# Fall 2024 CS4641/CS7641 Homework 1

Dr. Mahdi Roozbahani

Deadline: Friday, September 20th, 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 [28pts]

- 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: 17pts + 2% 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 [26pts]

- 5.1 Mutual Information and Entropy [16pts]
- 5.2 Entropy Proofs [10pts]

## Q6: Ethical Implications on Decision-Making [5 pts]

## Q7: Programming [5pts]

## Q8: Bonus for All [8%]

## Points Totals:

- **Total Base:** 100 pts
- **Total Undergrad Bonus:** 6%
- **Total Bonus for All:** 10%
- **Total Possible Assignment Grade (Undergrad):** 116%
- **Total Possible Assignment Grade (Grad):** 110%

# 1 Linear Algebra [10pts + 18pts]

## 1.1 Determinant and Inverse of Matrix [10pts]

Given a matrix $M$:

$$M = \begin{bmatrix} 3 & 1 & 4 \\ r & 2 & -4 \\ 0 & -3 & 5 \end{bmatrix}$$

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

$$|\boldsymbol{M}| = 3((2*5) - (-4*-3)) - r((1*5) - (-3*4)) + 0 = -6 - 17r$$

(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]

The inverse cannot exist when $|\boldsymbol{M}| = 0$ so setting the determinant to 0 and solving for r gives a value of $-\frac{6}{17}$. When r is this value and the determinant is 0 this means that matrix has no inverse. The matrix is singular because it has no inverse so it loses its full rank of 3 because the columns or rows are linearly dependent.

(c) Find the mathematical equation that describes the relationship between the determinant of $\boldsymbol{M}$ and the determinant of $\boldsymbol{M}^{-1}$. [3pts]

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

$$M = \begin{bmatrix} 3 & 1 & 4 \\ 0 & 2 & -4 \\ 0 & -3 & 5 \end{bmatrix}$$

Then

$$|\boldsymbol{M}| = 3((2*5) - (-4*-3)) - 0 + 0 = -6$$

$$M^{-1} = \begin{bmatrix} \frac{1}{3} & \frac{17}{6} & 2 \\ 0 & -\frac{5}{2} & -2 \\ 0 & -\frac{3}{2} & -1 \end{bmatrix}$$

Then

$$|\boldsymbol{M}| = \frac{1}{3}((-\frac{5}{2}*-1) - (-2*-\frac{3}{2})) - 0 + 0 = -\frac{1}{6}$$

Finally

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

## 1.2 Eigenvalues and Eigenvectors [5+15pts]

### 1.2.1 Eigenvalues [5pts]

Given the following matrix $\boldsymbol{A}$, find an expression for the eigenvalues $\lambda$ of $\boldsymbol{A}$ in terms of $a$, $b$, and $c$. [5pts]

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

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$det(A - \lambda I) = det(\begin{bmatrix} a-\lambda & b \\ b & c-\lambda \end{bmatrix}) = 0$$

$$(a-\lambda)(c-\lambda) - b^2 = 0$$
$$\lambda^2 - (a+c)\lambda + (ac - b^2) = 0$$

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

### 1.2.2 Eigenvectors [15pts]

Given a matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{bmatrix} 11 & 4 \\ 4 & 5 \end{bmatrix}$$

(a) Calculate the eigenvalues of $\boldsymbol{A}$. [3pts]

Using the above formula:

$$\lambda_1 = \frac{(11+5)+\sqrt{(11+5)^2-4(11*5-4^2)}}{2} \text{ and } \lambda_2 = \frac{(11+5)-\sqrt{(11+5)^2-4(11*5-4^2)}}{2}$$

$$\lambda_1 = \frac{(16)+\sqrt{(16)^2-4(55-4^2)}}{2} \text{ and } \lambda_2 = \frac{(16)-\sqrt{(16)^2-4(55-4^2)}}{2}$$

$$\lambda_1 = 13 \text{ and } \lambda_2 = 3$$

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

For $\lambda_1$ to get the eigenvector we solve the equation $(A - 13I)x = 0$

$$\begin{bmatrix} 11-13 & 4 \\ 4 & 5-13 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

