APMA1690: Homework # 7 (Due by 11pm on November 9)

1 Review

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

1.1 Markov Chains vs. Markov Chain Monte Carlo

When we were talking about (homogeneous[1]) Markov chains, we were in a situation where a transition probability $p(x, y)$ of a Markov chain was given; we were asked to derive a/the stationary distribution π associated with $p(x, y)$.

We are now learning the theory of Markov Chain Monte Carlo (MCMC), and the situation is reversed — a distribution π is given, and we are asked to derive a transition probability $p(x, y)$ satisfying the following requirements

- the Markov chain associated with $p(x, y)$ is irreducible and aperiodic;
- the given distribution π is the stationary distribution of the Markov chain. (Since the Markov chain is irreducible, we are safe to use the phrase "**the** stationary distribution.")

1.2 Asymptotic Behaviors of Markov Chains

The MCMC method needs some asymptotic results about Markov chains as its theoretical foundations.

Theorem 1.1. *Let $\{X_n\}_{n=0}^{\infty}$ be a homogeneous Markov chain taking values in a finite state space $\mathcal{X} = \{x_1, \ldots, x_S\}$. If this Markov chain is **irreducible** and **aperiodic**, we have*

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{P}(X_n = x_j \,|\, X_0 = x_i) = \pi(x_j), \quad \textit{equivalently} \\ &\lim_{n\to\infty} (\boldsymbol{P}^n)_{ij} = \pi(x_j), \quad \textit{for all } i, j \in \{1, 2, \ldots, S\}, \end{aligned} \tag{1.1}$$

where π is the unique stationary distribution of the Markov chain.

Theorem 1.1 further implies the following

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}(X_n = x_j) &= \sum_{i=1}^{S} \Big[\lim_{n\to\infty} \mathbb{P}(X_n = x_j \,|\, X_0 = x_i) \Big] \cdot \mathbb{P}(X_0 = x_i) \\ &= \pi(x_j) \cdot \sum_{i=1}^{S} \mathbb{P}(X_0 = x_i) \\ &= \pi(x_j), \quad \text{for all } j = 1, \ldots, S, \end{aligned}$$

that is, X_n looks like a π-distributed random variable when n is sufficiently large.

[1] All the Markov chains referred to in APMA 1690 are homogeneous Markov chains.

1.3 Generating a Markov Chain from a Transition Probability

For any transition probability function $p(x, y)$, when x is fixed, the function $y \mapsto p(x, y)$ of y is a PMF.

Suppose what we know is a transition probability function p, instead of a Markov chain. With p, using the following conceptual algorithm, we can generate a Markov chain whose transition probability is the given p.

Algorithm 1 : Generating Markov Chains

Input: (i) transition probability p, and (ii) initialization x_0.
Output: a Markov chain $\{X_n\}_{n=0}^{\infty}$ whose transition probability is p.

1: $X_0 \leftarrow x_0$.
2: **for all** $n = 1, 2, \ldots$ **do**
3: Generate X_n from the PMF $p(X_{n-1}, \cdot)$.
4: **end for**

1.4 Main Theme of MCMC

In many applications, the distribution π of interest is defined in an extremely high-dimensional space. For example, if π is the random 256-by-256 binary-valued pictures (i.e., each picture has 256×256 pixels, and each pixel takes its value in the binary set $\{-1, 1\}$), the π is a PMF defined on a state space containing 2^{65536} elements.[2] It is **infeasible** to generate random variables **exactly** following such a high-dimensional distribution.

Suppose we can generate a Markov chain $\{X_n\}_{n=0}^{\infty}$ satisfying the following requirements

- $\{X_n\}_{n=0}^{\infty}$ is irreducible and aperiodic;
- the given distribution π is the stationary distribution of $\{X_n\}_{n=0}^{\infty}$.

Then, we have the following

$$\lim_{n\to\infty} \mathbb{P}(X_n = x) = \pi(x), \quad \text{for all } x.$$

Hence, when n is large, X_n **approximately** follows the given distribution π; we may approximately view X_n as a random variable generated from π.

