**Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019**
**Indian Institute of Technology Kanpur**
**Homework Assignment Number 1**

# QUESTION 1

*Student Name:* Piyush Kumar Gaurav
*Roll Number:* 20104442
*Date:* February 27, 2021

**My solution to problem 1**

Consider x be a scalar Random Variable drawn from a Gaussian Distribution ($\mathcal{N}(x|0,\eta)$). Here $\eta$ is again sampled from an Exponential Distribution ($exp(\eta|\frac{\gamma^2}{2})$ where $\gamma > 0$). Thus we have got

$$x \sim \mathcal{N}(x|0,\eta)$$
$$\eta \sim exp\left(\eta|\frac{\gamma^2}{2}\right) = \frac{\gamma^2}{2}\exp\left(-\frac{\gamma^2}{2}\eta\right)$$

Marginal Distribution of $x$ by integrating out $\eta$ can be denoted as

$$\begin{aligned} MarginalDistribution &= \int p(x|\eta)p(\eta|\gamma)d\eta \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\eta}}\exp\left(-\frac{x^2}{2\eta}\right)\frac{\gamma^2}{2}\exp\left(-\frac{\gamma^2}{2}\eta\right)d\eta \end{aligned} \tag{1}$$

In order to solve the above integral consider the Moment Generating Function (MGF) of Marginal Distribution. In general, MGF can be defined as follows:

$$M_x(t) = \mathbb{E}\left[e^{tx}\right] = \int_{-\infty}^{\infty} e^{tx}f(x)dx$$

where $f(x)$ is the probability distribution of x.
Moment Generating function for the Marginal likelihood will be

$$\begin{aligned} M_x(t) &= \int_{-\infty}^{\infty} e^{tx}\int_0^\infty \frac{1}{\sqrt{2\pi\eta}}\exp\left(-\frac{x^2}{2\eta}\right)\frac{\gamma^2}{2}\exp\left(-\frac{\gamma^2}{2}\eta\right)d\eta\, dx \\ &= \int_0^\infty \left[\int_{-\infty}^{\infty}\frac{e^{tx}}{\sqrt{2\pi\eta}}\exp\left(-\frac{x^2}{2\eta}\right)dx\right]\frac{\gamma^2}{2}\exp\left(-\frac{\gamma^2}{2}\eta\right)d\eta \end{aligned} \tag{2}$$

In the above equation the term in square bracket is nothing but MGF of a Gaussian Distribution with $Mean = 0$ and $Variance = \eta$. We know that,

$$MGF(\mathcal{N}(x|0,\eta)) = e^{\left(\frac{\eta t^2}{2}\right)} \tag{3}$$

Substituting Eqn. 3 in Eqn. 2 We get,

$$
\begin{aligned}
M_x(t) &= \int_0^\infty \left[ e^{\left(\frac{\eta t^2}{2}\right)} \right] \frac{\gamma^2}{2} \exp\left(-\frac{\gamma^2}{2}\eta\right) d\eta \\
&= \frac{\gamma^2}{2} \int_0^\infty \exp\left(\frac{t^2-\gamma^2}{2}\eta\right) d\eta \\
&= \frac{-1}{\left(1-\frac{t^2}{\gamma^2}\right)} \left[ \exp\left(-\eta\left(\frac{\gamma^2-t^2}{2}\right)\right)\right]_0^\infty \\
&= \frac{e^0}{\left(1-\frac{t^2}{\gamma^2}\right)}
\end{aligned} \tag{4}
$$

The above expression has a similar form as MGF of Laplace Distribution ($\mathcal{L}(\mu, b)$)

$$
MGF(\mathcal{L}(\mu, b)) = \frac{e^{t\mu}}{1-b^2t^2} \tag{5}
$$

By Comparing Eqn. 4 and Eqn. 5 we can say that Marginal is a Laplace Distribution ($\mathcal{L}(0, \frac{1}{\gamma})$) with $Mean = 0$ and $Variance = \frac{1}{\gamma}$

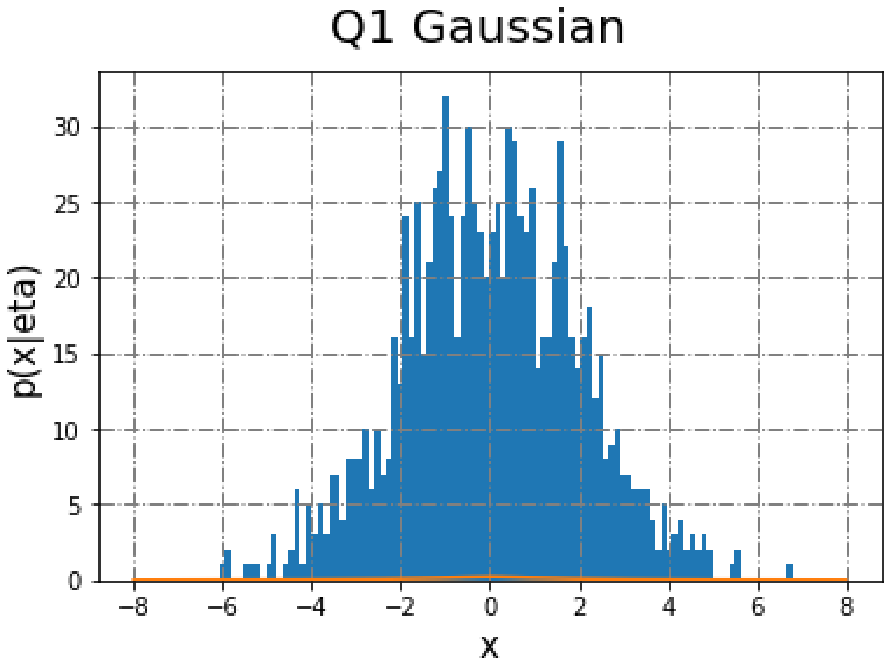

Figure 1: PDF of x w.r.t. $\eta$

In spite of the fact that both these distribution have same variance parameter, the Laplace distribution is more peaked near the Mean as compared to Gaussian Distribution (which is broader)

