Fall 2023 CS4641/CS7641 Homework 1

Dr. Mahdi Roozbahani

Deadline: Friday, September 22nd, 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.
- Please answer each question on a new page. It makes it more organized to map your answers on GradeScope. When submitting your assignment, 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 [38pts]

- 1.1 Determinant and Inverse of a Matrix [15pts]
- 1.2 Characteristic Equation [8pts]
- 1.3 Eigenvalues and Eigenvectors [15pts]

Q2: Expectation, Co-variance and Statistical Independence [9pts]

- 2.1 Covariance [5pts]
- 2.2 Correlation [4pts]

Q3: Optimization [19pts: 15pts + 4pts Bonus for All]

Q4: Maximum Likelihood [25pts: 10pts + 15pts Bonus for Undergrads]

- 4.1 Discrete Example [10pts]
- 4.2 Poisson Distribution [15pts Bonus for Undergrads]

Q5: Information Theory [32pts]

- 5.1 Marginal Distribution [6pts]
- 5.2 Mutual Information and Entropy [16pts]
- 5.3 Entropy Proofs [10pts]

Q6: Programming [5pts]

Q7: Bonus for All [20pts]

1 Linear Algebra [15pts + 8pts + 15pts]

1.1 Determinant and Inverse of Matrix [15pts]

Given a matrix M:

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

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

$$|\boldsymbol{M}| = -2r + 14$$

Solution: $-2r + 28$

Using diagonal method, write this matrix as so:

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

this gives us:
$(10r - 8 + 15) - (12r - 25 + 4) = 10r + 7 - 12r + 21 = -2r + 28$

(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? [3pts]

Solution: r = 14

For the inverse of a matrix to not have an inverse, it must be a singular matrix, meaning it has a determinant of 0. This also means that the rank of the matrix is smaller than n since the columns are not independent, which in this case is 3.

(c) Will all values of r found in part (b) allow for a column (or row) to be expressed as a linear combination of the other columns (or rows)? **If yes**, provide

- **either** the linear equation of the third column $\boldsymbol{C_3}$ as a linear combination of the first column $\boldsymbol{C_1}$ and second column $\boldsymbol{C_2}$
- **or** the linear equation of the second row $\boldsymbol{R_2}$ as a linear combination of the first row $\boldsymbol{R_1}$ and third row $\boldsymbol{R_3}$.

If no, explain why. [3pts]

Start here: $C_3 = \frac{a}{b}C_1 + \frac{c}{d}C_2$, or $R_2 = \frac{a}{b}R_1 + \frac{c}{d}R_3$.

Solution: $R_2 = 3 * R_1 + 4 * R_3$

solve this as a system:
5x-4y = -1
-2x + 5y = 14
3x - 2y = 1
where x is what we multiply R_1 by and y is what we multiply R_3 by
this gives us x=3 and y = 4, meaning our solution is:
$R_2 = 3 * R_1 + 4 * R_3$

(d) Write down $\boldsymbol{M}^{-1}$ for $r = 0$ (calculation process is **NOT** required). [2pts]

$$M^{-1} = \begin{bmatrix} \frac{5}{28} & \frac{-11}{28} & \frac{-1}{14} \\ \frac{3}{14} & \frac{-1}{14} & \frac{-2}{7} \\ \frac{5}{28} & \frac{17}{28} & \frac{-1}{14} \end{bmatrix}$$

(e) 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 $r = 0$.

Solution: They are inverses of each other.

$|M| = 28$
$|M|^{-1} = \frac{1}{28}$
we know that this is true because the inverse of the determinant is used to calculate the inverse

1.2 Characteristic Equation [8pts]

Consider the eigenvalue problem:

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}, \boldsymbol{x} \neq 0$$

where $\boldsymbol{x}$ is a non-zero eigenvector and λ is an eigenvalue of $\boldsymbol{A}$. Prove the determinant $|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0$.

