APMA1690: Homework # 8 (Due by 11pm on November 16)

1 Review

I would suggest you go through the review section before going to the problem set.

1.1 Metropolis-Hastings Algorithm

Metropolis and Ulam (1949) and Metropolis et al. (1953) were the first to describe Markov chain simulation of probability distributions. The method was concluded as the "Metropolis algorithm." Hastings (1970) generalized this algorithm.

Throughout the problem set, we assume the following:

- The state space $\mathcal{X}$ is finite, i.e., $\#\mathcal{X} < \infty$.
- π is the distribution of interest, and it is strictly positive, i.e.,

$$\pi(x) > 0, \quad \text{for all } x \in \mathcal{X}.$$

1.1.1 Metropolis Algorithm

Suppose we have a transition probability function $q(x, y)$ in hand, and $q(x, y)$ satisfies the following conditions

- we know how to generate a Markov chain from $q(x, y)$ in an efficient way;
- $q(x, y)$ is symmetric, i.e., $q(x, y) = q(y, x)$ for all $x, y \in \mathcal{X}$.

The Metropolis algorithm generates a Markov chain with the following transition probability[1]

$$p(x,y) = \begin{cases} q(x,y) \cdot \min\left\{1, \dfrac{\pi(y)}{\pi(x)}\right\}, & \text{if } x \neq y, \\ 1 - \displaystyle\sum_{z:z\neq x} p(x,z), & \text{if } x = y, \end{cases} \tag{1.1}$$

for all $x, y \in \mathcal{X}$, where "$\sum_{z:z\neq x}$" denotes the sum across all $z \in \mathcal{X}$ that are not equal to x. Algorithm 1 can generate a Markov chain whose transition probability is the $p(x, y)$ defined in Eq. (1.1).

[1] The logic of Eq. (1.1) is the following: we first define $p(x, y)$ for all $x \neq y$, then we define $p(x, x) = 1 - \sum_{z:z\neq x} p(x, z)$.

Algorithm 1 : Metropolois Algorithm

Input: (i) the distribution π of interest satisfying $\pi(x) > 0$ for all $x \in \mathcal{X}$; (ii) a jumping distribution — a symmetric transition probability q of an irreducible and aperiodic Markov chain; (iii) an initial starting point x_0; (iv) a large integer n^*.

Output: The first n^* components of a Markov chain $\{X_n\}_{n=0}^{\infty}$ with π as its stationary distribution.

1: Initialize $X_0 \leftarrow x_0$.
2: **for all** $n = 1, 2, \ldots, n^*$ **do**
3: Sample a proposal X^* from the PMF $q(X_{n-1}, \cdot)$.
4: Compute the ratio $r \leftarrow \frac{\pi(X^*)}{\pi(X_{n-1})}$. (Remark: Since we assume that $\pi(x) > 0$ for all $x \in \mathcal{X}$, the ratio r is always well-defined.)
5: Generate $Y \sim \text{Bernoulli}(\min\{r, 1\})$, i.e., $\mathbb{P}(Y = 1) = \min\{r, 1\}$.
6: $X_n \leftarrow Y \cdot X^* + (1 - Y) \cdot X_{n-1}$.
7: **end for**

1.1.2 Metropolis-Hastings Algorithm

The Metropolis algorithm requires $q(x, y)$ to be symmetric. **This symmetry condition can be removed** by modifying Eq. (1.1) to the following form, which results in the Metropolis-Hastings algorithm.

$$p(x,y) := \begin{cases} q(x,y) \cdot \min\left\{1, \dfrac{\pi(y) \cdot q(y,x)}{\pi(x) \cdot q(x,y)}\right\}, & \text{if } x \neq y \text{ and } q(x,y) > 0, \\ 0, & \text{if } x \neq y \text{ and } q(x,y) = 0, \\ 1 - \displaystyle\sum_{z: z \neq x} p(x,z), & \text{if } x = y, \end{cases} \tag{1.2}$$

for all $x, y \in \mathcal{X}$, where "$\sum_{z:z\neq x}$" denotes the sum across all z's that are not equal to x. **The only difference between Eq. (1.1) and Eq. (1.2) is that the $q(x, y)$ in Eq. (1.2) is no longer required to be symmetric.**

