Homework 1

Gaurav Batra
gbatra3@wisc.edu

Instructions: This is a background self-test on the type of math we will encounter in class. If you find many questions intimidating, we suggest you drop 760 and take it again in the future when you are more prepared. You can use this latex file as a template to develop your homework. Submit your homework on time as a single pdf file to **Gradescope**. There is no need to submit the latex source or any code.

1 Vectors and Matrices [6 pts]

Consider the matrix X and the vectors $\mathbf{y}$ and $\mathbf{z}$ below:

$$X = \begin{pmatrix} 3 & 2 \\ -7 & -5 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \qquad \mathbf{z} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

1. Compute $\mathbf{y}^T X \mathbf{z}$

   Step 1: Compute $X\mathbf{z}$
   First, let's multiply the matrix X with the vector $\mathbf{z}$:

   $$X\mathbf{z} = \begin{pmatrix} 3 & 2 \\ -7 & -5 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 3(1) + 2(-1) \\ -7(1) + (-5)(-1) \end{pmatrix} = \begin{pmatrix} 3-2 \\ -7+5 \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

   Step 2: Compute $\mathbf{y}^T(X\mathbf{z})$
   Next, we compute the dot product $\mathbf{y}^T$ and the result of $X\mathbf{z}$:

   $$\mathbf{y}^T = \begin{pmatrix} 2 & 1 \end{pmatrix}, \quad X\mathbf{z} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

   Now, perform the dot product:

   $$\mathbf{y}^T X\mathbf{z} = 2(1) + 1(-2) = 2 - 2 = 0$$

   Thus, the value of $\mathbf{y}^T X\mathbf{z}$ is 0.

2. Is X invertible? If so, give the inverse, and if no, explain why not.

   A square matrix is invertible if and only if its determinant is non-zero.
   Given the matrix X:

   $$X = \begin{pmatrix} 3 & 2 \\ -7 & -5 \end{pmatrix}$$

   The determinant of a 2×2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is given by:

   $$\det(X) = ad - bc$$

In this case:

$$\det(X) = (3)(-5) - (2)(-7) = -15 + 14 = -1$$

Since the determinant of X is -1, which is non-zero, X is invertible.

Now, the inverse of a 2×2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is given by:

$$X^{-1} = \frac{1}{\det(X)} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Substitute the values for $a = 3$, $b = 2$, $c = -7$, and $d = -5$:

$$X^{-1} = \frac{1}{-1} \begin{pmatrix} -5 & -2 \\ 7 & 3 \end{pmatrix} = \begin{pmatrix} 5 & 2 \\ -7 & -3 \end{pmatrix}$$

Thus, the inverse of X is:

$$X^{-1} = \begin{pmatrix} 5 & 2 \\ -7 & -3 \end{pmatrix}$$

2 Calculus [3 pts]

1. If $y = e^{-x} + \arctan(z)x^{6/z} - \ln \frac{x}{x+1}$, what is the partial derivative of y with respect to x?

   Given:

   $$y = e^{-x} + \arctan(z)x^{6/z} - \ln\left(\frac{x}{x+1}\right)$$

   We will differentiate each term with respect to x, treating z as a constant.

   Term 1: e^{-x}
   The derivative of e^{-x} is:

   $$\frac{\partial}{\partial x}\left(e^{-x}\right) = -e^{-x}$$

   Term 2: $\arctan(z)x^{6/z}$
   Using the power rule:

   $$\frac{\partial}{\partial x}\left(x^{6/z}\right) = \frac{6}{z}x^{(6/z)-1}$$

   Thus, the derivative of the second term is:

   $$\frac{\partial}{\partial x}\left(\arctan(z)x^{6/z}\right) = \arctan(z) \cdot \frac{6}{z}x^{(6/z)-1}$$

   Term 3: $-\ln\left(\frac{x}{x+1}\right)$
   Simplify:

   $$\ln\left(\frac{x}{x+1}\right) = \ln(x) - \ln(x+1)$$

   Thus, the derivative is:

   $$\frac{\partial}{\partial x}\left(-\ln\left(\frac{x}{x+1}\right)\right) = -\left(\frac{1}{x} - \frac{1}{x+1}\right)$$

   Final Result:
   Summing the derivatives:

   $$\frac{\partial y}{\partial x} = -e^{-x} + \arctan(z) \cdot \frac{6}{z}x^{(6/z)-1} - \left(\frac{1}{x} - \frac{1}{x+1}\right)$$

3 Probability and Statistics [10 pts]

Consider a sequence of data $S = (1, 1, 1, 0, 1)$ created by flipping a coin x five times, where 0 denotes that the coin turned up heads and 1 denotes that it turned up tails.

1. (2.5 pts) What is the probability of observing this data, assuming it was generated by flipping a biased coin with $p(x = 1) = 0.6$?

   The probability of flipping heads is $p(x = 0) = 1 - 0.6 = 0.4$. The probability of observing the entire sequence is the product of the probabilities of each individual outcome, assuming independent flips.

   $$P(S) = p(x_1 = 1) \cdot p(x_2 = 1) \cdot p(x_3 = 1) \cdot p(x_4 = 0) \cdot p(x_5 = 1)$$

   Substituting the probabilities:

   $$P(S) = (0.6) \cdot (0.6) \cdot (0.6) \cdot (0.4) \cdot (0.6)$$

   $$P(S) = (0.6)^4 \cdot (0.4) = 0.1296 \cdot 0.4 = 0.05184$$

   Thus, the probability of observing the sequence S is:

   $$P(S) = 0.05184$$

2. (2.5 pts) Note that the probability of this data sample could be greater if the value of $p(x = 1)$ was not 0.6, but instead some other value. What is the value that maximizes the probability of S? Please justify your answer.