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

Solution: Let us start with the equation $\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$. Start by subtracting $\lambda \boldsymbol{x}$ over to the left to get $\boldsymbol{A}\boldsymbol{x} - \lambda \boldsymbol{x} = 0$
We want to factor out x, but to do this, we must write this equation as $\boldsymbol{A}\boldsymbol{x} - \lambda \boldsymbol{I}\boldsymbol{x} = 0$. This is because in order for the subtraction of A - I to be possible, they both must be matrices of the same dimensions, so we multiply the eigenvector by the identity matrix to get a matrix of the same dimensions.
Now we can factor out the vector x: $(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{x} = 0$
With this new formula lets consider the scenario that $(\boldsymbol{A} - \lambda \boldsymbol{I})$ is a non-singular matrix, aka the inverse exists. This is a problem because thiss means that the only solution is the trivial solution when we multiply the inverse: $(\boldsymbol{A} - \lambda \boldsymbol{I})^{-1}(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{x} = (\boldsymbol{A} - \lambda \boldsymbol{I})^{-1}0$, which becomes $x = 0$. This is a problem because x should be a non-zero eigenvector, yet we will always get the trivial solution as a solution if the inverse of $\boldsymbol{A} - \lambda \boldsymbol{I}$ exists. So the only way to make sure that x is not the zero vector is if $\boldsymbol{A} - \lambda \boldsymbol{I}$ does not have an inverse, which by the property of inverses means that the determinant of $\boldsymbol{A} - \lambda \boldsymbol{I}$ is 0.

1.3 Eigenvalues and Eigenvectors [5+10pts]

1.3.1 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. [5pts]

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

Solution: $\frac{-(-a-c)\pm\sqrt{(-a-c)^2-4*(-b^2+ac)}}{2}$ to find eigenvalues, you must find:

$$|\begin{bmatrix} a & b \\ b & c \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}|$$

which is:

$$|\begin{bmatrix} a-\lambda & b \\ b & c-\lambda \end{bmatrix}|$$

which is $(a-\lambda)(c-\lambda)-b^2$
which is $\lambda^2 - a\lambda - c\lambda - b^2 + ac$ when expanded
plug this into the quadratic formula:
$\frac{-(-a-c)\pm\sqrt{(-a-c)^2-4*(-b^2+ac)}}{2}$

1.3.2 Eigenvectors [10pts]

Given a matrix $\boldsymbol{A}$:

$$\boldsymbol{A} = \begin{bmatrix} x & 4 \\ 16 & x \end{bmatrix}$$

(a) Calculate the eigenvalues of $\boldsymbol{A}$ as a function of x (calculation process required). [3pts]

Solution: $\lambda_1 = x - 8$ and $\lambda_2 = x + 8$

Our solution to get eigenvalues is:

$$|\begin{bmatrix} x-\lambda & 4 \\ 16 & x-\lambda \end{bmatrix}|$$

Which gives us a formula of:

$$(x-\lambda)^2 - 64 = 0$$

$$(x-\lambda)^2 = 64$$

$$(x-\lambda) = \pm 8$$

$$\lambda_1 = x - 8, \lambda_2 = x + 8$$

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

Solution:

$$\begin{bmatrix} \frac{1}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} \end{bmatrix}$$

,

$$\begin{bmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix}$$

we essentially want to solve for $\boldsymbol{AX} = \lambda \boldsymbol{x}$ Using our work from part A, we can solve using the eigenvalues and then normalize the vectors:

$\lambda_1 = x - 8$:

$$\begin{bmatrix} x & 4 \\ 16 & x \end{bmatrix} * \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = (x-8) * \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

This can be written as the following system:

$$v_1 x + 4v_2 = (x-8)v_1$$

$$16v_1 + v_2 x = (x-8)v_2$$

Solving this sytem gives us the following relationship:
$-2v_1 = v_2$, giving us an eigenvector of:

$$\begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

and then normalize by dividing the magnitude:

$$\begin{bmatrix} \frac{1}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} \end{bmatrix}$$

$\lambda_1 = x + 8$:

$$\begin{bmatrix} x & 4 \\ 16 & x \end{bmatrix} * \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = (x+8) * \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

This can be written as the following system:

$$v_1 x + 4v_2 = (x+8)v_1$$

$$16v_1 + v_2 x = (x+8)v_2$$

Solving this sytem gives us the following relationship:

$2v_1 = v_2$, giving us an eigenvector of:

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

and then normalize by dividing the magnitude:

$$\begin{bmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{bmatrix}$$

2 Expectation, Co-variance and Statistical Independence [5pts + 4pts]

2.1 Covariance [5pts]

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

$$p(x) = \begin{cases} 0.3 & x = c \\ 0.7 & x = -c \end{cases}$$

where c is a nonzero constant. Let Y obey the Standard Normal (Gaussian) Distribution, which can be written as $Y \sim N(0,1)$. X and Y are statistically independent (i.e. $P(X|Y) = P(X)$). Meanwhile, let $Z = X - Y$.

