HOMEWORK 0

>>Matej Popovski<<
>>popovski<<

Instructions: This is a background self-test on the type of mathematics and computer science 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. This template can be generated by compiling the `hw0.tex` file, available under the "Files" tab on Canvas; please use it 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. 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:

$$\mathbf{X} = \begin{pmatrix} 6 & 7 \\ 8 & 9 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \qquad \mathbf{z} = \begin{pmatrix} 7 \\ 6 \end{pmatrix}$$

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

$$\mathbf{y}^\top \mathbf{X}\mathbf{z} = (2,3) \begin{bmatrix} 6 & 7 \\ 8 & 9 \end{bmatrix} \begin{bmatrix} 7 \\ 6 \end{bmatrix} = (2,3) \begin{bmatrix} 84 \\ 110 \end{bmatrix} = 498.$$

2. Is $\mathbf{X}$ invertible? If so, give the inverse, and if no, explain why not. Yes. $\det(\mathbf{X}) = 6 \cdot 9 - 7 \cdot 8 = -2 \neq 0$, so $\mathbf{X}$ is invertible. Its inverse is

$$\mathbf{X}^{-1} = \frac{1}{\det(\mathbf{X})} \begin{bmatrix} 9 & -7 \\ -8 & 6 \end{bmatrix} = \begin{bmatrix} -\frac{9}{2} & \frac{7}{2} \\ 4 & -3 \end{bmatrix}.$$

2 Calculus [3 pts]

1. If $y = e^x + \tan(z)x^{6z} - \ln(\frac{7x+z}{x^4})$, what is the partial derivative of y with respect to x?

$$\frac{\partial y}{\partial x} = e^x + 6z \tan(z)\, x^{6z-1} + \frac{21x + 4z}{x(7x + z)}.$$

3 Probability and Statistics [10 pts]

Consider a sequence of data $S = (0, 1, 1, 0, 1, 1, 1)$ created by flipping a coin x seven 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 $p(x = 1) = 0.7$? There are 5 ones and 2 zeros in S. Since $p(x = 1) = 0.7$ and $p(x = 0) = 0.3$, the probability of observing S is

$$P(S) = (0.7)^5(0.3)^2.$$

2. (2.5 pts) Note that the probability of this data sample could be greater if the value of $p(x = 1)$ was not 0.7, but instead some other value. What is the value that maximizes the probability of S? Please justify your answer.
Let $p = P(x = 1)$. The probability of observing S is

$$P(S) = p^5(1-p)^2.$$

Maximizing this is equivalent to maximizing the log-likelihood

$$\ell(p) = 5\ln(p) + 2\ln(1-p).$$

Differentiating and setting to zero:

$$\frac{d\ell}{dp} = \frac{5}{p} - \frac{2}{1-p} = 0 \quad \Rightarrow \quad 5(1-p) = 2p \quad \Rightarrow \quad p = \tfrac{5}{7}.$$

Thus, the probability is maximized when

$$p(x = 1) = \frac{5}{7}.$$

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	0.4
0	1	0.3
1	0	0.2
1	1	0.1

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

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

From the table, $P(A = 0, B = 1) = 0.3$ and $P(B = 1) = 0.3 + 0.1 = 0.4$. Thus,

$$P(A = 0 \mid B = 1) = \frac{0.3}{0.4} = 0.75.$$

(b) What is $P(A = 0 \vee B = 0)$?
We want

$$P(A = 0 \vee B = 0) = 1 - P(A = 1, B = 1).$$

From the table, $P(A = 1, B = 1) = 0.1$. Thus,

$$P(A = 0 \vee B = 0) = 1 - 0.1 = 0.9.$$

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. $f(n) = \frac{n}{2}$, $g(n) = \log_2(n)$.
Here $f(n) = \frac{n}{2}$ and $g(n) = \log_2 n$. Since $\log n = o(n)$,

$$\lim_{n\to\infty} \frac{g(n)}{f(n)} = \lim_{n\to\infty} \frac{2\log_2 n}{n} = 0.$$

Thus $g(n) = O(f(n))$ and $f(n) \neq O(g(n))$.

2. $f(n) = \ln(n), g(n) = \log_2(n)$.
Here $f(n) = \ln n$ and $g(n) = \log_2 n$. They differ by a constant factor:

$$\ln n = (\ln 2)\ \log_2 n.$$

Therefore $f(n) = \Theta(g(n))$, i.e., both $f(n) = O(g(n))$ and $g(n) = O(f(n))$.

