# CS5691: Pattern Recognition and Machine Learning
## Assignment #1

**Topics:** Regression, Classification, Density Estimation **Deadline:** 04 Oct 2021, 11:55 PM

**Teammate 1:** Sanjanaa G V **Roll number:** CS21M057
**Teammate 2:** Gudivada Harsha Vardhan **Roll number:** CS21M021

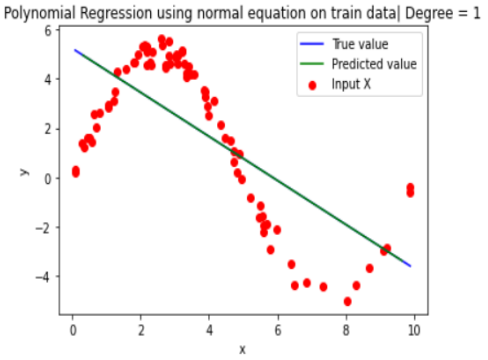
- This assignment has to be completed in teams of 2. Collaborations outside the team are strictly prohibited.
- Be precise with your explanations. Unnecessary verbosity will be penalized.
- Check the Moodle discussion forums regularly for updates regarding the assignment.
- Type your solutions in the provided LaTeX template file.
- For coding questions you will be required to upload the code in a zipped file to Moodle as well as embed the result figures in your LaTeX solutions.
- Attach a **README** with your code submission which gives a brief overview of your approach and a single command-line instruction for each question to read the data and generate the test results and figures.
- We highly recommend using `Python 3.6+` and standard libraries like `numpy, Matplotlib, pandas`. You can choose to use your favourite programming language however the TAs will only be able to assist you with doubts related to Python.
- You are supposed to write your own algorithms, any library functions which implement these directly are strictly off the table. Using them will result in a straight zero on coding questions, `import` wisely!
- **Please start early and clear all doubts ASAP.**
- Please note that the TAs will **only** clarify doubts regarding problem statements. The TAs won't discuss any prospective solution or verify your solution or give hints.
- Post your doubt only on Moodle so everyone is on the same page.

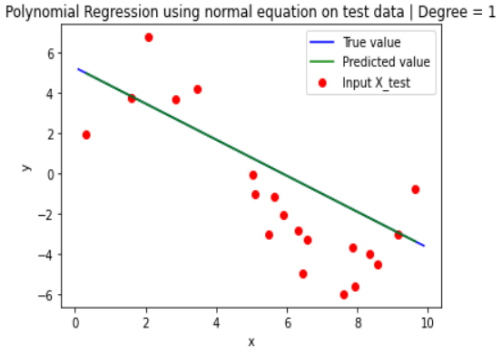
1. [**Regression**] You will implement linear regression as part of this question for the dataset provided. For each sub-question, you are expected to report the following - (i) plot of the best fit curve, (ii) equation of the best fit curve along with coefficients, (iii) value of final least squared error over the test data and (iv) scatter plot of model output vs expected output and for both train and test data. You can also generate a `.csv` file with your predictions on the test data which we should be able to reproduce when we run your command-line instruction.

   Note that you can only regress over the points in the train dataset and you are not supposed to fit a curve on the test dataset. Whatever solution you get for the train data, you have to use that to make predictions on the test data and report results.

(a) (2 marks) Use standard linear regression to get the best fit curve. Vary the maximum degree term of the polynomial to arrive upon an optimal solution.

**Solution:** 1. Standard Linear regression.
(i) Best Fit Curve:
Best fit curve is of the form $y = w_0 + w_1 x$




(ii) Equation of the Best Fit curve:
The weights obtained from the algorithm are:

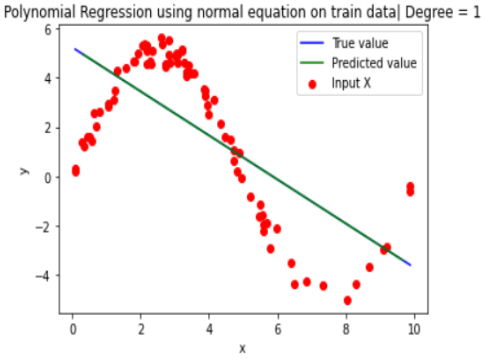
$$w_0 = 5.2405633 \; and \; w_1 = -0.89377393$$

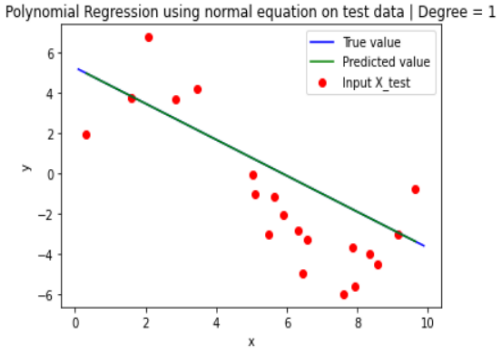
Therefore, the equation of the best fit curve along with the coefficients is given by:

$$y = 5.2405633 + (-0.89377393)x = 5.2405633 - 0.89377393x$$

(iii) Value of the final least square error over the test data: 6.52

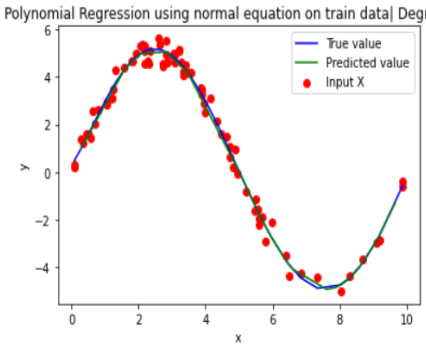
(iv) Scatter plot of model output vs expected output and for both train and test data

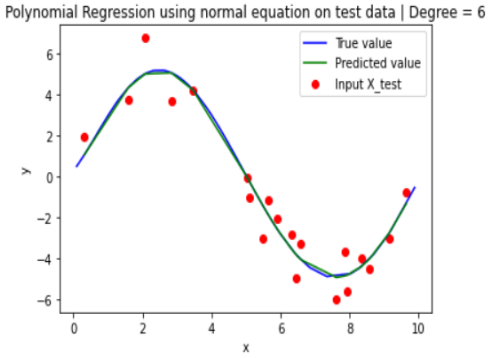



2. Optimal solution at degree 6.
Optimal Solution is obtained by varying the degrees and noting the least square errors. We have observed that at degree 6, we have obtained the least error.

(i) Best Fit Curve:
Best fit curve is of the form $y = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + w_5x^5 + w_6x^6$




(ii) Equation of the Best Fit curve:
The weights obtained from the algorithm are:

$$w_0 = 2.83080367e-01, w_1 = 2.12270911e+00, w_2 = 1.21262018e+00,$$

$$w_3 = -7.74050945e-01, w_4 = 1.23251829e-01,$$

$$w_5 = -7.70753015e-03 \; and \; w_6 = 1.69354639e-04$$

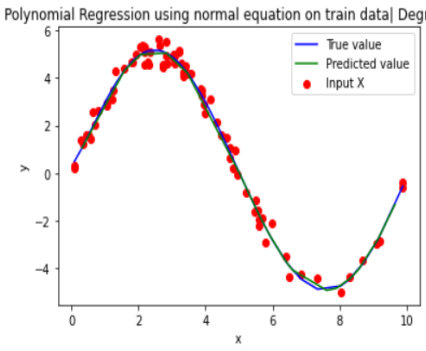
Therefore, the equation of the best fit curve along with the coefficients is given by:

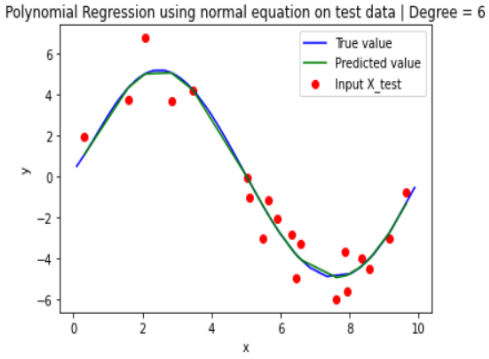
$$y = 2.83080367e-01+2.12270911e+00x+1.21262018e+00x^2-7.74050945e-01x^3+$$

$$1.23251829e-01x^4 - 7.70753015e-03x^5 + 1.69354639e-04x^6$$

(iii) Value of the final least square error over the test data: 0.834

(iv) Scatter plot of model output vs expected output and for both train and test data