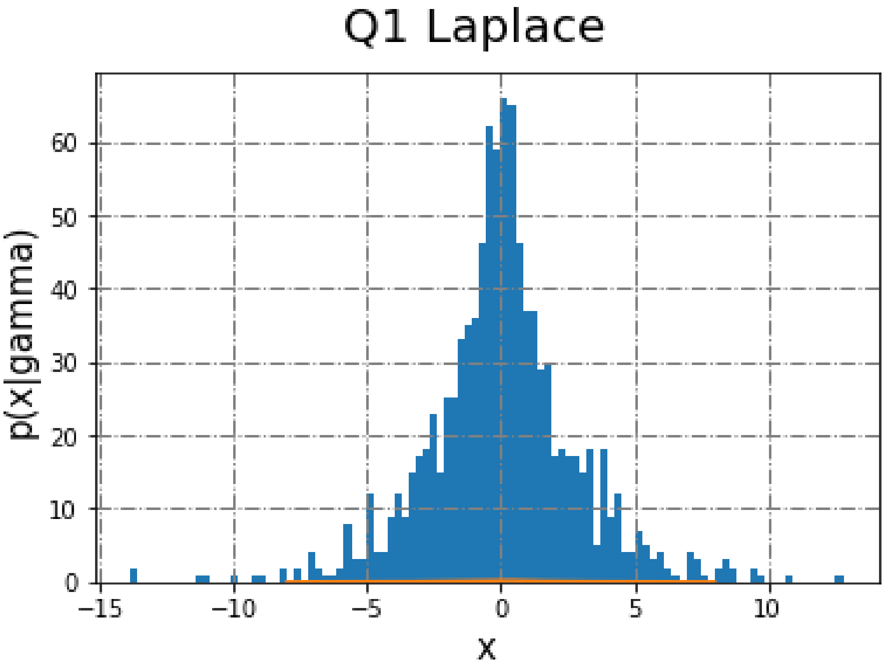

Figure 2: PDF of x w.r.t. $\gamma$

**Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019**
**Indian Institute of Technology Kanpur**
**Homework Assignment Number 1**

# QUESTION 2

*Student Name:* Piyush Kumar Gaurav
*Roll Number:* 20104442
*Date:* February 27, 2021

---

**My solution to problem 2**

For the case of Linear Regression, We compute the Likelihood function from data. The Posterior is computed using Prior and Likelihood. Further Posterior Predictive Distribution is computed utilising the Posterior Distribution and Likelihood function.
Likelihood (when we have N no. of observations):

$$p(\mathbf{y}|\mathbf{X},\mathbf{w},\beta) = \prod_{\mathbf{n=1}}^{\mathbf{N}} \mathcal{N}(\mathbf{w^T x_n}, \beta^{\mathbf{-1}}) = \mathcal{N}(\mathbf{w^T X}, \beta^{\mathbf{-1}}\mathbf{I_N})$$

Consider Prior Distribution as:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \lambda^{\mathbf{-1}}\mathbf{I_d})$$

The Posterior will thus be,

$$\mathcal{N}(\mu_N, \Sigma_N)$$

where $\Sigma_N = (\beta\sum_{n=1}^{N} x_n x_n^T + \lambda\mathbf{I_D})^{\mathbf{-1}} = (\beta\mathbf{X^T X} + \lambda\mathbf{I_d})^{\mathbf{-1}}$ and $\mu_N = \beta^{-1}(\beta X^T X + \lambda\mathbf{I_d})^{\mathbf{-1}}\mathbf{X^T y}$ also Let us define $A = {\Sigma_N}^{-1} = (\beta\sum_{n=1}^{N} x_n x_n^T + \lambda\mathbf{I_D})$ which we shall use later in the proof

The Posterior Predictive Distribution (PPD) would therefore be

$$\mathcal{N}(\mu_N^T x_*, x_*^T \Sigma_N x_*)$$

Now if we consider the Variance term for PPD that would be

$$Var(PPD) = \beta^{-1} + x_*^T(\beta\mathbf{X^T X} + \lambda\mathbf{I_d})^{\mathbf{-1}})\mathbf{x}_*$$

Consider a case when there is no data at all that mean $\mathbf{X} = \mathbf{0}$. In such a case Variance of PPD will be:

$$Var(PPD) = \beta^{-1} + x_*^T(\lambda\mathbf{I_d})^{\mathbf{-1}})\mathbf{x}_*$$

Now we add one data i.e. $x_0$ and recompute the $Var(PPD)$

$$Var(PPD) = \beta^{-1} + x_*^T(\beta x_0^T x_0 + \lambda\mathbf{I_d})^{\mathbf{-1}})\mathbf{x}_*$$

