Fall 2022 CS4641/CS7641 Homework 1

Dr. Mahdi Roozbahani

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] DONE
- 1.2 Characteristic Equation [8pts] DONE: check
- 1.3 Eigenvalues and Eigenvectors [20pts] DONE

Q2: Covariance, Correlation, and Independence [9pts]

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

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

DONE: some parts wrong, debug

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

- 4.1 Discrete Example [10pts] DONE: check
- 4.2 Weibull Distribution [15pts Bonus for Undergrads] DONE: check

Q5: Information Theory [35pts]

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

Q6: Bonus for All [15pts]

TODO

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]

Let M_{ij} be the determinant of the sub-matrix when row i and column j are removed from M.
$|M| = \sum_{j=1}^{d} (-1)^{i+j} a_{ij} M_{ij}$
$|M| = 4(r * 1 - 2 * 7) - 2(-3 * 1 - 2 * 0) + (-3 * 7 - r * 0)$
$|M| = 4r - 71$

(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]

M^{-1} does not exist when $|M| = 0$. Setting $|M| = 4r - 71 = 0$, we find M^{-1} does not exist when $r = \frac{71}{4}$. At this value of r, M is rank deficient (and therefore non-invertible/singular, since M is square). This is illustrated in part *(c)*, since a row/column of M may then be represented as a linear combination of the other rows/columns.

(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]

Yes. We can show this by reducing M to a row echelon form:

- Beginning with M for $r = \frac{71}{4}$:
  $M = \begin{bmatrix} 4 & 2 & 1 \\ -3 & \frac{71}{4} & 2 \\ 0 & 7 & 1 \end{bmatrix}$
- Multiplying the second row of M by $\frac{4}{3}$ and adding the first row, we obtain:
  $M = \begin{bmatrix} 4 & 2 & 1 \\ 0 & \frac{77}{3} & \frac{11}{3} \\ 0 & 7 & 1 \end{bmatrix}$
- Multiplying the second row by $\frac{3}{11}$ and subtracting the second row from the third row, we obtain:
  $M = \begin{bmatrix} 4 & 2 & 1 \\ 0 & 7 & 1 \\ 0 & 0 & 0 \end{bmatrix}$

Since the third row of M has no pivot (is all zeros), M is rank deficient for $r = \frac{71}{4}$. We can equivalently say that any row/column of M can be expressed as a linear combination of the other two rows/columns of M. For example, we have:
$R_2 = -\frac{3}{4} R_1 + \frac{11}{4} R_3$

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

$M^{-}1 = \begin{bmatrix} \frac{14}{71} & -\frac{5}{71} & -\frac{4}{71} \\ -\frac{3}{71} & -\frac{4}{71} & \frac{11}{71} \\ \frac{21}{71} & \frac{28}{71} & -\frac{6}{71} \end{bmatrix}$

(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]

$|M| = 4(0) - 71$
As expected:
$|M^{-1}| = \frac{1}{|M|=} - \frac{1}{71}$

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 $|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0$.

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

$Ax = \lambda x \ , \quad x \neq 0$
$0 = (A - \lambda I)x \ , \quad x \neq 0$

Proof by contradiction:
If $A - \lambda I$ is invertible/non-singular, then it is full rank and its null space is trivial, i.e. $\ker(A - \lambda I) = \{0\}$. Therefore, the equation $(A - \lambda I)x = 0$ has only the trivial solution, $x = 0$.
However, $x \neq 0$, so $A - \lambda I$ must be non-invertible/singular. The determinant of a matrix is non-zero if and only if the matrix is invertible.
Therefore:
$|A - \lambda I| = 0 \quad \square$

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]

$0 = |\boldsymbol{A} - \lambda \boldsymbol{I}|$
$0 = (a - \lambda)(c - \lambda) - b^2$
$0 = \lambda^2 - (a + c)\lambda + ac - b^2$
$\lambda = \frac{(a+c) \pm \sqrt{(a+c)^2 - 4(ac - b^2)}}{2}$
$\lambda = \frac{(a+c) \pm \sqrt{a^2 + 4b^2 + c^2 - 2ac)}}{2}$

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

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

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: $\boldsymbol{v}_i^T \boldsymbol{v}_j \neq \boldsymbol{v}_i \boldsymbol{v}_j^T$

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

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

Since the eigenvectors v_n are orthonormal to each other, $\boldsymbol{V}$ is an orthonormal basis. Therefore:
$\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$