   The sequence S contains 4 tails (denoted by 1) and 1 head (denoted by 0). The probability of observing this sequence is:

   $$P(S) = p^4(1 - p)$$

   where p^4 represents the probability of getting 4 tails, and $(1 - p)$ represents the probability of getting 1 head.

   To maximize $P(S)$, we take the derivative of $P(S)$ with respect to p and set it equal to 0. Let:

   $$P(p) = p^4(1 - p)$$

   Differentiate $P(p)$ using the product rule:

   $$\frac{dP(p)}{dp} = \frac{d}{dp}\left(p^4 \cdot (1 - p)\right) = \frac{d}{dp}\left(p^4\right) \cdot (1 - p) + p^4 \cdot \frac{d}{dp}(1 - p)$$

   This simplifies to:

   $$\frac{dP(p)}{dp} = 4p^3(1 - p) - p^4$$

   Set the derivative equal to 0:

   $$4p^3(1 - p) - p^4 = 0$$

   Factor the equation:

   $$p^3(4 - 5p) = 0$$

   This gives two solutions: $p = 0$ and $p = \frac{4}{5}$. Since $p = 0$ is not realistic in this context, the value that maximizes the probability is:

   $$p = \frac{4}{5}$$

   Thus, the value of p that maximizes the probability of observing the sequence S is $p = \frac{4}{5}$.

A	B	$P(A,B)$
0	0	0.3
0	1	0.1
1	0	0.1
1	1	0.5

3. (5 pts) Consider the following joint probability table where both A and B are binary random variables:

(a) What is $P(A=0|B=1)$?

$$P(A=0 \mid B=1) = \frac{P(A=0,B=1)}{P(B=1)}$$

From the table:

$$P(A=0,B=1) = 0.1$$

Next, we compute $P(B=1)$:

$$P(B=1) = P(A=0,B=1) + P(A=1,B=1) = 0.1 + 0.5 = 0.6$$

Thus, the conditional probability is:

$$P(A=0 \mid B=1) = \frac{0.1}{0.6} = \frac{1}{6}$$

(b) What is $P(A=1 \vee B=1)$?

$$P(A=1 \vee B=1) = P(A=1) + P(B=1) - P(A=1,B=1)$$

From the table:

$$P(A=1) = P(A=1,B=0) + P(A=1,B=1) = 0.1 + 0.5 = 0.6$$

We already know that $P(B=1) = 0.6$, and from the table:

$$P(A=1,B=1) = 0.5$$

Thus:

$$P(A=1 \vee B=1) = 0.6 + 0.6 - 0.5 = 0.7$$

4 Big-O Notation [6 pts]

For each pair (f, g) of functions below, list which of the following are true: $f(n) = O(g(n))$, $g(n) = O(f(n))$, both, or neither. Briefly justify your answers.

1. $f(n) = \ln(n)$, $g(n) = \log_2(n)$.

$$\log_2(n) = \frac{\ln(n)}{\ln(2)}$$

Thus, $\ln(n)$ and $\log_2(n)$ differ by a constant factor $\ln(2)$, which does not affect asymptotic growth.
$f(n) = \ln(n) = O(\log_2(n))$

Since $\ln(n)$ is a constant multiple of $\log_2(n)$, we have $f(n) = O(g(n))$.
$g(n) = \log_2(n) = O(\ln(n))$

Similarly, $\log_2(n)$ is a constant multiple of $\ln(n)$, so $g(n) = O(f(n))$.
Therefore, both $f(n) = O(g(n))$ and $g(n) = O(f(n))$ are true.

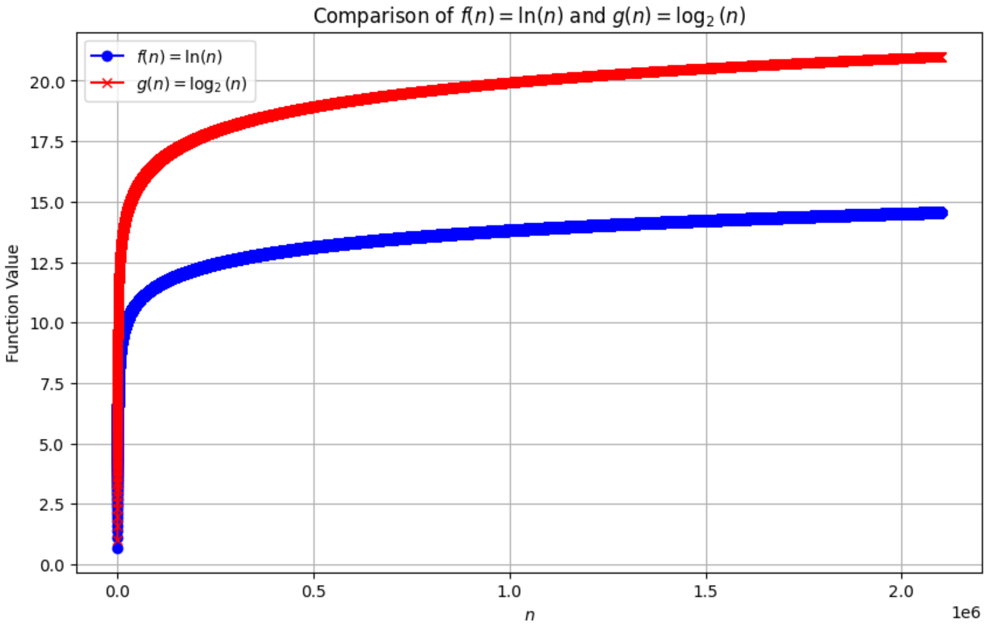

Figure 1: Comparison of $f(n) = \ln(n)$ and $g(n) = \log_2(n)$. The plot shows that $f(n)$ and $g(n)$ grow at the same rate up to a constant factor.