This system gives the solution $x_1 = 2x_2$

The eigenvector for $\lambda_1$ is $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$

We then find the magnitude to normalize $\sqrt{2^2+1^2} = \sqrt{5}$. So the normalized eigenvector is $\begin{bmatrix} \frac{2\sqrt{5}}{5} \\ \frac{\sqrt{5}}{5} \end{bmatrix}$

For $\lambda_2$ to get the eigenvector we solve the equation $(A - 3I)x = 0$

$$\begin{bmatrix} 11-3 & 4 \\ 4 & 5-3 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

This system gives the solution $x_1 = -\frac{1}{2}x_2$

The eigenvector for $\lambda_1$ is $\begin{bmatrix} -\frac{1}{2} \\ 1 \end{bmatrix}$

We then find the magnitude to normalize $\sqrt{(-\frac{1}{2})^2 + 1^2} = \frac{\sqrt{5}}{2}$. So the normalized eigenvector is $\begin{bmatrix} -\frac{\sqrt{5}}{5} \\ \frac{2\sqrt{5}}{5} \end{bmatrix}$

(c) If done correctly, the normalized eigenvectors from part (b) and the matrix $(\boldsymbol{A} - \lambda \boldsymbol{I})$ are both nonzero. Despite both being nonzero, we still have $(\boldsymbol{A} - \lambda \boldsymbol{I})x = 0$ (where $x$ is an eigenvector). What are some properties of the matrix $(\boldsymbol{A} - \lambda \boldsymbol{I})$ which allow for this? Additionally, why is the determinant $|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0$? [5pts]

The $(\boldsymbol{A} - \lambda \boldsymbol{I})x = 0$ comes from the property that $\boldsymbol{A}x = \lambda \boldsymbol{I}$. From this we get that the $(\boldsymbol{A} - \lambda \boldsymbol{I})$ is singular because it transforms x into a zero vector for a certain eigenvalue $\lambda$. When the $\lambda$ is an eigenvalue of the matrix $A$ that means $(\boldsymbol{A} - \lambda \boldsymbol{I})$ has a null space making it possible for $(\boldsymbol{A} - \lambda \boldsymbol{I})x = 0$. This $|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0$ is true because a determinant represents whether the matrix can be inverted. Since the matrix is singular for each eigenvalue it means that the matrix $(\boldsymbol{A} - \lambda \boldsymbol{I})$ does not have full rank. Since there is an eigenvector in the null space of $(\boldsymbol{A} - \lambda \boldsymbol{I})$ that means the determinant of $(\boldsymbol{A} - \lambda \boldsymbol{I})$ is 0. Property of eigenvalue and eigenvector.

**NOTE:** There are many ways to solve this problem. You are allowed to use linear algebra properties as part of your solution.

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

Suppose $X$, $Y$, and $Z$ are three different random variables. Let $X$ obey a two point Distribution. The probability mass function for $X$ is:

$$p(x) = \begin{cases} 0.9 & x = c \\ 0.1 & x = -c \end{cases}$$

where $c$ is a nonzero constant. The distribution of $Y$ is not known, but it is provided $Var(Y) = 1.44c^2$. $X$ and $Y$ are statistically independent (i.e. $P(X|Y) = P(X)$). Meanwhile, let $Z = 4X + 2Y$.

Calculate the correlation coefficient defined as $\rho(X, Z) = \frac{Cov(X,Z)}{\sqrt{Var(X)Var(Z)}}$. Round your answer to 3 decimal places or simplified radical form.

