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]

$$|\boldsymbol{M}| = a_{11}C_{11} + a_{12}C_{12} + a_{13}C_{13} = 3 \times \det \begin{bmatrix} 2 & -4 \\ -3 & 5 \end{bmatrix} - 1 \times \det \begin{bmatrix} r & -4 \\ 0 & 5 \end{bmatrix} + 4 \times \det \begin{bmatrix} r & 2 \\ 0 & -3 \end{bmatrix}$$

$$= 3 \times (2 \cdot 5 - (-4 \cdot -3)) - 1 \times (5 \cdot r - (-4 \cdot 0)) + 4 \times (r \cdot -3 - (0 \cdot 2))$$

$$= 3 \times (10 - (12)) - (5r - 0) + 4 \times (-3r - 0)$$

$$= 3 \times -2 - (5r) + 4 \times (-3r)$$

$$= -6 - 5r - 12r$$

$$-17r - 6$$

(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]

We set $-17r - 6 = det(M) = 0$, since if the determinant is 0, the matrix is not invertible. Using algebra, we solve for r and see that $r = -6/17$ causing the matrix to be invertible. If the matrix is not invertible, then it's a singular matrix, meaning some, but not all, rows and columns are linearly independent, and it's rank is less than the order of M.

(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 $r = 0$.

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

1.2 Eigenvalues and Eigenvectors [5+15pts]

1.2.1 Eigenvalues [5pts]

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

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

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

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

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

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

$$\det(\mathbf{A} - \lambda\mathbf{I}) = (a-\lambda)(c-\lambda) - b^2$$

$$\det(\mathbf{A} - \lambda\mathbf{I}) = ac - a\lambda - c\lambda + \lambda^2 - b^2$$

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

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

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]

$$\lambda^2 - (11+5) + (11 \cdot 5 - 4^2) = 0$$

$$\lambda^2 - 16\lambda + 39 = 0$$

$$\lambda_1 = 3$$

$$\lambda_2 = 13$$

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

$$(A - \lambda I)\mathbf{x} = 0$$

Eigenvector for $\lambda_1 = 3$

$$\mathbf{A} - \lambda_1\mathbf{I} = \begin{bmatrix} 8 & 4 \\ 4 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 8 & 4 \\ 4 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$8x_1 + 4x_2 = 0$$
$$4x_1 + 2x_2 = 0$$

Performing RREF on A:

$$A = \begin{bmatrix} 1 & 1/2 \\ 0 & 0 \end{bmatrix}$$

$$x_2 = s_1$$

$$x_1 + \frac{1}{2}s_1 = 0 \rightarrow x_1 = -\frac{1}{2}s_1$$

$$\mathbf{x}_1 = \begin{bmatrix} -\frac{1}{2}s_1 \\ s_1 \end{bmatrix} \rightarrow s_1 \begin{bmatrix} -\frac{1}{2} \\ 1 \end{bmatrix}$$

Eigenvector for $\lambda_2 = 13$

$$\mathbf{A} - \lambda_2 \mathbf{I} = \begin{bmatrix} -2 & 4 \\ 4 & -8 \end{bmatrix}$$

$$\begin{bmatrix} -2 & 4 \\ 4 & -8 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$-2y_1 + 4y_2 = 0$$
$$4y_1 - 8y_2 = 0$$

Performing RREF on A:

$$A = \begin{bmatrix} 1 & -2 \\ 0 & 0 \end{bmatrix}$$

$$x_2 = s_1$$

$$x_1 - 2s_1 = 0 \rightarrow x_1 = 2s_1$$

$$\mathbf{x}_2 = \begin{bmatrix} 2s_1 \\ s_1 \end{bmatrix} \rightarrow s_1 \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Normalized Eigenvector Formula:

$$\frac{1}{\|\mathbf{x}\|}\mathbf{x}$$

Magnitude Formula:

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2}$$

Normalized Eigenvectors:

$$\|\mathbf{x}_1\| = \sqrt{-\frac{1}{2}^2 + 1^2} = \frac{\sqrt{5}}{2} = 1.11803$$

$$\mathbf{v_1} = \frac{1}{1.11803}\begin{bmatrix} -\frac{1}{2} \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} -0.447215 \\ 0.89443 \end{bmatrix}$$

$$\|\mathbf{x}_2\| = \sqrt{2^2 + 1^2} = \sqrt{5} = 2.23607$$

$$\mathbf{v_2} = \frac{1}{2.23607}\begin{bmatrix} 2 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 0.89443 \\ 0.447213 \end{bmatrix}$$