Calculate the covariance of Y and Z (i.e. $Cov(Y, Z)$). Do values of c affect the covariance between Y and Z? [5pts]

Solution: -Var(Y) = -1

$$\begin{aligned}
Cov(Y, Z) &= E(YZ) - E[Y] * E[Z] \\
&= E(Y(X - Y)) - E[Y] * E[X - Y] \\
&= E(YX - Y^2) - E[Y] * E[X - Y] \\
&= E[YX] - E[Y^2] - E[Y] * (E[X] - E[Y]) \\
&= E[Y] * E[X] - E[Y^2] - E[Y] * E[X] + E[Y]^2 \\
&= -E[Y^2] + E[Y]^2 \\
&= -(E[Y^2] - E[Y]^2) \\
&= -var(Y)
\end{aligned}$$

Since Y is a standard Normal Gaussian, $-var(y) = -1$

Since covariance is only affected by scalar quantities, values of c do not make a difference.

2.2 Correlation Coefficient [4pts]

Let X and Y be statistically independent random variables with $Var(X) = 4$ and $Var(Y) = 10$. We do not know $E[X]$ or $E[Y]$. Let $Z = 2X + 9Y$. Calculate the correlation coefficient defined as $\rho(X,Z) = \frac{Cov(X,Z)}{\sqrt{Var(X)Var(Z)}}$. If applicable, please round your answer to 3 decimal places. [4pts]

Solution: 0.139

$$cov(X,Z) = cov(X, 2X + 9Y)$$
$$= cov(2X, X) + cov(9Y, X)$$
$$= 2 * cov(X, X) + 9 * cov(Y, X)$$
$$= 2 * var(x) + 9 * cov(Y, X)$$
$$= 2 * var(x) + 9 * cov(X, Y)$$

Since the covariance of two independent random variables is 0:

$$= 2 * var(x)$$
$$cov(X,Z) = 2 * 4$$

Now we need var(z):

$$var(Z) = var(2X + 9Y)$$
$$= var(2X) + var(9Y)$$
$$= 4var(X) + 81var(Y)$$
$$var(z) = 4 * 4 + 81 * 10$$

Plug the values we found into our formula:

$$\frac{8}{\sqrt{4 * (4 * 4 + 81 * 10)}} = 0.139$$

3 Optimization [15pts + 4pts 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:

$$\begin{aligned} \min_{x,y} \quad & f(x,y) = xy - 4x \\ \text{s.t.} \quad & g_1(x,y) = x^2 + 3y^2 \leq 18 \\ & g_2(x,y) = y \leq 1 \end{aligned}$$

(a) Write the Lagrange function for the minimization problem. Now change the minimum function to a maximum function (i.e. $\underset{x,y}{max} \;\; f(x,y) = xy - 4x$) and provide the Lagrange function for the maximization problem with the same constraints g_1 and g_2. [2pts]

NOTE: The maximization problem is only for part (a).

> **Solution:** min: $L(x, y, \lambda_1, \lambda_2) = xy - 4x + \lambda_1(x^2 + 3y^2 - 18) + \lambda_2(y - 1)$
> max: $L(x, y, \lambda_1, \lambda_2) = -(xy - 4x) - \lambda_1(x^2 + 3y^2 - 18) - \lambda_2(y - 1)$
> To maximize a legrage function, we make $f(x, y)$ negative so that the lambda functions help maximize and not reduce the problem.

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

> **Solution:** These are all partial derivatives:
>
> (a) Stationary Condition:
> $\frac{\partial L}{\partial y} = x + 6y\lambda_1 + \lambda_2 = 0$
> $\frac{\partial L}{\partial x} = y - 4 + 2x\lambda_1 = 0$
> $\frac{\partial L}{\lambda_1} = (x^2 + 3y^2 - 18) = 0$
> $\frac{\partial L}{\lambda_2} = (y - 1) = 0$
>
> (b) Primal Feasibility
> $g_1 \leq 18$
> $g_2 \leq 1$
>
> (c) Dual Feasibility
> $\lambda_1, \lambda_2 > 0$
>
> (d) Complementary slackness
> $\lambda_1(x^2 + 3y^2 - 18) = 0$
> $\lambda_2(y - 1) = 0$