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]

$0 = |\boldsymbol{A} - \lambda \boldsymbol{I}|$
$0 = (x - \lambda)^2 - 25$
$0 = \lambda^2 - 2x\lambda + x^2 - 25$
$\lambda = \frac{2x \pm \sqrt{4x^2 - 4(x^2 - 25)}}{2}$
$\lambda = x \pm 5$

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

For $\lambda_1 = x + 5$:

$(\boldsymbol{A} - \lambda_1 \boldsymbol{I})\boldsymbol{v_1} = 0$

$$\begin{bmatrix} x - (x+5) & 5 \\ 5 & x - (x+5) \end{bmatrix} \boldsymbol{v_1} = 0$$

Resulting in the system:
$-5v_{1,1} + 5v_{1,2} = 0 \quad$ and $\quad 5v_{1,1} - 5v_{1,2} + = 0$

We get the relationship: $v_{1,2} = v_{1,1}$

Let's choose $v_{1,1} = v_{1,2} = 1$ and normalize by a factor of $\sqrt{1^2 + 1^2} = \sqrt{2}$:

$$\boldsymbol{v_1} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

For $\lambda_2 = x - 5$:

$(\boldsymbol{A} - \lambda_1 \boldsymbol{I})\boldsymbol{v_2} = 0$

$$\begin{bmatrix} x - (x-5) & 5 \\ 5 & x - (x-5) \end{bmatrix} \boldsymbol{v_2} = 0$$

Resulting in the system:
$5v_{2,1} + 5v_{2,2} = 0 \quad$ and $\quad 5v_{2,1} + 5v_{2,2} + = 0$

We get the relationship: $v_{2,2} = -v_{2,1}$

Let's choose $v_{2,1} = -v_{2,2} = 1$ and normalize by a factor of $\sqrt{1^2 + (-1)^2} = \sqrt{2}$:

$$\boldsymbol{v_2} = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

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]

$E[X] = 0.6c - 0.4c = 0.2c \quad E[Y] = 0 \quad E[Z] = E[X]\cancelto{0}{E[Y]} = 0$ since X and Y are independent
$cov(Y, Z) = E[YZ] - \cancelto{0}{E[Y]}\cancelto{0}{E[Z]}$
$cov(Y, Z) = E[YZ]$
$cov(Y, Z) = E[XY^2]$
$cov(Y, Z) = E[X]E[Y^2]$ since X and Y are independent
$cov(Y, Z) = E[X](var(Y) + E[Y^2])$
$cov(Y, Z) = 0.2c(1 + 0)$
$cov(Y, Z) = 0.2c$
Yes, values of c ***do*** affect 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]

$var(Z) = var(3X) + var(2Y) + 2\cancelto{0}{cov(3X, 2Y)}$ covariance goes to zero, since X and Y are independent
$var(Z) = 9(5) + 4(15) = 105$

$\rho(X, Z) = \frac{cov(X,Z)}{\sqrt{var(X)var(Z)}}$
$\rho(X, Z) = \frac{cov(X,3X+2Y)}{\sqrt{var(X)var(Z)}}$
$\rho(X, Z) = \frac{3cov(X,X)+2\cancelto{0}{cov(X,Y)}}{\sqrt{var(X)var(Z)}}$
$\rho(X, Z) = \frac{3var(X)}{\sqrt{var(X)var(Z)}}$
$\rho(X, Z) = \frac{(3)(5)}{\sqrt{(5)(105)}}$
$\rho(X, Z) \approx 0.655$

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. $\min_{x,y} \quad f(x,y) = -4y + xy$) 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).

Setting the constraints as equalities (assuming $\lambda > 0$) :

For the maximization problem: $g_1(x,y) = 2x^2 + y^2 - 12$
$g_2(x,y) = x - 1$

$L(x,y,\lambda) = f(x,y) - \lambda_1 g_1(x) - \lambda_2 g_2(x)$
$L(x,y,\lambda_1,\lambda_2) = -4y + xy - \lambda_1(2x^2 + y^2 - 12) - \lambda_2(x-1)$

For the minimization problem: $g_1(x,y) = 2x^2 + y^2 - 12$
$g_2(x,y) = x - 1$

$L(x,y,\lambda) = f(x,y) + \lambda_1 g_1(x) + \lambda_2 g_2(x)$
$L(x,y,\lambda_1,\lambda_2) = -4y + xy + \lambda_1(2x^2 + y^2 - 12) + \lambda_2(x-1)$

By convention, the λ values are positive, so we just flip the signs to change the minimization problem to a maximization problem.

