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

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

**Teammate 1:** Manish Kumar **Roll number:** CS21M033
**Teammate 2:** Pranab Kumar Rout **Roll number:** CS21M045

---

- 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.

---

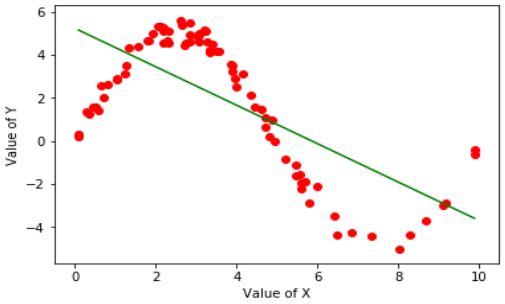
1. [**Regression**] You will implement linear regression as part of this question for the dataset provided. For each sub-question, you are expected to report the following - (i) plot of the best fit curve, (ii) equation of the best fit curve along with coefficients, (iii) value of final least squared 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.

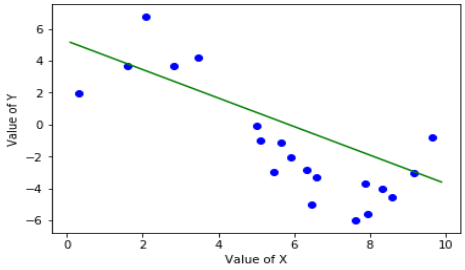
**Solution:** For Train Data Set :
Equation best fit line: y = 5.240563302374699 + ( -0.893773932197468 )x
( slope = -0.893773932197468,bias = 5.240563302374699)
Square Error over Train data: 389.95560454513736

```
Square Error over Train data:  389.95560454513736
```



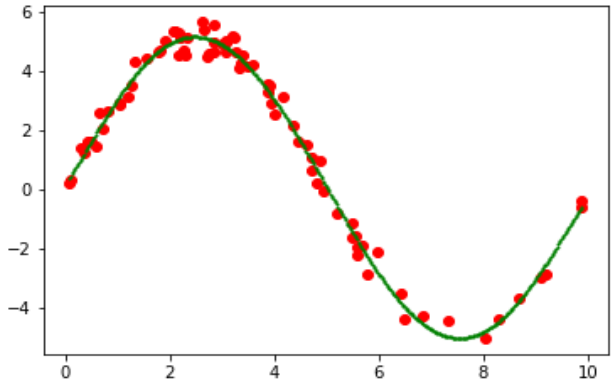

For Test Set Data:
Square Error over test data: 130.41418785131765

```
Square Error over test data:  130.41418785131765
Out[24]: Text(0,0.5,'Value of Y')
```



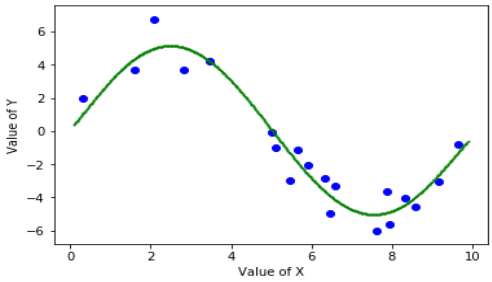

Optimal at Degree 6 for Train Data set:
Equation best fit curve:
$Y = W[0] + (W[1] * X) + (W[2] * X^2) + (W[3] * X^3) + (W[4] * X^4) + (W[5] * X^5) + (W[6] * X^6)$
Square Error over Train data: 8.558650493910909
Coefficient are $(W_0, W_2, W_2, .....W_6)$

```
[ 1.21867224e-01  2.78941392e+00  5.33713745e-01 -4.92997129e-01
  6.86379156e-02 -2.75075998e-03]
Square Error over Train data:  8.558650493910909
```



Optimal at Degree 6 for Test Data set:
Square Error over test data: 16.859264663941552

```
Square Error over test data:  16.859264663941552
Out[36]: Text(0,0.5,'Value of Y')
```

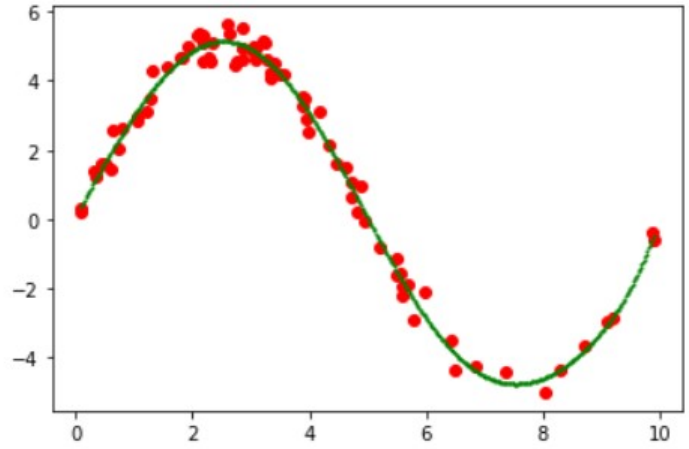



(b) (1 mark) In the above problem, increase the maximum degree of the polynomial such that the curve overfits the data.