In order to compute inverse in the 2nd term of RHS, following Identity is used:

$$(M + vv^T) = M^{-1} - \frac{(M^{-1}v)(v^T M^{-1})}{1 + v^T M^{-1} v} \tag{6}$$

such that $M = \lambda\mathbf{I_d}$ and $v = x_0$ The new Variance of PPD will be:

$$\begin{aligned}
Var(PPD) &= \beta^{-1} + x_*^T\left((\lambda\mathbf{I_d})^{\mathbf{-1}} - \frac{(\lambda\mathbf{I_d})^{\mathbf{-1}}\mathbf{x_0 x_0^T}(\lambda\mathbf{I_d})^{\mathbf{-1}}}{\mathbf{1} + \mathbf{x_0^T}(\lambda\mathbf{I_d})^{\mathbf{-1}}\mathbf{x_0}}\right)x_* \\
&= \beta^{-1} + x_*^T(\lambda\mathbf{I_d})^{\mathbf{-1}})\mathbf{x}_* - \mathbf{x_*^T}\left(\frac{(\lambda\mathbf{I_d})^{\mathbf{-1}}\mathbf{x_0 x_0^T}(\lambda\mathbf{I_d})^{\mathbf{-1}}}{\mathbf{1} + \mathbf{x_0^T}(\lambda\mathbf{I_d})^{\mathbf{-1}}\mathbf{x_0}}\right)\mathbf{x}_* \\
&= \beta^{-1} + x_*^T(\lambda\mathbf{I_d})^{\mathbf{-1}})\mathbf{x}_* - \mathbf{x_*^T}\left(\frac{(\lambda\mathbf{I_d})^{\mathbf{-1}}\mathbf{x_0 x_0^T}(\lambda\mathbf{I_d})^{\mathbf{-1}}}{\mathbf{1} + \mathbf{x_0^T}(\lambda\mathbf{I_d})^{\mathbf{-1}}\mathbf{x_0}}\right)\mathbf{x}_* \\
&= \beta^{-1} + x_*^T(\lambda\mathbf{I_d})^{\mathbf{-1}})\mathbf{x}_* - \left(\frac{\left[(\mathbf{x_*^T}\lambda\mathbf{I_d})^{\mathbf{-1}}\mathbf{x_0}\right]\left[\mathbf{x_0^T}(\lambda\mathbf{I_d})^{\mathbf{-1}}\mathbf{x}_*\right]}{\mathbf{1} + \mathbf{x_0^T}(\lambda\mathbf{I_d})^{\mathbf{-1}}\mathbf{x_0}}\right)
\end{aligned} \tag{7}$$

The RHS of the equation has three term, the third term is the effect due to additional data point and the negative sign signifies that it helps the overall value to shrink.

Now, In order to generalise consider we have N data points. Also as we assumed before, Let

$$A_N = (\beta \sum_{n=1}^{N} x_n x_n^T + \lambda \mathbf{I_D}) = (\beta \mathbf{X^T X} + \lambda \mathbf{I_d})$$

(Note that since $\beta$ and $\lambda$ are positive "$A$" matrix will be positive definite.)
Similarly,

$$\begin{aligned} A_{N+1} &= (\beta \sum_{n=1}^{N+1} x_n x_n^T + \lambda \mathbf{I_D}) \\ &= (\beta \sum_{n=1}^{N} x_n x_n^T n + \lambda \mathbf{I_D}) \\ &= (\beta \sum_{n=1}^{N} x_n x_n^T + \lambda \mathbf{I_D}) + \beta \mathbf{x_{n+1} x_{n+1}}^{\mathbf{T}} \\ &= A_N + \beta x_{n+1} {x_{n+1}}^T \end{aligned} \tag{8}$$

Using Eqn. 6 we can write,

$$\begin{aligned} {A_{N+1}}^{-1} &= (A_N + \beta x_{n+1} {x_{n+1}}^T)^{-1} \\ &= (A_N)^{-1} - \frac{(A_N)^{-1} x^{n+1} {x^{n+1}}^T (A_N)^{-1}}{\beta + {x^{n+1}}^T (A_N)^{-1} x^{n+1}} \end{aligned} \tag{9}$$

Now, Since Variance of PPD after having N+1 data can be written as $Var_{N+1}(PPD) = \beta^{-1} + x_*^T (A_N + 1)^{-1} x_*$

$$\begin{aligned} Var_{N+1}(PPD) &= \beta^{-1} + x_*^T \left( (A_N)^{-1} - \frac{(A_N)^{-1} x_{n+1} {x_{n+1}}^T (A_N)^{-1}}{\beta + {x_{n+1}}^T (A_N)^{-1} x_{n+1}} \right) x_* \\ &= \beta^{-1} + x_*^T (A_N)^{-1} x_* - x_*^T \left( \frac{(A_N)^{-1} x_{n+1} {x_{n+1}}^T (A_N)^{-1}}{\beta + {x_{n+1}}^T (A_N)^{-1} x_{n+1}} \right) x_* \end{aligned} \tag{10}$$

In the RHS of above expression $\beta$, $A_N$ are positive (positive definite) which implies that the first two terms are positive while third term is always negative. Also note that the first two term are the effect of prior and N data points while the third term came into existence due to the $(N+1)^{th}$ point. This proves that as the training set size $N$ increases, the variance of the predictive posterior shrinks in general.

**[Reference]** *Qazaz, C. S., C. K. I. Williams, and C. M. Bishop (1997). An upper bound on the Bayesian error bars for generalized linear regression. In S. W. Ellacott, J. C. Mason, and I. J. Anderson (Eds.), Mathematics of Neural Networks: Models, Algorithms and Applications, pp. 295–299. Kluwer*

**Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019**
**Indian Institute of Technology Kanpur**
**Homework Assignment Number 1**

*Student Name:* Piyush Kumar Gaurav
*Roll Number:* 20104442
*Date:* February 27, 2021

# QUESTION 3

**My solution to problem 3**

Consider N scalar-valued observations $x_1$ , . . . , $x_N$ drawn i.i.d. from $\mathcal{N}(\mu, \sigma^2)$. Their empirical mean will be $\bar{x} = \sum_{n=1}^{N} x_N$.

