APMA 1690: Homework Index

- MCMC
- Asymptotic behavior of Markov Chain
- Generating Markov chains from a transition probability
- Metropolis-Hastings
- Multivariate normal

HW 8 (p. 70-78)
- Metropolis-Hastings
- Gibbs sampling

HW 9 (p. 79-94)
- Graph theory
- Periodic lattices
- Ising Model
- Curie Temperature
- Conditional distributions
- Neighborhoods and cliques

HW 10 (p. 95-
- Manifold learning/ PCA
- Covariance sampling
- Matrix diagonalization

APMA1690: Homework # 1 (Due by 11 pm Sept 21)

1 Review

I would suggest you read the Review section before going to the problem set. - Mike

1.1 Three Building Blocks of Probability Theory

1.1.1 Sample Spaces

Definition 1.1. *1. The **sample space** of an experiment is the collection of all possible outcomes of the experiment. A sample space is usually denoted by Ω.*

*2. Any subset A of Ω (allowed to be empty $\emptyset$) is called an **event**, and $\emptyset$ is called the/an **impossible event**. The sample space Ω, as a subset of itself, is called the **inevitable event**.*

1.1.2 Random Variables

Definition 1.2. *Let Ω be a sample space and $\mathbb{R}^d$ denote d-dimensional space.*

- *Any map $X : \Omega \to \mathbb{R}^d$, $\omega \mapsto X(\omega)$ is called a ($\mathbb{R}^d$-valued) **random variable**; when $d \geq 2$, the $\mathbb{R}^d$-valued random variable is also referred to as a **random vector**.*
- *If there exists a fixed $x \in \mathbb{R}^d$ such that $X(\omega) = x$ for all $\omega \in \Omega$, i.e., $X(\omega)$ is a constant function of ω, we call X **deterministic**.*
- *If random variable X is not deterministic, we call X **truly random**.*

1.1.3 Probabilities

Definition 1.3. *Let Ω be a sample space. Suppose $\mathbb{P}$ is a real-valued function of subsets of Ω, i.e.,*

$$\begin{aligned} \mathbb{P}: \ & \{\text{subsets of } \Omega\} \to \mathbb{R}, \\ & A \mapsto \mathbb{P}(A). \end{aligned}$$

*If $\mathbb{P}$ satisfies the following three axioms, $\mathbb{P}$ is called a **probability**, and the pair $(\Omega, \mathbb{P})$ is called a **probability space***

1. *$\mathbb{P}(A) \geq 0$ for any subset $A \subseteq \Omega$;*
2. *$\mathbb{P}(\Omega) = 1$;*
3. *For any infinitely long sequence of disjoint subsets $\{A_i\}_{i=1}^{\infty}$, i.e., $A_i \cap A_j = \emptyset$ if $i \neq j$, we have*

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

1.2 Properties of Probabilities

Theorem 1.1. *Let $(\Omega, \mathbb{P})$ be a probability space. Then, we have the following*

1. $\mathbb{P}(\emptyset) = 0$, *i.e., the probability of the impossible event is zero;*
2. *if two events E_1 and E_2 satisfy $E_1 \cap E_2 = \emptyset$, we have $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2)$;*
3. *suppose $A, B \subseteq \Omega$. If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$;*
4. $0 \leq \mathbb{P}(A) \leq 1$ *for any subsets $A \subseteq \Omega$;*
5. *for any $A, B \subseteq \Omega$, we have $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$;*
6. *for any sequence of subsets $\{A_n\}_{n=1}^{\infty}$, we have $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.*

1. Let $A_1 = \Omega$ and $A_n = \emptyset$ for all $n \geq 2$. Then, $\{A_n\}_{n=1}^{\infty}$ is a sequence of disjoint sets. We have

$$A_1 = \Omega = \Omega \cup \emptyset \cup \emptyset \cdots \cup \emptyset \cup \cdots = \bigcup_{n=1}^{\infty} A_n,$$

   which implies $\mathbb{P}(A_1) = \mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n) = \mathbb{P}(A_1) + \mathbb{P}(\emptyset) + \mathbb{P}(\emptyset) + \cdots + \mathbb{P}(\emptyset) + \cdots$. We cancel $\mathbb{P}(A_1)$ and get the following

$$0 = \mathbb{P}(\emptyset) + \mathbb{P}(\emptyset) + \cdots + \mathbb{P}(\emptyset) + \cdots.$$

   Since the definition of probability enforce $\mathbb{P}(\emptyset) \geq 0$, we have $\mathbb{P}(\emptyset) = 0$.

2. Let $A_1 = E_1$, $A_2 = E_2$, and $A_n = \emptyset$ for $n \geq 3$. Then, $\{A_n\}_{n=1}^{\infty}$ is a sequence of disjoint sets. We have

$$E_1 \cup E_2 = A_1 \cup A_2 \cup \emptyset \cup \emptyset \cup \cdots \cup \emptyset \cup \cdots = \bigcup_{n=1}^{\infty} A_n,$$

   which implies

$$\begin{aligned} \mathbb{P}(E_1 \cup E_2) &= \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) + \sum_{n=1}^{\infty} \mathbb{P}(A_n) \\ &= \mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(\emptyset) + \mathbb{P}(\emptyset) + \cdots + \mathbb{P}(\emptyset) + \cdots \\ &= \mathbb{P}(E_1) + \mathbb{P}(E_2). \end{aligned}$$

3. $B = A \cup (B - A)$. Since $A \cap (B - A) = \emptyset$, we have $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B - A)$. Because $\mathbb{P}(B - A) \geq 0$, we have $\mathbb{P}(B) \geq \mathbb{P}(A)$.

The proofs of other results are left for homework.

1.3 Indicator Functions

Let A be a subset of $\mathbb{R}^d$. The **indicator function** $\mathbf{1}_A$ of A is defined as

$$\mathbf{1}_A(x) := \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

The function $\mathbf{1}_A(x)$ is sometimes represented as $\mathbf{1}(x \in A)$.

1.4 Cumulative Distribution Functions (CDFs)

Definition 1.4. *Let X be an $\mathbb{R}$-valued random variable defined on an underlying probability space $(\Omega, \mathbb{P})$. The function F_X defined as follows*

$$F_X(t) := \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \leq t\}\right) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in (-\infty, t]\}\right), \quad \text{for all } t \in \mathbb{R}, \tag{1.2}$$

*is called the **cumulative distribution function** (CDF) of X, which is denoted as $X \sim F_X$. (F_X is sometimes briefly denoted by F.)*

Remark: $F_X(t)$ is defined for **all** real numbers $t \in \mathbb{R}$.

2 Problem Set

1. (2 points) Suppose $\Omega = \{1, 2, \cdots, n\}$ is the sample space of interest, where $n < +\infty$ is a positive integer. For any subset (i.e., event) $A \subseteq \Omega$, we define

$$\mathbb{P}(A) := \frac{\#A}{n}$$

where $\#A$ denotes the number of elements in A. Please verify that the $\mathbb{P}$ defined above is a probability.

To be a probability, $\mathbb{P}$ must satisfy three axioms:

(a) $\mathbb{P}(A) \geq 0 \quad A \subset \Omega$
To see that this is true, observe that $A \subset \Omega$ so $0 \leq \#A \leq n$. Hence,

$$\frac{0}{n} \leq \frac{\#A}{n} \leq \frac{n}{n}$$

By definition of $\mathbb{P}$,

$$0 \leq \mathbb{P}(A) \leq 1$$

which is a stronger condition than $\mathbb{P}(A) \geq 0$

(b) $\mathbb{P}(\Omega)$ By definition of $\mathbb{P}$,

$$\mathbb{P}(\Omega) = \frac{\#\Omega}{n} = \frac{n}{n} = 1$$

(c) For any sequence of disjoint subsets, $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$
Let $A := \bigcup_{i=1}^{m} A_i$ be a sequence of m disjoint events in Ω. Then $\#A =$ so

$$\mathbb{P}\left(\bigcup_{i=1}^{m} A_i\right) = \mathbb{P}(A) = \frac{\#A}{n} = \frac{m}{n}$$

But as m is a positive integer,

$$\frac{m}{n} = \sum_{i=1}^{n} \frac{1}{n}$$

Then as each A_i is disjoint, $\mathbb{P}(A_i) = \frac{1}{n}$ so

$$\frac{m}{n} = \sum_{i=1}^{m} \mathbb{P}(A_i) = \mathbb{P}\left(\bigcup_{i=1}^{m} A_i\right)$$

However, this does not depend on the finiteness of m so

$$\sum_{i=1}^{\infty} \mathbb{P}(A_i) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right)$$

Then, as $\mathbb{P}$ satisfies all three requirements, it is a probability. ■

2. Please prove the results iv), v), and vi) of Theorem 1.1 (see the Review section), i.e.,

- (1 point) $0 \leq \mathbb{P}(A) \leq 1$ for any subsets $A \subseteq \Omega$;
  As $(\mathbb{P}, \Omega)$ is a probability space, $\mathbb{P}(A) \geq 0$. But as $A \subseteq \Omega$, $\mathbb{P}(A) \leq \mathbb{P}(\Omega)$. Thus by the definition of a probability,

$$0 \leq A \leq \mathbb{P}(\Omega) = 1 \implies 0 \leq A \leq 1 \quad \blacksquare$$

- (1 point) for any $A, B \subseteq \Omega$, we have $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$;
  For any $A, B \subseteq \Omega$, $A \cap (B \cap A^c) = \emptyset$ so from Property 2,

$$\mathbb{P}(A \cup (B \cap A^c)) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$$

  Additionally,

$$(B \cap A^c) \cap (B \cap A) = \emptyset$$

  and partition Ω. Therefore,

$$\mathbb{P}(B) = \mathbb{P}(B \cap A^c) + \mathbb{P}(B \cap A)$$

  Rearranging,

$$\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(B \cap A)$$

  so together with the first equation,

$$\begin{aligned} \mathbb{P}(A \cup (B \cap A^c)) &= \mathbb{P}(A) + \mathbb{P}(B \cap A^c) \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(B \cap A) \end{aligned}$$

  Finally, observe that

$$\begin{aligned} A \cup (B \cap A^c) &= (A \cup B) \cap (A \cup A^c) \\ &= (A \cup B) \cap \Omega \\ &= A \cup B \end{aligned}$$

  so

$$\mathbb{P}(A \cup (B \cap A^c)) = \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(B \cap A) \quad \blacksquare$$

- (1 point) for any sequence of subsets $\{A_n\}_{n=1}^{\infty}$, we have $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.
  In the case where $\{A_n\}_{n=1}^{\infty}$ is mutually disjoint, the equality follows trivially from Axiom 3. When the sequence is not mutually disjoint, we observe that

$$\mathbb{P}(A_1 \cup A_2) \leq \mathbb{P}(A) + \mathbb{P}(A_2)$$

  because $\mathbb{P}(A \cap A_2) \geq 0$.

Now to establish the inductive step, we see that

$$\begin{aligned}\mathbb{P}(A_1 \cup A_2 \cup A_3) &= \mathbb{P}(A_1 \cup A_2) + \mathbb{P}(A_3) - \mathbb{P}((A_1 \cup A_2) \cap A_3) \\ &= \mathbb{P}(A_1) + \mathbb{P}(B_2) + \mathbb{P}(A_3) - \mathbb{P}(A_1 \cap A_2) - \mathbb{P}((A_1 \cup A_2) \cap A_3)\end{aligned}$$

with $\mathbb{P}(A_1 \cap A_2) \geq 0$ and $\mathbb{P}((A_1 \cup A_2) \cap A_3) \geq 0$ so

$$\mathbb{P}(A_2 \cup A_2 \cup A_3) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3)$$

That is, for $n \geq 2$,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = -\tilde{P} + \sum_{i=1}^{n} \mathbb{P}(A_i)$$

where $\tilde{P} \geq 0$ is the sequence of intersections of earlier n. Thus,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i) \quad \blacksquare$$

Since the results *i)*, *ii)*, and *iii)* have been proved in the Review section, you can directly apply these results (i.e., results *i)*, *ii)*, and *iii)*) in your proofs.

3. Suppose the sample space of interest is $\Omega = [0,1] = \{\text{all the real numbers that are } \geq 0 \text{ and } \leq 1\}$. For any subset (i.e., event) $A \subseteq [0,1] = \Omega$, we define

$$\mathbb{P}(A) = \int_0^1 \mathbf{1}_A(x)\, dx,$$

where $\mathbf{1}_A(x)$ is the indicator function of A (see Eq. (1.1)). The $\mathbb{P}$ defined above is a probability (you do not need to prove this fact).

Let X be a random variable defined by

$$X(\omega) = \omega + 1,$$

for all $\omega \in \Omega = [0,1]$.

(a) (1 point) Consider the event

$$A = \{\omega \in \Omega \,|\, X(\omega) = 1.5\}. \tag{2.1}$$

Please calculate the probability of the event A defined in Eq. (2.1), i.e., $\mathbb{P}(A)$.

$$X = 1.5 \implies \omega = 0.5$$

$$\begin{aligned}
\mathbb{P}(A) &= \int_0^1 \mathbb{1}_A(x)\, dx \\
&= \int_0^{0.5} \mathbb{1}_A(x)\, dx + \int_{0.5}^{0.5} \mathbb{1}_A(x)\, dx + \int_{0.5}^1 \mathbb{1}_A(x)\, dx \\
&= \int_0^{0.5} 0\, dx + \int_{0.5}^{0.5} 1\, dx + \int_{0.5}^1 0\, dx \\
&= 0 + 0 + 0 = \boxed{0}
\end{aligned}$$

(b) (0.5 points) Is the event A defined in Eq. (2.1) an impossible event?
Despite the fact that $\mathbb{P}(A) = 0$, the event $\boxed{\text{is not impossible}}$. $X(\omega) = 1.5 = 0.5 + 1$ occurs when $\omega = 0.5 \in [0,1]$ so the event can occur.

(c) (0.5 points) Consider the event

$$B = \{\omega \in \Omega \,|\, X(\omega) = 0.5\}. \tag{2.2}$$

Please calculate the probability of the event B defined in Eq. (2.2), i.e., $\mathbb{P}(B)$.

$$X = 0.5 \implies \omega = -0.5$$

But $-0.5 \notin [0,1]$ so

$$\mathbb{P}(A) = \int_0^1 \mathbb{1}_A(x)\, dx = \int_0^1 0\, dx = \boxed{0}$$

(d) (0.5 points) Is the event B defined in Eq. (2.2) an impossible event?
$\omega \notin [0,1] = \Omega$ so it is $\boxed{\text{impossible}}$.

(e) (0.5 points) Please calculate the CDF of X.

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\omega + 1 \leq x) = \begin{cases} 0 & x < 1 \\ 1 & 1 \leq x \leq 2 \\ 0 & x > 2 \end{cases}$$

4. (2 points) Let $(\Omega, \mathbb{P})$ be a probability space and X a random variable defined on Ω. Please prove that the CDF $F_X(t)$ of X is a non-decreasing function, i.e., $F_X(t_1) \leq F_X(t_2)$ if $t_1 \leq t_2$.

By definition,

$$F_X(t_1) = \mathbb{P}(X \leq t_1) = \mathbb{P}(X \in (-\infty, t_1])$$
$$F_X(t_2) = \mathbb{P}(X \leq t_2) = \mathbb{P}(X \in (-\infty, t_2])$$

However, if $t_1 \leq t_2$,

$$\begin{aligned} F_X(t_2) &= \mathbb{P}(X \in (-\infty, t_1] \cup (t_1, t_2]) \\ &= \mathbb{P}(-\infty < X \leq t_1) + \mathbb{P}(t_1 < X \leq t_2) \\ &= F_X(t_1) + \mathbb{P}(t_1 < X \leq t_2) \end{aligned}$$

and by the fact that $\mathbb{P}(A) \geq 0 \quad A \in \Omega$, $\mathbb{P}(t_1 < X \leq t_2) \geq 0$ so

$$F_X(t_2) - F_X(t_1) \geq 0$$

and

$$F_X(t_2) \geq F_X(t_1) \quad \blacksquare$$

APMA1690: Homework # 2 (Due by 11pm Sep 28)

1 Question 1

1.1 Review before Question 1: the Law of Large Numbers (LLN)

Theorem 1.1 (Etemadi's strong LLN, 1981, see Etemadi (1981) or Theorem 5.17 of Klenke (2020))**.** *Let $Z_1, Z_2, \ldots$ be real-valued random variables defined on probability space $(\Omega, \mathbb{P})$. Suppose they are identically distributed and pairwise independent (i.e., Z_i and Z_j are independent for all i, j with $i \neq j$). If $\mathbb{E}Z_1$ exists, then $\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^n Z_i\right) = \mathbb{E}Z_1$ with probability one, i.e.,*

$$\mathbb{P}\left\{\omega \in \Omega : \lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^n Z_i(\omega)\right) = \mathbb{E}Z_1\right\} = 1. \tag{1.1}$$

1.2 Setup of Question 1

Let $(\Omega, \mathbb{P})$ be a probability space as follows

- The set Ω is defined by

$$\Omega = \underbrace{\{0,1\} \times \{0,1\} \times \cdots \times \{0,1\} \times \cdots}_{\text{Cartesian product, infinitely many } \{0,1\}},$$

  where each element $\omega \in \Omega$ is an infinitely long vector of the form $\omega = \underbrace{(\omega_1, \omega_2, \ldots, \omega_n, \ldots)}_{\text{infinitely many entries}}$,

  and $\omega_i \in \{0,1\}$ for any positive integer i.

- $\mathbb{P}$ is a probability and satisfies the following: for any positive integer i, we have

$$\mathbb{P}\{\omega = (\omega_1, \omega_2, \ldots, \omega_n, \ldots) \in \Omega \,|\, \omega_i = 1\} = \mathbb{P}\{\omega = (\omega_1, \omega_2, \ldots, \omega_n, \ldots) \in \Omega \,|\, \omega_i = 0\} = \frac{1}{2};$$

  furthermore, for any positive integer k and any vector

$$\boldsymbol{v} = (v_1, v_2, \ldots, v_k) \in \underbrace{\{0,1\} \times \cdots \times \{0,1\}}_{\text{Cartesian product, } k \text{ times}},$$

  we have $\mathbb{P}\{\omega = (\omega_1, \omega_2, \ldots, \omega_n, \ldots) \in \Omega \,|\, \omega_i = v_i \text{ for all } i = 1, 2, \ldots, k\} = (1/2)^k$.

We define the following two subsets of Ω

$$\Omega_1 = \left\{\omega = (\omega_1, \omega_2, \ldots, \omega_n, \ldots) \in \Omega \,\middle|\, \lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^n \omega_i\right] = \frac{1}{2}\right\},$$

$$\Omega_0 = \left\{\omega = (\omega_1, \omega_2, \ldots, \omega_n, \ldots) \in \Omega \,\middle|\, \lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^n \omega_i\right] \neq \frac{1}{2}\right\};$$

1.3 Question 1

a (1 points) For the following two elements $\omega^{(1)}$ and $\omega^{(2)}$ in Ω, which of them belongs to Ω_1? Which of them belongs to Ω_0? Explain (rather than mathematically prove) your answer.

$$\omega^{(1)} = (1,0,0,1,0,0,1,0,0,\ldots), \quad \text{(repetition of the "1,0,0" pattern)},$$
$$\omega^{(2)} = (1,0,1,0,1,0,1,0,\ldots), \quad \text{(repetition of the "1,0,1,0" pattern)}.$$

Both elements are members of Ω_1. By the law of large numbers, the average should tend towards the expected value which is $1/2$ because having a 1 or a 0 is equally likely at each index.

b (2 points) Prove the following using the LLN in a rigorous way

$$\mathbb{P}(\Omega_1) = 1 \quad \text{and} \quad \mathbb{P}(\Omega_0) = 0.$$

Let $\{X_i\}_{i=1}^{\infty}$ be an infinite sequence of iid random variables such that $X_1 \sim \text{Bernoulli}(\frac{1}{2})$. Then

$$\mathbb{E}X_1 = 0 \cdot \mathbb{P}(X=0)1 \cdot \mathbb{P}(X=1) = \mathbb{P}(X=1) = \frac{1}{2}$$

and

$$\begin{aligned}\mathbb{P}(\Omega_1) &= \mathbb{P}\left(\{\vec{\omega} \in \Omega \middle| \lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=0}^{n} X_i\right] = \frac{1}{2}\}\right) \\ &= \mathbb{P}\left(\{\vec{\omega} \in \Omega \middle| \lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=0}^{n} X_i\right] = \mathbb{E}X_1\}\right) \\ &\overset{LLN}{=} 1\end{aligned}$$

Then because Ω_1 and Ω_0 partition Ω, we know that

$$\mathbb{P}(\Omega) = \mathbb{P}(\Omega_1) + \mathbb{P}(\Omega_0) \implies 1 = 1 + \mathbb{P}(\Omega_0) \implies \mathbb{P}(\Omega_0) = 0 \quad \blacksquare$$

2 Question 2

2.1 Review before Question 2

We have the following two versions of the Law of the Iterated Logarithm

- (Heuristic/sloppy version.) Suppose $X_1, \ldots, X_n, \ldots$ are iid random variables defined on the probability space $(\Omega, \mathbb{P})$. The mean is $\mu = \mathbb{E}X_1$ and the standard deviation is $\sigma = \sqrt{Var(X_1)}$. Then, we "approximately" (rather than "exactly") have the following inequality when the sample size n is very large, where "log" is the natural logarithm.

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i(\omega) - \mu \right| \leq \sigma \cdot \sqrt{\frac{2 \cdot \log(\log n)}{n}}.$$

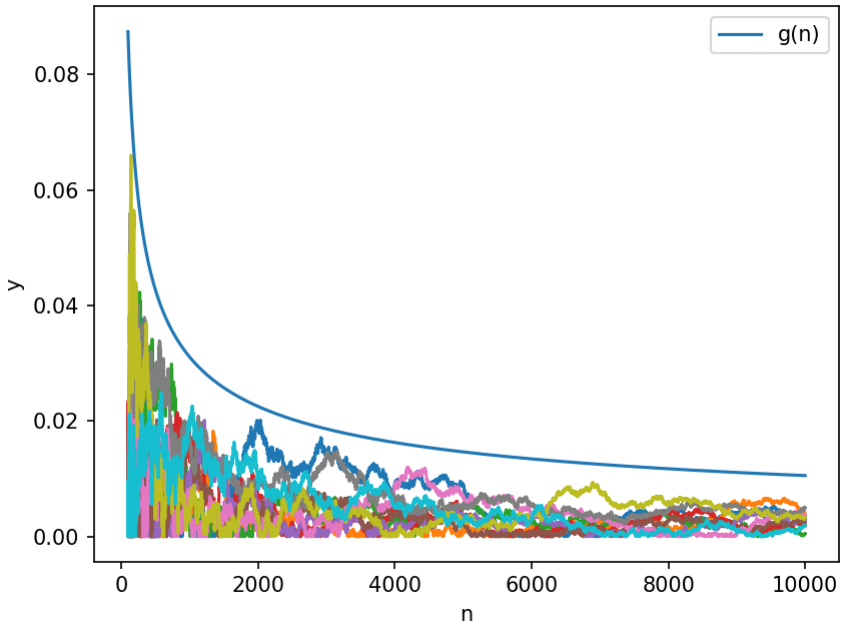
- (Mathematical/rigorous version, see Theorem 22.11 of Klenke (2020); not required) Suppose $X_1, \ldots, X_n, \ldots$ are iid random variables defined on the probability space $(\Omega, \mathbb{P})$. The mean is $\mu = \mathbb{E}X_1$ and the standard variance is $\sigma = \sqrt{Var(X_1)}$. Then, we have

$$\mathbb{P}\left\{ \omega \in \Omega \,\middle|\, \limsup_{n \to \infty} \left[\frac{\left| \frac{1}{n} \sum_{i=1}^{n} X_i(\omega) - \mu \right|}{\sigma \cdot \sqrt{\frac{2 \cdot \log(\log n)}{n}}} \right] = 1 \right\} = 1.$$

2.2 Question 2

Use any programming language of your choice to execute the following steps:
Step 1: Generate 10,000 random numbers $x_1, x_2, \ldots, x_{10000}$ from the distribution Bernoulli(0.5).
Step 2: For each n in the range 100 to 10,000, compute y_n using the formula $y_n = \left| \frac{1}{n} \sum_{i=1}^{n} x_i - 0.5 \right|$.
Step 3: Plot the graph of "y_n versus n" (with n on the horizontal axis).
Step 4: Repeat Steps 1-3 nine more times, and overlay all ten graphs on the same plot.
Step 5: Define a function $g(n) = 0.5 \cdot \sqrt{\frac{2 \cdot \log(\log n)}{n}}$. Then, add the graph of "$g(n)$ versus n" to the plot from Step 4.

Display the picture obtained in Step 5 and provide the code used to generate it (2 points). Explain the picture from Step 5 (1 point).




```
import matplotlib.pyplot as plt
from scipy.stats import bernoulli
from math import sqrt, log

def avg(lst):
    return sum(lst)/len(lst)

len_rands = 10000
x = range(100, len_rands)

for plot in range(10):
    rands = bernoulli.rvs(0.5, size=len_rands)
    running_avg = 0
    y = []
    for n in x:
        running_avg = ((running_avg * (n-1)) + (rands[n] - 0.5))/n
        y.append(abs(running_avg))
    plt.plot(x, y)

g = list(map(lambda n: 0.5 * sqrt((2 * log(log(n)))/n), x))
plt.plot(x, g, label="g(n)")

plt.xlabel("n")
plt.ylabel("y")
plt.legend()
plt.show()
```

The plot shown above illustrates the law of the iterated logarithm. First, I generated a series of iid random variables $\{X_i\}_{i=1}^{10000} \sim \text{Bernoulli}\left(\frac{1}{2}\right)$ with $\mu = \mathbb{E}X_1 = \frac{1}{2}$. Then for larger and larger n, I plotted the mean absolute error

$$e_n = |\overline{X} - \mu|$$

for ten different series of random variables to illustrate that they all share a similar shape but are random. Finally, the function $g(n)$ is the quantity $\sqrt{\text{Var}X_1} \cdot \sqrt{\frac{2 \log \log n}{n}}$. The final graph confirms the heuristic that the running error is less than the quantity $g(n)$.