(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]

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

Properties that allow for this include linear dependence and singularity of the matrix $(\mathbf{A} - \lambda \mathbf{I})$. All columns of the matrix $(\mathbf{A} - \lambda \mathbf{I})$ are linearly dependent and introducing the eigenvector to expresssion and setting it equal to zero means the eigenvector lies in the null space of the matrix. The eigenvector is mapped to the zero vector by this matrix expression, even though both are non-zero. Additionally, the matrix is singular because there's a non-trivial solution to the matrix expression $(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = 0$ (the eigenvector forming a basis for the matrix's null space).

The determinant is set equal to zero in the expression $\det\mathbf{A} - \lambda\mathbf{I} = 0$, because this expression has a non-trivial solution, requiring a singular matrix. We can only find these solutions when the determinant is zero.

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

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

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

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

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

Because both X and Y are independent, the covariance term drops to zero:

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

Plugging into this equation, we obtain:

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

We use the following covariance property to obtain:

$$Cov(X, Z) = Cov(X, 4X + 2Y) = 4Cov(X, X) + 2Cov(X, Y)$$
$$= 4Var(X) + 2Cov(X, Y)$$

Because X and Y are independent, the covariance reduces to 0 once more:

$$= 4Var(X)$$

Solving, we get:

$$Cov(X, Z) = 4 \cdot 0.36c^2 = 1.44c^2$$

We can now finally solve for the correlation coefficient:

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

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. $\min_{x,y} \; f(x,y) = 2x + 3xy$) 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).

(b) List the names of all 4 groups of KKT conditions and their 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 when applicable. Candidate points must satisfy all the conditions mentioned in part b). [8pts]

(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]

(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).

(a) Maximization Problem:

$$L(m,s,\boldsymbol{\lambda}) = f(x,y) - \lambda_1 g_1(x,y) - \lambda_2 g_2(x,y) = 2x + 3xy - \lambda_1(x^2 + 4y^2 - 9) - \lambda_2(y - \frac{1}{2})$$

Minimization Problem:

$$L(m,s,\boldsymbol{\lambda}) = f(x,y) + \lambda_1 g_1(x,y) + \lambda_2 g_2(x,y) = 2x + 3xy + \lambda_1(x^2 + 4y^2 - 9) + \lambda_2(y - \frac{1}{2})$$

(b) Because we have 2 constraints, we need to consider 2 sets of the following 4 KKT conditions:
Stationarity:

$$\frac{\partial L}{\partial x} = 2 + 3y - 2\lambda_1 x = 0$$
$$\frac{\partial L}{\partial y} = 3y - 8\lambda_1 y - \lambda_2 = 0$$

(Note: I did my partial derivative wrong, and didn't realize one of the y's was supposed to be an x until much later. Please spare me.)

Primal Feasibility:

$$g_1(x, y) = x^2 + 4y^2 - 9 \leq 0$$
$$g_2(x, y) = y - \frac{1}{2} \leq 0$$

Dual Feasibility:

$$\lambda_1 \geq 0$$
$$\lambda_2 \geq 0$$

Complementary Slackness:

$$\lambda_1(x^2 + 4y^2 - 9) = 0$$
$$\lambda_2(y - \frac{1}{2}) = 0$$

(c) To find the values of $\boldsymbol{\lambda}$ and $\mathbf{x}$, we have to solve the systems of equations formed by the 4 above conditions for both the active and inactive constraints. 4 equations (Stationarity and Complementary Slackness) and 4 unknowns.
Let's first consider where both constraints are binding:

$$g_1(x, y) = x^2 + 4y^2 = 9$$
$$\lambda_1 > 0$$

$$g_2(y) = y = \frac{1}{2}$$
$$\lambda_2 > 0$$

$$g_1(x, \frac{1}{2}) = x^2 + 4(\frac{1}{2})^2 = 9 \rightarrow x^2 - 8 = 0 \rightarrow x_1 = -2\sqrt{2}, x_2 = 2\sqrt{2}$$