The following theorem is Theorem 18.15 of Klenke (2020) and provides the theoretical foundation for the Metropolis-Hastings algorithm

Theorem 1.1. *Assume that q is irreducible and that for any $x, y \in \mathcal{X}$, we have $q(x, y) > 0$ if and only if $q(y, x) > 0$. Then, the transition probability p defined in Eq. (1.2) is irreducible with unique stationary distribution π. If, in additon, q is aperiodic, then p is aperiodic.*

Algorithm 2 can generate a Markov chain whose transition probability is the $p(x, y)$ defined in Eq. (1.2).

1.2 Gibbs Sampling

"Gibbs sampling is named after the physicist Josiah Willard Gibbs, in reference to an analogy between the sampling algorithm and statistical physics. The algorithm was described by brothers Stuart and Donald Geman in 1984, some eight decades after the death of Gibbs (Geman and Geman, 1984) and became popularized in the statistics community for calculating marginal probability distribution, especially the posterior distribution." (see the Wikipedia page on Gibbs sampling.)

Algorithm 2 : Metropolois-Hastings Algorithm

Input: (i) the distribution π of interest satisfying $\pi(x) > 0$ for all $x \in \mathcal{X}$; (ii) a jumping distribution — a transition probability q (not necessarily symmetric) of an irreducible and aperiodic Markov chain; (iii) an initial starting point x_0; (iv) a large integer n^*.

Output: The first n^* components of a Markov chain $\{X_n\}_{n=0}^{\infty}$ with π as its stationary distribution.

1: Set $X_0 \leftarrow x_0$.
2: **for all** $n = 1, 2, \ldots, n^*$ **do**
3: Sample a proposal X^* from the PMF $q(X_{n-1}, \cdot)$.
4: Compute the ratio $r \leftarrow \frac{\pi(X^*) \cdot q(X^*, X_{n-1})}{\pi(X_{n-1}) \cdot q(X_{n-1}, X^*)}$. (Remark: If $q(X_{n-1}, X^*) = 0$, then it is almost impossible to sample X^* in the preceding step. So, the ratio r is well-defined with probability one.)
5: Generate $Y \sim \text{Bernoulli}(\min\{r, 1\})$.
6: $X_n \leftarrow Y \cdot X^* + (1 - Y) \cdot X_{n-1}$.
7: **end for**

1.3 2-dimensional Gibbs Sampling

Firstly, let us review the 2-dimensional Gibbs sampling. Let $\pi(\boldsymbol{x}) = \pi(\xi_1, \xi_2)$ be a joint PMF on a two-dimensional state space $\mathcal{X}$, where $\boldsymbol{x} = (\xi_1, \xi_2)$. We assume π is strictly positive, i.e.,

$$\pi(\boldsymbol{x}) > 0, \quad \text{for all } \boldsymbol{x} \in \mathcal{X}. \tag{1.3}$$

We define the following marginal distributions

$$\pi_1(\xi_1) = \sum_{\xi_2} \pi(\xi_1, \xi_2), \quad \pi_2(\xi_2) = \sum_{\xi_1} \pi(\xi_1, \xi_2).$$

Then, we have the following conditional distributions

$$\pi_{1|2}(\xi_1|\xi_2) = \frac{\pi(\xi_1, \xi_2)}{\pi_2(\xi_2)}, \text{ which is a function of } \xi_1, \text{ and } \xi_2 \text{ is viewed as a fixed parameter;}$$
$$\pi_{2|1}(\xi_2|\xi_1) = \frac{\pi(\xi_1, \xi_2)}{\pi_1(\xi_1)}, \text{ which is a function of } \xi_2, \text{ and } \xi_1 \text{ is viewed as a fixed parameter.}$$

Because of Eq. (1.3), all the marginal and conditional distributions defined above are strictly positive.