3 Question 3

3.1 Review before Question 3

- (Empirical CDF.) Let $X_1, X_2, \ldots, X_n$ be random variables defined on the probability space $(\Omega, \mathbb{P})$. For each fixed ω, consider the indicator function $\mathbb{1}\{X_i(\omega) \le t\}$, which is a function of t. The function $F_n(\omega, t)$, defined as follows, is called the empirical CDF based on the random variables $X_1, X_2, \ldots, X_n$

$$F_n(\omega, t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i(\omega) \le t\} = \frac{\text{number of } \{X_i(\omega)\}_{i=1}^n \text{ such that } X_i(\omega) \le t}{n}, \quad \text{for all } t \in \mathbb{R}.$$

- The following is the Glivenko-Cantelli theorem.

  Theorem 3.1 (Glivenko-Cantelli, 1933; see Theorem 5.23 of Klenke (2020))**.** *Let $F(t)$ be the CDF shared by iid random variables $X_1, \ldots, X_n, \ldots$, and $F_n(\omega, t)$ is the empirical CDF based on $\{X_i(\omega)\}_{i=1}^n$. Then, the empirical CDF converges to $F(t)$ uniformly with probability one, i.e.,*[1]

$$\mathbb{P}\left\{\omega \in \Omega : \lim_{n\to\infty}\left(\sup_{t\in\mathbb{R}} |F_n(\omega, t) - F(t)|\right) = 0\right\} = 1. \tag{3.1}$$

3.2 You may directly apply the following results to solve Question 3

Please feel free to utilize the following two results without proving them.

(i) (Dvoretzky-Kiefer-Wolfowitz inequality, 1956, see Dvoretzky et al. (1956))
Let $F(t)$ be the CDF shared by iid random variables $X_1, \ldots, X_n, \ldots$. The function $F_n(\omega, t)$ represents the empirical CDF based on $\{X_i(\omega)\}_{i=1}^n$. Then, there exists a positive constant C (not depending on F) such that

$$\mathbb{P}\left\{\omega \in \Omega \,\middle|\, \sup_{t\in\mathbb{R}} |F_n(\omega, t) - F(t)| > \epsilon\right\} \le Ce^{-2n\epsilon^2},$$

for any $\epsilon > 0$ and positive integer n.

(ii) (See Theorem 1.8 of Shao (2003)) Let $Z(\omega), Z_1(\omega), \ldots, Z_n(\omega), \cdots$ be random variables. If, for every $\epsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}\left\{\omega \in \Omega \,\middle|\, Z_n(\omega) \ge \epsilon\right\} < \infty,$$

then Z_n converges to 0 with probability one, i.e.,

$$\mathbb{P}\left\{\omega \in \Omega \,\middle|\, \lim_{n\to\infty} Z_n(\omega) = 0\right\} = 1.$$

[1] The "sup" in Eq. (3.1) denotes the "supremum." If you are not familiar with the "sup," please feel free to just view it as maximum, i.e., "max."

3.3 Question 3

(2 points) Prove the Glivenko-Cantelli theorem (i.e., Theorem 3.1) using the two results presented above.

We let $X_1, \ldots, X_n$ be iid random variables sharing the CDF $F(t)$ which form the basis for the empirical CDF $F_n(\omega, t)$.

We note that the sup function fixes a value of t so the quantity $\sup_{t\in\mathbb{R}} \mid F_n(\omega,t) - F(t) \mid$ is a random variable. We call it $Z_n(\omega)$.

By the Dvoretzky-Kiefer-Wolfowitz inequality,

$$\mathbb{P}(Z_n > \varepsilon) \leq Ce^{-2n\varepsilon^2}$$

Thus, we calculate

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}(Z_n \geq \varepsilon) &= \sum_{n=1}^{\infty} \mathbb{P}(Z_n = \varepsilon) + \sum_{n=1}^{\infty} \mathbb{P}(Z_n > \varepsilon) \\ &\leq \sum_{n=1}^{\infty} \mathbb{P}(Z_n = \varepsilon) + \sum_{n=1}^{\infty} Ce^{-2n\varepsilon^2}\end{aligned}$$

Because ε can be any positive real number, Z_n must be continuous or it can not hit every value of ε. Either way, $\mathbb{P}(Z_n = \varepsilon) = 0$ for any particular ε. Thus,

$$\sum_{n=1}^{\infty} \mathbb{P}(Z_n \geq \varepsilon) \leq \sum_{n=1}^{\infty} Ce^{-2n\varepsilon^2}$$

Looking at the RHS, we observe that the series $a_n = Ce^{-2\varepsilon^2}$ is positive (because C is positive) and monotonically decreasing so we can apply the integral convergence test (using the properties of the Gaussian Integral):

$$\int_1^{\infty} Ce^{-2x\varepsilon^2}\, dx \overset{a=2\varepsilon^2}{=} \int_1^{\infty} Ce^{-ax^2}\, dx < \int_0^{\infty} Ce^{-ax^2} = \frac{1}{2}\sqrt{\frac{\pi}{a}} < \infty$$

Then by Theorem 1.8 of Shao 2003,

$$\mathbb{P}\left(\left\{\omega \in \Omega \middle| \lim_{n\to\infty} Z_n(\omega) = 0\right\}\right) = \mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} \left(\sup_{t\in\mathbb{R}} \left|F_n(\omega,t) - F(t)\right|\right) = 0\right\}\right) = 1$$

Which is precisely the Glivenko-Cantelli theorem. ■

4 Question 4

4.1 Review before Question 4

Definition 4.1. *Suppose g is a real-valued function defined on $\mathbb{R}$.*

1. *Let X be a discrete random variable whose CDF is $\sum_{k=0}^{K} p_k \cdot \mathbb{1}_{[x_k,\infty)}(x)$. If $\sum_{k=0}^{K} |g(x_k)| \cdot p_k < \infty$, then the following sum is called the mean/expected value of $g(X)$ and denoted as $\mathbb{E}[g(X)]$*

$$\boxed{\mathbb{E}\left[g(X)\right] \stackrel{\text{def}}{=} \sum_{k=0}^{K} g(x_k) \cdot p_k.}$$

   If $\sum_{k=0}^{K} |g(x_k)| \cdot p_k = \infty$, we say the expected value of $g(X)$ does not exist.

2. *Let X be a continuous random variable with PDF $p_X(x)$. If $\int_{-\infty}^{+\infty} |g(x)| \cdot p_X(x)dx < \infty$, we call the following integral as the mean/expected value of $g(X)$ and denote it as $\mathbb{E}[g(X)]$*

$$\boxed{\mathbb{E}[g(X)] \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} g(x) \cdot p_X(x)dx.}$$

   If $\int_{-\infty}^{+\infty} |g(x)| \cdot p_X(x)dx = \infty$, we say the expected value of X does not exist.

4.2 Question 4

Let X be a continuous random variable whose PDF is the following[2]

$$p_X(x) = \frac{1}{\pi\left(1+x^2\right)}, \quad \text{for all } x \in \mathbb{R}. \tag{4.1}$$

a (1 point) Question: Does the expected value $\mathbb{E}X$ of X exist? Please prove your answer.

$$\int_{-\infty}^{\infty} \frac{|x|}{\pi(1+x^2)}\,dx = \int_{-\infty}^{0} \frac{-x}{\pi(1+x^2)} + \int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\,dx = 2\int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\,dx$$

But

$$\frac{1}{\pi}\int_{0}^{\infty} \frac{2x}{1+x^2}\,dx = \frac{1}{\pi}\int_{0}^{\infty} \frac{1}{u}\,du =$$
$$frac1\pi \left[\ln\ |1+x^2|\ \right]_0^{\infty} = \infty - 0 = \infty$$

So the integral does not converge.

Thus, the expected value

$$\mathbb{E}X = \int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)}\,dx$$

does not exist. ■

[2]The PDF corresponds to a Cauchy distribution.

b Suppose we have done the following

- For each positive integer n, we generated random numbers $\xi_1, \xi_2, \ldots, \xi_n$ from the distribution whose PDF is the one presented in Eq. (4.1) (i.e., the Cauchy distribution with parameters "$(x_0 = 0, \gamma = 1)$"; see the Wikipedia page about Cauchy distribution).
- We computed the sample average $\overline{\xi}_n = \frac{1}{n}\sum_{i=1}^{n} \xi_i$.
- We plotted the "sample average $\overline{\xi}_n = \frac{1}{n}\sum_{i=1}^{n} \xi_i$ vs. sample size n" curves.
- We repeated this procedure 1000 times and got 1000 such curves. The 1000 curves are presented in Figure 1.

The R code for generating the figure is provided as follows.

```
        m=1000
        X=rcauchy(m)
        X_bar=cumsum(X)/(1:m)
        plot(1:m, X_bar, type = "l", xlab = "Sample size n", ylab = "Sample
    average",
             ylim = c(-50, 50), lwd=0.2,
             main = "Sample average vs. sample size")
        for (i in 1:999) {
        X=rcauchy(m)
        X_bar=cumsum(X)/(1:m)
        lines(1:m, X_bar,
                lwd=0.2)
        }
        abline(h=0, lty=2, col="red", lwd=2)

```

(1 point) Question: When $n \to \infty$, does $\overline{\xi}_n = \frac{1}{n}\sum_{i=1}^{n} \xi_i$ converge to anything with probability one? Please heuristically (rather than mathematically/rigorously) explain your answer using both Figure 1 and the Law of Large Numbers.

The law of large numbers says that the sample average of a sequence of n independently and identically distributed random variables converges to the expected value of the corresponding CDF with probability 1 if the expected value exists. In the case of the Cauchy distribution, however, the expected value does not exist so the LLN does not apply. Experimental results (as in Fig. 1) suggest that there is no uniformity whatsoever in the convergence pattern of $\overline{\xi}_n$.

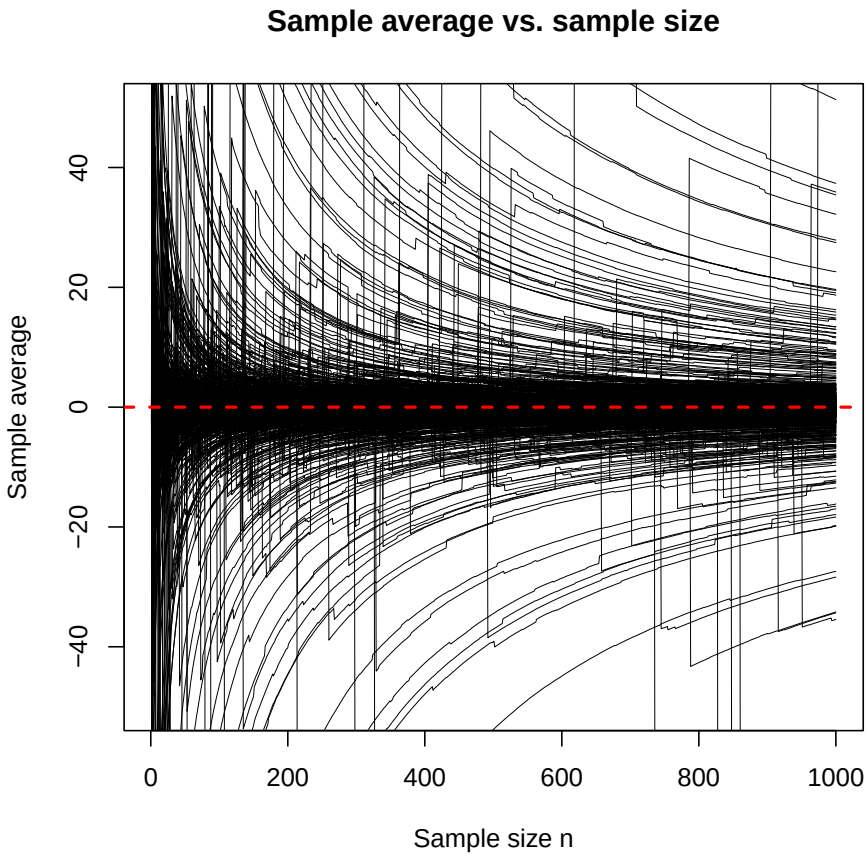


Figure 1: The 1000 "sample average $\overline{\xi}_n = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ vs. sample size n" curves.

References

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

N. Etemadi. An elementary proof of the strong law of large numbers. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 55(1):119–122, 1981.

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

J. Shao. *Mathematical statistics*. Springer Science & Business Media, 2003.

APMA1690: Homework # 3 (Due by 11pm Oct 5)

1 Review

- (**Multiplicative Congruential Generator**)
  For "properly chosen" positive integers m and a (e.g., $m = 2^{31} - 1$ and $a = 7^5$), we have the multiplicative congruential generator (MCG) presented by the following algorithm

---

Algorithm 1 : Multiplicative Congruential Generator

Input: (i) positive integers m and a; (ii) a seed $s \in \{1, 2, \ldots, m-1\}$; (iii) sample size n.

Output: pseudo-random numbers $g(1), g(2), \ldots, g(n)$. (These numbers look like iid from $\text{Unif}(\{1, 2, \ldots, m-1\})$.)

1: Initialization: $g(1) \leftarrow s$.
2: **for all** $i = 1, 2, \ldots, n-1$, **do**
3: $g(i+1) \leftarrow (a \cdot g(i) \bmod m)$ = the remainder of $\frac{a \cdot g(i)}{m}$.
4: **end for**

The "remainder" referred to in the algorithm above is the remainder in integer division.

- (**Fundamental theorem for the "inverse CDF method"**)
  Let $F(t)$ be a CDF of interest. We define the function $G(u)$ on the open interval $(0, 1)$ by

$$G(u) = \inf\{t \in \mathbb{R} : F(t) \geq u\}, \quad \text{for all } u \in (0, 1), \tag{1}$$

  where "inf" denotes the infimum operation (you may view it as "min" for simplicity). The function G in Eq. (1) is the "generalized inversion" of F. Let U be a random variable defined on the probability space $(\Omega, \mathbb{P})$ and following the continuous uniform distribution on $(0, 1)$, i.e., $U \sim \text{Unif}(0, 1)$, and we define a new random variable X by

$$X(\omega) = G(U(\omega)), \quad \text{for all } \omega \in \Omega.$$

  Then, the CDF of X is the CDF $F(t)$ of interest.
  Remarks: (i) For any given real number $0 < u < 1$, the notation "$\{t \in \mathbb{R} : F(t) \geq u\}$" denotes the collection of the real numbers t such that $F(t) \geq u$.
  (ii) "$\inf\{t \in \mathbb{R} : F(t) \geq u\}$" denotes the smallest number in the collection $\{t \in \mathbb{R} : F(t) \geq u\}$.
  (iii) If the inverse F^{-1} of F exists, then $G = F^{-1}$.

- Algorithm 2 algorithm provides the procedures for implementing the inverse CDF method.

Algorithm 2 : Inverse CDF Method

Input: (i) The CDF F of interest; (ii) sample size n.

Output: A sequence of iid (pseudo) random numbers $x_1, \ldots, x_n$ following the distribution F.

1: Generate (pseudo) iid random variables $u_1, u_2, \ldots, u_n$ from Unif$(0, 1)$.
2: **for all** $i = 1, \ldots, n,$ **do**
3: Compute $x_i \leftarrow G(u_i) = \inf\{t \in \mathbb{R} : F(t) \geq u_i\}$.
4: **end for**

2 Problem Set

1. This question helps you better understand the MCG. Let g be the MCG in Algorithm 1.

   (a) (1 point) Write the computer code for the MCG as described in Algorithm 1 using your preferred programming language. Please include your code in your submission.

```python
def MCG(s, n):
    m = 2**(31) - 1
    a = 7**5
    g = [s]

    for i in range(1, n):
        g.append((a * g[i - 1]) % m)

    return g
```

   (b) (1 point) Let $m = 2^{31} - 1$ and $a = 7^5$. Initialize the seed by $g(1) \leftarrow 1690$ and generate $g(1), g(2), \ldots, g(10)$ using your code. Show all the ten numbers $g(1), g(2), \ldots, g(10)$ and the code for generating these numbers.

n	$g(n)$
1	1690
2	28403830
3	641801176
4	2089489798
5	254955595
6	808809400
7	88100290
8	1085341247
9	604240711
10	23463114

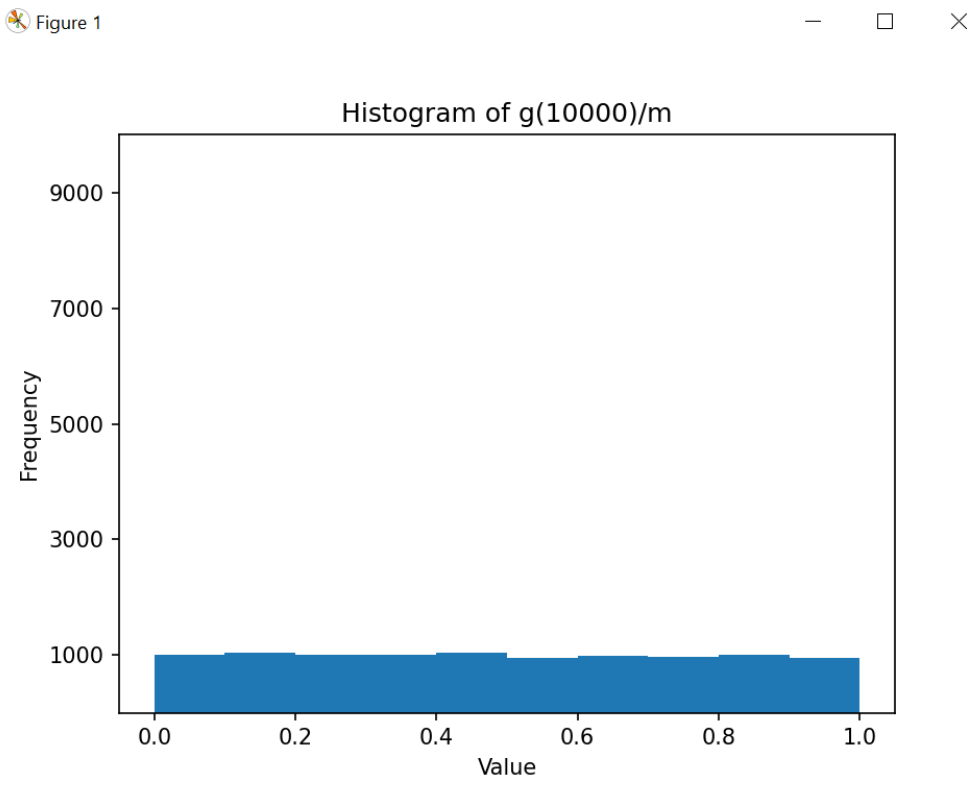
```python
def Q1B():
    return MCG(s=1690, n=10)
```

   (c) (1 point) Let $m = 2^{31} - 1$ and $a = 7^5$. Initialize the seed by $g(1) \leftarrow 1690$ and generate $g(1), g(2), \ldots, g(10000)$ using your code. Plot and show the histogram of the following

10000 values

$$\left\{\frac{g(1)}{m}, \frac{g(2)}{m}, \ldots, \frac{g(10000)}{m}\right\}.$$

Show the code for generating this histogram.




```python
def Q1F(plot=True):
    n = 10000
    m = 2**(31) - 1
    s = 1690

    y = list(map(lambda x: math.log(1/(1 - (x/m))), MCG(s=s, n=n)))

    if plot:
        plt.title("Random from Exp(1)")
        plt.ylabel("Frequency")
        plt.xlabel("Value")
        plt.ylim(0, 10000)
        plt.yticks(range(1000, 10000, 2000), range(1000, 10000, 2000))
        plt.hist(y)
        plt.show()

    return y
```

(d) (1 point) Please heuristically (rather than mathematically/rigorously) explain the relationship between the histogram you obtained in the preceding question and Unif$(0,1)$.

The sequence of random numbers $\left\{\frac{g(1)}{m}, \ldots, \frac{g(n)}{m}\right\}$ "look like" random numbers taken iid from Unif$(0,1)$.

(e) (0.5 points) Let $F(t) = (1 - e^{-t}) \cdot \mathbb{1}(t > 0)$, which is the CDF of the exponential distribution Exp(1). Compute an explicit express of the $G(u)$ defined in Eq. (1) for all $0 < u < 1$.

By the Fundamental Theorem of the Inverse CDF, we are looking for an explicit expression for

$$G(u) = \inf\{t \in \mathbb{R} : F(t) \geq u\} \quad \forall u \in (0,1)$$

Heuristically, we ask "what is the smallest t for which $F(t)$ is greater than or equal to u, for all $u \in (0,1)$?" First, we observe

$$F(t) = (1 - e^{-t}) \cdot \mathbb{1}(t > 0) = \begin{cases} 1 - e^{-t} & t > 0 \\ 0 & t \leq 0 \end{cases}$$

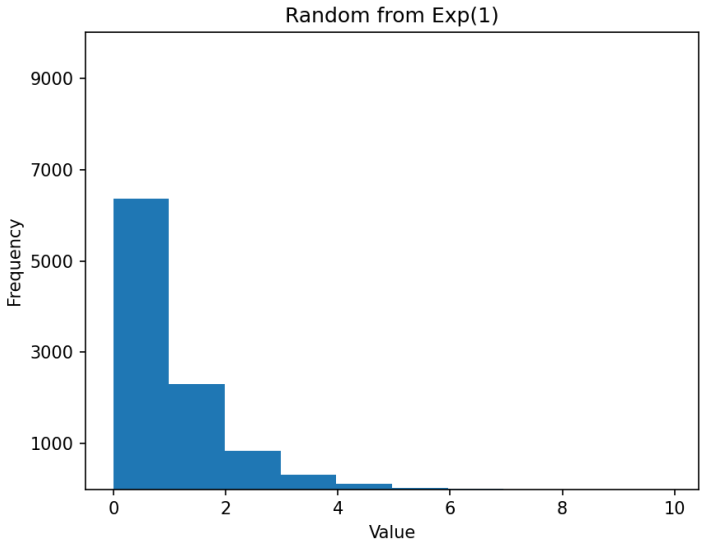
Immediately, we see that $G(u)$ is not defined for $t \leq 0$ and for $t > 0$:

$$\begin{aligned} G(u) &= \inf\{t \in \mathbb{R} : 1 - e^{-t} \geq u\} \\ &= \inf\{t \in \mathbb{R} : 1 - u \geq e^{-t}\} \\ &= \inf\{t \in \mathbb{R} : \log(1-u) \geq -t\} \\ &= \inf\{t \in \mathbb{R} : -\log(1-u) \leq t\} \\ &= \inf\{t \in \mathbb{R} : \log(\frac{1}{1-u}) \leq t\} \\ &= \boxed{\log(\frac{1}{1-u})} \end{aligned}$$

(f) (0.5 points) Using the values $g(1), g(2), \ldots, g(10000)$ generated in the preceding question, plot and show the histogram of the following 10000 values

$$\left\{\log\left(\frac{1}{1 - \frac{g(1)}{m}}\right), \log\left(\frac{1}{1 - \frac{g(2)}{m}}\right), \ldots, \log\left(\frac{1}{1 - \frac{g(10000)}{m}}\right)\right\},$$

where "log" denotes the natural logarithm and $m = 2^{31} - 1$. Please heuristically (rather than mathematically/rigorously) explain the relationship between the histogram you obtained here and the probability density function of the exponential distribution Exp(1).




```python
def Q1F(plot=True):
    n = 10000
    m = 2**(31) - 1
    s = 1690

    y = list(map(lambda x: math.log(1/(1 - (x/m))), MCG(s=s, n=n)))

    if plot:
        plt.title("Random from Exp(1)")
        plt.ylabel("Frequency")
        plt.xlabel("Value")
        plt.ylim(0, 10000)
        plt.yticks(range(1000, 10000, 2000), range(1000, 10000, 2000))
        plt.hist(y)
        plt.show()

    return y
```

By the inverse CDF method, this sequence of values are random numbers which "look like" random numbers generated iid from the distribution Exp(1).

2. Let X be a random variable defined on the probability space $(\Omega, \mathbb{P})$, and X satisfies

$$\mathbb{P}\{\omega \in \Omega : X(\omega) = i\} = \frac{1}{n}, \quad \text{for all } i = 1, 2, \ldots, n.$$

where n is a given positive integer. Define a new random variable Y by

$$Y(\omega) = \frac{X(\omega)}{n}, \quad \text{for all } \omega \in \Omega.$$

(a) (0.5 point) Show that the CDF of Y is the following

$$\begin{aligned} F_n(t) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\frac{i}{n} \le t\}} \\ &= \frac{\text{the number of integers } i \text{ such that } \frac{i}{n} \le t}{n}, \end{aligned} \tag{2}$$

for all $t \in \mathbb{R}$.

We first note that from the definition of a discrete CDF and $\mathbb{P}(X = i) = \frac{1}{n}$, the CDG od X is

$$F_X(t) = \sum_{i=1}^{n} \frac{1}{n} \cdot \mathbb{1}_{(i \le t)}$$

Then, because $Y(\omega) = \frac{X(\omega)}{n}$,

$$F_Y(t) = \mathbb{P}(Y \le t) = \mathbb{P}(\frac{X}{n} \le t) = \sum_{i=1}^{n} p_i \cdot \mathbb{1}_{(y_i \le t)} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(\frac{i}{n} \le t)} \quad \blacksquare$$

(b) (1 point) Let $F_n(t)$ be the CDF defined in Eq (2). Use the definition of Riemann integrals to prove the following

$$\begin{aligned} \lim_{n\to\infty} F_n(t) = &\text{ the CDF of } \text{Unif}(0,1) \\ &= \begin{cases} 0 & \text{if } t < 0 \\ t & \text{if } 0 \le t < 1 \\ 1 & \text{if } 1 \le t. \end{cases} \end{aligned}$$

From Eq 2,

$$\lim_{n\to\infty} F_n(t) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\frac{i}{n} \le t\}}$$

Since $1 \le i \le n$, $\frac{1}{n} \le \frac{i}{n} \le 1$ so we know that $\mathbb{1}_{(x \le t)} = 0$ for $t < 0$. Similarly, for $t \ge 1$, $\mathbb{1}_{(x \le 1)} = 1$ for all x.