Condition 1: $x = -2\sqrt{2}$:

$$2 + 3y - 2\lambda_1 x = 0 \rightarrow 2 + 3(\frac{1}{2}) - 2\lambda_1(-2\sqrt{2}) = 0 \rightarrow \lambda_1 = \frac{-7\sqrt{2}}{16}$$

which does not satisfy the dual feasibility.
Condition 2: $x = 2\sqrt{2}$:

$$2 + 3y - 2\lambda_1 x = 0 \rightarrow 2 + 3(\frac{1}{2}) - 2\lambda_1(2\sqrt{2}) = 0 \rightarrow \lambda_1 = \frac{7\sqrt{2}}{16}$$

which does satisfy the dual feasibility.
Now we check to see if this lambda satisfies the second condition:

$$3y - 8\lambda_1 y - \lambda_2 = 0 \rightarrow 3(\frac{1}{2}) - 8(\frac{7\sqrt{2}}{16})(\frac{1}{2}) - \lambda_2 = 0 \rightarrow \lambda_2 = \frac{3}{2} - \frac{7\sqrt{2}}{4} < 0$$

which does not satisfy the dual feasibility.

A candidate point was unable to be located when both constraints are active.

Let's now try when constraint 1 is active, constraint 2 inactive:

$$g_1(x,y) = x^2 + 4y^2 = 9$$
$$\lambda_1 > 0$$
$$y = \pm\sqrt{\frac{9-x^2}{4}}$$

$$g_2(y) = y < \frac{1}{2}$$
$$\lambda_2 = 0$$

Because $\lambda_2 = 0$, we have for the partial derivative of the second Lagrange equation:

$$\frac{\partial L}{\partial y} = 3y - 8\lambda_1 y - \lambda_2 = 0 \rightarrow \frac{\partial L}{\partial y} = 3y - 8\lambda_1 y = 0$$

Solving for λ_1, we get:

$$3\left(\sqrt{\frac{9-x^2}{4}}\right) - 8\lambda_1\left(\sqrt{\frac{9-x^2}{4}}\right) = 0 \rightarrow 3\left(\frac{\sqrt{9-x^2}}{2}\right) - 8\lambda_1\left(\frac{\sqrt{9-x^2}}{2}\right) = 0$$

$$\rightarrow 3\left(\frac{\sqrt{9-x^2}}{2}\right) = 8\lambda_1\left(\frac{\sqrt{9-x^2}}{2}\right)$$

Dividing $(\frac{\sqrt{9-x^2}}{2})$ on both sides we get:

$$\lambda_1 = \frac{3}{8}$$

which satisfies dual feasibility.
Now we solve for x:

$$\frac{\partial L}{\partial x} = 2 + 3y - 2\lambda_1 x = 0 \rightarrow 2 + 3\left(\frac{\sqrt{9-x^2}}{2}\right) - 2\left(\frac{3}{8}\right)x = 0$$

This requires a lot of algebraic manipulation to solve, but the goal is to get this equation in quadratic form to solve for the roots. Doing so, we obtain:

$$45x^2 - 45x - 260 = 0$$

with the only valid root being:

$$x = 2.995$$

So for the positive root case of y, we have $x = 2.995$. Solving for y via $\frac{\sqrt{9-x^2}}{2}$, we obtain:

$$y = 0.087$$

So our first potential candidate point is:

$$(x,y) = (2.995, 0.087)$$

However, this point violates the first stationarity condition, so it's not a valid candidate point.
For the negative root case of y, we have the following equation:

$$\frac{\partial L}{\partial x} = 2 + 3y - 2\lambda_1 x = 0 \rightarrow 2 - 3\left(\frac{\sqrt{9-x^2}}{2}\right) - 2\left(\frac{3}{8}\right)x = 0$$

Again, we obtain a quadratic equation of the following form:

$$45x^2 - 48x - 260 = 0$$

with the only valid root being:

$$x = -1.929$$

So for the negative root case of y, we have $x = -1.929$. Solving for y via $-\frac{\sqrt{9-x^2}}{2}$, we obtain:

$$y = -1.1488$$

So our next potential candidate point is:

$$(x, y) = (-1.929, -1.1488)$$

This point holds for all conditions. Therefore, our first candidate point is:

$$(x^*, y^*) = (-1.929, -1.1488)$$

Two candidate points were considered and only one was located when constraint 1 is active, constraint 2 inactive.