2. $f(n) = \log_2 \log_2(n)$, $g(n) = \log_2(n)$.

1. Is $f(n) = O(g(n))$?
To show $f(n) = O(g(n))$, we need to verify if:

$$\log_2 \log_2(n) \leq C \cdot \log_2(n)$$

for some constant C and sufficiently large n. Since $\log_2 \log_2(n)$ grows much slower than $\log_2(n)$, this inequality will hold true. Thus:

$$\log_2 \log_2(n) = O(\log_2(n))$$

2. Is $g(n) = O(f(n))$?
To show $g(n) = O(f(n))$, we need to check if:

$$\log_2(n) \leq C \cdot \log_2 \log_2(n)$$

for some constant C and sufficiently large n. Since $\log_2(n)$ grows faster than $\log_2 \log_2(n)$, there does not exist such a constant C that satisfies this inequality. Thus:

$$\log_2(n) \neq O(\log_2 \log_2(n))$$

Therefore, $f(n) = O(g(n))$ but $g(n) \neq O(f(n))$.

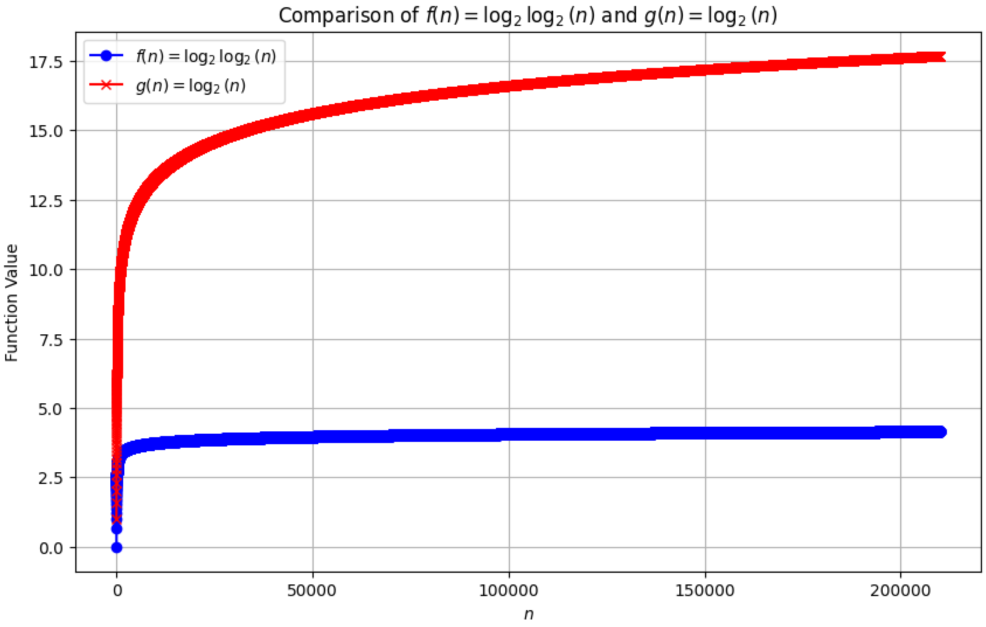

Figure 2: Comparison of $f(n) = \log_2 \log_2(n)$ and $g(n) = \log_2(n)$. The plot shows that $f(n)$ grows much slower than $g(n)$.

3. $f(n) = n!$, $g(n) = 2^n$.

1. Is $f(n) = O(g(n))$?
To determine if $f(n) = O(g(n))$, we need to check if there exists a constant C such that:

$$n! \leq C \cdot 2^n$$

Using Stirling's approximation:

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

We compare:

$$\frac{n!}{2^n} \approx \sqrt{2\pi n} \left(\frac{n}{2e}\right)^n$$

For large n, $\left(\frac{n}{2e}\right)^n$ grows without bound, so $n!$ grows faster than 2^n. Hence, $n! \neq O(2^n)$.
2. Is $g(n) = O(f(n))$?
To determine if $g(n) = O(f(n))$, we need to check if:

$$2^n \leq C \cdot n!$$

Since $n!$ grows faster than 2^n, there will always be some constant C such that this inequality holds for sufficiently large n. Thus:

$$2^n = O(n!)$$

Therefore, $f(n) = n!$ is not $O(g(n))$, but $g(n) = 2^n$ is $O(f(n))$.