The empirical mean can also be represented as

$$\bar{x} = (1/N).x_1 + (1/N).x_2 + \ldots + (1/N).x_N$$

The above equation can be written as linear equation:

$$\bar{x} = \mathbf{W}^T\mathbf{X}$$

where $\mathbf{W} = [1/N\, 1/N\, \ldots\, 1/N]^T$ and $\mathbf{X} = [x_1 x_2 \ldots x_N]^T$ are $NX1$ vectors

Since each component of $\mathbf{X}$ are i.i.d. and belongs to a Gaussian Distribution with mean $\mu$ and variance $\sigma^2$, the vector $\mathbf{X}$ also follows a Multivariate Gaussian Distribution with following parameters:

$$Mean = \mathbb{E}\left[\mathbf{X}\right] = [\mu\, \mu\, \ldots\, \mu]^T$$
$$Covariance(\mathbf{X}) = \sigma^2\mathbf{I}$$

where $\mathbf{I}$ is a $NXN$ Identity matrix.

The Mean of $\bar{x}$ will be

$$\begin{aligned}\mathbb{E}\left[\bar{x}\right] &= \mathbb{E}\left[\mathbf{W}^T\mathbf{X}\right] = \mathbf{W}^T\mathbb{E}\left[\mathbf{X}\right] \\ &= \sum_{n=1}^{N}(1/N)\mu = \mu\end{aligned} \tag{11}$$

The Variance of $\bar{x}$ will be

$$\begin{aligned}Var(\bar{x}) &= Var(\mathbf{W}^T\mathbf{X}) \\ &= \mathbf{W}^T\, Cov(\mathbf{X})\, \mathbf{W} \\ &= \sigma^2\, (\mathbf{W}^T\, \mathbf{I}\, \mathbf{W}) \\ &= \sum_{n=1}^{N}(\sigma^2/N^2) \\ &= (\sigma^2/N)\end{aligned} \tag{12}$$

Hence, the probability distribution of $\bar{x}$ is $\mathcal{N}(\mu, (\sigma^2/N))$

Why this make intuitive sense !

Since the mean of x is $\mu$, it is intuitive that the even the average of "empirical mean" of x will lie at $\mu$ as there is no scope of extra bias. Also if we have less number of x, its empirical mean may not coincide with the value of actual $\mu$ which implies that the uncertainty is high. But as we take more and more x to find its average, the empirical mean will tend to coincide the actual $\mu$ more and more precisely. This is explained by the fact that Variance of "empirical mean" is $\sigma^2/N$

**Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019**
**Indian Institute of Technology Kanpur**
**Homework Assignment Number 1**

*Student Name:* Piyush Kumar Gaurav
*Roll Number:* 20104442
*Date:* February 27, 2021

# QUESTION 4

### My solution to problem 4.1

As mentioned in the question, Let us assume that $\mu_0$ and $\sigma_0^2$ are the hyper parameter of the prior distribution. The posterior distribution of $\mu_m$ can be written as:

$$p(\mu_m|\mathbf{x}^{(m)}, \sigma^2) = \frac{p(\mathbf{x}^{(m)}|\mu_m, \sigma^2)p(\mu_m)}{\int p(\mathbf{x}^{(m)}|\mu_m, \sigma^2)p(\mu_m)d\mu_m}$$

Derivation of Posterior Distribution:

$$\begin{aligned} p(\mu_m|\mathbf{x}^{(m)}, \sigma^2) &\propto p(\mathbf{x}^{(\mathbf{m})}|\boldsymbol{\mu}_{\mathbf{m}}, \boldsymbol{\sigma}^{\mathbf{2}})\mathbf{p}(\boldsymbol{\mu}_{\mathbf{m}}) \\ &= \left[\prod_{n=1}^{N_m} \mathcal{N}(\mathbf{x}_n^{(m)}|\mu_m, \sigma^2)\right] \mathcal{N}(\mu_m|\mu_o, \sigma_o^2) \\ &= \left[\prod_{n=1}^{N_m} \exp\left(\frac{-(\mathbf{x}_n^{(m)} - \mu_m)^2}{2\sigma^2}\right)\right] \exp\left(\frac{-(\mu_m - \mu_o)^2}{2\sigma_o^2}\right) \\ &= \exp\left(\frac{-\sum_{n=1}^{N_m}(\mathbf{x}_n^{(m)} - \mu_m)^2}{2\sigma^2}\right) \exp\left(\frac{-(\mu_m - \mu_o)^2}{2\sigma_o^2}\right) \end{aligned} \tag{13}$$

As derived in the lecture (by completing the squares), the posterior distribution of $\mu_m$ can now be written as:

$$p(\mu_m|\mathbf{x}^{(m)}, \sigma^2) = \mathcal{N}(\mu_m|\mu_d, \sigma_d^2)$$

where

$$\mu_d = \left(\frac{\sigma^2}{\sigma^2 + N_m\sigma_o^2}\right)\mu_o + \left(\frac{N_m\sigma_o^2}{\sigma^2 + N_m\sigma_o^2}\right)\bar{x}^{(m)}$$

$$\frac{1}{\sigma_d^2} = \frac{N_m}{\sigma^2} + \frac{1}{\sigma_o^2}$$

$$\bar{x}^{(m)} = \frac{1}{N_m}\sum_{n=1}^{N_m} x_n^{(m)}$$

**My solution to problem 4.2**

The marks of each student in $m$th School can be denoted as:

$$x^{(m)} = \mu_m + \epsilon$$

where $\mu_m \sim \mathcal{N}(\mu_0, \sigma_0^2)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Since the above equation of $x^{(m)}$ is a linear sum of two Gaussian distribution, the probability distribution for $x^{(m)}$ will also be Gaussian. In order to determine the Gaussian probability distribution we need to compute the two parameters i.e. Mean and Variance of $x^{(m)}$

$$\begin{aligned} \mathbb{E}\left[x^{(m)}\right] &= \mathbb{E}\left[\mu_m\right] + \mathbb{E}\left[\epsilon\right] \\ &= \mu_0 + 0 \\ &= \mu_0 \\ \text{Var}\left[x^{(m)}\right] &= \text{Var}\left[\mu_m\right] + \text{Var}\left[\epsilon\right] \\ &= \sigma_0^2 + \sigma^2 \end{aligned} \tag{14}$$

