Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 1

Student Name: Jinang Shah
Roll Number: 17807649
Date: February 26, 2021

Given, for a scalar random variable x, $p(x|\eta) = \mathcal{N}(0, \eta)$; $p(\eta|\gamma) = Exp(\eta|\gamma^2/2)$

Marginal Distribution $p(x|\gamma) = \int_\eta p(x|\eta)p(\eta|\gamma)d\eta = \int_\eta \frac{1}{\sqrt{2\pi\eta}} exp(\frac{-x^2}{2\eta})\ \frac{\gamma^2}{2} exp(\frac{-\gamma^2\eta}{2})d\eta$

Moment Generating Function of $p(x|\gamma) : M_X(t) = \int_x e^{tx}(\int_\eta \frac{1}{\sqrt{2\pi\eta}} exp(\frac{-x^2}{2\eta})\ \frac{\gamma^2}{2} exp(\frac{-\gamma^2\eta}{2})d\eta)dx$
$M_X(t) = \int_\eta \frac{\gamma^2}{2} exp(\frac{-\gamma^2\eta}{2})(\int_x \frac{1}{\sqrt{2\pi\eta}} e^{tx} exp(\frac{-x^2}{2\eta})dx)d\eta$

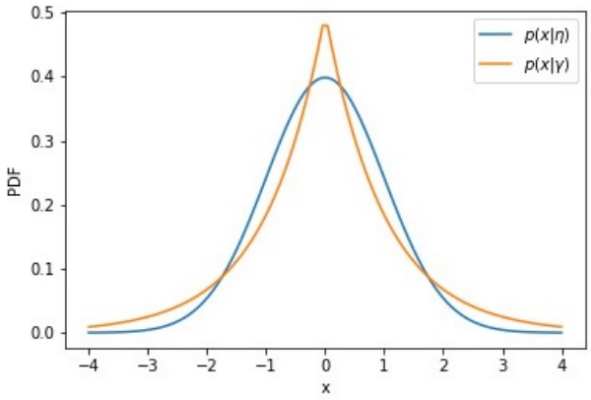
We are aware that MGF of $\mathcal{N}(0, \eta)$ which is, $exp(\frac{\eta t^2}{2}) \implies \int_x \frac{1}{\sqrt{2\pi\eta}} e^{tx} exp(\frac{-x^2}{2\eta})dx = exp(\frac{\eta t^2}{2})$

$M_X(t) = \int_{\eta=0}^{\infty} \frac{\gamma^2}{2} exp(\frac{-\gamma^2\eta}{2}) exp(\frac{\eta t^2}{2})d\eta = \int_{\eta=0}^{\infty} \frac{\gamma^2}{2} exp(\frac{\eta(t^2-\gamma^2)}{2})d\eta$

For $|t| < |\gamma|,\ M_X(t) = \frac{\gamma^2}{\gamma^2 - t^2}$, which is also MGF for Laplace distribution with $\mu = 0,\ b = 1/\gamma$

Therefore, $p(x|\gamma) = Laplace(x|0, 1/\gamma) = \frac{\gamma}{2} exp(-\gamma|x|)$

The marginal distribution $p(x|\gamma)$ is obtained by marginalizing over η i.e. integrating out a η. It gives us distribution for our observation which is only conditioned on the prior parameters (γ). This is also sometimes referred to as "model evidence" as it provides evidence for the given observation with the chosen prior belief.