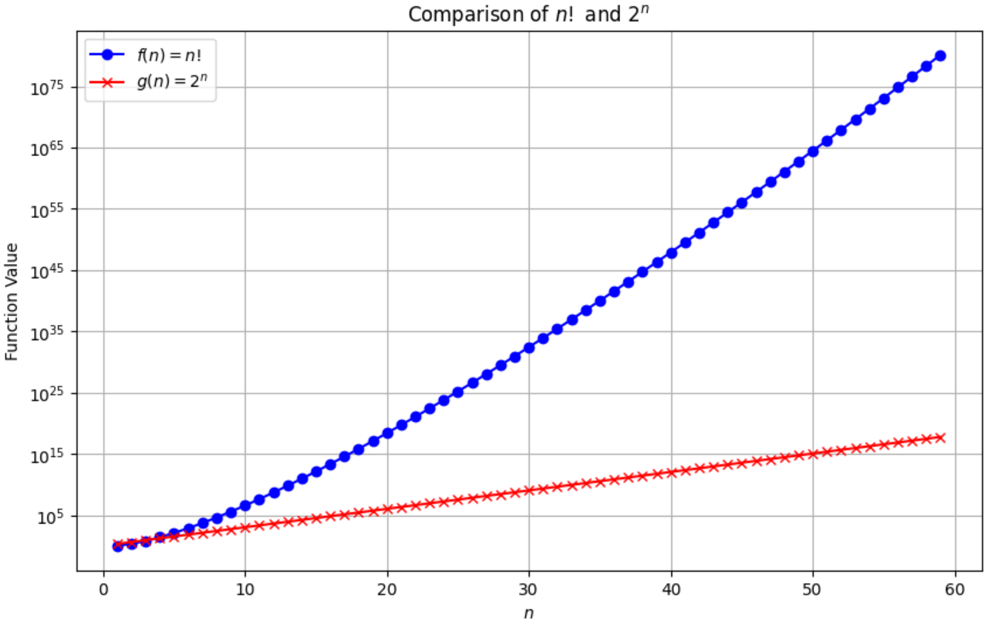

Figure 3: Comparison of $n!$ and 2^n. The plot shows that $n!$ grows much faster than 2^n for larger values of n.

5 Probability and Random Variables

5.1 Probability [12.5 pts]

State true or false. Here Ω denotes the sample space and A^c denotes the complement of the event A.

1. For any $A, B \subseteq \Omega$, $P(A|B)P(A) = P(B|A)P(B)$.

   The statement is **false**.

   Explanation: The correct form of this relation is Bayes' Theorem, which states:

   $$P(A|B)P(B) = P(B|A)P(A)$$

   This equation indicates that $P(A|B)$ is the conditional probability of A given B, and it is equal to the conditional probability of B given A, scaled by their respective unconditional probabilities $P(A)$ and $P(B)$.

2. For any $A, B \subseteq \Omega$, $P(A \cup B) = P(A) + P(B) - P(B \cap A)$.

   The statement is **true**.

   Explanation: This is the formula for the probability of the union of two events, which accounts for the overlap between the events. The inclusion-exclusion principle is applied here to avoid double-counting the probability of the intersection $P(A \cap B)$, which appears in both $P(A)$ and $P(B)$. Therefore, the correct expression for the probability of $A \cup B$ is:

   $$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

3. For any $A, B, C \subseteq \Omega$ such that $P(B \cup C) > 0$, $\frac{P(A \cup B \cup C)}{P(B \cup C)} \geq P(A|B \cup C)P(B)$.

   The statement is **true**.

4. For any $A, B \subseteq \Omega$ such that $P(B) > 0, P(A^c) > 0$, $P(B|A^C) + P(B|A) = 1$.

   This statement is **false**.

   Counterexample:

   Consider the following probabilities:

   - $P(B) = 0.8$
   - $P(A) = 0.3$
   - $P(A \cap B) = 0.2$

   We will calculate whether $P(B|A^c) + P(B|A) = 1$ holds:

   (a) Calculate $P(B|A)$:
   $$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{0.2}{0.3} = \frac{2}{3} \approx 0.6667$$

   (b) Calculate $P(A^c)$:
   $$P(A^c) = 1 - P(A) = 1 - 0.3 = 0.7$$

   (c) Find $P(B \cap A^c)$:
   $$P(B \cap A^c) = P(B) - P(B \cap A) = 0.8 - 0.2 = 0.6$$

(d) Calculate $P(B|A^c)$:

$$P(B|A^c) = \frac{P(B \cap A^c)}{P(A^c)} = \frac{0.6}{0.7} \approx 0.8571$$

(e) Check if $P(B|A^c) + P(B|A) = 1$:

$$P(B|A) + P(B|A^c) \approx 0.6667 + 0.8571 = 1.5238$$

Since $1.5238 \neq 1$, the statement $P(B|A^c) + P(B|A) = 1$ does not hold. Therefore, the statement is **false**.

5. If A and B are independent events, then A^c and B^c are independent.

This statement is **true**.

Proof:

Start with the independence of A and B. By definition, if A and B are independent, then:

$$P(A \cap B) = P(A) \cdot P(B)$$

Using De Morgan's law:

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

Thus:

$$P(A^c \cap B^c) = 1 - P(A \cup B)$$

Calculate $P(A^c)$ and $P(B^c)$:

$$P(A^c) = 1 - P(A)$$
$$P(B^c) = 1 - P(B)$$

Compute $P(A^c) \cdot P(B^c)$:

$$P(A^c) \cdot P(B^c) = (1 - P(A)) \cdot (1 - P(B))$$
$$P(A^c) \cdot P(B^c) = 1 - P(A) - P(B) + P(A) \cdot P(B)$$
$$P(A^c) \cdot P(B^c) = 1 - P(A \cup B) = P(A^c \cap B^c)$$

Thus, $P(A^c \cap B^c) = P(A^c) \cdot P(B^c)$, proving that A^c and B^c are independent.

5.2 Discrete and Continuous Distributions [12.5 pts]

Match the distribution name to its probability density / mass function.

