Homework 1

Devansh Goenka
Student ID: 908 335 1354

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.
Use this latex file as a template to develop your homework. Submit your homework on time as a single pdf file to Canvas. There is no need to submit the latex source or any code. Please check Piazza for updates about the homework.

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 $\mathbf{y}^T X \mathbf{z}$

   $\mathbf{y}^T = \begin{pmatrix} 2 & 1 \end{pmatrix}$
   $\mathbf{y}^T X = \begin{pmatrix} -1 & -1 \end{pmatrix}$
   $\mathbf{y}^T X Z = 0\ [Answer]$

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

   The columns of X are linearly independent, hence we can say that X is invertible (or non-singular).
   One more way of proving this is to check if $\det X \neq 0$

   $\det X = \begin{vmatrix} 3 & 2 \\ -7 & -5 \end{vmatrix} = -1$
   Thus, the inverse of X exists.
   $\mathbf{X}^{-1} = \frac{1}{-1}\begin{pmatrix} -5 & -2 \\ 7 & 3 \end{pmatrix} = \begin{pmatrix} 5 & 2 \\ -7 & -3 \end{pmatrix}\ [Answer]$

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?
   $y = e^{-x} + \arctan(z)x^{6/z} - (\ln x - \ln(x+1))$
   $y = e^{-x} + \arctan(z)x^{6/z} + \ln(x+1) - \ln x$
   $\frac{\partial y}{\partial x} = (-1) * e^{-x} + \arctan(z) * \frac{6}{z}x^{\frac{6}{z}-1} + \frac{1}{x+1} - \frac{1}{x}$
   $\frac{\partial y}{\partial x} = -e^{-x} + \frac{6}{z}\arctan(z)x^{\frac{6-z}{z}} + \frac{1}{x+1} - \frac{1}{x} [Answer]$

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 $p(x = 1) = 0.6$?

With the biased coin,
$p(x = 1) = 0.6$
The probability of observing this data becomes : 0.6 * 0.6 * 0.6 * 0.4 * 0.6 = 0.05184 [Answer]

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.
Let us say that the value of p(x=1) which maximizes the above sequence is x. Clearly, $0 < x < 1$.
The probability of the above sequence occurring then becomes:
$x * x * x * (1 - x) * x = \mathbf{x}^4 - \mathbf{x}^5$
As we know, to find the maxima of this polynomial function, we need to equate the first derivative to 0.
$f'(x) = 4\mathbf{x}^3 - 5\mathbf{x}^4 = 0$
$\rightarrow 4\mathbf{x}^3 = 5\mathbf{x}^4$
$\rightarrow x = \frac{4}{5} = 0.8$
To verify if this is indeed the maxima, we need to check that the second order derivative should be negative.
$f''(x) = 12\mathbf{x}^2 - 20\mathbf{x}^3$
At x = 0.8, $f''(x) = -2.56$
Thus, the value for $p(x = 1)$ which maximises the sequence is $0.8[Answer]$.

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

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

(a) What is $P(A = 0|B = 1)$?
From the chain rule of probability, we have :
$P(A|B) = \frac{P(A \cap B)}{P(B)}$
Thus, $P(A = 0|B = 1) = \frac{P(A = 0 \cap B = 1)}{P(B = 1)} = \frac{0.1}{0.1 + 0.5} = 0.1667\ [Answer]$

(b) What is $P(A = 1 \vee B = 1)$?
We know:
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Thus,
$P(A = 1 \vee B = 1) = P(A = 1 \cup B = 1) = P(A = 1) + P(B = 1) - P(A = 1 \cap B = 1)$
$\rightarrow 0.6 + 0.6 - 0.5 = 0.7\ [Answer]$

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. $f(n) = \ln(n)$, $g(n) = \log_2(n)$.
Both are true.
$g(n) = \ln(2) * f(n)$
Thus, we can re-write the above functions as $f(n) = k * g(n)$ and $g(n) = k' * f(n)$
Therefore, $f(n) = O(g(n)$ and $g(n) = O(f(n))\ [Answer]$

2. $f(n) = \log_2 \log_2(n)$, $g(n) = \log_2(n)$.
Only $f(n) = O(g(n))$ is true.
For sufficiently large n, we can easily see that $f(n) < g(n)$
Therefore, we can say that $f(n) = O(g(n))\ [Answer]$

3. $f(n) = n!$, $g(n) = 2^n$.
Only $g(n) = O(f(n))$ is true.

For sufficiently large n, $g(n) < f(n)$
Therefore, we can say that $g(n) = O(f(n))$ $[Answer]$

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 $A, B \subseteq \Omega$, $P(A|B)P(A) = P(B|A)P(B)$.
   False
2. For any $A, B \subseteq \Omega$, $P(A \cup B) = P(A) + P(B) - P(B \cap A)$.
   True
3. For any $A, B, C \subseteq \Omega$ such that $P(B \cup C) > 0$, $\frac{P(A \cup B \cup C)}{P(B \cup C)} \geq P(A|B \cup C)P(B)$.
   True
4. For any $A, B \subseteq \Omega$ such that $P(B) > 0, P(A^c) > 0$, $P(B|A^C) + P(B|A) = 1$.
   False
5. If A and B are independent events, then A^c and B^c are independent.
   True

5.2 Discrete and Continuous Distributions [12.5 pts]

Match the distribution name to its probability density / mass function. Below, $|\boldsymbol{x}| = k$.

