Fall 2022 CS4641/CS7641 Homework 1

Bipin Koirala (9037.15.285)

Deadline: Friday, September 23rd, 11:59 pm AOE

- No unapproved extension of the deadline is allowed. Late submission will lead to 0 credit.
- 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.
- 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

- This assignment has no programming, only written questions.
- 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.
- 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 must correctly map pages of your PDF to each question/subquestion to reflect where they appear. Make sure to map the whole solution for each question/subquestion and NOT just the first page. **Improperly mapped questions may not be graded correctly or may receive point deductions.**
- 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 [43pts]

- 1.1 Determinant and Inverse of a Matrix [15pts]
- 1.2 Characteristic Equation [8pts]
- 1.3 Eigenvalues and Eigenvectors [20pts]

Q2: Covariance, Correlation, and 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 Weibull Distribution [15pts Bonus for Undergrads]

Q5: Information Theory [35pts]

- 5.1 Marginal Distribution [6pts]
- 5.2 Mutual Information and Entropy [19pts]
- 5.3 Entropy Proofs [10pts]

Q6: Bonus for All [15pts]

1 Linear Algebra [15pts + 8pts + 20pts]

1.1 Determinant and Inverse of Matrix [15pts]

Given a matrix M:

$$M = \begin{bmatrix} 4 & 2 & 1 \\ -3 & r & 2 \\ 0 & 7 & 1 \end{bmatrix}$$

(a) Calculate the determinant of M in terms of r. (Calculation process is required) [4pts]

(b) For what value(s) of r does M^{-1} not exist? Why? What does it mean in terms of rank and singularity for these values of r? [3pts]

(c) Will all values of r found in part b allow for a column (row) to be expressed as a linear combination of the other columns (rows) respectively? If yes, provide the linear combination of C_3 for column or the linear combination of R_2 for row; if no, explain why. [3pts]

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

(e) Find the determinant of M^{-1} for $r = 0$. What is the relationship between the determinant of M and the determinant of M^{-1}? [3pts]

(a) $$\det\begin{pmatrix} 4 & 2 & 1 \\ -3 & r & 2 \\ 0 & 7 & 1 \end{pmatrix} = \begin{vmatrix} 4 & 2 & 1 \\ -3 & r & 2 \\ 0 & 7 & 1 \end{vmatrix} = 4\begin{vmatrix} r & 2 \\ 7 & 1 \end{vmatrix} - 2\begin{vmatrix} -3 & 2 \\ 0 & 1 \end{vmatrix} + 1\begin{vmatrix} -3 & r \\ 0 & 7 \end{vmatrix} = 4\ (r-14) - 2(-3+0) + 1(-21+0)$$

$\therefore \det(M) = 4r - 56 + 6 - 21 = \mathbf{4r - 71}$

(b) M^{-1} does not exist when $\det(M) = 0$. Determinant of a matrix $\mathbb{R}^{n\times n}$ is given by;

$$M^{-1} = \frac{1}{det(M)}\Big[adjoint\ of\ M\Big]$$

Where *adjoint of M* is the transpose of a co-factor matrix of M. When $\det(M) = 0$; M^{-1} is undefined. Determinant of M does not exist when $\mathbf{r = 71/4}$
When r = 71/4; it means that any column of M can be expressed as the linear combination of other two columns of M. For example;

$$\begin{bmatrix} 2 \\ 71/4 \\ 7 \end{bmatrix} = \frac{-5}{4}\begin{bmatrix} 4 \\ -3 \\ 0 \end{bmatrix} + 7\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

This means that the columns of M are not linearly independent which in turn means that the matrix M is not a full rank matrix.

Also, the matrix M is said to be non-singular when its determinant is non-zero. Matrix M (if $r \neq 71/4$) represents a linear transformation $\mathbb{R}^3 \to \mathbb{R}^3$ and the mapping is either one-to-one or onto. When $r = 71/4$ the linear transformation is neither one-to-one nor onto and the matrix M is said to be singular.

(c) Since M is a singular matrix (if $r \neq 71/4$), we can represent each column(row) as a linear combination of other columns(rows). Expressing C_3 as a linear combination of C_1 and C_2:

$$\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = a_1\begin{bmatrix} 4 \\ -3 \\ 0 \end{bmatrix} + a_2\begin{bmatrix} 2 \\ 71/4 \\ 7 \end{bmatrix}$$

A system of equation is obtained from above:

$$\begin{aligned} 2a_1 + 2a_2 &= 1 \\ -3a_1 + \frac{71}{4}a_2 &= 2 \\ 0a_1 + 7a_2 &= 1 \end{aligned}$$

Here; $a_2 = 1/7$ and $a_1 = 5/28$

$$\therefore \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = \frac{5}{28}\begin{bmatrix} 4 \\ -3 \\ 0 \end{bmatrix} + \frac{1}{7}\begin{bmatrix} 2 \\ 71/4 \\ 7 \end{bmatrix}$$

(d) $\det(M) =$ -71 (if r = 0)

$$\begin{aligned} M^{-1} &= \frac{1}{\det(M)}\Big[adjoint\ of\ M\Big] \\ &= \frac{-1}{71}\begin{bmatrix} -14 & 3 & -21 \\ 5 & 4 & -28 \\ 4 & -11 & 6 \end{bmatrix}^T \\ &= \frac{-1}{71}\begin{bmatrix} -14 & 5 & 4 \\ 3 & 4 & -11 \\ -21 & -28 & 6 \end{bmatrix} \end{aligned}$$

(e) Determinant of M^{-1} is given by;

