Spring 2025 CS4641/CS7641 Homework 1

Dr. Mahdi Roozbahani

Deadline: Friday, February 7th, 11:59 pm EST

- No unapproved extension of the deadline is allowed. For late submissions, please refer to the course website.
- Discussion is encouraged on Ed as part of the Q/A. However, all assignments should be done individually.
- Plagiarism is a **serious offense**. You are responsible for completing your own work. You are not allowed to copy and paste, or paraphrase, or submit materials created or published by others, as if you created the materials. All materials submitted must be your own. This also means you may not submit work created by generative models as your own.
- All incidents of suspected dishonesty, plagiarism, or violations of the Georgia Tech Honor Code will be subject to the institute's Academic Integrity procedures. If we observe any (even small) similarities/plagiarisms detected by Gradescope or our TAs, **WE WILL DIRECTLY REPORT ALL CASES TO OSI**, which may, unfortunately, lead to a very harsh outcome. **Consequences can be severe, e.g., academic probation or dismissal, grade penalties, a 0 grade for assignments concerned, and prohibition from withdrawing from the class**.

Instructions

- We will be using Gradescope for submission and grading of assignments.
- **Unless a question explicitly states that no work is required to be shown, you must provide an explanation, justification, or calculation for your answer.** Basic arithmetic can be combined (it does not need to each have its own step); your work should be at a level of detail that a TA can follow it.
- Your write-up must be submitted in PDF form, you may use either Latex, markdown, or any word processing software. We will **NOT** accept handwritten work. Make sure that your work is formatted correctly, for example submit $\sum_{i=0} x_i$ instead of sum_{i=0} x_i.
- **A useful video tutorial on LaTeX has been created by our TA team** and can be found here and an Overleaf document with the commands can be found here.
- When submitting your assignment on Gradescope, **you are required to correctly map pages of your PDF to each question/ subquestion to reflect where they appear.** Improperly mapped questions will not be graded correctly.
- All assignments should be done individually, each student must write up and submit their own answers.
- **Graduate Students**: You are required to complete any sections marked as Bonus for Undergrads

Point Distribution

Q1: Linear Algebra [30pts]

- 1.1 Determinant and Inverse of a Matrix [10pts]
- 1.2 Eigenvalues and Eigenvectors [20pts]

Q2: Expectation, Co-variance and Statistical Independence [7pts]

Q3: Optimization [17pts + 3% Bonus for All]

- 3.1 KKT [17pts]
- 3.2 Primal and Dual Form [3% Bonus for All]

Q4: Maximum Likelihood [20pts: 10pts + 10 pts Grad/6% Bonus for Undergrads]

- 4.1 Discrete Example [10pts]
- 4.2 Poisson Distribution [10pts Grad / 6% Bonus for Undergrads]

Q5: Information Theory [31pts]

- 5.1 Mutual Information and Entropy [21pts]
- 5.2 Entropy Proofs [10pts]

Q6: Ethical Implications on Decision-Making [10 pts]

- 6.1 Loan Eligibility [5pts]
- 6.2 Voting and Probabilistic Models [5pts]

Q7: Programming [5pts]

Q8: Bonus Questions [7% Bonus for All]

- 8.1 Marginal Probability Density Functions [2% Bonus for All]
- 8.2 Coin Toss Game [2% Bonus for All]
- 8.3 Dice Roll Expectation [3% Bonus for All]

Points Totals:

- **Total Programming Points for All:** 5 pts
- **Total Written Points for Grad:** 115 pts
- **Total Written Points for Undergrad:** 105 pts

1 Linear Algebra [30pts]

1.1 Determinant and Inverse of Matrix [10pts]

Given a matrix M:

$$M = \begin{bmatrix} 3 & -2 & 4 \\ r & 1 & -1 \\ 0 & 2 & 2 \end{bmatrix}$$

(a) Calculate the determinant of $\boldsymbol{M}$ in terms of r (a calculation process is required). [4pts]

$$|\boldsymbol{M}| = 3 \cdot \begin{vmatrix} 1 & -1 \\ 2 & 2 \end{vmatrix} - r \cdot \begin{vmatrix} -2 & 4 \\ 2 & 2 \end{vmatrix}$$

$$\begin{vmatrix} 1 & -1 \\ 2 & 2 \end{vmatrix} = (1)(2) - (-1)(2) = 2 + 2 = 4$$

$$\begin{vmatrix} -2 & 4 \\ 2 & 2 \end{vmatrix} = (-2)(2) - (4)(2) = -4 - 8 = -12$$

$$|\boldsymbol{M}| = 3(4) - r(-12)$$

$$|\boldsymbol{M}| = 12 + 12r$$

(b) For what value(s) of r does $\boldsymbol{M}^{-1}$ not exist? Why doesn't $\boldsymbol{M}^{-1}$ exist in this case? What does it mean in terms of rank and singularity for these values of r? *This question can be answered in less than 7 lines.* [3pts]

Solution:

When the determinant is zero it means the matrix is singular and therefore is not invertible. 12 + 12r = 0

12r = -12

r = -1

(c) Find the mathematical equation that describes the relationship between the determinant of $\boldsymbol{M}$ and the determinant of $\boldsymbol{M}^{-1}$. You do **NOT** need to show any work. [3pts]

NOTE: It may be helpful to find the determinant of $\boldsymbol{M}$ and $\boldsymbol{M}^{-1}$ for $r = 0$.