Hence we deduce from above result that $x^{(m)} \sim \mathcal{N}(\mu_0, \sigma_0^2 + \sigma^2)$. Since this expression of probability distribution is free from $\mu_m$, it is Marginal likelihood distribution of $x^{(m)}$ w.r.t. $\mu_m$. Consequently we can write,

$$p(x^{(m)}|\mu_0, \sigma^2, \sigma_0^2) = \mathcal{N}(\mu_0, \sigma_0^2 + \sigma^2)$$

Marginal likelihood distribution of all the students will be,

$$\prod_{m=1}^{M}\prod_{n=1}^{N_m} p(x_n^{(m)}|\mu_0, \sigma^2, \sigma_0^2) = \prod_{m=1}^{M}\prod_{n=1}^{N_m} \mathcal{N}(x_n^{(m)}|\mu_0, \sigma_0^2 + \sigma^2) \tag{15}$$

Taking log both side,

$$log\left(\prod_{m=1}^{M}\prod_{n=1}^{N_m} p(x_n^{(m)}|\mu_0, \sigma^2, \sigma_0^2)\right) = log\left(\prod_{m=1}^{M}\prod_{n=1}^{N_m} \mathcal{N}\left(x_n^{(m)}|\mu_0, \sigma_0^2 + \sigma^2\right)\right) \tag{16}$$

In order to get MLE - II estimate of $\mu_0$ we need to differentiate above log - marginal likelihood equation w.r.t. $\mu_0$. Now as per the property of MLE - II (Hyper parameters has to be based on the data / observations) and the equation above it is intuitive and straightforward to write the following results:

$$\mu_0 = \frac{1}{M}\sum_{m=1}^{M} \bar{x}^{(m)} = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{N_m}\sum_{n=1}^{N_m} x_n^{(m)}\right)$$

So the MLE - II estimate $\mu_0$ is average of the ”empirical mean of every school”

**My solution to problem 4.3**

If we substitute MLE-II estimate of $\mu_0$ in the result of 4.1 we get the following parameter

$$\mu_d = \left(\frac{\sigma^2}{\sigma^2 + N_m\sigma_o^2}\right)\left(\frac{1}{M}\sum_{m=1}^{M}\bar{x}^{(m)}\right) + \left(\frac{N_m\sigma_o^2}{\sigma^2 + N_m\sigma_o^2}\right)\bar{x}^{(m)}$$

This makes intuitive sense that if we utilize the observations (i.e. the test scores of students) to estimate prior for the mean, it has to be the empirical average of the empirical mean score of every school The benefit in deriving estimate of $\mu_0$ through MLE-II is that: The prior mean will now be more realistic (rather than being vague) as it is derived from the practical data and would lie near to the empirical value .

**Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019**
**Indian Institute of Technology Kanpur**
**Homework Assignment Number 1**

*Student Name:* Piyush Kumar Gaurav
*Roll Number:* 20104442
*Date:* February 27, 2021

# QUESTION 5

### My solution to problem 5

The Likelihood for the $m^{th}$ school is denoted by:

$$p(\mathbf{y^{(m)}}|\mathbf{X^{(m)}}\mathbf{w_{(m)}}, \beta^{-1}\mathbf{I_N}) \sim \mathcal{N}(\mathbf{y^{(m)}}|\mathbf{X^{(m)}}\mathbf{w_{(m)}}, \beta^{-1}\mathbf{I_N})$$

Hence, the total Likelihood will be

$$\prod_{m=1}^{M} p(\mathbf{y^{(m)}}|\mathbf{X^{(m)}}\mathbf{w_{(m)}}, \beta^{-1}\mathbf{I_N}) = \prod_{\mathbf{m=1}}^{\mathbf{M}} \mathcal{N}(\mathbf{y^{(m)}}|\mathbf{X^{(m)}}\mathbf{w_{(m)}}, \beta^{-1}\mathbf{I_N})$$

The prior would be

$$p(\mathbf{w_{(m)}}|\mathbf{w_{(0)}}, \lambda^{-1}\mathbf{I_D}) \sim \mathcal{N}(\mathbf{w_{(m)}}|\mathbf{w_{(0)}}, \lambda^{-1}\mathbf{I_D})$$

Since Computation of Marginal Likelihood is similar to that of PPD (both integrates out $\mathbf{w_m}$), except that Marginal is computed using Prior rather than Posterior. Hence using the PPD formula (and replacing Posterior's Mean and Variance by Prior's Mean and Variance) we get, the Marginal Likelihood for $m^{th}$ school:

$$\begin{aligned} MarginalLikelihood^{(m)} &= \int p(\mathbf{y^{(m)}}|\mathbf{X^{(m)}}\mathbf{w_{(m)}}, \beta^{-1}\mathbf{I_N})\mathbf{p}(\mathbf{w_{(m)}}|\mathbf{w_{(0)}}, \lambda^{-1}\mathbf{I_D})\mathbf{dw_{(m)}} \\ &= \mathcal{N}(\mathbf{y^{(m)}}|\mathbf{X^{(m)}}\mathbf{w_0}, \mathbf{X^{(m)}}(\lambda^{-1}\mathbf{I_D})\mathbf{X^{(m)}}^{\mathbf{T}} + \beta^{-1}\mathbf{I_N}) \end{aligned} \tag{17}$$