$\frac{-1}{71}\det\begin{pmatrix} -14 & 5 & 4 \\ 3 & 4 & -11 \\ -21 & -28 & 6 \end{pmatrix} = \frac{-1}{71}\begin{vmatrix} -14 & 5 & 4 \\ 3 & 4 & -11 \\ -21 & -28 & 6 \end{vmatrix} = \left(14/71\begin{vmatrix} -4/71 & 11/71 \\ 28/71 & -6/71 \end{vmatrix} + 5/71\begin{vmatrix} -3/71 & 11/71 \\ 21/71 & -6/71 \end{vmatrix} - 4/71\begin{vmatrix} -3/71 & -4/71 \\ 21/71 & 28/71 \end{vmatrix}\right) =$
$\frac{14}{71}\left(\frac{-4}{71}\right) + \frac{5}{71}\left(\frac{-3}{71}\right) + 0$
$\therefore \det(\mathbf{M}) = \mathbf{\frac{-1}{71}}$

If r = 0;

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

1.2 Characteristic Equation [8pts]

Consider the eigenvalue problem:

$$Ax = \lambda x, x \neq 0$$

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

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

Solution;

$$Ax = \lambda x \ \forall x \text{ s.t. } x \neq 0 \text{ ——(i)}$$

Consider the following expression:

$$\begin{aligned}(A - \lambda I_n)x &= Ax - \lambda I_n x \\ &= Ax - \lambda x \\ &= \lambda x - \lambda x \qquad \because \ from \quad (i) \\ \therefore (A - \lambda I_n)x &= 0 \qquad\qquad (ii)\end{aligned}$$

A vector x s.t. $x \neq 0$ is the kernel of $(A - \lambda I_n)$. i.e. $\ker(A - \lambda I_n) \neq \overrightarrow{\{0\}}$
If $(A - \lambda I_n)x$ exists;

$$\begin{aligned}(A - \lambda I_n)x &= 0 \\ (A - \lambda I_n)^{-1}(A - \lambda I_n)x &= (A - \lambda I_n)^{-1}0 \\ x &= 0\end{aligned}$$

This contradicts our original supposition that $x \neq 0$.
Then the followings are equivalent:

- $\det(A - \lambda I_n)^{-1}$ does not exist
- $|A - \lambda I_n| = 0$
- columns of $A - \lambda I_n$ are linearly dependent

1.3 Eigenvalues and Eigenvectors [5+5+10pts]

1.3.1 Eigenvalues [5pts]

Given a matrix A:

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

(a) Find an expression for the eigenvalues (λ) of $\boldsymbol{A}$ and solve for λ in the terms given. [4pts]

(b) Find a simple expression for the eigenvalues if $c = a$. [1pt]

(a) To get eigenvalues of A: $\det(A - \lambda I_n) = 0$

$$|A - \lambda I_n| = 0$$

$$\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = 0$$

$$\begin{vmatrix} a - \lambda & b \\ c & d - \lambda \end{vmatrix} = 0$$

$$\lambda^2 - (a + d)\lambda + ad - bc = 0$$

λ is given by;

$$\lambda = \frac{(a + d) \pm \sqrt{(a + d)^2 - 4 \times 1 \times (ad - bc)}}{2}$$

$$\lambda = \frac{(a + d) \pm \sqrt{a^2 - 2ad + 4bc + d^2}}{2}$$

(b) If c = a;

$$\lambda = \frac{(a + d) \pm \sqrt{a^2 + 2a(2b - d) + d^2}}{2}$$

1.3.2 Trace and Eigenvectors [5pts]

A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ can be decomposed as

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T = \sum_{n=1}^{N} \lambda_n \boldsymbol{v}_n \boldsymbol{v}_n^T$$

Where $\boldsymbol{V}$ is a matrix whose columns are the eigenvectors of $\boldsymbol{A}$, $\boldsymbol{v}_n$ are the columns of $\boldsymbol{V}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix whose elements are the eigenvalues of $\boldsymbol{A}$. The eigenvectors are orthonormal to each other, i.e., $\boldsymbol{v}_i^T \boldsymbol{v}_j = \begin{cases} 1, i = j \\ 0, i \neq j \end{cases}$.

(a) Show that $\text{trace}(\boldsymbol{A}) = \sum_{n=1}^{N} \lambda_n$ [3pts]
NOTE: $\boldsymbol{v}_i^T \boldsymbol{v}_j \neq \boldsymbol{v}_i \boldsymbol{v}_j^T$

(b) What is the result of the multiplication $\boldsymbol{V}^T\boldsymbol{V}$? Show your work or present an argument. [2pts]

(a) Here;

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\boldsymbol{T}}$$

Taking trance on both sides

$$\begin{aligned}
trace(\boldsymbol{A}) &= trace(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\boldsymbol{T}}) \\
&= trace \sum_{n=1}^{N} \lambda_n \boldsymbol{v}_n \boldsymbol{v}_n^T \\
&= \sum_{n=1}^{N} trace(\lambda_n \boldsymbol{v}_n \boldsymbol{v}_n^T) \\
&= \sum_{n=1}^{N} trace(\boldsymbol{v}_n^T \lambda_n \boldsymbol{v}_n) \quad \because trace(\boldsymbol{X}\boldsymbol{Y}\boldsymbol{Z}) = trace(\boldsymbol{Z}\boldsymbol{X}\boldsymbol{Y}) \\
&= \sum_{n=1}^{N} \lambda_n trace(\boldsymbol{v}_n^T \boldsymbol{v}_n) \quad \because trace(\boldsymbol{c}\boldsymbol{X}) = \boldsymbol{c}\ trace(\boldsymbol{X}) \\
&= \sum_{n=1}^{N} \lambda_n \times 1 \quad \because \boldsymbol{v_i}^T \boldsymbol{v_i} = 1 \\
\therefore trace(\boldsymbol{A}) &= \sum_{n=1}^{N} \lambda_n
\end{aligned}$$