(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). [7pts]

Solution: The cases are combinations of lambdas being active/inactive

CASE 1: λ_1, λ_2 are active

$$x^2 + 3y^2 = 18, \lambda_1 > 0, y = 1, \lambda_2 > 0$$

$$x^2 + 3 = 18$$

$$x^2 = 15$$

$$x = \pm\sqrt{15}$$

Plug this x value into the partial derivative of the legrange equation with respect to x to solve for λ_1

$$y - 4 + 2x\lambda_1 = 0$$

$$1 - 4 + 2 * \sqrt{15}\lambda_1 = 0$$

$$2\sqrt{15}\lambda_1 = 3$$

$$\lambda_1 = \frac{3}{2\sqrt{15}}$$

Plug this into the second partial derivative to get λ_2

$$x + 6y\lambda_1 + \lambda_2 = 0$$

$$\sqrt{15} + 6 * 1 * \frac{3}{2\sqrt{15}} + \lambda_2 = 0$$

$$\sqrt{15} + \frac{9}{\sqrt{15}} + \lambda_2 = 0$$

$$\lambda_2 = -\sqrt{15} - \frac{9}{\sqrt{15}}$$

Since lambda cant be negative, we test the other value of x: $-\sqrt{15}$

$$y - 4 + 2x\lambda_1 = 0$$

$$1 - 4 - 2\sqrt{15}\lambda_1 = 0$$

$$-3 - 2\sqrt{15}\lambda_1 = 0$$

$$-2\sqrt{15}\lambda_1 = 3$$

$$\lambda_1 = \frac{3}{-2\sqrt{15}}$$

Since lambda cant be negative, we know this cannot be case 1

CASE 2: λ_1 is active and λ_2 is not

$$x^2 + 3y^2 = 18, \lambda_1 > 0, y < 1, \lambda_2 = 0$$

Since λ_2 is 0, these are the equations for this case:

$$\frac{\partial L}{\partial y} = x + 6y\lambda_1 = 0$$
$$\frac{\partial L}{\partial x} = y - 4 + 2x\lambda_1 = 0$$
$$\frac{\partial L}{\lambda_1} = x^2 + 3y^2 - 18 = 0$$

Start by solving for x and y:

Put the first equation in terms of λ_1

$$x + 6y\lambda_1 = 0$$

$$\lambda_1 = \frac{-x}{6y}$$

Put the second equation in terms of λ_2

$$y - 4 + 2x\lambda_1 = 0$$

$$2x\lambda_1 = 4 - y$$

$$\lambda_1 = \frac{4 - y}{2x}$$

set the λ_1 equal to each other and solve for x:

$$\frac{-x}{6y} = \frac{4 - y}{2x}$$

$$-2x^2 = 24y - 6y^2$$
$$x^2 = -12y + 3y^2$$

Plug this into the third equation:

$$-12y + 3y^2 + 3y^2 - 18 = 0$$

$$6y^2 - 12y - 18 = 0$$

This gives us y = 3 or y = -1

Since our constraint says $y < 1$, we know y = -1(using y = 3 also yields a negative λ_1 value)

With y = -1, we can solve back for x in the third equation:

$$x^2 + 3 - 18 = 0$$

$$x^2 - 15 = 0$$

$$x = \pm\sqrt{15}$$

With $x = \sqrt{15}$ and y = -1, we can solve for λ_1

$$\sqrt{15} + 6 * -1 * \lambda_1 = 0$$

$$\sqrt{15} - 6\lambda_1 = 0$$

$$\lambda_1 = \frac{\sqrt{15}}{6}$$

CASE 3: λ_1 is active and λ_2 is not

$$x^2 + 3y^2 < 18, \lambda_1 = 0, y = 1, \lambda_2 > 0$$

Since λ_1 is zero, these are the equations for this case:

$x + \lambda_2 = 0$

$y - 4 = 0$

$y - 1 = 0$

This case is not possible since we get y = 4 and y = 1, which is impossible to solve this system with.