Total Marginal Likelihood will thus be:

$$\begin{aligned} TotalMarginalLikelihood &= \prod_{m=1}^{M} \int p(\mathbf{y}^{(m)}|\mathbf{X^{(m)}}\mathbf{w_{(m)}}, \beta^{-1}\mathbf{I_N})\, \mathbf{p}(\mathbf{w_{(m)}}|\mathbf{w_{(0)}}, \lambda^{-1}\mathbf{I_D})\mathbf{dw_{(m)}} \\ &= \prod_{m=1}^{M} \mathcal{N}(\mathbf{y^{(m)}}|\mathbf{X^{(m)}}\mathbf{w_0}, \mathbf{X^{(m)}}(\lambda^{-1}\mathbf{I_D})\mathbf{X^{(m)}}^{\mathbf{T}} + \beta^{-1}\mathbf{I_N}) \end{aligned} \tag{18}$$

Taking log of Marginal likelihood

$$\begin{aligned} log(TML) &= log\left(\prod_{m=1}^{M} \mathcal{N}(\mathbf{y^{(m)}}|\mathbf{X^{(m)}}\mathbf{w_0}, \mathbf{X^{(m)}}(\lambda^{-1}\mathbf{I_D})\mathbf{X^{(m)}}^{\mathbf{T}} + \beta^{-1}\mathbf{I_N})\right) \\ &= \sum_{m=1}^{M}\left(log(\mathcal{N}(\mathbf{y^{(m)}}|\mathbf{X^{(m)}}\mathbf{w_0}, \mathbf{X^{(m)}}(\lambda^{-1}\mathbf{I_D})\mathbf{X^{(m)}}^{\mathbf{T}} + \beta^{-1}\mathbf{I_N}))\right) \\ &\propto \sum_{m=1}^{M}\left(\mathbf{y^{(m)}} - \mathbf{X^{(m)}}\mathbf{w_0}\right)^{T}\left(\mathbf{X^{(m)}}(\lambda^{-1}\mathbf{I_D})\mathbf{X^{(m)}}^{\mathbf{T}} + \beta^{-1}\mathbf{I_N}\right)^{-1}\left(\mathbf{y^{(m)}} - \mathbf{X^{(m)}}\mathbf{w_0}\right) \end{aligned} \tag{19}$$

The above expression needs to be optimized w.r.t. $\mathbf{w_0}$ in order to find the MLE-II value of $\mathbf{w_0}$.

**Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019**
**Indian Institute of Technology Kanpur**
**Homework Assignment Number 1**

*Student Name:* Piyush Kumar Gaurav
*Roll Number:* 20104442
*Date:* February 27, 2021

# QUESTION 6

**My solution to problem 6.1**

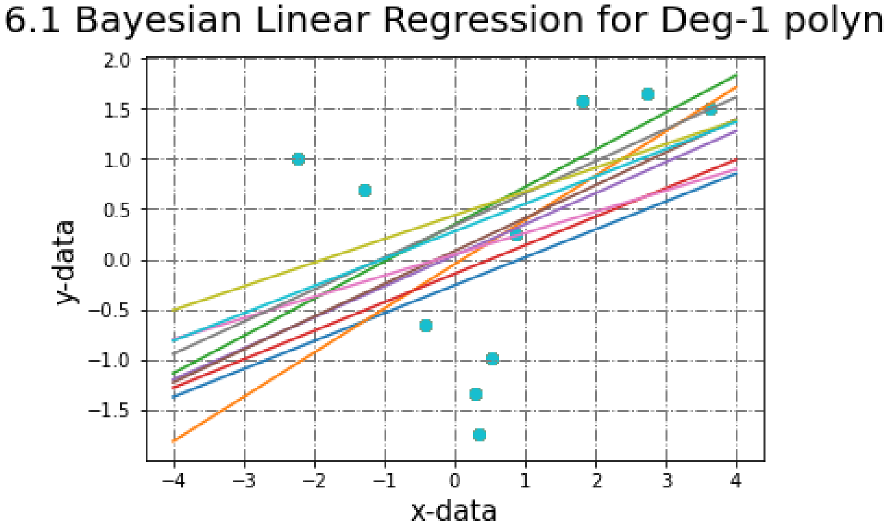

Figure 3: Plot showing 10 random functions drawn from the inferred posterior for degree 1 polynomial

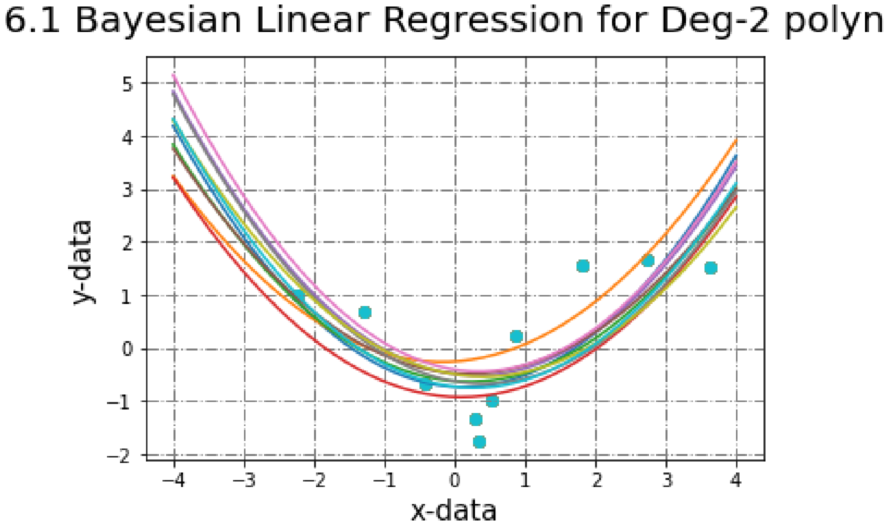

Figure 4: Plot showing 10 random functions drawn from the inferred posterior for degree 2 polynomial

