebra Refresher and Hints (Q6, Q8)

- Let A be any matrix. v is an eigenvector of A if $Av = \lambda v$ for some $\lambda \in \mathbb{R}$. λ is called the eigenvalue associated with v.
- Vectors u, v are orthogonal when $\langle u, v \rangle = 0$
- A matrix U is orthogonal when all its rows are pairwise orthogonal, and all its columns are pairwise orthogonal
- For a square orthogonal matrix U, $UU^T = U^T U = I$
- The operator norm of a matrix A is defined as:

$$\|A\|_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

Linear Algebra Refresher and Hints (Q6, Q8)

Properties of a $n \times n$ **symmetric** matrix A:

- Exhibits an eigendecomposition:

$$A = UDU^T = \sum_{i=1}^{n} \lambda_i \underbrace{v_i v_i^T}_{n \times x}$$

  where U orthogonal, $D = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$, λ_i eigenvalues, v_i eigenvector of unit norm corresponding to λ_i.

- All eigenvectors v_i are orthogonal:

$$\langle v_i, v_j \rangle = 0 \qquad \forall\, i \neq j$$

- The largest eigenvalue λ_1 of A is given by

$$\lambda_1 = \max_{\|x\|_2 = 1} x^T A x.$$

- The second largest eigenvalue λ_2 is given by

$$\lambda_2 = \max_{\|x\|_2 = 1, \langle x, v_1 \rangle = 0} x^T A x.$$

Markov Chain Mixing Times

Markov Chain Mixing Times (HW2 Q8)

- In theory:

$$\lim_{t \to \infty} \mathbf{T}^t \mathbf{x} = \boldsymbol{\pi}. \tag{1}$$

- In practice: how long does it take for my Markov Chain to reach a stationary distribution?
  Reach means:

$$\|\mathbf{T}^k \mathbf{x} - \boldsymbol{\pi}\|_{TV} < 1/4 \tag{2}$$

Markov Chain Mixing Times (HW2 Q8)

- In theory:

$$\lim_{t\to\infty} \mathbf{T}^t \mathbf{x} = \boldsymbol{\pi}. \tag{1}$$

- In practice: how long does it take for my Markov Chain to reach a stationary distribution?
  Reach means:

$$\|\mathbf{T}^k \mathbf{x} - \boldsymbol{\pi}\|_{TV} < 1/4 \tag{2}$$

- Linear algebraic view: you will learn how to bound the mixing time in terms of the difference between the two largest eigenvalues in magnitude of $\mathbf{T}$

Markov Chain Mixing Times (HW2 Q8)

High level overview:

- For a connected d-regular transition matrix $\mathbf{T}$, you will show its largest eigenvalue is 1
- Let $\lambda_{\max}$ denote the next largest eigenvalue. You will show that the number of steps k required to mix is

$$k \geq \frac{\log n}{1 - \lambda_{\max}}.$$

- Asymptotically, $1 - \lambda_{\max}$ could be $O(1)$ (clique), $O(1/n)$, $O(1/n^2)$ (cycle), etc, so mixing time could vary a lot.

Markov Chain Mixing Times (HW2 Q8)

- $S \subseteq V$ set of vertices in the graph, $E(S, \overline{S})$: set of edges that are cut between the two partitions S and $\overline{S}$.
- Conductance of a cut S:

$$\Phi(S) = \frac{|E(S, \overline{S}|)}{d \cdot |S|}, \tag{3}$$

- Conductance of the graph represented by $\mathbf{T}$:

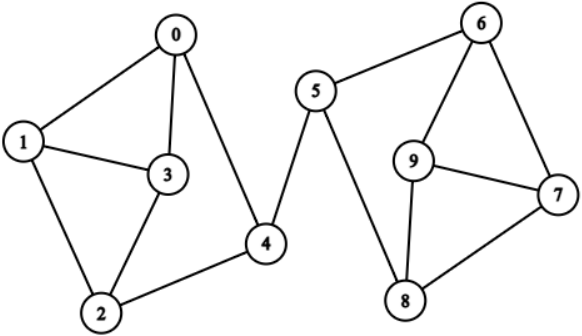
$$\Phi_{\mathbf{T}} = \min_{S, |S| \leq |\overline{S}|} \frac{|E(S, \overline{S}|)}{d \cdot |S|}. \tag{4}$$

- ☹ Let's do some examples!