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

- Stationary Condition: $\nabla L(x,y,\lambda) = 0$
  - $\frac{\partial L(x,y,\lambda)}{\partial x} = y - 2\lambda_1 x - \lambda_2 = 0$
  - $\frac{\partial L(x,y,\lambda)}{\partial y} = -4 - 2\lambda_1 y = 0$
  - $\frac{\partial L(x,y,\lambda)}{\partial \lambda_1} = -2x^2 + y^2 - 12 = 0$
  - $\frac{\partial L(x,y,\lambda)}{\partial \lambda_2} = 1 - x = 0$
- Primal Feasibility: $g_i(x^*) \leq 0 \quad \text{for} \quad i = 1, ..., m$
  - $g_1(x) = 2x^2 + y^2 - 12 \leq 0$
  - $g_2(x) = x - 1 \leq 0$
- Dual Feasibility: $\lambda \geq 0$
- Complimentary Slackness: $\lambda_i g_i(x) = 0 \quad \text{for} \quad i = 1, ..., m$
  - $\lambda_1 g_1(x) = \lambda_1(2x^2 + y^2 - 12) = 0$
  - $\lambda_2 g_2(x) = \lambda_2(x - 1) = 0$

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

- Both constraints are binding:
  $\lambda_1 > 0, \quad g_1(x) = 2x^2 + y^2 = 12, \quad \lambda_2(x) > 0, \quad g_2(x) = x = 1$
  - $g_1(x) = 2(1)^2 + y^2 = 12$
    $y = \pm\sqrt{10}$
    Case 1: $y = \sqrt{10}$:
    - $-4 - 2\lambda_1 y = -4 - 2\lambda_1(\sqrt{10}) = 0$
      $\lambda_1 = -2/\sqrt{10}$
    - $y - 2\lambda_1 x - \lambda_2 = \sqrt{10} - 2\lambda_1 - \lambda_2 = 0$
      $\lambda_2 = \sqrt{10} + 4/\sqrt{10}$

    Case 2: $y = -\sqrt{10}$:
    - $-4 - 2\lambda_1 y = -4 - 2\lambda_1(-\sqrt{10}) = 0$
      $\lambda_1 = 2/\sqrt{10}$
    - $y - 2\lambda_1 x - \lambda_2 = -\sqrt{10} - 2\lambda_1 - \lambda_2 = 0$
      $\lambda_1 = -\sqrt{10} - 4/\sqrt{10} > 0$

  This combination cannot satisfy the KKT conditions

- Constraint 1 Binding, Constraint 2 Slack:
  $\lambda_1 > 0, \quad g_1(x) = 2x^2 + y^2 = 12, \quad \lambda_2(x) = 0, \quad g_2(x) = x < 1$
  - $g_1(x) = 2x^2 + y^2 = 12$
    $x = \cancel{\pm}\sqrt{12 - y^2}$ because $x < 1$

    Case 1: $x = -\sqrt{12 - y^2}$:
    - $y - 2\lambda_1 x - \cancel{\lambda_2}^{0} = 0$
      $\lambda_1 = \frac{y}{2x} = -\frac{y}{2\sqrt{12-y^2}}$
    - $-4 - 2\lambda_1 y = 0$
      $-4 + \frac{2y^2}{2\sqrt{12-y^2}} = 0$
      $y^4 + 16y^2 - (16)(12) = 0$
      $y = \pm 2\sqrt{2}, \pm 2\sqrt{6}i$
      λ_1 is positive for $y = -2\sqrt{2}$, but the other candidate points are rejected for generating negative or imaginary values of λ_1

  Candidate points found for this combination!
- Constraint 1 Slack, Constraint 2 Binding:
  $\lambda_1 = 0, \quad g_1(x) = 2x^2 + y^2 < 12, \quad \lambda_2(x) > 0, \quad g_2(x) = x = 1$
  - $-4 - \cancel{2\lambda_1 y}^{0} = 0$

  This combination cannot satisfy the KKT conditions
- Constraint 1 Slack, Constraint 2 Slack:
  $\lambda_1 = 0, \quad g_1(x) = 2x^2 + y^2 < 12, \quad \lambda_2(x) = 0, \quad g_2(x) = x < 1$
  - $-4 - \cancel{2\lambda_1 y}^{0} = 0$