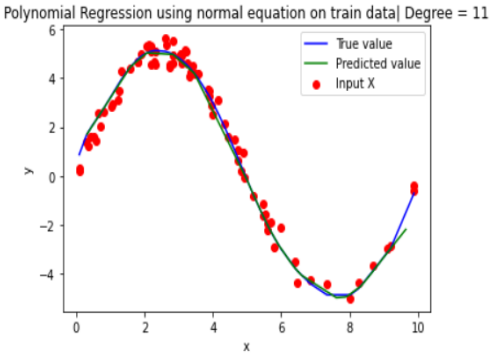



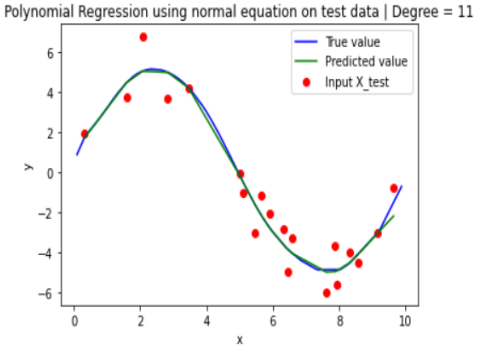

(b) (1 mark) In the above problem, increase the maximum degree of the polynomial such that

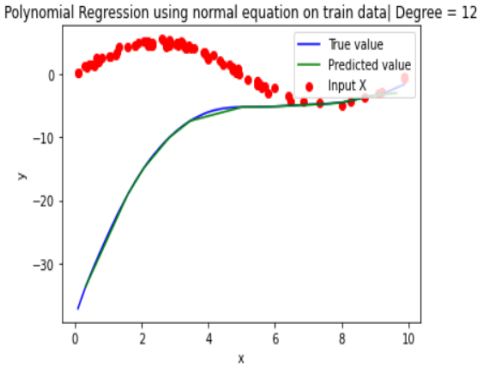
the curve overfits the data.

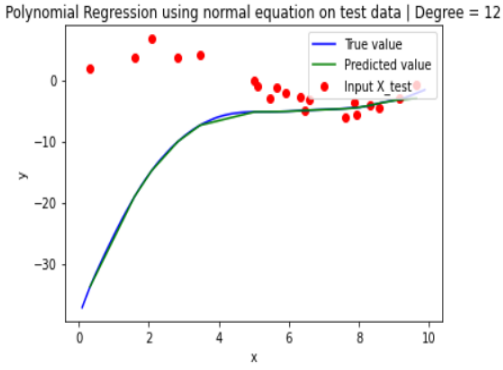
**Solution:** We have observed the overfit case by increasing the degree of the polynomial. At degree 12, the curve overfits. This is conformed in the value for least square error. In the initial degrees between 6 to 11, the least square error has minor changes. But, once we reach degree 12, our least square error showed very high values. (LSE at degree 11 = 0.85 but LSE at degree 12 = 133.64)

(i) Overfit curve at degree 12:
Overfit curve is of the form $y = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + w_5x^5 + w_6x^6 + w_7x^7 + w_8x^8 + w_9x^9 + w_{10}x^{10} + w_{11}x^{11} + w_{12}x^{12}$




(ii) Equation of the OverFit curve:
The weights obtained from the algorithm are:

$$w_0 = -3.88596135e+01, w_1 = 1.89702731e+01, w_2 = -1.37157056e+01,$$

$$w_3 = 1.56973334e+01, w_4 = -1.09922116e+01, w_5 = 4.51746934e+00,$$

$$w_6 = -1.12414032e+00, w_7 = 1.63294229e-01, w_8 = -1.09448216e-02,$$

$$w_9 = -3.21290006e-04, w_{10} = 1.14580338e-04,$$

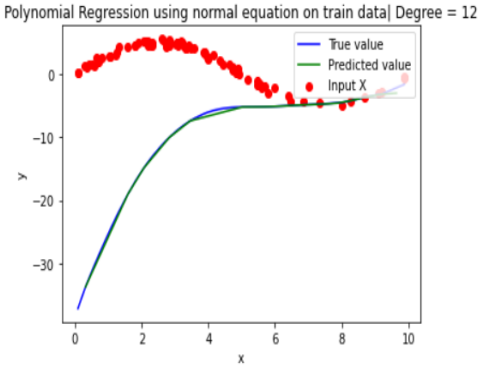
$$w_{11} = -7.81906160e-06, w_{12} = 1.86771742e-07$$

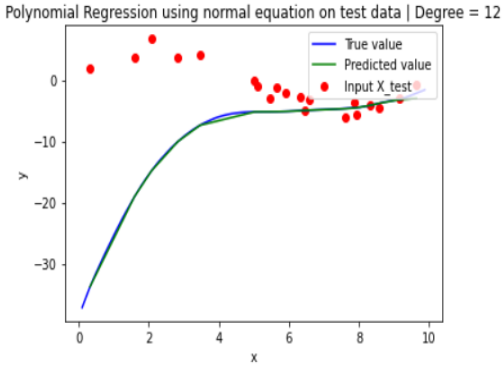
Therefore, the equation of the overfit curve along with the coefficients is given by:

$$y = -3.88596135e{+}011 + 1.89702731e{+}01x - 1.37157056e{+}01x^2 + 1.56973334e{+}01x^3$$
$$-1.09922116e{+}01x^4 + 4.51746934e{+}00x^5 - 1.12414032e{+}00x^6 + 1.63294229e{-}01x^7$$
$$-1.09448216e-02x^8 - 3.21290006e-04x^9 + 1.14580338e-04x^{10}$$
$$-7.81906160e-06x^{11} + 1.86771742e-07x^{12}$$

(iii) Value of the final least square error over the test data: 133.642

(iv) Scatter plot of model output vs expected output and for both train and test data



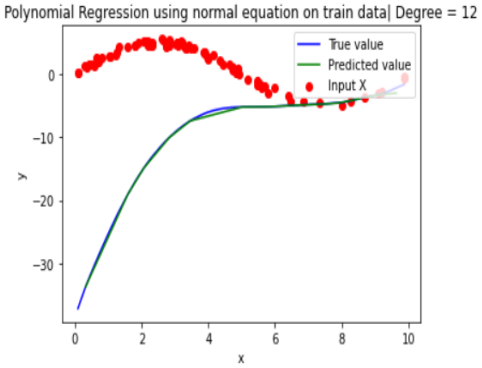

(c) (2 marks) Use ridge regression to reduce the overfit in the previous question, vary the value of lambda ($\lambda$) to arrive at the optimal value. Report the optimal $\lambda$ along with other deliverables previously mentioned.

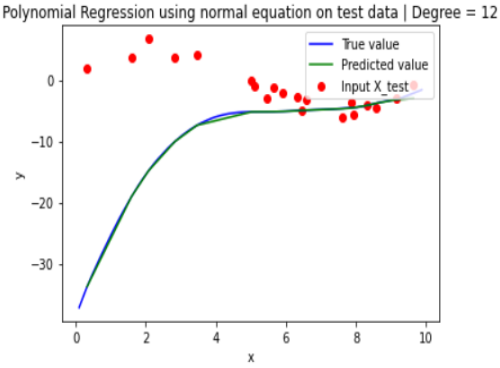
**Solution:** Using ridge regression reduces the overfit significantly. We have observed that we get the least least square error at $\lambda = 3$. Some of the least square errors for different $\lambda$s are as follows:

$$\lambda = 1, \quad LSE = 0.860$$
$$\lambda = 2, \quad LSE = 0.843$$
$$\lambda = 3, \quad LSE = 0.842$$
$$\lambda = 4, \quad LSE = 0.867$$
$$\lambda = 5, \quad LSE = 0.869$$

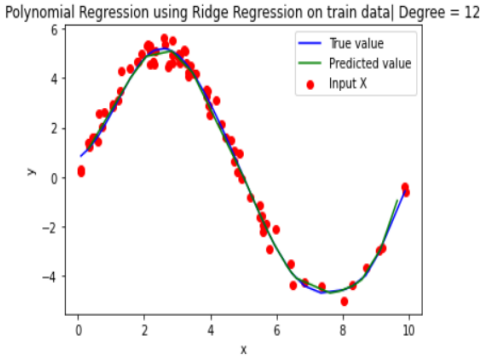
So, optimal $\lambda$ to arrive at the solution is $\lambda = 3$ .
(i) Best fit curve at degree 12:
Best fit curve is of the form $y = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 + w_5x^5 + w_6x^6 + w_7x^7 + w_8x^8 + w_9x^9 + w_{10}x^{10} + w_{11}x^{11} + w_{12}x^{12}$