For a given distribution π, two widely adopted methods of generating such a Markov chain are "Metropolis-Hastings algorithm" and "Gibbs sampling."

1.5 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 is concluded as the "Metropolis algorithm." Hastings (1970) generalized this algorithm.

Suppose π is the distribution of interest, and it is strictly positive, i.e.,

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

[2] Recall that $2^{10} = 1023 \approx 10^3$.

1.5.1 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[3][4]

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

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.

The following claims show that the $p(x, y)$ defined in Eq. (1.2) works.

Claim 1.2. *If $q(x, y)$ is irreducible and aperiodic[5], then $p(x, y)$ is irreducible and aperiodic.*

Proof: See the problem set section.

Claim 1.3. *The $p(x, y)$ defined by Eq. (1.2) satisfies the following equation[6]*

$$\pi(x)p(x,y) = \pi(y)p(y,x), \quad \text{for all } x, y \in \mathcal{X}. \tag{1.3}$$

Proof: See the problem set section.

Claim 1.4. *π is a stationary distribution of p.*

Proof: It comes from Claim 1.3. Specifically, Eq. (1.3) implies

$$\sum_{y\in\mathcal{X}} \pi(x)p(x,y) = \sum_{y\in\mathcal{X}} \pi(y)p(y,x),$$

where the left-hand side is equal to $\pi(x)\sum_{y\in\mathcal{X}} p(x, y) = \pi(x)$. Therefore, $\pi(x) = \sum_{y\in\mathcal{X}} \pi(y)p(y, x)$.

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

---

[3] The logic of Eq. (1.2) is the following: we first define $p(x, y)$ for all $x \neq y$, then we define $p(x, x) = 1 - \sum_{z:z\neq x} p(x, z)$.

[4] You may ask, *why is $p(x, x)$ defined separately?* Answer: If we simply define $p(x, x) = q(x, x)$, then p would not be a transition probability. The transition probability assumption requires $1 = \sum_{y\in\mathcal{X}} p(x, y)$. However, $\frac{\pi(y')}{\pi(x)}$ can be strictly smaller than 1 for some $y' \neq x$. In this scenario, if we did not define $p(x, x)$ separately — that is, we define $p(x, x) = q(x, x)$, we will have

$$1 = \sum_{y\in\mathcal{X}} p(x,y) = \sum_{y\in\mathcal{X}} q(x,y)\cdot\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} = q(x,x) + \sum_{y\neq x,y'} q(x,y)\cdot\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} + q(x,y')\cdot\min\left\{1, \frac{\pi(y')}{\pi(x)}\right\} < \sum_{y\in\mathcal{X}} q(x,y) = 1,$$

which is a contradiction. Hence, we need to define $p(x, x)$ separately to correct for the "loss of mass" in the term $q(x, y') \cdot \min\left\{1, \frac{\pi(y')}{\pi(x)}\right\}$.

[5] It actually means, "the Markov chain associated with $q(x, y)$ is irreducible and aperiodic;" for $p(x, y)$, similarly.

[6] If π satisfies Eq. (1.3), it is called a reversible distribution of the transition probability p (see Example 6.5.4 of Durrett (2010) or Eq. (18.8) of Klenke (2020)).

Algorithm 2 : Metropolis 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**

1.5.2 Metropolis-Hastings Algorithm

The Metropolis algorithm requires $q(x, y)$ to be symmetric. This symmetry condition can be removed by modifying Eq. (1.2) to the following form, which results in the Metropolis-Hastings algorithm.

$$p(x,y) := \begin{cases} q(x,y) \cdot \min\left\{1, \dfrac{\pi(y) \cdot q(y,x)}{\pi(x) \cdot q(x,y)}\right\}, & \text{if } x \neq y \text{ and } q(x,y) > 0, \\ 0, & \text{if } x \neq y \text{ and } q(x,y) = 0, \\ 1 - \displaystyle\sum_{z: z \neq x} p(x,z), & \text{if } x = y, \end{cases} \tag{1.4}$$

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. (1.2) and Eq. (1.4) is that the $q(x, y)$ in Eq. (1.4) 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 1.5. *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. (1.4) is irreducible and has the unique stationary distribution π. If, in additon, q is aperiodic, then p is aperiodic as well.*

Algorithm 3 can generate a Markov chain whose transition probability is the $p(x, y)$ defined in Eq. (1.4).