Considering variance to be 1 for both $p(x|\gamma)$ *and* $p(x|\eta)$, we get the above graph. Here, it is clearly observable that the peak of $p(x|\gamma)$ is much sharper than the peak of $p(x|\eta)$.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 2

Student Name: Jinang Shah
Roll Number: 17807649
Date: February 26, 2021

For Bayesian Linear Regression with $p(w) = \mathcal{N}(0, \lambda^{-1}I)$, $p(y|x, w) = \mathcal{N}(w^T x, \beta^{-1})$

$p(y_*|x_*) = \mathcal{N}(\mu_N^T x_*, \beta^{-1} + x_*^T \Sigma_N x_*); \mu_N = \sum(\beta \sum_{n=1}^{N} y_n x_n); \Sigma_N = (\beta \sum_{n=1}^{N} x_n x_n^T + \lambda I)^{-1}$

Now, let us consider variance of posterior predictive distribution for $N+1$ samples, for that,
$\Sigma_{N+1} = (\beta \sum_{n=1}^{N+1} x_n x_n^T + \lambda I)^{-1} = (\Sigma_N^{-1} + (\beta^{0.5} x_{N+1})(\beta^{0.5} x_{N+1})^T)^{-1}$

Since, we are aware of the relationship given in the problem statement, we get,
$\Sigma_{N+1} = (\Sigma_N^{-1} + (\beta^{0.5} x_{N+1})(\beta^{0.5} x_{N+1})^T)^{-1} = \Sigma_N - \beta \frac{\Sigma_N x_{N+1} x_{N+1}^T \Sigma_N}{1+\beta(x_{N+1}^T \Sigma_N x_{N+1})}$

Now obtaining *var* of posterior predictive distribution for $N+1$ samples,
$var_{N+1} = \beta^{-1} + x_*^T \Sigma_{N+1} x_* = (\beta^{-1} + x_*^T \Sigma_N x_*) - x_*^T (\beta \frac{\Sigma_N x_{N+1} x_{N+1}^T \Sigma_N}{1+\beta(x_{N+1}^T \Sigma_N x_{N+1})}) x_*$
$var_{N+1} = var_N - (\beta \frac{x_*^T \Sigma_N x_{N+1} x_{N+1}^T \Sigma_N x_*}{1+\beta(x_{N+1}^T \Sigma_N x_{N+1})})$

Since Σ_N is a positive definite symmetric matrix, $x_*^T \Sigma_N x_{N+1} x_{N+1}^T \Sigma_N x_* = (x_*^T \Sigma_N x_{N+1})(x_*^T \Sigma_N x_{N+1})^T$, which implies that it will be too positive. Furthermore, since $\beta > 0$, we can also safely imply
$1 + \beta(x_{N+1}^T \Sigma_N x_{N+1}) > 0 \implies \mathbf{var_{N+1} < var_N}$

Hence, it can be observed that the variance of the predictive posterior decreases as we increase the training sample size.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 3

Student Name: Jinang Shah
Roll Number: 17807649
Date: February 26, 2021

Given N scalar observations sampled from $\mathcal{N}(\mu, \sigma^2)$
We have empirical Mean as $x' = \sum_{n=1}^{N} x_n/N$

We can define a random variable z such that x' can be written as its linear transformation,
$x' = A^T z;\ Here\ A = [1/N, 1/N, ..., 1/N]^T_{NX1};\ z = [x_1, ..., x_N]^T_{NX1}$

Here z can be assumed as a multivariate random variable, and furthermore, it can clearly be observed that since $x_1, ..., x_N$ are drawn i.i.d. from $\mathcal{N}(\mu, \sigma^2)$, we can safely say that z is a gaussian multivariate random variable with $E[z] = [\mu, ..., \mu]^T_{NX1}$ and $cov[z] = \sigma^2 I_N$.

Now, since x' is derived from linear transformation of z we can safely predict that x' will too follow gaussian distribution and its corresponding parameters can be calculated with the already known results in the following manner:

$E[x'] = E[A^T z] = A^T E[z] = \sum_{n=1}^{N} \frac{1}{N}\mu = \mu$

$cov[x'] = cov[A^T z] = A^T cov[z] A = A^T(\sigma^2 I_N)A = \sigma^2 A^T A = \sigma^2 \sum_{n=1}^{N} \frac{1}{N}\frac{1}{N} = \frac{\sigma^2}{N}$

Therefore probability distribution of x' here is $\mathcal{N}(\mu, \frac{\sigma^2}{N})$

This also makes an intuitive sense as all the observations are drawn i.i.d. that means all of them can be assumed to be following the same distribution even independently. Now, considering each of them as a separate variable will bring you to the same result.