3. $f(n) = n^{100}, g(n) = 100^n$.
Here $f(n) = n^{100}$ and $g(n) = 100^n$. Any exponential dominates any polynomial, and

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = \lim_{n\to\infty} \frac{n^{100}}{100^n} = 0.$$

Hence $f(n) = O(g(n))$ and $g(n) \neq O(f(n))$.

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 $A, B \subseteq \Omega$, $P(A|B)P(B) = P(B|A)P(A)$.
True (when defined). Since $P(A \mid B)P(B) = P(A \cap B) = P(B \mid A)P(A)$, the equality holds whenever $P(A), P(B) > 0$.

2. For any $A, B \subseteq \Omega$, $P(A \cup B) = P(A) + P(B) - P(A|B)$.
False. The correct formula is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Replacing $P(A \cap B)$ with $P(A \mid B)$ is incorrect.

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

$$\frac{P(A \cup B \cup C)}{P(B \cup C)} \geq 1.$$

Also $P(A \mid B \cup C)P(B \cup C) = P(A \cap (B \cup C)) \leq P(B \cup C) \leq 1$. Hence LHS $\geq$ RHS.

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. By the law of total probability,

$$P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c),$$

which does *not* imply $P(B \mid A^c) + P(B \mid A) = 1$ in general (take $P(B \mid A) = P(B \mid A^c) = \frac{1}{2}$ as a simple counterexample where the sum is 1, but it can be $\neq 1$ for other values).

5. For any n events $\{A_i\}_{i=1}^n$, if $P(\bigcap_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$, then $\{A_i\}_{i=1}^n$ are mutually independent.
False. Mutual independence requires $P(\cap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i)$, not $\sum_{i=1}^n P(A_i)$. For $n = 2$ with independent A, B and $P(A) = P(B) = \frac{1}{2}$, we have $P(A \cap B) = \frac{1}{4} \neq 1 = P(A) + P(B)$.

5.2 Discrete and Continuous Distributions [12.5 pts]

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

(a) Laplace h
(b) Multinomial i
(c) Poisson l
(d) Dirichlet k
(e) Gamma j

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

(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) $f(\mathbf{x}; n, \boldsymbol{\alpha}) = \frac{n!}{\Pi_{i=1}^k x_i!} \Pi_{i=1}^k \alpha_i^{x_i}$ 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) $f(\mathbf{x}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1}$ 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?
   For $X \sim \text{Binomial}(n, p)$, the mean is
   $$\mathbb{E}[X] = np.$$

   (b) What is the variance of the random variable?
   For $X \sim \text{Binomial}(n, p)$, the variance 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) $\mathbb{E}[3X]$
   Since expectation is linear,
   $$\mathbb{E}[3X] = 3\mathbb{E}[X] = 3 \cdot 1 = 3.$$

   (b) $\text{Var}(3X)$
   For variance, $\text{Var}(aX) = a^2\text{Var}(X)$, so
   $$\text{Var}(3X) = 9 \cdot \text{Var}(X) = 9 \cdot 1 = 9.$$

   (c) $\text{Var}(X + 3)$
   Variance is unaffected by adding a constant:
   $$\text{Var}(X + 3) = \text{Var}(X) = 1.$$

5.4 Mutual and Conditional Independence [12 pts]

1. (3 pts) If X and Y are independent random variables, show that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.
   Since X and Y are independent, their joint density (or pmf) factorizes: $f_{X,Y}(x, y) = f_X(x) f_Y(y)$. Then
   $$\mathbb{E}[XY] = \iint xy\, f_{X,Y}(x, y)\, dx\, dy = \left(\int x f_X(x)\, dx\right)\left(\int y f_Y(y)\, dy\right) = \mathbb{E}[X]\mathbb{E}[Y].$$
   (The same argument holds for sums in the discrete case.)

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

Using $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X,Y)$ and independence $\Rightarrow \text{Cov}(X,Y) = 0$, we get

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

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?

Since the dice are independent, the result of the first die gives no information about the result of the second die. Thus the two results are independent random variables.

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?

If the first die is a 1 and we are told that the sum of the two dice is even, then the second die must also be odd. This means the outcome of the second die is now restricted (it cannot be even). Hence, under this conditioning, the result of the second die is *not* independent of the first die.

5.5 Central Limit Theorem [3 pts]

Provide one line explanation. No calculation needed.