  This combination cannot satisfy the KKT conditions

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

Candidate Points found in previous part: $x = -\sqrt{12 - y^2} = -2, y = -2\sqrt{2}$

(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]

I will proceed with the one candidate point I found, $x = -2, y = -2\sqrt{2}$:

$\lambda_1 = \frac{y}{2x} = \frac{\sqrt{2}}{2}$

$\lambda_2 = 0$

$H_L = \begin{bmatrix} \frac{\partial^2 L}{\partial x^2} & \frac{\partial^2 L}{\partial x \partial y} \\ \frac{\partial^2 L}{\partial y \partial x} & \frac{\partial^2 L}{\partial y^2} \end{bmatrix} = \begin{bmatrix} -4\lambda_1 & 1 \\ 1 & -2\lambda_1 \end{bmatrix} = \begin{bmatrix} -2\sqrt{2} & 1 \\ 1 & -\sqrt{2} \end{bmatrix}$

$|H_L| = (2\sqrt{2}(2) - 1) = 3$ Showing that $L(x, y)$ is concave at this point

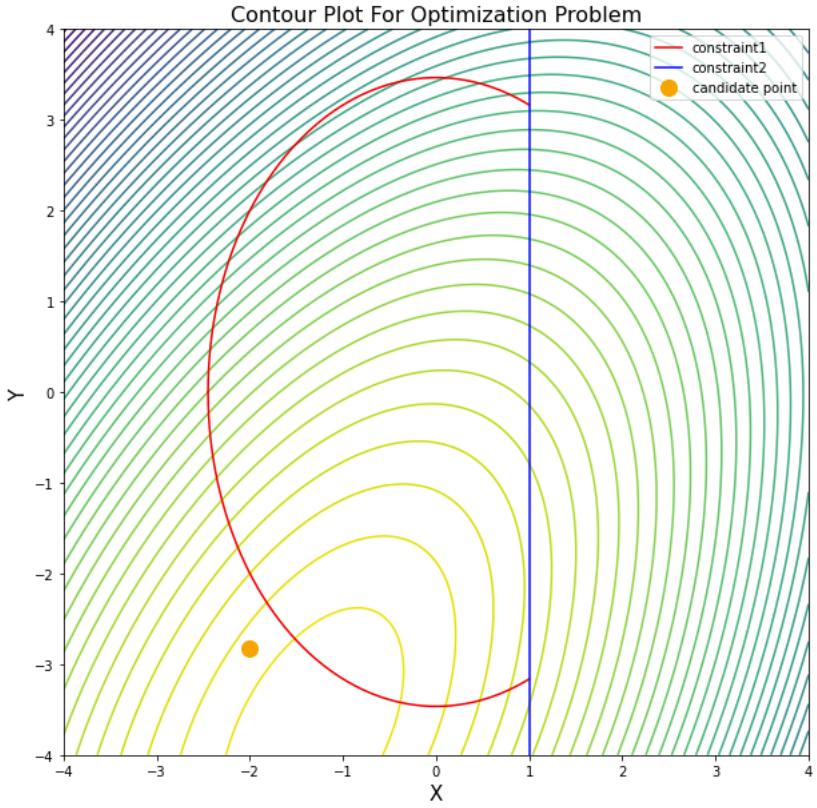
$\frac{\partial^2 L}{\partial x^2} = -2\sqrt{2} < 0$ So we know our candidate point is a local maxima

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).

(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



My plot and/or candidate point don't look quite right here, but I ran out of time to debug my results.
COPY AND PASTE THIS TEXT INTO HW1 FOR CREDIT TODO 1 objective$_f unction = lambda x, y : -4 * y + x * y - np.sqrt(2)/2 * (2 * x * *2 + y * *2 - 12)$

TODO 2 x$_s tart = -2.449489742783178 x_e nd = 1$

TODO 3 constraint1$_f unction_n egative = lambda x : -np.sqrt(12-2*x**2) constraint1_f unction_p ositive = lambda x : np.sqrt(12 - 2 * x * *2)$

TODO 4 constraint$_{2p} osition = 1$

TODO 5 x$_v als_c andidate = [-2] y_v als_c andidate = [-2.8284271247461903]$

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

Roll 6-sided die 6 times. Shreeya wins if she gets anything except 3 the first 5 rolls and 3 on the 6th roll. Marion wins otherwise. Shreeya can tune θ, the likelihood of getting a 3.