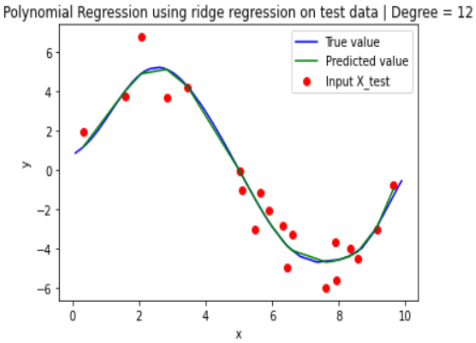
Without regression,




With regression,




(ii) Equation of the OverFit curve:
The weights obtained from the algorithm are:

$$w_0 = 8.74783269e-01, w_1 = 7.96699614e-01, w_2 = 6.59228830e-01,$$

$$w_3 = 3.41966623e-01, w_4 = -3.45704229e-021, w_5 = -1.87674751e-01,$$

$$w_6 = 6.09960267e-02, w_7 = 4.75513828e-03, w_8 = -5.88465205e-03,$$

$$w_9 = 1.31438123e-03, w_{10} = -1.41808349e-04,$$

$$w_{11} = 7.73575383e-06, w_{12} = -1.71260488e-07$$

Therefore, the equation of the overfit curve along with the coefficients is given by:

$$y = 8.74783269e-01 + 7.96699614e-01x + 6.59228830e-01x^2 + 3.41966623e-01x^3$$

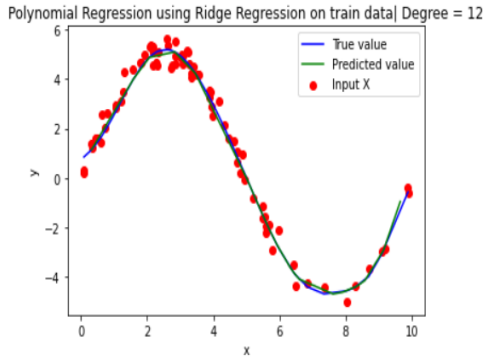
$$-3.45704229e-02x^4 - 1.87674751e-01x^5 + 6.09960267e-02x^6 + 4.75513828e-03x^7$$

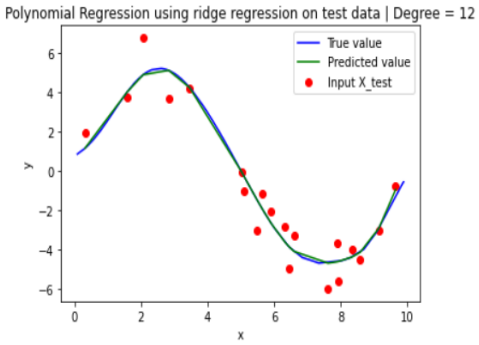
$$-5.88465205e-03x^8 + 1.31438123e-03x^9 - 1.41808349e-04x^{10}$$

$$+7.73575383e-06x^{11} + -1.71260488e-07x^{12}$$

(iii) Value of the final least square error over the test data: 0.8417

(iv) Scatter plot of model output vs expected output and for both train and test data




2. **[Classification]** You will implement classification algorithms that you have seen in class as part of this question. You will be provided train and test data as before, of which you are only supposed to use the train data to come up with a classifier which you will use to just make predictions on the test data. For each sub-question below, plot the test data along with your classification boundary and report confusion matrices on both train and test data. Again, your code should generate a `.csv` file with your predictions on the test data as before.

   (a) (2 marks) Implement the Preceptron learning algorithm with starting weights as $\mathbf{w} = [0, 1]^T$ for $\mathbf{x} = [x, y]^T$ and with a margin of 1.

   **Solution:**

   (b) (1 mark) Calculate (code it up!) a Discriminant Function for the two classes assuming Normal distribution when the covraiance matrices for both the classes are equal and $C_1 = C_2 = \sigma^2 I$ for some $\sigma$.

   **Solution:**

   (c) (1 mark) Calculate a Discriminant Function for the two classes assuming Normal distribution when both $C_1$ and $C_2$ are full matrices and $C_1 = C_2$.

   **Solution:**

   (d) (1 mark) Calculate a Discriminant Function for the two classes assuming Normal distribution when both $C_1$ and $C_2$ are full matrices and $C_1 \neq C_2$.

   **Solution:**

3. **[Probability]** In this question, you are required to verify if the following probability mass functions over their respective supports $S$ follow the following properties:

   1. $P(X = x) \geq 0 \quad \forall x \in S$, and
   2. $\sum_{x \in S} P(X = x) = 1$.

   In addition, find the expectation, $\mathbb{E}(X)$ and variance, $Var(X)$ in the following cases.

   (a) (2 marks) A discrete random variable X is said to have a Geometric distribution, with parameter $p \in (0, 1]$ over the support $S = \{1, 2, 3, \dots\}$ if it has the following probability mass function:

$$P(X = x) = (1 - p)^{x-1} p$$

**Solution:** 1. $P(X = x) \geq 0 \quad \forall x \in S$

This property is used to verify that the probability of any value inside the sample space S taken by a discrete random variable is greater than or equal to 0. Given, Probability Mass Function(PMF) of the Geometric distribution as

$$P(X = x) = (1 - p)^{x-1} p$$

To verify this property, note that $p \in (0, 1] \implies p > 0$ and $x \in \{1, 2, 3, \cdots\} \implies (x - 1) \in \{0, 1, 2, \cdots\}$. Therefore, $(1 - p) \geq 0$ and hence $(1 - p)^{x-1} \geq 0$. This mean the individual components in the given PMF is greater than 0 and hence, the product of the two is also $\geq 0$. Therefore,

$$P(X = x) = (1 - p)^{x-1} p \geq 0 \quad \forall x \in \{1, 2, 3, \cdots\} \; and \; p \in (0, 1]$$

2. $\sum_{x \in S} P(X = x) = 1$.

The 'Additivity theorem' states that - "If the sample space has an infinite number of elements and $A_1, A_2 ...$ is a sequence of disjoint events, then the probability of their union satisfies:

$$P(A_1 \cup A_2 \cup ...) = P(A_1) + P(A_2) + ...$$

And the 'Normalization theorem' states that - The probability of the entire sample space is equal to 1, that is $P(S) = 1$.

For the Geometric distribution, as $x$ ranges over all the possible values of $X$, the events $\{X = x\}$ are disjoint and divides the sample space. Hence, from the additivity and normalization theorems, we can verify this property, as $x$ ranges over all possible numerical values of $X$ and hence, the summation of all components of Geometric distribution in the sample space $S$ is 1.

Further, using the given Probability Mass Function(PMF) of the Geometric distribution, $P(X = x) = (1 - p)^{x-1} p$, this property can be written as,

$$LHS = \sum_{x \in S} P(X = x) = \sum_{x \in S} (1 - p)^{x-1} p$$

$$= \sum_{x=1}^{\infty} (1-p)^{x-1} p$$

$$= p \sum_{x=1}^{\infty} (1-p)^{x-1}$$

Let, $(1-p) = q$.

$$= p \sum_{x=1}^{\infty} q^{x-1}$$

$$= p(q^0 + q^1 + q^2 + q^3 + \cdots) = p \sum_{x=0}^{\infty} q^x$$

The above equation is an infinite geometric series. We know that the sum of an infinite geometric series is given by $\Sigma_{k=0}^{\infty} ar^k = a\frac{1}{1-r}$. Using this in the above equation,

$$\sum_{x \in S} P(X = x) = p\frac{1}{1-q} = p\frac{1}{1-(1-p)} = 1 = RHS$$

Hence, the second property is verified for a geometric distribution.

3. Expectation, $E[x]$.

We know that the Expectation, $E[x] = \sum_{x_i \in S} x_i p_i(x)$.
Given, $P(X = x) = (1-p)^{x-1} p$.

$$E[x] = \sum_{x_i \in S} x_i p_i(x) = \sum_{x_i \in S} x_i (1-p)^{x_i - 1} p$$

Let, $(1-p) = q$.

$$E[x] = p \sum_{i \in S} x_i q^{x_i - 1}$$

$$= p \sum_{x_i \in S} \frac{\partial}{\partial q} q^{x_i}$$

$$= p \frac{\partial}{\partial q} \sum_{x_i \in S} q^{x_i}$$

$$= p \frac{\partial}{\partial q} (\sum_{x_i=1}^{\infty} q^{x_i})$$

The above equation is an infinite geometric series. We know that the sum of an infinite geometric series is given by $\Sigma_{k=0}^{\infty} ar^k = a\frac{1}{1-r}$. Using this in the above equation,

$$E[x] = p \frac{\partial}{\partial q} (\frac{1}{1-q})$$

$$= p\frac{\partial}{\partial q}(\frac{1}{1-q}) = p\frac{1}{(1-q)^2}$$
$$= p\frac{1}{(1-(1-p))^2} = p\frac{1}{p^2} = \frac{1}{p}$$

Therefore, the expectation, $E(x)$ for Geometric distribution is equal to $\frac{1}{p}$ *where* $p \in (0, 1]$

4. Variance, $Var(X)$.

We know that the Variance, $Var(X) = E[X^2] - (E[X])^2$.
Given, $P(X = x) = (1-p)^{x-1}p$.