**Solution:**

For training Set: overfits at degree 7. The coefficients are:

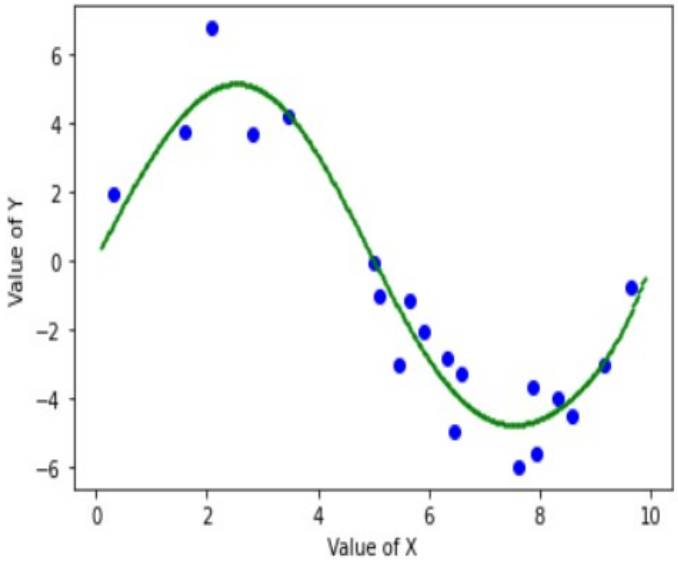
```
[ 4.02766839e-02  3.43467545e+00 -5.97694768e-01  2.66685670e-01
 -1.73672182e-01  3.65890248e-02 -3.13386583e-03  9.70465609e-05]
Square Error over Train data:  8.081459371700225
```




For testing Set:

```
Square Error over test data:  17.13207106270889

Text(0, 0.5, 'Value of Y')
```




(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:**

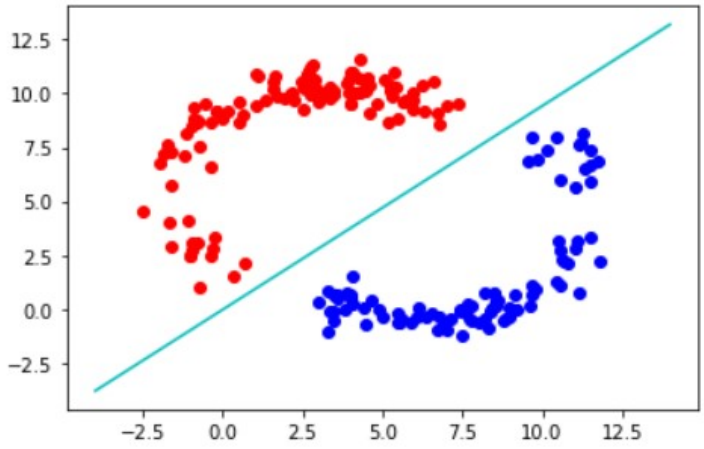
2. **[Classification]** You will implement classification algorithms that you have seen in class as part of this question. You will be provided train and test data as before, of which you are only supposed to use the train data to come up with a classifier which you will use to just make predictions 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:**

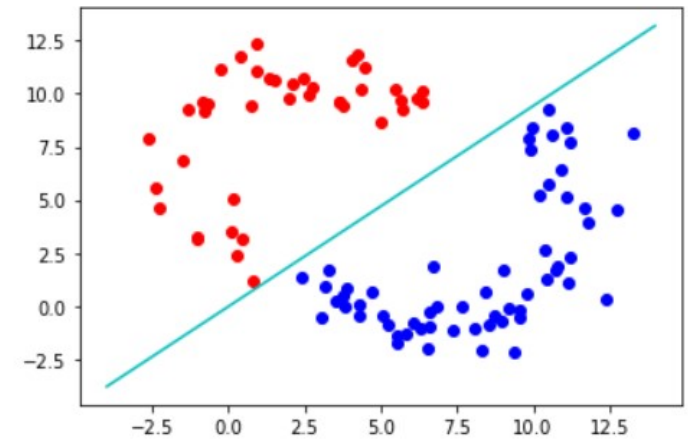
The $w = [1.0, -108.37644158723957, 114.97877958994734]$