With these marginal and conditional distributions, the transition probability of the Gibbs sampler from state $\boldsymbol{x} = (\xi_1, \xi_2)^\intercal$ to $\boldsymbol{y} = (\eta_1, \eta_2)^\intercal$ is the following

$$p(\boldsymbol{x}, \boldsymbol{y}) = \pi_{1|2}(\eta_1|\xi_2) \cdot \pi_{2|1}(\eta_2|\eta_1). \tag{1.4}$$

Based on Eq. (1.4), the 2-dimensional Gibbs sampling algorithm is implemented as follows

- Step 1: Sample $\xi_1^{(1)} \sim \pi_{1|-1}(\xi_1|\xi_2^{(0)})$.
- Step 2: Sample $\xi_2^{(1)} \sim \pi_{2|-2}(\xi_2|\xi_1^{(1)})$; then, we have $\boldsymbol{X}^{(1)} = (\xi_1^{(1)}, \xi_2^{(1)})^\intercal$.
- Repeat...

The following theorem is the corner stone of the 2-dimensional Gibbs sampling

Theorem 1.2. *Let π be a PMF satisfying Eq. (1.3). The transition probability $p(\boldsymbol{x}, \boldsymbol{y})$ defined in Eq. (1.4) has the following properties*

- *(Irreducibility) p is irreducible.*
- *(Aperiodicity) p is aperiodic.*
- *(Stationarity) π is the unique stationary distribution of p.*

The proof of Theorem 1.2 was given in class.

1.3.1 d-dimensional Gibbs Sampling with $d \geq 2$

Suppose we focus on a **strictly positive** d-dimensional PMF $\pi(\xi_1, \ldots, \xi_{i-1}, \xi_i, \xi_{i+1}, \ldots, \xi_d)$. Then, we define the following marginal and conditional distributions for each fixed index $i \in \{1, 2, \ldots, d\}$

$$\pi_{-i}(\xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_d) \stackrel{\text{def}}{=} \sum_{\xi_i} \pi(\xi_1, \ldots, \xi_i, \ldots, \xi_d),$$

$$\pi_{i|-i}(\xi_i \,|\, \xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_d) \stackrel{\text{def}}{=} \frac{\pi(\xi_1, \ldots, \xi_i, \ldots, \xi_d)}{\pi_{-i}(\xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_d)}$$

We provide the following algorithmic version of the Gibbs sampling

Algorithm 3 : Gibbs Sampling

Input: (i) the given distribution $\pi(\xi_1, \ldots, \xi_d)$; (ii) a starting point $\boldsymbol{x}^{(0)} = (\xi_1^{(0)}, \ldots, \xi_d^{(0)})^\intercal$.

Output: A homogeneous Markov chain $\{\boldsymbol{X}^{(n)} = (\xi_1^{(n)}, \ldots, \xi_d^{(n)})^\intercal\}_{n=0}^{\infty}$ with π as its stationary distribution.

1: Set $\boldsymbol{X}^{(0)} \leftarrow \boldsymbol{x}^{(0)}$.
2: **for all** $n = 1, 2, \ldots,$ **do**
3: Sample $\xi_1^{(n)} \sim \pi_{1|-1}(\xi_1 | \xi_2^{(n-1)}, \ldots, \xi_d^{(n-1)})$
4: **for all** $j = 2, \ldots, d$ **do**
5: Sample $\xi_j^{(n)} \sim \pi_{j|-j}(\xi_i | \xi_1^{(n)}, \ldots, \xi_{j-1}^{(n)}, \xi_{j+1}^{(n-1)}, \ldots, \xi_d^{(n-1)})$
6: **end for** $\boldsymbol{X}^{(n)} \leftarrow (\xi_1^{(n)}, \xi_2^{(n)}, \ldots, \xi_d^{(n)})^\intercal$
7: **end for**

We take the scenario $d = 2$ as an example:

- Step 1: Sample $\xi_1^{(1)} \sim \pi_{1|-1}(\xi_1 | \xi_2^{(0)})$.
- Step 2: Sample $\xi_2 \sim \pi_{2|-2}(\xi_2 | \xi_1^{(1)})$; then, we have $\boldsymbol{X}^{(1)} = (\xi_1^{(1)}, \xi_2^{(1)})^\intercal$.
- Repeat...