Let's now try when constraint 1 is inactive, constraint 2 active:

$$g_1(x, y) = x^2 + 4y^2 < 9$$

$$\lambda_1 = 0$$

$$g_2(y) = y = \frac{1}{2}$$

$$\lambda_2 > 0$$

Because $\lambda_1 = 0$, we have for the partial derivative of the second Lagrange equation:

$$\frac{\partial L}{\partial y} = 3y - 8\lambda_1 y - \lambda_2 = 0 \rightarrow \frac{\partial L}{\partial y} = 3y - \lambda_2 = 0$$

Solving for λ_2, we get:

$$3\left(\frac{1}{2}\right) - \lambda_2 = 0 \rightarrow \lambda_2 = \frac{3}{2}$$

which satisfies dual feasibility. However, in solving for x, we notice that because λ_1, which is 0, is multiplied by the only stationarity condition that contains an x variable, causing the entire term to be zero by default. This trivially invalidates the stationarity condition, which means no candidate points exist in this scenario.

Onto the last case where both constraints are inactive, we have:

$$g_1(x, y) = x^2 + 4y^2 < 9$$

$$\lambda_1 = 0$$

$$g_2(y) = y < \frac{1}{2}$$

$$\lambda_2 = 0$$

We have for the partial derivative of the second Lagrange equation:

$$\frac{\partial L}{\partial y} = 3y - 8\lambda_1 y - \lambda_2 = 0 \rightarrow \frac{\partial L}{\partial y} = 3y = 0$$

Which means we have:

$$y = 0$$

which also satisfies the second constraint. However, we run into another problem with the stationarity condition when solving for x, which reflects what occurred in the previous case:

$$2 + 3(0) - 2(0)x = 0 \rightarrow 2 + 0 - 0 = 0 \rightarrow 2 = 0$$

which is false, so this doesn't work either.

(d) The only candidate point for this set of constraints given this objective function is:

$$(x^*, y^*) = (-1.929, -1.1488)$$

(e) The only candidate point is $(x^*, y^*) = (-1.929, -1.1488)$, where each second derivative can be calculated as the following, where $\lambda_1 = \frac{3}{8}$:

$$\frac{\partial^2 L}{\partial x^2} = -2\lambda_1 = -\frac{3}{4}$$

$$\frac{\partial^2 L}{\partial y^2} = 3 - 8\lambda_1 = -5.375$$

$$\frac{\partial^2 L}{\partial xy} = 3$$

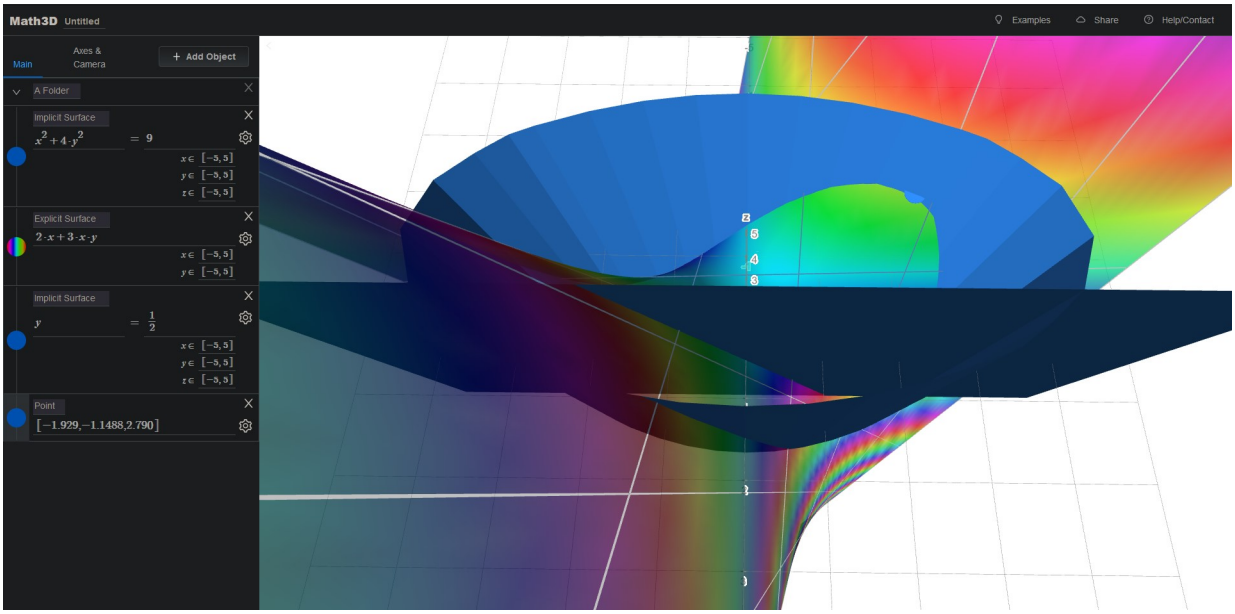
$$\frac{\partial^2 L}{\partial yx} = 0$$