Markov Chain Mixing Times (HW2 Q8)

Conductance of a cut S:

$$\Phi(S) = \frac{|E(S, \overline{S}|)}{d \cdot |S|}, \tag{5}$$

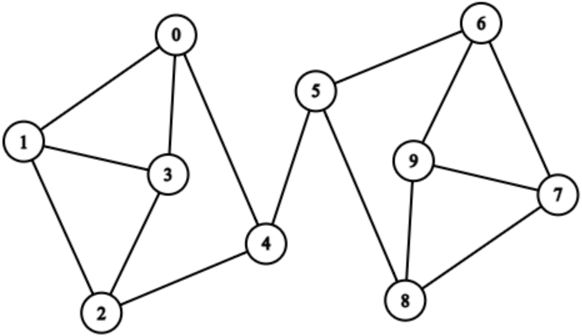


What is $\Phi(\{1, 2\})$?

Markov Chain Mixing Times (HW2 Q8)

Conductance of a cut S:

$$\Phi(S) = \frac{|E(S, \overline{S}|)}{d \cdot |S|}, \tag{5}$$

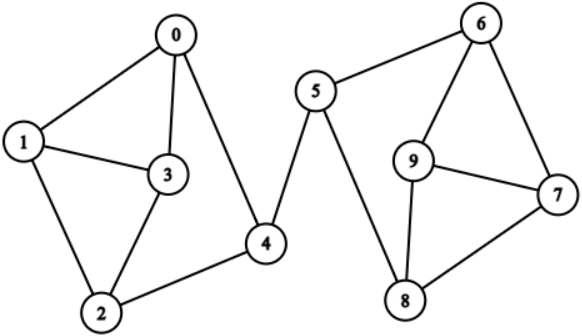


What is $\Phi(\{1, 2\})$? Ans: $1/2$

Markov Chain Mixing Times (HW2 Q8)

Conductance of the entire graph:

$$\Phi_{\mathbf{T}} = \min_{S, |S| \leq |\overline{S}|} \frac{|E(S, \overline{S}|)}{d \cdot |S|}. \tag{6}$$

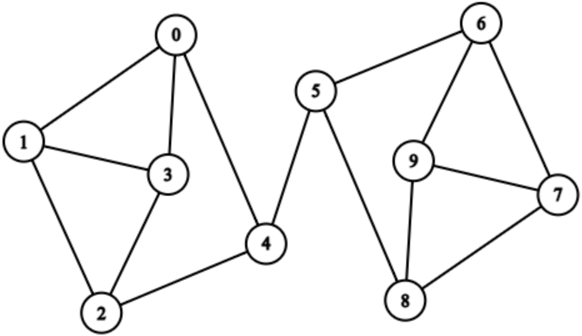


What is $\Phi_{\mathbf{T}}$?

Markov Chain Mixing Times (HW2 Q8)

Conductance of the entire graph:

$$\Phi_{\mathbf{T}} = \min_{S, |S| \le |\overline{S}|} \frac{|E(S, \overline{S}|)}{d \cdot |S|}. \tag{6}$$



What is $\Phi_{\mathbf{T}}$? Take everything on the left side, 1/15

Markov Chain Mixing Times (HW2 Q8)

High level overview, continued:

- In practice, hard to find/characterize $1 - \lambda_{\max}$ for a family of graphs
- You will use the conductance $\Phi_{\mathbf{T}}$ of the graph represented by $\mathbf{T}$, and use it to bound $1 - \lambda_{\max}$. You will prove the LHS of the following result:

$$\frac{1 - \lambda_{\max}}{2} \leq \Phi_{\mathbf{T}} \leq \sqrt{2 \cdot (1 - \lambda_{\max})}$$

Markov Chain Mixing Times (HW2 Q8)

Other hints:

- We only consider graphs that are connected, d-regular, and distribute transition probabilities uniformly among its neighbors. This implies that $\mathbf{T}$ is symmetric.
- Q8(a): After you show that $\lambda_1 = 1$ and $v_1 = \frac{1}{\sqrt{n}}\vec{1}$, this fact is very important and will be used many times
- Q8(h): When relaxing from discrete to continuous constraints, the solution can only get better, i.e

$$\min_{x \in \{0,1\}^n} f(x) \geq \min_{x \in \mathbb{R}^n} f(x)$$

- Lots of hints included in problem ☺

Homework Overview

What's Next?

- You have learnt enough material to do Q1-Q5, Q8, and B.1(a) and B.1(b) of the programming homework
- Next week we will cover annealed importance sampling and Hamiltonian Monte Carlo, and you will have everything you need
- Start early ☺

Thank you for coming to recitation and good luck on the homework!