(a) Gamma : j

(b) Multinomial : i

(c) Laplace : h

(d) Poisson : l

(e) Dirichlet : k

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

(g) $f(x; n, \alpha) = \binom{n}{x} \alpha^x (1 - \alpha)^{n-x}$ for $x \in \{0, \ldots, n\}$; 0 otherwise

(h) $f(x; b, \mu) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$

(i) $f(\boldsymbol{x}; n, \boldsymbol{\alpha}) = \frac{n!}{\Pi_{i=1}^k x_i!} \Pi_{i=1}^k \alpha_i^{x_i}$ for $x_i \in \{0, \ldots, n\}$ and $\sum_{i=1}^k x_i = n$; 0 otherwise

(j) $f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ for $x \in (0, +\infty)$; 0 otherwise

(k) $f(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1}$ for $x_i \in (0, 1)$ and $\sum_{i=1}^k x_i = 1$; 0 otherwise

(l) $f(x; \lambda) = \lambda^x \frac{e^{-\lambda}}{x!}$ for all $x \in Z^+$; 0 otherwise

5.3 Mean and Variance [10 pts]

1. Consider a random variable which follows a Binomial distribution: $X \sim \text{Binomial}(n, p)$.

(a) What is the mean of the random variable?

$$E[X] = \sum_{k=0}^{n} k \cdot P(X = k)$$

For a Binomial random variable, the probability mass function is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Substituting this into the expectation formula, we get:

$$E[X] = \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k}$$

Using the identity $k \cdot \binom{n}{k} = n \cdot \binom{n-1}{k-1}$, we rewrite the sum as:

$$E[X] = \sum_{k=0}^{n} n \cdot \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

Factoring out n:

$$E[X] = n \sum_{k=1}^{n} \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

Changing the index of summation from k to $j = k - 1$, we get:

$$E[X] = np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j}$$

This is the binomial expansion of $(p + (1-p))^{n-1}$:

$$E[X] = np \cdot 1 = np$$

Thus, the mean of X is:

$$E[X] = np$$

(b) What is the variance of the random variable?

The variance of X is given by:

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

We know the following:

$$E[X] = np$$

Also:

$$E[X^2] = E[X(X-1)] + E[X]$$

To find the expectation $E[X(X-1)]$, where $X \sim \text{Binomial}(n, p)$. The binomial random variable X counts the number of successes in n independent Bernoulli trials, each with success probability p. Using indicator variables, we can express X as:

$$X = I_1 + I_2 + \cdots + I_n$$

where $I_i = 1$ if trial i is a success, and $I_i = 0$ otherwise.
Thus, $X(X-1)$ can be expanded as:

$$X(X-1) = \sum_{i=1}^{n} \sum_{j \neq i} I_i I_j$$

The expectation $E[I_i I_j]$ for independent trials i and j is:

$$E[I_i I_j] = p^2$$

since $P(\text{both trials are successes}) = p^2$.
There are $n(n-1)$ pairs of distinct trials, so:

$$E[X(X-1)] = n(n-1)p^2$$

Now, using this in the expression for $E[X^2]$:

$$E[X^2] = E[X(X-1)] + E[X] = n(n-1)p^2 + np = n^2p^2 + np(1-p)$$

Finally, substituting into the formula for variance:

$$\text{Var}(X) = E[X^2] - (E[X])^2 = n^2p^2 + np(1-p) - (np)^2 = np(1-p)$$

Thus, the variance of X is:

$$\text{Var}(X) = np(1-p)$$

2. Let X be a random variable and $\mathbb{E}[X] = 1, \text{Var}(X) = 1$. Compute the following values:

   (a) $\mathbb{E}[5X]$

   $$\mathbb{E}[aX] = a \cdot \mathbb{E}[X]$$

   Here, $a = 5$, so:

   $$\mathbb{E}[5X] = 5 \cdot \mathbb{E}[X]$$
   $$\mathbb{E}[5X] = 5 \cdot 1 = 5$$

   (b) $\text{Var}(5X)$

   $$\text{Var}(aX) = a^2 \cdot \text{Var}(X)$$

   Here, $a = 5$, so:

   $$\text{Var}(5X) = 5^2 \cdot \text{Var}(X)$$
   $$\text{Var}(5X) = 25 \cdot 1 = 25$$

   (c) $\text{Var}(X+5)$

   $$\text{Var}(X+c) = \text{Var}(X)$$

   Here, $c = 5$, so:

   $$\text{Var}(X+5) = \text{Var}(X) = 1$$

5.4 Mutual and Conditional Independence [12 pts]

1. (3 pts) If X and Y are independent random variables, show that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{X,Y}(x,y)\, dx\, dy$$

where $f_{X,Y}(x,y)$ is the joint probability density function of X and Y.