(b) We have;

$$\boldsymbol{V} = \begin{bmatrix} | & | & \cdots\cdots & | \\ \boldsymbol{v_1} & \boldsymbol{v_2} & \cdots\cdots & \boldsymbol{v_n} \\ | & | & \cdots\cdots & | \end{bmatrix} \text{ and } \boldsymbol{V^T} = \begin{bmatrix} \text{———} & \boldsymbol{v_1} & \text{———} \\ \text{———} & \boldsymbol{v_2} & \text{———} \\ \text{———} & \boldsymbol{v_n} & \text{———} \end{bmatrix}$$

Now;

$$\boldsymbol{V^T V} = \begin{bmatrix} \boldsymbol{v_1^T v_1} & \cdots & \boldsymbol{v_1^T v_n} \\ \vdots & \ddots & \vdots \\ \boldsymbol{v_n^T v_1} & \cdots & \boldsymbol{v_n^T v_n} \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} = \boldsymbol{I} \ (\because \text{eigenvectors are orthonormal})$$

1.3.3 Eigenvalue and Eigenvector Calculations [10pts]

Given a matrix

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

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

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

(a) To get eigenvalues of A: $\det(A - \lambda I_n) = 0$

$$\begin{aligned} |A - \lambda I_n| &= 0 \\ \det\left(\begin{bmatrix} x & 5 \\ 5 & x \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) &= 0 \\ \begin{vmatrix} x-\lambda & 5 \\ 5 & x-\lambda \end{vmatrix} &= 0 \\ (x-\lambda)^2 &= 25 \\ x - \lambda &= \pm 5 \\ i.e.\ \boldsymbol{\lambda = x \pm 5} & \end{aligned}$$

(b) Let $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$ be the eigenvectors associated with $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$ respectively.
When, $\boldsymbol{\lambda_1 = x + 5}$

$$\begin{aligned} (A - \lambda_1 I)\boldsymbol{v_1} &= 0 \\ \begin{bmatrix} x-\lambda_1 & 5 \\ 5 & x-\lambda_1 \end{bmatrix} \begin{bmatrix} \boldsymbol{v_{11}} \\ \boldsymbol{v_{12}} \end{bmatrix} &= 0 \\ \begin{bmatrix} -5 & 5 \\ 5 & -5 \end{bmatrix} \begin{bmatrix} \boldsymbol{v_{11}} \\ \boldsymbol{v_{12}} \end{bmatrix} &= 0 \\ \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \boldsymbol{v_{11}} \\ \boldsymbol{v_{12}} \end{bmatrix} &= 0 \end{aligned}$$

A system of equation is obtained from above:

$$\begin{aligned} -\boldsymbol{v_{11}} + \boldsymbol{v_{12}} &= 0 \\ \boldsymbol{v_{11}} - \boldsymbol{v_{12}} &= 0 \end{aligned}$$

i.e. $\boldsymbol{v_{11} = v_{12}}$. When, $\boldsymbol{v_{11} = v_{12} = t}$; $\boldsymbol{v_1} = \begin{bmatrix} \boldsymbol{t} \\ \boldsymbol{t} \end{bmatrix} = \boldsymbol{t} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$\therefore$ Normalized $\boldsymbol{v_1} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

When, $\boldsymbol{\lambda_1 = x - 5}$

$$(A - \lambda_2 I)\boldsymbol{v_2} = 0$$

$$\begin{bmatrix} x - \lambda_2 & 5 \\ 5 & x - \lambda_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{v_{21}} \\ \boldsymbol{v_{22}} \end{bmatrix} = 0$$

$$\begin{bmatrix} 5 & 5 \\ 5 & 5 \end{bmatrix} \begin{bmatrix} \boldsymbol{v_{21}} \\ \boldsymbol{v_{22}} \end{bmatrix} = 0$$

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{v_{21}} \\ \boldsymbol{v_{22}} \end{bmatrix} = 0$$

A system of equation is obtained from above:

$$\boldsymbol{v_{21}} + \boldsymbol{v_{22}} = 0$$

i.e. $\boldsymbol{v_{21} = -v_{22}}$. When, $\boldsymbol{v_{22} = t}$; $\boldsymbol{v_2} = \begin{bmatrix} \boldsymbol{-t} \\ \boldsymbol{t} \end{bmatrix} = \boldsymbol{t} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

$\therefore$ Normalized $\boldsymbol{v_2} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

2 Expectation, Co-variance and 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 distribution function is

$$p(x) = \begin{cases} 0.6 & x = c \\ 0.4 & x = -c \end{cases}$$

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 independent. Meanwhile, let $Z = XY$.

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

$$\begin{aligned} Cov(Y,Z) &= \mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z] \\ &= \mathbb{E}[YXY] - \mathbb{E}[Y]\mathbb{E}[XY] \\ &= \mathbb{E}[XY^2] - \mathbb{E}[Y]\mathbb{E}[XY] \\ &= \mathbb{E}[X]\mathbb{E}[Y^2] - \mathbb{E}[X](\mathbb{E}[Y])^2 \\ &= \mathbb{E}[X]\ (\mathbb{E}[Y^2] - (E[Y])^2) \\ &= \mathbb{E}[X]Var(Y) \end{aligned}$$

Here; $\mathbb{E}[X] = 0.6 \times (c) + 0.4 \times (-c) = 0.2c$ and $Var(Y) = 1$

$\therefore Cov(Y,Z) = 0.2c$

The values of 'c' indeed affects the covariance between Y and Z.