But we notice that the definition of Riemann Integrals,

$$\int_0^1 H(x)\, dx = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} H(\frac{i}{n})$$

is quite similar to $\lim_{n\to\infty} F_n(t)$, especially since we only need to check $0 \le t < 1$. Letting $H(\frac{i}{n}) = \mathbb{1}_{\frac{i}{n} \le t}$, we have

$$\begin{aligned}\lim_{n\to\infty} F_n(t) &= \int_0^1 \mathbb{1}_{(x\le t)}\, dx \\ &= \int_0^t 1\, dx + \int_t^1 0\, dx \\ &= t\end{aligned}$$

All together,

$$\lim_{n\to\infty} F_n(t) = \begin{cases} 0 & t < 0 \\ t & 0 \le t < 1 \\ 1 & \le t \end{cases} \qquad \blacksquare$$

3. Let $F(t)$ denote the CDF of the Bernoulli($\frac{1}{3}$) distribution.

   (a) (1 point) Compute an explicit express of the $G(u)$ defined in Eq. (1) for all $0 < u < 1$.

$$F(t) = \frac{2}{3} \cdot \mathbb{1}_{(t \geq 0)} + \frac{1}{3} \cdot \mathbb{1}_{(t \geq 1)}$$

   So

$$\begin{aligned} G(u) &= \inf\{t \in \mathbb{R} : F(t) \geq u\} \\ &= \inf\{t \in \mathbb{R} : \frac{2}{3} \cdot \mathbb{1}_{(t \geq 0)} + \frac{1}{3} \cdot \mathbb{1}_{(t \geq 1)} \geq u\} \\ &= \boxed{\begin{cases} 0 & \text{if } 0 < u \leq \frac{2}{3} \\ 1 & \text{if } \frac{2}{3} < u < 1 \end{cases}} \end{aligned}$$

   (b) (1 point) Explain why $G(0) = \inf\{t \in \mathbb{R} : F(t) \geq 0\}$ is ill-defined.
   In Eq 2, u is defined on the *open* interval $(0, 1)$ so $u = 0$ is not in the domain of the inverse.

   (c) (0.5 point) Let $g(1), g(2), \ldots, g(10)$ be the values you generated in Question 1 (b). Compute and show the following ten values

$$\left\{ G\left(\frac{g(1)}{m}\right), G\left(\frac{g(2)}{m}\right), \ldots, G\left(\frac{g(10)}{m}\right) \right\},$$

   where $m = 2^{31} - 1$ and G is the function in Question 3 (a).

$$\{g(i)\}_{i=1}^{10} = \{0, 0, 0, 1, 0, 0, 0, 0, 0, 0\}$$

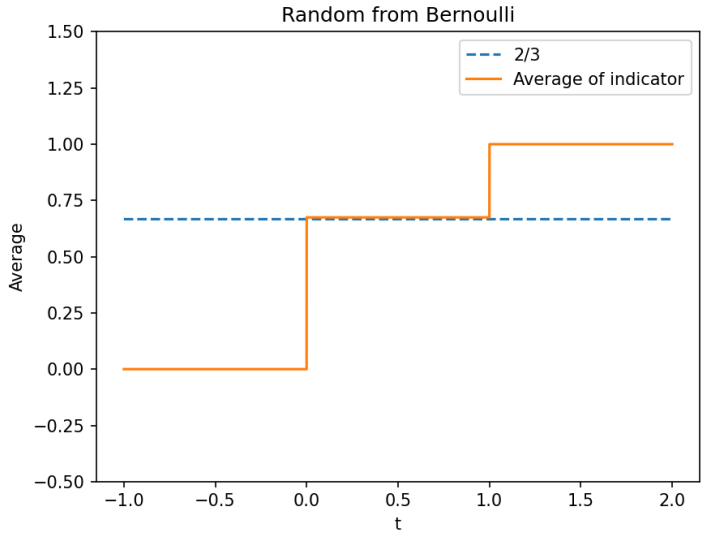
```
def Q3C():
    m = 2**(31) - 1
    g = Q1B()

    def G(n):
        if n < (2 / 3):
            return 0
        else:
            return 1

    return list(map(lambda x: G(x/m), g))
```

   (d) (1 point) Let $g(1), g(2), \ldots, g(10000)$ be the values you generated in Question 1 (c). Denote

$$x_i = G\left(\frac{g(i)}{m}\right), \quad \text{for all } i = 1, 2, \ldots, 10000,$$

where $m = 2^{31} - 1$. Plot and show the graph of the following function of t

$$\frac{1}{10000} \sum_{i=1}^{10000} \mathbf{1}\{x_i \leq t\}.$$




```python
def Q3D(plot=True):
    m = 2**(31) - 1
    g = Q1C(False)

    def G(n):
        if n < (2 / 3):
            return 0
        else:
            return 1

    def ind_avg(x):
        y = []
        total = 0
        n = range(len(x))

        for index in n:
            total += x[index]
            y.append(total)
        y = list(map(lambda x: x/len(n), y))

        return y

    t = np.linspace(-1, 2, 10000)
    y = ind_avg(list(map(lambda x: G(x), g)))

    if plot:
        plt.title("Random from Bernoulli")
        plt.ylabel("Average")
        plt.xlabel("t")
        plt.ylim(0, 0.5)
        plt.plot(t, np.linspace(1/3, 1/3, 10000), linestyle='dashed', label='1/3')
        plt.plot(t, y, label='Average of indicator')

        plt.legend()
        plt.show()

    return y
```

APMA1690: Homework # 4 (Due by 11pm Oct 19)

1 Review

Please read the review section before delving into the problem set.

1.1 Random Variables

Let $\boldsymbol{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(d)})$ be a $\mathbb{R}^d$-valued random variable defined on the probability space $(\Omega, \mathbb{P})$, i.e.,

$$\begin{aligned} \boldsymbol{X}: \ & \Omega \to \mathbb{R}^d, \\ & \omega \mapsto \boldsymbol{X}(\omega) = \Big(X^{(1)}(\omega), X^{(2)}(\omega), \ldots, X^{(d)}(\omega)\Big), \end{aligned}$$

where each $X^{(i)}$ is a $\mathbb{R}^1$-valued random variable. When $d > 1$, $\boldsymbol{X}$ is also referred to as a "random vector."

Suppose H is a d-variable function which takes values in $\mathbb{R}$, i.e.,

$$\begin{aligned} H: \ & \mathbb{R}^d \to \mathbb{R}, \\ & \boldsymbol{x} = (x_1, x_2, \ldots, x_d) \mapsto H(\boldsymbol{x}) = H(x_1, x_2, \ldots, x_d). \end{aligned}$$

Then, we have the $\mathbb{R}^1$-valued random variable $H(\boldsymbol{X})$ defined as follows

$$\begin{aligned} H(\boldsymbol{X}): \ & \Omega \to \mathbb{R}^1, \\ & \omega \mapsto H\left(\boldsymbol{X}(\omega)\right) = H\Big(X^{(1)}(\omega), X^{(2)}(\omega), \ldots, X^{(d)}(\omega)\Big). \end{aligned}$$

1.2 Setup

Suppose our goal is to compute the following multiple integral[1]

$$v = \int H(\boldsymbol{x})\, d\boldsymbol{x} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} H(x_1, x_2, \ldots, x_d)\, dx_1 dx_2 \cdots dx_d,$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ and $d\boldsymbol{x} = dx_1 dx_2 \cdots dx_d$. This integral can be represented as follows

$$\boxed{v = \int H(\boldsymbol{x})\, d\boldsymbol{x} = \int \frac{H(\boldsymbol{x})}{f(\boldsymbol{x})} \cdot f(\boldsymbol{x})\, d\boldsymbol{x} = \mathbb{E}\left[\frac{H(\boldsymbol{X})}{f(\boldsymbol{X})}\right],}$$

where $f(\boldsymbol{x})$ is a d-dimensional PDF, and the random vector $\boldsymbol{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(d)}) \sim f(\boldsymbol{x})$. Additionally, the PDF $f(\boldsymbol{x})$ satisfies the following conditions

[1] We assume that all means and variances utilized herein do exist.

1. $\{\boldsymbol{x} \in \mathbb{R}^d \,|\, H(\boldsymbol{x}) \neq 0\} \subset \{\boldsymbol{x} \in \mathbb{R}^d \,|\, f(\boldsymbol{x}) \neq 0\}$;
2. We know how to generate random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} f(\boldsymbol{x})$;
3. $f(\boldsymbol{x})$ is similar to the "optimal" PDF $\frac{1}{\int H(\boldsymbol{x}')d\boldsymbol{x}'} \cdot H(\boldsymbol{x})$. This similarity makes $\text{Var}\left(\frac{H(\boldsymbol{X}_1)}{f(\boldsymbol{X}_1)}\right)$ small. (Since the integral $v = \int H(\boldsymbol{x}')d\boldsymbol{x}'$ is unavailable at this point, the optimal PDF is not achievable.)

1.3 Importance Sampling

We generate random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \overset{iid}{\sim} f(\boldsymbol{x})$ and compute the following estimator of v

$$\boxed{\widehat{v}_n = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{H(\boldsymbol{X}_i)}{f(\boldsymbol{X}_i)}\right].} \tag{1.1}$$

Then, we have

1. Law of large numbers $\Longrightarrow$ $\widehat{v}_n \approx v$ when the sample size n is sufficiently large;
2. Law of the iterated logarithm $\Longrightarrow$ $|\widehat{v}_n - v| \leq \sqrt{\text{Var}\left(\frac{H(\boldsymbol{X}_1)}{f(\boldsymbol{X}_1)}\right)} \cdot \sqrt{\frac{2\log(\log n)}{n}}$, where "$\leq$" holds in an approximate way.

A good reference for importance sampling is Chapter 7 of Wang (2012).

1.4 Markov Chains

Roughly speaking, a Markov chain is a sequence of random variables $\{X_n\}_{n=0}^{\infty}$ satisfying the Markov property.[2] In APMA 1690, we assume that all random variables take values in a generic countable set $\mathcal{X}$, i.e., $\mathcal{X}$ can be expressed as $\mathcal{X} = \{\xi_0, \xi_1, \ldots, \xi_n, \ldots\}$. The countability assumption of $\mathcal{X}$ heavily simplifies the theory of Markov chains. The following is the definition[3] of Markov chains.

Definition 1.1. • *A sequence of random variables $\{X_n\}_{n=0}^{\infty}$ taking values in $\mathcal{X}$ is called a **Markov chain** if this sequence satisfies*

$$\boxed{\mathbb{P}(X_{n+1} = y \,|\, X_n = x, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \mathbb{P}(X_{n+1} = y \,|\, X_n = x)} \tag{1.2}$$

*for all $n = 0, 1, \ldots$, all $y \in \mathcal{X}$, and all the $x_0, x_1, \ldots, x_{n-1}, x \in \mathcal{X}$ such that $\mathbb{P}(X_n = x, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) > 0$. The property in Eq. (1.2) is called the **Markov property**.*

[2] For Markov chains in this course, we always let the index n go from 0 instead of 1. It is just a convention.

[3] The definition herein works only for the scenario where a Markov chain takes values in a countable space $\mathcal{X} = \{\xi_0, \xi_1, \ldots, \xi_n, \ldots\}$. It is one of the reasons that we assume $\mathcal{X}$ is countable. For the general definition of Markov chains and the relevant details, see Definition 17.1 and Remark 17.2 of Klenke (2020). Since the materials of general Markov chains involve too much real analysis knowledge (see APMA 2110), we skip the general Markov chains in this course.

- *Furthermore, if there exists a bivariate function* $p : \mathcal{X} \times \mathcal{X} \to [0, 1]$ *such that*

$$\boxed{p(x, y) = \mathbb{P}(X_{n+1} = y \,|\, X_n = x) \quad \textit{(this function does not depend on } n\textit{)},}$$

  for all $n = 0, 1, \ldots$, *then* $\{X_n\}_{n=0}^{\infty}$ *is called a* ***homogeneous*** *Markov chain, and the function* $p(\cdot, \cdot)$ *is called the* ***transition probability*** *of this Markov chain.*

- *In literature,* $\mathcal{X}$ *is usually referred to as the* ***state space*** *of the Markov chain* $\{X_n\}_{n=0}^{\infty}$*; each element in* $\mathcal{X}$ *is referred to as a* ***state****.*

Throughout this course, all the Markov chains will be homogeneous. Hence, **we will omit the word "homogeneous" hereafter**. In Eq. (1.2), if we call X_n as "the present," X_{n+1} as "the future," and $X_{n-1}, \ldots, X_0$ as "the past," the Markov property means, "given the present, the future does not depend on the past" — the condition $X_{n-1} = x_{n-1}, \ldots, X_0 = x_0$ in Eq. (1.2) does not play any role!

The value of $p(x, y)$ is the probability of "transiting from x to y," so it is called a transition probability. For each fixed $x \in \mathcal{X}$, the univariate function $p(x, \cdot) : y \mapsto p(x, y)$ is a PMF.

2 Problem Set

1. The unit ball in d-dimensional space is defined as follows

$$\mathbf{B}^d = \left\{(x_1, x_2, \ldots, x_d) \in \mathbb{R}^d \,\middle|\, x_1^2 + x_2^2 + \cdots + x_d^2 < 1\right\}.$$

The boundary of this unit ball is the following $(d-1)$-dimensional sphere

$$\mathbb{S}^{d-1} = \left\{(x_1, x_2, \ldots, x_d) \in \mathbb{R}^d \,\middle|\, x_1^2 + x_2^2 + \cdots + x_d^2 = 1\right\}.$$

(1 point) Please explain (not rigorously prove) the claim, "***when the dimension*** d ***is large, most of the volume of the unit ball*** $\mathbf{B}^d$ ***is concentrated near the sphere*** $\mathbb{S}^{d-1}$***.***"

As d gets large, each individual x_i^2 must get smaller for their sum to be less than 1. This means, however, that for large d, the volume added by each successive x_i gets smaller and smaller ($x_{10000} \approx x_{10001}$) as the sum tends to 1.

Heuristically, one can imagine the 3-sphere which has volume $\frac{4}{3}\pi r^3$. Small increases in radius lead to very large increases in volume because of the cubic factor; elements towards the edge of the boundary contribute more to the total volume.

2. Suppose we are interested in the following multiple integral

$$v_d = \int H(\boldsymbol{x}) d\,\boldsymbol{x}, \quad \text{where } H(\boldsymbol{x}) = \mathbb{1}_{\{x_1^2+x_2^2+\cdots+x_d^2<1\}},$$

and we want to estimate the v_d using the importance sampling approach. To do so, we need to choose a d-variable PDF which is simiar to $\frac{1}{v_d} \cdot \mathbb{1}_{\{x_1^2+x_2^2+\cdots+x_d^2<1\}}$.

We restrict our attention to the following collection of multivariable normal PDFs indexed by $\sigma > 0$

$$f_\sigma(\boldsymbol{x}) = (2\pi\sigma^2)^{-\frac{d}{2}} \cdot \exp\left\{-\frac{x_1^2+\cdots+x_d^2}{2\sigma^2}\right\},$$

and we need to choose a proper parameter σ^* such that $f_{\sigma^*}(\boldsymbol{x})$ is similar to $\frac{1}{v_d}\cdot\mathbb{1}_{\{x_1^2+x_2^2+\cdots+x_d^2<1\}}$ in some way.

Because of the claim, "*when the dimension d is large, most of the volume of the unit ball $\mathbf{B}^d$ is concentrated near the sphere $\mathbb{S}^{d-1}$*," we choose a parameter σ^* such that

$$\mathbb{E}\left[\left(X^{(1)}\right)^2 + \left(X^{(2)}\right)^2 + \cdots + \left(X^{(d)}\right)^2\right] = 1, \tag{2.1}$$
$$\text{where } \boldsymbol{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(d)}) \sim f_{\sigma^*}(\boldsymbol{x}).$$

That is, σ^* ensures that the average mass of the PDF $f_{\sigma^*}(\boldsymbol{x})$ lies in the region near the sphere $\mathbb{S}^{d-1}$. In this manner, sample points from $f_{\sigma^*}(\boldsymbol{x})$ concentrate in the region of importance, i.e., the vicinity of the unit sphere $\mathbb{S}^{d-1}$.

Questions:

(a) (3 points) Show that the "good choice" σ^* satisfying Equation (2.1) is

$$\sigma^* = \frac{1}{\sqrt{d}}.$$

(Hint: You may consider using the Chi-squared distribution.)

We seek to find a $\sigma^* > 0$ for which random variables $X_1, \ldots, X_n$ sampled from

$$f_{\sigma^*}(x) = (2\pi(\sigma^*)^2)^{-\frac{d}{2}} \cdot \exp\left(-\frac{1}{2(\sigma^*)^2}\sum_{i=1}^{d} x_i^2\right)$$

satisfy

$$\mathbb{E}[\sum_{i=1}^{d} X_i^2] = 1$$

We immediately notice that the RV $\vec{X} = (X_1, X_2, \ldots, X_n)$ is drawn from a normal distribution and we can define a new random variable by

$$Q = \sum_{i=1}^{d} X_i^d$$

If $Q \sim \chi^2$, then $\mathbb{E}Q = 1$ which is precisely the condition we would like f_{σ^*} to have.
When does a sum of squares follow a chi-squared distribution? Precisely when the random variables $X_1, X_2, \ldots, X_n$ are independent standard normal random variables. For the multivariate normal distribution, being "standard" means that each random variable has zero mean and unit variance.
That is, with

$$f(\vec{x}) = \frac{1}{\sqrt{(2\pi)^k \det \Sigma}} \cdot \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right)$$

we have $\vec{\mu} = \vec{0}$ and $\Sigma_{ij} = c$, i.e.,

$$f(\vec{x}) = \frac{1}{\sqrt{(2\pi)^k}} \cdot \exp\left(-\frac{1}{2}x^T cIx\right) = \frac{1}{\sqrt{(2\pi)^k}} \cdot \exp\left(-\frac{c}{2}\sum_{i=1}^{k} x_i^2\right)$$

Now looking back at $f_{\sigma^*}(x)$, we have

$$f_{\sigma^*}(x) = \frac{1}{\sqrt{(2\pi(\sigma^*)^2)^d}} \cdot \exp\left(-\frac{1}{2(\sigma^*)^2}\sum_{i=1}^{d} x_i^2\right)$$

Setting the two equal, we can solve for σ^*:

$$\frac{c}{\sqrt{(2\pi)^k}} \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{k} x_i^2\right) = \frac{1}{\sqrt{(2\pi(\sigma^*)^2)^d}} \cdot \exp\left(-\frac{1}{2(\sigma^*)^2}\sum_{i=1}^{d} x_i^2\right)$$

So we let $c = d$ and $\sigma^* = \frac{1}{\sqrt{d}}$. ■

(b) (2 points) Compute v_{100} (i.e., the volume of the unit ball in 100-dimensional space) using the importance sampling method (see Equation (1.1)) and the multivariable normal PDF $f_{\sigma^*}(\boldsymbol{x})$ satisfying Equation (2.1). Provide your estimated value and the code for generating the value. You may use the sample size $n = 100000$. (Please feel free to use your preferred programming languages. You can also use any built-in functions of these programming languages that generate random numbers/vectors.)
Unfortunately, Python (even with extra libraries) is not very good at ultra-high precision floating point math. My code generated a volume estimate of

$$2.391122628395068972331284844e-40$$

which is quite far from the true value of

$$1.0329397732669577e-64$$

It does, however, correctly approximate the volume of a 2-ball and 3-ball, leading me to conclude that the problem indeed lies with Python and not my own implementation.

Homework > HW 4 > HW4.py > ...

```python
import numpy as np
import scipy
import math
from decimal import Decimal

def H(vec_x):
    sum_squares = sum(list(map(lambda x: x**2, vec_x)))
    if sum_squares < 1:
        return 1
    else:
        return 0

def f(vec_x):
    d = len(vec_x)
    sum_squares = sum(list(map(lambda x: x**2, vec_x)))

    return Decimal((2*math.pi/d)**(-d/2))*Decimal(-sum_squares/(2/d)).exp()

def volume(dim, n):
    summation = 0
    for i in range(1, n + 1):
        vec_x = scipy.stats.multivariate_normal.rvs(mean=None, cov=(1/dim), size=dim)
        summation += Decimal(H(vec_x))/Decimal(f(vec_x))

        if i % 1000 == 0:
            print(f"On iteration {i}/{n}\r", end="")

    return Decimal(summation)/Decimal(n)


print(f"Estimated volume: {volume(100, 100000)}")

```

3. (4 points) Let $\{X_n\}_{n=0}^{\infty}$ be a homogeneous Markov chain with a discrete state space $\mathcal{X}$. Let μ denote the PMF of X_0, i.e., $\mu(x) = \mathbb{P}(X_0 = x)$, for all $x \in \mathcal{X}$; furthermore, $p(x, y)$ denotes the transition probability of the Markov chain, i.e.,

$$p(x, y) = \mathbb{P}(X_{n+1} = y \,|\, X_n = x) = \mathbb{P}(X_1 = y \,|\, X_0 = x), \quad \text{for all } x, y \in \mathcal{X}.$$

Prove the following identity

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = \mu(x_0) \cdot p(x_0, x_1) \cdot p(x_1, x_2) \ldots p(x_{n-1}, x_n),$$

for all $n = 1, 2, \ldots$ and $x_0, x_1, \ldots, x_n \in \mathcal{X}$. (Hint: use the law of total probability/definition of conditional probability, and the Markov property in Eq. (1.2).)

By the law of total probability and the Markov property,

$$\begin{aligned}
\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) &= \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, \ldots, X_0 = x) \cdot \mathbb{P}(X_{n-1} = x_{n-1}, \ldots, X_0 = x) \\
&= \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}) \cdot \mathbb{P}(X_{n-1} = x_{n-1}, \ldots, X_0 = x) \\
&= p(x_{n-1}, x_n) \cdot \mathbb{P}(X_{n-1} = x_{n-1}, \ldots, X_0 = x) \\
&= p(x_{n-1}, x_n) \cdot p(x_{n-1}, x_{n-2}) \ldots \mathbb{P}(X_1 = x_1, X_0 = x_0) \\
&= p(x_{n-1}, x_n) \cdot p(x_{n-1}, x_{n-2}) \ldots \mathbb{P}(X_1 = x_1 \mid X_0 = x_0) \cdot \mathbb{P}(X_0 = x_0) \\
&= p(x_{n-1}, x_n) \cdot p(x_{n-2}, x_{n-1}) \ldots p(x_0, x_1) \cdot \mu(x_0) \\
&= \mu(x_0) \cdot p(x_0, x_1) \cdot p(x_1, x_2) \ldots p(x_{n-1}, x_n) \quad \blacksquare
\end{aligned}$$

References

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

H. Wang. *Monte Carlo simulation with applications to finance.* CRC Press, 2012.

APMA1690: Homework # 5 (Due by 11pm Oct 26)

"*A drunk man will eventually find his way home, but a drunk bird may get lost forever.*"

— Shizuo Kakutani

1 Review

Please read the review section before delving into the problem set.

1.1 Notations

- $\mathbb{Z}$ = the collection of all integers.
- $\mathbb{Z}^d = \underbrace{\mathbb{Z} \times \mathbb{Z} \times \cdots \times \mathbb{Z}}_{\text{Cartesian product, } d \text{ times}}$.
- For a Markov chain[1] (MC) $\{X_n\}_{n=0}^{\infty}$, the subscript n is conventionally referred to as "time." All the components X_n of the MC are random variables defined on an underlying probability space $(\Omega, \mathbb{P})$, i.e., $X_n : \Omega \to \mathcal{X}$.
- Let $\boldsymbol{A}$ be a matrix. The transpose of $\boldsymbol{A}$ is denoted as $\boldsymbol{A}^\intercal$.

1.2 Simple Random Walks

The following is the definition of the d-dimensional simple random walk (SRW), where $d \in \{1, 2, 3, \ldots\}$.

Definition 1.1. *Let $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ be $\mathbb{Z}^d$-valued random variables, i.e.,*

$$\begin{aligned} \xi_i : \ & \Omega \to \mathbb{Z}^d, \\ & \omega \mapsto \xi_i(\omega), \end{aligned}$$

for all $i = 1, 2, \ldots$. Suppose the random variables $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ are iid and satisfy

$$\mathbb{P}(\xi_i = \boldsymbol{e}_k) = \mathbb{P}(\xi_i = -\boldsymbol{e}_k) = \frac{1}{2d}, \quad \text{for all } k = 1, \ldots, d, \tag{1.1}$$

where $\boldsymbol{e}_k$ is the k-th axis of the d-dimensional space, i.e.,

$$\boldsymbol{e}_k = (0, \ldots, 0, \underbrace{1}_{\uparrow\, k^{th}}, 0, \ldots, 0).$$

[1] All Markov chains utilized throughout this semester are assumed to be homogeneous Markov chains (HMCs). The "homogeneous" will sometimes be suppressed for simplicity.

Particularly, when $d = 1$, we have $e_1 = 1$ and $-e_1 = -1$.

We define the sequence $\{X_n\}_{n=0}^{\infty}$ of random variables taking values in $\mathbb{Z}^d$ as the following

$$X_0(\omega) = x_0 \in \mathbb{Z}^d \quad \text{for all } \omega \in \Omega,$$
$$X_n(\omega) = x_0 + \sum_{i=1}^{n} \xi_i(\omega), \quad \text{for all } n = 1, 2, \ldots,$$

where $x_0 \in \mathbb{Z}^d$ is a fixed point as an initialization. The sequence $\{X_n\}_{n=0}^{\infty}$ is called the d-dimensional ***simple random walk****.*

Each random variable ξ_i can be viewed as one step of the drunk man ($d = 2$) or the drunk bird ($d = 3$), and the initialization x_0 can be viewed as the "home" of the man/bird. The cumulation of the steps is the walking/flying path of the man/bird. Since the man/bird is drunk, each step ξ_i is totally random; thus, we have the "drunk step distribution" in Eq. (1.1).

