Mathematics for Machine Learning (AI 512): MCMC Sampling

Amit Chattopadhyay

IIIT-Bangalore

MCMC Sampling

Introduction

- **Markov Chain Monte Carlo (MCMC):** Only available method for sampling from distributions on **high dimensional** domain
- **SIAM News Survey:** MCMC is among top 10 important algorithms of 20th century
- **Metropolis:** 1953, **Hasting:** 1970
- **Key Idea:** 1. Construct a Markov chain on the "**state space**" ("**vertices**" of a lattice graph in domain) whose stationary distribution is the target distribution

  2. A "**random walk**" on the "**state space**" generates the required samples

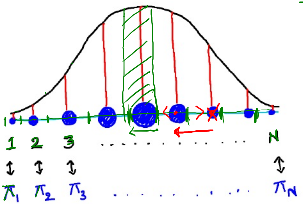
MCMC Sampling: 1D Case

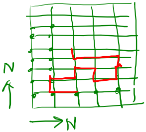
Problem:

Given a p.d.f. $p(x)$, generate samples $\{x_1, x_2, \ldots, x_n\}$ that follow $p(x)$.

Goal: To design a Markov chain s.t.: $\underline{\pi} = \underline{\pi}\,\mathbb{P}$ with $\boxed{\pi_i = \boxed{p(x_i)}}$

known





N^2

d-dim dom

N^d grid pts.

Step I: Constructing (target) stationary distribution $\underline{\pi}$

1. Start with a 1D lattice graph $G = (V, E)$ on the domain of $p(x)$
2. Discretize $p(x)$ at the grid points to obtain: $\underline{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$

- Each interior vertex of the lattice graph G has $2d$ edges ($d = 1$)
- Let r be the maximum degree of any vertex in G

Step II: Constructing transition matrix $\mathbb{P}$ (Metropolis-Hasting Algo.)

1. At any state 'i', select a neighbor 'j' with probability $\frac{1}{r}$

   Since degree of 'i' can be $< r$ the random walk can remain at 'i' with some positive probability.

2. If a neighbor j ($\neq i$) is selected

   I $\pi_j \geq \pi_i$: $\min\left\{1, \frac{\pi_j}{\pi_i}\right\} = 1$.

$$p_{i,j} = \boxed{\frac{1}{r}} \min\left\{1, \frac{\pi_j}{\pi_i}\right\}$$

   II $\pi_j < \pi_i$: $\min\left\{1, \frac{\pi_j}{\pi_i}\right\} = \frac{\pi_j}{\pi_i}$

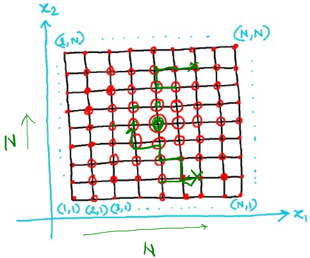
3. $p_{i,i} = 1 - \sum_{j \neq i} p_{i,j}$

MCMC Sampling: 2D Case

Problem:

Given a p.d.f. $p(\underline{x})$, generate samples $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\}$ that follows $p(\underline{x})$.

Goal: To design a Markov chain to satisfy: $\underline{\pi} = \underline{\pi}\,\mathbb{P}$ with $\pi_i = p(\underline{x}_i)$



Step I: Constructing (target) stationary distribution $\underline{\pi}$

1. Start with a 2D lattice (mesh or grid) graph G on the domain of $p(\underline{x})$
2. Discretize $p(\underline{x})$ at the grid points to obtain: $\underline{\pi} = (\pi_1, \pi_2, \ldots, \pi_{N^2})$

MCMC Sampling: 2D Case

- Each interior vertex of the lattice graph G has $2d$ edges ($d = 2$)
- Let r be the max degree of any vertex in G

Step II: Constructing transition matrix $\mathbb{P}$ (Metropolis-Hasting Algo.)

1. At any state 'i', select a neighbor 'j' with probability $\frac{1}{r}$

   Since degree of 'i' can be $< r$ the random walk can remain at 'i' with some positive probability.

2. If a neighbor j ($\neq i$) is selected

$$p_{i,j} = \frac{1}{r} \min\left\{1, \frac{\pi_j}{\pi_i}\right\}$$

3. $p_{i,i} = 1 - \sum_{j \neq i} p_{i,j}$ ($\neq 0$, for some i's)

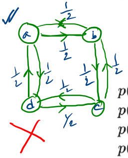
Observations

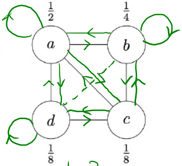
- According to property of Markov chain sample generated at $t+1$ depends on sample generated at t
- **Drawback:** Although the Markov chain eventually converges to the desired distribution, the initial samples may follow a very different distribution, especially if the starting point is in a region of low density. As a result, a **burn-in period** may be long.

$$\begin{array}{c|cccc} & a & b & c & d \\ a & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ b & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ c & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ d & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{array}$$

$\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$ Compute the transition matrix $\mathbb{P}$ for the given probability distribution using MH algorithm.