(a) Gamma (j)

(b) Multinomial (i)

(c) Laplace (h)

(d) Poisson (l)

(e) Dirichlet (k)

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

(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?
   np
   (b) What is the variance of the random variable?
   $np(1-p)$
2. Let X be a random variable and $\mathbb{E}[X] = 1, \text{Var}(X) = 1$. Compute the following values:
   (a) $\mathbb{E}[5X]$
   We know, $\mathbb{E}[\alpha X] = \alpha\mathbb{E}[X]$
   Thus, $\mathbb{E}[5X] = 5\mathbb{E}[X] = 5$ $[Answer]$
   (b) $\text{Var}(5X)$
   We know, $\text{Var}(\alpha X) = \alpha^2\text{Var}(X)$
   Thus, $\text{Var}(5X) = 25\text{Var}(X) = 25$ $[Answer]$

(c) $\text{Var}(X + 5)$
The variance is not affected by addition of a constant as the distribution still remains the same. Hence, $\text{Var}(X + 5) = 1$ $[Answer]$

5.4 Mutual and Conditional Independence [12 pts]

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

   Let us consider that both random variables are continuous.
   We know that:
   $\mathbb{E}[g(x, y)] = \int\int g(x, y)f_{xy}(x, y)dydx$
   Here, $g(x, y) = XY$
   $\mathbb{E}(XY) = \int\int xyf_{xy}(x, y)dydx$

   Since X & Y are independent,the joint probability function can factor,
   $\mathbb{E}(XY) = \int\int xyf_x(x)f_y(y)dydx$
   $\mathbb{E}(XY) = \int xf_x(x)\int yf_y(y)dydx$
   $\mathbb{E}(XY) = \mathbb{E}(Y) \cdot \int xf_x(x)dx$
   $\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$ $[Answer]$

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

   We begin with the expansion:
   $\text{Var}(X + Y) = \text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y)$
   We know that $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
   Since X & Y are independent, $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$
   →Cov(X, Y) = 0
   →Var(X+Y) = Var(X) + 2*0 + Var(Y)
   →Var(X+Y) = Var(X) + Var(Y) [Answer]

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?

   No, the rolls are independent of each other.

   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?

   Now, the outcome of the second die becomes dependent on the first die.

5.5 Central Limit Theorem [3 pts]

Prove the following result.

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

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

   Here, $X_1, X_2, ...X_n$ are $i.i.d$
   According to the Law of Large Numbers:
   Whenever $n \to \infty$, and $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \mathbb{E}(X_n) = \mu$, and we have
   $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
   Then $\bar{X}$ converges to μ
   We have :
   $\mathbb{E}(\sqrt{n}\bar{X}) = \sqrt{n}\mathbb{E}(\bar{X})$
   Since $\bar{X}$ converges to μ, which is 0 here,
   $\mathbb{E}(\sqrt{n}\bar{X}) = 0$
   Similarly,
   $\text{Var}(\sqrt{n}\bar{X}) = n\text{Var}(\bar{X})$
   $\text{Var}(\bar{X}) = \dfrac{\sigma^2}{n} = \dfrac{1}{n}$

$\rightarrow \mathrm{Var}(\sqrt{n}\bar{X}) = \frac{n}{n} = 1$
Moreover, from the Central Limit Theorem, we know that whenever $i.i.d$ variables are added, their normalized sum converges to a Normal Distribution.

Thus, we can establish that

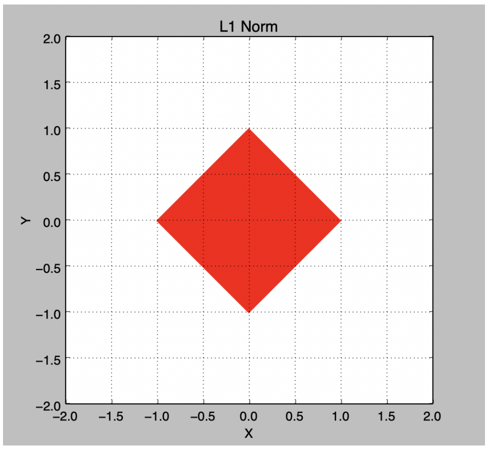
$$\sqrt{n}\bar{X} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)\ [Answer]$$

6 Linear algebra

6.1 Norms [5 pts]

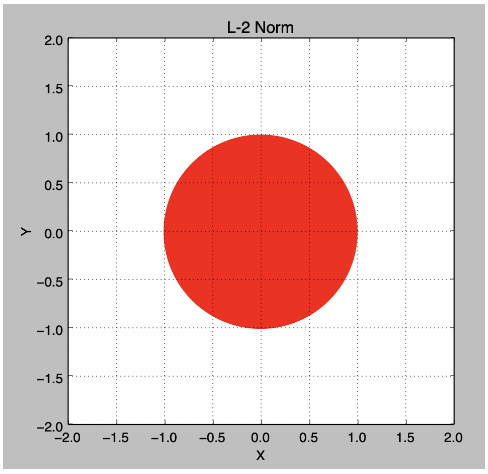
Draw the regions corresponding to vectors $\mathbf{x} \in \mathbb{R}^2$ with the following norms:

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



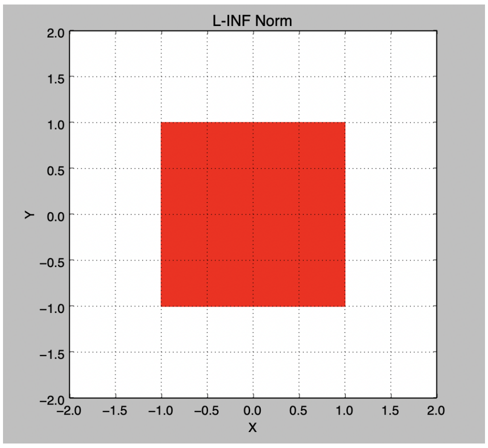
The shaded region indicates all vectors having L-1 norm ≤ 1

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



The shaded region indicates all vectors having L-2 norm ≤ 1

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



The shaded region indicates all vectors having L-∞ norm ≤ 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 L2 norm or spectal norm is equal to the largest singular value in the SVD of M
   The largest singular value after performing the SVD of M is 7
   Therefore, $||M||_2 = 7\ [Answer]$

5. $||M||_F$ (Frobenius norm)
   $||M||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}$
   $||M||_F = 9.1104\ [Answer]$

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 $\mathbf{w} \neq 0$.

   Let us consider a point x on the hyperplane. To find the smallest distance to the hyperplane from the origin, we need to calculate the projection of the vector from x to the origin (o) on w.
   Thus:
   distance = $||proj_w(x)||$
   $\rightarrow ||\frac{x \cdot w}{w \cdot w} w||$
   $\rightarrow \frac{|x \cdot w|}{||w||_2}$
   Now, As $w \cdot x = -b$
   distance $= \frac{|b|}{||w||_2}\ [Answer]$

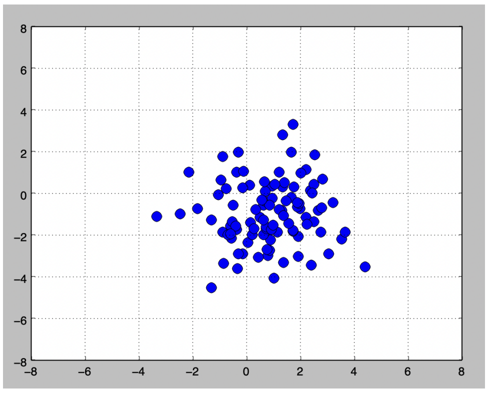
2. The Euclidean distance between two parallel hyperplane $\mathbf{w}^T\mathbf{x} + b_1 = 0$ and $\mathbf{w}^T\mathbf{x} + b_2 = 0$ is $\frac{|b_1 - b_2|}{||\mathbf{w}||_2}$ (Hint: you can use the result from the last question to help you prove this one).

   Using the previous solution, we can say that:
   $d_1 =$ Distance from origin to hyperplane $1 = \frac{|b_1|}{||\mathbf{w}||_2}$
   $d_2 =$ Distance from origin to hyperplane $2 = \frac{|b_2|}{||\mathbf{w}||_2}$
   Now, the distance between these two hyperplanes is:
   $d = d_1 - d_2 = \frac{|b_1|}{||\mathbf{w}||_2} - \frac{|b_2|}{||\mathbf{w}||_2} = \frac{|b_1 - b_2|}{||\mathbf{w}||_2}\ [Answer]$

7 Programming Skills [10 pts]

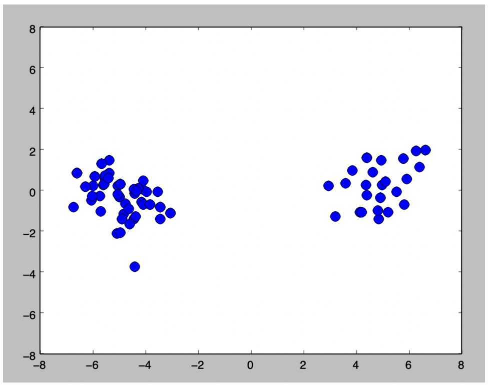
Sampling from a distribution. For each question, submit a scatter plot (you will have 3 plots in total). Make sure the axes for all plots have the same ranges.

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 $\mathbb{R}^{2\times 2}$.



100 samples drawn from the above described Gaussian distribution

2. Make a scatter plot by drawing 100 items from a mixture distribution $0.3N\left((5,0)^T, \begin{pmatrix} 1 & 0.25 \\ p.25 & 1 \end{pmatrix}\right) + 0.7N\left((-5,0)^T, \begin{pmatrix} 1 & -0.25 \\ -0.25 & 1 \end{pmatrix}\right)$.



100 samples drawn from the above described Mixture Distribution