1.3 Recurrence and Transience

Let $\{X_n\}_{n=0}^{\infty}$ be a homogeneous Markov chain taking values in the discrete state space $\mathcal{X}$, which can be either finite or infinite. For this Markov chain and each element $y \in \mathcal{X}$, we define the following random variable

$$\begin{aligned} T_y(\omega) &\stackrel{\text{def}}{=} \min\{n > 0 \mid X_n(\omega) = y\} \\ &= \text{ the time at which the sequence } \{X_n(\omega)\}_{n=1}^{\infty} \text{ first visits } y. \end{aligned}$$

Here, we explicitly write down ω to emphasize that T_y is a random variable. We denote the following probability, which will be used to define "recurrence/transience" and "irreducibility."

$$\begin{aligned} \rho_{xy} &\stackrel{\text{def}}{=} \mathbb{P}(T_y < \infty \mid X_0 = x) \\ &= \text{the conditional probability that MC will visit } y \text{ at least once, given that it starts from } x. \end{aligned}$$

With the notations $\{\rho_{xy}\}_{x,y\in\mathcal{X}}$, we can define reccurrence and transience as follows.

Definition 1.2. *Suppose we have a homogeneous Markov chain $\{X_n\}_{n=0}^{\infty}$. We provide the following two versions of the same definition.*

1. *(Rigorous/mathematical version) A state $y \in \mathcal{X}$ is said to be* ***recurrent*** *for this Markov chain if $\rho_{yy} = 1$; otherwise (i.e., $\rho_{yy} < 1$), the state y is said to be* ***transient*** *for this Markov chain.*
2. *(Heuristic version) If the Markov chain starting from y (i.e., $X_0 = y$) will almost surely (i.e., with probability one) return to y, the state y is said to be recurrent for this Markov chain; otherwise, y is said to be transient.*

The following theorem gives an approach to identifying recurrence.

Theorem 1.1. *Let $\{X_n\}_{n=0}^{\infty}$ be a homogeneous Markov chain taking values in the discrete state space $\mathcal{X}$. For any $y \in \mathcal{X}$, we have the following*

1. *The state y is recurrent for the Markov chain if and only if*

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n = y \,|\, X_0 = y) = \infty.$$

2. *If y is not recurrent, we have the following for all $x \in \mathcal{X}$*

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n = y \,|\, X_0 = x) < \infty.$$

The proof of Theorem 1.1 involves the so-called "strong Markov property," which is beyond the scope of this course. Hence, the proof of Theorem 1.1 is omitted to avoid a cheating proof.

The following theorem shows that recurrence is "contagious."

Theorem 1.2. *Let $\{X_n\}_{n=0}^{\infty}$ be a homogeneous Markov chain taking values in the discrete state space $\mathcal{X}$, which can be either finite or infinite. If x is recurrent (i.e., $\rho_{xx} = 1$) and $\rho_{xy} > 0$, then we have*

1. *$\rho_{yy} = 1$, i.e., the state y is recurrent;*

2. *$\rho_{xy} = \rho_{yx} = 1$.*

That is, $\rho_{xx} = 1$ and $\rho_{xy} > 0$ implies $\rho_{xx} = \rho_{yy} = \rho_{xy} = \rho_{yx} = 1$.

The proof of Theorem 1.2 is given in my lecture notes (proof of Theorem 3.2.3 therein).

1.4 Irreducibility

Definition 1.3. *Let $\{X_n\}_{n=0}^{\infty}$ is a homogeneous Markov chain taking values in the discrete state space $\mathcal{X}$ and having transition probability p.*

1. *The Markov chain $\{X_n\}_{n=0}^{\infty}$ or transition probability p is said to be **recurrent** if all states of $\mathcal{X}$ are recurrent for this Markov chain.*

2. *The Markov chain $\{X_n\}_{n=0}^{\infty}$ or transition probability p is said to be **irreducible** if $\rho_{xy} > 0$ for all $x, y \in \mathcal{X}$, i.e., we have a positive probability of transiting between any two states.*

Because of the following theorem, the scenario where the state space $\mathcal{X}$ is finite is important in the Markov chain theory.

Theorem 1.3. *Let $\{X_n\}_{n=0}^{\infty}$ be a Markov chain taking values in the state space $\mathcal{X}$. If $\mathcal{X}$ is finite (i.e., $\#\mathcal{X} < \infty$), we have*

1. *$\mathcal{X}$ has at least one state that is recurrent for $\{X_n\}_{n=0}^{\infty}$;*

2. *furthermore, if $\{X_n\}_{n=0}^{\infty}$ is irreducible, $\{X_n\}_{n=0}^{\infty}$ is recurrent, i.e., all states in $\mathcal{X}$ are recurrent for $\{X_n\}_{n=0}^{\infty}$.*

1.5 Transition Matrices

Let $\{X_n\}_{n=0}^{\infty}$ be a homogeneous Markov chain taking values in $\mathcal{X} = \{x_1, x_2, \ldots, x_S\}$ and having the transition probability p. We define the following **transition matrix** of the Markov chain

$$\boldsymbol{P} = \begin{pmatrix} p(x_1, x_1) & p(x_1, x_2) & \cdots & p(x_1, x_S) \\ p(x_2, x_1) & p(x_2, x_2) & \cdots & p(x_2, x_S) \\ \vdots & \vdots & \ddots & \vdots \\ p(x_S, x_1) & p(x_S, x_2) & \cdots & p(x_S, x_S) \end{pmatrix}, \tag{1.2}$$

which is an $S \times S$ matrix.

2 Problem Set

1. Prove that the 1-dimensional simple random walk is not only a Markov chain but also a homogeneous Markov chain. Please derive the transition probability of the 1-dimensional simple random walk.

   The 1-dimensional simple random walk is defined by

$$X_n(\omega) = \begin{cases} x_0 & n = 0 \\ x_0 + \sum_{i=1}^{n} \xi_i(\omega) & n = 1, 2, \dots \end{cases}$$

   where

$$\xi_1, \xi_2, \dots, \ \xi_n \overset{iid}{\sim} \text{Unif}\left(\{-1, 1\}\right) \implies \mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = \frac{1}{2}$$

   Observe:

$$X_{n-1}(\omega) = x_0 + \sum_{i=1}^{n-1} \xi_i(\omega)$$

$$X_n(\omega) = x_0 + \sum_{i=1}^{n} \xi_i(\omega) = X_{n-1}(\omega) + \xi_n(\omega)$$

   Then for two particular states $x, y \in \mathfrak{X}$, we can say

$$\mathbb{P}(X_n = y \mid X_{n-1} = x) = \mathbb{P}(\xi_n = y - x)$$

   But since ξ_i are drawn iid, the probability does not depend on n – just on x and y:

$$\mathbb{P}(\xi_n = y - x) = \mathbb{P}(y - x = 1 \mid y - x \in \{-1, 1\}) = \mathbb{P}(y - x = -1 \mid y - x \in \{-1, 1\}) = \frac{1}{2}$$

   Heuristically, this makes sense: by the recursive formula above, the Markov chain changes by ± 1 every time step with equal probability. So the outcome of being in a particular state one time step after being in an earlier state represents half of the state space.

   And in fact, the above calculation is exactly the way to calculate the transition probability:

$$p(x, y) = \mathbb{P}(X_n = y \mid X_{n-1} = x) = \frac{1}{2}$$

   Since this function does not depend on n, the MC is homogenenous. ■

2. Prove that the point $0 \in \mathbb{Z}$ is recurrent for the 1-dimensional simple random walk.

   You may consider applying the following results: (You do not need to prove any of the following results.)

   - Theorem 1.1.
   - Let $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ be two sequences of nonnegative numbers. Suppose $\lim_{n\to\infty}(a_n/b_n) = 1$. Then, we have $\sum_{n=1}^{\infty} a_n < \infty$ if and only if $\sum_{n=1}^{\infty} b_n < \infty$.
   - (Stirling's formula).

$$\lim_{n\to\infty} \frac{n!}{n^n \cdot e^{-n} \cdot \sqrt{2\pi n}} = 1.$$

   By Theorem 1.1 the state $0 \in \mathbb{Z}$ is recurrent for the 1-d simple random walk if and only if

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n = 0 \mid X_0 = 0) = \infty$$

   By the definition of this 1-d SRW

$$X_n(\omega) = \sum_{i=1}^{n} \xi_i(\omega)$$

   with $\xi_i(\omega) \overset{iid}{\sim} \text{Unif}(\{-1, 1\})$. So

$$\mathbb{P}(X_n = 0) = \mathbb{P}\left(\sum_{i=1}^{n} \xi_i(\omega) = 0\right)$$

   So when does $\sum_{i=1}^{n} \xi_i(\omega) = 0$? Clearly, it must be after an even number of steps. So, we need to calculate the probability that out of the $2n$ steps of the chain, exactly n of them are in one particular direction (WLOG say $\xi = 1$). Thus,

$$\mathbb{P}\left(\sum_{i=1}^{n} \xi_i(\omega) = 0\right) = \frac{1}{2^{2n}} \binom{2n}{n} = \frac{1}{2^{2n}} \cdot \frac{(2n)!}{n!(2n-n)!} = \frac{1}{2^{2n}} \cdot \frac{(2n)!}{n! \cdot n!}$$

   By Sterling's formula,

$$\frac{1}{2^{2n}} \cdot \frac{(2n)!}{n! \cdot n!} \approx \frac{1}{2^{2n}} \cdot \frac{(2n)^{2n} e^{-2n} \sqrt{4\pi n}}{[n^n e^{-n} \sqrt{2\pi n}]^2} \approx \frac{1}{\sqrt{\pi n}}$$

   (for large n).

   So to review, we have shown that

$$\mathbb{P}(X_n = 0 \mid X_0 = 0) = \mathbb{P}\left(\sum_{i=1}^{n} \xi_i(\omega) = 0\right) = \frac{1}{\sqrt{\pi n}}$$

So

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n = 0 \mid X_0 = 0) = \sum_{n=1}^{\infty} \frac{1}{\sqrt{\pi n}}$$

But by the integral convergence test,

$$\int_1^{\infty} \frac{1}{\sqrt{\pi n}}\, dn = \frac{1}{\sqrt{\pi}} \int_1^{\infty} n^{-\frac{1}{2}}\, dn = \frac{1}{\sqrt{\pi}} [2\sqrt{n}]_1^{\infty} = \infty$$

So the sum is ∞. Thus by Theorem 1.1, $0 \in \mathbb{Z}$ is recurrent. ■

3. (1 point) Let $\{X_n\}_{n=0}^{\infty}$ be a homogeneous Markov chain taking values in the state space $\mathcal{X} = \{x_1, x_2, \ldots, x_S\}$ and having transition probability p.

   - Let $\boldsymbol{P}$ denote the transition matrix of the Markov chain (see Eq. (1.2)).
   - $\pi : \mathcal{X} \to [0, 1]$ is a probability mass function. Denote the column vector $(\pi(x_1), \pi(x_2), \ldots, \pi(x_S))^\intercal$ as $\boldsymbol{\pi}$.

   Prove that the matrix equation $\boldsymbol{\pi} = \boldsymbol{P}^\intercal \boldsymbol{\pi}$ is equivalent to the following

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

$$\begin{aligned}
\vec{\pi} &= P^T \vec{\pi} \\
&= \begin{pmatrix} p(x_1,x_1) & p(x_2,x_1) & \cdots & p(x_S,x_1) \\ p(x_1,x_2) & p(x_2,x_2) & \cdots & p(x_S,x_2) \\ \vdots & \vdots & \ddots & \vdots \\ p(x_1,x_S) & p(x_2,x_S) & \cdots & p(x_S,x_S) \end{pmatrix} \begin{pmatrix} \pi(x_1) \\ \pi(x_2) \\ \vdots \\ \pi(x_S) \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^S p(x_i,x_1) \cdot \pi(x_1) \\ \sum_{i=1}^S p(x_i,x_2) \cdot \pi(x_2) \\ \vdots \\ \sum_{i=1}^S p(x_i,x_S) \cdot \pi(x_S) \end{pmatrix}
\end{aligned}$$

Then since $\mathfrak{X} = \{x_1, x_2, \ldots, X_S\}$,

$$\vec{\pi}_i = \pi(x_i) = \sum_{j=1}^{S} p(x_i, x_j) \cdot \pi(x_i) = \sum_{y \in \mathfrak{X}} \pi(y) \cdot p(y, x_i)$$

So $\forall x \in \mathfrak{X}$,

$$\pi(x) = \sum_{y \in \mathfrak{X}} \pi(y) \cdot p(y, x) \quad \blacksquare$$

APMA1690: Homework # 6 (Due by 11pm on November 2)

1 Review

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

1.1 Notations

- $\mathbb{Z}$ = the collection of all integers.
- $\mathbb{Z}^d = \underbrace{\mathbb{Z} \times \mathbb{Z} \times \cdots \times \mathbb{Z}}_{\text{Cartesian product, } d \text{ times}}$.
- For a Markov chain (MC) $\{X_n\}_{n=0}^{\infty}$, the subscript n is conventionally referred to as "time."
- Let $\boldsymbol{A}$ be a matrix. The transpose of $\boldsymbol{A}$ is denoted as $\boldsymbol{A}^\intercal$.
- Let $\boldsymbol{P}$ be an S-by-S matrix. $\boldsymbol{P}^n = \underbrace{\boldsymbol{P}\boldsymbol{P}\cdots\boldsymbol{P}}_{\text{matrix multiplication, } n \text{ matrices}}$
- $(\boldsymbol{P}^n)_{ij}$ = the entry in the i^{th} row and j^{th} column of the matrix $\boldsymbol{P}^n$.

1.2 Notation ρ_{xy}

Let $\{X_n\}_{n=0}^{\infty}$ be a homogeneous Markov chain taking values in the discrete state space $\mathcal{X}$, which can be either finite or infinite. Recall that each X_n, for a fixed n, is a random variable, i.e.,

$$\begin{aligned} X_n : \ & \Omega \to \mathcal{X}, \\ & \omega \mapsto X_n(\omega). \end{aligned}$$

For the Markov chain and each element $y \in \mathcal{X}$, we define the following random variable

$$\begin{aligned} T_y(\omega) &\stackrel{\text{def}}{=} \min\{n > 0 \mid X_n(\omega) = y\} \\ &= \text{ the time at which the sequence } \{X_n(\omega)\}_{n=1}^{\infty} \text{ first visits } y. \end{aligned}$$

Here, we explicitly write down ω to emphasize that T_y is a random variable. We denote the following probability, which will be used to define "recurrence/transience" and "irreducibility."

$$\begin{aligned} \rho_{xy} &\stackrel{\text{def}}{=} \mathbb{P}(T_y < \infty \mid X_0 = x) \\ &= \text{the conditional probability that MC will visit } y \text{ at least once, given that it starts from } x. \end{aligned}$$

1.3 Transition Matrices

Let $\{X_n\}_{n=0}^{\infty}$ be a homogeneous Markov chain whose state space is $\mathcal{X} = \{x_1, x_2, \ldots, x_S\}$ (where $S < \infty$) and having transition probability p. We define the following **transition matrix** of the Markov chain

$$\boldsymbol{P} = \begin{pmatrix} p(x_1, x_1) & p(x_1, x_2) & \cdots & p(x_1, x_S) \\ p(x_2, x_1) & p(x_2, x_2) & \cdots & p(x_2, x_S) \\ \vdots & \vdots & \ddots & \vdots \\ p(x_S, x_1) & p(x_S, x_2) & \cdots & p(x_S, x_S) \end{pmatrix}. \tag{1.1}$$

It is straightforward that $\sum_{j=1}^{S} \boldsymbol{P}_{ij} = \sum_{j=1}^{S} p(x_i, x_j) = 1$ for all $i \in \{1, 2, \ldots, S\}$.

1.4 Irreducibility

Definition 1.1. *Let $\{X_n\}_{n=0}^{\infty}$ is a homogeneous Markov chain taking values in the discrete state space $\mathcal{X}$ and having transition probability p.*

1. *The Markov chain $\{X_n\}_{n=0}^{\infty}$ or transition probability p is said to be **recurrent** if all states of $\mathcal{X}$ are recurrent for this Markov chain.*
2. *The Markov chain $\{X_n\}_{n=0}^{\infty}$ or transition probability p is said to be **irreducible** if $\rho_{xy} > 0$ for all $x, y \in \mathcal{X}$, i.e., we have a positive probability of transiting between any two states.*

Because of the following theorem, the scenario where the state space $\mathcal{X}$ is finite is of importance in the Markov chain theory.

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

1. *$\mathcal{X}$ has at least one state that is recurrent for $\{X_n\}_{n=0}^{\infty}$;*
2. *furthermore, if $\{X_n\}_{n=0}^{\infty}$ is irreducible, then $\{X_n\}_{n=0}^{\infty}$ is recurrent, i.e., all states in $\mathcal{X}$ are recurrent for $\{X_n\}_{n=0}^{\infty}$.*

1.5 Stationary Distributions

Definition 1.2. *Let $\{X_n\}_{n=0}^{\infty}$ be a homogeneous Markov chain taking values in the discrete state space $\mathcal{X}$ and having transition probability p. If a PMF π defined on $\mathcal{X}$ (i.e., $\pi : \mathcal{X} \to [0, 1]$) satisfies the following equation,*

$$\pi(x) = \sum_{y \in \mathcal{X}} \pi(y) \cdot p(y, x), \quad \text{for all } x \in \mathcal{X}, \tag{1.2}$$

*the PMF π is called a **stationary distribution** or **invariant distribution** for $\{X_n\}_{n=0}^{\infty}$.*

Generally, for a given transition probability p, the existence and uniqueness of a stationary distribution π satisfying Eq. (1.2) are not guaranteed and not trivial. For example, simple random walks taking values in $\mathbb{Z}^d$ do not have a stationary distribution. The general theory of the existence and uniqueness of stationary distributions is sort of complicated (Durrett, 2010, Section 6.5).

When the state space $\mathcal{X}$ is finite, the existence and uniqueness of the stationary distribution of an irreducible Markov chain are crystal clear and easy, which are presented in the following theorem and follow from the "irreducible non-negative matrices version" of the Perron-Frobenius theorem in linear algebra (Meyer, 2000, Section 8.3).

Theorem 1.2. *Let $\boldsymbol{P} = (p(x_i, x_j))_{1\le i,j\le S}$ be the transition matrix of a homogeneous Markov chain $\{X_n\}_{n=0}^{\infty}$ taking values in a finite state space $\mathcal{X} = \{x_1, x_2, \ldots, x_S\}$. If the Markov chain $\{X_n\}_{n=0}^{\infty}$ is **irreducible**, this chain **has** a **unique** stationary distribution on $\mathcal{X}$, i.e., $\pi(x) = \sum_{y\in\mathcal{X}} \pi(y)\cdot p(y, x)$, for all $x \in \mathcal{X}$; equivalently,*

$$\boldsymbol{\pi}^\intercal = \boldsymbol{\pi}^\intercal \boldsymbol{P}, \tag{1.3}$$

where $\boldsymbol{\pi} = (\pi(x_1), \ldots, \pi(x_S))^\intercal$. Furthermore, $\pi(x_i) > 0$ for all $i = 1, \ldots, S$.

1.6 Directed Graphs

Definition 1.3. *For the transition matrix $\boldsymbol{P}$ of a homogeneous Markov chain taking values in $\mathcal{X} = \{x_1, \ldots, x_S\}$, we define the **directed graph** $G(\boldsymbol{P}) = (V, E)$ **for** $\boldsymbol{P}$ as follows*

1. *The collection of vertices of the graph is $V = \{x_1, \ldots, x_S\}$;*
2. *The collection of directed edges of the graph is E, and the directed edge $(x_i \to x_j) \in E$ if and only if $\boldsymbol{P}_{ij} = p(x_i, x_j) > 0$.*

The following example helps you get familiar with the definition above. Consider the following transition probability matrix

$$\boldsymbol{P} = \begin{pmatrix} 0.3 & 0 & 0 & 0 & 0.70 & 0 & 0 \\ 0.1 & 0.2 & 0.3 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 & 0 \\ 0.6 & 0 & 0 & 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}. \tag{1.4}$$

The directed graph $G(\boldsymbol{P})$ associated with the $\boldsymbol{P}$ is the one in Figure 1.

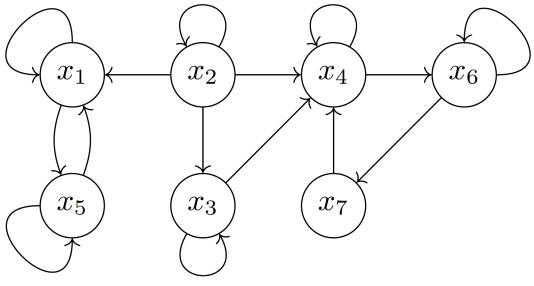


Figure 1: The directed graph $G(\boldsymbol{P})$, where $\boldsymbol{P}$ is the one defined in Eq. (1.4).

Definition 1.4. *Let $G(\boldsymbol{P})$ be the directed graph defined in Definition 1.3. For two given vertices $x, y \in V$, we say that there exists a directed path going from x to y, if there exist finitely many vertices, say $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(l)}$, such that the following conditions are satisfied*

- $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(l)} \in V$;
- *the directed edges $(x \to \xi^{(1)}), (\xi^{(1)} \to \xi^{(2)}), (\xi^{(2)} \to \xi^{(3)}), \ldots, (\xi^{(l-1)} \to \xi^{(l)}), (\xi^{(l)} \to y)$ belong to E.*

The collection $\{(x \to \xi^{(1)}), (\xi^{(1)} \to \xi^{(2)}), (\xi^{(2)} \to \xi^{(3)}), \ldots, (\xi^{(l-1)} \to \xi^{(l)}), (\xi^{(l)} \to y)\}$ of directed edges is called a path going from x to y.

For example, in Figure 1, $\{(x_2 \to x_3), (x_3 \to x_4)\}$ is a directed path going from x_2 to x_4.

Theorem 1.3. *Let $\boldsymbol{P}$ be the transition matrix of a homogeneous Markov chain taking values in the state space $\mathcal{X} = \{x_1, \ldots, x_S\}$. The Markov chain is irreducible if and only if $G(\boldsymbol{P})$ satisfies the following: for each pair of vertices x_i and x_j, there exist at least one directed path going from x_i to x_j and one directed path going from x_j to x_i.*

The proof of Theorem 1.3 is left as a homework question.

1.7 Asymptotic Theorems

Theorem 1.4. *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*

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{P}(X_n = x_j \,|\, X_0 = x_i) = \pi(x_j) \quad \text{for all } i, j \in \{1, \ldots, S\}, \quad \text{equivalently} \\ &\lim_{n\to\infty} (\boldsymbol{P}^n)_{ij} = \pi(x_j) \quad \text{for all } i, j \in \{1, \ldots, S\}, \end{aligned} \tag{1.5}$$

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

Theorem 1.5 further implies the following

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

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

Theorem 1.5. *Let $\{X_n\}_{n=0}^{\infty}$ be a homogeneous Markov chain taking values in a finite state space $\mathcal{X}$. If this chain is irreducible, for any function $f : \mathcal{X} \to \mathbb{R}$ such that $\sum_{x\in\mathcal{X}} |f(x)| \cdot \pi(x) < \infty$, we have*

$$\lim_{n\to\infty} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right\} = \sum_{x\in\mathcal{X}} f(x) \cdot \pi(x), \quad \text{with probability one,} \tag{1.6}$$

where π is the stationary of $\{X_n\}_{n=0}^{\infty}$.

The "with probability one" in Eq. (1.6) means the following

$$\mathbb{P}\left\{\omega \in \Omega \,\middle|\, \lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n} f\left(X_i(\omega)\right)\right] = \sum_{x\in\mathcal{X}} f(x)\cdot\pi(x)\right\} = 1.$$

2 Problem Set

1. (2 points) Prove Theorem 1.3 (see the Review section).

   Theorem 1.3 claims that a finite Markov Chain is irreducible iff its directed graph has as least one directed path from $x_i \to x_j$ and at least one from $x_j \to x_i$ for every pair of vertices (x_i, x_j).

   If the MC is irreducible, then for any states x_i, x_j

   $$\rho_{x_i,x_j} > 0$$

   i.e., there is a non-zero probability of eventually reaching x_i, starting at x_j. Equivalently, there is a path from x_j to x_i. Similarly, since x_i and x_j are arbitrary, $\rho_{x_j,x_i} > 0$. Thus, we have a path in the directed graph from any point x_i to x_j and vice versa.

   The proof of the other direction, is practically the same. If there exist edges in the directed graph, then $p(x_i, x_j) > 0$ and $p(x_j, x_i) > 0$ so $\rho_{x_i,x_j} > 0$ and $\rho_{x_j,x_i} > 0$.d Then as these points are arbitrary, the MC is irreducible. ■

2. (3 points) In this question, you will be asked to provide a partial proof for Theorem 1.2. By solving this problem, you will observe how algebraic topology (MATH 2410, Hatcher (2002)), linear algebra, and probability theory intersect.

   We first state a generalized version of the Brouwer fixed-point theorem, which is a result in algebraic topology:

   Theorem 2.1 (generalized Brouwer fixed-point theorem)**.** *Let $K \subset \mathbb{R}^S$ be convex, bounded, and closed. Every continuous function $f : K \to K$ has a fixed point, i.e., there exists $x \in K$ such that $f(x) = x$.*

   (Notice: To use the Brouwer fixed-point theorem, you have to verify that $f(K) \subset K$, i.e., $f(\boldsymbol{\xi}) \in K$ for all $\boldsymbol{\xi} \in K$.)

   Please assume that Theorem 2.1 is true and apply it to show the following:
   Let $\boldsymbol{P} = (p(x_i, x_j))_{1 \le i,j \le S}$ be the transition matrix of a homogeneous Markov chain $\{X_n\}_{n=0}^{\infty}$ taking values in a finite state space $\mathcal{X} = \{x_1, x_2, \ldots, x_S\}$. Then, there exists $\boldsymbol{\pi} \in \Delta$ such that $\boldsymbol{P}^\intercal \boldsymbol{\pi} = \boldsymbol{\pi}$, where

$$\Delta = \left\{ \boldsymbol{\xi} = (\xi_1, \ldots, \xi_S)^\intercal \in \mathbb{R}^S \;\middle|\; \sum_{i=1}^{S} \xi_i = 1 \text{ and } \xi_i \ge 0 \text{ for all } i = 1, 2, \ldots, S \right\}.$$

   Furthermore, if all entries of $\boldsymbol{P}$ are positive, i.e., $p(x_i, x_j) > 0$ for all i and j, all entries of the vector $\boldsymbol{\pi}$ are positive.