2.2 Correlation Coefficient [4pts]

Let X and Y be independent random variables with $var(X) = 5$ and $var(Y) = 15$. We do not know $E[X]$ or $E[Y]$. Let $Z = 3X + 2Y$. What is the correlation coefficient $\rho(X, Z) = \frac{cov(X,Z)}{\sqrt{var(X)var(Z)}}$? If applicable, please round your answer to 3 decimal places. [4pts]

Solution;

$$
\begin{aligned}
Var(Z) &= Var(3X + 2Y) \\
&= 3^2 Var(X) + 2^2 Var(Y) + 2 \times 3 \times 2\ Cov(X, Y) \\
&= 3^2 Var(X) + 2^2 Var(Y) && \because \text{(X, Y independent)} \\
&= 9 \times 5 + 4 \times 15 \\
&= 105
\end{aligned}
$$

$$
\begin{aligned}
Cov(X, Z) &= \mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z] \\
&= \mathbb{E}[X(3X + 2Y)] - \mathbb{E}[X]\mathbb{E}[3X + 2Y] \\
&= \mathbb{E}[3X^2 + 2XY] - \mathbb{E}[X]\mathbb{E}[3X + 2Y] \\
&= \mathbb{E}[3X^2] + \mathbb{E}[2XY] - \mathbb{E}[X]\Big(\mathbb{E}[3X] + \mathbb{E}[2Y]\Big) \\
&= 3\mathbb{E}[X^2] + 2\mathbb{E}[X][Y] - 3(\mathbb{E}[X])^2 - 2\mathbb{E}[X]\mathbb{E}[Y] \\
&= 3\Big(\mathbb{E}[X^2] - \mathbb{E}[X]^2\Big) + 0 \\
&= 3Var(X) \\
&= 15
\end{aligned}
$$

$$
\begin{aligned}
\rho(X, Z) &= \frac{Cov(X, Z)}{\sqrt{Var(X)\ Var(Z)}} \\
&= \frac{15}{\sqrt{5 \times 105}} \\
&= 0.655
\end{aligned}
$$

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:

$$\begin{aligned} \max_{x,y} \quad & f(x,y) = -4y + xy \\ \text{s.t.} \quad & g_1(x,y) = 2x^2 + y^2 \leq 12 \\ & g_2(x,y) = x \leq 1 \end{aligned}$$

(a) Write the Lagrange function for the maximization problem. Now change the maximum function to a minimum function (i.e. $\min_{x,y} f(x,y) = -4y+xy$) 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).

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

(c) Solve for 4 possibilities formed by each constraint being active or inactive. Do not forget to check the inactive constraints for each point. Candidate points must satisfy the inactive constraints. [5pts]

(d) List the candidate point(s) (there may be 0, 1, 2, or any number of candidate points) [4pts]

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

HINT 1: Click here for an example maximization problem.
HINT 2: Click here to determine how to set up the problem for minimization in part (a).

(a) Lagrange function for the maximization problem.

$$\begin{aligned} \mathcal{L}(x,y) &= f(x,y) - \lambda_1 g_1(x,y) - \lambda_2 g_2(x,y) \\ &= xy - 4y - \lambda_1(2x^2 + y^2 - 12) - \lambda_2(x-1) \qquad \text{(where } \lambda_1, \lambda_2 \geq 0\text{)} \end{aligned}$$

Lagrange function for the minimization problem.

$$\begin{aligned} \min_{x,y} \quad & f(x,y) = -4y + xy \\ \text{s.t.} \quad & g_1(x,y) = 2x^2 + y^2 > 12 \\ & g_2(x,y) = x > 1 \end{aligned}$$

$$\begin{aligned} \mathcal{L}(x,y) &= f(x,y) + \lambda_1 g_1(x,y) + \lambda_2 g_2(x,y) \\ &= xy - 4y + \lambda_1(2x^2 + y^2 - 12) + \lambda_2(x-1) \qquad \text{(where } \lambda_1, \lambda_2 \geq 0\text{)} \end{aligned}$$

(b) KKT Conditions are as follows:

- Stationary conditions

$$\frac{\partial \mathcal{L}}{\partial x} = 0 \qquad i.e. \qquad y - 4\lambda_1 x - \lambda_2 = 0$$
$$\frac{\partial \mathcal{L}}{\partial y} = 0 \qquad i.e. \qquad x - 4 - 2\lambda_1 y = 0$$

- Primal Feasibility

$$g_1(x,y) \le 0 \qquad\qquad g_2(x,y) \le 0$$
$$i.e.\ \ 2x^2 + y^2 - 12 \le 0 \qquad\qquad i.e\ \ x - 1 \le 0$$

- Dual Feasibility

$$\lambda_1 \ge 0 \qquad \text{and} \qquad \lambda_2 \ge 0$$

- Complementary Slackness

$$\lambda_1 g_1(x,y) = 0 \qquad \text{and} \qquad \lambda_2 g_2(x,y) = 0$$
$$\text{if}\ \ g_1(x,y) = 0;\ \ \lambda_1 \ne 0 \qquad\qquad \text{if}\ \ g_2(x,y) = 0;\ \ \lambda_2 \ne 0$$
$$\text{if}\ \ g_1(x,y) \ne 0;\ \ \lambda_1 = 0 \qquad\qquad \text{if}\ \ g_2(x,y) \ne 0;\ \ \lambda_2 = 0$$

(c) Four possibilities formed by each constraint being active or inactive are as follows:

(i) When both constraints are active

$$2x^2 + y^2 = 12 \qquad\qquad x = 1$$
$$\lambda_1 > 0 \qquad\qquad \lambda_2 > 0$$

From these equations we have; $y = \pm\sqrt{10}$. Thus, the possible solutions are $(x,y) = (1,\sqrt{10})$ and $(1,-\sqrt{10})$

When $y = \sqrt{10}$, according to second stationary condition;
$1 - 4 - 2\lambda_1\sqrt{10} = 0$ i.e. $\lambda_1 < 0$ (not feasible)

When $y = -\sqrt{10}$, according to second stationary condition;
$1 - 4 - 2\lambda_1\sqrt{10} = 0$ i.e. $\lambda_1 = \frac{3}{2\sqrt{10}}$ (feasible)
Plugging λ_1 in the first stationary condition we get $\lambda_2 < 0$ (not feasible).

Therefore, the solution is not feasible when both constraints are active.