Solution:

Since the determinant of M = 12
and the determinant of $M^{-1} = \frac{1}{12}$
then the relationship between these is:

$$\det(M^{-1}) = \frac{1}{\det(M)}$$

1.2 Eigenvalues and Eigenvectors [20pts]

1.2.1 Eigenvalues [5pts]

Given the following matrix $\boldsymbol{A}$, find an expression for the eigenvalues λ of $\boldsymbol{A}$ in terms of a, b, and c. (Simplify your answer into the form $\lambda = ...$). [5pts]

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

Solution:

$$\det(A - \lambda I) = 0,$$

$$A - \lambda I = \begin{bmatrix} a-\lambda & b \\ b & c-\lambda \end{bmatrix}.$$

$$\lambda = \frac{(a+c) \pm \sqrt{(a+c)^2 - 4(ac-b^2)}}{2}.$$

1.2.2 Eigenvectors [15pts]

Given a matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{bmatrix} -7 & 2 \\ 6 & 4 \end{bmatrix}$$

(a) Calculate the eigenvalues of $\boldsymbol{A}$. Simplify your answer into the form $\lambda =$ numbers [3pts]
Solution:

$$\det(A - \lambda I) = \begin{vmatrix} -7-\lambda & 2 \\ 6 & 4-\lambda \end{vmatrix}$$

$$= (-7-\lambda)(4-\lambda) - (6)(2).$$

$$= \lambda^2 + 3\lambda - 40.$$

$$\lambda = 5, -8.$$

(b) Find the normalized eigenvectors of matrix $\boldsymbol{A}$ (calculation process required). [7pts]
Solution:

$$A - 5I = \begin{bmatrix} -7-5 & 2 \\ 6 & 4-5 \end{bmatrix} = \begin{bmatrix} -12 & 2 \\ 6 & -1 \end{bmatrix}.$$

$$-12x + 2y = 0$$

$$y = 6x.$$

x = 1, y = 6

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 6 \end{bmatrix}.$$

Now we can normalize

$$\|\mathbf{v}_1\| = \sqrt{1^2 + 6^2} = \sqrt{37}.$$

$$\mathbf{v}_1 = \frac{1}{\sqrt{37}} \begin{bmatrix} 1 \\ 6 \end{bmatrix}.$$

$$A - (-8I) = \begin{bmatrix} -7+8 & 2 \\ 6 & 4+8 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 6 & 12 \end{bmatrix}.$$

$$x + 2y = 0$$

$$x = -2y.$$

y = 1, x = -2

$$\mathbf{v}_2 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}.$$

now we can normalize for v2:

$$\|\mathbf{v}_2\| = \sqrt{(-2)^2 + 1^2} = \sqrt{5}.$$

$$\mathbf{v}_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} -2 \\ 1 \end{bmatrix}.$$

(c) When calculating the eigenvectors, were the columns of the matrix $(\boldsymbol{A} - \lambda \boldsymbol{I})$ linearly independent or linearly dependent?
Now, consider the linearly independent matrix

$$\boldsymbol{B} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

Solve for the vector $\boldsymbol{x}$ which satisfies the equation $\boldsymbol{Bx} = 0$.
Hint: Use row reduction on

$$\begin{bmatrix} 1 & 2 & | & 0 \\ 2 & 1 & | & 0 \end{bmatrix}$$

Afterwards, recall that matrices can be interpreted as a transformation on a vector. For example,

$$\text{scaling} = \begin{bmatrix} k & 0 \\ 0 & k \end{bmatrix}, \text{flip across } x_2 = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

So, consider you have several vectors close to each other, and you want to apply a transformation to separate them. If you want to make sure the vectors keep their non-zero values after the transformation, what important property must the transformation matrix have [5pts]?

Solution:

For the first question since we had a non-trivial solution we know that the columns of the matrix were linearly dependent.

Since the matrix is linearly independent we know that the solution is trivial where x_1 and $x_2 = 0$. If you want to make sure that the vectors keep their non-zero values after the transformation the transformation matrix must be invertible meaning the determinant must not be zero.

2 Expectation, Co-variance, and Statistical Independence [7pts]

Suppose X, Y, and Z are three different real-valued random variables.

Let X obey a discrete binary distribution. The probability mass function for X is:

$$p(x) = \begin{cases} 0.8 & x = c \\ 0.2 & x = -c \end{cases}$$

where c is some nonzero constant. The distribution of Y is not known, but it is provided that $Var(Y) = 0.94c^2$. Additionally, X and Y are statistically independent (i.e. $P(X|Y) = P(X)$). Finally, let $Z = 9X + 3Y$.