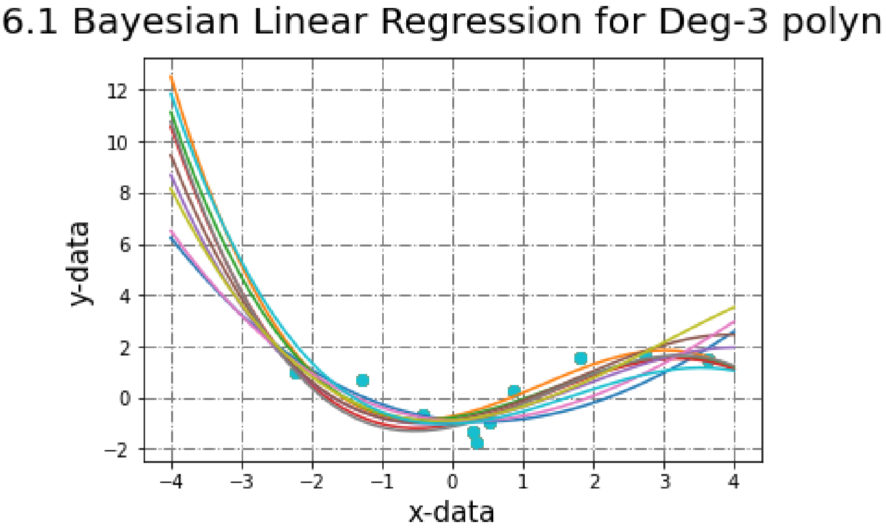

Figure 5: Plot showing 10 random functions drawn from the inferred posterior for degree 3 polynomial

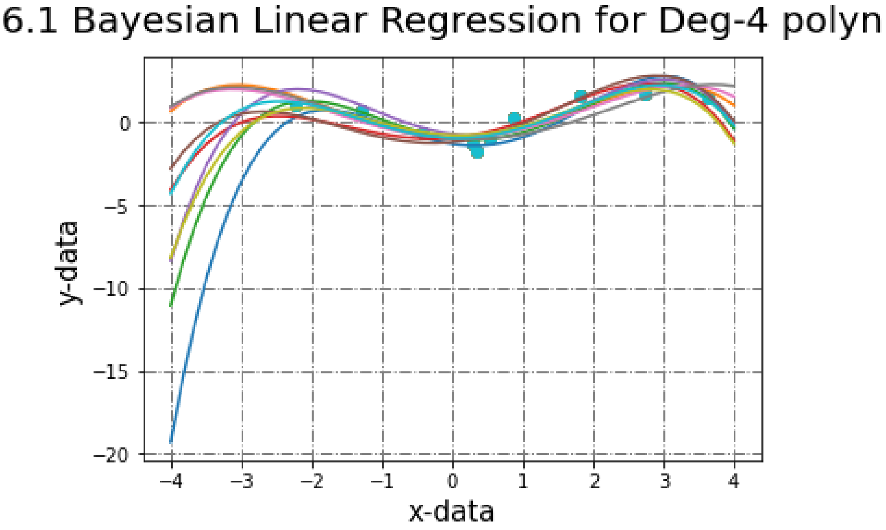

Figure 6: Plot showing 10 random functions drawn from the inferred posterior for degree 4 polynomial

**My solution to problem 6.2**

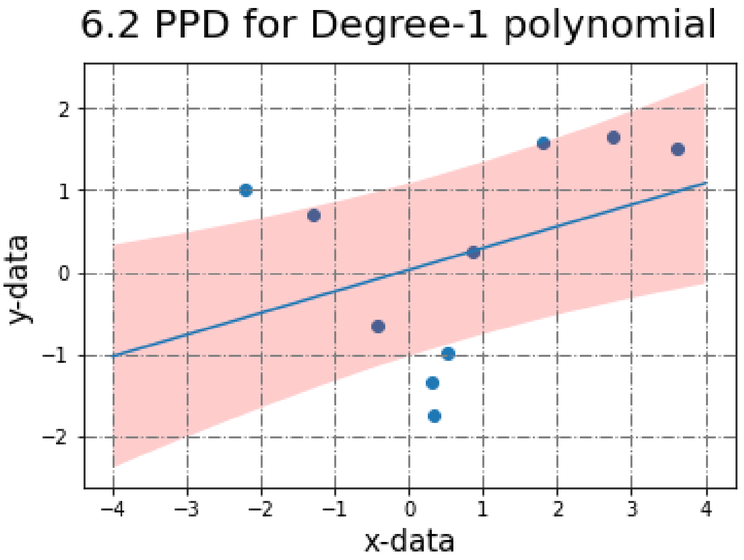

Figure 7: Mean of the PPD along with +/- 2 times standard deviation for deg 1 polynomial