For training Set:




```
confusion matrix for train data:
[[110.   0.]
 [  0.  90.]]
```

For Testing Set:




```
confusion matrix for test data:
[[41.  0.]
 [ 0. 59.]]
```

(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:** $P(X = x) = (1 - p)^{x-1} p$ and $\mathbb{E}(X) = \sum_{x=1}^{\infty} xp(X = x)$

$$\begin{aligned} E(X) &= \sum_{x=1}^{\infty} xp(1-p)^{x-1} \\ &= p \sum_{x=1}^{\infty} x(1-p)^{x-1} \\ &= \frac{p}{1-p} \sum_{x=1}^{\infty} x(1-p)^{x} \end{aligned}$$

We know that, $\sum_{x=1}^{\infty} x(1-p)^x = \frac{1-p}{p^2}$ [$\because$ AGP infinite summation]

$$\mathbb{E}(X) = \frac{p}{1-p}\left(\frac{1-p}{p^2}\right) = \mathbf{\frac{1}{p}}$$

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

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

$$= \frac{p}{1-p} \sum_{x=1}^{\infty} x^2 (1-p)^x = \left(\frac{p}{1-p}\right) M$$

where $M = \sum_{x=1}^{\infty} x^2 (1-p)^x$

$$pM - p(1-p)M = p \sum_{x=1}^{\infty} x^2 (1-p)^x - p(1-p) \sum_{x=1}^{\infty} x^2 (1-p)^x$$

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

$$M = \frac{1-p}{p^2} + \frac{2(1-p)^2}{p^3}$$

$$\implies E[X^2] = \left(\frac{p}{1-p}\right)\left(\frac{1-p}{p^2} + \frac{2(1-p)^2}{p^3}\right)$$

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

$$
\begin{aligned}
Var[X] &= E[X^2] - (E[X])^2 \\
&= \frac{p}{p^2} + \frac{2-2p)}{p^2} - \frac{1}{p^2} \\
&= \frac{1-p}{p^2}
\end{aligned}
$$

(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:**

$$
\begin{aligned}
\mathbb{E}(X) &= \sum_{x=0}^{\infty} e^{-\lambda} \frac{x\lambda^x}{x!} \\
&= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\
&= \lambda e^{-\lambda} \left( \frac{1}{0!} + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + ... \right) \\
&= \lambda e^{-\lambda} e^{\lambda} = \lambda
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}(X^2) &= \sum_{x=0}^{\infty} x^2 \frac{1}{x!} \lambda^x e^{-\lambda} \\
&= \lambda e^{-\lambda} \sum_{x=1}^{\infty} x \frac{\lambda^{x-1}}{(x-1)!} \\
&= \lambda e^{-\lambda} \left( \sum_{x=1}^{\infty} \frac{x-1}{(x-1)!} \lambda^{x-1} + \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \right) \\
&= \lambda e^{-\lambda} \left( \lambda \sum_{x=2}^{\infty} \frac{1}{(x-2)!} \lambda^{x-2} + e^{\lambda} \right) \\
&= \lambda e^{-\lambda} \left( \lambda e^{\lambda} + e^{\lambda} \right) \\
&= \lambda^2 + \lambda
\end{aligned}
$$

Now, $Var[X] = E[X^2] - (E[X])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$

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:** $q_j > 0$

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

As per Matrix,

$$E(w) = \frac{1}{2}(Xw - Y)^T Q(Xw - Y) = \frac{1}{2}(w^T X^T QXw - w^T X^T QY - Y^T QXw + y^T QY)$$

Derivative w.r.t ,$w$ and equate to zero

$\implies \frac{d}{dw}E(w) = X^T QXw - Y^T QX = 0$

$\implies w^* = (X^T QX)^{-1} Y^T QX$

(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:** In Ridge Regression,We Put constraints on the parameter w, instead of just minimizing the residual sum of Squares we have penalty term on w.
Such that $\sum_{j=1}^{P} W_i^2 \leq t$ (For some $t > 0$)
Here Objective Function becomes

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

$RSS(w) = (t - \phi w)^T(t - \phi w)$
Differentiate w.r.to $w$
$\frac{d}{dw}RSS(w) = \frac{d}{dw}(t^T t - \phi^T tw - w^T\phi t + w^T(\phi^T\phi w)$