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

From the expectation of the geometric distribution derived in the above section, $E[X] = \frac{1}{p}$.

$$Var(x) = E[X(X-1)] + \frac{1}{p} - \frac{1}{p^2} \tag{1}$$

We know that, $E[x] = \sum_{x_i \in S} x_i p_i(x)$.

$$E[X(X-1)] = \sum_{x_i \in S} x_i(x_i - 1)p_i(x)$$

$$= \sum_{x_i \in S} x_i(x_i - 1)(1-p)^{x_i - 1}p$$

Let $q = (1-p)$.

$$= p\sum_{x_i \in S} x_i(x_i - 1)q^{x_i - 1}$$

$$= p\sum_{x_i \in S} (x_i - 1)\frac{\partial}{\partial q}q^{x_i}$$

$$= p\sum_{x_i \in S} \frac{\partial}{\partial q}(x_i - 1)q^{x_i}$$

$$= p\frac{\partial}{\partial q}(\sum_{x_i=1}^{\infty}(x_i - 1)q^{x_i})$$

$$= pq^2\frac{\partial}{\partial q}(\sum_{x_i=2}^{\infty}(x_i - 1)q^{x_i - 2})$$

$$= pq^2\frac{\partial}{\partial q}(\sum_{x_i=2}^{\infty}\frac{\partial}{\partial q}q^{x_i - 1})$$

$$= p\frac{\partial}{\partial q}(q^2\frac{\partial}{\partial q}(\sum_{x_i=1}^{\infty} q^{x_i}))$$

The above equation has an infinite geometric series. We know that the sum of an infinite geometric series is given by $\Sigma_{k=0}^{\infty} ar^k = a\frac{1}{1-r}$. Using this in the above equation,

$$E[X(X-1)] = p\frac{\partial}{\partial q}(q^2\frac{\partial}{\partial q}(\frac{1}{1-q}))$$

$$= p\frac{\partial}{\partial q}(q^2\frac{1}{(1-q)^2})$$

$$= p\frac{\partial}{\partial q}(\frac{q^2}{(1-q)^2})$$

$$= p[\frac{2q}{(1-q)^2} + \frac{q^2}{(1-q)^3}(-2)(-1)]$$

$$= p[\frac{2q}{(1-q)^2} + \frac{2q^2}{(1-q)^3}]$$

But $q = (1-p)$.

$$= p[\frac{2q}{(1-(1-p))^2} + \frac{2q^2}{(1-(1-p))^3}]$$

$$= p[\frac{2q}{p^2} + \frac{2q^2}{p^3}]$$

$$= p(\frac{2q}{p^2})(1 + \frac{(1-p)}{p})$$

$$= (\frac{2q}{p})(\frac{p+1-p}{p})$$

$$E[X(X-1)] = \frac{2(1-p)}{p^2}$$

Using this in (1),

$$Var(x) = E[X(X-1)] + \frac{1}{p} - \frac{1}{p^2}$$

$$= \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2}$$

$$= \frac{2-2p+p-1}{p^2}$$

$$Var(x) = \frac{1-p}{p^2}$$

Therefore, the variance, $Var(x)$ for Geometric distribution is equal to $\frac{1-p}{p^2}$ *where* $p \in (0, 1]$.

(b) (2 marks) A discrete random variable X is said to have a Poisson distribution, with parameter $\lambda > 0$ over the support $S = \{0, 1, 2, \dots\}$ if it has the following probability mass function:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

**Solution:** 1. Checking Property 1: $P(X = x) \geq 0 \quad \forall x \in S$

First axiom of probability, called the non-negativity axiom, states that "The probability of an event is a non-negative real number". Hence, from this axiom, we can verify that the probability of any value inside the sample space S taken by the random variable is greater than or equal to 0.

We can also verify this by using the given Probability Mass Function(PMF) of the Poisson distribution.

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

It is give that the parameter $\lambda > 0$. Hence, $\lambda^x > 0$. Similarly as $e > 0, e^{-\lambda} > 0$. We know that the factorial of any number is a positive number. Hence, $x! > 0$. So, all the parts in the given PMF is greater than 0. So, though the random variable may take any value $x$, its probability given by the PMF is greater than 0.

Therefore, $P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \geq 0 \quad \forall x \in S$

2. Checking Property 2: $\sum_{x \in S} P(X = x) = 1$.

One of the basic theorems of probability, called the 'additivity theorem' states that - "If the sample space has an infinite number of elements and $A_1, A_2...$ is a sequence of disjoint events, then the probability of their union satisfies:

$$P(A_1 \cup A_2 \cup ...) = P(A_1) + P(A_2) + ...$$

Another basic theorem of probability, called the 'Normalization theorem' states that - The probability of the entire sample space is equal to 1, that is $P(S) = 1$.

In the given property to be verified, as $x$ ranges over all the possible values of $X$, the events $\{X = x\}$ are disjoint and forms partition of the sample space. Hence, from the additivity and normalization theorems, we can verify that the summation above, where $x$ ranges over all possible numerical values of $X$ is 1.

We can also verify this by using the given Probability Mass Function(PMF) of the Poisson distribution.

$\sum_{x \in S} P(X = x)$

$= \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!}$

$= e^{-\lambda}(\sum_{x=0}^{\infty} \frac{\lambda^x}{x!})$

$= e^{-\lambda}(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + ...)$

We know that $(1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ...)$ is the Maclaurin series expansion for the exponential function $e^x$.

So, $= e^{-\lambda} e^{\lambda}$

$= 1$

3. Calculating the expectation - $E[X]$ of the Poisson distribution:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Expectation $= E[X] = \sum_x x p_X(x)$

$= \sum_x x P(X = x)$

$= \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!}$

When $x = 0$, the term is zero.

$= \sum_{x=1}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!}$

$= \sum_{x=1}^{\infty} \frac{\lambda^x e^{-\lambda}}{(x-1)!}$

$= \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!}$

Let $y = x - 1$. Then,

$= \lambda \sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!}$

This summation is equal to the equation in the 2nd property we proved above whose sum is equal to 1. This summation is just the summation of the PMF over the entire sample space, which is equal to 1. So,

$= \lambda.1$

$= \lambda$

4. Calculating the variance - $Var(X)$ of the Poisson distribution:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

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

$E[X]$ was already calculated above as $\lambda$.

So, $E[X^2] = \sum_{x=0}^{\infty} x^2 \frac{\lambda^x e^{-\lambda}}{x!}$

When $x = 0$, the term is zero.

$= \sum_{x=1}^{\infty} x^2 \frac{\lambda^x e^{-\lambda}}{x!}$

$= \sum_{x=1}^{\infty} x \frac{\lambda^x e^{-\lambda}}{(x-1)!}$

$= \lambda \sum_{x=1}^{\infty} x \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!}$

Let $y = x - 1$. Then,

$= \lambda \sum_{y=0}^{\infty} (y+1) \frac{\lambda^y e^{-\lambda}}{y!}$

$= \lambda(\sum_{y=0}^{\infty} y. \frac{\lambda^y e^{-\lambda}}{y!} + \sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!})$.

First summation inside the brackets is the formula for expectation of Poisson distribution. This, we showed above to be equal to $\lambda$. Second summation is just the summation of the PMF over the entire sample space, which is equal to 1. So,

$= \lambda(E[X] + 1)$

$= \lambda(\lambda + 1)$

$= \lambda^2 + \lambda$

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

$= \lambda^2 + \lambda - \lambda^2$

$= \lambda$

4. **[Linear Regression]** Recall the closed form solution for linear regression that we derived in class, the following questions are a follow-up to the same.

   (a) (2 marks) Say we have a dataset where every datapoint has a weight identified with it. Then we have the error function (sum of squares) given by

$$E(w) = \sum_{j=1}^{N} \frac{q_j(y_j - w^T x_j)^2}{2}$$

   where $q_j$ is the weight associated with each of the datapoints ($q_j > 0$). Derive the closed form solution for $w^*$.