(ii) When first constraint is active and second constraint is inactive

$$2x^2 + y^2 = 12 \qquad\qquad x < 1$$
$$\lambda_1 > 0 \qquad\qquad \lambda_2 = 0$$

From these equations we have; $x = \pm\sqrt{\frac{12-y^2}{2}}$

Taking above values for x and after solving for y using stationary conditions we get $y = \pm\sqrt{10}$ which does not satisfy the constraint $x < 1$ so; it is not a feasible solution.

(iii) When first constraint is inactive and second constraint is active

$$2x^2 + y^2 < 12 \qquad\qquad x = 1$$
$$\lambda_1 = 0 \qquad\qquad \lambda_2 > 0$$

From stationary condition 1 we have $y - 4\lambda_1 - \lambda_2 = 0$ i.e. $y = \lambda_2$
From condition 2 we get $1 - 4 - 2\lambda_1\lambda_2 = 0$ i.e. $1 = 4$ which is not possible. So, the solution is not feasible.

(iv) When both the constraints are inactive.

$$2x^2 + y^2 < 12 \qquad\qquad x < 1$$
$$\lambda_1 = 0 \qquad\qquad \lambda_2 = 0$$

This gives $x = 1$ and $y = \pm\sqrt{10}$. When plugged into stationary conditions it yields; $\sqrt{10} = 0$ and $-3 = 0$ which is not possible. Hence the solution is not feasible.

(f) **BONUS FOR ALL:** Make a contour plot of objective function $f(x, y)$ and constraints g_1 and g_2 using the template Google Colab code. Mark the maximum candidate point and include a screenshot of your plot. Also include the text output from the last cell in the Google Colab for grading purposes. Lastly, briefly explain why your plot makes sense in one sentence. [4pts]

Note 1: Points on a line in the contour plot have equal values of the objective function. Keeping this in mind, you should be able to figure out the approximate location of the maximum.

Note 2: To use the Google Colab notebook, click "Copy to Drive" upon initial opening

4 Maximum Likelihood [10pts + 15pts Bonus for Undergrads]

4.1 Discrete Example [10pts]

Marion and Shreeya are arguing over which course they should take in Fall 2022. Marion's argument is that they should take CS-7650 NLP because Professor Roozbahani will teach it. Shreeya's argument is that they should take CS-7641 ML because it would be difficult to take NLP without having introductory knowledge of Machine Learning.

To resolve this conflict, their other friend Nicole makes a proposition that they should leave it to chance to decide which course they should take. Marion then proposes that Shreeya will toss a 6-sided die 6 times, and Shreeya must get anything except 3 during the first 5 times and must get 3 during the 6th time. Any other combination will make Marion the winner. But Shreeya is also allowed to tamper with the die in any manner she likes to increase her odds.

Now, Shreeya needs you to help her have her way. If the probability of getting a 3 is θ and the probability of landing on 1 is double of that of landing on 2, 4, 5, and 6, what value of θ is most likely to ensure that they will have to take CS-7641 ML? Use your expertise of Maximum Likelihood Estimation and probability distribution function to convince Shreeya.

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;
Let 'X' be a random variable that represents the probability of getting any one side of biased die.
$\mathcal{P}(X = 3) = \theta$. Since, the total probability of events in a sample space is 1. We get $\mathcal{P}(X \neq 3) = 1 - \theta$

To maximize the chance to get Shreeya win the duel is given by log-likelihood function below;

$$\begin{aligned}\mathcal{L}(\theta) &= \prod_{i=1}^{6} \mathcal{P}(X = x_i) \\ &= \mathcal{P}(X \neq 3)^5 \mathcal{P}(X = 3) \\ &= (1-\theta)^5 \ \theta\end{aligned}$$

Taking log on both sides we get;

$$\begin{aligned}log\Big(\mathcal{L}(\theta)\Big) &= log\Big((1-\theta)^5 \ \theta\Big) \\ &= log\Big((1-\theta)^5\Big) \ log(\theta) \\ &= 5log(1-\theta) + log(\theta)\end{aligned}$$

Differentiating with respect to θ and setting it to zero;

$$\begin{aligned}\frac{\partial log\Big(\mathcal{L}(\theta)\Big)}{\partial \theta} &= \frac{\partial\Big[5log(1-\theta) + log(\theta)\Big]}{\partial \theta} \\ 0 &= \frac{-5}{1-\theta} + \frac{1}{\theta} \\ \therefore \ \theta &= \frac{1}{6}\end{aligned}$$

4.2 Weibull distribution [15pts Bonus for Undergrads]

The Weibull distribution is defined as

$$P(X = x; \lambda, k) = \frac{k}{\lambda}(\frac{x}{\lambda})^{k-1} e^{-(x/\lambda)^k},\ x \geq 0$$

(a) Assume we have one observed data x_1, and $X_1 \sim Weibull(\lambda)$, what is the likelihood given λ and k? [2 pts]

(b) Now, assume we are given n such values $(x_1, ..., x_n)$, $(X_1, ..., X_n) \sim Weibull(\lambda)$. Here $X_1, ..., X_n$ are i.i.d. random variables. What is the likelihood of this data given λ and k? You may leave your answer in product form. [3 pts]

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

(a) The likelihood function is in the event of one observation (x_1) given by;

$$\mathcal{L}(\lambda, k) = \frac{k}{\lambda}\Big(\frac{x_1}{\lambda}\Big)^{k-1} e^{-(x_1/\lambda)^k}$$

(b) The likelihood function in the event when we have 'n' observations is given by;

$$\begin{aligned}
\mathcal{L}(\lambda, k) &= \prod_{i=1}^{n} \frac{k}{\lambda}\Big(\frac{x_i}{\lambda}\Big)^{k-1} e^{-(x_i/\lambda)^k} \\
&= \prod_{i=1}^{n} \frac{k}{\lambda^k} x_i^{k-1} e^{-(x_i/\lambda)^k} \\
&= \Big(\frac{k}{\lambda^k}\Big)^n \prod_{i=1}^{n} x_i^{k-1} \times e^{-(\Sigma_{i=1}^{n} x_i/\lambda)^k}
\end{aligned}$$