Computing the Hessian, we obtain:

$$H = -\frac{3}{4}(-5.375) - 3 \cdot 0 = 4.03125 > 0$$

which means that this point is either a maximum or minimum if $f_x x > 0$ or $f_x x < 0$, respectively. Because $\frac{\partial^2 L}{\partial x^2} = -\frac{3}{4}$ is less than 0, we see that this candidate point is a maximum point and is therefore concave.

(f) Here, we see that given the bounds provided by both constraint functions, the blue point plotted visually demonstrates that there is in fact only a maximum located at $(-1.929, -1.1488, 2.790)$

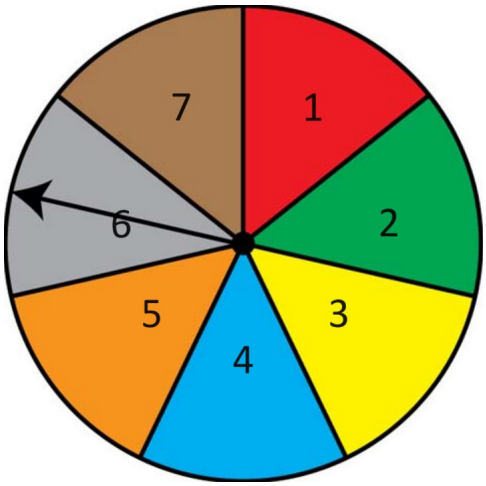


4 Maximum Likelihood [10pts + 10pts Grads / 6% Bonus for Undergrads]

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 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



Let $Y :=$ Spinner lands on 1,2,...,6 and then on 7

Let $X_i :=$ Spinner lands on ith digit where $1 \leq i \leq 7$

Let $\theta :=$ Probability of spinner landing on section 1

Then we obtain the following joint probability of the data set:

$$P_Y(X_i, \theta) = \left(\prod_{i=1}^{6} P_X(x_i) \right) \cdot \theta$$

We then observe that $P_X(x_i) = 1 - \theta$, because our problem resolves strictly around the probability of the spinner landing on section 1, therefore simplifying our joint probability of the data set to the following likelihood function:

$$L(\theta) = P_Y(\theta) = (1 - \theta)^6 \cdot \theta$$

Applying the log function to both sides yields:

$$\log(L(\theta)) = \log((1 - \theta)^6 \cdot \theta) = 6 \log(1 - \theta) + \log(\theta)$$

Taking the derivative of the log-likelihood function wrt θ and setting it equal to 0 allows us to solve for a maximum likelihood estimator:

$$\frac{\partial}{\partial\theta}6\log(1-\theta)+\frac{\partial}{\partial\theta}\log(\theta)=0$$

$$=6\frac{\partial}{\partial\theta}\log(1-\theta)+\frac{1}{\ln(2)\theta}$$

$$=-6\left(\frac{1}{\ln(2)(1-\theta)}\right)+\frac{1}{\ln(2)\theta}$$

$$=-\left(\frac{6}{\ln(2)(1-\theta)}\right)+\frac{1}{\ln(2)\theta}$$

$$=\frac{1}{\ln(2)}\left(-\frac{6}{(1-\theta)}+\frac{1}{\theta}\right)$$

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

$$=\theta(1-\theta)\left(-\frac{6}{(1-\theta)}+\frac{1}{\theta}\right)$$

$$=-6\theta+(1-\theta)$$

$$-6\theta+1-\theta=0$$

$$\theta(-6-1)+1=0$$

$$-7\theta+1=0$$

$$-7\theta=-1$$

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

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 λ given x_1 is an observed value of X_1 ?[2 pts / 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 λ given this data? You may leave your answer in product form. [2 pts / 1%]

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

(a)

$$L(\lambda) = L(\lambda|x_1) = p_X(X = x_1|\lambda) = \frac{\lambda^{x_1} e^{-\lambda}}{x_1!}$$

(b) The likelihood function can be written as:

$$L(\lambda) = L(\lambda|\boldsymbol{x}) = \prod_{i=1}^{n} p_X(X = x_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum i=1^n x_i}}{\prod_{i=1}^n x_i!}$$