We define a correlational measure γ:

$$\gamma(X, Z) = \frac{Cov(X, Z)}{\sqrt{Var(X) \cdot Var(Z)}} + Var(X + Z)$$

Evaluate $\gamma(X, Z)$ in terms of c. Remember to show your work to receive credit. Round the values in your final answer to 3 decimal places, but do not round in intermediate steps.

HINT: Review the probability and statistics lecture slides for relevant formulae.

We can start off by calculating the individual components in our formula.

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

$$\mathbb{E}[X] = P(X = c)c + P(X = -c)(-c) = 0.8c - 0.2c = 0.6c$$
$$\mathbb{E}[X^2] = P(X = c)c^2 + P(X = -c)c^2 = 0.8c^2 + 0.2c^2 = c^2$$
$$\text{Var}(X) = c^2 - (0.6c)^2 = c^2 - 0.36c^2 = 0.64c^2$$

$$\text{Var}(Z) = a^2\text{Var}(X) + b^2\text{Var}(Y), \quad Z = aX + bY, \quad a = 9,\, b = 3$$
$$\text{Var}(Z) = 9^2\text{Var}(X) + 3^2\text{Var}(Y)$$
$$\text{Var}(Z) = 81(0.64c^2) + 9(0.94c^2)$$
$$\text{Var}(Z) = 51.84c^2 + 8.46c^2 = 60.3c^2$$

$$Z = 9X + 3Y \implies \text{Cov}(X, Z) = \text{Cov}(X, 9X + 3Y)$$
$$\text{Cov}(X, Z) = 9\text{Cov}(X, X) + 3\text{Cov}(X, Y)$$

Since X and Y are independent, $\text{Cov}(X, Y) = 0$, so:

$$\text{Cov}(X, Z) = 9\text{Var}(X) = 9(0.64c^2) = 5.76c^2$$

$$\text{Var}(X + Z) = \text{Var}(X) + \text{Var}(Z) + 2\text{Cov}(X, Z)$$
$$\text{Var}(X + Z) = 0.64c^2 + 60.3c^2 + 2(5.76c^2)$$
$$\text{Var}(X + Z) = 0.64c^2 + 60.3c^2 + 11.52c^2 = 72.46c^2$$

$$\gamma(X, Z) = \frac{\text{Cov}(X, Z)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Z)}} + \text{Var}(X + Z)$$

Substitute the values:

$$\frac{\text{Cov}(X, Z)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Z)}} = \frac{5.76c^2}{\sqrt{(0.64c^2)(60.3c^2)}} = \frac{5.76c^2}{\sqrt{38.592c^4}}$$

$$= \frac{5.76c^2}{6.2182c^2} = \frac{5.76}{6.2182} \approx 0.9272$$

For $\mathrm{Var}(X + Z)$:

$$\mathrm{Var}(X + Z) = 72.46c^2$$

$$\gamma(X, Z) = 0.9272 + 72.46c^2$$

$$\gamma(X, Z) = 0.9272 + 72.46c^2$$

3 Optimization [17pts + 3% Bonus for All]

3.1 KKT [17pts]

Optimization problems are related to minimizing a function (usually termed loss, cost or error function) or maximizing a function (such as the likelihood) with respect to some variable x. The Karush-Kuhn-Tucker (KKT) conditions are first-order conditions for a solution in nonlinear programming to be optimal, provided that some regularity conditions are satisfied.

In this question, you will be solving the following optimization problem. In this problem, you are tasked with helping professor Mahdi optimise the type and number of GPU's to buy. You must balance cost and availability. You have the total compute defined by function f(x,y), and the constraints with respect to cost and availability.

$$\begin{aligned} \max_{x,y} \quad & f(x,y) = 7x^2 + 4y^2 \\ \text{s.t.} \quad & g_1(x,y) = 15x + 6y \leq 250 \\ & g_2(x,y) = x \leq 8 \end{aligned}$$

(a) Write the Lagrange function for the maximization problem. Now change the maximum function to a minimum function (i.e. $\underset{x,y}{min} \quad f(x,y) = 7x^2 + 4y^2$) and provide the Lagrange function for the minimization problem with the same constraints g_1 and g_2. [2pts]
NOTE: The minimization problem is only for part (a).

(a) Write the Lagrange function for the maximization problem. Now change the maximum function to a minimum function (i.e. $\min_{x,y} \; f(x,y) = 7x^2 + 4y^2$) and provide the Lagrange function for the minimization problem with the same constraints g_1 and g_2. [2pts]
NOTE: The minimization problem is only for part (a).

$$\mathcal{L}(x,y) = f(x,y) - \lambda_1 g_1 - \lambda_2 g_2$$
$$\mathcal{L}(x,y) = 7x^2 + 4y^2 - \lambda_1(15x + 6y - 250) - \lambda_2(x - 8) \quad \text{(maximization)}$$
$$\mathcal{L}(x,y) = 7x^2 + 4y^2 + \lambda_1(15x + 6y - 250) + \lambda_2(x - 8) \quad \text{(minimization)}$$

(b) List the names of all 4 groups of KKT conditions and their corresponding mathematical equations or inequalities for this specific maximization problem. Be sure to simplify completely, and calculate the derivative. [2pts]
Solution:
KKT Conditions:

(a) **Stationary Condition:**

$$\frac{\partial \mathcal{L}}{\partial x}, \frac{\partial \mathcal{L}}{\partial y}$$
$$\frac{\partial \mathcal{L}}{\partial x} = 14x - 15\lambda_1 - \lambda_2 = 0 \quad \rightarrow \quad (1)$$
$$\frac{\partial \mathcal{L}}{\partial y} = 8y - 6\lambda_1 = 0 \quad \rightarrow \quad (2)$$

(b) **Complementary Slackness:**

$$\lambda_1(15x + 6y - 250) = 0 \quad \rightarrow \quad (3)$$
$$\lambda_2(x - 8) = 0 \quad \rightarrow \quad (4)$$

(c) **Primal Feasibility:**

$$15x + 6y - 250 \leq 0 \quad \rightarrow \quad (5)$$
$$x - 8 \leq 0 \quad \rightarrow \quad (6)$$

(d) **Dual Feasibility:**

$$\lambda_1, \lambda_2 \geq 0 \quad \rightarrow \quad (7)$$

(c) Solve for 4 possibilities formed by each constraint being active or inactive. Do not forget to check the inactive constraints for each point when applicable. Candidate points must satisfy all the conditions mentioned in part b) (Quick note: If a constraint is binding its corresponding lambda should be greater than or equal to zero). [8pts]

(c) 4 possibilities formed by constraints:

(a) **Case 1:** $g_1 \rightarrow \text{Active} \rightarrow \lambda_1 \geq 0$
$g_2 \rightarrow \text{Active} \rightarrow \lambda_2 \geq 0$
$x - 8 \leq 0 \rightarrow x = 8$
Substituting $x = 8$ into (5):
$15x + 6y - 250 = 0$
$15(8) + 6y - 250 = 0$
$120 + 6y - 250 = 0$
$6y = 130 \rightarrow y = \frac{65}{3}$
Candidate Point: $(8, \frac{65}{3})$
Now solve for λ_1 and λ_2:
From (2): $8y - 6\lambda_1 = 0 \rightarrow \lambda_1 = \frac{4}{3}y$
$\lambda_1 = \frac{4}{3} \cdot \frac{65}{3} = \frac{260}{9}$
From (1): $14x - 15\lambda_1 - \lambda_2 = 0$
$14(8) - 15 \cdot \frac{260}{9} - \lambda_2 = 0$
$112 - \frac{1300}{3} - \lambda_2 = 0$
$\lambda_2 = -321.33$ (Not Feasible)

(b) **Case 2:** $g_1 \rightarrow \text{Active} \rightarrow \lambda_1 \geq 0$
$g_2 \rightarrow \text{Inactive} \rightarrow \lambda_2 = 0$
Substituting $x = 8$ into (5):
$15x + 6y - 250 = 0$
$15(13.0208) + 6y - 250 = 0$
Solving x and y, we get:
$(x, y) = (13.0208, 9.1145)$
Now solve for λ_1 and λ_2:
From (2): $8y - 6\lambda_1 = 0 \rightarrow \lambda_1 = \frac{8y}{6} = \frac{4}{3}y$
From (1): $\lambda_2 = 0$ (Feasible)
Not a Candidate Point: $(13.0208, 9.1145)$
since x = 18.0208 ¿ 8

(c) **Case 3:** $g_1 \rightarrow \text{Inactive} \rightarrow \lambda_1 = 0$
$g_2 \rightarrow \text{Active} \rightarrow \lambda_2 \geq 0$
Substituting $x = 8$:
$14x - 15\lambda_1 - \lambda_2 = 0$
$8y - 6\lambda_1 = 0$
Solving: $(x, y) = (8, 0)$
Candidate Point: $(8, 0)$

(d) **Case 4:** Both g_1 and $g_2 \rightarrow$ Inactive:
$\lambda_1 = 0, \lambda_2 = 0$
Candidate Point: $(0, 0)$

(d) List the candidate point(s) (there is at least 1) obtained from part c). Please round answers to 3 decimal points and use that answer for calculations in further parts. This part can be completed in one line per candidate point. [2pts]
Final Candidate Points: $(8, 0), (0, 0)$

(e) Find the **one** candidate point for which $f(x, y)$ is largest. Check if $L(x, y)$ is concave, convex, or neither at this point by using the Hessian in the second partial derivative test. [3pts]

HINT: Read the Example_optimization_problem.pdf in Canvas Files for HW1 to see an example with some explanations.

HINT: Watch this video walking you through how to solve a similar problem.

(e) One candidate point for which $f(x,y)$ is the largest:

$$f(x,y) = 7x^2 + 4y^2$$

$$(8,0) \quad \rightarrow \quad f(8,0) = 448 \quad \text{(Largest)}$$
$$(0,0) \quad \rightarrow \quad f(0,0) = 0$$

Hessian:

$$f(x,y) = 7x^2 + 4y^2$$
$$|H| = \begin{vmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial y \partial x} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{vmatrix} = \begin{bmatrix} 14 & 0 \\ 0 & 8 \end{bmatrix}$$

Eigenvalues:

$$\lambda_1 = 14, \quad \lambda_2 = 8 \quad \text{(both } > 0)$$

Conclusion:
Since all eigenvalues are positive, $f(x,y)$ is **convex**.