(c) Maximum likelihood estimator for λ:

$$\begin{aligned}
\log\Big(\mathcal{L}(\lambda, k)\Big) &= \log\left[\Big(\frac{k}{\lambda^k}\Big)^n \prod_{i=1}^{n} x_i^{k-1} \times e^{-(\Sigma_{i=1}^{n} x_i/\lambda)^k}\right] \\
&= \log\Big(\frac{k}{\lambda^k}\Big)^n + \log\left[\prod_{i=1}^{n} x_i^{k-1}\right] - \sum_{i=1}^{n}(x_i/\lambda)^k \\
&= n\log\Big(\frac{k}{\lambda^k}\Big) + (k-1)\log\prod_{i=1}^{n} x_i - \sum_{i=1}^{n}(x_i/\lambda)^k \\
&= n\log(k) - nk\log(\lambda) + (k-1)\sum_{i=1}^{n}\log(x_i) - \sum_{i=1}^{n}(x_i/\lambda)^k
\end{aligned}$$

Now, differentiating above with respect to λ and setting it to zero.

$$
\begin{aligned}
\frac{\partial \log(\mathcal{L}(\lambda, k))}{\partial \lambda} &= \frac{\partial}{\partial \lambda}\left[n\log(k) - nk\log(\lambda) + (k-1)\sum_{i=1}^{n}\log(x_i) - \sum_{i=1}^{n}(x_i/\lambda)^k \right] \\
0 &= 0 - \frac{nk}{\lambda} + 0 - \frac{\partial}{\partial \lambda}\left(\lambda^{-k}\sum_{i=1}^{n} x_i \right) \\
0 &= -\frac{nk}{\lambda} + k\lambda^{-k-1}\sum_{i=1}^{n} x_i \\
0 &= -nk + k\lambda^{-k}\sum_{i=1}^{n} x_i \\
n &= \lambda^{-k}\sum_{i=1}^{n} x_i \\
\therefore \quad \lambda &= \sqrt[k]{\frac{\sum_{i=1}^{n} x_i}{n}}
\end{aligned}
$$

5 Information Theory [6pts + 19pts + 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	$\frac{1}{16}$	$\frac{1}{4}$
1	$\frac{3}{16}$	$\frac{1}{2}$

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

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

(a) Marginal distribution of X and Y are as follows;

P(X,Y)	Y=0	Y=1	P(X=x)
X = 0	1/16	1/4	5/16
X = 1	3/16	1/2	11/16
P(Y=y)	4/16	3/4	1

(b) We know the mutual information $I(X,Y)$ is given by;

$$
\begin{aligned}
I(X,Y) &= H(X) - H(X|Y) \\
&= \sum_{x,y} \mathcal{P}(x,y) \log \frac{\mathcal{P}(x,y)}{\mathcal{P}(x)\mathcal{P}(y)} \\
&= \mathcal{P}(0,0) \log \frac{\mathcal{P}(0,0)}{\mathcal{P}(0)\mathcal{P}(0)} + \mathcal{P}(0,1) \log \frac{\mathcal{P}(0,1)}{\mathcal{P}(0)\mathcal{P}(1)} + \mathcal{P}(1,0) \log \frac{\mathcal{P}(1,0)}{\mathcal{P}(1)\mathcal{P}(0)} + \mathcal{P}(1,1) \log \frac{\mathcal{P}(1,1)}{\mathcal{P}(1)\mathcal{P}(1)} \\
&= \frac{1}{16} \log \left(\frac{1/16}{(5/16)(4/16)} \right) + \frac{1}{4} \log \left(\frac{1/4}{(5/16)(3/4)} \right) + \frac{3}{16} \log \left(\frac{3/16}{(11/16)(4/16)} \right) + \frac{1}{2} \log \left(\frac{1/2}{(11/16)(3/4)} \right) \\
&= \frac{1}{16} \log(0.8) + \frac{1}{4} \log(16/15) + \frac{3}{16}(12/11) + \frac{1}{2}(32/33) \\
&= 0.0045
\end{aligned}
$$

5.2 Mutual Information and Entropy [19pts]

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	**Age Group** (x_1)	**Vaccine Doses** (x_2)	**Wears Mask?** (x_3)	**Underlying Conditions** (x_4)	**Symptomatic** (Y)
1	Y	H	F	T	T
2	Y	H	F	F	F
3	A	H	F	T	T
4	S	M	F	T	T
5	S	L	T	T	T
6	S	L	T	F	F
7	A	L	T	F	T
8	Y	L	F	T	F
9	Y	L	T	T	F
10	S	M	T	T	T

Table 1: Age Groups: {(Y)outh, (A)dult, (S)enior}, Vaccine Doses: {(H) booster, (M) 2 doses, (L) 1 dose}

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

(b) Find conditional entropy $H(Y|x_2)$, $H(Y|x_4)$, respectively, to at least 3 decimal places. [8pts]

(c) Find mutual information $I(x_2, Y)$ and $I(x_4, Y)$ and determine which one (x_2 or x_4) is more informative. [4pts]

(d) Find joint entropy $H(Y, x_3)$ to at least 3 decimal places. [4pts]

(a) Here, $\mathcal{P}(Y = T) = 0.6$ and $\mathcal{P}(Y = F) = 0.4$

$$\begin{aligned} H(Y) &= \sum I(Y)\mathcal{P}(Y) \\ &= -\mathcal{P}(Y = T)\log_2 \mathcal{P}(Y = T) - \mathcal{P}(Y = F)\log_2 \mathcal{P}(Y = F) \\ &= 0.971 \end{aligned}$$