$$p(X|\theta) = \left\{ \begin{array}{ll} (1-\theta)/3 & \text{if } x = 1 \\ (1-\theta)/6 & \text{if } x = 2 \\ \theta & \text{if } x = 3 \\ (1-\theta)/6 & \text{if } x = 4 \\ (1-\theta)/6 & \text{if } x = 5 \\ (1-\theta)/6 & \text{if } x = 6 \end{array} \right\} \tag{1}$$

Since we only care about whether we roll a 3 or not a 3, we can instead treat the die like a Bernoulli RV!

$$p(X|\theta) = \left\{ \begin{array}{ll} \theta & \text{if } x = 3 \\ (1-\theta) & \text{if } x \neq 3 \end{array} \right\} \tag{2}$$

Since our "Bernoulli-like" RV is not actually binary, I use the indicator function for notation below:

$$\mathbb{1}_3(X_i) = \left\{ \begin{array}{ll} 1 & \text{if } X_i = 3 \\ 0 & \text{if } X_i \neq 3 \end{array} \right\} \tag{3}$$

Now, we can perform our MLE for θ:
$L(\theta) = \prod_{i=1}^{6} p(X_i|\theta)$
$L(\theta) = \prod_{i=1}^{6} \theta^{\mathbb{1}_3(X_i)}(1-\theta)^{1-\mathbb{1}_3(X_i)}$

$\hat{\theta} = \text{argmax}_\theta L(\theta)$
$LL(\theta) = \log L(\theta)$
$LL(\theta) = \log \prod_{i=1}^{6} p(X_i|\theta)$
$LL(\theta) = \sum_{i=1}^{6} \log p(X_i|\theta)$
$LL(\theta) = \sum_{i=1}^{6} \mathbb{1}_3(X_i) \log \theta + (1 - \mathbb{1}_3(X_i)) \log(1-\theta)$
$LL(\theta) = \log \theta + 5 \log(1-\theta)$
$\frac{\partial LL(\theta)}{\partial \theta} = \frac{1}{\theta} - \frac{5}{1-\theta} = 0$
$(1-\theta) - 5\theta = 0$
$\hat{\theta} = \frac{1}{6}$

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]

$L(x_1, \lambda, k) = \prod_{i=1}^{1} P(X = x_i|\lambda, k)$
$L(x_1, \lambda, k) = \frac{k}{\lambda}(\frac{x_1}{\lambda})^{k-1} e^{-(x_1/\lambda)^k}$

(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]

$L(x_i, \lambda, k) = \prod_{i=1}^{n} P(X = x_i|\lambda, k)$
$L(x_i, \lambda, k) = \prod_{i=1}^{n} \frac{k}{\lambda}(\frac{x_i}{\lambda})^{k-1} e^{-(x_i/\lambda)^k}$

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

$\hat{\lambda} = \text{argmax}_\lambda L(\lambda, k)$
$LL(\lambda, k) = \log L(\lambda)$
$LL(\lambda, k) = \log \prod_{i=1}^{n} \frac{k}{\lambda}(\frac{x_i}{\lambda})^{k-1} e^{-(x_i/\lambda)^k}$
$LL(\lambda, k) = n \log k - nk \log \lambda - \frac{1}{\lambda^k} \sum_{i=1}^{n} x_i^k + (k-1) \sum_{i=1}^{n} \log x_i$

$\frac{\partial LL(\lambda,k)}{\partial \lambda} = -\frac{nk}{\lambda} + \frac{k}{\lambda^{k+1}} \sum_{i=1}^{n} x_i^k = 0$

$\hat{\lambda} = \sqrt[k]{(\frac{1}{n} \sum_{i=1}^{n} x_i^k)}$

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]

$$p(X) = \sum_j p(x_i, y_j) = \left\{ \begin{array}{ll} \frac{5}{16} & \text{if } x = 0 \\ \frac{11}{16} & \text{if } x = 1 \end{array} \right\} \tag{4}$$

$$p(Y) = \sum_i p(x_i, y_j) = \left\{ \begin{array}{ll} \frac{1}{4} & \text{if } y = 0 \\ \frac{3}{4} & \text{if } y = 1 \end{array} \right\} \tag{5}$$

(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]