3.2 Primal and Dual Form [3% Bonus for All]

Convex optimization problems involve minimizing a function given a constraint. The Lagrangian function includes a penalty for violating this constraint. A maximum is taken over the penalty because this is the opposite of what we are trying to accomplish which is minimization.

Under certain conditions, which are satisfied in the following problem, maximizing a variable followed by minimizing over another variable is equivalent to minimizing over the latter followed by maximizing over the former. In literature, this is referred to as transforming a problem from the primal form into the dual form.

For the following problem, write out the primal form, switch it to the dual form, and then solve for the pair (x, y).
NOTE: The following video does a great job at visualizing this concept. Additionally, for linear constraints, there is no "inactive" state for the constraint.

$$\begin{aligned} \min_{x,y} \quad & f(x,y) = x^2 + y^2 \\ \text{s.t.} \quad & g_1(x,y) = x + y = 4 \end{aligned}$$

4 Maximum Likelihood [20pts: 10pts + 10pts Grad / 6% Bonus for Undergrads]

4.1 Discrete Example [10pts]

Mastermind Mahdi decides to give a challenge to his students for their MLE Final. He provides a spinner with 10 sections, each numbered 1 through 10. The students can change the sizes of each section, meaning that they can select the probability the spinner lands on a certain section. Mahdi then proposes that the students will get a 100 on their final if they can spin the spinner 10 times such that it doesn't land on section 1 during the first 9 spins and lands on section 1 on the 10th spin. If the probability of the spinner landing on section 1 is θ, what value of θ should the students select to most likely ensure they get a 100 on their final? Use your knowledge of Maximum Likelihood Estimation to get a 100 on the final.

NOTE: You must specify the log-likelihood function and use MLE to solve this problem for full credit. You may assume that the log-likelihood function is concave for this question

2025-1Spring/Solution/spinner_10.png

Solution:

- 10 sections: $1 \to 10$
- Students will get 100 if they spin 9 times without landing on section 1 but land on section 1 on the 10th spin.
- $P(s_1) = \theta \quad$ (probability of landing on section 1)
- $P(\neg s_1) = 1 - \theta \quad$ (probability of not landing on section 1)
- $P(\text{correct sequence}) = (1-\theta)^9 \cdot \theta$

Likelihood Function

$$L(\theta) = (1-\theta)^9 \cdot \theta$$

Log-Likelihood Function

$$\ell(\theta) = \log L(\theta) = 9\log(1-\theta) + \log(\theta)$$

Maximizing the Log-Likelihood Function

Maximize by taking the partial derivative with respect to θ:

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{-9}{1-\theta} + \frac{1}{\theta} = 0$$

Solve for θ:

$$\frac{1}{\theta} = \frac{9}{1-\theta} \quad \to \quad 1-\theta = 9\theta \quad \to \quad 1 = 10\theta \quad \to \quad \theta = \frac{1}{10}$$

Solve for MLE

$$L(\theta) = \left(1 - \frac{1}{10}\right)^9 \cdot \frac{1}{10} = \left(\frac{9}{10}\right)^9 \cdot \frac{1}{10}$$

4.2 Normal distribution [10 pts Grad / 6% Bonus for Undergrads]

The Normal distribution is defined as:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

(a) Let $(X_1,\dots,X_n) \sim \mathcal{N}(\mu,\sigma^2)$, where $X_1,\dots,X_n$ are i.i.d. random variables, and let $x_1,\dots,x_n$ be the observed values of $X_1,\dots,X_n$. What is the likelihood of (μ,σ^2) given this data? Express your answer in product form. [4 pts / 2%]
Solution:

$x_1, x_2, \dots, x_n$ is normally distributed and independent of each other.

$$f(x \mid \mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Likelihood function:

$$L(\mu,\sigma^2 \mid x_1,x_2,\dots,x_n) = \prod_{i=1}^{n} f(x_i \mid \mu,\sigma^2)$$

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$L(\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$L(\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right)$$

(b) What are the maximum likelihood estimators (MLEs) for μ and σ^2 (Hint the MLEs of σ^2 is in terms of μ) ? [6 pts / 4%]
Solution:
Use log-likelihood function:

$$\ell(\mu,\sigma^2) = \log L(\mu,\sigma^2)$$

$$\ell(\mu,\sigma^2) = \log\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right)\right)$$

$$\ell(\mu,\sigma^2) = n\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Simplify:

$$\ell(\mu,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Differentiate with respect to μ:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

Set to zero:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \quad \to \quad \sum_{i=1}^{n} x_i - n\mu = 0 \quad \to \quad \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Thus equivalent to the sample mean formula:

$$\mu = \bar{x}$$

Differentiate with respect to σ^2:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

Set to zero:

$$-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

Solve for σ^2:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

5 Information Theory [31pts]

5.1 Mutual Information and Entropy [21pts]

A recent study has shown symptomatic infections are responsible for higher transmission rates. Using the data collected from positively tested patients, we wish to determine which feature(s) have the greatest impact on whether or not someone will present with symptoms. To do this, we will compute the entropies, conditional entropies, and mutual information of select features. Please use base 2 when computing logarithms.