For independent random variables X and Y, the joint probability density function factors into the product of the marginal density functions:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

Substituting this into the expectation formula:

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_X(x)f_Y(y)\, dx\, dy$$

We can separate the integrals since $f_X(x)$ and $f_Y(y)$ are functions of x and y respectively:

$$\mathbb{E}[XY] = \left(\int_{-\infty}^{\infty} xf_X(x)\, dx\right)\left(\int_{-\infty}^{\infty} yf_Y(y)\, dy\right)$$

The first integral is the definition of the expectation of X:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf_X(x)\, dx$$

The second integral is the definition of the expectation of Y:

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} yf_Y(y)\, dy$$

Thus:

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

2. (3 pts) If X and Y are independent random variables, show that $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$.
Hint: $\text{Var}(X+Y) = \text{Var}(X) + 2\text{Cov}(X,Y) + \text{Var}(Y)$

$$\text{Var}(X+Y) = \mathbb{E}[(X+Y)^2] - (\mathbb{E}[X+Y])^2$$

$$(X+Y)^2 = X^2 + 2XY + Y^2$$

Thus:

$$\mathbb{E}[(X+Y)^2] = \mathbb{E}[X^2 + 2XY + Y^2]$$

Using the linearity of expectation:

$$\mathbb{E}[X^2 + 2XY + Y^2] = \mathbb{E}[X^2] + \mathbb{E}[2XY] + \mathbb{E}[Y^2]$$

$$\mathbb{E}[X^2 + 2XY + Y^2] = \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2]$$

To compute $\mathbb{E}[XY]$ for independent random variables X and Y:

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

Thus:

$$\mathbb{E}[2XY] = 2\mathbb{E}[X] \cdot \mathbb{E}[Y]$$

Substitute this into our previous equation:

$$\mathbb{E}[X^2 + 2XY + Y^2] = \mathbb{E}[X^2] + 2\mathbb{E}[X] \cdot \mathbb{E}[Y] + \mathbb{E}[Y^2]$$

Now compute $(\mathbb{E}[X + Y])^2$:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$(\mathbb{E}[X + Y])^2 = (\mathbb{E}[X] + \mathbb{E}[Y])^2$$

Expanding this:

$$(\mathbb{E}[X] + \mathbb{E}[Y])^2 = \mathbb{E}[X]^2 + 2\mathbb{E}[X] \cdot \mathbb{E}[Y] + \mathbb{E}[Y]^2$$

Thus:

$$\text{Var}(X + Y) = \mathbb{E}[X^2] + 2\mathbb{E}[X] \cdot \mathbb{E}[Y] + \mathbb{E}[Y^2] - (\mathbb{E}[X]^2 + 2\mathbb{E}[X] \cdot \mathbb{E}[Y] + \mathbb{E}[Y]^2)$$

$$\text{Var}(X + Y) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

This proves that if X and Y are independent random variables, then $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.

3. (6 pts) If we roll two dice that behave independently of each other, will the result of the first die tell us something about the result of the second die?

Two random variables X and Y are said to be independent if and only if the joint probability distribution factors into the product of the marginal distributions. Mathematically, this is expressed as:

$$P(X \cap Y) = P(X) \cdot P(Y)$$

In the context of rolling two dice, let X be the outcome of the first die and Y be the outcome of the second die. The dice are independent, which means:

$$P(X = x \text{ and } Y = y) = P(X = x) \cdot P(Y = y)$$

For independent random variables, the conditional probability of Y given X is equal to the marginal probability of Y:

$$P(Y = y \mid X = x) = P(Y = y)$$

This indicates that the outcome of the first die does not influence or change the probability distribution of the outcome of the second die. In other words, knowing the result of the first die does not give us any information about the result of the second die.

Therefore, when rolling two independent dice, the result of the first die will not tell us anything about the result of the second die.

If, however, the first die's result is a 1, and someone tells you about a third event — that the sum of the two results is even — then given this information is the result of the second die independent of the first die?

Let's denote the outcomes of the first and second dice as X and Y, respectively. We are given that $X = 1$ and the event E that the sum $X + Y$ is even. We need to determine if Y is independent of X given E.

First, we need to determine which values of Y are possible given that $X + Y$ is even and $X = 1$.

If $X = 1$, then Y must be such that $1 + Y$ is even. Thus, Y must be odd. The possible outcomes for Y are $1, 3, 5$.

The probability of $Y = y$ given $X = 1$ and E can be computed as:

$$P(Y = y \mid X = 1, E) = \frac{P(X = 1 \text{ and } Y = y \text{ and } E)}{P(E)}$$

For our problem, we need to find the probability distribution of Y when $X = 1$ and the sum is even.

Thus, the distribution of Y changes based on the information that $X + Y$ is even, and this dependence implies that knowing Y is constrained by X. Therefore, the result of Y is not independent of X given this additional information.

In summary, given that the sum of the two dice is even, the result of the second die is not independent of the first die. The result of the second die is constrained by the value of the first die and the requirement that the sum be even.

5.5 Central Limit Theorem [3 pts]

Prove the following result.

1. Let $X_1, \ldots, X_n$ are iid, $X_i \sim \mathcal{N}(0, 1)$, and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, then the distribution of $\bar{X}$ satisfies

$$\sqrt{n}\bar{X} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0, 1)$$

To prove this result, we will use the Central Limit Theorem (CLT) and properties of normal distributions.

Since $X_i \sim \mathcal{N}(0, 1)$, the mean and variance of each X_i are:

$$\mathbb{E}[X_i] = 0 \quad \text{and} \quad \text{Var}(X_i) = 1.$$

The sample mean $\bar{X}$ is given by:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

By linearity of expectation:

$$\mathbb{E}[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\cdot n \cdot 0 = 0.$$

For the variance of $\bar{X}$:

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\cdot n = \frac{1}{n}.$$

To find the distribution of $\sqrt{n}\bar{X}$, we standardize $\bar{X}$:

$$\sqrt{n}\bar{X} = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i.$$

According to the CLT, if $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2, then:

$$\frac{\frac{1}{n}\sum_{i=1}^{n}X_i - \mu}{\frac{\sigma}{\sqrt{n}}} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1).$$