**HINT:** Review the probability and statistics lecture slides

$$Var(X) = E[X^2] - (E[X])^2$$

$$E[X]) = (0.9c) + (-0.1c) = 0.8c$$

$$E[X^2] = (0.9c^2) + (0.1(-c)^2) = c^2$$

$$Var(X) = c^2 - (0.8c)^2 = 0.36c^2$$

$$Var(Z) = Var(4X + 2Y) = 16Var(X) + 4Var(Y)$$

$$Var(Z) = 16 * 0.36c^2 + 4 * 1.44c^2 = 2 * 5.76c^2 = 11.52c^2$$

$$Cov(X, Z) = Cov(X, 4X + 2Y) = 4Cov(X, X) + 2Cov(X, Y)$$
$$Cov(X, Z) = 4Var(X) = 4 * 0.36c^2 = 1.44c^2 \text{ due to independence of X and Y}$$

$$\rho(X, Z) = \frac{Cov(X,Z)}{\sqrt{Var(X)Var(Z)}} = \frac{1.44c^2}{\sqrt{0.36c^2 * 11.52c^2}}$$

$$\rho(X, Z) = \frac{1.44}{\sqrt{0.36 * 11.52}} = 0.707$$

# 3 Optimization [17pts + 2% Bonus for All]

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:

$$\begin{aligned} \max_{x,y} \quad & f(x,y) = 2x + 3xy \\ \text{s.t.} \quad & g_1(x,y) = x^2 + 4y^2 \leq 9 \\ & g_2(x,y) = y \leq \frac{1}{2} \end{aligned}$$

(a) Write the Lagrange function for the maximization problem. Now change the maximum function to a minimum function (i.e. $\underset{x,y}{min}\ f(x,y) = 2x+3xy$) 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).

Lagrange for Maximization:

$$L(x,y,\lambda) = 2x + 3xy - \lambda_1(x^2 + 4y^2 - 9) - \lambda_2(y - 0.5)$$

Lagrange for Minimization:

$$L(x,y,\lambda) = 2x + 3xy + \lambda_1(x^2 + 4y^2 - 9) + \lambda_2(y - 0.5)$$

(b) List the names of all 4 groups of KKT conditions and their corresponding mathematical equations or inequalities for this specific maximization problem. [2pts]

Stationarity Condition

$$\frac{\partial L}{\partial x} = 2 + 3y - \lambda_1(2x) = 0$$

$$\frac{\partial L}{\partial y} = 3x - 8\lambda_1(y) - \lambda_2 = 0$$

Primal Feasibility:

$$g_1(x,y) = x^2 + 4y^2 - 9 \leq 0$$

$$g_2(x,y) = x - 0.5 \leq 0$$

Dual Feasibility:

$$\lambda_1 \geq 0$$

$$\lambda_2 \geq 0$$

Complementary Slackness:

$$\lambda_1(x^2 + 4y^2 - 9) = 0$$

$$\lambda_2(y - 0.5) = 0$$

(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). [8pts]

binding for $g_1$ and $g_2$

$$x^2 + 4y^2 - 9 = 0$$

$y - 0.5 = 0$ means $y = 0.5$
$(x)^2 + 4 * (0.5)^2 - 9 = 0$ and $x = \mp\sqrt{8}$

$$(\mp\sqrt{8}, 0.5)$$

satisfies the Complementary Slackness and Primal Feasibility for $(x, y)$

For the Stationarity Condition when $x = \sqrt{8}$:

$$2 + 3(0.5) - \lambda_1(2 * \sqrt{8}) = 0$$

$$\lambda_1 = \frac{3.5}{2\sqrt{8}} > 0$$

$$2 + 3(0.5) - \frac{3.5}{2\sqrt{8}}(2 * \sqrt{8}) = 0$$

$$3\sqrt{8} - 8\lambda_1(0.5) - \lambda_2 = 0$$

$$\lambda_2 = 6.01 > 0$$

Dual Feasibility is satisfied so when $(x, y) = (\sqrt{8}, 0.5)$ it is a candidate point
For the Stationarity Condition when $x = -\sqrt{8}$:

$$2 + 3(0.5) - \lambda_1(2 * -\sqrt{8}) =$$

$$\lambda_1 = -\frac{3.5}{2\sqrt{8}} < 0$$

Dual Feasibility is not satisfied because of $\lambda_1$

inactive for $g_1$ and $g_2$
$x^2 + 4y^2 - 9 < 0$ and $\lambda_1 = 0$ for Complementary Slackness
$y - 0.5 < 0$ and $\lambda_2 = 0$ for Complementary Slackness