**Solution:** Given, the error function as sum of squares,

$$E(w) = \sum_{j=1}^{N} \frac{q_j(y_j - w^T x_j)^2}{2}$$

We need a $w$ that optimises or minimises the above squared loss.

$$w^* = argmin_w \;\; E(w) = argmin_w \;\; \sum_{j=1}^{N} \frac{q_j(y_j - w^T x_j)^2}{2}$$

To find the minima of $E(w)$, we should take the first derivative of $E(w)$ and set it to zero.

$$\frac{\partial}{\partial w} E(w) = 0$$

$$\implies \frac{\partial}{\partial w}(\sum_{j=1}^{N} \frac{q_j(y_j - w^T x_j)^2}{2}) = 0$$

$$\implies \sum_{j=1}^{N} \frac{\partial}{\partial w} \frac{q_j(y_j - w^T x_j)^2}{2} = 0$$

$$\implies \sum_{j=1}^{N} \frac{2q_j(y_j - w^T x_j)}{2} \frac{\partial}{\partial w}(y_j - w^T x_j) = 0$$

We know that, $\frac{\partial}{\partial w} w\top x_j = \frac{\partial}{\partial w} x_j \cdot w = x_j$ (Partial derivative of dot product with respect to each element in $w$). We also know that $w\top x_j = x_j\top w$. Using these two properties in the above equation,

$$\implies \sum_{j=1}^{N} q_j(y_j - x_j^\top w)x_j = 0$$

$$\implies \sum_{j=1}^{N} (q_j y_j x_j - q_j x_j x_j^\top w) = 0$$

$$\implies \sum_{j=1}^{N} q_j y_j x_j - w \sum_{j=1}^{N} q_j x_j x_j^\top = 0$$

$$\implies \sum_{j=1}^{N} q_j y_j x_j = w \sum_{j=1}^{N} q_j x_j x_j^\top$$

Solving for $w$,

$$w^* = (\sum_{j=1}^{N} q_j x_j x_j^\top)^{-1} (\sum_{j=1}^{N} q_j y_j x_j)$$

$$w^* = (X^\top Q X)^{-1} (X^\top Q Y)$$

where $Q$ is a $N \times N$ diagonal matrix containing weights $w_i$ as the diagonal elements. Therefore, the closed form solution for $w*$ is given by $w^* = (X^\top Q X)^{-1}(X^\top Q Y)$.

(b) (1 mark) We saw in class that the error function in case of ridge regression is given by:

$$\frac{1}{2} \sum_{n=1}^{N} (t_n - w^T \phi(x_n))^2 + \frac{\lambda}{2} w^T w$$

Show that this error is minimized by :

$$w^* = (\lambda I + \phi^T \phi)^{-1} + \phi^T t$$

Also show that $(\lambda I + \phi^T \phi)$ is invertible for any $\lambda > 0$.

**Solution:** Error function in the case of ridge regression is given by :

$\frac{1}{2} \sum_{n=1}^{N} (t_n - w^T \phi(x_n))^2 + \frac{\lambda}{2} w^T w$

We know that transpose of a scalar results in the same scalar value. Hence the term, $w^T \phi(x_n)$, a scalar value can also be written as its transpose, which is equal to $\phi(x_n)^T w$

So, Error function $E(w, \lambda) = \frac{1}{2} \sum_{n=1}^{N} (t_n - \phi(x_n)^T w)^2 + \frac{\lambda}{2} w^T w$

This error function can be written in the matrix form as:

$E(w, \lambda) = \frac{1}{2} (t - \phi w)^T (t - \phi w) + \frac{\lambda}{2} w^T w$

$= \frac{1}{2} [t^T t - (\phi w)^T t - t^T \phi w + (\phi w)^T (\phi w)] + \frac{\lambda}{2} w^T w$

$= \frac{1}{2} [t^T t - w^T \phi^T t - t^T \phi w + w^T \phi^T \phi w] + \frac{\lambda}{2} w^T w$

We know that transpose of a scalar results in the same scalar value. Hence the term, $t^T\phi w$, a scalar value can also be written as its transpose, which is equal to $w^T\phi^T t$

$E(w,\lambda) = \frac{1}{2}[t^Tt - w^T\phi^Tt - w^T\phi^Tt + w^T\phi^T\phi w] + \frac{\lambda}{2}w^Tw$

$= \frac{1}{2}[t^Tt - 2w^T\phi^Tt + w^T\phi^T\phi w] + \frac{\lambda}{2}w^Tw$

$\frac{\lambda}{2}w^Tw$ can also be written as $\frac{1}{2}w^T\lambda Iw$ where $I$ is a $KXK$ identity matrix, and K is the number of parameters.

$E(w,\lambda) = \frac{1}{2}[t^Tt - 2w^T\phi^Tt + w^T\phi^T\phi w] + \frac{1}{2}w^T\lambda Iw$

$= \frac{1}{2}[t^Tt - 2w^T\phi^Tt] + \frac{1}{2}w^T[\phi^T\phi + \lambda I]w$

We need to find $w^*$ such that the error function is minimized.

$w^* = \arg\min_w E(w,\lambda)$

For obtaining minimum, gradient(first order derivative) with respect to $w$ must be equal to 0. First, let us calculate the gradient.

$\frac{\partial}{\partial w}E(w,\lambda)$

$= \frac{\partial}{\partial w}\frac{1}{2}[t^Tt - 2w^T\phi^Tt] + \frac{\partial}{\partial w}\frac{1}{2}w^T[\phi^T\phi + \lambda I]w$

$= \frac{1}{2}\frac{\partial}{\partial w}[t^Tt - 2w^T\phi^Tt] + \frac{1}{2}\frac{\partial}{\partial w}w^T[\phi^T\phi + \lambda I]w$

$= \frac{1}{2}\frac{\partial}{\partial w}(t^Tt) - \frac{\partial}{\partial w}(w^T\phi^Tt) + \frac{1}{2}\frac{\partial}{\partial w}w^T[\phi^T\phi + \lambda I]w$

Evaluating each term on the l;eft side:

$\frac{\partial}{\partial w}(t^Tt) = 0$

$\frac{\partial}{\partial w}(w^T\phi^Tt) = \phi^Tt$

We know $\frac{\partial}{\partial w}(x^TAx) = 2Ax$

$\frac{\partial}{\partial w}w^T[\phi^T\phi + \lambda I]w = 2(\phi^T\phi + \lambda I)w$

Substutuing all the above values

$0 - \phi^Tt + (\phi^T\phi + \lambda I)w$

$\frac{\partial}{\partial w}E(w,\lambda) = (\phi^T\phi + \lambda I)w - \phi^T t$

Equating this first order derivative to 0 gives the point of extreme.

$w^* = (\lambda I + \phi^T\phi)^{-1}\phi^T t$

So, by using the first order derivative (gradient) condition, we got a extreme point. Now, we need to check that this extreme point is indeed a point of global minimum.

A property of positive-definite matrices states: Given a function of several real variables that is twice differentiable, then if its Hessian matrix (matrix of its second partial derivatives) is positive-definite at a point p, then the function is convex near p.

So, now we need to calculate the second partial derivative (Hessian matrix) and check its value is positive-definite at $w^*$. If it is positive-definite, then $w^*$ is indeed the point of global minimum.

Calculating second order derivative of the error function:

$\frac{\partial}{\partial w}(\frac{\partial}{\partial w}(E(w,\lambda)))$

Above, we already calculated $\frac{\partial}{\partial w}(E(w,\lambda)) = (\phi^T\phi + \lambda I)w - \phi^T t$

$\frac{\partial}{\partial w}(\frac{\partial}{\partial w}(E(w,\lambda)))$

$= \frac{\partial}{\partial w}((\phi^T\phi + \lambda I)w - \phi^T t)$

$= \frac{\partial}{\partial w}((\phi^T\phi + \lambda I)w) - \frac{\partial}{\partial w}(\phi^T t))$

$= \phi^T\phi + \lambda I$

We have to prove that this value of second order derivative, the Hessian matrix is positive definite to confirm that the previously calculated point $w^*$ is indeed the point of minimum.

Definition of positive definite: A symmetric matrix $M$ with real entries is positive-definite if the real number $z^T Mz$ is positive for every nonzero real column vector $z$, where $z^T$ is the transpose of $z$

Let $M = \phi^T\phi + \lambda I$

$z^T Mz$

$= z^T(\phi^T\phi + \lambda I)z$

$= z^T\phi^T\phi z + \lambda z^T z$

$= (\phi z)^T\phi z + \lambda z^T z$

We know $A^TA \geq 0$. So, $(\phi z)^T\phi z \geq 0$

By definition, $z \neq 0$. So, $z^Tz > 0 \implies \lambda z^Tz > 0$

So, $z^TMz > 0$ which implies $M = \phi^T\phi + \lambda I$ is positive definite. So, the value of the second order derivative at $w^*$ is positive definite. Hence, $w^* = (\lambda I + \phi^T\phi)^{-1} + \phi^T t$ is the point at which the error function is minimized.

2. Showing $(\lambda I + \phi^T\phi)$ is invertible for any $\lambda > 0$ :

There exists a condition that if a matrix is positive definite, it is invertible. We already proved in the 1st part that $(\lambda I + \phi^T\phi)$ is positive-definite. Hence, it is invertible.