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

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

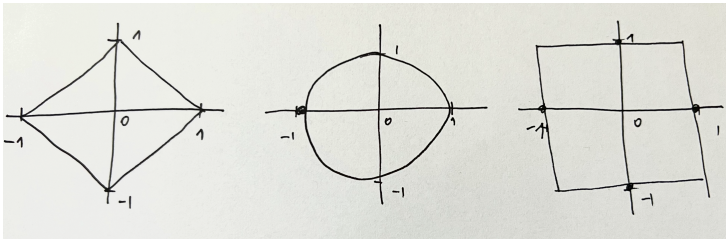
By the Central Limit Theorem, the standardized sample mean of i.i.d. random variables converges in distribution to a standard normal as $n \to \infty$.

6 Linear algebra

6.1 Norms [3 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 $||\mathbf{x}||_\infty = \max_i |x_i|$)



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 $\mathbf{w}^\top \mathbf{x} + b = 0$ is $\frac{|b|}{\|\mathbf{w}\|_2}$. You may assume $\mathbf{w} \neq 0$.

We want to minimize the distance from the origin to the hyperplane:

$$\min \|x\|_2 \quad \text{s.t. } w^\top x + b = 0, \quad w \neq 0.$$

It is more convenient to minimize $\frac{1}{2}\|x\|_2^2$ (same minimizer, differentiable). Using Lagrange multipliers:

$$\mathcal{L}(x, \lambda) = \tfrac{1}{2}\|x\|_2^2 + \lambda(w^\top x + b).$$

Stationarity condition:

$$\nabla_x \mathcal{L} = x + \lambda w = 0 \quad \Rightarrow \quad x^\star = -\lambda w.$$

Constraint condition:

$$\frac{\partial \mathcal{L}}{\partial \lambda} = w^\top x + b = 0 \quad \Rightarrow \quad w^\top(-\lambda w) + b = 0 \quad \Rightarrow \quad -\lambda\|w\|_2^2 + b = 0.$$

Thus

$$\lambda = \frac{b}{\|w\|_2^2}, \quad x^\star = -\frac{b}{\|w\|_2^2} w.$$

Finally, the minimum distance is

$$\text{dist}(0, \{x : w^\top x + b = 0\}) = \|x^\star\|_2 = \frac{|b|}{\|w\|_2}.$$

Geometrically, the minimizer lies along the direction of the normal vector w, which is why the perpendicular direction gives the shortest distance.

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

Let $H_1 : w^\top x + b_1 = 0$ and $H_2 : w^\top x + b_2 = 0$ be two parallel hyperplanes. Pick any $x_1 \in H_1$ so that $w^\top x_1 + b_1 = 0$. The perpendicular distance from x_1 to H_2 is

$$\text{dist}(x_1, H_2) = \frac{|w^\top x_1 + b_2|}{\|w\|_2}.$$

Since $w^\top x_1 + b_1 = 0$, we have

$$w^\top x_1 + b_2 = (w^\top x_1 + b_1) + (b_2 - b_1) = b_2 - b_1.$$

Thus

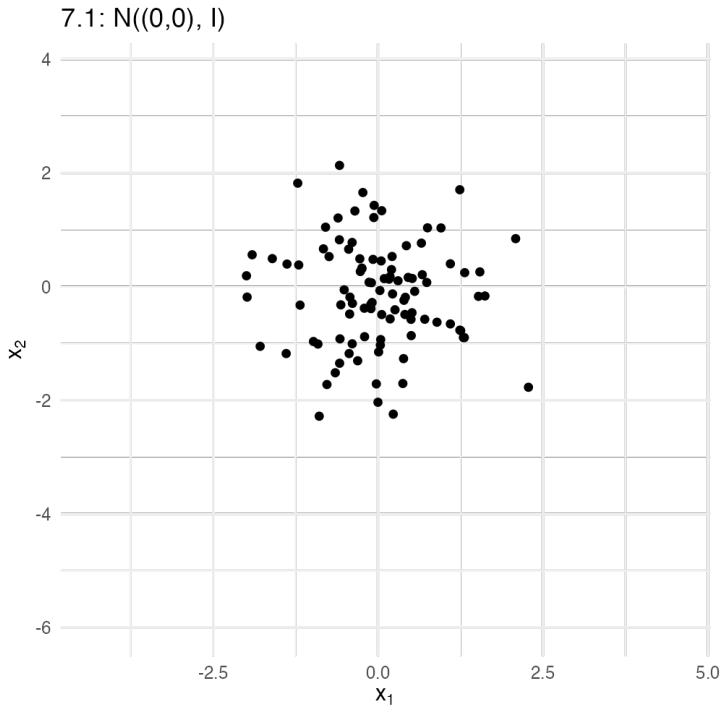
$$\text{dist}(x_1, H_2) = \frac{|b_2 - b_1|}{\|w\|_2}.$$

This distance is independent of the choice of $x_1 \in H_1$ because all points in H_1 are parallel-shifted copies of one another relative to H_2. The shortest path is along the common normal w, so this is indeed the distance between H_1 and H_2.