$\implies 0 - 2\phi^T t + 2\phi^T \phi w + 2\lambda w = 0$
$\implies (\phi^T \phi + \lambda I) w^* = \phi^T t$
$\implies w^* = (\phi^T \phi + \lambda I)\phi^T t$

Now,
$x^T x$ is a psd $\implies$ for $(\lambda > 0)$, $x^T x + \lambda I$ is PD
$x^T x$ has all Eigenvalues non-negative as it is a PSD. If C is a Eigenvalue of $x^T x$ Then $c + \lambda$ is an eigen value of $x^T x + \lambda I$, Where $(\lambda > 0)$
$\implies x^T x + \lambda I$ is PD ,so 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 $X^T X w = X^T y$ such that the Euclidean norm of the solution $w^*$ is minimum.

**Solution:**
$X = \begin{bmatrix} -2 & 6 \\ -1 & 3 \end{bmatrix} \quad y = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$
$X^T y = \begin{bmatrix} -2 & -1 \\ 6 & 3 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \begin{bmatrix} -5 \\ 15 \end{bmatrix} \qquad X^T X = \begin{bmatrix} -2 & -1 \\ 6 & 3 \end{bmatrix} \begin{bmatrix} -2 & 6 \\ -1 & 3 \end{bmatrix} = \begin{bmatrix} 5 & -15 \\ -15 & 45 \end{bmatrix}$
Now let W= $\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$
$X^T X W = X^T y$
$\implies \begin{bmatrix} 5 & -15 \\ -15 & 45 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} -5 \\ 15 \end{bmatrix}$
$\begin{bmatrix} 5 & -15 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} -5 \\ 0 \end{bmatrix}$ (Row2=row2+3*Row1)
$5 * w_0 - 15 * w_1 = -5$
$w_0 - 3 * w_1 = -1$
$w_0 = 3 * w_1 - 1$
we need the euclidean norm to be minimum for w $\implies \sqrt{w_0^2 + w_1^2}$ is minimum possible
$\sqrt{(3 * w_1 - 1)^2 + w_1^2}$
lets take derivative w.r.to $w_1$
$\frac{d}{dw_1}(\sqrt{((3 * w_1 - 1)^2 + w_1^2)}) = 0$
$\frac{d}{dw_1}(\sqrt{10w_1^2 - 6w_1 - 1}) = 0$
$20w_1 - 6 = 0$

$$w_1 = \frac{3}{10} \qquad w_0 = \frac{-1}{10} \quad \implies w^* = \begin{bmatrix} \frac{3}{10} \\ \\ \frac{-1}{10} \end{bmatrix}$$

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:** As per the scenario given in the question:

$$p(x|C_k) = \prod_{m=1}^{M} \prod_{l=1}^{L} \mu_{kml}^{x_{ml}}$$

We are given the fact that one-hot encoding hs been done $\implies$ one out the L is 1 and rest are 0. Even the M components of the input $x$ are assumed to be independent conditioned on the class $C_k$.

$$a_k = \ln p(\mathbf{x}|C_k)p(C_k)$$

$$a_k = \sum_{m=1}^{M} \sum_{l=1}^{L} x_{ml}\ ln(\mu_{kml}) + ln(p(C_k))$$

This is evident that, $a_k$ is a linear function of the components of $x$.

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:**

$$\mathcal{N}(\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$P(y=1) = 0.5 \implies P(y=0) = 0.5$

At the decision boundary, the risk should be same $\implies$

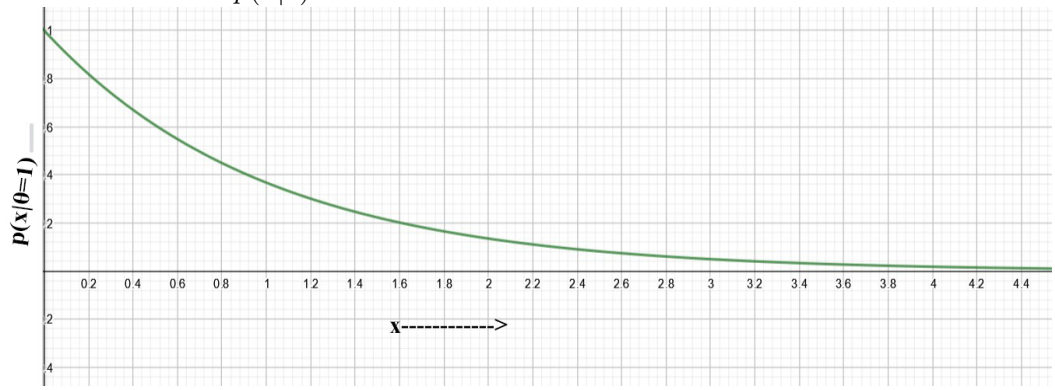
$$\begin{aligned}
P(y=0|x) &= \sqrt{2}P(y=1|x) \\
P(y=0)P(x|y=0) &= \sqrt{2}p(y=1)P(x|y=1) \\
P(y=0)P(x|y=0) &= \sqrt{2}p(y=1)P(x|y=1) \\
\frac{1}{\sqrt{2\pi(\frac{1}{4})^2}} e^{-\frac{x^2}{2(\frac{1}{4})^2}} &= \frac{\sqrt{2}}{\sqrt{2\pi(\frac{1}{2})^2}} e^{-\frac{x^2}{2(\frac{1}{2})^2}} \\
4e^{-8x^2} &= 2\sqrt{2}e^{-2x^2} \\
e^{6x^2} &= \sqrt{2} \\
6x^2 &= ln(\sqrt{2}) \\
x &= \sqrt{\frac{ln(\sqrt{2})}{6}}
\end{aligned}$$