In our case, $\mu = 0$ and $\sigma^2 = 1$. Therefore:

$$\frac{\bar{X}-0}{\frac{1}{\sqrt{n}}} = \sqrt{n}\bar{X} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1).$$

Thus,

$$\sqrt{n}\bar{X} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1).$$

6 Linear algebra

6.1 Norms [5 pts]

Draw the regions corresponding to vectors $\mathbf{x} \in \mathbb{R}^2$ with the following norms:

1. $||\mathbf{x}||_1 \leq 1$ (Recall that $||\mathbf{x}||_1 = \sum_i |x_i|$)

   The region corresponding to $||\mathbf{x}||_1 \leq 1$ is a diamond-shaped region in the $\mathbb{R}^2$ plane. This region can be described by the inequalities:

$$|x_1| + |x_2| \leq 1.$$
$$x_1 + x_2 = 1,$$
$$x_1 - x_2 = 1,$$
$$-x_1 + x_2 = 1,$$
$$-x_1 - x_2 = 1.$$

   These lines form a diamond with vertices at $(1,0)$, $(0,1)$, $(-1,0)$, and $(0,-1)$.

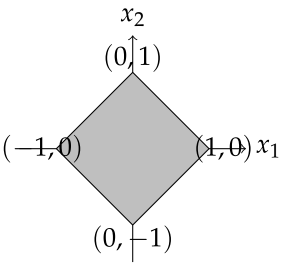

Figure 4: Region where $||\mathbf{x}||_1 \leq 1$.

2. $||\mathbf{x}||_2 \leq 1$ (Recall that $||\mathbf{x}||_2 = \sqrt{\sum_i x_i^2}$)

   The region corresponding to $||\mathbf{x}||_2 \leq 1$ is a circle centered at the origin with radius 1. This region can be described by the inequality:

$$\sqrt{x_1^2 + x_2^2} \leq 1,$$

   or equivalently:

$$x_1^2 + x_2^2 \leq 1.$$

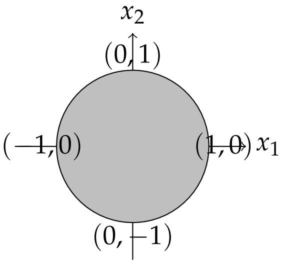

Figure 5: Region where $||\mathbf{x}||_2 \leq 1$.

3. $||\mathbf{x}||_\infty \leq 1$ (Recall that $||\mathbf{x}||_\infty = \max_i |x_i|$)

   The region corresponding to $||\mathbf{x}||_\infty \leq 1$ is a square centered at the origin with side length 2. This region can be described by the inequalities:

$$-1 \leq x_1 \leq 1$$
$$-1 \leq x_2 \leq 1$$

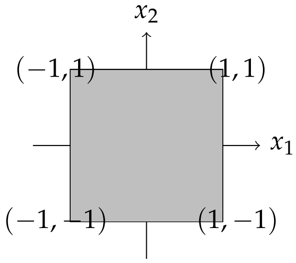

Figure 6: Region where $||\mathbf{x}||_\infty \leq 1$.

For $M = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 3 \end{pmatrix}$, Calculate the following norms.

4. $||M||_2$ (L2 norm)

   The L_2 norm of a matrix M is given by the largest singular value of M. For a diagonal matrix like M, the singular values are the absolute values of the diagonal entries. Thus, the L_2 norm is the largest diagonal entry.

$$||M||_2 = \max\{|5|, |7|, |3|\} = 7$$

5. $||M||_F$ (Frobenius norm)

   The Frobenius norm of a matrix M is computed as the square root of the sum of the absolute squares of its entries. For a diagonal matrix, this is:

$$||M||_F = \sqrt{5^2 + 7^2 + 3^2} = \sqrt{25 + 49 + 9} = \sqrt{83}$$

6.2 Geometry [10 pts]

Prove the following. Provide all steps.

1. The smallest Euclidean distance from the origin to some point $\mathbf{x}$ in the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ is $\frac{|b|}{||\mathbf{w}||_2}$. You may assume $\mathbf{w} \neq 0$.

   The Euclidean distance from a point $\mathbf{x}$ to the origin is given by $d = \|\mathbf{x}\|_2$. To minimize this distance, we find the perpendicular (shortest) distance between the origin and the hyperplane.

   Parametrize the line passing through the origin and orthogonal to the hyperplane The normal vector of the hyperplane is $\mathbf{w}$. Any line perpendicular to the hyperplane must be parallel to $\mathbf{w}$. Therefore, we can parametrize any point on this line as:

$$\mathbf{x}(t) = t\mathbf{w}$$

where t is a scalar.

Substitute $\mathbf{x}(t) = t\mathbf{w}$ into the equation of the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$:

$$\mathbf{w}^T(t\mathbf{w}) + b = 0$$

$$t\mathbf{w}^T\mathbf{w} + b = 0$$

$$t\|\mathbf{w}\|_2^2 = -b$$

$$t = \frac{-b}{\|\mathbf{w}\|_2^2}$$

Substitute $t = \frac{-b}{\|\mathbf{w}\|_2^2}$ back into the parametrized equation $\mathbf{x}(t) = t\mathbf{w}$:

$$\mathbf{x} = \frac{-b}{\|\mathbf{w}\|_2^2}\mathbf{w}$$

$$d = \|\mathbf{x}\|_2 = \left\| \frac{-b}{\|\mathbf{w}\|_2^2}\mathbf{w} \right\|_2$$

$$d = \frac{|b|}{\|\mathbf{w}\|_2}$$

2. The Euclidean distance between two parallel hyperplane $\mathbf{w}^T\mathbf{x} + b_1 = 0$ and $\mathbf{w}^T\mathbf{x} + b_2 = 0$ is $\frac{|b_1 - b_2|}{\|\mathbf{w}\|_2}$ (Hint: you can use the result from the last question to help you prove this one).