The log-likelihood can be obtained via the following steps:

$$= \ln(L(\lambda)) = \ln\left(e^{-n\lambda} \frac{\lambda^{\sum i=1^n x_i}}{\prod_{i=1}^n x_i!}\right)$$

$$= \ln\left(e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}\right) - \ln\left(\prod_{i=1}^{n} x_i!\right)$$

$$= \ln\left(e^{-n\lambda}\right) + \ln\left(\lambda^{\sum_{i=1}^n x_i}\right) - \ln\left(\prod_{i=1}^{n} x_i!\right)$$

$$= -n\lambda + \left(\sum_{i=1}^{n} x_i\right) \ln(\lambda) - \ln\left(\prod_{i=1}^{n} x_i!\right)$$

(c) Taking the derivative wrt λ and setting it equal to zero and solving for λ, we have:

$$\frac{\partial}{\partial\lambda}\left(-n\lambda + \left(\sum_{i=1}^{n} x_i\right) \ln(\lambda) - \ln\left(\prod_{i=1}^{n} x_i!\right)\right) = 0$$

$$= \frac{\partial}{\partial\lambda}(-n\lambda) + \frac{\partial}{\partial\lambda}\left(\left(\sum_{i=1}^{n} x_i\right) \ln(\lambda)\right) - \frac{\partial}{\partial\lambda}\left(\ln\left(\prod_{i=1}^{n} x_i!\right)\right)$$

$$= -n + \left(\sum_{i=1}^{n} x_i\right) \frac{1}{\lambda}$$

$$\left(\sum_{i=1}^{n} x_i\right) \frac{1}{\lambda} = n$$

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

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

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]

$$H(Y) = E(I(y)) = \sum P(Y=y)\log_2\left(\frac{1}{P(y)}\right) = P(Y=T)\log_2\left(\frac{1}{p(y)}\right) + P(Y=F)\log_2\left(\frac{1}{p(y)}\right)$$

$$= \frac{1}{2}\log_2\frac{1}{0.5} + \frac{1}{2}\log_2\frac{1}{0.5}$$

$$= 1$$

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

$$H(Y|X_1) = \sum_{x \in X_1} p(X_1) H(Y|X_1 = x)$$

First we find $H(Y|X_1 = L)$:

$$H(Y|X_1 = L) = E(I(Y|X_1 = L)) = \sum_{y \in Y} P(Y|X_1 = L)\log_2\frac{1}{P(Y|X_1 = L)}$$

$$= P(Y = T|X_1 = L)\log_2\frac{1}{P(Y = T|X_1 = L)} + P(Y = F|X_1 = L)\log_2\frac{1}{P(Y = L|X_1 = L)}$$

$$= \frac{2}{5}\cdot\log_2\frac{5}{2} + \frac{3}{5}\log_2\frac{5}{3}$$

$$= 0.971$$

Next we find $H(Y|X_1 = M)$:

$$H(Y|X_1 = M) = E(I(Y|X_1 = M)) = \sum_{y \in Y} P(Y|X_1 = M)\log_2\frac{1}{P(Y|X_1 = M)}$$

$$= P(Y = T|X_1 = M) \log_2 \frac{1}{P(Y = T|X_1 = M)} + P(Y = F|X_1 = M) \log_2 \frac{1}{P(Y = L|X_1 = M)}$$
$$= \frac{1}{2} \cdot \log_2 \frac{2}{1} + \frac{1}{2} \log_2 \frac{2}{1}$$
$$= 1$$

Finally we find $H(Y|X_1 = H)$:

$$H(Y|X_1 = H) = E(I(Y|X_1 = H)) = \sum_{y \in Y} P(Y|X_1 = H) \log_2 \frac{1}{P(Y|X_1 = H)}$$

$$= P(Y = T|X_1 = H) \log_2 \frac{1}{P(Y = T|X_1 = H)} + P(Y = F|X_1 = H) \log_2 \frac{1}{P(Y = L|X_1 = H)}$$
$$= \frac{2}{3} \cdot \log_2 \frac{3}{2} + \frac{1}{3} \log_2 \frac{3}{1}$$
$$= 0.918$$