CASE 4: λ_1 is not active and λ_2 is not.

$$x^2 + 3y^2 < 18, \lambda_1 = 0, y < 1, \lambda_2 = 0$$

Since λ_1 and λ_2 are zero, these are the equations for this case:

$x = 0$

$y - 4 = 0$

$x^2 + 3y^2 - 18 = 0$

$y - 1 = 0$ This case is again not possible because we get values of x = 0, y = 1, or y = 4, and no combination of those values gives a valid system of equations.

(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]

Solution: There is one candidate point: ($\sqrt{15}$, -1) = (3.873, -1)

(e) Find the **one** candidate point for which $f(x, y)$ is smallest. Check if $L(x, y)$ is concave or convex at this point by using the Hessian in the second partial derivative test. [2pts]

Solution: Since there is only one candidate, it is (3.873, -1), which is convex.

Let us begin by calculating the hessian:
Our 4 partial derivatives are as follows:

$$\frac{\partial L}{\partial xx} = 2\lambda_1$$

$$\frac{\partial L}{\partial xy} = 1$$

$$\frac{\partial L}{\partial yy} = 6\lambda_1$$

$$\frac{\partial L}{\partial yx} = 1$$

This gives us the following hesian matrix, after plugging in $\lambda_1 = \frac{\sqrt{15}}{6}$:

$$\begin{bmatrix} \frac{\sqrt{15}}{3} & 1 \\ 1 & \sqrt{15} \end{bmatrix}$$

The determinant is: 5-1 = 4, meaning this is a local min or max.
This indicates that it might be a convex matrix, but lets check the eigenvalues just to be safe:

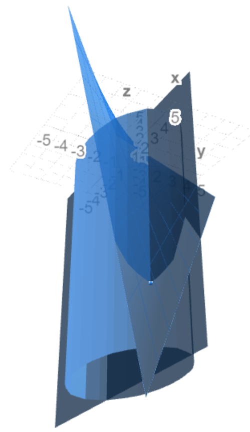
$$(\frac{\sqrt{15}}{3} - \lambda)(\sqrt{15} - \lambda) - 1 = 0$$
$$= 5 - \frac{\sqrt{15}}{3}\lambda - \sqrt{15}\lambda + \lambda^2 - 1 = 0$$
$$= 15 - \sqrt{15}\lambda - 3\sqrt{15}\lambda + 3\lambda^2 - 3 = 0$$
$$= 3\lambda^2 - 4\sqrt{15}\lambda + 12 = 0$$

This gives us eigenvalues 0.9 and 4.2. Since both eigenvalues are positive we know that this is a positive definite matrix, meaning that this is a convex function.

(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 minimum 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. [4pts]

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

Solution:



This makes sense because the candidate point is at the smallest point on f(x,y) but also less than or equal to the constraints defined by g_1 (must be inside the cylinder to be less than 18) and g_2 (the point must be below 1 on the x,y plane).

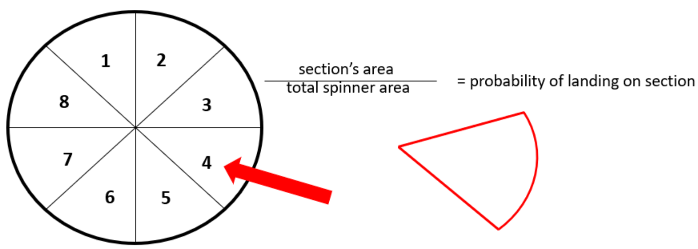
HINT: Read the Example_optimization_problem.pdf in Canvas Files for HW1 to see an example with some explanations.
HINT: Click here for an example maximization problem. It's recommended to only watch up until 23:14.
HINT: Click here to determine how to set up the problem for minimization in part (a) and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

4 Maximum Likelihood [10pts + 15pts 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 8 sections, each numbered 1 through 8. 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 8 times such that it doesn't land on section 1 during the first 7 spins and lands on section 1 on the 8th 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