For the Stationarity Condition:

$$2 + 3(y) - 0(x) = 0$$

$$y = -\frac{2}{3}$$

For the second Stationarity condition:

$$3x - 8 * 0(y) - 0 = 0$$

$$x = 0$$

Using the point $(0, -\frac{2}{3})$ for $g_2 = -\frac{2}{3} - \frac{1}{2} < 0$ so Primal Feasibility and Complementary Slackness are satisfied.

$$2 + 3(-\frac{2}{3}) - \lambda_1(2 * 0) = 0$$

$$\lambda_1 = 0 \leq 0$$

$$3 * 0 - 8 * 0 * (-\frac{2}{3}) - \lambda_2 = 0$$

$$\lambda_2 = 0 \leq 0$$

Dual Feasibility is satisfied so $(0, -\frac{2}{3})$ is a candidate point

(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]

Candidate Point is $(\sqrt{8}, 0.5)$

(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]

(f) **BONUS FOR ALL:** Make a 3D plot of the objective function $f(x, y)$ and constraints $g_1$ and $g_2$ using Math3d. Mark the maximum candidate point and include a screenshot of your plot. Briefly explain why your plot makes sense in one sentence. Although this is bonus, this is **VERY HELPFUL** in understanding what was accomplished in this problem. [2%]

**NOTE:** Use an explicit surface for the objective function, implicit surfaces for the constraints, and a point for the minimum candidate point.

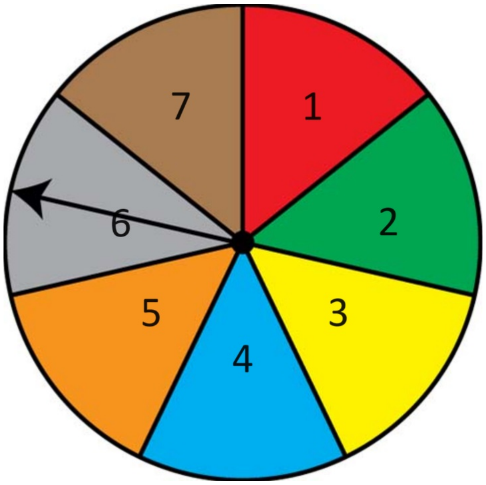
**HINT:** Read the Example_optimization_problem.pdf in Canvas Files for HW1 to see an example with some explanations.
**HINT:** Click here for a video explaining the intuition behind KKT problems.
**HINT:** Click here for an example maximization problem. It's recommended to only watch up until 23:14.
**HINT:** Click here to determine how to set up the problem for minimization in part (a) and for KKT conditions in part (b).

# 4 Maximum Likelihood [10pts + 10pts Grads / 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 7 sections, each numbered 1 through 7. 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 7 times such that it doesn't land on section 1 during the first 6 spins and lands on section 1 on the 7th spin. If the probability of the spinner landing on section 1 is $\theta$, what value of $\theta$ 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



$$f(\theta) = (1-\theta)^6 * \theta$$

$$\log(f(\theta)) = log((1-\theta)^6 * \theta) = \log(1-\theta)^6 + \log(\theta) = 6\log(1-\theta) + \log(\theta)$$

$$(log(f(\theta)))' = 6 * \frac{-1}{(1-\theta)(ln(5))} + \frac{1}{\theta ln(5)}$$

$$(log(f(\theta)))' = -\frac{6}{(1-\theta)(ln(5))} + \frac{1}{\theta ln(5)}$$

$$0 = -\tfrac{6}{(1-\theta)(ln(5))} + \tfrac{1}{\theta ln(5)}$$

$$\frac{6}{(1-\theta)(ln(5))} = \frac{1}{\theta ln(5)}$$

$$6\theta ln(5) = (1-\theta)(ln(5))$$

$$6\theta = 1-\theta$$

$$\theta = \frac{1}{7}$$

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

The Poisson distribution is defined as:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} (k = 0, 1, 2, ...).$$