1.6 Applications of MCMC

When you were learning the theoretical foundation of MCMC (i.e., Theorem 1.1), you assumed that state spaces are finite. In applications, people directly use these MCMC formulas/algorithms for general scenarios, e.g., state spaces are infinite and continuous. It usually works because of the following

Algorithm 3 : Metropolois-Hastings Algorithm

Input: (i) the distribution π of interest; (ii) a jumping distribution — a transition probability q of an irreducible and aperiodic Markov chain; (iii) an initial starting point x_0.

Output: A Markov chain $\{X_n\}_{n=0}^{\infty}$ with π as its stationary distribution.

1: Set $X_0 \leftarrow x_0$.
2: **for all** $n = 1, 2, \ldots$ **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**

- The theoretical foundation for the general scenarios exists, although it involves very advanced mathematical tools and requires more conditions (e.g., see Tierney (1994) and Athreya et al. (1996)).
- In computers, everything is discrete and finite.

We will also adopt the application convention and apply the Metropolis-Hastings algorithm/Gibbs sampling to general scenarios, e.g., continuous distributions defined on continuous state spaces.

2 Problem Set

1. (3 points) Suppose $q(x, y)$ is the transition probability of a Markov chain, and $q(x, y)$ is symmetric, i.e., $q(x, y) = q(y, x)$ for all $x, y \in \mathcal{X}$. Let $p(x, y)$ be the function defined by Eq. (1.2). **Prove Claim 1.2.**

   Claim: If $q(x, y)$ is irreducible and aperiodic, then $p(x, y)$ is irreducible and aperiodic.

   Because $q(x, y) = q(y, x) \quad \forall x, y \in \mathfrak{X}$, it is clear that if $q(x, y) > 0$, then $q(y, x) > 0$ and vice versa. Then since q is irreducible and aperiodic, then by Theorem 1.5, the MC generated from

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

   is also irreducible and aperiodic.

   However, this formula can be simplified. By the symmetry of q,

$$\min\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\} = \min\{1, \frac{\pi(y)}{\pi(x)}\}$$

   In this version, if $q(x, y) = q(y, x) = 0$, there is no risk of division by 0 so we can combine the cases $q = 0$ and $q > 0$. Thus, we have that the MC generated by

$$p(x, y) = \begin{cases} q(x, y) \cdot \min\{1, \frac{\pi(y)}{\pi(x)}\} & q(x, y) \geq 0 \\ 1 - \sum_{z : z \neq x} p(x, z) & x = y \end{cases}$$

   is irreducible and aperiodic, which is exactly what we were trying to prove! ■

2. (3 points) Suppose $q(x, y)$ is the transition probability of a Markov chain, and $q(x, y)$ is symmetric, i.e., $q(x, y) = q(y, x)$ for all $x, y \in \mathcal{X}$. Let $p(x, y)$ be the function defined by Eq. (1.2). **Prove Claim 1.3.**

Claim: $\pi(x)p(x, y) = \pi(y)p(y, x) \qquad \forall x, y \in \mathfrak{X}$ with

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

Proof:

If $y = x$, then clearly

$$\pi(x)p(x, y) = \pi(y)p(y, x)$$

and we are done.

Otherwise,

$$\pi(x)p(x, y) = \pi(x)q(x, y) \cdot \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$$

This gives us two more cases.