Proof that if a matrix is positive-definite, it is invertible:

We will prove this by contradiction. Let $M = \phi^T\phi + \lambda I$ be positive-definite. Assume that $M$ is not invertible. But, if $M$ is not invertible, $\exists x \neq 0$ satisfying $Mx = 0$. Then, we will have $x^TMx = 0$. But, as M is positive-definite, which means $z^TMz > 0$. So, the final conclsuion is false. Hence, our assumption that $M$ is not invertible is false. Thus, if a matrix is positive-definite, it is invertible.

(c) (1 mark) Given

$$X = \begin{bmatrix} -2 & 6 \\ -1 & 3 \end{bmatrix} \quad y = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

Solve $X^TXw = X^Ty$ such that the Euclidean norm of the solution $w^*$ is minimum.

**Solution:** Given,

$$X = \begin{bmatrix} -2 & 6 \\ -1 & 3 \end{bmatrix} \quad y = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

$$X^\top X = \begin{bmatrix} -2 & -1 \\ 6 & 3 \end{bmatrix} \begin{bmatrix} -2 & 6 \\ -1 & 3 \end{bmatrix} = \begin{bmatrix} 5 & -1 \\ -15 & 45 \end{bmatrix}$$

$$X^\top y = \begin{bmatrix} -2 & -1 \\ 6 & 3 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \begin{bmatrix} -5 \\ 15 \end{bmatrix}$$

Let $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ Substituting the above matrices in $X^T X w = X^T y$,

$$\begin{bmatrix} 5 & -15 \\ -15 & 45 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -5 \\ 15 \end{bmatrix}$$

The above matrix is of the form $Ax = b$. Performing Gaussian Elimination, $R_2 = R_2 + 3R_1$,

$$\begin{bmatrix} 5 & -15 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -5 \\ 0 \end{bmatrix}$$

In $A$, there is only one non-zero pivot, therefore, 2nd variable, $w_2$ is a free variable and this system of equation has infinite solutions. Let $w_2 = 0$,

$$\implies 5w_1 - 15w_2 = -5 \implies w_1 = -1$$

$\therefore w_{particular} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$ is one of the infinite solutions of the given equation.

Further, solving $Ax = 0$, gives the null space solution of the given system.

$$\begin{bmatrix} 5 & -15 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Let the pivot variable, $w_2$, be equal to 1, that is, $w_2 = 1$.

$$\implies 5w_1 - 15w_2 = 0 \implies w_1 = 3$$

$\therefore w_{nullspace} = \alpha \begin{bmatrix} 3 \\ 1 \end{bmatrix} \quad \forall \alpha \in \mathbf{R}$ is the nullspace solution of the equation.

We know that, the complete solution for a system with infinite solutions is given by,

$$w_{complete} = w_{particular} + w_{nullspace}$$

$$\implies w_{complete} = \begin{bmatrix} -1 \\ 0 \end{bmatrix} + \alpha \begin{bmatrix} 3 \\ 1 \end{bmatrix} \quad \forall \alpha \in \mathbf{R}$$

$$\implies w_{complete} = \begin{bmatrix} -1 + 3\alpha \\ \alpha \end{bmatrix} \quad \forall \alpha \in \mathbf{R}$$

To find the Euclidean norm of the solution such that $w^*$ is minimum, let us take Euclidean norm on $w_{complete}$.

$$||w_{complete}||_2 = \sqrt{w_{complete}^\top w_{complete}}$$

$$= \sqrt{\begin{bmatrix} -1 + 3\alpha & \alpha \end{bmatrix} \begin{bmatrix} -1 + 3\alpha \\ \alpha \end{bmatrix}}$$

$$= \sqrt{1 - 6\alpha + 10\alpha^2}$$

To minimise $w_{complete}$, $\frac{\partial}{\partial\alpha}||w_{complete}||_2 = 0$.

$$\frac{\partial}{\partial\alpha}||w_{complete}||_2 = 0$$

$$\frac{\partial}{\partial\alpha}\sqrt{1-6\alpha+10\alpha^2} = 0$$

From chain rule,

$$\frac{1}{2}(1-6\alpha+10\alpha^2)^{\frac{-1}{2}}(20\alpha-6) = 0$$

$$\implies \frac{10\alpha-3}{\sqrt{1-6\alpha+10\alpha^2}} = 0$$

Solving for $\alpha$,

$$\alpha = \frac{3}{10}$$

Substituting $\alpha$ in $w_{complete}$ to obtain $w^*$ which is minimum,

$$w^* = \begin{bmatrix} -1+3\frac{3}{10} \\ \frac{3}{10} \end{bmatrix} = \begin{bmatrix} -1/10 \\ 3/10 \end{bmatrix}$$

Therefore, the solution for $X^TXw = X^Ty$ such that the Euclidean norm of the solution is minimum is given by, $w^* = \begin{bmatrix} -1/10 \\ 3/10 \end{bmatrix}$.

5. (2 marks) **[Naive Bayes]** For multiclass classification problems, $p(C_k|\mathbf{x})$ can be written as:

$$p(C_k|\mathbf{x}) = \frac{exp(a_k)}{\sum_j exp(a_j)}$$

where $a_k = \ln p(\mathbf{x}|C_k)p(C_k)$. The above form is called the normalized exponential or softmax function. Now, consider a $K$ class classification problem for which the feature vector $\mathbf{x}$ has $M$ components. Each component is a categorical variable and takes one of $L$ possible values. Let these components be represented using one-hot encoding. Let us also make the naive Bayes assumption that the features are independent given the class. Show that the quantities $a_k$ are linear functions of the components of $\mathbf{x}$.

**Solution:** $a_k = \ln p(\mathbf{x}|C_k)p(C_k)$ where $\mathbf{x}$ is the feature vector

$= \ln p(x_1, x_2, ..., x_M|C_k)p(C_k)$

Using the naive bayes assumption that features $x_1, x_2, ...x_M$ are independent given the class.

$= \ln p(x_1|C_k)p(x_2|C_k)....p(x_M|C_k)p(C_k)$

$= \ln p(C_k) + \sum_{i=1}^{M} \ln p(x_i|C_k)$

To substitute a value into the class conditional distribution $\ln p(x_i|C_k)$, we need to understand the type of its distribution. The given distribution can be compared to a 'L' sided die rolled 'M' times. And each roll is independent. And each roll's result can be any of the 'L' values. Hence, multinomial distribution gives the probability of any particular combination of values for various rolls (here components).

Hence, given a class $C_k$, the feature vector $x = (x_1, x_2, ..x_n)$ is then a histogram, with $x_i$ counting the number of times event $i$ was observed. So, likelihood of observing $x_i$ in a multinomial naive Bayes model with parameters is given by: $p(x_i|C_k) = \prod_{j=1}^{L} \mu_{kij}^{x_{ij}}$

$= \ln p(C_k) + \sum_{i=1}^{M} \ln \prod_{j=1}^{L} \mu_{kij}^{x_{ij}}$

We know that log of a product is equal to the sum of individual logs, $log(AB...) = log(A)log(B)....$

$= \ln p(C_k) + \sum_{i=1}^{M} \sum_{j=1}^{L} \ln \mu_{kij}^{x_{ij}}$

$= \ln p(C_k) + \sum_{i=1}^{M} \sum_{j=1}^{L} x_{ij} \ln \mu_{kij}$

So, $a_k$ are linear functions of the components of $\mathbf{x}$

6. (2 marks) **[Naive Bayes]** Consider a Gaussian Naive Bayes classifier for a dataset with single attribute x and two classes 0 and 1. The parameters of the Gaussian distributions are:

$$p(x|y=0) \sim \mathcal{N}(0, 1/4)$$

$$p(x|y=1) \sim \mathcal{N}(0, 1/2)$$

$$P(y=1) = 0.5$$

Find the decision boundary for this classifier if the loss matrix is L = $\begin{bmatrix} 0 & \sqrt{2} \\ 1 & 0 \end{bmatrix}$

**Solution:** We know that Gaussian distribution, $\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$.
Given,

$$p(x|y=0) \sim \mathcal{N}(0, 1/4) = \frac{1}{\sqrt{2\pi * 1/4}} e^{\frac{-(x-0)^2}{2*1/4}} = \frac{1}{\sqrt{\pi/2}} e^{-2x^2}$$

$$p(x|y=1) \sim \mathcal{N}(0, 1/2) = \frac{1}{\sqrt{2\pi * 1/2}} e^{\frac{-(x-0)^2}{2*1/2}} = \frac{1}{\sqrt{\pi}} e^{-x^2}$$