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

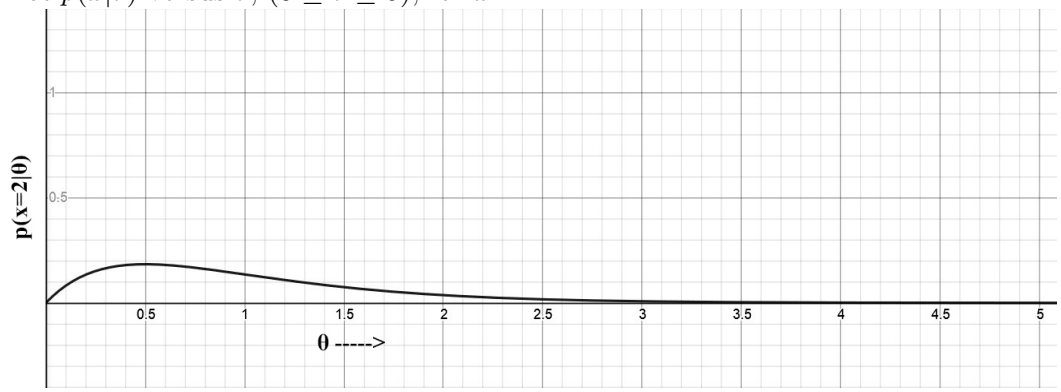
$$p(x|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

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



Plot $p(x|\theta)$ 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:**

$\mathbb{L}(\theta) = p(X|\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n \left( e^{-\theta \sum_{i=1}^{n} x_i} \right)$

Taking Log on both sides:

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

For maximum, equate the derivate to 0,

$$\frac{d}{d\theta}\left(n\ ln(\theta) - \theta \sum_{i=1}^{n} x_i\right) = 0$$
$$\implies \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0$$
$$\implies \hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i} = 1/\bar{X}$$

(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:**

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:**

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

$\mathbb{L}(\alpha) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)}\lambda^{\alpha}x_i^{\alpha-1}e^{-\lambda x_i} = \frac{\lambda^{n\alpha}}{\Gamma([\alpha)]^n}\left(\prod_{i=1}^{n} x_i^{\alpha-1}\right)e^{-\lambda\sum_{i=1}^{n} x_i}$

Taking log on both sides:

$$ln(\mathbb{L}(\alpha)) = n\alpha\, ln(\lambda) - n\, ln(\Gamma(\alpha)) + (\alpha - 1)\sum_{i=1}^{n} ln(x_i) - \lambda\sum_{i=1}^{n} x_i$$

Derive w.r.t $\alpha$ and equate with 0,

$$\frac{d}{d\alpha}\left(n\alpha\, ln(\lambda) - n\, ln(\Gamma(\alpha)) + (\alpha - 1)\sum_{i=1}^{n} ln(x_i) - \lambda\sum_{i=1}^{n} x_i\right) = 0$$

$$\implies n\, ln(\lambda) - \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} ln(x_i) = 0$$

Derive w.r.t $\lambda$ and equate with 0,

$$\frac{d}{d\lambda}\left(n\alpha\, ln(\lambda) - n\, ln(\Gamma(\alpha)) + (\alpha - 1)\sum_{i=1}^{n} ln(x_i) - \lambda\sum_{i=1}^{n} x_i\right) = 0$$

$$\implies \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\implies \hat{\lambda} = \frac{\hat{\alpha}}{\bar{X}} \text{ or } \hat{\alpha}/\bar{X} \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$