Solution: $l(\theta|X) = \log\theta + 7log(1-\theta)$, $\theta = \frac{1}{8}$

we can treat this as a bernoulli distribution, with 8 trials
there are 7 '0's or F, and 1 '1' or passes.
Our trials look like this [0,0,0,0,0,0,0,1]
We can define our objective function as $L(\theta|X) = \theta^{\sum x_i}(1-\theta)^{\sum 1-x_i}$ Since in this formula, we define x_i as the ith value from the spinner, but since we know the exact order for the distribution (since we only care about pass or fail), we set this value to our fail and pass values, giving us the objective function:
$L(\theta|X) = \theta^1(1-\theta)^7$
Take the log of both sides.
$\log L(\theta|X) = \log\theta^1 + log(1-\theta)^7$
$l(\theta|X) = \log\theta + 7log(1-\theta)$
We need to optimize theta, meaning we want the max theta so that we can most likely ensure the students get a 100:
do this by taking the derivative (with respect to theta) and setting to 0, solving for theta:

$$\frac{\partial l}{\partial\theta} = \frac{1}{\theta\ln(2)} - \frac{7}{(1-\theta)\ln(2)} = 0$$
$$\frac{1}{\theta\ln(2)} = \frac{7}{(1-\theta)\ln(2)}$$
$$7\theta\ln(2) = (1-\theta)\ln(2)$$
$$\theta = \frac{1}{8}$$

4.2 Poisson distribution [15 pts]: Bonus for undergrads

The Poisson distribution is defined as:

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

(a) Let $X_1 \sim Poisson(\lambda)$, and x_1 be an observed value of X_1. What is the likelihood given λ? [2 pts]

> **Solution:** $\mathrm{P}(x|\lambda) = \frac{\lambda^{x_1} e^{-\lambda}}{x_1!}$

(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 this data given λ? You may leave your answer in product form. [3 pts]

(c) What is the maximum likelihood estimator of λ? [10 pts]

5 Information Theory [6pts + 16pts + 10pts]

5.1 Marginal Distribution [6pts]

Suppose the joint probability distribution of two binary random variables X and Y are given as follows. X are the rows, and Y are the columns.

X \ Y	0	1
0	$\frac{6}{16}$	$\frac{3}{16}$
1	$\frac{1}{4}$	$\frac{3}{16}$

(a) Show the marginal distribution of X and Y, respectively. [3pts]

Solution: you just add the probabilities across the row and columns:

X: $\frac{9}{16}$ when 0, $\frac{7}{16}$ when 1

Y: $\frac{10}{16}$ when 0, $\frac{6}{16}$ when 1

(b) Find mutual information $I(X, Y)$ for the joint probability distribution in the previous question to at least 5 decimal places (please use base 2 to compute logarithm) [3pts]

Solution: 0.00686

$$P(0,0) = \frac{6}{16} * \log(\frac{\frac{6}{16}}{\frac{9}{16} * \frac{10}{16}})$$

$$P(1,0) = \frac{1}{4} * \log(\frac{\frac{1}{4}}{\frac{7}{16} * \frac{10}{16}})$$

$$P(0,1) = \frac{3}{16} * \log(\frac{\frac{3}{16}}{\frac{9}{16} * \frac{6}{16}})$$

$$P(1,1) = \frac{3}{16} * \log(\frac{\frac{3}{16}}{\frac{7}{16} * \frac{6}{16}})$$

Sum all of these values to get 0.00686

5.2 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	H	F	T	T
2	H	F	F	F
3	H	F	T	F
4	M	F	T	F
5	L	T	T	T
6	L	T	F	F
7	L	T	F	T
8	L	T	T	T
9	L	T	T	T
10	M	T	T	T

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: 0.971

6/10 are true, 4/10 are false:
$-\frac{6}{10} * \log(\frac{6}{10}) - \frac{4}{10} * \log(\frac{4}{10}) = 0.971$

(b) Find conditional entropy $H(Y|X_1)$ and $H(Y|X_3)$ to at least 3 decimal places. [7pts]

Solution: 0.836, 0.880

H($Y|X_1$):

Begin by constructing a table:

$Y\|X_1$	H	M	L
Y = F	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$
Y = T	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{4}{10}$

We now must do a summation on the possible combinations of Y and X_1

$$p(T,H) = \frac{1}{10} * \log(\frac{\frac{3}{10}}{\frac{1}{10}})$$

$$p(F,H) = \frac{2}{10} * \log(\frac{\frac{3}{10}}{\frac{2}{10}})$$

$$p(T,M) = \frac{1}{10} * \log(\frac{\frac{2}{10}}{\frac{1}{10}})$$

$$p(F,M) = \frac{1}{10} * \log(\frac{\frac{2}{10}}{\frac{1}{10}})$$

$$p(T,L) = \frac{4}{10} * \log(\frac{\frac{5}{10}}{\frac{4}{10}})$$

$$p(F,L) = \frac{1}{10} * \log(\frac{\frac{5}{10}}{\frac{1}{10}})$$

Add the results of all of these to get: 0.836

H($Y|X_3$)

For this, we must construct another table.