Also, $p(y=1) = 0.5 \implies p(y=0) = 0.5$
From Baye's rule,

$$p(y = C_1|x) = \frac{p(x|y = C_1)p(y = C_1)}{p(x)}$$

$$p(x) = p(x|y=0)p(y=0) + p(x|y=1)p(y=1)$$

$$\implies p(x) = \frac{0.5}{\sqrt{\pi/2}}e^{-2x^2} + \frac{0.5}{\sqrt{\pi}}e^{-x^2}$$

$$\implies p(y=0|x) = \frac{p(x|y=0)p(y=0)}{p(x)} = \frac{\frac{1}{\sqrt{\pi/2}}e^{-2x^2} * 0.5}{\frac{0.5}{\sqrt{\pi/2}}e^{-2x^2} + \frac{0.5}{\sqrt{\pi}}e^{-x^2}} = \frac{\sqrt{2}e^{-2x^2}}{\sqrt{2}e^{-2x^2} + e^{-x^2}}$$

$$\implies p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x)} = \frac{\frac{1}{\sqrt{\pi}}e^{-x^2} * 0.5}{\frac{0.5}{\sqrt{\pi/2}}e^{-2x^2} + \frac{0.5}{\sqrt{\pi}}e^{-x^2}} == \frac{e^{-x^2}}{\sqrt{2}e^{-2x^2} + e^{-x^2}}$$

Given, the loss matrix $L$,

$$L = \begin{bmatrix} \lambda_{11} & \lambda_{21} \\ \lambda_{12} & \lambda_{22} \end{bmatrix} = \begin{bmatrix} 0 & \sqrt{2} \\ 1 & 0 \end{bmatrix}$$

The conditional risk or the expected loss with the taking action $\alpha_i$ corresponding to each row of our loss matrix $L$ is given by,

$$R(\alpha_i|x) = \Sigma_{j=1}^{c} \lambda(\alpha_i|y_j)p(y_j|x)$$

where $c$ is the total number of classes availabl and $\lambda(\alpha_i|y_j)$ is the loss incurred while determining $y_i$ when $y_j$ is the true value.
In our case,

$$R(\alpha_1|x) = \lambda_{11}p(y=0|x) + \lambda_{12}p(y=1|x) = 0 + \frac{e^{-x^2}}{\sqrt{2}e^{-2x^2} + e^{-x^2}}$$

$$R(\alpha_2|x) = \lambda_{21}p(y=0|x) + \lambda_{22}p(y=1|x) = \sqrt{2}\frac{\sqrt{2}e^{-2x^2}}{\sqrt{2}e^{-2x^2} + e^{-x^2}} + 0$$

The space where $R(\alpha_1|x) = R(\alpha_2|x)$ gives the decision boundary.

$$\frac{e^{-x^2}}{\sqrt{2}e^{-2x^2} + e^{-x^2}} = \frac{2e^{-2x^2}}{\sqrt{2}e^{-2x^2} + e^{-x^2}}$$

$$\implies e^{-x^2} = 2e^{-2x^2}$$

$$\implies e^{-x^2} - 2e^{-2x^2} = 0$$

$$\implies e^{-2x^2}(e^{x^2} - 2) = 0$$

$$\implies e^{x^2} = 2$$

Taking log on both sides,

$$\implies x^2 = ln(2)$$

$$\implies x = \pm\sqrt{ln(2)} = \pm 0.832554$$

Therefore, the decision boundary of the Gaussian Baye's classifier is given by the equations, $x = 0.832554$ *and* $-0.832554$.

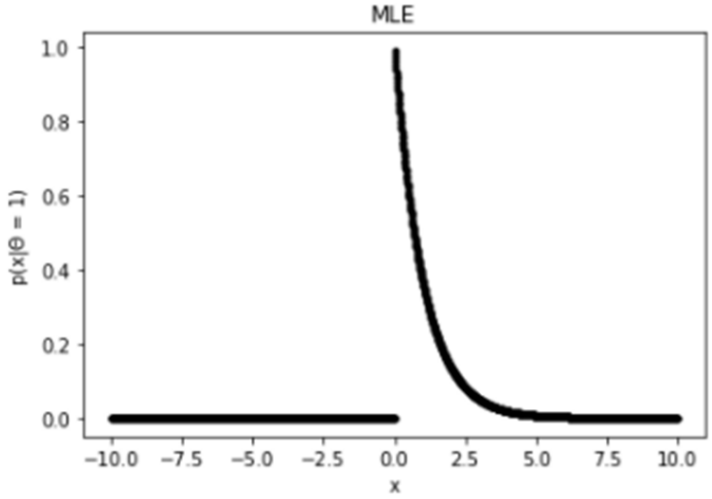
7. **[MLE]** Let $x$ have an exponential density

$$p(x|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

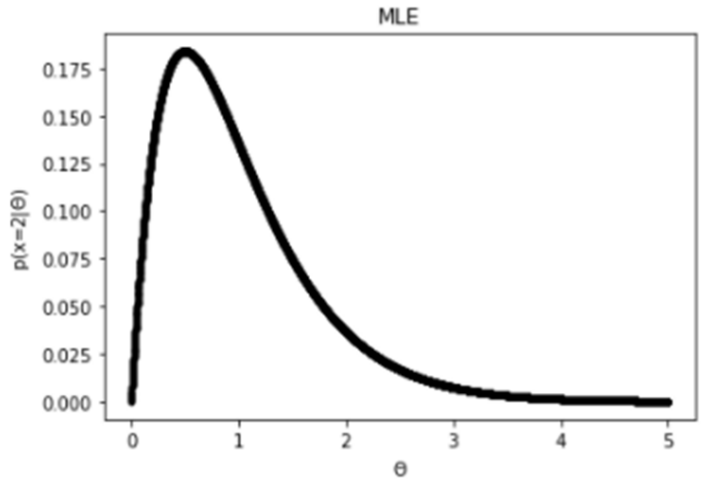
(a) (2 marks) Plot $p(x|\theta)$ versus $x$ for $\theta = 1$. Plot $p(x|\theta)$ versus $\theta$, $(0 \leq \theta \leq 5)$, for $x = 2$.

**Solution:** Using Python's matplotlib, both are plotted.

Plot of $p(x|\theta)$ versus $x$ for $\theta = 1$ :




Plot of $p(x|\theta)$ versus $\theta$, $(0 \leq \theta \leq 5)$, for $x = 2$ :




(b) (1 mark) Suppose that $n$ samples $x_1, ..., x_n$ are drawn independently according to $p(x|\theta)$. Give the maximum likelihood estimate for $\theta$.

**Solution:** $p(x|\theta) = \theta e^{-\theta x}$

$p(x|\theta) = p(x_1, x_2, ..., x_n|\theta)$

As samples $x_1, ..., x_n$ are drawn independently:

$p(\mathbf{x}|\theta) = p(x_1, x_2, ..., x_n|\theta) = p(x_1|\theta)p(x_2|\theta), \cdots p(x_n|\theta)$

$p(\mathbf{x}|\theta) = \Pi_{i=1}^{n} p(x_i|\theta)$

Taking ln on both sides,

$\ln(p(\mathbf{x}|\theta)) = \ln(\Pi_{i=1}^{n} p(x_i|\theta))$

We know that log of a product is equal to the sum of individual logs,
$\ln(AB...) = ln(A)ln(B)....$

$\ln(p(\mathbf{x}|\theta)) = \sum_{i=1}^{n} \ln(p(x_i|\theta))$

$= \sum_{i=1}^{n} \ln(\theta e^{-\theta x_i})$

$= \sum_{i=1}^{n} [\ln \theta + \ln e^{-\theta x_i}]$

$= \sum_{i=1}^{n} \ln \theta + \sum_{i=1}^{n} (-\theta x_i)$

$= n \ln \theta - \theta \sum_{i=1}^{n} x_i$

The maximum likelihood estimation, $\theta_{MLE}$, is given by

$$\theta_{MLE} = argmax_\theta \quad \ln(p(\mathbf{x}|\theta))$$

Maximizing $\ln(p(\mathbf{x}|\theta))$ with respect to $\theta$

$\frac{\partial}{\partial\theta}(\ln(p(\mathbf{x}|\theta))) = 0$

$\frac{\partial}{\partial\theta}(n \ln \theta - \theta \sum_{i=1}^{n} x_i) = 0$

$\frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0$

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

Thus, Maximum Likelihood Estimate for $\theta$ is $\frac{n}{\sum_{i=1}^{n} x_i}$

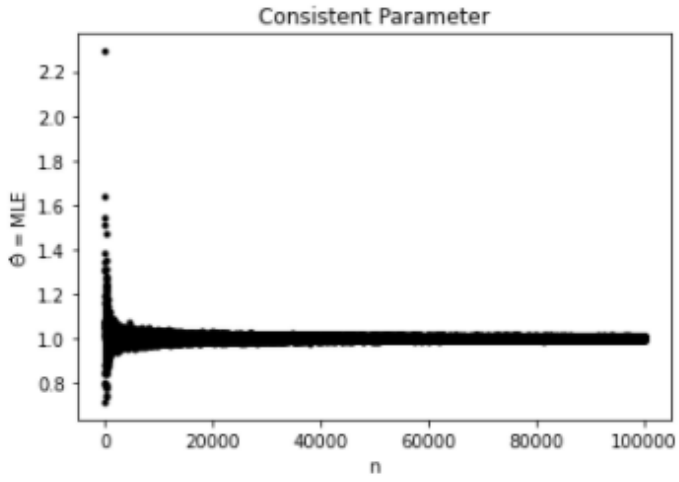
(c) (2 marks) On the graph generated with $\theta = 1$ in part (a), mark the maximum likelihood estimate $\hat{\theta}$ for large $n$. Write down your observations.