   Hint: Apply Theorem 2.1 by letting $K = \Delta$ and $f(\boldsymbol{\xi}) = \boldsymbol{P}^\intercal \boldsymbol{\xi}$. In your proof, you do not need to show that $K = \Delta$ is bounded and closed, which might be outside the scope of APMA 1690. But you need to show that $K = \Delta$ is convex (see the definition of convex sets by clicking the link).

   Remark: Here, we assume all entries of $\boldsymbol{P}$ are strictly positive. To remove the strict positivity condition, we need the irreducibility of the Markov chain. In addition, this question does not involve the uniqueness of $\boldsymbol{\pi}$. The uniqueness needs the the irreducibility structure. The set Δ defined above is usually referred to as the **probability simplex**, which is a fundamental building block of the simplicial homology theory in algebraic topology.

   Consider the probability simplex

$$\Delta = \left\{ \boldsymbol{\xi} = (\xi_1, \ldots, \xi_S)^\intercal \in \mathbb{R}^S \;\middle|\; \sum_{i=1}^{S} \xi_i = 1 \text{ and } \xi_i \ge 0 \text{ for all } i = 1, 2, \ldots, S \right\}.$$

   We assume the space is bounded and closed. To see that it is convex, observe that for $x, y \in \Delta$,

$$(1-t)x + ty \in \Delta \quad t \in [0, 1]$$

   because

$$(1-t)\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_S \end{pmatrix} + t\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_S \end{pmatrix} = \begin{pmatrix} x_1 - t(x_1 - y_1) \\ x_2 - t(x_2 - y_2) \\ \vdots \\ x_S - t(x_S - y_S) \end{pmatrix}$$

and

$$\begin{aligned}\sum_{i=1}^{S} x_i - t(x_i - y_i) &= \sum_{x=1}^{S} x_i - \sum_{i=1}^{S} t(x_i - y_i) \\ &= 1 - t\left(t\sum_{i=1}^{S} x_i - \sum_{i=1}^{S} y_i\right) \\ &= 1 - t(1-1) \\ &= 1\end{aligned}$$

To see the second condition, observe that x_i, y_i, t are all ≥ 0 so

$$x_1 - t(x_1 - y_1) = x_1 - tx_1 + ty_1$$

and so with $t = 0$, we have $x_1 \geq 0$ and $t = 1$, we have $y_1 \geq 0$ so $(1-t)x + ty \in \Delta$. Thus, the probability simplex fits the conditions of the Brouwer fixed-point theorem. Now consider $f(\vec{\xi}) = P^T\vec{\xi}$ where $\vec{\xi} \in \Delta$:

$$\begin{aligned}f(\vec{\xi}) &= \begin{pmatrix} p(x_1,x_1) & p(x_2,x_1) & \dots & p(x_S,x_1) \\ p(x_1,x_2) & p(x_2,x_2) & \dots & p(x_S,x_2) \\ \vdots & \vdots & \ddots & \vdots \\ p(x_1,x_S) & p(x_2,x_S) & \dots & p(X_S,x_S) \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_S \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^{S} \xi_i \cdot p(x_i,x_1) \\ \sum_{i=1}^{S} \xi_i \cdot p(x_i,x_2) \\ \vdots \\ \sum_{i=1}^{S} \xi_i \cdot p(x_i,x_S) \end{pmatrix}\end{aligned}$$

Clearly, this is a vector in $\mathbb{R}^S$ and further,

$$\begin{aligned}\sum_{j=1}^{S}\sum_{i=1}^{S} \xi_i \cdot p(x_i,x_j) &= \sum_{i=1}^{S}\sum_{j=1}^{S} \xi_i \cdot p(x_i,x_j) \\ &= \sum_{i=1}^{S} \xi_i \sum_{j=1}^{S} \cdot p(x_i,x_j) \\ &= \sum_{i=1}^{S} \xi_i \sum_{j=1}^{S} \cdot p(x_i,x_j) \\ &= \sum_{i=1}^{S} \xi_i \cdot 1 \\ &= 1\end{aligned}$$

Then because $0 \leq p(x_i,x_j)$ and $\xi_i \geq 0$, every entry in the vector will be a sum of strictly non-negative terms so the entries themselves will be non-negative. Thus, $f(\Delta) \subset \Delta$.

Then by the Brouwer fixed point theorem, there exists $\pi \in \Delta$ such that $f(\pi) = P^T \vec{\pi} = \vec{\pi}$.

Finally, if all entries of P are strictly positive, then each entry of π will be of the form

$$\pi(x_i) = \sum_{i=1}^{S} p(x_j, x_i) \cdot \pi(x_i)$$

But since we already know $\vec{\pi} \in \Delta$,

$$\sum_{I=1}^{S} \pi(x_i) = 1$$

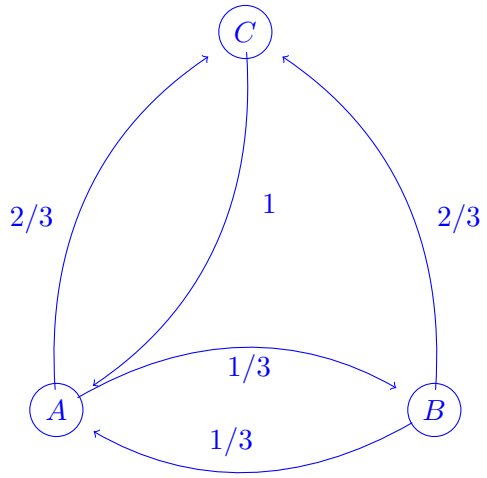
which means that at least one $\{\pi(x_i)\}_{i=1}^{S}$ must be positive and none of them can be negative. So the sum $\sum_{i=1}^{S} p(x_j, x_i) \cdot \pi(x_i)$ must be positive. ■

3. Suppose $\{X_n\}_{n=0}^{\infty}$ is a homogeneous Markov chain taking values in the state space $\mathcal{X} = \{x_1, x_2, x_3\}$, and its transition probability matrix is the following

$$\boldsymbol{P} = \begin{pmatrix} 0 & 1/3 & 2/3 \\ 1/3 & 0 & 2/3 \\ 1 & 0 & 0 \end{pmatrix}.$$

(a) (1 point) Prove that the Markov chain $\{X_n\}_{n=0}^{\infty}$ is irreducible. (Hint: You may use Theorem 1.3.)

The directed graph of P can be represented



Notice that there is a path from every node to every other node (A and B are trivial and $C \to A \to B$). Thus, by theorem 1.3, P is irreducible. ■

(b) (1 point) Prove that the Markov chain $\{X_n\}_{n=0}^{\infty}$ is aperiodic.

The period of a Markov chain is given by $d = \gcd(I_i)$ where $I_i = \{n \geq 1 \mid (P^n)_{ii} > 0\}$.

We calculate the first few powers:

$$P^2 = \begin{pmatrix} 7/9 & 0 & 2/9 \\ 6/9 & 1/9 & 2/9 \\ 0 & 3/9 & 6/9 \end{pmatrix}$$

$$P^3 = \begin{pmatrix} 2/9 & 7/27 & 14/27 \\ 7/27 & 2/9 & 14/27 \\ 7/9 & 0 & 2/9 \end{pmatrix}$$

And see that both $2 \in I_i$ and $3 \in I_i$ for $i = \{1, 2, 3\}$ because both main diagonals are positive. But 2 and 3 are already coprime so no matter what other values are in I_i, the GCD is 1 and the MC is aperiodic. ■

(c) (1 point) Compute the stationary distribution of the Markov chain $\{X_n\}_{n=0}^{\infty}$.

By Theorem 1.2, as the MC is irreducible, its unique stationary distribution is given by $\vec{\pi}^T = \vec{\pi}^T P$

$$\begin{aligned}\begin{pmatrix}\pi(x_1) & \pi(x_2) & \pi(x_3)\end{pmatrix} &= \begin{pmatrix}\pi(x_1) & \pi(x_2) & \pi(x_3)\end{pmatrix}\begin{pmatrix}0 & 1/3 & 2/3\\ 1/3 & 0 & 2/3\\ 1 & 0 & 0\end{pmatrix}\\ &= \begin{pmatrix}\frac{1}{3}\pi(x_1)+\pi(x_3) & \frac{1}{3}\pi(x_1) & \frac{2}{3}\pi(x_1)+\frac{2}{3}\pi(x_2)\end{pmatrix}\end{aligned}$$

Which gives a system we can put in terms of $\pi(x_1)$:

$$\begin{cases}\pi(x_1) = \frac{1}{3}\pi(x_1)+\pi(x_3)\\ \pi(x_2) = \frac{1}{3}\pi(x_1)\\ \pi(x_3) = \frac{2}{3}\pi(x_1)+\frac{2}{3}\pi(x_2)\end{cases} = \begin{cases}\pi(x_1) = \pi(x_1)\\ \pi(x_2) = \frac{1}{3}\pi(x_1)\\ \pi(x_3) = \frac{2}{3}\pi(x_1)+\frac{2}{9}\pi(x_1) = \frac{8}{9}\pi(x_1)\end{cases}$$

Further, we have that

$$\pi(x_1)+\pi(x_2)+\pi(x_3) = 1$$

so

$$\pi(x_1)+\frac{1}{3}\pi(x_1)+\frac{8}{9}\pi(x_1) = \frac{20}{9}\pi(x_1) = 1 \implies \pi(x_1) = \frac{9}{20}$$

Substituting back in,

$$\begin{cases}\pi(x_1) = \frac{9}{20}\\ \pi(x_2) = \frac{1}{3}\cdot\frac{9}{20} = \frac{3}{20}\\ \pi(x_3) = \frac{8}{9}\cdot\frac{9}{20} = \frac{8}{20}\end{cases}$$

so

$$\boxed{\vec{\pi}^T = \begin{pmatrix}9/20\\ 3/20\\ 2/5\end{pmatrix}}$$

(d) (1 point) Suppose the function $f : \mathcal{X} \to \mathbb{R}$ is defined as follows

$$f(x_k) = k^2, \quad \text{for all } k = 1, 2, 3.$$

Compute the following value.

$$\sum_{x\in\mathcal{X}} f(x)\cdot\pi(x),$$

where π is the stationary distribution you derived in part (c).

$$\begin{aligned}\sum_{x\in\mathfrak{X}} f(x)\cdot\pi(x) &= f(x_1)\cdot\pi(x_1)+f(x_2)\cdot\pi(x_2)+f(x_3)\cdot\pi(x_3)\\ &= 1\cdot\frac{9}{20}+4\cdot\frac{3}{20}+9\cdot\frac{2}{5}\\ &= \frac{9+12+72}{20}\\ &= \boxed{\frac{93}{20}}\end{aligned}$$

(e) (1 point) Compute the following limit

$$\lim_{n\to\infty} \boldsymbol{P}^n.$$

(Hint: the limit is a 3-by-3 matrix.)

Since the MC is irreducible and aperiodic, we can apply Theorem 1.4:

$$\lim_{n\to\infty} P^n = \begin{pmatrix} \pi(x_1) & \pi(x_2) & \pi(x_3) \\ \pi(x_1) & \pi(x_2) & \pi(x_3) \\ \pi(x_1) & \pi(x_2) & \pi(x_3) \end{pmatrix}$$

$$= \boxed{\begin{pmatrix} 9/20 & 3/20 & 2/5 \\ 9/20 & 3/20 & 2/5 \\ 9/20 & 3/20 & 2/5 \end{pmatrix}}$$

References

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

A. Hatcher. *Algebraic topology*. New York : Cambridge University Press, 2002.

C. D. Meyer. *Matrix analysis and applied linear algebra*, volume 71. Siam, 2000.

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.)

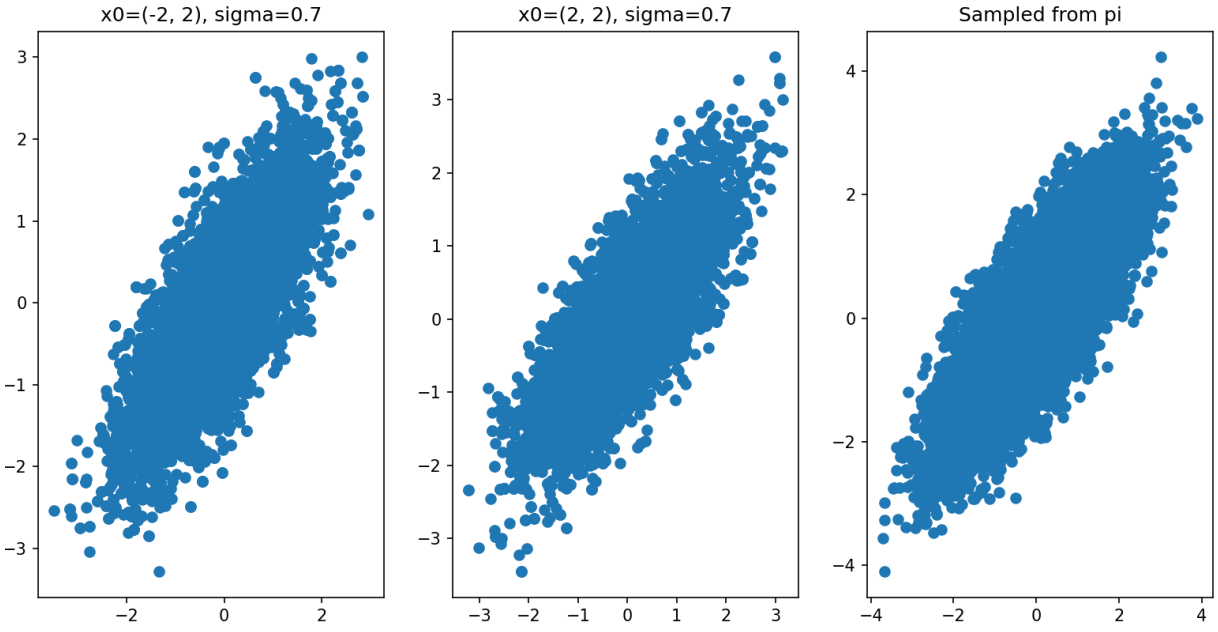
(b) (0.5 points) Replace $\boldsymbol{x}_0$ with $(2, 2)$ and repeat part (a).

(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:

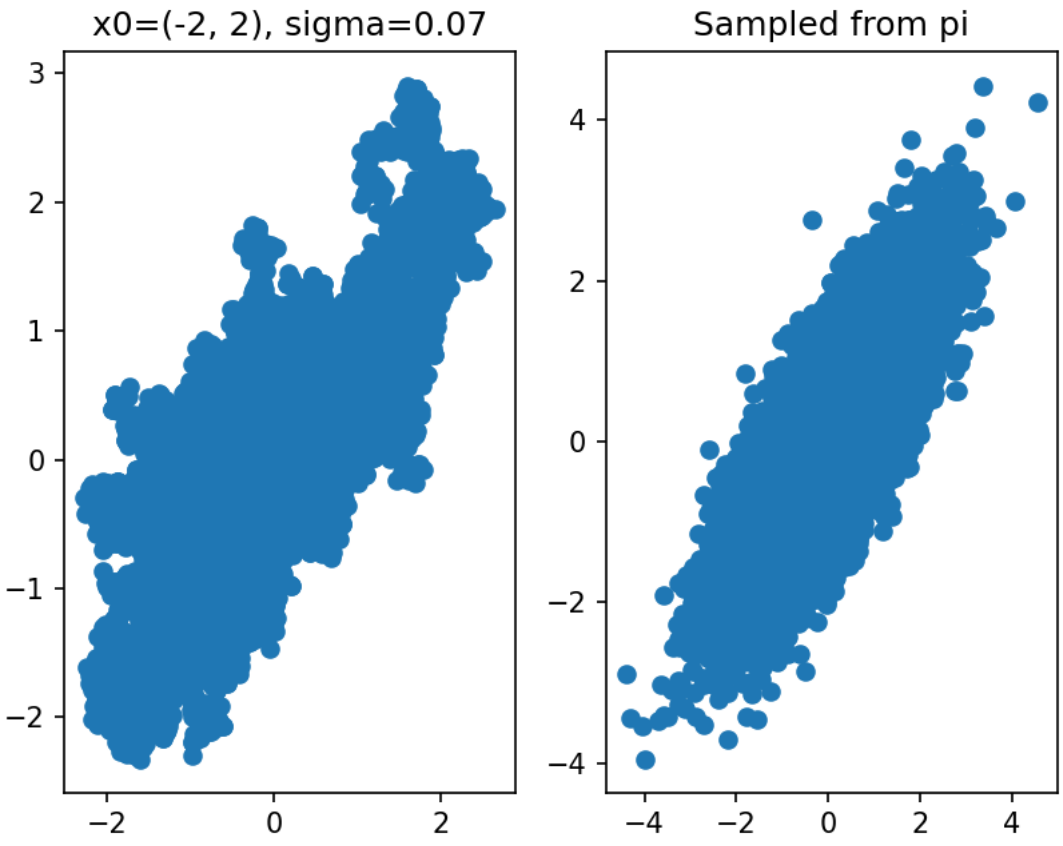
```
def abc():
    fig, (p1, p2, p3) = plt.subplots(1, 3)

    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.

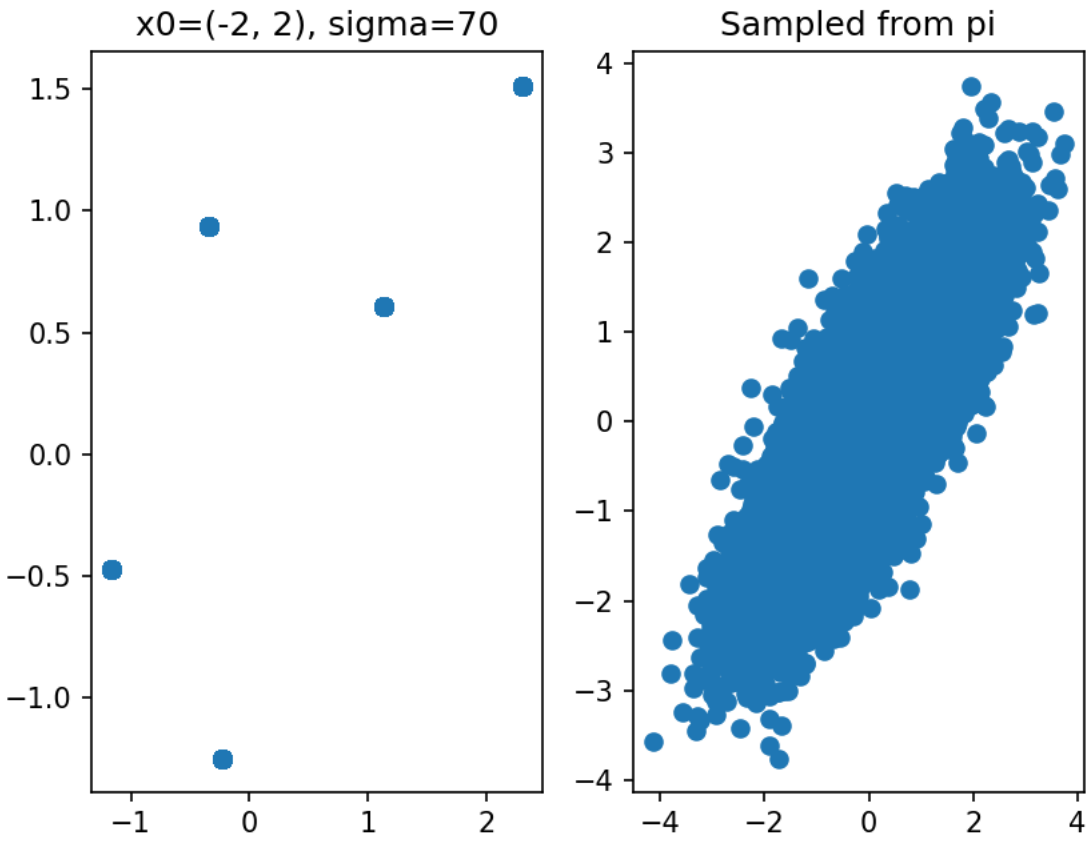
```
def d():
    fig, (p1, p3) = plt.subplots(1, 2)

    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.

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 $\{X_{50}, ..., X_{100}\}$.

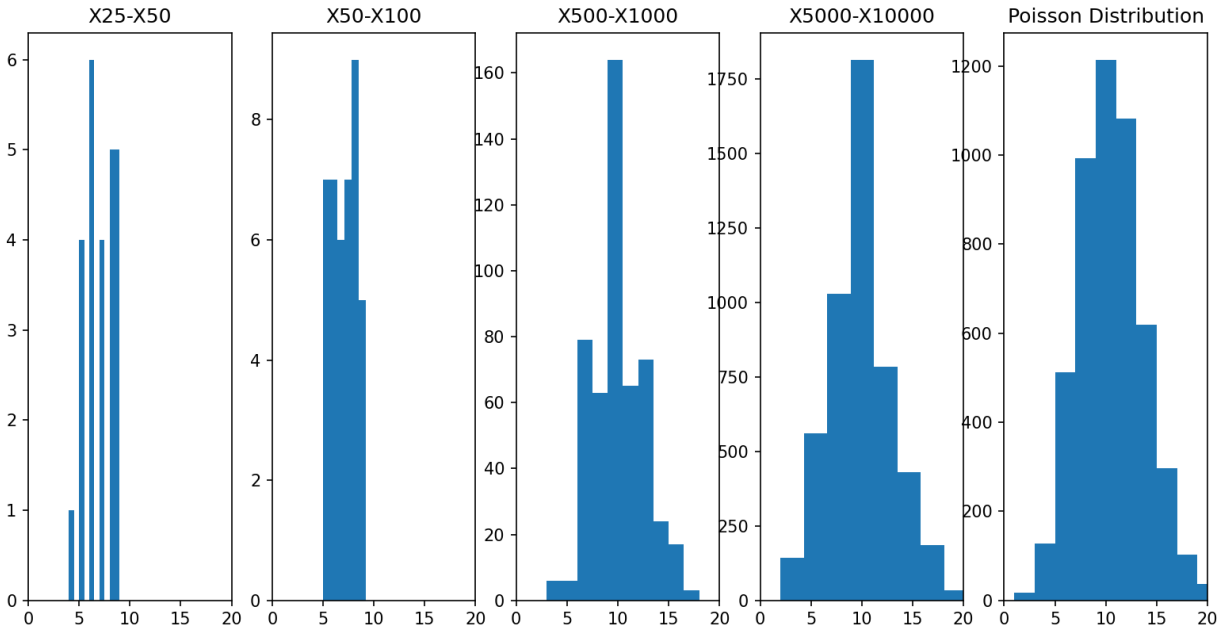
(c) Draw the histogram of $\{X_{500}, ..., X_{1000}\}$.

(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.

APMA1690: Homework # 9 (Due by 11pm on November 30)

"*He was too simple to wonder when he had attained humility. But he knew he had attained it, and he knew it was not disgraceful and it carried no loss of true pride.*"

— sentences from *The Old Man and the Sea*, by Ernest Hemingway

1 Review

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

1.1 Some Concepts in Graph Theory

We list as follows the concepts needed for graphical models

- A **graph** is an ordered pair $G = (V, E)$ comprising:
  - V is a set of **vertices**; the elements of V are usually indexed by positive integers.
  - E is a subset of $\{(i, j) \,|\, i, j \in V \text{ and } i \neq j\}$, i.e., pairs of vertices; if we view the pairs as unordered, i.e., $(i, j) = (j, i)$ for all i and j, this graph is called an **undirected graph**, otherwise, it is called a **directed graph**. (We focus on undirected graphs when talking about graphical models, i.e., we will keep assuming $(i, j) = (j, i)$.)
- For a given graph $G = (V, E)$, two vertices i and j are called **adjacent** if there is an edge between them, i.e., $(i, j) = (j, i) \in E$.
- For a given graph $G = (V, E)$ and a vertex $i \in V$, the **neighborhood** of i is defined as the collection of vertices adjacent to i, i.e., $\mathcal{N}(i) := \{j \in V \,|\, (i, j) \in E\}$. Since we do not allow 'self connecting edges' (i.e., vertices i and j should be different if then form an edge $(i, j) \in E$), we have $i \notin \mathcal{N}(i)$.
- For a given graph $G = (V, E)$, a set of vertices (i.e., a subset $c \subset V$) is called a **clique** if every pair of vertices in this set are adjacent. Furthermore, we denote $\mathcal{C}(G) := \{\text{all cliques of graph } G\}$. In addition, we adopt the convention that each vertex itself is a clique, i.e., $\{i\} \in \mathcal{C}(G)$ for all $i \in V$.

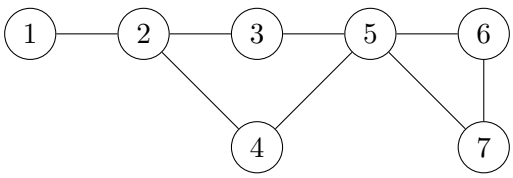


Figure 1: A graph $G = (V, E)$ with $V = \{1, 2, \ldots, 7\}$
and $E = \{(1, 2), (2, 3), (2, 4), (3, 5), (4, 5), (5, 6), (5, 7), (6, 7)\}$.

Figure 1 provides an example of a graph. In this graph, for example, vertices 1 and 2 are adjacent, vertices 2 and 3 are adjacent. Additionally, the neighborhood of vertex 5 is $\mathcal{N}(5) = \{3, 4, 6, 7\}$. We list all the elemnts of $\mathcal{C}(G)$ as follows

$$\begin{aligned} &\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \\ &\{1,2\}, \{2,3\}, \{2,4\}, \{3,5\}, \{4,5\}, \{5,6\}, \{5,7\}, \{6,7\}, \\ &\{5,6,7\}. \end{aligned} \tag{1.1}$$

1.2 Notations

We also adopt the following notations

- Let $G = (V, E)$ be a graph with $V = \{1, 2, \ldots, d\}$ and $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^\intercal$ a vector in the product space $\mathcal{X} = \mathcal{X}_1 \times \cdots \mathcal{X}_d$.
- For any subset $c = \{i_1, \ldots, i_k\} \subseteq V$ with $i_1 < i_2 < \cdots < i_k$, we denote $x_c := (x_{i_1}, x_{i_2}, \ldots, x_{i_k})^\intercal$ and $\mathcal{X}_c := \mathcal{X}_{i_1} \times \cdots \times \mathcal{X}_{i_k}$. For example, let $V = \{1, \ldots, 7\}$ and $c = \{1, 3, 5\}$, then $x_c = (x_1, x_3, x_5)^\intercal$ and $\mathcal{X}_c = \mathcal{X}_1 \times \mathcal{X}_3 \times \mathcal{X}_5$.