More than that this result also makes sense as if you have more and more samples, with monte-carlo approximation you will be more nearer to the actual mean (μ) of the distribution. And that we can see here, as number of observation increases the covariance for empirical mean decreases as now it is highly likely that empirical mean will be much closer to the actual mean of distribution than it was before.

Mean of empirical mean distribution will remain same μ as even with higher number of observation as the variance decreases with higher number of observations, eventually it should converge to μ, the actual mean of the distribution.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 4

Student Name: Jinang Shah
Roll Number: 17807649
Date: February 26, 2021

Given, the scores of students in school m are drawn independently as $x_n^m \sim \mathcal{N}(\mu_m, \sigma^2)$; and $\mu_m \sim \mathcal{N}(\mu_0, \sigma_0^2)$ for *all* $m \in [1, M]$.

Posterior distribution of μ_m : $p(\mu_m|x^m, \sigma^2, \mu_0, \sigma_0^2) \propto p(x^m|\mu_m, \sigma^2)p(\mu_m|\mu_0, \sigma_0^2)$
$p(\mu_m|x^m, \sigma^2, \mu_0, \sigma_0^2) \propto (\prod_{n=1}^{N_m} \mathcal{N}(x_n^m|\mu_m, \sigma^2))\mathcal{N}(\mu_m|\mu_0, \sigma_0^2) \propto (\prod_{n=1}^{N_m} exp(\frac{-(x_n^m-\mu_m)^2}{2\sigma^2}))exp(\frac{-(\mu_m-\mu_0)^2}{2\sigma_0^2})$
$\mathbf{p}(\mu_\mathbf{m}|\mathbf{x^m}, \boldsymbol{\sigma^2}, \boldsymbol{\mu_0}, \boldsymbol{\sigma_0^2}) = \mathcal{N}(\mu_{N_m}, \sigma_{N_m}^2)$

Here, $\mu_{N_m} = \frac{\sigma^2}{\sigma^2+N_m\sigma_0^2}\mu_0 + \frac{N_m\sigma_0^2}{\sigma^2+N_m\sigma_0^2}\tilde{x^m}$; $\frac{1}{\sigma_{N_m}^2} = \frac{N_m}{\sigma^2} + \frac{1}{\sigma_0^2}$; $\tilde{x^m} = \frac{1}{N_m}\sum_{n=1}^{N_m} x_n^m$

For Marginal Likelihood: $p(x|\sigma^2, \mu_0, \sigma_0^2) = \prod_m p(x|\sigma^2, \mu_0, \sigma_0^2) = \prod_m \frac{p(x^m|\mu_m,\sigma^2)p(\mu_m|\mu_0,\sigma_0^2)}{p(\mu_m|x^m,\sigma^2,\mu_0,\sigma_0^2)}$
$p(x|\sigma^2, \mu_0, \sigma_0^2) = \prod_m \frac{(\prod_{n=1}^{N_m} \mathcal{N}(x_n^m|\mu_m,\sigma^2))\mathcal{N}(\mu_m|\mu_0,\sigma_0^2)}{\mathcal{N}(\mu_{N_m},\sigma_{N_m}^2)}$

Now using MLE-II to estimate μ_0: $\mu_0 = argmax_{\mu_0}\ log(\prod_m \frac{(\prod_{n=1}^{N_m} \mathcal{N}(x_n^m|\mu_m,\sigma^2))\mathcal{N}(\mu_m|\mu_0,\sigma_0^2)}{\mathcal{N}(\mu_{N_m},\sigma_{N_m}^2)})$
$\mu_0 = argmin_{\mu_0}\ L = argmin_{\mu_0}\ \sum_m(\frac{(\mu_m-\mu_0)^2}{2\sigma_0^2} - \frac{(\mu_m-\mu_{N_m})^2}{2\sigma^2})$