2 Problem Set

1. (5 points) (An application of the Metropolis-Hastings algorithm) Suppose we wish to sample from the Poisson distribution with parameter 10. That is, the target distribution is

$$\pi(x) = e^{-10}\frac{10^x}{x!}, \qquad x = 0, 1, 2, \ldots.$$

(The state space $\mathcal{X} = \{0, 1, 2, \ldots\}$ is infinite, but the Metropolis-Hastings algorithm still works.) Suppose the proposal transition probability $q(x, y)$ is defined as follows

$$\text{for all } x \geq 1, \quad q(x, y) = \begin{cases} 0.5 & \text{if } y = x \pm 1 \\ 0 & \text{otherwise;} \end{cases}$$

and

$$\text{for } x = 0, \quad q(0, y) = \begin{cases} 0.5 & \text{if } y = 1 \text{ or } 0 \\ 0 & \text{otherwise.} \end{cases}$$

Use the Metropolis-Hastings algorithm (Algorithm 2) to generate $\{X_0, X_1, \ldots, X_n\}$, starting from $X_0 = 0$, with $n = 10^4$.

   (a) Draw the histogram of $\{X_{25}, ..., X_{50}\}$.

   (b) Draw the histogram of $\{X_{50}, ..., X_{100}\}$.

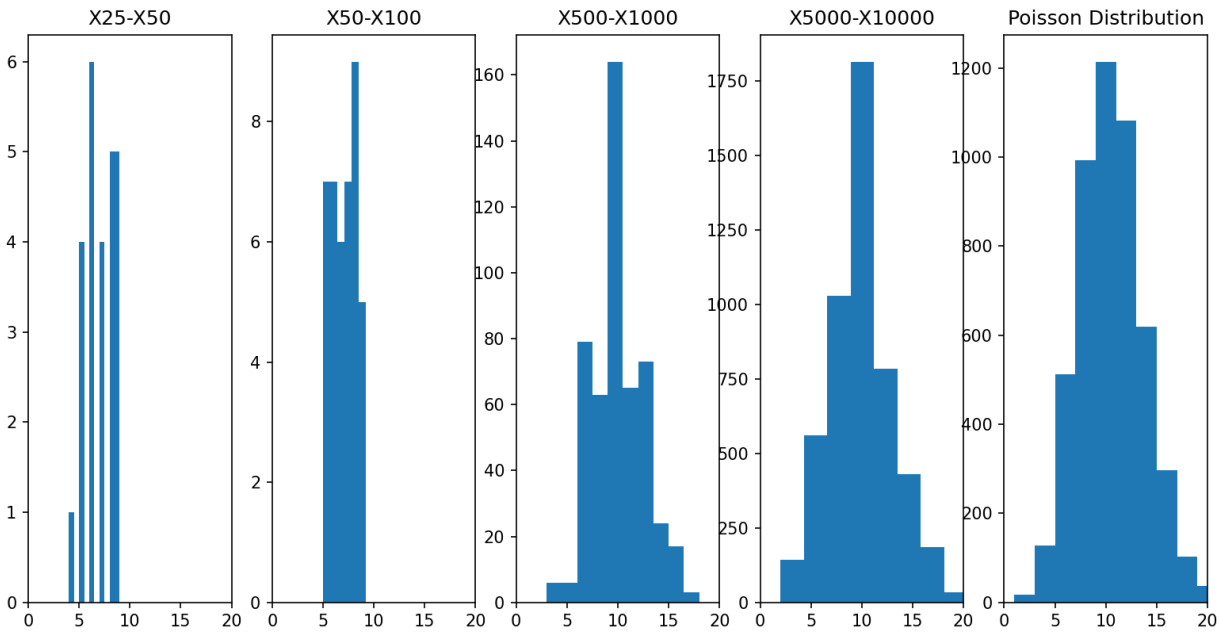
   (c) Draw the histogram of $\{X_{500}, ..., X_{1000}\}$.

   (d) Draw the histogram of $\{X_{5000}, ..., X_{10000}\}$.