ID	**Vaccine Doses** (X_1)	**Wears Mask?** (X_2)	**Underlying Conditions** (X_3)	**Symptomatic** (Y)
1	L	T	F	F
2	M	F	F	T
3	L	F	F	F
4	H	T	F	F
5	L	F	T	T
6	H	F	T	T
7	L	T	T	F
8	M	F	F	T
9	H	T	T	F
10	M	T	F	F

Table 1: Vaccine Doses: {(H) booster, (M) 2 doses, (L) 1 dose, (T) True, (F) False}

(a) Find entropy $H(Y)$ to at least 3 decimal places. [3pts]
Solution:

$$\begin{aligned}
(a)\quad H(Y) &= -\sum P(y_i)\log_2 P(y_i) \\
Y_{\text{true}} = 4 \quad &\Rightarrow \quad \frac{4}{10} \\
Y_{\text{false}} = 6 \quad &\Rightarrow \quad \frac{6}{10} \\
P(Y = \text{true}) = 0.4, &\quad P(Y = \text{false}) = 0.6 \\
H(Y) &= -\left[0.4\log_2(0.4) + 0.6\log_2(0.6)\right] \\
H(Y) &= 0.971
\end{aligned}$$

(b) Find the average conditional entropy $H(Y|X_1)$ and $H(Y|X_2)$ to at least 3 decimal places. [7pts]
Solution:

$$H(Y|X) = \sum P(x_i) H(Y|x_i)$$

$$H(Y|X):$$

$$\begin{aligned} x = L: & \\ P(x = L) &= 0.4 \\ P(x = M) &= 0.3 \\ P(x = H) &= 0.3 \end{aligned}$$

$$\begin{aligned} P(Y = T|L) &= 0.25 \\ P(Y = F|L) &= 0.75 \end{aligned}$$

$$\begin{aligned} H(Y|L) &= -\left[0.25 \log_2(0.25) + 0.75 \log_2(0.75)\right] \\ &= 0.811128 \end{aligned}$$

$$\begin{aligned} x = M: & \\ P(Y = T|M) &= \frac{2}{3}, \quad P(Y = F|M) = \frac{1}{3} \end{aligned}$$

$$\begin{aligned} H(Y|M) &= -\left[\frac{2}{3} \log_2\left(\frac{2}{3}\right) + \frac{1}{3} \log_2\left(\frac{1}{3}\right)\right] \\ &= 0.918 \end{aligned}$$

$$x = H: \quad Y = F, T, F = M \quad = 0.918$$

$$\begin{aligned} H(Y|X) &= P(L)H(Y|L) + P(M)H(Y|M) + P(H)H(Y|H) \\ &= (0.4)(0.811) + (0.3)(0.918) + (0.3)(0.918) \\ &= 0.875 \end{aligned}$$

$H(Y|X_2):$

$$P(X_2 = T) = P(X_2 = F) = 0.5$$

$$\begin{aligned} X_2 &= T: \\ P(Y = F|T) = 1, &\quad P(Y = T|T) = 0 \\ H(Y|T) &= -[1\log_2(1)] = 0 \end{aligned}$$

$$\begin{aligned} X_2 &= F: \\ P(Y = T|F) &= \frac{4}{5} = 0.8 \\ P(Y = F|F) &= \frac{1}{5} = 0.2 \end{aligned}$$

$$\begin{aligned} H(Y|F) &= -[0.8\log_2(0.8) + 0.2\log_2(0.2)] \\ &= 0.7220 \end{aligned}$$

$$\begin{aligned} H(Y|X_2) &= (0.5)(0) + (0.5)(0.7220) \\ &= 0.3610 \end{aligned}$$

$$H(Y|X_1) = 0.875, \quad H(Y|X_2) = 0.361$$

(c) Find mutual information $I(X_1, Y)$ and $I(X_2, Y)$ to at least 3 decimal places and determine which one (X_1 or X_2) is more informative. [3pts]
Solution:

$$I(X_1, Y) = H(Y) - H(Y|X_1)$$

We already solved these values, so we can plug in:

$$\begin{aligned} H(Y) &= 0.971 \\ H(Y|X_1) &= 0.875 \\ H(Y|X_2) &= 0.361 \end{aligned}$$

$$I(X_1, Y) = (0.971) - (0.875) = 0.096$$

Now compute $I(X_2, Y):$

$$I(X_2, Y) = (0.971) - (0.361) = 0.610$$

Since $I(X_2, Y) > I(X_1, Y), \quad X_2$ is more informative.