$\frac{1}{2} \min\{1, 1\}$





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

$$p_{i,j} = \frac{1}{r} \min\left\{1, \frac{\pi_j}{\pi_i}\right\}$$

$$p_a p_{a,b} = p_b p_{b,a} \quad \forall a, b$$

$$r = 3$$

$$p_{a,b} = \frac{1}{3} \min\left\{1, \frac{\frac{1}{4}}{\frac{1}{2}}\right\} = \frac{1}{6}$$

$$p_{a,c} = \frac{1}{3} \min\left\{1, \frac{\frac{1}{8}}{\frac{1}{2}}\right\} = \frac{1}{12}$$

$$p_{a,d} = \frac{1}{3} \min\left\{1, \frac{\frac{1}{8}}{\frac{1}{2}}\right\} = \frac{1}{12}$$

$$p_{a,a} = 1 - \left(\frac{1}{6} + \frac{1}{12} + \frac{1}{12}\right) = 1 - \frac{4}{12} = \frac{2}{3}$$

$$p_{b,a} = \frac{1}{3} \min\left\{1, \frac{\frac{1}{2}}{\frac{1}{4}}\right\} = \frac{1}{3}.$$

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

$$\left(\frac{1}{2}\ \ \frac{1}{4}\ \ \frac{1}{8}\ \ \frac{1}{8}\right)\begin{bmatrix} & \\ & \end{bmatrix} = \left(\frac{1}{2}\ \ \frac{1}{4}\ \ \frac{1}{8}\ \ \frac{1}{8}\right)$$

Markov Chain: Reversible

Definition:

Let $(X_0, X_1, \ldots)$ be a Markov chain with state space $S = (s_1, \ldots, s_N)$ and transition matrix $\mathbb{P} = (p_{i,j})_{N \times N}$.

A probability distribution $\underline{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ on S is said to be **reversible** for the chain (or for $\mathbb{P}$) if for all $i, j \in \{1, 2, \ldots, N\}$ we have

$$\boxed{\pi_i \, p_{i,j} = \pi_j \, p_{j,i}.} \text{ (\textbf{detailed balance})}$$

A Markov chain is said to be **reversible** if there exists a reversible distribution for it.

Properties of Transition Matrix $\mathbb{P}$

1. **Detailed balance condition**: $\pi_i p_{i,j} = \pi_j p_{j,i}$ $(\forall\ i,j)$
2. **Global balance condition**: $\underline{\pi} = \underline{\pi}\mathbb{P}$
3. $\underline{\pi}$ is unique stationary distribution for $\mathbb{P}$.

1. $\pi_i p_{i,j} = \pi_i \frac{1}{r} \min\left\{1, \frac{\pi_j}{\pi_i}\right\} = \frac{1}{r}\min\{\pi_i, \pi_j\} = \frac{1}{r}\pi_j \min\left\{\frac{\pi_i}{\pi_j}, 1\right\}$

$= \pi_j p_{j,i} \quad \forall i,j$

2. $\sum_{i=1}^{N} \pi_i p_{i,j} = \sum_{i=1}^{N} \pi_j p_{j,i} = \pi_j \sum_{i=1}^{N} p_{j,i} = \pi_j$

$\Rightarrow \pi_j = \sum_{i=1}^{N} \pi_i p_{i,j} \quad \forall j = 1, 2, \ldots N$

$\boxed{\underline{\pi} = \underline{\pi}\mathbb{P}}$

3. $\mathbb{P}$ is irreducible & aperiodic.

$\Rightarrow \pi$ is unique

Detailed Balance $\Rightarrow$ Global Balance

Proposition

If $\underline{\pi}$ is a reversible distribution for the Markov chain, then it is also a stationary distribution for it.

- MH algorithm can be implemented even when $\underline{\pi}$ is known only up to a constant, i.e.,

$$\pi(\underline{x}) = c\, f(\underline{x})$$

where $f(\underline{x})$ is known, but c is unknown,

since, MH algorithm depends on $\underline{\pi}$ only through ratio.

- **HW:** Implement MH algorithm to generate samples following the distribution:

$$p(x) = exp(-x), \;\; x \geq 0.$$

$$\sum_{i=1}^{N} \pi_i = 1.$$