-	X_3 =F	X_3 =T
Y = F	$\frac{2}{10}$	$\frac{2}{10}$
Y = T	$\frac{1}{10}$	$\frac{5}{10}$

We must now do a summation on the possible combinations of Y and X_3

$$p(y = F, X_3 = F) = \frac{2}{10} \log(\frac{\frac{3}{10}}{\frac{2}{10}})$$

$$p(y = F, X_3 = T) = \frac{2}{10} \log(\frac{\frac{7}{10}}{\frac{2}{10}})$$

$$p(y = T, X_3 = F) = \frac{1}{10} \log(\frac{\frac{3}{10}}{\frac{1}{10}})$$

$$p(y = T, X_3 = T) = \frac{5}{10} \log(\frac{\frac{7}{10}}{\frac{5}{10}})$$

Adding all of these values we get: 0.880

(c) Find mutual information $I(X_1, Y)$ and $I(X_3, Y)$ to at least 3 decimal places and determine which one (X_1 or X_3) is more informative. [3pts]

Solution: I(X_1, Y) = H(Y) - H($Y|X_1$) = 0.971 - 0.836 = 0.135

I(X_3, Y) = H(Y) - H($Y|X_3$) = 0.971 - 0.880 = 0.091

Since I(X_3, Y) has a greater decrease in entropy, which is more informative.

(d) Find joint entropy $H(Y, X_2)$ to at least 3 decimal places. [3pts]

Solution: 1.685

-	X_2 =F	X_2 =T
Y = F	$\frac{3}{10}$	$\frac{1}{10}$
Y = T	$\frac{1}{10}$	$\frac{5}{10}$

$$p(y = F, X_3 = F) = \frac{3}{10}\log(\frac{1}{\frac{3}{10}})$$

$$p(y = F, X_3 = T) = \frac{1}{10}\log(\frac{1}{\frac{1}{10}})$$

$$p(y = T, X_3 = F) = \frac{1}{10}\log(\frac{1}{\frac{1}{10}})$$

$$p(y = T, X_3 = T) = \frac{5}{10}\log(\frac{1}{\frac{5}{10}})$$

Sum all of these to get 1.685

5.3 Entropy Proofs [10pts]

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

Solution: H(x) $= \sum_x p_x \log \frac{1}{p_x}$

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

(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) [7pts]

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

Solution:

$$I(x,y) = H(x) - H(x|y)$$

Plug in the values from part a:

$$I(x,y) = \sum_x p_x \log \frac{1}{p_x} - \sum p(x,y) \log \frac{p(x)}{p(x,y)}$$

$$I(x,y) = \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)}$$

$$I(x,y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Apply bayes rule, p(x,y) for two independent variables is p($y|x$)p(x)
Plug this into our formula for p(x,y):

$$I(x,y) = \sum_{x,y} p(x,y) \log \frac{p(y|x)p(x)}{p(x)p(y)}$$

by definition, if x and y are independent, $p(y|x) =$ p(y)

$$I(x,y) = \sum_{x,y} p(x,y) \log \frac{p(y)p(x)}{p(x)p(y)}$$

$$I(x,y) = \sum_{x,y} p(x,y) \log 1$$

$$I(x,y) = \sum_{x,y} p(x,y) * 0$$

$$I(x,y) = 0$$

6 Programming [5 pts]

See the Programming subfolder in Canvas.

7 Bonus for All [20 pts]

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

$$P = 5X + 3XY^2$$
$$Q = X$$

Calculate the variance $Var(P+Q)$. *(This question may take substantial work to support, e.g. 25 to 30 lines)* [10pts]

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.
HINT: $Var(P+Q) = Var(P) + Var(Q) + 2Cov(P,Q)$ may be a good starting point.

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

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

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)* [5 pts]

$f_X(x) =$ Start your answer here.

(c) A person decides to toss a biased coin with $P(heads) = 0.2$ repeatedly until he gets a head. He will make at most 5 tosses. Let the random variable Y denote the number of heads. Find the probability distribution of Y. Then, find the variance of Y. *(It is possible to thoroughly support your answer to this question in 5 to 10 lines)* [5 pts]