Now we can finally compute $H(Y|X_1)$ via the following calculation:

$$= p(X_1 = L)H(Y|X_1 = L) + p(X_1 = M)H(Y|X_1 = M) + p(X_1 = H)H(Y|X_1 = H)$$
$$= \frac{1}{2} \cdot 0.971 + \frac{1}{5} \cdot 1 + \frac{3}{10} \cdot 0.918$$
$$= 0.961$$

$$H(Y|X_2) = \sum_{x \in X_2} p(X_2)H(Y|X_2)$$

First we find $H(Y|X_2 = T)$:

$$H(Y|X_2 = T) = E(I(Y|X_2 = T)) = \sum_{y \in Y} P(Y|X_2 = T) \log_2 \frac{1}{P(Y|X_2 = T)}$$

$$= P(Y = T|X_2 = T) \log_2 \frac{1}{P(Y = T|X_2 = T)} + P(Y = F|X_2 = T) \log_2 \frac{1}{P(Y = F|X_2 = T)}$$
$$= \frac{1}{4} \cdot \log_2 \frac{4}{1} + \frac{3}{4} \log_2 \frac{4}{3}$$
$$= 0.811$$

Then we find $H(Y|X_2 = F)$:

$$H(Y|X_2 = F) = E(I(Y|X_2 = F)) = \sum_{y \in Y} P(Y|X_2 = F) \log_2 \frac{1}{P(Y|X_2 = F)}$$

$$= P(Y = T|X_2 = F) \log_2 \frac{1}{P(Y = T|X_2 = F)} + P(Y = F|X_2 = F) \log_2 \frac{1}{P(Y = F|X_2 = F)}$$
$$= \frac{2}{3} \cdot \log_2 \frac{3}{2} + \frac{1}{3} \log_2 \frac{3}{1}$$
$$= 0.918$$

Now we can finally compute $H(Y|X_2)$ via the following calculation:

$$= p(X_2 = T)H(Y|X_2 = T) + p(X_2 = F)H(Y|X_2 = F)$$
$$= \frac{2}{5} \cdot 0.811 + \frac{3}{5} \cdot 0.918$$
$$= 0.875$$

(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]

$$I(X_1, Y) = H(Y) - H(Y|X_1) = 1 - 0.961 = 0.039$$

$$I(X_2, Y) = H(Y) - H(Y|X_2) = 1 - 0.875 = 0.125$$

Because $I(X_2, Y) > I(X_1, Y)$ and knowing that mutual information is symmetric in that $I(X_i, Y) = I(Y, X_i)$, it can be said that we become more informed about Y after seeing feature X_2.

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

$$H(Y, X_3) = \sum_{y, x_3} P(Y = y, X_3 = x) \log_2 \frac{1}{P(Y = y, X_3 = x)}$$

$$= P(Y = T, X_3 = T) \log_2 \frac{1}{P(Y = T, X_3 = T)} + P(Y = T, X_3 = F) \log_2 \frac{1}{P(Y = T, X_3 = F)} +$$

$$P(Y = F, X_3 = T) \log_2 \frac{1}{P(Y = F, X_3 = T)} + P(Y = F, X_3 = F) \log_2 \frac{1}{P(Y = F, X_3 = F)}$$

$$= \frac{3}{10} \log_2 \frac{10}{3} + \frac{1}{5} \log_2 \frac{5}{1} + \frac{1}{10} \log_2 \frac{10}{1} + \frac{2}{5} \log_2 \frac{5}{2}$$

$$= 1.847$$

5.2 Entropy Proofs [10pts]

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

$$H(Y) = E(I(Y)) = \sum P(Y = y) \log_2 \left(\frac{1}{P(y)} \right)$$

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

(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. You may use any theorem/ proof from the slides without having to re-prove it). [7pts]

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

Introduction: It can be shown via direct proof that $I(X, Y) = 0$ if X and Y are statistically independent, which can be rewritten as *if X and Y are statistically independent, then $I(X, Y) = 0$.*

Body: If X and Y are statistically independent, then no information about X tells you about Y, and vice versa. This means their mutual probability relation can be written as the following:

$$P(X, Y) = P(X \cap Y) = P(X)P(Y)$$

This relation holds for their respective entropies in the following form:

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

From this point forward, we can observe that the formula for mutual information can be written as both of the following based on the symmetric property:

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

and

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

thus completes the proof because $H(X|Y)$ becomes $H(X)$, as well as for Y.

Conclusion: It was proven directly that If X and Y are statistically independent, then no information about X tells you about Y based on the fact that independence of both variables in the mutual information formula causes the information value to be reduced to zero. This makes sense, because if the joint relation of both variables is insignificant, then no information can possibly be extracted by their relation.