   You will have four plots. Please provide the four plots and the code for generating $\{X_0, X_1, \ldots, X_n\}$ and the plots.

```python
import numpy as np
from scipy.stats import bernoulli, poisson
import matplotlib.pyplot as plt
import math

def pi(x: int) -> float:
    return math.exp(-10) * (10**x) / math.factorial(x)


def q(x: int, y: int) -> float:
    if x == 0:
        if (y == 1) or (y == 0):
            return 0.5
        else:
            return 0
    else:
        if (y == x+1) or (y == x-1):
            return 0.5
        else:
            return 0

def sample(old_state: int) -> int:
    if old_state == 0:
        return bernoulli.rvs(0.5, size=1)[0].item()
    else:
        return old_state + np.random.choice([-1, 1], size=1)[0].item()


def metropolis_hastings(n: int, x0: float) -> list[int]:
    X = []
    X.append(x0)

    for i in range(1, n):
        X_star = sample(X[i-1])
        r = (pi(X_star) * q(X_star, X[i-1])) / (pi(X[i-1]) * q(X[i-1], X_star))
        try:
            Y = bernoulli.rvs(min([r, 1]), size=1)[0].item()
        except:
            print(f"i: {i}, X_star: {X_star}, pi: {pi(X_star)}")

        Xn = Y*X_star + (1 - Y)* X[i-1]

        X.append(Xn)
        if i % 1000 == 0: print(i)

    return X
```

```
fig, (p1, p2, p3, p4, p5) = plt.subplots(1, 5)

x = metropolis_hastings(10**4, 0)
counts1, bins1 = np.histogram(x[25:50], 10)
p1.set_xlim(0, 20)
p1.hist(bins1[:-1], bins1, weights=counts1)
p1.set_title("X25-X50")

counts2, bins2 = np.histogram(x[50:100], 10)
p2.set_xlim(0, 20)
p2.hist(bins1[:-1], bins2, weights=counts2)
p2.set_title("X50-X100")

counts3, bins3 = np.histogram(x[500:1000], 10)
p3.set_xlim(0, 20)
p3.hist(bins3[:-1], bins3, weights=counts3)
p3.set_title("X500-X1000")

counts4, bins4 = np.histogram(x[5000:10000], 10)
p4.set_xlim(0, 20)
p4.hist(bins4[:-1], bins4, weights=counts4)
p4.set_title("X5000-X10000")

true_counts, true_bins = np.histogram(poisson(10).rvs(size=5000), 10)
p5.set_xlim(0, 20)
p5.hist(true_bins[:-1], true_bins, weights=true_counts)
p5.set_title("Poisson Distribution")

plt.show()
```



2. (5 points) Suppose we have a 3-dimensional PMF $\pi(\xi_1, \xi_2, \xi_3)$, and it is strictly positive. Please provide a transition probability $p(\boldsymbol{x}, \boldsymbol{y})$ with $\boldsymbol{x} = (\xi_1, \xi_2, \xi_3)$ and $\boldsymbol{y} = (\eta_1, \eta_2, \eta_3)$ such that π is the stationary distribution of p, i.e., π and p satisfy

$$\sum_{\xi_1}\sum_{\xi_2}\sum_{\xi_3} \pi(\xi_1, \xi_2, \xi_3) \cdot p\Big((\xi_1, \xi_2, \xi_3), (\eta_1, \eta_2, \eta_3)\Big) = \pi(\eta_1, \eta_2, \eta_3), \tag{2.1}$$

which is equivalent to $\sum_{\boldsymbol{x}} \pi(\boldsymbol{x}) p(\boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{y})$. Specifically, please provide the formula of the $p(\boldsymbol{x}, \boldsymbol{y})$ and prove Eq. (2.1).

(There is more than one correct answer. You only need to provide one answer.)

Hint: The Gibbs sampling algorithm (Algorithm 3) with dimensionality $d = 3$ is a hint.