(b) To calculate $I(x_2, Y)$

$P(x_2, y)$	Y=T	Y=F	$P(x_2 = x)$
$x_2 = H$	0.2	0.1	0.3
$x_2 = M$	0.2	0	0.2
$x_2 = L$	0.2	0.3	0.5
P(Y=y)	0.6	0.4	1

$$
\begin{aligned}
I(x_2, Y) &= H(x_2) - H(x_2|Y) \\
&= \sum_{x_2,y} \mathcal{P}(x_2, y) \log \frac{\mathcal{P}(x_2, y)}{\mathcal{P}(x_2)\mathcal{P}(y)} \\
&= \mathcal{P}(H,T) \log \frac{\mathcal{P}(H,T)}{\mathcal{P}(H)\mathcal{P}(T)} + \mathcal{P}(H,F) \log \frac{\mathcal{P}(H,F)}{\mathcal{P}(H)\mathcal{P}(F)} + \mathcal{P}(M,T) \log \frac{\mathcal{P}(M,T)}{\mathcal{P}(M)\mathcal{P}(T)} + \mathcal{P}(M,F) \log \frac{\mathcal{P}(M,F)}{\mathcal{P}(M)\mathcal{P}(F)} \\
&+ \mathcal{P}(L,T) \log \frac{\mathcal{P}(L,T)}{\mathcal{P}(L)\mathcal{P}(T)} + \mathcal{P}(L,F) \log \frac{\mathcal{P}(L,F)}{\mathcal{P}(L)\mathcal{P}(F)} \\
&= 0.2 \log \left(\frac{0.2}{0.3 \times 0.6} \right) + 0.1 \log \left(\frac{0.1}{0.3 \times 0.4} \right) + 0.2 \log \left(\frac{0.2}{0.2 \times 0.6} \right) + 0 + 0.2 \log \left(\frac{0.2}{0.5 \times 0.6} \right) \\
&+ 0.3 \log \left(\frac{0.3}{0.5 \times 0.4} \right) \\
&= 0.210
\end{aligned}
$$

Now;

$$
\begin{aligned}
H(Y|x_2) &= H(Y) - I(x_2, Y) \\
&= 0.971 - 0.210 \\
&= 0.761 \qquad \text{☺}
\end{aligned}
$$

To calculate $I(x_4, Y)$

$P(x_4, y)$	Y=T	Y=F	$P(x_4 = x)$
$x_4 = T$	0.5	0.2	0.7
$x_4 = F$	0.1	0.2	0.3
P(Y=y)	0.6	0.4	1

$$
\begin{aligned}
I(x_4, Y) &= H(x_4) - H(x_4|Y) \\
&= \sum_{x_4,y} \mathcal{P}(x_4, y) \log \frac{\mathcal{P}(x_4, y)}{\mathcal{P}(x_4)\mathcal{P}(y)} \\
&= \mathcal{P}(T,T) \log \frac{\mathcal{P}(T,T)}{\mathcal{P}(T)\mathcal{P}(T)} + \mathcal{P}(T,F) \log \frac{\mathcal{P}(T,F)}{\mathcal{P}(T)\mathcal{P}(F)} + \mathcal{P}(F,T) \log \frac{\mathcal{P}(F,T)}{\mathcal{P}(F)\mathcal{P}(T)} + \mathcal{P}(F,F) \log \frac{\mathcal{P}(F,F)}{\mathcal{P}(F)\mathcal{P}(F)} \\
&= 0.5 \log \left(\frac{0.5}{0.7 \times 0.6} \right) + 0.2 \log \left(\frac{0.2}{0.7 \times 0.4} \right) + 0.1 \log \left(\frac{0.1}{0.3 \times 0.6} \right) + 0.2 \log \left(\frac{0.2}{0.3 \times 0.4} \right) \\
&= 0.0913
\end{aligned}
$$

Now;

$$
\begin{aligned}
H(Y|x_4) &= H(Y) - I(x_4, Y) \\
&= 0.971 - 0.0913 \\
&= 0.880 \qquad \text{☺}
\end{aligned}
$$

(c) From previous part [5.2(b)]; $I(x_2, Y) = 0.210$ and $I(x_4, Y) = 0.0913$
Mutual information quantifies the reduction in uncertainty in Y after seeing feature X_i. The more reduction in entropy; the more informative a feature so x_2 is more informative than x_4.

(d) To calculate $H(Y, x_3)$

$P(x_3, y)$	Y=T	Y=F	$P(x_3 = x)$
$x_3 = T$	0.3	0.2	0.5
$x_3 = F$	0.3	0.2	0.5
P(Y=y)	0.6	0.4	1

$$
\begin{aligned}
H(Y, x_3) &= -\sum_{x_3, y} \mathcal{P}(y, x_3) \log_2 \mathcal{P}(y, x_3) \\
&= \Big(-0.3 \times \log(0.3) - 0.2 \times \log(0.2) \Big) \times 2 \\
&= 1.971 \qquad \text{☺}
\end{aligned}
$$

5.3 Entropy Proofs [10pts]

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

(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 independent. (Note: you must provide a mathematical proof and cannot use the visualization shown in class found here)

Start from $I(X;Y) = H(X) - H(X|Y)$ [7pts]

(a)

$$
\begin{aligned}
H(X|Y) &\equiv \sum_{y\in Y} p(y)H(X|Y=y) \\
&= -\sum_{y\in Y} p(y) \sum_{x\in X} \log_2 p(x|y) \\
&= -\sum_{y\in Y} \sum_{x\in X} p(y,x) \log_2 p(x|y) \\
&= -\sum_{x\in X, y\in Y} p(y,x) \log_2 \frac{p(y,x)}{p(y)}
\end{aligned}
$$

$$
\begin{aligned}
H(X) &= \sum I(X)\mathcal{P}(X) \\
&= -\sum_{k=1}^{K} \mathcal{P}(X=k) \log_2 \mathcal{P}(X=k)
\end{aligned}
$$