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 $\mathbf{y} = \mathbf{xA}$. $\mathbf{y}$ 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 $\mathbf{x_D}$:

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

The original equation can be rewritten as:

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

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

(a) 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.

(b) 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.

B, D

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%]

$$Var(P+Q) = Var(P) + Var(Q) + 2Cov(P,Q)$$
$$= [Var(2X+5XY^2)] + [Var(X)] + [2Cov(2X+5XY^2, X)]$$
$$= [2^2 Var(X) + 5^2 Var(XY^2) + 2\cdot 2\cdot 5 Cov(XY^2, X)] + [Var(X)] + [2(2Cov(X,X) + 5Cov(XY^2, X)]$$
$$= 4Var(X) + 25Var(XY^2) + 20Cov(XY^2, X) + Var(X) + 4Cov(X,X) + 10Cov(XY^2, X)$$
$$= 4Var(X) + 25Var(XY^2) + 20Cov(XY^2, X) + Var(X) + 4Var(X) + 10Cov(XY^2, X)$$
$$= 9Var(X) + 25Var(XY^2) + 30Cov(XY^2, X)$$
$$= 9\cdot 1 + 25(E[(XY^2)^2] - E[XY^2]^2) + 30(E[XY^2\cdot X] - E[XY^2]E[X])$$
$$= 9 + 25E[X^2Y^4] - 25E[XY^2]^2 + 30E[X^2Y^2] - 30E[XY^2]\cdot 0$$
$$= 9 + 25E[X^2Y^4] - 25E[XY^2]^2 + 30E[X^2Y^2]$$

Note that $E(X^2) = Var(X) = 1$, and the moment equation for a standard normal random variable is $E[(X-\mu)^p] = \sigma^p \cdot (p-1)$ if p is even and 0 if odd, and that both X, Y are independent:

$$E[(X-0)^4] = \sigma^4(4-1) = 1^2\cdot 3 = 3$$
$$E[Y^4] = 3$$
$$E[X] = 0$$
$$E[Y^2] = 1$$
$$E[XY^2] = E[X]E[Y^2] = 0$$
$$E[X^2Y^2] = E[X^2]E[Y^2]$$
$$(E[XY^2])^2 = 0$$

Now we continue the simplification:

$$= 9 + 25E[X^2]E[Y^4] - 25E[XY^2]^2 + 30E[X^2Y^2]$$
$$= 9 + 25\cdot 3 - 25\cdot 0 + 30\cdot 1\cdot 1$$
$$= 9 + 75 + 30$$
$$= 114$$

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.

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

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

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%]

$$f_X(x) = \int_0^\infty \frac{1}{24}xe^{-\frac{1}{3}y}dy$$
$$= \frac{1}{24}x\int_0^\infty e^{-\frac{1}{3}y}dy$$
$$= \frac{1}{24}x \cdot \lim_{a\to\infty}\left(\int_0^a e^{-\frac{1}{3}y}dy\right)$$
$$= \frac{1}{24}x \cdot \lim_{a\to\infty}\left(-\frac{3}{e^{\frac{1}{3}a}} + 3\right)$$
$$= \frac{1}{24}x \cdot 3$$
$$= \frac{1}{8}x$$
$$f_Y(y) = \int_0^4 \frac{1}{24}xe^{-\frac{1}{3}y}dx$$
$$= \frac{1}{24}e^{-\frac{1}{3}y}\int_0^4 xdx$$
$$= \frac{1}{24}e^{-\frac{1}{3}y} \cdot 8$$
$$= \frac{1}{3}e^{-\frac{1}{3}y}$$

(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%]

$P(tails) = 1 - P(heads) = 0.75$
The probability of getting no heads in 6 tosses:

$$P(Y = 0) = P(tails)^6 \approx 0.178$$

The probability of getting the first head on the k-th toss is a geometric distribution:

$$P(\text{First head on } k\text{th toss}) = (0.75)^{k-1} \cdot 0.25$$

$$P(Y = 1) = \sum_{k=1}^{6}(0.75)^{k-1} \cdot 0.25$$
$$P(Y = 1) = 1 - (0.75)^6 = 0.822$$
$$P(Y = 0) = 0.822$$

Y follows a Bernoulli distribution with $p = P(Y = 1) = 0.822$ with variance $Var(Y) = p(1-p) = 0.146$

References

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