(d) Find joint entropy $H(Y, X_3)$ to at least 3 decimal places. [3pts]
Solution:

$$
\begin{aligned}
P(Y=F, X_3=F) &= 0.4, \quad P(Y=T, X_3=F) = 0.2 \\
P(Y=F, X_3=T) &= 0.2, \quad P(Y=T, X_3=T) = 0.2
\end{aligned}
$$

$$
\begin{aligned}
H(Y|X_3) &= -\sum P(y_i, x_3) \log_2 \left(P(y_i, x_3)\right) \\
&= -\Big(0.4 \log_2(0.4) + 0.2 \log_2(0.2) + 0.2 \log_2(0.2) + 0.2 \log_2(0.2)\Big) \\
&= 1.922
\end{aligned}
$$

(e) Find the conditional entropy $H(Y|X_1, X_2)$. [5 pts]
Solution:

$$
\begin{aligned}
L, T &= \frac{2}{10}, \quad H(Y|L,T) = 0 \\
L, F &= \frac{2}{10}, \quad H(Y|L,F) = -\left[0.5 \log_2(0.5) + 0.5 \log_2(0.5)\right] = 1 \\
M, T &= \frac{1}{10}, \quad H(Y|M,T) = 0 \\
M, F &= \frac{2}{10}, \quad H(Y|M,F) = 0 \\
H, T &= \frac{2}{10}, \quad H(Y|H,T) = 0 \\
H, F &= \frac{1}{10}, \quad H(Y|H,F) = 0 \\
H(Y|X_1, X_2) &= 0.2
\end{aligned}
$$

5.2 Entropy Proofs [10pts]

Given the mathematical definition of $H(X)$ and $H(X|Y)$ below, prove that $I(X,Y) = 0$ if X and Y are statistically independent. (Note: you must provide a mathematical proof and cannot use the visualization shown in class found here. You may use any theorem/ proof from the slides without having to re-prove it). [10pts]

$$H(X) = -\sum_x P(x)\log_2 P(x)$$

$$H(X|Y) = -\sum_{x,y} P(x,y)\log_2 \frac{P(x,y)}{P(y)}$$

Start from: $I(X,Y) = H(X) - H(X|Y)$

Solution:

$$\begin{aligned}
I(X,Y) &= H(X) - H(X|Y) \\
H(X) &= -\sum P(x)\log_2 P(x) \\
H(X|Y) &= -\sum P(x|y)\log_2 P(x|y) \\
&= -\sum P(x|y)\log_2 \frac{P(x,y)}{P(y)} \\
\text{Since } X \text{ and } Y \text{ are independent:}\quad P(x,y) &= P(x)P(y) \\
H(X|Y) &= -\sum P(x)P(y)\log_2 \frac{P(x)P(y)}{P(y)} \\
&= -\sum P(x)P(y)\log_2 P(x) \\
H(X|Y) &= -\sum \left(P(x)\log_2 P(x)\sum P(y)\right), \quad \sum P(y) = 1 \\
H(X|Y) &= -\sum P(x)\log_2 P(x), \quad \text{which is } H(X) \\
\text{So:}\quad H(X|Y) &= H(X)
\end{aligned}$$

Now solve:

$$\begin{aligned}
I(X,Y) &= H(X) - H(X|Y) \\
&= H(X) - H(X) \\
&= 0
\end{aligned}$$

$\therefore$ This proves that X and Y are statistically independent.

6 Ethical Implications on Decision-Making [10 pts]

6.1 Loan Eligibility [5pts]

Real-world Implications

Loan eligibility determines who can receive a loan, typically based on financial history and demographics. It is a difficult problem, and often uses algorithms to make loan decisions. Often, this can result in reinforcing inequality and bias [**oneil**].

Suppose we're using a matrix to represent the attributes of individuals for loan approval. Each attribute (like income, credit score, years of employment, etc.) constitutes a column in our matrix. Here's a hypothetical toy example:

	Annual Income	Debt-to-Income Ratio	Employment History (years)	Credit Score
Candidate 1	50,000	0.2	5	700
Candidate 2	51,000	0.21	5.1	710
Candidate 3	45,000	0.19	4.9	690
Candidate 4	100,000	0.05	10	780

One algorithm used to predict credit score is linear regression, formulated as $\mathbf{y} = \mathbf{x}\mathbf{A}$. $\mathbf{y}$ are the target variables, $\mathbf{x}$ are the input features, and $\mathbf{A}$ is a matrix trained with an existing dataset. Training data $(\mathbf{x_D}, \mathbf{y_D})$ are taken from the training dataset D, $(\mathbf{x_D}, \mathbf{y_D}) \in D$. If $\mathbf{x_D}$ is linearly independent, $\mathbf{A}$ can be trained by simply inverting $\mathbf{x_D}$:

$$\begin{aligned} \mathbf{y_D} &= \mathbf{x_D}\mathbf{A} \\ \mathbf{x_D}^{-1}\mathbf{y_D} &= \mathbf{A} \end{aligned}$$

The original equation can be rewritten as:

$$\begin{aligned} \mathbf{y} &= \mathbf{x}\mathbf{A} \\ &= \mathbf{x}\mathbf{x_D}^{-1}\mathbf{y_D} \end{aligned}$$