Differentiating w.r.t. μ_0 and equating to 0: $\frac{dL}{d\mu_0} = 0$

$-\sum_m \frac{N_m\mu_0-\sum_n=1^{N_m}x_n^m}{N_m\sigma_0^2+\sigma^2} = 0 \implies \mu_0^{MLE} = \frac{\sum_{m=1}^M \frac{N_m}{\sigma^2+N_m\sigma_0^2}\tilde{x^m}}{\sum_{m=1}^M \frac{N_m}{\sigma^2+N_m\sigma_0^2}}$

Now new estimate of $\mu_{N_m} = \frac{\sigma^2}{\sigma^2+N_m\sigma_0^2}(\frac{\sum_{m=1}^M \frac{N_m}{\sigma^2+N_m\sigma_0^2}\tilde{x^m}}{\sum_{m=1}^M \frac{N_m}{\sigma^2+N_m\sigma_0^2}}) + \frac{N_m\sigma_0^2}{\sigma^2+N_m\sigma_0^2}\tilde{x^m}$

The benefit of using MLE-II here is that we are now not just using a randomly or our believed prior but instead the beauty is that we are learning these hyperparameters from the data itself. This way we can avoid the personal bias and can choose the hyperparameter objectively which fit the data more perfectly. In this way we can incorporate data of all schools, which will further help our model, as now it will likely generalise well compared to before when we had a prior with a personal bias i.e. known value.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Jinang Shah
Roll Number: 17807649
Date: February 26, 2021

QUESTION 5

For Bayesian Linear Regression, $p(y^m|x^m, w_m) = \mathcal{N}(y^m|x^m w_m, \beta^{-1} I_{N_m}); p(w_m) = \mathcal{N}(w_0, \lambda^{-1} I_D)$

Now, we can write y^m as, $y^m = x^m w_m + \epsilon$; where $p(\epsilon) = \mathcal{N}(0, \beta^{-1} I_{N_m})$

Results and equation of PPD for bayesian linear regression can be used to find marginal likelihood. Where, only difference will be that prior terms will replace the respective posterior terms.

$p(y^m|x^m, w_0) = \mathcal{N}(x^m w_0, \beta^{-1} I_{N_m} + \lambda^{-1} x^m x^{mT})$

$\implies p(y^m|x^m, w_0) = \prod_{m=1}^{M} \mathcal{N}(x^m w_0, \beta^{-1} I_{N_m} + \lambda^{-1} x^m x^{mT})$

log of MLE II objective for estimating w_0:

$argmin_{w_0} \ \sum_{m=1}^{M} (y^m - x^m w_m)^T (\beta^{-1} I_{N_m} + \lambda^{-1} x^m x^{mT})^{-1} (y^m - x^m w_m)$

Just as mentioned in the problem 4, im my opinion using MLE II will be much better here. As we might not have a whole idea about the data and its nature while deciding a prior but using data to approximate the prior, keeps away our personal bias and also learn those values objectively.

More than that because of using MLE II for estimation of prior, we can safely assume that now this prior will help the better learning of the school specific weights and will help us fit our model to the data in relatively better way than by just using a "known" value of the prior. As data of these schools might have different distribution that the schools where we are taking this prior as reference from, so I think yes, it is always better to use such data based estimation of hyperparameters, as it will find better values for school specific weights according to the observations provided.

Topics in Probabilistic Modeling & Inference (CS698X), Spring 2021
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION
6

Student Name: Jinang Shah
Roll Number: 17807649
Date: February 26, 2021

6.3)

k = 1, log marginal likelihood: -32.352015280445244
k = 2, log marginal likelihood: -22.77215317878222
k = 3, log marginal likelihood: -22.079070642234182
k = 4, log marginal likelihood: -22.386776180355803

According to this data, model 3 seems to be explaining the data best as it has the highest log marginal likelihood.