   From the previous problem, we know that the shortest distance from the origin to the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ is given by

$$d = \frac{|b|}{\|\mathbf{w}\|_2}$$

Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be points on the two hyperplanes, respectively. Since the hyperplanes are parallel, the normal vector $\mathbf{w}$ is the same for both hyperplanes. The Euclidean distance between the two hyperplanes is simply the difference in their distances from the origin.

Using the result from the previous problem, the distance from the origin to the hyperplane $\mathbf{w}^T\mathbf{x} + b_1 = 0$ is:

$$d_1 = \frac{|b_1|}{\|\mathbf{w}\|_2}$$

Similarly, the distance from the origin to the hyperplane $\mathbf{w}^T\mathbf{x} + b_2 = 0$ is:

$$d_2 = \frac{|b_2|}{\|\mathbf{w}\|_2}$$

$$d = \left| \frac{|b_1|}{\|\mathbf{w}\|_2} - \frac{|b_2|}{\|\mathbf{w}\|_2} \right|$$

$$d = \frac{|b_1 - b_2|}{\|\mathbf{w}\|_2}$$

Thus, the Euclidean distance between the two parallel hyperplanes $\mathbf{w}^T\mathbf{x} + b_1 = 0$ and $\mathbf{w}^T\mathbf{x} + b_2 = 0$ is:

$$d = \frac{|b_1 - b_2|}{\|\mathbf{w}\|_2}$$

This distance is minimized because we are measuring along the normal direction, which gives the shortest possible distance between the hyperplanes.

7 Programming Skills [10 pts]

Sampling from a distribution. For ea ch question, submit a scatter plot (you will have 2 plots in total). Make sure the axes for all plots have the same ranges.

1. Make a scatter plot by drawing 100 items from a two dimensional Gaussian $N((1,-1)^T, 2I)$, where I is an identity matrix in $\mathbb{R}^{2\times 2}$.

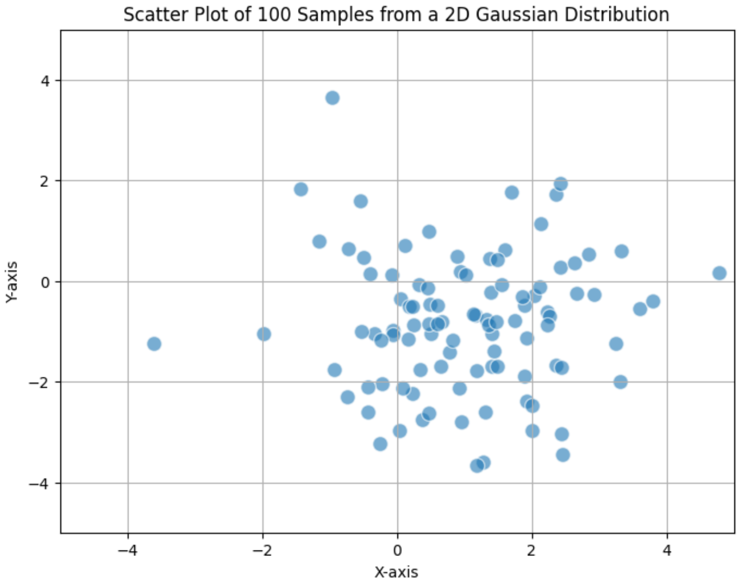

Figure 7: Scatter Plot of 100 Samples from a Gaussian Distribution

2. Make a scatter plot by drawing 100 items from a mixture distribution $0.3N\left((5,0)^T, \begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix}\right) + 0.7N\left((-5,0)^T, \begin{pmatrix} 1 & -0.25 \\ -0.25 & 1 \end{pmatrix}\right)$.

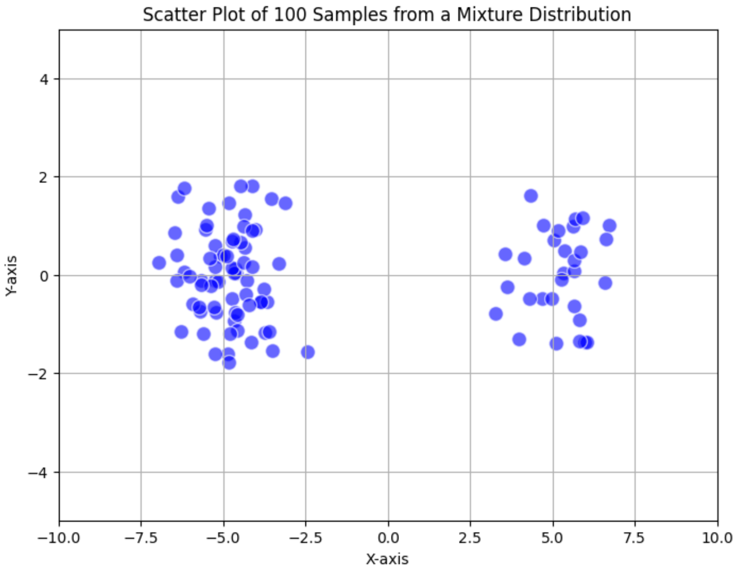

Figure 8: Scatter Plot of 100 Samples from a Mixture Distribution