1.3 Application of Gibbs Sampling to the 2-Dimensional Ising Model

This subsection helps discover a classical result from **?**[1] in a numerical way.

1.3.1 Periodic Lattices

`(This subsection on periodic lattices is only for HW 9 and will not be required for the final exam.)`

We use periodic lattices, following the convention in the standard literature on the Ising model. The periodic structure will make the coding part easier. Without the periodic structure, we would have to deal with the boundaries of nonperiodic lattices separately.

A 3-by-3 periodic lattice Λ_3 is presented in Figure 2. An N-by-N periodic lattice Λ_N with a generic size N is defined in a similar way. Specifically, for the vertex in the i-th row and j-th column of an N-by-N periodic lattice, its neighbors are the following four vertices[2]

- the vertex in the k-th row and l-th column, where $k \in \{1, \ldots, N\}$ with $k - 1 \equiv i - 1 \mod (N)$ and $l \in \{1, \ldots, N\}$ with $l - 1 \equiv j - 2 \mod (N)$.
- the vertex in the k-th row and l-th column, where $k \in \{1, \ldots, N\}$ with $k - 1 \equiv i - 1 \mod (N)$ and $l \in \{1, \ldots, N\}$ with $l - 1 \equiv j \mod (N)$.
- the vertex in the k-th row and l-th column, where $k \in \{1, \ldots, N\}$ with $k - 1 \equiv i \mod (N)$ and $l \in \{1, \ldots, N\}$ with $l - 1 \equiv j - 1 \mod (N)$.

[1] Professor Onsager taught statistical mechanics at Brown University and did the research work important enough to gain him the unshared Nobel Prize in Chemistry in 1968. However, "the Great Depression limited Brown's ability to support a faculty member who was only useful as a researcher and not a teacher; he was let go by Brown, being hired by Yale University" (see Wikipedia).

[2] For integers a and b, the notation '$a \equiv b \mod (N)$' denotes the following: N is a divisor of $a - b$, i.e., there exists an integer k (not necessarily positive) such that $a - b = kN$.

- the vertex in the k-th row and l-th column, where $k \in \{1, \ldots, N\}$ with $k - 1 \equiv i - 2 \mod (N)$ and $l \in \{1, \ldots, N\}$ with $l - 1 \equiv j - 1 \mod (N)$.

Neighborhoods of 5 and 3 of the 3-by-3 periodic lattice Λ_3 (see Figure 2) are presented in Figures 3 and 4, respectively. The two figures visually interpret the complicated modular arithmetic notation 'mod (N)' above.

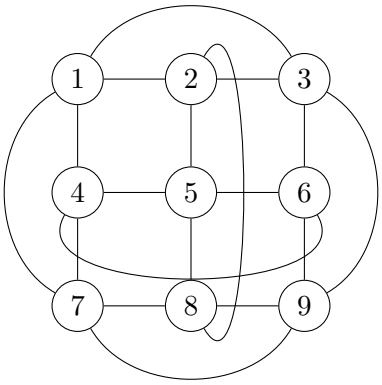


Figure 2: A 3-by-3 periodic lattice Λ_3, which is a graph.

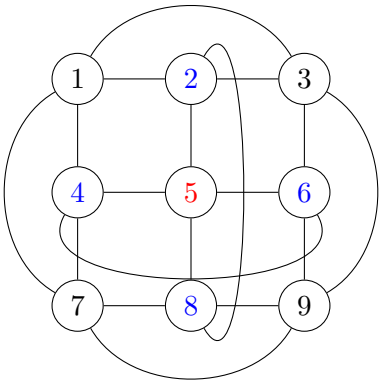


Figure 3: The **neighborhood** of vertex 5 (red) is $\mathcal{N}(5) = \{2, 4, 6, 8\}$, presented in blue.

1.3.2 Ising Model PMF

Suppose Λ_N is an N-by-N periodic lattice. Each vertex (also called a 'site,' i.e., Section 18.3 of **?**) is labeled by a positive integer, e.g., vertices of a 3-by-3 periodic lattice (see Figure 2) are labeled by $\{1, 2, \ldots, 9\}$.

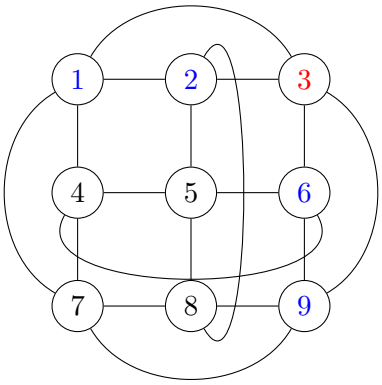


Figure 4: The **neighborhood** of vertex 3 (red) is $\mathcal{N}(3) = \{2, 1, 6, 9\}$, presented in blue.

At each vertex $k \in \Lambda_N$, there exists an atom with spin $s_k \in \{-1, 1\}$. A **spin configuration** $\boldsymbol{s}$ is a vector of spin values assigned to the N^2 vertices, i.e.,

$$\boldsymbol{s} = (s_k)_{k\in\Lambda} = (s_1, s_2, \ldots, s_{N^2}) .$$

Let $\mathcal{X}$ denote the collection of all possible spin configurations $\boldsymbol{s}$, i.e.,

$$\mathcal{X} = \underbrace{\{-1, 1\} \times \{-1, 1\} \times \cdots \times \{-1, 1\}}_{\text{Cartesian product of } N^2 \text{ sets}} . \tag{1.2}$$

It is straightforward that $\#\mathcal{X} = 2^{N^2}$.

Because of the perturbation from heat, the spin values are not deterministic, i.e, $\boldsymbol{s}$ is a random vector. The Ising model gives PMFs describing the randomness of $\boldsymbol{s}$. A simplified Ising model-based PMF is the following one characterizing a "ferromagnetic zero-field" model[3]

$$\begin{aligned} \pi_{N,\beta}(\boldsymbol{s}) &= \frac{1}{Z_\beta} \cdot \exp\left\{ \beta \cdot \sum_{\langle i,j\rangle} s_i s_j \right\} \\ &= \frac{1}{Z_\beta} \cdot \exp\left\{ \beta \cdot \left(\sum_{\text{all edges } (i,j) \text{ on the } N\text{-by-}N \text{ periodic lattice}} s_i s_j \right) \right\} , \end{aligned} \tag{1.3}$$

where $\langle i, j\rangle$ denotes that i and j are neighbors. The PMF formula in Eq. (1.3) is called the 2-dimensional Ising model[4]. The notations above are explained as follows:

- $\sum_{\langle i,j\rangle}$ denotes the sum over all pairs (i, j) such that i and j are neighbors. It is the standard notation in the Ising model literature, e.g., the Wikipedia page on the Ising Model.

[3]The PMF $\pi_{N,\beta}(\boldsymbol{s}) = \frac{1}{Z_\beta} \cdot \exp\left\{-\beta \cdot \sum_{\langle i,j\rangle} s_i s_j\right\}$ with an extra negative sign represents an "antiferromagnetic zero-field" model, which is not of interest.

[4]German pronunciation is like the English word "easing" (see the remark in Example 18.16 of **?**).

- β is a model parameter. Specifically, $\beta = 1/T$, where T indicates the temperature of the magnetic system.
- $Z_\beta = \sum_{\boldsymbol{s}\in\mathcal{X}} \exp\left\{\beta \cdot \sum_{\langle i,j\rangle} s_i s_j\right\}$ is the normalizing constant. As a function of β, the quantity Z_β is the partition function of the model. The letter Z stands for the German word *Zustandssumme*, "sum over states," and $\sum_{\boldsymbol{s}\in\mathcal{X}}$ is the sum over all possible spin configurations.
- $H_N(\boldsymbol{s}) = -\sum_{\langle i,j\rangle} s_i s_j$ presents the energy of the magnetic system, and $\pi_{N,\beta}(\boldsymbol{s}) = \frac{1}{Z_\beta} \cdot \exp\{-\beta \cdot H_N(\boldsymbol{s})\}$ is referred to as the Boltzmann distribution of the magnetic system in the statistical mechanics literature. Notably, the negative sign in $H_N(\boldsymbol{s})$ and the negative sign in the Boltzmann distribution are canceled out. Forgetting to cancel out the negative sign will make a ferromagnetic model become an antiferromagnetic model.

1.3.3 Curie Temperature

The Curie temperature is the one above which certain materials lose their permanent magnetic properties. The Curie temperature is named after Pierre Curie. In this section, we describe the Curie temperature using the Ising model. The discussion in this section provides a nontrivial question that can be solved by the Gibbs sampling algorithm.

Suppose the distribution $\pi_{N,\beta}(\boldsymbol{s})$ of spin configurations $\boldsymbol{s}$ is given by the ferromagnetic zero-field model in Eq. (1.3). Macroscopically, the individual spins cannot be observed, but the following average magnetization can

$$m_N(\beta) := \sum_{\boldsymbol{s}\in\mathcal{X}} \pi_{N,\beta}(\boldsymbol{s}) \cdot \left|\frac{\sum_{i\in\Lambda_N} s_i}{N^2}\right| = \mathbb{E}_{\pi_{N,\beta}} \left|\frac{\sum_{i\in\Lambda_N} s_i}{N^2}\right|, \tag{1.4}$$

where "$\sum_{\boldsymbol{s}\in\mathcal{X}}$" means the sum over all possible spin configurations. If we consider a very large system, then we are close to the so-called thermodynamic limit

$$m(\beta) := \lim_{N\to\infty} m_N(\beta). \tag{1.5}$$

The interpretation of $m(\beta)$ is roughly presented as follows: when $m(\beta) > 0$, the material of interest is magnetic; when $m(\beta) = 0$, it is not magnetic. The magic of Curie Temperature comes from this limiting procedure in Eq. (1.5). **?** essentially showed that there exists a critical value $\beta_C = \frac{1}{T_C}$, where T_C is the Curie temperature of interest, such that

$$m(\beta) \begin{cases} > 0, & \text{if } \beta > \beta_C, \\ = 0, & \text{if } \beta < \beta_C. \end{cases} \tag{1.6}$$

That is, we have the Curie temperature T_C is represented as follows

$$T_C = \frac{1}{\beta_C}. \tag{1.7}$$

Below the Curie temperature, the material of interest is magnetic; above the Curie temperature, it is not. **?** provided the exactly value of β_C as follows

$$\boxed{\beta_C = \frac{\log(1+\sqrt{2})}{2} \approx 0.44.} \tag{1.8}$$

1.3.4 Estimation of β_C

Mathematically deriving the value β_C in Eq. (1.8) is nearly a Nobel prize-level question. Using the Gibbs sampling, we can estimate this value numerically. To estimate β_C, we need to estimate the $m(\beta)$ defined in Eq. (1.5). When the size N of the periodic lattice Λ_N is large, we have $m(\beta) \approx m_N(\beta)$.

Suppose we have a homogeneous Markov chain (HMC)

$$\left\{ \boldsymbol{X}^{(n)} = \left(X_1^{(n)}, X_2^{(n)}, \ldots, X_{N^2}^{(n)} \right) \right\}_{n=0}^{\infty} \tag{1.9}$$

whose state space is the $\mathcal{X}$ defined in Eq. (1.2); furthermore, this HMC is irreducible and apperiodic, and its stationary distribution is the $\pi_{N,\beta}(\boldsymbol{s})$ defined in Eq. (1.3). Visualizations of $\boldsymbol{X}^{(10000)}$ with lattice size $N = 100$ for different β values are presented in Figure 5.

Then, the ergodic theorem implies the following with probability one

$$\lim_{n\to\infty} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\sum_{k\in\Lambda_N} X_k^{(i)}}{N^2} \right| \right\} = m_N(\beta).$$

Therefore, when both the lattice size N and sample size n are large, we have

$$m(\beta) \approx m_N(\beta) \approx \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\sum_{k\in\Lambda_N} X_k^{(i)}}{N^2} \right| =: \widehat{v}_{N,n}(\beta). \tag{1.10}$$

Then, we can apply Eq. (1.10) (i.e., the estimator $\widehat{v}_{N,n}(\beta)$) to estimate $m(\beta)$, which help us estimate β_C. Now, it remains to get such an HMC in Eq. (1.9), which can be done through the Gibbs sampling.

1.3.5 Application of the Gibbs Sampling to Ising Model

For each vertex i, we define the neighborhood $\mathcal{N}(i)$ of i by the following

$$\mathcal{N}(i) := \{\text{the neighbors of site } i\}.$$

Suppose we are interested in the spin s_{i^*} at the vertex i^*. The Boltzmann distribution $\pi_{N,\beta}(\boldsymbol{s})$ in Eq. (1.3) can be represented as follows

$$\pi_{N,\beta}(\boldsymbol{s}) = \frac{1}{Z_\beta} \cdot \exp\left\{ \beta \left[\sum_{j\in\mathcal{N}(i^*)} s_{i^*} s_j \right] + \beta \left[\sum_{\langle i',j'\rangle \text{ and } i',j'\neq i^*} s_{i'} s_{j'} \right] \right\}. \tag{1.11}$$

1.4 Conditional Distributions

A key quantity in the Gibbs sampling (see Algorithm 1) is the conditional distribution of s_{i^*} given the values of other coordinates, i.e.,

$$\pi_{N,\beta}(s_{i^*} | \boldsymbol{s}_{-i^*}) = \pi_{N,\beta}(s_{i^*} | s_1, \ldots, s_{i^*-1}, s_{i^*+1}, \ldots, s_{N^2}).$$

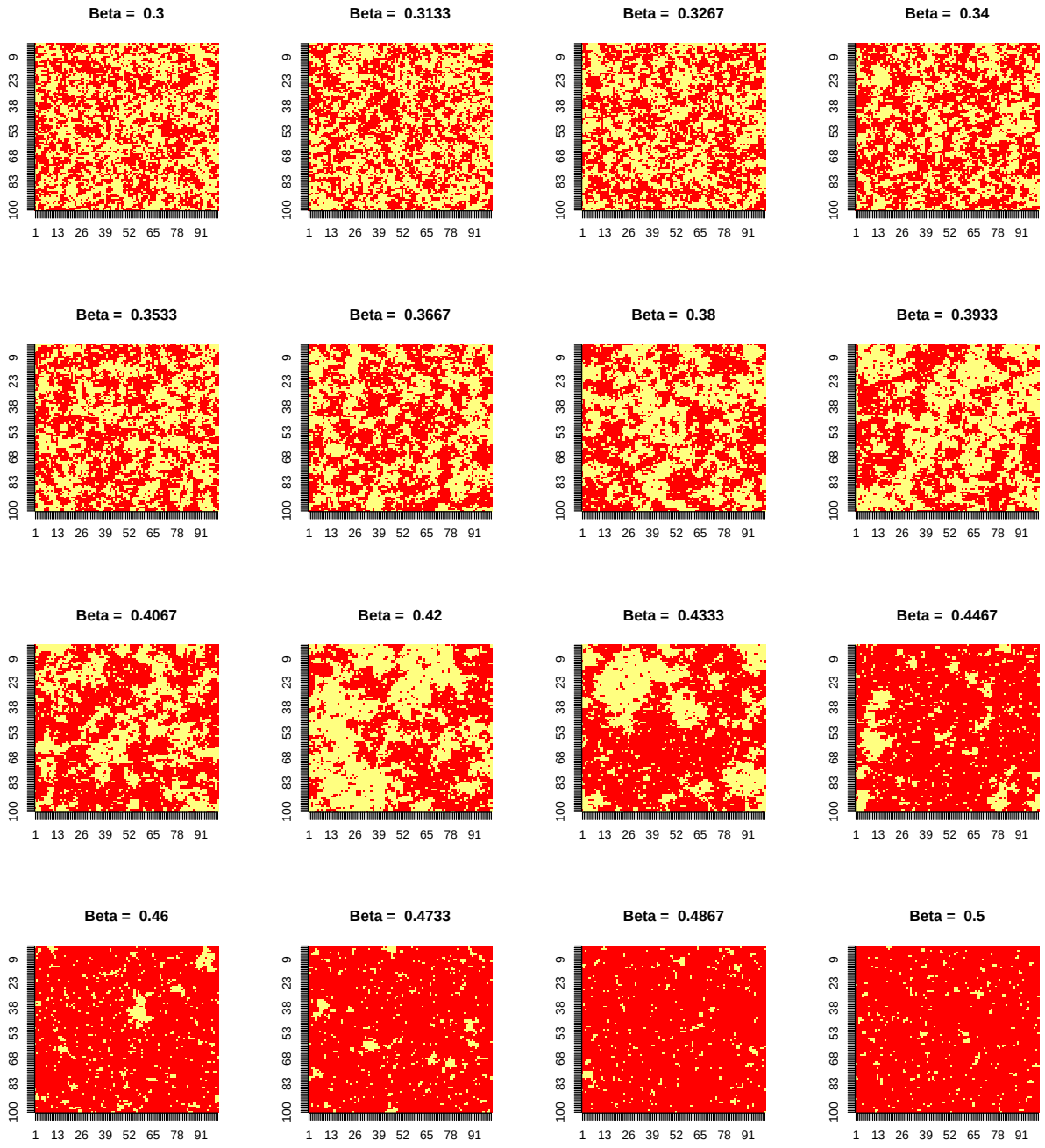


Figure 5: Visualizations of $\boldsymbol{X}^{(10000)}$ with lattice size $N = 100$ for different β values. Recall the critical value $\beta_C = \frac{\log(1+\sqrt{2})}{2} \approx 0.44$.

Using the representation in Eq. (1.11), the conditional distribution $\pi_{N,\beta}(s_{i^*}|\boldsymbol{s}_{-i^*})$ can be represented as follows

(1.12)
$$\begin{aligned}\pi_{N,\beta}(s_{i^*}|\boldsymbol{s}_{-i^*}) &= \frac{\pi_{N,\beta}(\boldsymbol{s})}{\sum_{s_{i^*}\in\{-1,1\}}\pi_{N,\beta}(\boldsymbol{s})}\\ &= \frac{\frac{1}{Z_\beta}\cdot\exp\left\{\beta\left[\sum_{\langle i',j'\rangle \text{ and } i',j'\neq i^*} s_{i'}s_{j'}\right]\right\}\cdot\exp\left\{\beta\left[\sum_{j\in\mathcal{N}(i^*)} s_{i^*}s_j\right]\right\}}{\frac{1}{Z_\beta}\cdot\exp\left\{\beta\left[\sum_{\langle i',j'\rangle \text{ and } i',j'\neq i^*} s_{i'}s_{j'}\right]\right\}\cdot\sum_{s_{i^*}\in\{-1,1\}}\exp\left\{\beta\left[\sum_{j\in\mathcal{N}(i^*)} s_{i^*}s_j\right]\right\}}\\ &= \frac{\exp\left\{\beta\left[\sum_{j\in\mathcal{N}(i^*)} s_{i^*}s_j\right]\right\}}{\sum_{s_{i^*}\in\{-1,1\}}\exp\left\{\beta\left[\sum_{j\in\mathcal{N}(i^*)} s_{i^*}s_j\right]\right\}}\\ &= \frac{\exp\left\{\beta\left[\sum_{j\in\mathcal{N}(i^*)} s_{i^*}s_j\right]\right\}}{\exp\left\{-\beta\left[\sum_{j\in\mathcal{N}(i^*)} s_j\right]\right\}+\exp\left\{\beta\left[\sum_{j\in\mathcal{N}(i^*)} s_j\right]\right\}},\end{aligned}$$

More precisely, we have the following

(1.13)
$$\begin{aligned}\pi_{N,\beta}(1\,|\boldsymbol{s}_{-i^*}) &= \frac{\exp\left\{\beta\left[\sum_{j\in\mathcal{N}(i^*)} s_j\right]\right\}}{\exp\left\{-\beta\left[\sum_{j\in\mathcal{N}(i^*)} s_j\right]\right\}+\exp\left\{\beta\left[\sum_{j\in\mathcal{N}(i^*)} s_j\right]\right\}},\\ \pi_{N,\beta}(-1\,|\boldsymbol{s}_{-i^*}) &= \frac{\exp\left\{-\beta\left[\sum_{j\in\mathcal{N}(i^*)} s_j\right]\right\}}{\exp\left\{-\beta\left[\sum_{j\in\mathcal{N}(i^*)} s_j\right]\right\}+\exp\left\{\beta\left[\sum_{j\in\mathcal{N}(i^*)} s_j\right]\right\}}.\end{aligned}$$

The factor $\frac{1}{Z_\beta}\cdot\exp\left\{\beta\left[\sum_{\langle i',j'\rangle \text{ and } i',j'\neq i^*} s_{i'}s_{j'}\right]\right\}$ in Eq. (1.12) is canceled out. This cancelation is extremely important as it reduces the redundant computation.

1.5 Sampling from Conditional Distributions

To sample a random variable $X_{i^*}^{(n)}$ from the conditional distribution $\pi_{N,\beta}(\,\cdot\,|\boldsymbol{s}_{-i^*})$ defined in Eq. (1.13), we can adopt the following procedure:

- Compute $a \leftarrow \exp\left\{\beta\left[\sum_{j\in\mathcal{N}(i^*)} s_j\right]\right\}$.
- Compute $b \leftarrow \exp\left\{-\beta\left[\sum_{j\in\mathcal{N}(i^*)} s_j\right]\right\}$.
- Compute $p \leftarrow \frac{a}{a+b}$.
- Generate $Z \sim \text{Bernoulli}(p)$, i.e., $\mathbb{P}(Z=1)=p$ and $\mathbb{P}(Z=0)=1-p$.
- $X_{i^*}^{(n)} \leftarrow 2\cdot Z-1$, i.e., $\mathbb{P}(X_{i^*}^{(n)}=1)=\frac{a}{a+b}$ and $\mathbb{P}(X_{i^*}^{(n)}=-1)=\frac{b}{a+b}$.

With the procedure above, we can sample the HMC in Eq. (1.9) through Gibbs sampling (Algorithm 1).

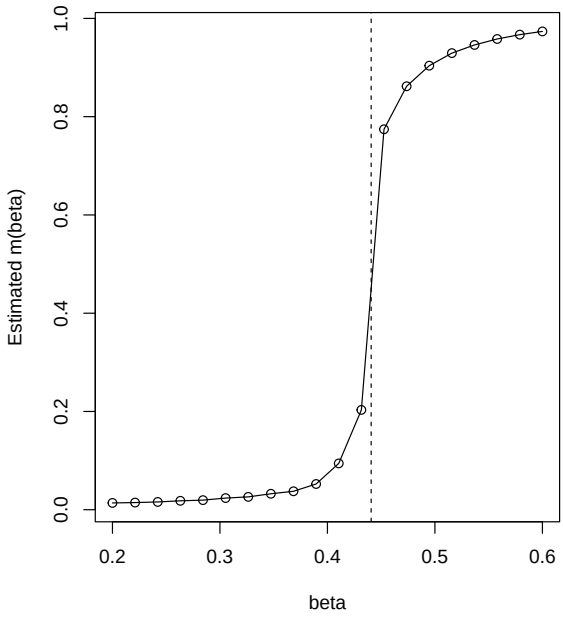


Figure 6: The '$\widehat{v}_{N,n}(\beta)$ vs. β' plot. The vertical dashed line indicates the critical value $\beta_C \approx 0.44$.

1.5.1 Numerical Experiment

As an ending section, we present a numerical experiment of the estimation $m(\beta) \approx \widehat{v}_{N,n}(\beta)$ presented in Eq. (1.10). In this experiment, we choose the lattice size to be $N = 100$ and the sample size to be $n = 10000$. The '$\widehat{v}_{N,n}(\beta)$ vs. β' plot is presented in Figure 6, where the vertical dashed line indicates the critical value $\beta_C \approx 0.44$. Figure 6 is compatible with Eq. (1.6).

1.5.2 Appendix: Gibbs Sampling

Algorithm 1 : 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 $\xi_j^{(n)} \sim \pi_{j|-j}(\xi_j | \xi_1^{(n)}, \ldots, \xi_{j-1}^{(n)}, \xi_{j+1}^{(n-1)}, \ldots, \xi_d^{(n-1)})$
6: **end for** $\boldsymbol{X}^{(n)} \leftarrow (\xi_1^{(n)}, \xi_2^{(n)}, \ldots, \xi_d^{(n)})^\intercal$
7: **end for**

2 Problem Set

1. (3 points) Let $G = (V, E)$ be a graph. Prove the following identity

$$\mathcal{N}(i) \bigcup \{i\} = \left(\bigcup_{c \in \mathcal{C}(G) \text{ and } i \in c} c \right), \tag{2.1}$$

where the union sign $\bigcup$ on the right hand side of Eq. (2.1) denotes the union over all cliques c such that $i \in c$. For the details of the union notation, please refer to the Wikipedia page on 'Union (set theory),' especially the "arbitrary unions" section therein.

Hint: see the definition of cliques.