(b)

$$
\begin{aligned}
I(X;Y) &= -H(X|Y) + H(X) \\
&= -\sum_{y\in Y} P(y)H(X|Y=y) - \sum_{x\in X} P(x)\log_2 P(x) \\
&= \sum_{y\in Y} P(y)\left(\sum_{x\in X} P_{X|Y=y}(x)\log_2 P_{X|Y=y}(x)\right) - \sum_{x\in X}\left(\sum_{y\in Y} P(x,y)\right)\log_2 P(x) \\
&= \sum_{x\in X, y\in Y} P(y)P_{X|y=y}(x)\log_2 P_{X|Y=y}(x) - \sum_{x\in X, y\in Y} P(x,y) log_2 P(x) \\
&= \sum_{x\in X, y\in Y} P(x,y) log_2 \frac{P(x,y)}{P(y)} - \sum_{x\in X, y\in Y} P(x,y) log_2 P(x) \\
&= \sum_{x\in X, y\in Y} P(x,y) log_2 P(x) - \sum_{x\in X, y\in Y} P(x,y) log_2 P(x) \qquad \because P(x,y) = P(x)\ P(y)\ \ independent \\
&= 0
\end{aligned}
$$

6 Bonus for All [15 pts]

(a) X, Y are two independent $N(0,1)$ random variables, and we have random variables P, Q defined as

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

then calculate the variance $Var(P+Q)$ [5pts]

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

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

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

(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 variance of Y. [5 pts]

(a) Here;

$$\begin{aligned} \mathbb{E}[X] &= 0 & \qquad \mathbb{E}[Y] &= 0 \\ Var(X) &= 1 & \qquad Var(Y) &= 1 \\ \mathbb{E}[X^2] &= 1 & \qquad \mathbb{E}[Y^2] &= 1 \end{aligned}$$

Now;

$$\begin{aligned} Var(P+Q) &= Var(P) + Var(Q) + 2\ Cov(P,Q) \\ &= Var(3X + XY^2) + Var(X) + 2Cov(3X + XY^2, X) \end{aligned}$$

$$\begin{aligned} Cov(3X + XY^2, X) &= \mathbb{E}[(3X + XY^2)(X)] - \mathbb{E}[3X + XY^2]\mathbb{E}[X] \\ &= \mathbb{E}[3X^2 + X^2Y^2] - \Big(3\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y^2]\Big)E[X] \\ &= 3\mathbb{E}[X^2] + \mathbb{E}[X^2]\mathbb{E}[Y^2] \\ &= 4 \end{aligned}$$

$$\begin{aligned} Var(3X + XY^2) &= \mathbb{E}[(3X + XY^2)^2] - \Big(\mathbb{E}[3X + XY^2]\Big)^2 \\ &= \mathbb{E}[9X^2 + 6X^2Y^2 + X^2Y^4] - \Big(3\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y^2]\Big)^2 \\ &= 9\mathbb{E}[X^2] + 6\mathbb{E}[X^2]\mathbb{E}[Y^2] + \mathbb{E}[X^2]\mathbb{E}[Y^4] - 0 \\ &= 9 + 6 + 1 \times \mathbb{E}[Y^4] \end{aligned}$$

$\mathbb{E}[Y^4]$ is given by;

$$\mathbb{E}[Y^4] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\ y^4 \exp(-\frac{y^2}{2})\ dy \quad = 3 \qquad \text{(calculations not shown)}$$

$$\therefore\ Var(P+Q) = 9 + 6 + 3 + 1 + 2 \times 4 = 27$$

(b) The marginal probability density function (pdf) for X is;

$$\begin{aligned} f_X(x) &= \int_{\mathbb{R}} f(x,y)\,dy \\ &= \int_{\mathbb{0}}^{\infty} 2e^{-2y}\,dy \\ &= \frac{-2}{2}e^{-2y}\Big|_0^{\infty} \end{aligned}$$

$$\text{i.e } f_X(x) = 1$$

The marginal probability density function (pdf) for Y is;

$$\begin{aligned} f_Y(y) &= \int_{\mathbb{R}} f(x,y)\,dx \\ &= \int_{\mathbb{0}}^{1} 2e^{-2y}\,dx \\ &= 2e^{-2y}x\Big|_0^{1} \end{aligned}$$

$$\text{i.e } f_X(x) = 2e^{-2y}$$

(c) We are asked to find the variance of Y i.e $Var(Y) = E[Y^2] - (\mathbb{E}[Y])^2$
Possible outcomes of this experiment are as follows (where $\mathcal{P}(H) = 0.2$ and $\mathcal{P}(T) = 0.8$):

$Y = 1$	H	w.p. P $= 1/5$
$Y = 1$	TH	w.p. P $= 4/25$
$Y = 1$	TTH	w.p. P $= 16/125$
$Y = 1$	$TTTH$	w.p. P $= 64/625$
$Y = 1$	$TTTTH$	w.p. P $= 256/3125$
$Y = 0$	$TTTTT$	w.p. P $= 1024/3125$

Now;

$$\begin{aligned} \mathbb{E}[Y] &= \sum_{y\in Y} y\mathcal{P}(Y=y) = 1\times\frac{1}{5} + 1\times\frac{4}{25} + 1\times\frac{16}{125} + 1\times\frac{64}{625} + 1\times\frac{256}{3125} + 0\times\frac{1024}{3125} \\ &= \frac{2101}{3125} \\ &= 0.672 \end{aligned}$$

$$\begin{aligned} \mathbb{E}[Y^2] &= \sum_{y\in Y} y^2\mathcal{P}(Y=y) = 1^2\times\frac{1}{5} + 1^2\times\frac{4}{25} + 1^2\times\frac{16}{125} + 1^2\times\frac{64}{625} + 1^2\times\frac{256}{3125} + 0\times\frac{1024}{3125} \\ &= \frac{2101}{3125} \\ &= 0.672 \end{aligned}$$

$$\therefore \quad Var(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = 0.22$$