If $\pi(x) > \pi(y)$, then $\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} = \frac{\pi(y)}{\pi(x)}$, so

$$\begin{aligned} \pi(x)p(x, y) &= \pi(x)q(x, y) \cdot \frac{\pi(y)}{\pi(x)} \\ &= \pi(y)q(x, y) \\ &= \pi(y)q(y, x) \end{aligned}$$

Since $q(y, x) = \frac{p(y,x)}{\min\{1, \frac{\pi(x)}{\pi(y)}\}}$ and $\pi(x) > \pi(y)$, the denominator is 1 and $q(y, x) = p(y, x)$ so

$$\pi(x)p(x, y) = \pi(y)p(y, x)$$

Finally, if $\pi(x) \leq \pi(y)$, then $\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} = 1$, so

$$\pi(x)p(x, y) = \pi(x)q(x, y) = \pi(x)q(y, x)$$

As above,

$$q(y, x) = \frac{p(y, x)}{\min\{1, \frac{\pi(x)}{\pi(y)}\}}$$

but the min function equals $\frac{\pi(x)}{\pi(y)}$ so

$$q(y, x) = \frac{\pi(y)}{\pi(x)} p(y, x)$$

and

$$\pi(x)q(y,x) = \pi(y)p(y,x)$$

So

$$\pi(x)p(x,y) = \pi(y)p(y,x)$$

for all $x, y \in \mathfrak{X}$. ■

3. (An application of the Metropolis algorithm) The distribution π of interest is the PDF of the following multivariate normal distribution

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.8\\0.8 & 1\end{pmatrix}\right).$$

The transition probability q (see Eq.(1.2)) is defined by the following: for each fixed $\boldsymbol{x} = (x_1, x_2)$, let $q(\boldsymbol{x}, \cdot)$ be the PDF of the following multivariate normal distribution

$$N\left(\begin{pmatrix}x_1\\x_2\end{pmatrix}, \begin{pmatrix}\sigma^2 & 0\\0 & \sigma^2\end{pmatrix}\right).$$

This code contains my implementation of the Metropolis algorithm and will be used in every part below:

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

def pi(x):
    return multivariate_normal.pdf(x, mean=[0, 0], cov=[[1, 0.8], [0.8, 1]])

def q(x, sigma):
    return multivariate_normal.rvs(mean=[x[0], x[1]], cov=[[sigma**2, 0], [0, sigma**2]], size=1)

def metropolis(n, x0, sigma):
    X = []
    X.append(x0)

    for i in range(1, n):
        X_star = q(X[i-1], sigma)
        r = pi(X_star) / pi(X[i-1])

        Y = bernoulli.rvs(min([r, 1]), size=1)

        Xn = np.dot(Y[0], X_star) + np.dot((1 - Y[0]), X[i-1])