Remark: The set identity in Eq. (2.1) partially implies the Hammersley–Clifford theorem.

$$\begin{aligned} \mathcal{N}(i) \cup \{i\} &= \{j \in V : (i,j) \in E\} \cup \{i\} \\ &= \{i, j_1, \ldots, j_k : (i,j_1), \ldots, (i,j_k) \in E\} \\ &= \{c \in \mathcal{C}(G) : i \in c\} \\ &= \bigcup_{\substack{c \in \mathcal{C}(G) \\ i \in c}} c \quad \blacksquare \end{aligned}$$

2. (7 points) In this question, we consider the Ising model on a 100-by-100 **periodic** lattice (i.e., the lattice size $N = 100$). We focus on the following 20 values of the parameter β

$$\beta \in \{0.2 + 0.02k \,|\, k = 1, 2, \ldots, 20\}\,. \tag{2.2}$$

Please do the following tasks using Gibbs sampling:

- For each of the β values in Eq. (2.2), generate the sequence

$$\left\{\boldsymbol{X}^{(i)} = \left(X_1^{(i)}, X_2^{(i)}, \ldots, X_{N^2}^{(i)}\right)\right\}_{i=0}^{1000}, \tag{2.3}$$

  which is the first 1000 (or 1001) components of an irreducible and aperiodic HMC with the $\pi_{N,\beta}$ in Eq. (1.3) as its stationary distribution.
- Plot $\boldsymbol{X}^{(1000)}$ for $\beta \in \{0.32, 0.4, 0.44, 0.6\}$. Each of your four plots should be similar to one of the panels in Figure 5.
- For each of the twenty β values in Eq. (2.2), compute

$$\widehat{v}(\beta) := \frac{1}{1000}\sum_{i=1}^{1000}\left|\frac{\sum_{k\in\Lambda_N} X_k^{(i)}}{N^2}\right|.$$

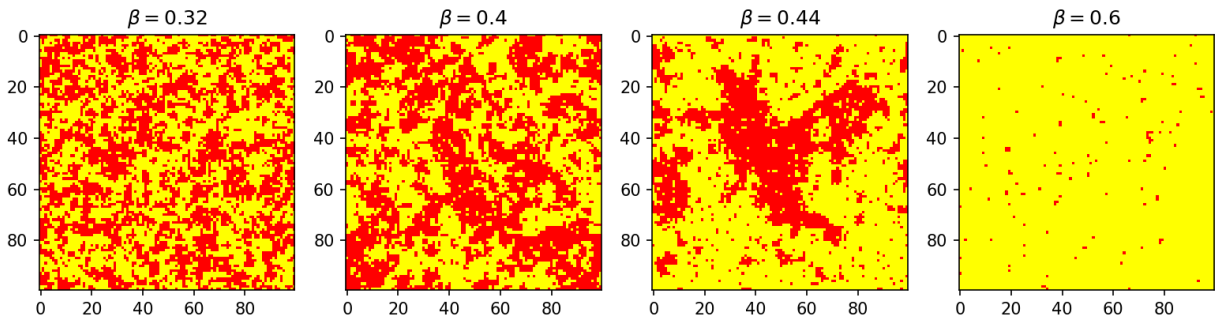
- List the following twenty values of $\widehat{v}(\beta)$

$$\left\{\widehat{v}(\beta)\,\middle|\,\beta = 0.2 + 0.02k \text{ for } k = 1, 2, \ldots, 20\right\}. \tag{2.4}$$

- Plot the twenty values in Eq. (2.4) in a '$\widehat{v}(\beta)$ vs. β' fashion, which should be similar to Figure 6.

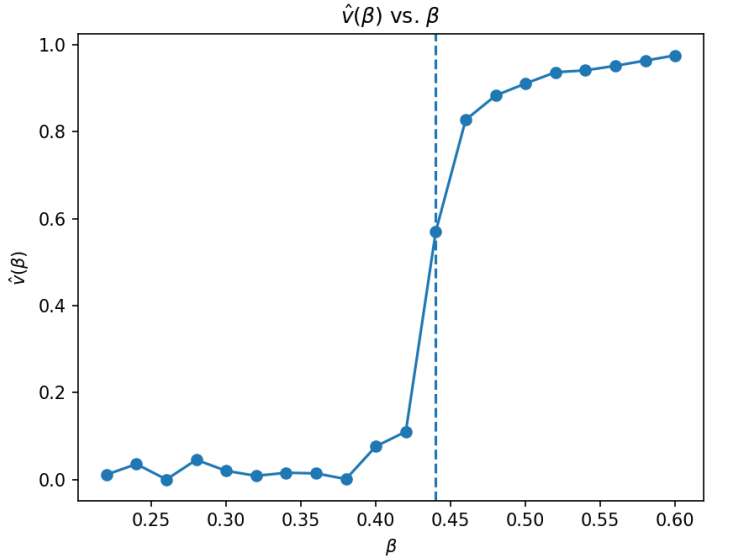
Please provide your code for completing each of the tasks.

Here are the plots for $\beta \in \{0.32, 0.4, 0.44, 0.6\}$:



The full list of values for $\widehat{v}(\beta)$ along with the plot of $\widehat{v}(\beta)$ vs. β are below:

β	$\widehat{v}(\beta)$
0.22	0.0116
0.24	0.0358
0.26	0.0004
0.28	0.0456
0.30	0.0202
0.32	0.0088
0.34	0.0157
0.36	0.0144
0.38	0.0014
0.40	0.0706
0.42	0.1104
0.44	0.5702
0.46	0.8272
0.48	0.8832
0.50	0.9108
0.52	0.9366
0.54	0.9410
0.56	0.9514
0.58	0.9634
0.60	0.9754



Finally, here is the full Python implementation of the model, plots, and estimator:

```python
import numpy as np
import matplotlib.pyplot as plt
from numba import njit

@njit
def bernoulli(p: float) -> int:
    if np.random.uniform() < p:
        return 1
    else:
        return 0

@njit
def modulo(coord : int) -> int:
    return coord % LATTICE_SIZE

@njit
def indexify(s : int) -> int:
    #Converts {-1, 1} to {0, 1} for Python indexing
    return (s+1)//2

@njit
def init_probability_matrix(beta : float) -> list[list[float]]:
    p = np.zeros((2, 2, 2, 2))
    for s1 in range(2):
        for s2 in range(2):
            for s3 in range(2):
                for s4 in range(2):
                    a = np.exp(beta * (2*(s1+s2+s3+s4)-4))
                    b = np.exp(-beta * (2*(s1+s2+s3+s4)-4))

                    p[s1, s2, s3, s4] = a/(a+b)
    return p
```

```python

@njit
def ising(n : int, beta : float, x0 : list[list[int]]) -> list[list[list[int]]]:
    #TRANSLATED AND MODIFIED FROM PROVIDED R CODE
    #Returns a list of n 2D arrays of size N x N whose entries are samples from $\pi_{N, \beta}$

    p = init_probability_matrix(beta)

    X = [x0]
    for k in range(n):
        Xi = X[k]
        for j in range(LATTICE_SIZE):
            jp1 = modulo(j+1)
            jm1 = modulo(j-1)

            for i in range(LATTICE_SIZE):
                ip1 = modulo(i+1)
                im1 = modulo(i-1)

                pij = p[indexify(Xi[ip1,j]),
                        indexify(Xi[im1,j]),
                        indexify(Xi[i,jp1]),
                        indexify(Xi[i,jm1])]

                Z = bernoulli(pij)
                Xi[i, j] = 2*Z - 1

        X.append(Xi)
    return X

```

```python
@njit
def estimator(beta: float, x0 : list[list[int]]) -> float:
    print("Estimating with beta =", beta)
    X = ising(CHAIN_LENGTH, beta, x0)

    total = 0
    for i in range(CHAIN_LENGTH):
        total += abs(np.sum(X[i]) / (LATTICE_SIZE**2))

    return total/CHAIN_LENGTH

def plot_X1000(x0 : list[list[int]]) -> None:
    ax1 = plt.subplot(1, 4, 1)
    ax1.imshow(ising(CHAIN_LENGTH, 0.32, x0)[-1], cmap='autumn')
    ax1.set_title(r'$\beta = 0.32$')
    print("Plotting beta = 0.32")

    ax2 = plt.subplot(1, 4, 2)
    ax2.imshow(ising(CHAIN_LENGTH, 0.4, x0)[-1], cmap='autumn')
    ax2.set_title(r'$\beta = 0.4$')
    print("Plotting beta = 0.45")

    ax3 = plt.subplot(1, 4, 3)
    ax3.imshow(ising(CHAIN_LENGTH, 0.44, x0)[-1], cmap='autumn')
    ax3.set_title(r'$\beta = 0.44$')
    print("Plotting beta = 0.44")

    ax4 = plt.subplot(1, 4, 4)
    ax4.imshow(ising(CHAIN_LENGTH, 0.6, x0)[-1], cmap='autumn')
    ax4.set_title(r'$\beta = 0.6$')
    print("Plotting beta = 0.6")
```

```python

def plot_estimates(x0):
    betas = np.linspace(0.22, 0.6, 20)
    estimates = [estimator(beta, x0) for beta in betas]

    print("\nEstimates:")
    for beta, estimate in zip(betas, estimates):
        print(f"Beta: {round(beta, 4)}, Estimator: {round(estimate, 4)}")

    plt.figure(2)
    plt.plot(betas, estimates, marker='o')
    plt.axvline(0.44, linestyle='--')
    plt.title(r'$\hat{v}(\beta)$' + " vs. " r'$\beta$')
    plt.xlabel(r'$\beta$')
    plt.ylabel(r'$\hat{v}(\beta)$')

LATTICE_SIZE = 100
CHAIN_LENGTH = 1000

x0 = np.full((LATTICE_SIZE, LATTICE_SIZE), -1)

plot_X1000(x0)
plot_estimates(x0)

plt.show()
```

References

APMA1690: Homework # 10 (Due by 11pm December 8)

"*Subtle is the Lord, but malicious he is not.*"

— Albert Einstein

1 Review

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

1.1 Manifold Hypothesis

Every science is based on one or several assumptions. Manifold learning (also known as 'dimension reduction') is not an exception — manifold[1] learning is based on the so-called 'manifold hypothesis:' *high dimensional data tend to lie in the vicinity of a low dimensional manifold* (Fefferman et al., 2016).[2] The 'high dimension' is usually referred to as the 'extrinsic dimension,' and the 'low dimension' is usually referred to as the 'intrinsic dimension.'

1.2 Notations

- d and D are two positive integers satisfying $d < D$. Hereafter, d denotes an intrinsic dimension, and D denotes an extrinsic dimension.
- For any matrix $\boldsymbol{A}$, its transpose is denoted as $\boldsymbol{A}^\intercal$.
- For any $\boldsymbol{x} = (x_1, \ldots, x_D)^\intercal \in \mathbb{R}^D$, its norm is defined as $\|\boldsymbol{x}\| := \sqrt{\sum_{k=1}^D x_k^2}$. Furthermore, if the column vector $\boldsymbol{x}$ is viewed as a D-by-1 matrix, we have

$$\|\boldsymbol{x}\|^2 = \boldsymbol{x}^\intercal \boldsymbol{x}. \tag{1.1}$$

- All vectors in this problem set are viewed as column vectors, which is also a convention widely adopted in the literature.
- $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ are identically and independently distributed (iid) $\mathbb{R}^D$-valued random variables (i.e., D-dimensional random vectors).

[1] The precise definition of manifolds is quite technical and beyond the scope of APMA 1690. You may just roughly think of a manifold as a surface/curve/hyperplane/line.

[2] The first author Charles Fefferman achieved a full professorship at the University of Chicago at the age of 22, making him the youngest full professor ever appointed in the United States. Fefferman entered the University of Maryland at age 14, graduated with degrees in mathematics and physics at 17, and earned his PhD in mathematics three years later from Princeton University.

1.3 A Glimpse of Manifold Learning

The manifold hypothesis described in Section 1.1 can be mathematically represented as follows:

- Latent d-dimensional random vectors $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n$ are generated by 'the Lord' in an iid way. They are not available to us.
- There exists a function $g : \mathbb{R}^d \to \mathbb{R}^D$. The image of g, i.e.,

$$\mathcal{M} := \left\{ g(z) \,:\, z \in \mathbb{R}^d \right\}, \tag{1.2}$$

  is called a manifold. Neither g nor $\mathcal{M}$ is available to us.
- The data points $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ that we observe in the high-dimensional space $\mathbb{R}^D$ are generated by the following mechanism

$$\boldsymbol{X}_i = g(\boldsymbol{Z}_i) + \boldsymbol{\varepsilon}_i, \qquad \text{for all } i = 1, 2, \ldots, n, \tag{1.3}$$

  where $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are iid D-dimensional random vectors playing the role of random noise.

The ultimate goal of manifold learning is to learn the low-dimensional manifold $\mathcal{M}$ defined in Eq. (1.2). Then, the observed high-dimensional data points $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ are reduced to their projections $\pi_{\mathcal{M}}(\boldsymbol{X}_1), \ldots, \pi_{\mathcal{M}}(\boldsymbol{X}_n)$ on the low-dimensional manifold $\mathcal{M}$. See Figure 1 for illustrations.

1.4 Two Branches of Manifold Learning

Dimension reduction/manifold learning refers to a collection of methods reducing the dimensionality of data while preserving most information in the data. There are the following two branches of dimension reduction:

- Linear dimension reduction, e.g., principal component analysis (PCA, Pearson, 1901).
- Nonlinear dimension reduction, e.g., principal curves (Hastie and Stuetzle, 1989), Isomap (Tenenbaum et al., 2000), local linear embedding (Roweis and Saul, 2000; Wu and Wu, 2018), Laplacian eigenmaps (Belkin and Niyogi, 2001), diffusion map (Coifman et al., 2005; Coifman and Lafon, 2006), principal manifolds (Smola et al., 2001; Meng and Eloyan, 2021).

While nonlinear dimension reduction is still developing, linear dimension reduction is relatively well-developed. We focus on the most widely used linear dimension reduction approach — PCA.

1.5 Mathematical Preparations

Materials in this subsection are from Sections 1.4 and 1.5 of Seber and Lee (2012).

For any random vector $\boldsymbol{X} = (X_1, \ldots, X_n)^\intercal$ with any dimension $n = 1, 2, \ldots$, we define its **covariance matrix** $\mathbb{V}(\boldsymbol{X})$ as follows (Seber and Lee, 2012, Definition 1.3)

$$\mathbb{V}(\boldsymbol{X}) := \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Cov}(X_n, X_n) \end{pmatrix} = \Big(\mathrm{Cov}(X_i, X_j) \Big)_{1 \le i, j \le n}, \tag{1.4}$$

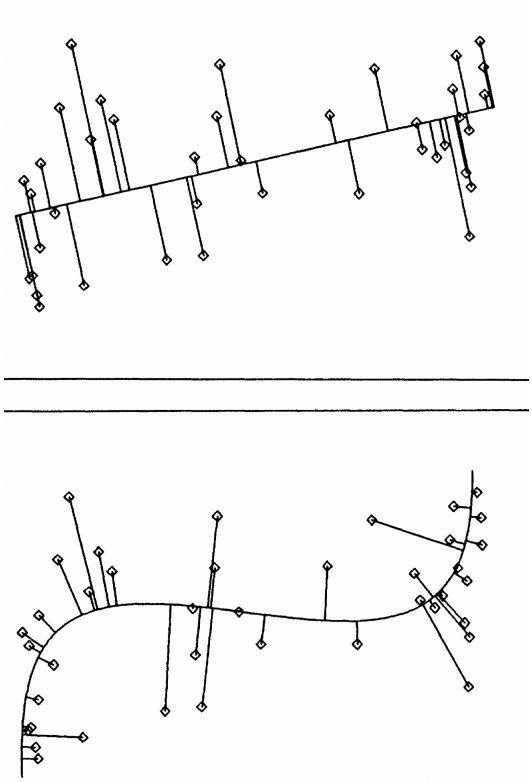


Figure 1: (2-dimensional) data points $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ and their projections $\pi_{\mathcal{M}}(\boldsymbol{X}_1), \ldots, \pi_{\mathcal{M}}(\boldsymbol{X}_n)$ on (1-dimensional) manifolds. The upper panel illustrates a linear manifold learning result, and the lower panel illustrates a nonlinear manifold learning result. This figure comes from Hastie and Stuetzle (1989).

which is an n-by-n matrix. Obviously, $\mathbb{V}(\boldsymbol{X})$ is a symmetric matrix. Furthermore, we have the following claims

Claim 1.1. *The covariance matrix* $\mathbb{V}(\boldsymbol{X})$ *defined in Eq.* (1.4) *is positive semi-definite* .

Proof: The proof is a homework problem.
One may apply the following claim to prove Claim 1.1.

Claim 1.2 (Theorem 1.3 of Seber and Lee (2012))**.** *Let* $\boldsymbol{A}$ *by any* m*-by-*n *deterministic matrix. Then, we have*

$$\mathbb{V}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\mathbb{V}(\boldsymbol{X})\boldsymbol{A}^\intercal.$$

Claim 1.3 (Theorem 1.5 of Seber and Lee (2012))**.** *Let* $\boldsymbol{A}$ *be any* n*-by-*n *symmetric matrix. Suppose* $\boldsymbol{\Sigma} := \mathbb{V}(\boldsymbol{X})$ *and* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\intercal = \mathbb{E}\boldsymbol{X}$*. Then, we have*

$$\mathbb{E}(\boldsymbol{X}^\intercal\boldsymbol{A}\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\intercal\boldsymbol{A}\boldsymbol{\mu}, \tag{1.5}$$

where $\mathrm{tr}(\cdot)$ *denotes the trace of a square matrix.*

1.6 Principal Component Analysis

Without loss of generality, we assume that the data points $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ have been centralized. That is, hereafter, we are under the following assumption

Assumption 1. *Data points* $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ *satisfy*

$$\mathbb{E}\boldsymbol{X}_i = (0, 0, \ldots, 0)^\intercal =: \boldsymbol{0}, \quad \text{for all } i = 1, 2, \ldots, n. \tag{1.6}$$

Assumption 1 is widely adopted in the PCA literature. It can easily be satisfied as $\mathbb{E}(\boldsymbol{X}_i - \frac{1}{n}\sum_{j=1}^n \boldsymbol{X}_j) = \boldsymbol{0}$, and we can always update $\boldsymbol{X}_i$ by $\boldsymbol{X}_i \leftarrow \boldsymbol{X}_i - \frac{1}{n}\sum_{j=1}^n \boldsymbol{X}_j$.

The PCA framework assumes that the $\mathcal{M}$ defined in Eq. (1.2) is a d-dimensional hyperplane embedded in $\mathbb{R}^D$ and going through the origin of $\mathbb{R}^D$.[3] Hereafter, we use V to denote d-dimensional hyperplanes.

1.6.1 PCA from the Viewpoint of Fitting Errors

Let V be a d-dimensional hyperplane embedded in $\mathbb{R}^D$ and going through the origin of $\mathbb{R}^D$. For any $\boldsymbol{x} \in \mathbb{R}^D$, its projection on the hyperplane V is denoted as $\boldsymbol{P}_V\boldsymbol{x}$.

PCA is the model claiming that the underlying manifold (see Eq. (1.2)) is the following hyperplane V^* that minimizes the fitting error $\mathbb{E}\left(\|\boldsymbol{X} - \boldsymbol{P}_V\boldsymbol{X}\|^2\right)$

$$V^* = \operatorname*{argmin}_V \left\{\mathbb{E}\left(\|\boldsymbol{X} - \boldsymbol{P}_V\boldsymbol{X}\|^2\right) : V \text{ is a } d\text{-dimensional hyperplane going through the origin of } \mathbb{R}^D\right\}. \tag{1.7}$$

Hereafter, all minimizations/minimizations are taken across all d-dimensional hyperplanes V going through the origin of $\mathbb{R}^D$.

1.6.2 PCA from the Viewpoint of Variance

By the Pythagorean theorem, we have $\|\boldsymbol{X} - \boldsymbol{P}_V\boldsymbol{X}\|^2 = \|\boldsymbol{X}\|^2 - \|\boldsymbol{P}_V\boldsymbol{X}\|^2$. Therefore,

$$\min_V \left\{\mathbb{E}\left(\|\boldsymbol{X} - \boldsymbol{P}_V\boldsymbol{X}\|^2\right)\right\} = \mathbb{E}\left(\|\boldsymbol{X}\|^2\right) - \max_V \left\{\mathbb{E}\left(\|\boldsymbol{P}_V\boldsymbol{X}\|^2\right)\right\}.$$

Therefore, we have the following representations

$$\boxed{V^* = \operatorname*{argmin}_V \left\{\mathbb{E}\left(\|\boldsymbol{X} - \boldsymbol{P}_V\boldsymbol{X}\|^2\right)\right\} = \operatorname*{argmax}_V \left\{\mathbb{E}\left(\|\boldsymbol{P}_V\boldsymbol{X}\|^2\right)\right\}.} \tag{1.8}$$

Eq. (1.8) provides the following two interpretations of the optimal hyperplane V^*:

- V^* is the hyperplane that minimizes the average fitting error $\mathbb{E}\left(\|\boldsymbol{X} - \boldsymbol{P}_V\boldsymbol{X}\|^2\right)$.
- V^* is the hyperplane that makes the projection $\boldsymbol{P}_{V^*}\boldsymbol{X}$ enjoy the largest variance.

Both interpretations indicate that the optimal hyperplane V^* goes through the 'middle' of data points (see the upper panel of Figure 1 for an illustration.).

[3] The 'going through the origin of $\mathbb{R}^D$' corresponds to Assumption 1.

1.6.3 Formulae of V^* and $\boldsymbol{P}_{V^*}$

Since the covariance matrix $\mathbb{V}(\boldsymbol{X})$ is positive semi-definite (see Claim 1.1), we have the following eigen-structure of $\mathbb{V}(\boldsymbol{X})$:

- eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq \ldots \geq \lambda_D \geq 0$;
- the corresponding eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_D$ satsifying $\mathbb{V}(\boldsymbol{X})\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$ for all $i = 1, 2, \ldots, D$ (the eigenvectors are viewed as column vectors).
- The eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_D$ are orthonormal, i.e.,

$$\boldsymbol{v}_i^\intercal \boldsymbol{v}_j = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases} \tag{1.9}$$

Then, we have the following formula of the optimal hyperplane V^* defined in Eq. (1.8)

$$\boxed{V^* = \operatorname{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d\} = \left\{ \sum_{k=1}^{d} \alpha_k \boldsymbol{v}_k \,:\, \alpha_1, \ldots, \alpha_d \in \mathbb{R} \right\},}$$

where span denotes the 'linear span.' We define a D-by-d matrix $\boldsymbol{W}$ as follows

$$\boldsymbol{W} := (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d),$$

i.e., the column vector $\boldsymbol{v}_i$ is the i-th column of the matrix $\boldsymbol{W}$. Then, the projection $\boldsymbol{P}_{V^*}$ can be represented as the following matrix

$$\boxed{\boldsymbol{P}_{V^*} = \boldsymbol{W}\boldsymbol{W}^\intercal,} \tag{1.10}$$

i.e., $\boldsymbol{P}_{V^*}\boldsymbol{X} = \boldsymbol{W}\boldsymbol{W}^\intercal\boldsymbol{X}$.

1.6.4 Choice of the Intrinsic Dimension d

Claim 1.4. *Let $\boldsymbol{X} = (X_1, \ldots, X_D)^\intercal$ be a D-dimensional random vector satisfying $\mathbb{E}X_i = 0$ for all $i = 1, 2, \ldots, D$. The projection matrix $\boldsymbol{P}_{V^*}$ is defined by Eq. (1.10). Then, we have*

1. $\mathbb{E}\left(\|\boldsymbol{X}\|^2\right) = \sum_{k=1}^{D} \lambda_k$.
2. $\mathbb{E}\left(\|\boldsymbol{P}_{V^*}\boldsymbol{X}\|^2\right) = \sum_{k=1}^{d} \lambda_k$.

Proof: The proof is a homework problem.

The following ratio is interpreted as 'the proportion of variance explained by the PCA.'

$$\boxed{r_d := \frac{\mathbb{E}\left(\|\boldsymbol{P}_V\boldsymbol{X}\|^2\right)}{\mathbb{E}\left(\|\boldsymbol{X}\|^2\right)} = \frac{\sum_{k=1}^{d}\lambda_k}{\sum_{k=1}^{D}\lambda_k}.}$$

This ratio r_d provides criteria for choosing the intrinsic dimension d in applications. A widely adopted criterion for choosing d is the following:[4]

- $r_d > 95\%$,
- and $r_{d-1} < 95\%$.

[4]The percentage 95% can be replaced with any other percentage you like.

1.6.5 Further Interpretation of r_d

Suppose the data $\boldsymbol{X}$ is generated by the following mechanism (which is a special case of Eq. (1.3)):

$$\boldsymbol{X} = \boldsymbol{L}\boldsymbol{Z} + \boldsymbol{\varepsilon},$$

- $\boldsymbol{Z}$ is a latent d-dimensional random vector, unavailable (available to 'the Lord' rather than us);
- $\boldsymbol{L}$ is a deterministic D-by-d matrix, unavailable to us;
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_D)^\intercal$ is a D-dimensional random vector, playing the role of random noise; we assume $\varepsilon_1, \ldots, \varepsilon_D \overset{iid}{\sim} N(0, \sigma^2)$, which implies

$$\mathbb{V}(\boldsymbol{\varepsilon}) = \begin{pmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{pmatrix}, \tag{1.11}$$

  which is a D-by-D diagonal matrix;
- The latent random variable $\boldsymbol{Z}$ and noise $\boldsymbol{\varepsilon}$ are independent;
- The D-dimensional random vector $\boldsymbol{X}$ represents the data we observe.

