HOMEWORK 1

Daniel Szabo
9074769625

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

$$\mathbf{y}^T X \mathbf{z} = \mathbf{y}^T \begin{pmatrix} 1 \\ -2 \end{pmatrix} = 0$$

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

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

2 Calculus [3 pts]

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

$$\frac{\delta y}{\delta x} = -e^{-x} + \frac{6}{z}\arctan(z)x^{6/z-1} + \frac{x+1}{x}\frac{1}{x^2}$$

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 $p(x = 1) = 0.6$?

   Assuming the sequence is ordered, the probability would be $.6^4 \cdot .4 = .05184$.

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.
   We want to maximize the likelihood $p^4(1-p)$, so taking the derivative we see this is equivalent to finding the root of $4p^3 - 5p^4$ in $(0, 1]$. This root is at $p = 0.8$, which is therefore the value that maximizes S. This makes intuitive sense as well, as 0.8 of the flips were heads in the sample.

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 $P(A = 0|B = 1)$?

$$P(A = 0|B = 1) = \frac{.1}{.1 + .5} = \frac{1}{6}$$

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

$$P(A = 1 \vee B = 1) = .1 + .1 + .5 = .7$$

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. $f(n) = \ln(n)$, $g(n) = \log_2(n)$.
   Because $\log_2(n) = \frac{\ln(n)}{\ln(e)}$, $f(n) = \Theta(g(n)) \implies f(n) = O(g(n))$ and $g(n) = O(f(n))$

2. $f(n) = \log_2 \log_2(n)$, $g(n) = \log_2(n)$.
   In this case, only $f(n) = O(g(n))$ because $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$.

3. $f(n) = n!$, $g(n) = 2^n$.
   Here $g(n) = O(f(n))$ by Stirling's approximation, which says $n! \sim \sqrt{n} \cdot n^n$ asymptotically, which is must faster than 2^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 $A, B \subseteq \Omega$, $P(A|B)P(A) = P(B|A)P(B)$.
   False.
   Multiplying both sides by $P(B)P(A)$ gives that $P(A)^3 = P(B)^3$, which is not always true.

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.
   This is equivalent to $P(A \cup B \cup C) \geq P(A)P(B)$. WLOG say $P(A) \geq P(B)$, which means $P(A \cup B \cup C) \geq P(A) \geq P(A)P(B)$.

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

5. If A and B are independent events, then A^c and B^c are independent.
   True.
   Applying De Morgan's law, $P(A^cB^c) = 1 - P(A \cup B) = 1 - P(A) - P(B) + P(A)P(B) = 1 - (1 - P(A))P(B) - P(A) = P(A^c) - P(A^c)P(B) = P(A^c)P(B^c)$.

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?

   $$\mathbb{E}[X] = np$$

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

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

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

   (a) $\mathbb{E}[5X]$
   $\mathbb{E}[5X] = 5$ by linearity of expectation.

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

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

5.4 Mutual and Conditional Independence [12 pts]

1. (3 pts) If X and Y are independent random variables, show that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.
   Expanding by the definition of expectation (assuming X, Y discrete for simplicity), we have

   $$\mathbb{E}[XY] = \sum_{x,y \in \Omega} xy \Pr[X = x, Y = y] = \sum_{x,y \in \Omega} xy \Pr[X = x] \Pr[Y = y]$$
   $$= \sum_{x \in \Omega} x \Pr[X = x] \sum_{y \in \Omega} y \Pr[Y = y] = \mathbb{E}[X]\mathbb{E}[Y].$$

The continuous case would be analogous with integrals replacing sums and density functions replacing probabilities.

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

   Directly applying the hint, $\text{Var}(X+Y) = \text{Var}(X) + 2\text{Cov}(X,Y) + \text{Var}(Y) = \text{Var}(X) + \text{Var}(Y)$ because the convariance of two independent random variables is 0.

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?

   It will not, because the two dice behave independently 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?

   It is no longer independent, the second die must be an odd number as well. The two dice are conditionally dependent on the third event.