**Solution:**
Plotting the maximum likelihood estimate $\hat{\theta}$ for large $n$ :

Method followed for plotting:
1. $\theta = 1$ in part (a). Hence, the exponential density will be:
$p(x|\theta = 1) = e^{-x}$

2. We will plot the maximum likelihood estimate ($\hat{\theta}$) calculated in 7(b) i.e., $\frac{n}{\sum_{i=1}^{n} x_i}$ on the Y-axis and 'n' - sample size in each iteration on the X-axis.We check how does the $\hat{\theta}$ change as n grows large.

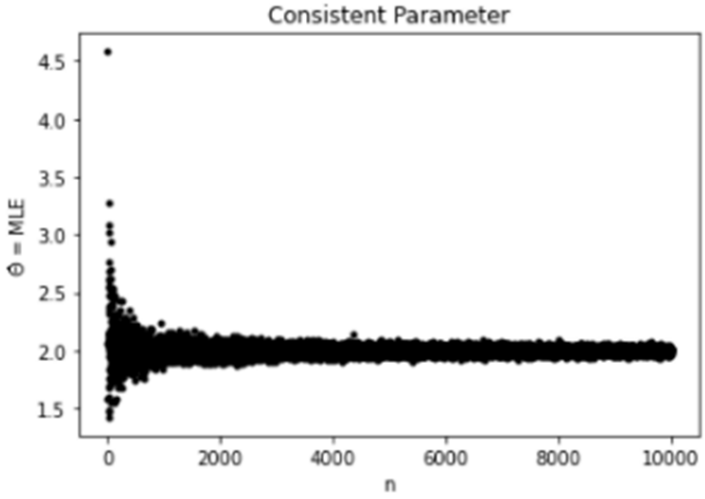
3. We use Python to generate random samples from exponential distribution. Then, we calculate the maximum likelihood estimate ($\hat{\theta}$) for each sample of size 'n' using the estimation/formula calculated in 7(b).




Observations from the plot of maximum likelihood estimate $\hat{\theta}$ for large $n$ :

1. As 'n' grows larger, the estimated values of the parameter $\theta$ which is given by $\frac{n}{\sum_{i=1}^{n} x_i}$ converges to 1 which is the actual parameter value as per 7(a).

If we set $\theta = 2$, and plot/mark maximum likelihood estimate $(\hat{\theta})$, it will converge to 2 as shown in the below 2nd plot.




ss

Hence, we observe that, as 'n' grows larger, as the number of data points used increases indefinitely, the resulting sequence of estimates converges in probability to the actual parameter. Such estimator is called a 'Consistent Estimator'.

2. Why does the estimated parameter converge probably to the true value of the

parameter as n becomes large?

Observations: 'Weak Law of Large Numbers" states that the sample average converges in probability towards the expected value. In this exponential distribution of random variable, $X_1, X_2, ..., X_n$, if a sample of N observations on variable X is taken from the population, the sample mean $\overline{X_n} \xrightarrow{P} \mu$ when $n \to \infty$

Therefore, $\frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{P} \mu$ when $n \to \infty$

For exponential distribution, mean $\mu = \frac{1}{\theta}$

$\frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{P} \frac{1}{\theta}$ when $n \to \infty$

As $f(x) = \frac{1}{x}$ is a continuous function for $x > 0$, by applying continuous mapping theorem:

$\frac{n}{\sum_{i=1}^{n} X_i} \xrightarrow{P} \theta$ when $n \to \infty$

Hence, the maximum likelihood estimate $\hat{\theta}$ converges in probability to the true value of the parameter for large $n$.

8. (3 marks) **[MLE]** Gamma distribution has a density function as follows

$$f(x|\alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}, \quad \text{with} \quad 0 \leq x \leq \infty$$

Suppose the parameter $\alpha$ is known, please find the MLE of $\lambda$ based on an i.i.d. sample $X_1, ..., X_n$.

**Solution:** Given, Gamma distribution with a density function,

$$f(x|\alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}, \quad \text{with} \quad 0 \leq x \leq \infty$$

Unknown parameter = $\lambda$. Let $\theta = \lambda$.

$$\implies f(x|\theta) = \frac{1}{\Gamma(\alpha)} \theta^{\alpha} x^{\alpha-1} e^{-\theta x}, \quad \text{with} \quad 0 \leq x \leq \infty$$

Given, data is independent and identically distributed random variable sample $\mathbf{x} = X_1, ..., X_n$.

$$\implies p(\mathbf{x}|\theta) = p(X_1, X_2, ..., X_n|\theta) = p(X_1|\theta)p(X_2|\theta), \cdots p(X_n|\theta)$$

$$\implies p(\mathbf{x}|\theta) = \Pi_{i=1}^{n} p(X_i|\theta)$$

Taking log on both sides (here, log refers to $log_e$),

$$\implies log(p(\mathbf{x}|\theta)) = log(\Pi_{i=1}^{n} p(X_i|\theta))$$

We know that log of a product is equal to the sum of individual logs, $log(AB...) = log(A)log(B).....$

$$\implies log(p(\mathbf{x}|\theta)) = \Sigma_{i=1}^{n} log(p(X_i|\theta))$$

From the Gamma distribution density function,

$$\implies log(p(\mathbf{x}|\theta)) = \Sigma_{i=1}^{n} log(\frac{1}{\Gamma(\alpha)}\theta^{\alpha}x_i^{\alpha-1}e^{-\theta x_i})$$

$$\implies log(p(\mathbf{x}|\theta)) = \Sigma_{i=1}^{n} [log(\frac{1}{\Gamma(\alpha)} + log(\theta^{\alpha}) + log(x_i^{\alpha-1}) + log(e^{-\theta x_i})]$$

The maximum likelihood estimation, $\theta_{MLE}$, is given by

$$\theta_{MLE} = argmax_{\theta} \;\; log(p(\mathbf{x}|\theta))$$

Maximizing $log(p(\mathbf{x}|\theta))$ with respect to $\theta$,

$$\frac{\partial}{\partial\theta}(log(p(\mathbf{x}|\theta))) = 0$$

$$\implies \Sigma_{i=1}^{n} (\frac{\partial}{\partial\theta}[log(\frac{1}{\Gamma(\alpha)} + log(\theta^{\alpha}) + log(x_i^{\alpha-1}) + log(e^{-\theta x_i}))]) = 0$$

$$\implies \Sigma_{i=1}^{n} (\frac{\partial}{\partial\theta} log\frac{1}{\Gamma(\alpha)} + \frac{\partial}{\partial\theta} log(\theta^{\alpha}) + \frac{\partial}{\partial\theta} log(x_i^{\alpha-1}) + \frac{\partial}{\partial\theta} log(e^{-\theta x_i})]) = 0$$

$\frac{\partial}{\partial\theta} log\frac{1}{\Gamma(\alpha)} = 0$, $\frac{\partial}{\partial\theta} log(x_i^{\alpha-1}) = 0$

$$\implies \Sigma_{i=1}^{n} (\frac{\partial}{\partial\theta}\alpha log(\theta) + \frac{\partial}{\partial\theta} - \theta x_i log(e)]) = 0$$

$$\implies \Sigma_{i=1}^{n} (\alpha\frac{\partial}{\partial\theta} log(\theta) - x_i\frac{\partial}{\partial\theta}\theta]) = 0$$

$$\implies \Sigma_{i=1}^{n} (\alpha\frac{1}{\theta} - x_i) = 0$$

$$\implies \Sigma_{i=1}^{n} \alpha\frac{1}{\theta} - \Sigma_{i=1}^{n} x_i = 0$$

$$\implies \alpha\frac{1}{\theta}\Sigma_{i=1}^{n} 1 = \Sigma_{i=1}^{n} x_i$$

$$\implies \theta = \alpha\frac{\Sigma_{i=1}^{n} 1}{\Sigma_{i=1}^{n} x_i}$$

Therefore, MLE based on the Gamma distribution and of $\lambda$ based on i.i.d sample is

$$\hat{\theta}_{MLE} = \lambda_{MLE} = \frac{\alpha n}{\Sigma_{i=1}^{n} x_i}$$