Claim 1.2, together with the independence between $\boldsymbol{Z}$ and $\boldsymbol{\varepsilon}$, implies

$$\mathbb{V}(\boldsymbol{X}) = \boldsymbol{L}\mathbb{V}(\boldsymbol{Z})\boldsymbol{L}^\intercal + \mathbb{V}(\boldsymbol{\varepsilon}). \tag{1.12}$$

We have the following matrix rank estimation

$$\operatorname{rank}(\boldsymbol{L}\mathbb{V}(\boldsymbol{Z})\boldsymbol{L}^\intercal) \le \operatorname{rank}(\mathbb{V}(\boldsymbol{Z})) \le d,$$

where the last inequality comes from that $\mathbb{V}(\boldsymbol{Z})$ is a d-by-d matrix. Then, the matrix $\boldsymbol{L}\mathbb{V}(\boldsymbol{Z})\boldsymbol{L}^\intercal$ has at most d non-zero eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_d > 0$. Specifically, we have the following eigen decomposition

$$\boldsymbol{L}\mathbb{V}(\boldsymbol{Z})\boldsymbol{L}^\intercal = \boldsymbol{U} \begin{pmatrix} \tilde{\lambda}_1 & & & & & \\ & \ddots & & & & \\ & & \tilde{\lambda}_d & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} \boldsymbol{U}^\intercal, \tag{1.13}$$

where $\boldsymbol{U}$ is an orthogonal matrix (i.e., $\boldsymbol{U}\boldsymbol{U}^\intercal = \boldsymbol{U}^\intercal\boldsymbol{U} = \boldsymbol{I}_D$ = the D-by-D identity matrix). Combining Eq. (1.11), (1.12), and (1.13), we have

$$\mathbb{V}(\boldsymbol{X}) = \boldsymbol{U} \begin{pmatrix} \tilde{\lambda}_1 + \sigma^2 & & & & & \\ & \ddots & & & & \\ & & \tilde{\lambda}_d + \sigma^2 & & & \\ & & & \sigma^2 & & \\ & & & & \ddots & \\ & & & & & \sigma^2 \end{pmatrix} \boldsymbol{U}^\intercal. \tag{1.14}$$

Therefore, the eigenvalues $\{\lambda_k\}_{k=1}^{D}$ of $\mathbb{V}(\boldsymbol{X})$ are the following

$$\begin{aligned}
&\lambda_1 = \tilde{\lambda}_1 + \sigma^2,\\
&\vdots\\
&\lambda_d = \tilde{\lambda}_d + \sigma^2,\\
&\lambda_{d+1} = \sigma^2,\\
&\vdots\\
&\lambda_D = \sigma^2.
\end{aligned} \tag{1.15}$$

If noise is very small (i.e., $\sigma^2 \approx 0$), we have

$$\lambda_1 \geq \ldots \geq \lambda_d \geq \lambda_{d+1} \approx \lambda_{d+2} \approx \ldots \approx \lambda_D \approx 0.$$

Furthermore, we have

$$r_d = \frac{\sum_{k=1}^{d} \lambda_k}{\sum_{k=1}^{D} \lambda_k} = \frac{\sum_{k=1}^{d} \tilde{\lambda}_k + d \cdot \sigma^2}{\sum_{k=1}^{d} \tilde{\lambda}_k + D \cdot \sigma^2}. \tag{1.16}$$

If the noise is very small (i.e., $\sigma^2 \approx 0$), Eq. (1.16) implies

$$1 \approx r_D \approx r_{D-1} \approx \ldots \approx r_{d+1} \approx r_d.$$

1.6.6 Computation in Applications

The only input for PCA is the covariance matrix $\mathbb{V}(\boldsymbol{X})$. However, the precise covariance matrix $\mathbb{V}(\boldsymbol{X})$ is unavailable in applications. In practical applications, when working with observed data points $\{\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,D})^\intercal\}_{i=1}^{n} \subseteq \mathbb{R}^D$, the following sample covariance matrix $\widehat{\mathbb{V}(\boldsymbol{X})}$ is a substitute for the precise covariance matrix (thanks to the Law of Large Numbers)

$$\begin{aligned}
\widehat{\mathbb{V}(\boldsymbol{X})} &:= \begin{pmatrix}
\widehat{\mathrm{Cov}}(X_1, X_1) & \widehat{\mathrm{Cov}}(X_1, X_2) & \cdots & \widehat{\mathrm{Cov}}(X_1, X_D)\\
\widehat{\mathrm{Cov}}(X_2, X_1) & \widehat{\mathrm{Cov}}(X_2, X_2) & \cdots & \widehat{\mathrm{Cov}}(X_2, X_D)\\
\vdots & \vdots & \ddots & \vdots\\
\widehat{\mathrm{Cov}}(X_D, X_1) & \widehat{\mathrm{Cov}}(X_D, X_2) & \cdots & \widehat{\mathrm{Cov}}(X_D, X_D)
\end{pmatrix},\\
\text{where } \widehat{\mathrm{Cov}}(X_i, X_j) &:= \frac{1}{n-1} \sum_{w=1}^{n} \left[\left(x_{i,w} - \frac{1}{n} \sum_{k=1}^{n} x_{i,k} \right) \left(x_{jw} - \frac{1}{n} \sum_{l=1}^{n} x_{j,l} \right) \right].
\end{aligned} \tag{1.17}$$

1.6.7 A Numerical Example

In this subsection, we provide a naive numerical example illustrating the aforementioned theoretical discussions.

- $Z \sim N(0, 1)$;
- $\boldsymbol{L} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$;

- $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$, where $\varepsilon_1, \varepsilon_2, \varepsilon_3 \overset{iid}{\sim} N(0,\, 0.04)$;
- $\boldsymbol{X} \overset{\text{def}}{=} \boldsymbol{L}Z + \boldsymbol{\varepsilon} = \begin{pmatrix} Z + \varepsilon_1 \\ Z + \varepsilon_2 \\ Z + \varepsilon_3 \end{pmatrix}$.

Then, we have

- $\mathbb{V}(\boldsymbol{L}Z) = \boldsymbol{L}\boldsymbol{L}^\intercal = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$, whose eigenvalues are $\tilde{\lambda}_1 = 3, \tilde{\lambda}_2 = \tilde{\lambda}_3 = 0$ (you have learned how to compute eigenvalues in your linear algebra class);
- Eq. (1.15) implies that the eigenvalues of $\mathbb{V}(\boldsymbol{X})$ are

$$\begin{aligned} \lambda_1 &= 3 + 0.04 = 3.04, \\ \lambda_2 &= \lambda_3 = 0.04. \end{aligned}$$

The results of a numerical experiment conducted using R are presented as follows, and they are compatible with our theoretical discussion.

```
> n=10000000
> sigma=0.2
>
> Z=rnorm(n)
> L=matrix(1, nrow = 3, ncol = 1)
> e=cbind(rnorm(n, sd=sigma), rnorm(n, sd=sigma), rnorm(n, sd=sigma))
> X=t(L%*%Z)+e
>
> covariance_matrix=var(X)
> eigenstructure=eigen(covariance_matrix)
> eigenstructure
eigen() decomposition
$values
[1] 3.04140759 0.04002077 0.03997760

$vectors
           [,1]       [,2]       [,3]
[1,] 0.5773425  0.1238912  0.8070481
[2,] 0.5772954 -0.7609286 -0.2961717
[3,] 0.5774129  0.6368977 -0.5108382
```

2 Problem Set

1. (2 points) Prove Claim 1.1.

   Claim: The covariance matrix

   $$\mathbb{V}(X) = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Cov}(X_n, X_n) \end{pmatrix} = \Big(\mathrm{Cov}(X_i, X_j)\Big)_{1 \le i,j \le n}$$

   is positive semi-definite.

   Proof: By definition, a matrix M is positive semi-definite if it is symmetric and if $z^T M z$ is nonnegative for every nonzero real column vector z.

   Since

   $$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[(Y - \mathbb{E}[Y])(X - \mathbb{E}[X])] = \mathrm{Cov}(Y, X)$$

   the covariance matrix is symmetric.

   Then let z be any nonzero real n-by-1 column vector. Then by Claim 1.2

   $$z^T \mathbb{V}(X) z = \mathbb{V}(z^T X)$$

   but since $z^T X$ is a scalar,

   $$\mathbb{V}(z^T X) = \mathrm{Cov}(z^T X, z^T x) = \mathrm{Var}(z^T X) \ge 0$$

   so $z^T \mathbb{V}(X) z \ge 0$ and we are done. ■

2. (3 points) Prove Claim 1.4. (Hint: apply Eq. (1.1), Claim 1.3, properties of trace, Eq. (1.9), and Eq. (1.10).)

   Claim: Let $\boldsymbol{X} = (X_1, \ldots, X_D)^\intercal$ be a D-dimensional random vector satisfying $\mathbb{E}X_i = 0$ for all $i = 1, 2, \ldots, D$. The projection matrix $\boldsymbol{P}_{V^*}$ is defined by Eq. (1.10). Then, we have

   (a) $\mathbb{E}\left(\|\boldsymbol{X}\|^2\right) = \sum_{k=1}^{D} \lambda_k$.

   (b) $\mathbb{E}\left(\|\boldsymbol{P}_{V^*}\boldsymbol{X}\|^2\right) = \sum_{k=1}^{d} \lambda_k$.

   Proof:

   By Eq. (1.1), if $\vec{x}$ is a D-by-1 column vector, we have

   $$\|\boldsymbol{x}\|^2 = \boldsymbol{x}^\intercal \boldsymbol{x}.$$

   so

   $$\mathbb{E}[||X||^2] = \mathbb{E}[X^T X]$$

   Claim 1.3 says, for any n-by-n symmetric matrix $\boldsymbol{A}$, $\boldsymbol{\Sigma} := \mathbb{V}(\boldsymbol{X})$, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\intercal = \mathbb{E}\boldsymbol{X}$, we have

   $$\mathbb{E}(\boldsymbol{X}^\intercal \boldsymbol{A} \boldsymbol{X}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\intercal \boldsymbol{A} \boldsymbol{\mu}, \tag{2.1}$$

Thus, we can write $X^TX = X^TIX$ and

$$\mathbb{E}[X^TX] = \mathbb{E}[X^TIX] = \text{tr}(I\Sigma) + \mu^TI\mu = \text{tr}(\Sigma) + \mu^T\mu = \text{tr}(\mathbb{V}(X)) + (\mathbb{E}X)^T(\mathbb{E}X)$$

By definition of trace,

$$\text{tr}(\mathbb{V}(X)) = \sum_{i=1}^{D} \lambda_i$$

and since $\mathbb{E}X = 0$,

$$(\mathbb{E}X)^T(\mathbb{E}X) = 0$$

so

$$\mathbb{E}\left(\|\boldsymbol{X}\|^2\right) = \sum_{i=1}^{D} \lambda_i$$

and we are done. □

Now for the second part, using Eq. (1.10),

$$\begin{aligned}\mathbb{E}[||P_V^*X||^2] &= \mathbb{E}[\left|\left|WW^TX\right|\right|^2] \\ &= \mathbb{E}[(WW^TX)^T(WW^TX)] \\ &= \mathbb{E}[X^TWW^TWW^TX]\end{aligned}$$

where W is a D-by-d matrix defined by $W = (v_1, \ldots, v_d)$.

Thus WW^T is a D-by-D matrix and clearly, it is symmetric:

$$(WW^T)^T = (W^T)^TW^T = WW^T$$

so we can apply claim 1.3 to get

$$\mathbb{E}[X^TWW^TWW^TX] = \text{tr}(WW^TWW^T\mathbb{V}(X)) + (\mathbb{E}X)^T(WW^T)^2(\mathbb{E}X) = \text{tr}(WW^TWW^T\mathbb{V}(X))$$

Then using the cyclic property of the trace, we have

$$\text{tr}(WW^TWW^T\mathbb{V}(X)) = \text{tr}(WW^T\mathbb{V}(X)WW^T)$$

We can calculate WW^T as:

$$\begin{aligned}WW^T &= \begin{pmatrix}\vec{v}_1 & \vec{v}_2 & \ldots & \vec{v}_d\end{pmatrix}\begin{pmatrix}\vec{v}_1^T \\ \vec{v}_2^T \\ \vdots \\ \vec{v}_d^T\end{pmatrix} \\ &= v_1v_1^T + v_2v_2^T + \cdots + v_dv_d^T \\ &= \sum_{i=1}^{d} v_iv_i^T\end{aligned}$$

And since $\mathbb{V}(X)$ is symmetric, we can write

$$\mathbb{V}(X) = Q\Lambda Q^T = \sum_{i=1}^{D} \lambda_i v_i v_i^T$$

where $Q = (v_1, v_2, \ldots, v_D)$ and Λ is a diagonal matrix of the eigenvalues of $\mathbb{V}(X)$.

So

$$\begin{aligned}
\text{tr}(WW^T\mathbb{V}(X)WW^T) &= \text{tr}(\left(\sum_{i=1}^{d} v_i v_i^T\right)\left(\sum_{j=1}^{D} \lambda_j v_j v_j^T\right)\left(\sum_{k=1}^{d} v_k v_k^T\right)) \\
&= \text{tr}(\left(\sum_{i=1}^{d}\sum_{j=1}^{D}\sum_{k=1}^{d} v_i v_i^T \lambda_j v_j v_j^T v_k v_k^T\right)) \\
&= \text{tr}(\left(\sum_{i=1}^{d}\sum_{j=1}^{D}\sum_{k=1}^{d} \lambda_j v_i v_i^T v_j v_j^T v_k v_k^T\right))
\end{aligned}$$

But since the eigenvectors $v_1, \ldots, v_D$ are orthonormal:

$$v_i^T v_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

All of these terms go to 0 unless $i = j = k$. However, when $d \leq j \leq D$, i and k are at most d so the only terms that survive are when $i = j = k$ and $1 \leq j \leq d$. Thus, we have

$$\text{tr}(WW^T\mathbb{V}(X)WW^T) = \begin{pmatrix} \lambda_1 & & & & & & \\ & \lambda_2 & & & & & \\ & & \ddots & & & & \\ & & & \lambda_d & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{pmatrix}$$

So

$$\mathbb{E}[||P_V^* X||^2] = \text{tr}(WW^T\mathbb{V}(X)WW^T) = \sum_{i=1}^{d} \lambda_i \quad \blacksquare$$

3. (5 points) Please conduct the following procedures:

   (a) Set $n = 100$.

   (b) Generate n random numbers $z_1, z_2, \ldots, z_n \overset{iid}{\sim} N(0, 1)$.

(c) Generate 2-dimensional random vectors $\boldsymbol{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \end{pmatrix}$ for all $i = 1, 2, \ldots, n$, where

$$\varepsilon_{1,1}, \varepsilon_{2,1}, \ldots, \varepsilon_{n,1}, \varepsilon_{1,2}, \varepsilon_{2,2}, \ldots, \varepsilon_{n,2} \overset{iid}{\sim} N(0, 0.04).$$

(d) Compute $\boldsymbol{x}_i = \begin{pmatrix} x_{i,1} \\ x_{i,2} \end{pmatrix} \overset{\text{def}}{=} \begin{pmatrix} z_i + \varepsilon_{i,1} \\ z_i + \varepsilon_{i,2} \end{pmatrix}$ for all $i = 1, 2, \ldots, n$.

(e) Compute the sample covariance matrix $\widehat{\mathbb{V}(\boldsymbol{X})}$ (see Eq. (1.17)) using $\{\boldsymbol{x}_i\}_{i=1}^n$. Present the matrix $\widehat{\mathbb{V}(\boldsymbol{X})}$.

(f) Compute eigenvalues λ_1, λ_2 and eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2$ of $\widehat{\mathbb{V}(\boldsymbol{X})}$, making the eigenvalues and eigenvectors satisfy the following

- $\lambda_1 \geq \lambda_2$,
- $\widehat{\mathbb{V}(\boldsymbol{X})}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$ for $i = 1, 2$, and
- eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2$ are orthonormal (see Eq. (1.9)).

Present the eigenvalues and eigenvectors. (Small numerical errors are allowed.)

(g) Compute the matrix

$$\boldsymbol{P} = \boldsymbol{v}_1 \boldsymbol{v}_1^\intercal, \tag{2.2}$$

which is a 2-by-2 matrix. Present this matrix.

(h) Plot the 2-dimensional data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$.

(i) Plot the following straight line

$$V^* = \{\alpha \boldsymbol{v}_1 \,:\, \alpha \in \mathbb{R}\}. \tag{2.3}$$

(j) Plot the 2-dimensional points $\boldsymbol{P}\boldsymbol{x}_1, \boldsymbol{P}\boldsymbol{x}_2, \ldots, \boldsymbol{P}\boldsymbol{x}_n$, where $\boldsymbol{P}$ is defined in Eq. (2.2).

(k) Plot the straight line segments[5] $(\boldsymbol{x}_1, \boldsymbol{P}\boldsymbol{x}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{P}\boldsymbol{x}_n)$.

Overlay all the plots. If you conduct all the procedures correctly, the plot you get should look like Figure 2. Please provide the code (in any programming language that you are comfortable with) for conducting the procedures.

[5] $(\boldsymbol{a}, \boldsymbol{b})$ denotes the straight line segment connecting points $\boldsymbol{a}$ and $\boldsymbol{b}$.

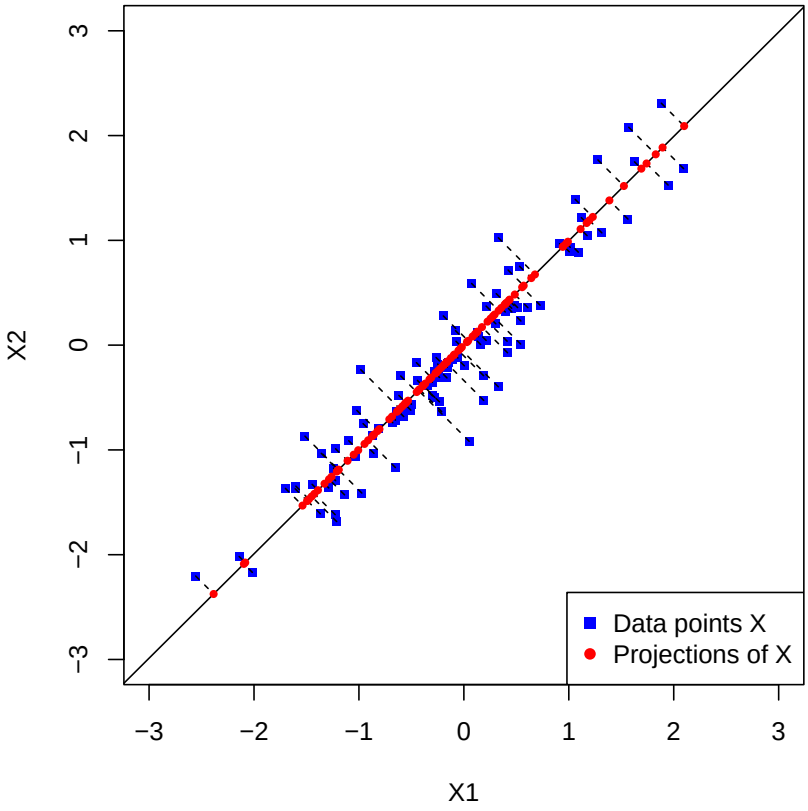
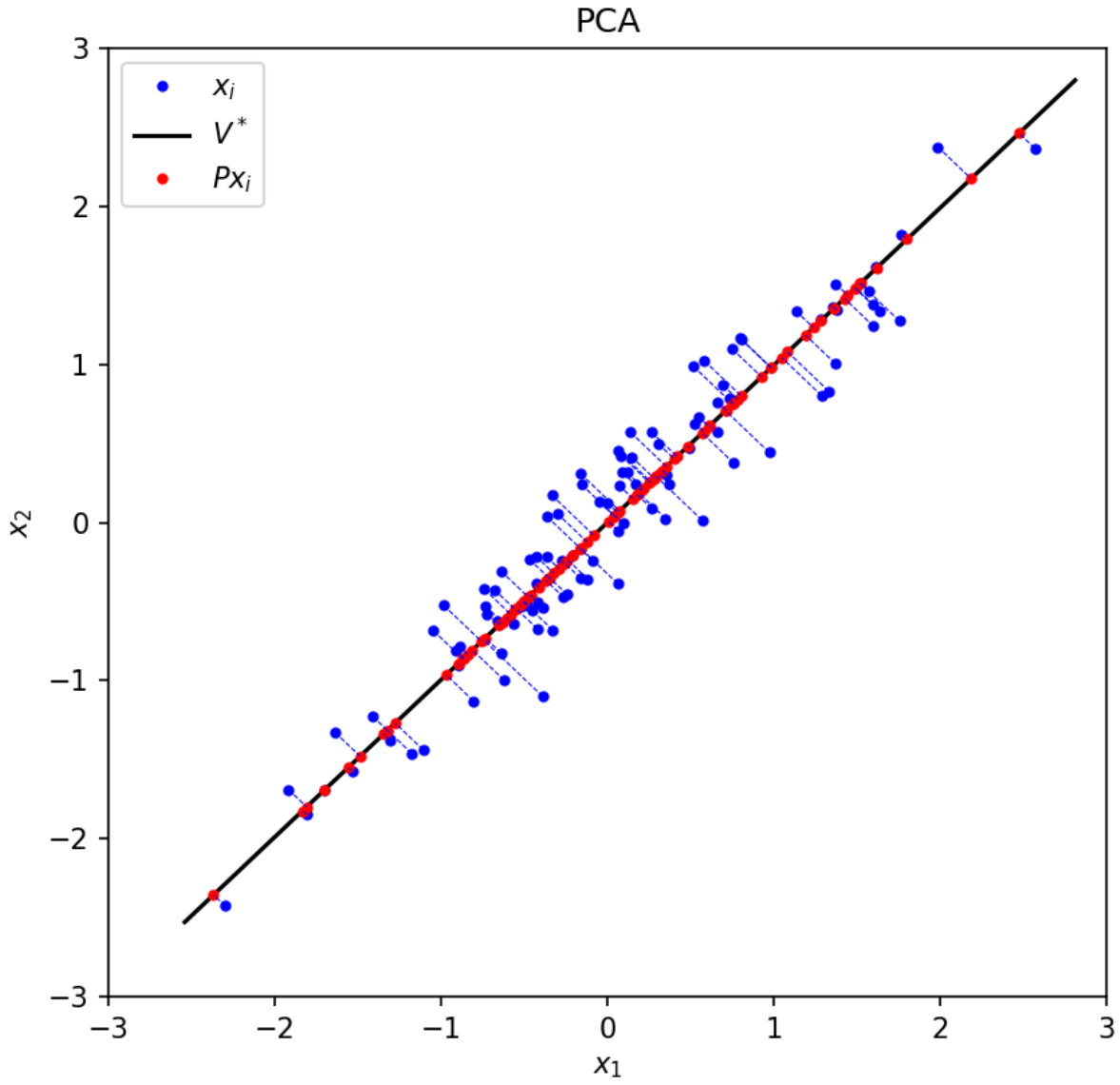


Figure 2: The blue squares present the data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, and the red dots present their projections $\boldsymbol{P}\boldsymbol{x}_1, \boldsymbol{P}\boldsymbol{x}_2, \ldots, \boldsymbol{P}\boldsymbol{x}_n$ on V^* (see Eq. (2.3)).

Here is the final plot:



Here are the numerical results

```
Covariance Matrix:
 [[0.94981143 0.93775809]
 [0.93775809 1.02122668]]

Eigenvalues:
 [0.04708138 1.92395673]

Eigenvectors:
 [[-0.72043392  0.69352358]
 [ 0.69352358  0.72043392]]

(Normality)
 v_1 norm:  1.0
 v_2 norm:  1.0

(Orthogonality)
 v_1 dot v_2:  0.0

P:
 [[0.48097496 0.49963792]
 [0.49963792 0.51902504]]
```

And here is the code I used to generate it. Note that I needed to generate the epsilon vector using the multivariate normal distribution to get a graph similar to the provided one. Generating 200 iid normal random variables and then reshaping them into a 100-by-2 matrix did work when $\varepsilon_i \sim N(0, 0.4)$ so I hypothesize the problem is simply one of scaling to be visible on the graph.

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

n = 100
D = 2
marker_size = 3

#Plots the data points
z = norm(0,1).rvs(size = n)
#epsilon = np.reshape(norm(0,0.04).rvs(size = D*n), (n,D))
epsilon = multivariate_normal([0,0], [[0.04, 0], [0, 0.04]]).rvs(size = n)

x = np.array([z, z]).T + epsilon

plt.plot(x[:,0], x[:,1], "o", markersize=marker_size, color='blue', label=r'$x_i$')


#Calculates the covariance matrix of the data points
cov = np.zeros((D, D))
for i in range(D):
    for j in range(D):
        cov[i, j] = np.sum((x[:,i] - np.mean(x[:,i])) * (x[:,j] - np.mean(x[:,j]))) / (n - 1)

#Calculate eigenvalues and eigenvectors of the covariance matrix
eig_vals, eig_vecs = np.linalg.eigh(cov)

eig_pairs = [(eig_vals[i], eig_vecs[:,i]) for i in range(len(eig_vals))]
eig_pairs.sort(key = lambda x: x[0], reverse = True)

lambda_1 = eig_pairs[0][0]
v_1 = np.array(eig_pairs[0][1]).reshape(D,1)

lambda_2 = eig_pairs[1]
v_2 = np.array(eig_pairs[1][1]).reshape(D,1)


#Calculate P matrix
P = v_1 * v_1.T

```

```
41
42  #Plotting V_star
43  slope = v_1[1]/v_1[0]
44  x_vals = np.array(plt.gca().get_xlim())
45  plt.plot(x_vals, (slope * x_vals), color='black', label=r'$V^*$')
46
47  #Plotting the projection of the data points onto V_star
48  proj = np.array([np.matmul(P, x[i]) for i in range(n)])
49
50  plt.plot([proj[i, 0] for i in range(n)], [proj[i, 1] for i in range(n)],
51          "o", markersize=marker_size, color='red', label=r'$Px_i$')
52
53  #Plots the straight lines from the data points to their projections
54  for i in range(n):
55      plt.plot([x[i,0], proj[i, 0]], [x[i,1], proj[i,1]],
56              color='blue', linestyle='dashed', linewidth=0.5)
57
58
59  #Final plots
60  plt.xlabel(r'$x_1$')
61  plt.xlim(-3,3)
62  plt.ylim(-3,3)
63  plt.ylabel(r'$x_2$')
64  plt.title('PCA')
65  plt.legend()
66  plt.show()
67
68
69  #Final Prints
70  print("Covariance Matrix: \n", cov, "\n")
71  print("Eigenvalues: \n", eig_vals, "\n")
72  print("Eigenvectors: \n", eig_vecs, "\n")
73  print("(Normality)\n v_1 norm: ", np.linalg.norm(v_1), "\n", "v_2 norm: ", np.linalg.norm(v_2), "\n")
74  print("(Orthogonality)\n v_1 dot v_2: ", np.dot(np.reshape(v_1, 2), np.reshape(v_2, 2)), "\n")
75  print("P: \n", P, "\n")
76
```

References

M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, 14, 2001.

R. R. Coifman and S. Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1): 5–30, 2006.

R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the national academy of sciences*, 102(21):7426–7431, 2005.

C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84 (406):502–516, 1989.

K. Meng and A. Eloyan. Principal manifold estimation via model complexity selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2):369–394, 2021.

K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

G. A. Seber and A. J. Lee. *Linear regression analysis*. John Wiley & Sons, 2012.

A. Smola, S. Mika, B. Schölkopf, R. Williamson, et al. Regularized principal manifolds. 2001.

J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

H.-T. Wu and N. Wu. Think globally, fit locally under the manifold setup: Asymptotic analysis of locally linear embedding. 2018.