6.4)

k = 1, log likelihood: -28.094004379075553
k = 2, log likelihood: -15.360663659052214
k = 3, log likelihood: -10.935846883615739
k = 4, log likelihood: -7.225291259028603

Now, seeing this data, the model with highest log likelihood is model 4, which is definitely not as our answer in 6.3 which found model 3 to be its best model.

I think the log marginal likelihood criteria seems better option here, as it doesn't take just one point estimate of a variable but instead integrates over all of its possbile values taking into account their uncertainty as well. Thus model will be more robust and will understand its position and its quality of prediction much much better than log likelihood which uses a point estimate, which doesn't give us any idea about how well a MAP estimate is. I mean we can't know from point estimate that how good or certain our estimate of parameter is either through MAP or through MLE.

Since, my best model is model 3, I will choose new data from the region [-4,-3]. As you can clearly see that model shows high variance there suggesting model is not certain about its prediction there. And thus to make model more certain where it's already not, I'll choose new samples in that region.

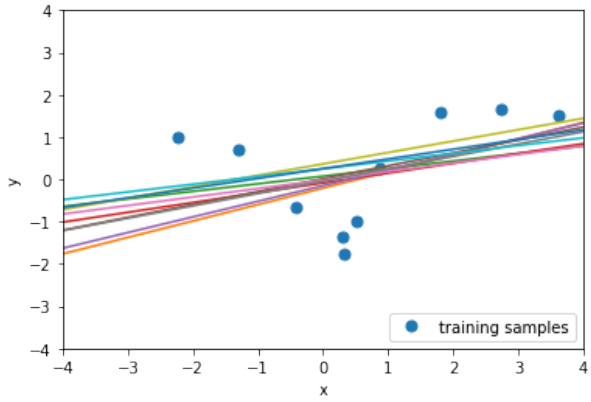


Figure 1: Caption

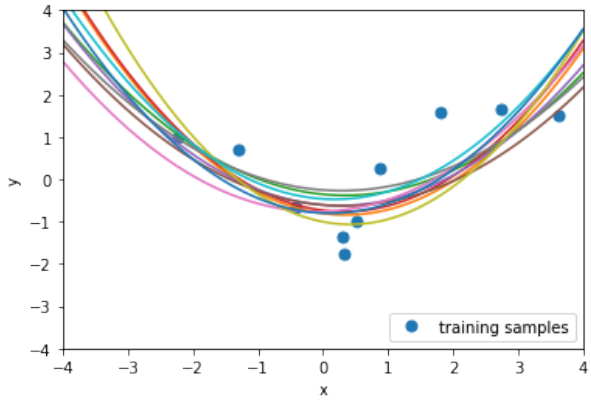


Figure 2: Caption

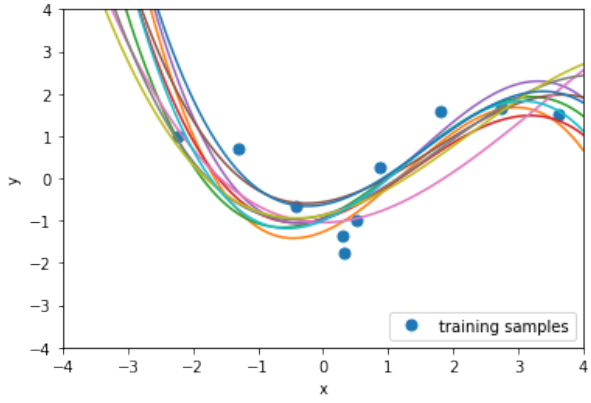


Figure 3: Caption

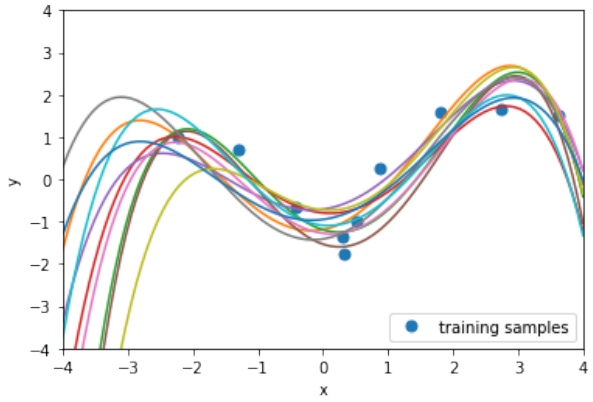


Figure 4: Caption

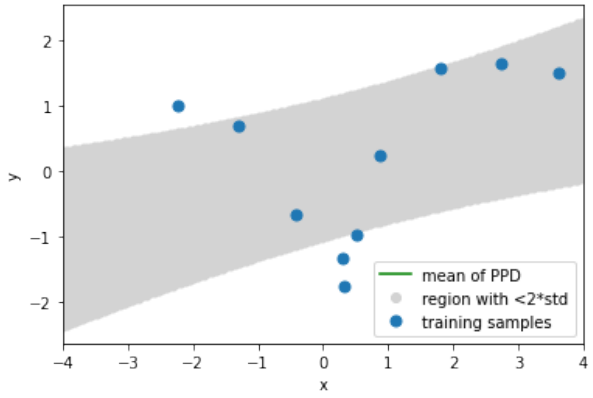


Figure 5: Caption

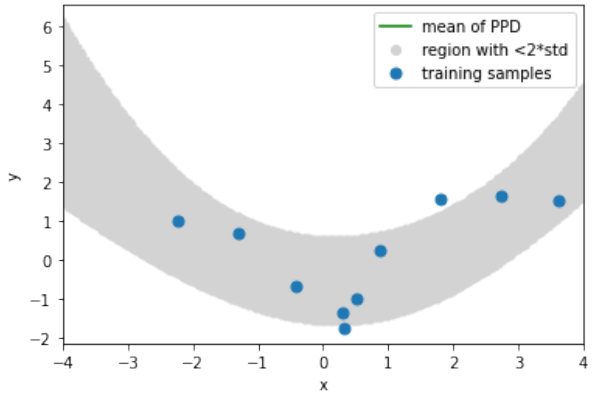


Figure 6: Caption

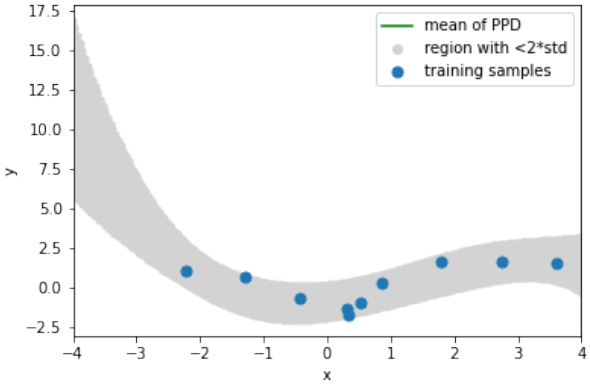


Figure 7: Caption

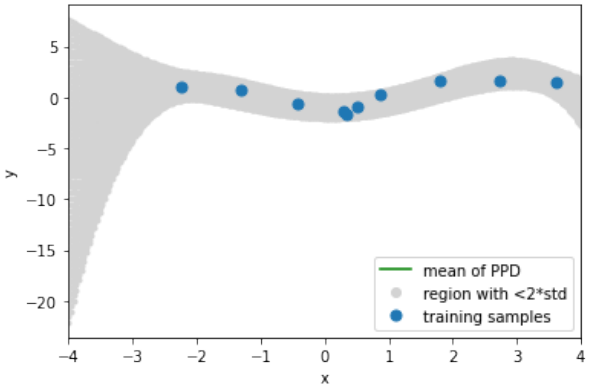


Figure 8: Caption