(a) Let $X_1 \sim Poisson(\lambda)$. What is the likelihood of $\lambda$ given $x_1$ is an observed value of $X_1$ ?[2 pts / 1%]

$$L(X_1|\lambda) = \tfrac{\lambda^{X_1} * e^{-\lambda}}{X_1!}$$

(b) Now, assume we are given $n$ such values. Let $(X_1, ..., X_n) \sim Poisson(\lambda)$ where $X_1, ..., X_n$ are i.i.d. random variables, and $x_1, ..., x_n$ be observed values of $X_1, ..., X_n$. What is the likelihood of $\lambda$ given this data? You may leave your answer in product form. [2 pts / 1%]

$$L(X_1, ..., X_n|\lambda) = \textstyle\prod_{i=1}^{n} \tfrac{\lambda^{X_i} * e^{-\lambda}}{X_i!}$$

(c) What is the maximum likelihood estimator of $\lambda$? [6 pts / 4%]

$$log(L(X_1, ..., X_n|\lambda)) = log(\prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!})$$

$$ln(L(X_1, ..., X_n|\lambda)) = ln(\prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!})$$

$$ln(L(X_1, ..., X_n|\lambda)) = \sum_{i=1}^{n} ln(\frac{\lambda^{X_i} e^{-\lambda}}{X_i!}) = \sum_{i=1}^{n} ln(\lambda^{X_i}) + ln(e^{-\lambda}) - ln(X_i!)$$

$$ln(L(X_1, ..., X_n|\lambda)) = \sum_{i=1}^{n} X_i ln(\lambda) - \lambda ln(e) - ln(X_i!) = \sum_{i=1}^{n} X_i ln(\lambda) - \lambda - ln(X_i!)$$

$$ln(L(X_1, ..., X_n|\lambda)) = -n\lambda + ln(\lambda) \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} ln(X_i!) = l(\lambda)$$

$$\frac{\partial l}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i = 0$$

$$\frac{1}{\lambda} \sum_{i=1}^{n} X_i = n$$

$$\lambda = \frac{1}{n} \sum_{i=1}^{n} X_i$$

# 5 Information Theory [16pts + 10pts]

## 5.1 Mutual Information and Entropy [16pts]

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 some 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 | L | F | T | T |
| 3 | L | F | F | F |
| 4 | H | T | F | F |
| 5 | L | F | T | T |
| 6 | H | F | T | T |
| 7 | L | F | T | F |
| 8 | M | F | F | T |
| 9 | H | T | F | T |
| 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]

$$H(Y) = \sum_{k}^{K} P(Y = k) * \log_2(\frac{1}{P(Y = k)})$$

$$P(Y = T) = 0.5$$

$$P(Y = F) = 0.5$$

$$H(Y) = 0.5 * \log_2(\frac{1}{0.5}) + 0.5 * \log_2(\frac{1}{0.5}) = 1$$

$$H(Y) = 1$$

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

| Y<br>$X_1$ | T | F |
|---|---|---|
| H | $\frac{2}{10}$ | $\frac{1}{10}$ |
| M | $\frac{1}{10}$ | $\frac{1}{10}$ |
| L | $\frac{3}{10}$ | $\frac{2}{10}$ |

Joint probability distribution of $X_1$ and $Y$

$$H(Y|X) = \sum_{x \in X, y \in Y} p(x, y) * \log_2(\frac{p(x)}{p(x, y)})$$

$$H(Y|X_1) = \frac{2}{10}\log_2(\frac{\frac{3}{10}}{\frac{2}{10}}) + \frac{1}{10}\log_2(\frac{\frac{3}{10}}{\frac{1}{10}}) + \frac{2}{10}\log_2(\frac{\frac{2}{10}}{\frac{1}{10}}) + \frac{3}{10}\log_2(\frac{\frac{5}{10}}{\frac{3}{10}}) + \frac{2}{10}\log_2(\frac{\frac{5}{10}}{\frac{2}{10}}) = 0.977$$

| Y<br>$X_2$ | T | F |
|---|---|---|
| T | $\frac{1}{10}$ | $\frac{3}{10}$ |
| F | $\frac{4}{10}$ | $\frac{2}{10}$ |