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

    return X

def split_vector(lst):
    x1 = list(map(lambda i: i[0], lst))
    x2 = list(map(lambda i: i[1], lst))
    return x1, x2
```

(a) (1 point) Let $\sigma = 0.7$. Using π, q, and $\boldsymbol{x}_0 = (-2, 2)$ as inputs, generate the first 20,000 components of a Markov chain, i.e., $\{X_n\}_{n=0}^{20000}$, using the Metropolis algorithm (Algorithm 2). Plot the second half of the sequence, i.e., $\{X_n\}_{n=10001}^{20000}$. Provide your code generating the plot. (Please feel free to use any code I uploaded to Canvas.)

(b) (0.5 points) Replace $\boldsymbol{x}_0$ with $(2, 2)$ and repeat part (a).

(c) (0.5 points) Generate 20,000 data points from π and plot these points. Provide your code generating the plot.

I generated the plots for A, B, and C together using this code:

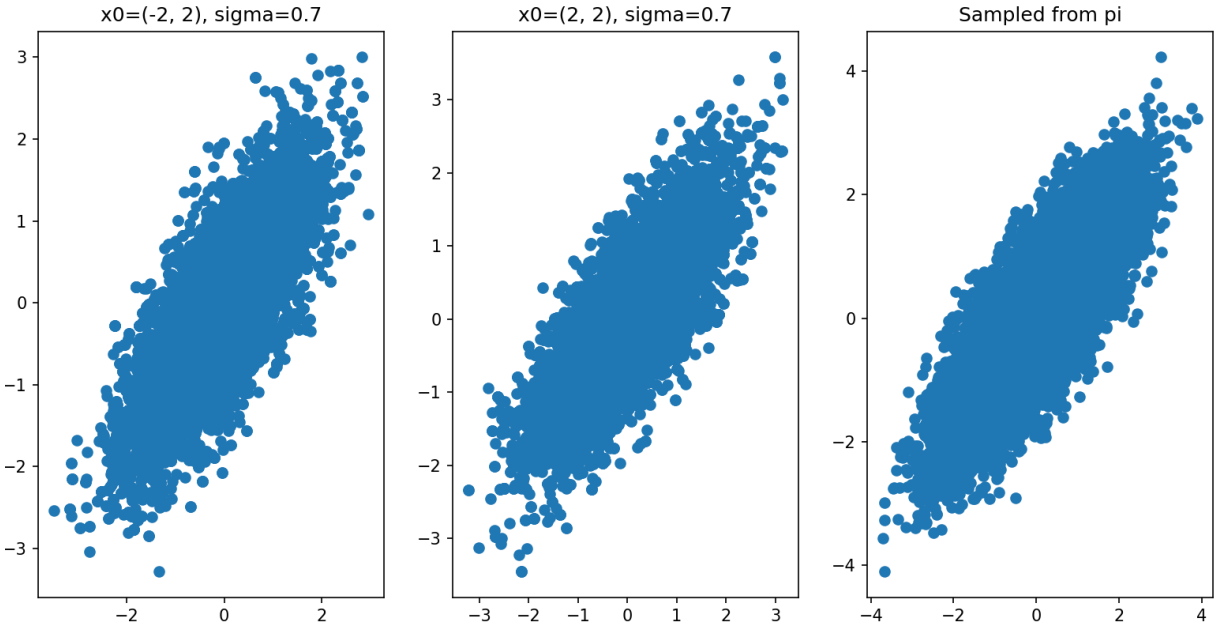
```
def abc():
    fig, (p1, p2, p3) = plt.subplots(1, 3)

    a1, a2 = split_vector(metropolis(20000, (-2, 2), 0.7)[10000:])
    p1.scatter(a1, a2)
    p1.set_title('x0=(-2, 2), sigma=0.7')

    b1, b2 = split_vector(metropolis(20000, (2, 2), 0.7)[10000:])
    p2.scatter(b1, b2)
    p2.set_title('x0=(2, 2), sigma=0.7')