Let $x = (\xi_1, \xi_2, \xi_3)^T$ and $y = (\eta_1, \eta_2, \eta_3)^T$ and then define

$$\boxed{p(x, y) := \pi_{3|-3}(\xi_1, \xi_2, \eta_3) \cdot \pi_{2|-2}(\xi_1, \eta_2, \eta_3) \cdot \pi_{1|-1}(\eta_1, \eta_2, \eta_3)}$$

Where

$$\pi_{3|-3}(\xi_1, \xi_2, \eta_3) = \frac{\pi(\xi_1, \xi_2, \eta_3)}{\pi_{-3}(\xi_1, \xi_2)}$$
$$\pi_{2|-2}(\xi_1, \eta_2, \eta_3) = \frac{\pi(\xi_1, \eta_2, \eta_3)}{\pi_{-2}(\xi_1, \eta_3)}$$
$$\pi_{1|-1}(\eta_1, \eta_2, \eta_3) = \frac{\pi(\eta_1, \eta_2, \eta_3)}{\pi_{-1}(\eta_2, \eta_3)}$$

So

$$
\begin{aligned}
&\sum_{\xi_1}\sum_{\xi_2}\sum_{\xi_3} \pi(\xi_1,\xi_2,\xi_3)\cdot p\Big((\xi_1,\xi_2,\xi_3),(\eta_1,\eta_2,\eta_3)\Big) \\
&\quad= \sum_{\xi_1}\sum_{\xi_2}\sum_{\xi_3} \pi(\xi_1,\xi_2,\xi_3)\cdot \pi_{3|-3}(\xi_1,\xi_2,\eta_3)\cdot \pi_{2|-2}(\xi_1,\eta_2,\eta_3)\cdot \pi_{1|-1}(\eta_1,\eta_2,\eta_3) \\
&\quad= \sum_{\xi_1}\sum_{\xi_2}\sum_{\xi_3} \pi(\xi_1,\xi_2,\xi_3)\cdot \frac{\pi(\xi_1,\xi_2,\eta_3)}{\pi_{-3}(\xi_1,\xi_2)}\cdot \frac{\pi(\xi_1,\eta_2,\eta_3)}{\pi_{-2}(\xi_1,\eta_3)}\cdot \frac{\pi(\eta_1,\eta_2,\eta_3)}{\pi_{-1}(\eta_2,\eta_3)} \\
&\quad= \sum_{\xi_1}\sum_{\xi_2} \pi_{-3}(\xi_1,\xi_2)\cdot \frac{\pi(\xi_1,\xi_2,\eta_3)}{\pi_{-3}(\xi_1,\xi_2)}\cdot \frac{\pi(\xi_1,\eta_2,\eta_3)}{\pi_{-2}(\xi_1,\eta_3)}\cdot \frac{\pi(\eta_1,\eta_2,\eta_3)}{\pi_{-1}(\eta_2,\eta_3)} \\
&\quad= \sum_{\xi_1}\sum_{\xi_2} \pi(\xi_1,\xi_2,\eta_3)\cdot \frac{\pi(\xi_1,\eta_2,\eta_3)}{\pi_{-2}(\xi_1,\eta_3)}\cdot \frac{\pi(\eta_1,\eta_2,\eta_3)}{\pi_{-1}(\eta_2,\eta_3)} \\
&\quad= \sum_{\xi_1} \pi_{-2}(\xi_1,\eta_3)\cdot \frac{\pi(\xi_1,\eta_2,\eta_3)}{\pi_{-2}(\xi_1,\eta_3)}\cdot \frac{\pi(\eta_1,\eta_2,\eta_3)}{\pi_{-1}(\eta_2,\eta_3)} \\
&\quad= \sum_{\xi_1} \pi(\xi_1,\eta_2,\eta_3)\cdot \frac{\pi(\eta_1,\eta_2,\eta_3)}{\pi_{-1}(\eta_2,\eta_3)} \\
&\quad= \pi_{-1}(\eta_2,\eta_3)\cdot \frac{\pi(\eta_1,\eta_2,\eta_3)}{\pi_{-1}(\eta_2,\eta_3)} \\
&\quad= \pi(\eta_1,\eta_2,\eta_3) \quad \blacksquare
\end{aligned}
$$

References

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

A. Klenke. *Probability theory: a comprehensive course, 3rd Edition*. Springer Science & Business Media, 2020.

N. Metropolis and S. Ulam. The monte carlo method. *Journal of the American statistical association*, 44(247):335–341, 1949.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.