Joint probability distribution of $X_2$ and $Y$

$$H(Y|X_2) = \frac{1}{10}\log_2(\frac{\frac{4}{10}}{\frac{1}{10}}) + \frac{3}{10}\log_2(\frac{\frac{4}{10}}{\frac{3}{10}}) + \frac{2}{10}\log_2(\frac{\frac{6}{10}}{\frac{2}{10}}) + \frac{4}{10}\log_2(\frac{\frac{6}{10}}{\frac{4}{10}}) = 0.895$$

(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]

$$I(X_1, Y) = H(Y) - H(Y|X_1) = 1 - 0.977 = 0.023$$

$$I(X_2, Y) = H(Y) - H(Y|X_2) = 1 - 0.895 = 0.105$$

$X_2$ has higher mutual information with Y so it is more informative

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

| Y<br>$X_3$ | T | F |
|---|---|---|
| T | $\frac{3}{10}$ | $\frac{1}{10}$ |
| F | $\frac{2}{10}$ | $\frac{4}{10}$ |

Joint probability distribution of $X_3$ and $Y$

$$H(Y|X_3) = \frac{3}{10}\log_2(\frac{\frac{4}{10}}{\frac{3}{10}}) + \frac{1}{10}\log_2(\frac{\frac{4}{10}}{\frac{1}{10}}) + \frac{2}{10}\log_2(\frac{\frac{6}{10}}{\frac{2}{10}}) + \frac{4}{10}\log_2(\frac{\frac{6}{10}}{\frac{4}{10}}) = 0.895$$

$$H(X_3) = (0.4)\log_2(\frac{1}{0.4}) + (0.6)\log_2(\frac{1}{0.6}) = 0.971$$

$$H(Y, X_3) = H(Y|X_3) + H(X_3) = 0.895 + 0.971 = 1.866$$

### 5.2 Entropy Proofs [10pts]

(a) Write the discrete case mathematical definition for $H(X|Y)$ and $H(X)$. [3pts]

$$H(X|Y) = \sum_{y \in Y} p(y) * H(X|Y = y) = \sum_{y \in Y, x \in X} p(x,y) \log(\frac{p(y)}{p(x,y)})$$

$$H(X) = -\sum_{k=1}^{K} P(x = k) * \log P(x = k) = \sum_{x \in X} p(x) \log(\frac{1}{p(x)})$$

(b) **Using the mathematical definition of $H(X)$ and $H(X|Y)$ from part (a)**, 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). [7pts]

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

Assume I(X, Y) = 0:

$$H(X) - H(X|Y) = 0$$

$$H(X|Y) = H(X)$$

I(X, Y) = 0 is true after we prove that $H(X|Y) = H(X)$:

$$\sum_{y \in Y, x \in X} p(x,y) * \log(\frac{p(y)}{p(x,y)}) = \sum_{x \in X} p(x) * \log(\frac{1}{p(x)})$$

Due to the X and Y being independent p(x, y) = p(x)p(y) we can rewrite the equation

$$\sum_{y \in Y, x \in X} p(x) * p(y) * \log(\frac{p(y)}{p(x)p(y)}) = \sum_{x \in X} p(x) * \log(\frac{1}{p(x)})$$

$$\sum_{y \in Y, x \in X} p(x) * p(y) * \log(\frac{1}{p(x)}) = \sum_{x \in X} p(x) * \log(\frac{1}{p(x)})$$

$$\sum_{y \in Y} \sum_{x \in X} p(x) * p(y) * \log(\frac{1}{p(x)}) = \sum_{x \in X} p(x) * \log(\frac{1}{p(x)})$$