Problems arise when the training data is close to linearly dependent. Recall that one way to invert a matrix is $\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\operatorname{adj}(\mathbf{A})$. As $\mathbf{A}$ becomes more linearly dependent and $\det(\mathbf{A}) \to 0$, $||\mathbf{A}^{-1}||$ can become so large it causes numerical errors. Rewriting the original equation:

$$\begin{aligned} \mathbf{y} &= \mathbf{x}\mathbf{x_D}^{-1}\mathbf{y_D} \\ &= \frac{1}{\det(\mathbf{x_D})}\mathbf{x}\operatorname{adj}(\mathbf{x_D})\mathbf{y_D} \end{aligned}$$

The errors caused by $\det(\mathbf{x_D}) \to 0$ propagate to $\mathbf{y}$, causing predictions to be wildly inaccurate anywhere outside of the original training set.

Practical Implications

1. Instability: With a small determinant, minor variations in the attributes can lead to significant variations in the results. So, a small difference in income might result in a disproportionate change in loan eligibility.
2. Poor Generalization: If the matrix is based on data with limited variation (like our small community example), it's essentially trained on a very narrow subset of potential applicants. If someone from outside this narrow subset applies (e.g., a person with a 2-year employment but a $70,000 income), the system may not process their application fairly or accurately because it's unfamiliar with such profiles.

Given that a matrix used for determining loan approvals has a determinant close to zero due to limited variation in applicants' attributes:
Which of the following implications might this have on the decision-making process? Choose all options that apply.

A) It ensures a more uniform scoring system since most applicants have similar attributes.

B) It can lead to unpredictable scores, where tiny variations in attributes yield vastly different outcomes.

C) The system is more resilient to errors because of the limited attribute variation.

D) It might not generalize well to broader populations, potentially leading to biases when applied to more diverse applicant groups.

Answer Here:
Solution:
B, D

6.2 Voting and Probabilistic Models [5pts]

A country uses a probabilistic model to predict election outcomes and determine where resources (such as campaign funding or polling stations) should be allocated. The model uses factors like historical voting patterns, demographic data, voter turnout rates, and regional economic indicators to predict the probability of a particular candidate or party winning in each region.

However, the data used to train the model is incomplete: it primarily reflects urban areas, leaving rural voting behaviors underrepresented; some ethnic and socioeconomic groups have historically low voter turnout rates, meaning their data is sparsely included; and the model relies on historical data that may not accurately reflect recent political, economic, or social changes.

This results in several key issues. The model prioritizes campaign resources and polling stations in regions with high probabilities of voter influence, often favoring well-represented demographics, while rural or underrepresented regions may receive fewer polling stations, making it harder for individuals in those areas to vote. If polling stations are removed from low-priority areas due to the model's predictions, it could discourage voting in already marginalized communities, further disenfranchising groups with historically low turnout and perpetuating cycles of political exclusion.

Which of the following is an ethical way to address the issue of bias in a probabilistic model used for allocating voting resources?

A) Collect and incorporate diverse and representative data to ensure the model accounts for voting patterns across all regions, demographics, and socioeconomic groups.

B) Prioritize resource allocation only in regions with historically high voter turnout, as these areas are statistically more likely to impact election outcomes.

C) Introduce fairness constraints in the model to guarantee equitable resource distribution, ensuring underrepresented regions and groups are not disadvantaged.

D) Exclude regions with low voter turnout from the model's predictions, as their voting behavior is less predictable and may skew results.

Answer Here:
Solution:
A,C

7 Programming [5 pts]

See the Programming subfolder in Canvas.

8 Bonus Questions [7% Bonus for All]

8.1 Marginal Probability Density Functions (2% Bonus for All)

Suppose that X and Y have joint pdf given by

$$f_{X,Y}(x,y) = \begin{cases} 2e^{-2y}, & 0 \leq x \leq 1, y \geq 0 \\ 0, & otherwise \end{cases}$$

What are the marginal probability density functions for X and Y? [5 pts]

8.2 Coin Toss Game (2% Bonus for All)

A person decides to toss a biased coin with $P(heads) = \frac{1}{3}$ repeatedly until he gets a head. He will make at most 6 tosses. Let the random variable Y denote the number of heads. Find the pmf of Y. Then, find the variance of Y. Round your answer to 3 decimal places. *(It is possible to thoroughly support your answer to this question in 5 to 10 lines)* [poin total here]

8.3 Dice Roll Expectation (3% Bonus for All)

Suppose you roll an 8-sided die. For each roll:

- You are paid the face value of the roll.
- If the roll gives $Y \in \{1, 2, 3, 4, 5\}$, the game stops.
- If the roll gives $Y \in \{6, 7, 8\}$, you can roll again.

The probabilities are uniform, and the payoff structure is:

- For $Y \in \{1, 2, 3, 4, 5\}$, the expected value of the payoff is 3.
- For $Y \in \{6, 7, 8\}$, the expected value of the payoff is 7 plus the extra roll's expected value.

What is the expected payoff for this game? [5pts]