    c1, c2 = split_vector(multivariate_normal.rvs(mean=[0, 0], cov=[[1, 0.8], [0.8, 1]], size=20000))
    p3.scatter(c1, c2)
    p3.set_title('Sampled from pi')
    plt.show()
```

which resulted in the following plot where the graph from part A is the lefmost subplot, part B is the center, and part C is the rightmost:

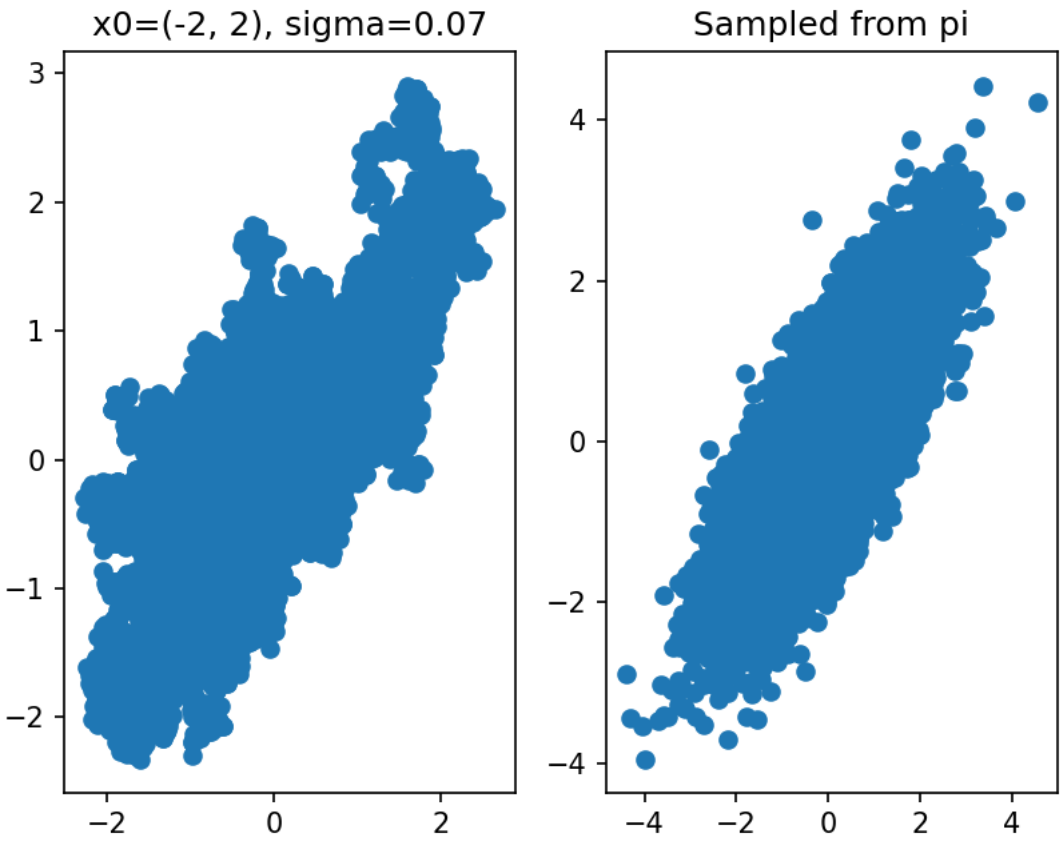


(d) (1 point) Replace σ with 0.07 and repeat part (a). Compare the plot you got in part (d) with the one you got in part (c). Are the two plots similar? If not, please explain the reason why they are not similar.

```
def d():
    fig, (p1, p3) = plt.subplots(1, 2)

    d1, d2 = split_vector(metropolis(20000, (-2, 2), 0.07)[10000:])
    p1.scatter(d1, d2)
    p1.set_title('x0=(-2, 2), sigma=0.07')