Since y is not dependent on x we can rewrite the equation

$$\sum_{y \in Y} p(y) * \sum_{x \in X} p(x) * \log(\frac{1}{p(x)}) = \sum_{x \in X} p(x) * \log(\frac{1}{p(x)})$$

$$\sum_{x \in X} p(x) * \log(\frac{1}{p(x)}) = \sum_{x \in X} p(x) * \log(\frac{1}{p(x)})$$

$$\sum_{x \in X} p(x) * \log(\frac{1}{p(x)}) = \sum_{x \in X} p(x) * \log(\frac{1}{p(x)})$$

H(X—Y) = H(X)

Finally we get $H(X) - H(X|Y) = 0$ with $I(X,Y) = 0$

# 6 Ethical Implications on Decision-Making [5 pts]

## 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 [1].

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 $\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\text{adj}(\mathbf{A})$. 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:

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

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 as all options that apply. Use "textbf{}" to select your answer.*

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:**

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

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

# 7 Programming [5 pts]

See the Programming subfolder in Canvas.

## 8 Bonus for All [8%]

(a) Let $X$, $Y$ be **two statistically independent** $N(0, 1)$ random variables, and $P$, $Q$ be random variables defined as:

$$P = 2X + 5XY^2$$
$$Q = X$$

Calculate the variance $Var(P + Q)$. *(This question may take substantial work to support, e.g. 25 to 30 lines)* [4%]

**HINT:** The following equality may be useful: $Var(XY) = E[X^2Y^2] - [E(XY)]^2$
**HINT:** $E[Y^4] = \int_{-\infty}^{\infty} y^4 f_Y(y)dy$ where $f_Y(y)$ is the probability density function of $Y$ (Wolfram alpha calculator or other similar calculators can be used)
**HINT:** $Var(P + Q) = Var(P) + Var(Q) + 2Cov(P, Q)$ may be a good starting point.

$$Var(P + Q) = Var(P) + Var(Q) + 2Cov(P, Q)$$

$$Var(P) = E[P^2] - E[P]^2$$
$$E[P] = E[2X + 5XY^2] = 2E[X] + 5E[X]E[Y] = 2(0) + 5(0)(0) = 0$$
$$E[P^2] = E[(2X + 5XY^2)^2] = E[4X^2 + 20X^2Y^2 + 25X^2Y]$$
$= 4E[X^2] + 20E[X^2Y^2] + 25E[X^2Y]$
$$= 4(Var(X) + E[X]^2) + 20(Var(X) + E[X]^2)(Var(Y) + E[Y]^2) + 25E[E[X]^2]E[Y]$$
$$= 4(1) + 20(1)(1) + 25E[X^2](0) = 24$$
$$Var(P) = 24 - 0 = 24$$
$$Var(Q) = Var(X) = E[Q^2] - E[Q]^2 = E[X^2] - E[X]^2 = 1$$
$$Cov(P, Q) = E[PQ] - E[P]E[Q]$$
$$E[P, Q] = E[2X^2 + 5X^2Y^2] = 2E[X^2] + 5E[X^2]E[Y^2] = 2(1) + 5(1)(1) = 7$$
$$Cov(P, Q) = 7 - (0)(0) = 7$$
$$Var(P + Q) = 24 + 1 + 2(7) = 39$$

(b) Suppose that $X$ and $Y$ have joint pdf given by:

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{24}xe^{-\frac{1}{3}y} & 0 \leq x \leq 4, y \geq 0 \\ 0 & otherwise \end{cases}$$

What are the marginal probability density functions for $X$ and $Y$? *(It is possible to thoroughly support your answer to this question in 8 to 10 lines)* [2%]

$$f_X(x) = \text{Start your answer here.}$$

(c) A person decides to toss a biased coin with $P(heads) = 0.25$ 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 probability distribution 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)* [2%]

## References

[1] Cathy O'Neil. *Weapons of Math Destruction.* Penguin Books, 2017.