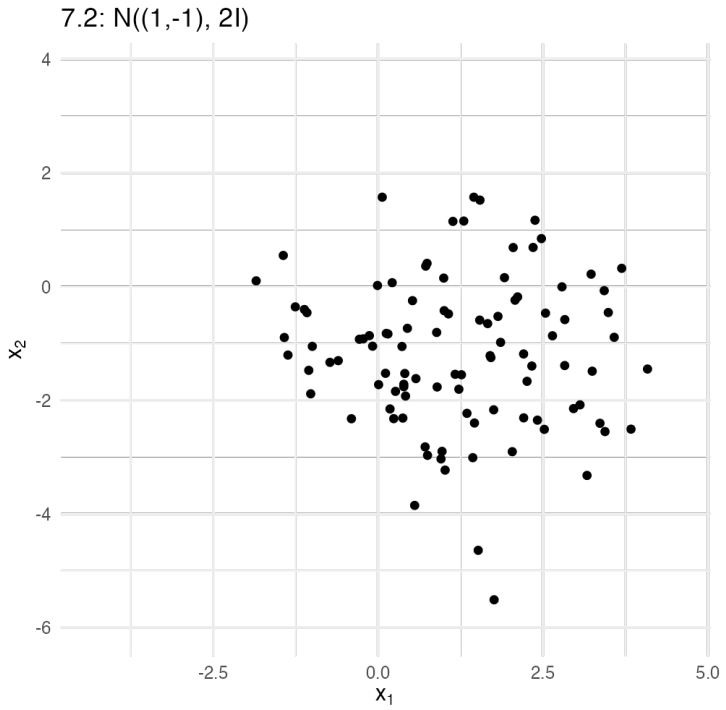
7 Programming Skills [12 pts]

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.

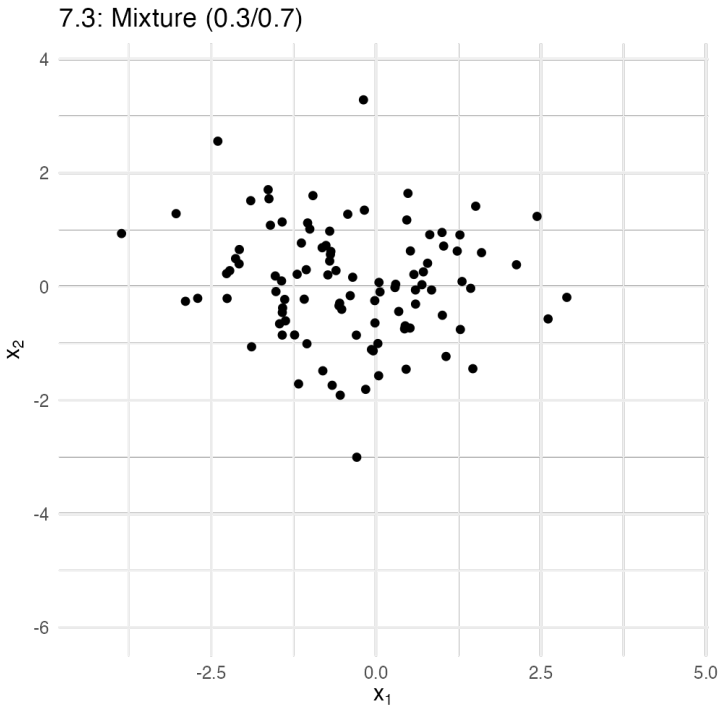
1. Draw 100 items $\mathbf{x} = [x_1, x_2]^\top$ from a 2-dimensional Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu} = (0, 0)^T$ and identity covariance matrix $\boldsymbol{\Sigma} = \mathbf{I}$, i.e., $p(\mathbf{x}) = \frac{1}{2\pi} \exp\left(-\frac{||\mathbf{x}||^2}{2}\right)$, and make a scatter plot (x_1 vs. x_2).



2. Make a scatter plot by drawing 100 items from $N(\boldsymbol{\mu} + (1, -1)^\top, 2\boldsymbol{\Sigma})$.



3. Make a scatter plot by drawing 100 items from a mixture distribution $0.3N\left((1,0)^\top, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}\right) + 0.7N\left((-1,0)^\top, \begin{pmatrix} 1 & -0.2 \\ -0.2 & 1 \end{pmatrix}\right)$.


7.3: Mixture (0.3/0.7)
4
2
0
-2
-4
-6
x_2
-2.5
0.0
2.5
5.0
x_1