    c1, c2 = split_vector(multivariate_normal.rvs(mean=[0, 0], cov=[[1, 0.8], [0.8, 1]], size=20000))
    p3.scatter(c1, c2)
    p3.set_title('Sampled from pi')
    plt.show()
```



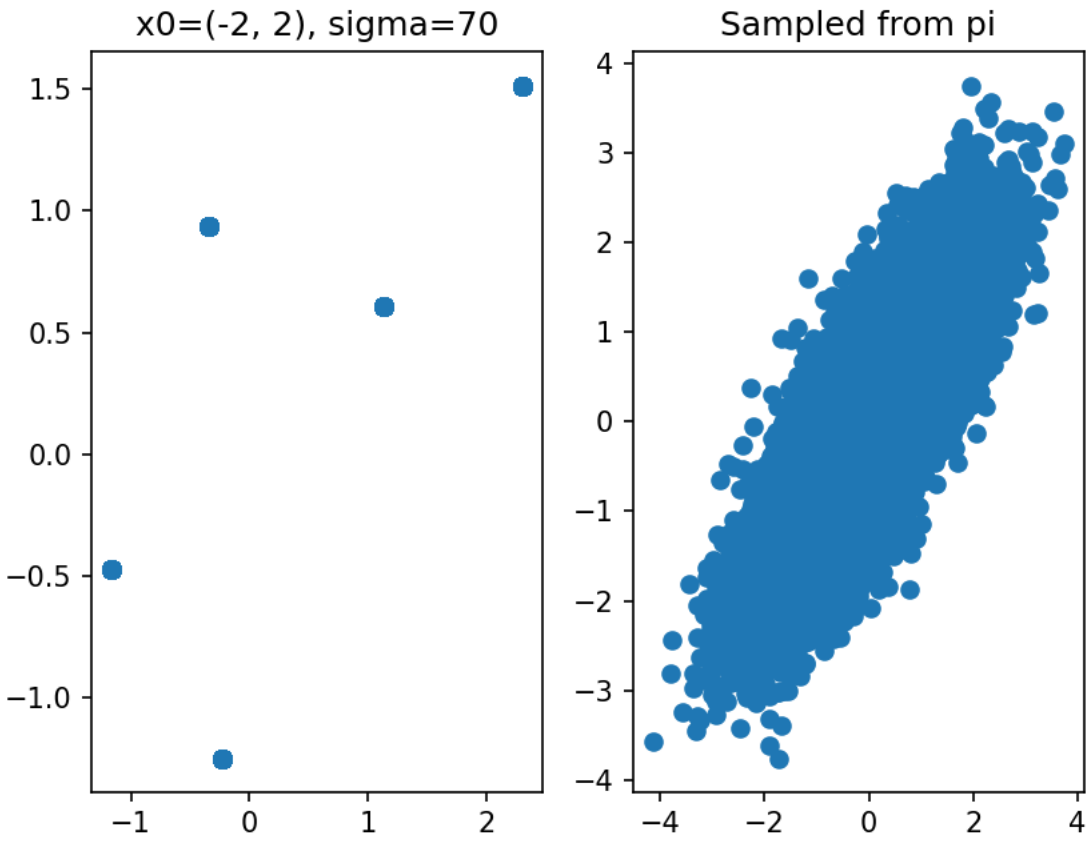
At first glance, the two plots look somewhat similar. On closer inspection, the plot limits do not align: the approximation in the left subplot ranges from $[-2, 2]$ in one dimension and $[-2, 3]$ in the other while the RVs sampled from π are in the range $x_1 \in [-4, 4], x_2 \in [-4, 4]$. With very small σ, the distance between any X^* and X_{n-1} is small so $r = \frac{\pi(X^*)}{\pi(X_{n-1})}$ tends to 1 so $Y \sim \text{Bernoulli}(\min\{1, 1\}) \to 1$ So $X_n \approx X^*$ and there is little change in the algorithm so it does not approximate π very well.

(e) (1 point) Replace σ with 70 and repeat part (a). Compare the plot you got in part (e) with the one you got in part (c). Are the two plots similar? If not, please explain the reason why they are not similar.

```
def e():
    fig, (p1, p3) = plt.subplots(1, 2)

    e1, e2 = split_vector(metropolis(20000, (-2, 2), 70)[10000:])
    p1.scatter(e1, e2)
    p1.set_title('x0=(-2, 2), sigma=70')

    c1, c2 = split_vector(multivariate_normal.rvs(mean=[0, 0], cov=[[1, 0.8], [0.8, 1]], size=20000))
    p3.scatter(c1, c2)
    p3.set_title('Sampled from pi')
    plt.show()
```



Like the above graph, this one does not resemble the plot of RVs drawn from π. When σ is very large, $r = \frac{\pi(X^*)}{\pi(X_{n-1})}$ tends to be small so Y tends to 0 and $X_n \approx X_{n-1}$. As a result, the model "gets stuck" and only a few points are fit very often.

For either part (d) or part (e), try your best to make your explanations convincing to the TAs who grade this question.

Hints for parts (d) and (e): You may need to consider the following quantities

- the distance between any two consecutive steps X_{n-1} and X_n,
- the value $r = \frac{\pi(X^*)}{\pi(X_{n-1})}$ in Algorithm 2 and the mechanism of $Y \sim \text{Bernoulli}(\min\{1, r\})$.

References

K. B. Athreya, H. Doss, and J. Sethuraman. On the convergence of the markov chain simulation method. *The Annals of Statistics*, 24(1):69–100, 1996.

R. Durrett. *Probability: theory and examples, 4th Edition*, volume 49. Cambridge university press, 2010.

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.

L. Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.