$I(X,Y) = \sum_i \sum_j p(x_i, y_j) \log_2(\frac{p(x_i,y_j)}{p(x_i)p(y_j)})$
$I(X,Y) = \frac{1}{16} \log_2(\frac{\frac{1}{16}}{\frac{5}{16}\frac{1}{4}}) + \frac{1}{4} \log_2(\frac{\frac{1}{4}}{\frac{5}{16}\frac{3}{4}}) + \frac{3}{16} \log_2(\frac{\frac{3}{16}}{\frac{11}{16}\frac{1}{4}}) + \frac{1}{2} \log_2(\frac{\frac{1}{2}}{\frac{11}{16}\frac{3}{4}})$
$I(X,Y) = 0.00449$

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]

$H(Y) = -\sum_{k=1}^{1} 0p(Y=k)\log_2 p(Y=k)$
$H(Y) = -0.6\log_2 0.6 - 0.4\log_2 0.4$
$H(Y) = 0.971$

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

$H(Y|x_2) = \sum_{x_2 \in H,M,L} p(x_2)H(Y|x_2)$
$H(Y|x_2) = p(x_2=H)H(Y|x_2=H) + p(x_2=M)H(Y|x_2=M) + p(x_2=L)H(Y|x_2=L)$
$H(Y|x_2) = \frac{3}{10}(-\frac{2}{3}\log_2(\frac{2}{3}) - \frac{1}{3}\log_2(\frac{1}{3})) + \frac{2}{10}(-\frac{2}{2}\log_2(\frac{2}{2})) + \frac{1}{2}(-\frac{2}{5}\log_2(\frac{2}{5}) - \frac{3}{5}\log_2(\frac{3}{5}))$
$H(Y|x_2) = 0.761$

$H(Y|x_4) = \sum_{x_4 \in T,F} p(x_4)H(Y|x_4)$
$H(Y|x_4) = p(x_4=T)H(Y|x_4=T) + p(x_4=F)H(Y|x_4=F)$
$H(Y|x_4) = \frac{7}{10}(-\frac{5}{7}\log_2(\frac{5}{7}) - \frac{2}{7}\log_2(\frac{2}{7})) + \frac{3}{10}(-\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3}))$
$H(Y|x_4) = 0.880$

(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]

$H(x_2, Y) = H(Y) - H(Y|x_2)$
$H(x_2, Y) = 0.971 - 0.761$
$H(x_2, Y) = 0.210$

$H(x_4, Y) = H(Y) - H(Y|x_4)$
$H(x_4, Y) = 0.971 - 0.880$
$H(x_4, Y) = 0.091$
We can say that feature x_2 is more informative because seeing x_2 results in a larger reduction in entropy (0.210) than x_4 (0.091).

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

$H(Y, x_3) = -\sum_{x_3 \in T,F; Y \in T,F} p(Y, x_3) \log_2 p(Y, x_3)$

$H(Y, x_3) = -p(Y{=}T, x_3{=}T) \log_2 p(Y{=}T, x_3{=}T) - p(Y{=}T, x_3{=}F) \log_2 p(Y{=}T, x_3{=}F) - p(Y{=}F, x_3{=}T) \log_2 p(Y{=}F, x_3{=}$
$p(Y{=}F, x_3{=}F) \log_2 p(Y{=}F, x_3{=}F)$
$H(Y, x_3) = -\frac{3}{10} \log_2 \frac{3}{10} - \frac{2}{10} \log_2 \frac{2}{10} - \frac{3}{10} \log_2 \frac{3}{10} - \frac{2}{10} \log_2 \frac{2}{10}$
$H(Y, x_3) = 1.971$

5.3 Entropy Proofs [10pts]

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

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

(b) **Using the mathematical definition of $H(X)$ and $H(X|Y)$ from part (a)**, prove that $I(X;Y) = 0$ if X and Y are 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]

$I(X;Y) = H(X) - H(X|Y)$
$I(X;Y) = -\sum_{x \in X} p(x) \log_2 p(x) + \sum_{x \in X} \sum_{y \in Y} p(x,y) log_2 p(x|y)$
$I(X;Y) = -\sum_{x \in X} p(x) \log_2 p(x) + \sum_{x \in X} \sum_{y \in Y} p(x,y) log_2 p(x)$ If $X \perp\!\!\!\perp Y$, $p(X|Y) = p(X)$
$I(X;Y) = -\sum_{x \in X} p(x) \log_2 p(x) + \sum_{x \in X} p(x) log_2 p(x)$ Marginalizing over Y
$I(X;Y) = 0$ $\square$

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]
TODO (last)

(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]
TODO (last)

(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]
TODO (last)