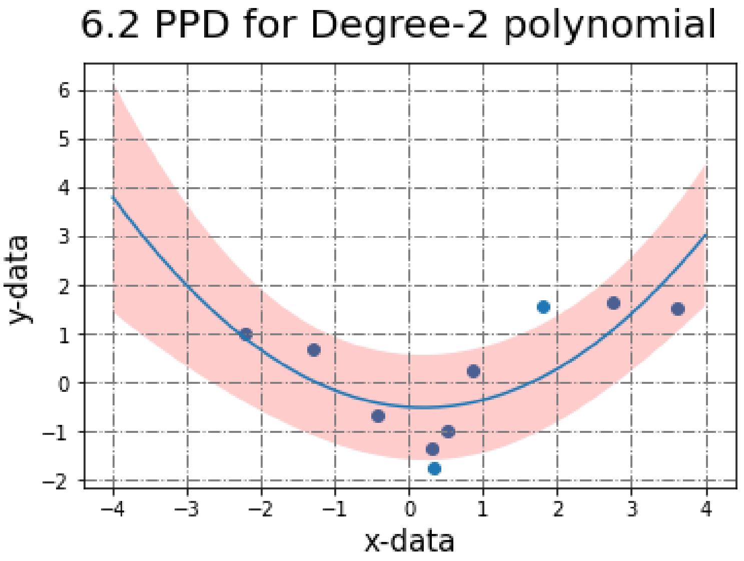

Figure 8: Mean of the PPD along with +/- 2 times the standard deviation for deg degree 2 polynomial

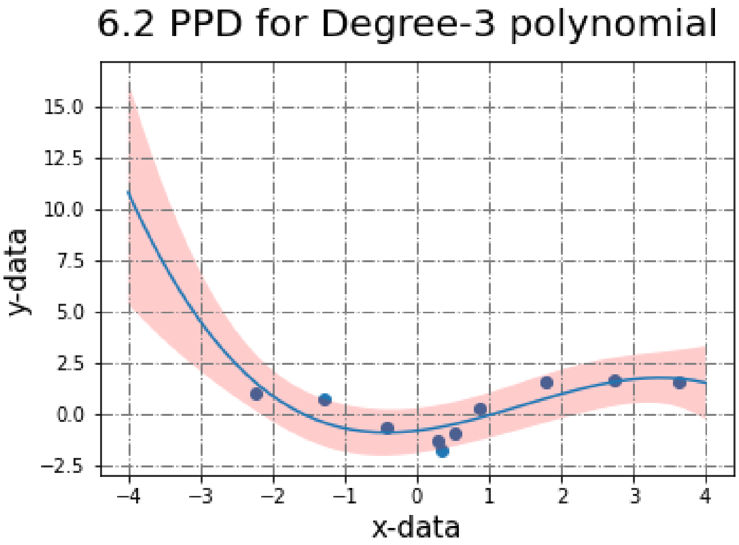

Figure 9: Mean of the PPD along with +/- 2 times the standard deviation for deg 3 polynomial

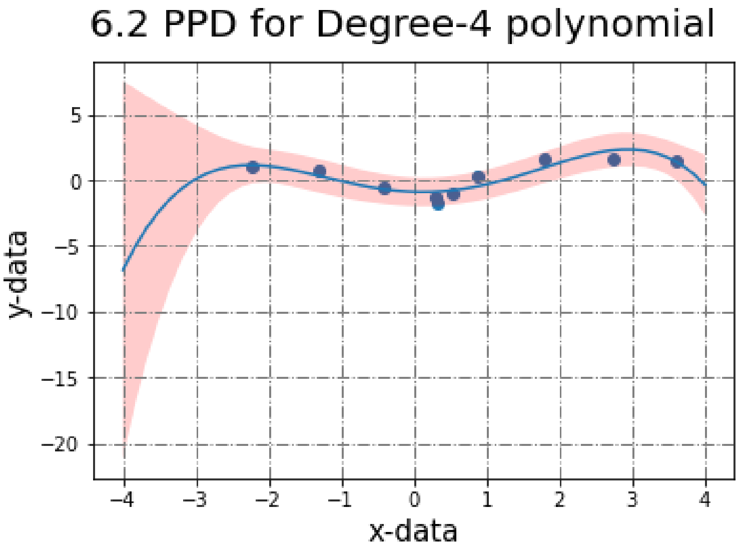

Figure 10: Mean of the PPD along with +/- 2 times the standard deviation for deg 4 polynomial

**My solution to problem 6.3**

Following are the Log marginal likelihood values

a) For Degree = 1 - Log marginal likelihood = -32.3520

b) For Degree = 2 - Log marginal likelihood = -22.7721

c) For Degree = 3 - Log marginal likelihood = -22.0790

d) For Degree = 4 - Log marginal likelihood = -22.3867

Model of Degree = 3 explains the data best both visually through graphs and also as per the log Marginal likelihood value.

**My solution to problem 6.4**

Following are the Log likelihood values

a) For Degree = 1 - Log likelihood = -28.0940

b) For Degree = 2 - Log likelihood = -15.3606

c) For Degree = 3 - Log likelihood = -10.9358

d) For Degree = 4 - Log likelihood = -7.2253

According to Log likelihood value Model of Degree = 4 is most suitable for the data.

log Marginal likelihood seems to be more sensible criteria to draw conclusion. Model of degree neither over fit nor under fit the data. This model also shows less uncertainty as compared to the other models.

**My solution to problem 6.5**

An additional training data in x range between -4 and -3 would certainly enhance the model performance as this regions lacks data points. The plots also signifies that the uncertainty in this region is high as compared to any other range of x. The additional data would narrow the shaded portions in this range.