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

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

   Look at the moment generating functions of $\sqrt{n}\bar{X}$:

$$M_{\sqrt{n}\bar{X}}(t) = \mathbb{E}[e^{t\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i}] = \mathbb{E}[e^{t\frac{1}{\sqrt{n}}X_1}]^n = M_{X_1}(t/\sqrt{n})^n.$$

   Using the Taylor expansion of $M_{X_1}(t/\sqrt{n})$ about $t = 0$, we see

$$\begin{aligned} M_{X_1}(t/\sqrt{n}) =& M_{X_1}(0) + M'_{X_1}(0)\frac{t}{\sqrt{n}} + M''_{X_1}(0)\frac{t^2}{2n} + o(t^2/n) \\ =& 1 + \mathbb{E}[X_1]\frac{t}{\sqrt{n}} + \mathbb{E}[X_1^2]\frac{t^2}{2n} + o(t^2/n) \\ =& 1 + \frac{t^2}{2n} + o(t^2/n). \end{aligned}$$

   Looking at the limit of $M_{\sqrt{n}\bar{X}}(t)$ as $n \to \infty$, we see that the higher order terms vanish which means

$$\lim_{n\to\infty} M_{\sqrt{n}\bar{X}}(t) = \lim_{n\to\infty}(1 + \frac{t^2}{2n} + o(t^2/n))^n = \lim_{n\to\infty}(1 + \frac{t^2}{2n})^n = e^{-\frac{t^2}{2}}.$$

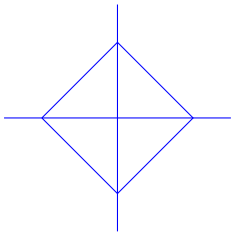
   Now, using the fact that the MGF of the normal distribution is $e^{-\frac{t^2}{2}}$, we see that the MGF of $\sqrt{n}\bar{X}$ converges to that of $\mathcal{N}(0,1)$, so the distribution converges as well. I used the proof in the wikipedia page for the CLT to help guide my proof.

6 Linear algebra

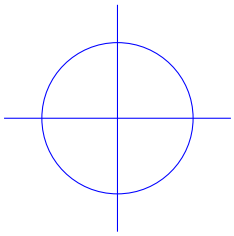
6.1 Norms [5 pts]

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

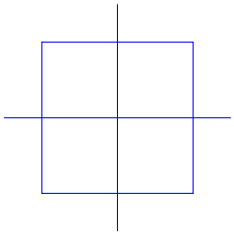
1. $||\mathbf{x}||_1 \leq 1$ (Recall that $||\mathbf{x}||_1 = \sum_i |x_i|$)



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



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



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 is just the maximum eigenvalue, so $||M||_2 = 7$.

5. $||M||_F$ (Frobenius norm)
   $||M||_F = \sqrt{83} = 9.1104335791443$.

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 $\mathbf{w} \neq 0$.
   Say the distance was $\frac{|b|}{||\mathbf{w}||_2} - \varepsilon$ for some $\varepsilon > 0$. Then there is some unit vector $\mathbf{u}$ such that

$$\begin{aligned} & \mathbf{w}^T(\frac{|b|}{||\mathbf{w}||_2} - \varepsilon)\mathbf{u} + b \\ =& \mathbf{w}^T \frac{|b|}{||\mathbf{w}||_2}\mathbf{u} + b - \varepsilon \mathbf{w}^T\mathbf{u} \\ \leq& \mathbf{w}^T \frac{|b|}{||\mathbf{w}||_2}\frac{\mathbf{w}}{\|\mathbf{w}\|_2} + b - \varepsilon \mathbf{w}^T \frac{\mathbf{w}}{\|\mathbf{w}\|_2} \\ =& |b| + b - \varepsilon\|\mathbf{w}\|_2 \\ <& 0 \quad \text{If } b \leq 0. \end{aligned}$$

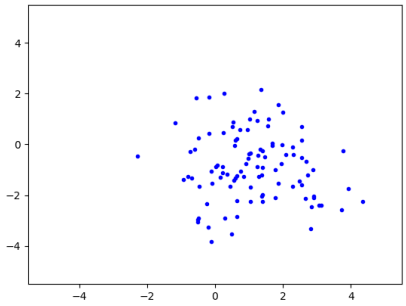
   If $b > 0$, we can repeat the same argument with $\mathbf{u} = -\frac{\mathbf{w}}{||\mathbf{w}||_2}$ to show $\mathbf{w}^T\mathbf{x} + b$ is bounded below by 0.

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).
   We can simply transform both planes by b_2 so that the origin lines on the second plane. We saw in problem 1. that the shortest distance from $\mathbf{w}^T\mathbf{x} + b_1 - b_2 = 0$ to the origin is $\frac{|b_1 - b_2|}{||\mathbf{w}||_2}$, and because these planes are parallel, that is the shortest distance at any point.

7 Programming Skills [10 pts]

Sampling from a distribution. For each 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 $\mathbb{R}^{2\times 2}$.



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