Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Atreya Goswami
Roll Number: 190201
Date: February 7, 2023

QUESTION 1

We are given a model m with parameters θ and hyperparameters λ. We assume prior $p(\theta \mid \lambda, m)$ on θ, $p(\lambda \mid m)$ on λ, and $p(m)$ on m. The likelihood is of the form $p(\mathbf{X} \mid \theta, \lambda, m)$ assuming that we have observed data $\mathbf{X}$.

1.1 Using Bayes rule,

$$p(\theta \mid \mathbf{X}, \lambda, m) = \frac{p(\mathbf{X} \mid \theta, \lambda, m)\, p(\theta \mid \lambda, m)}{p(\mathbf{X} \mid \lambda, m)} \tag{1.1.1}$$

Now, the marginal likelihood $p(\mathbf{X} \mid \lambda, m)$ can be obtained by marginalising the likelihood function $p(\mathbf{X} \mid \theta, \lambda, m)$ using the prior $p(\theta \mid \lambda, m)$.

$$p(\mathbf{X} \mid \lambda, m) = \int_\theta p(\mathbf{X} \mid \theta, \lambda, m)\, p(\theta \mid \lambda, m)\, d\theta \tag{1.1.2}$$

Using equations 1.1.1 and 1.1.2,

$$\boxed{p(\theta \mid \mathbf{X}, \lambda, m) = \frac{p(\mathbf{X} \mid \theta, \lambda, m)\, p(\theta \mid \lambda, m)}{\int_\theta p(\mathbf{X} \mid \theta, \lambda, m)\, p(\theta \mid \lambda, m)\, d\theta}} \tag{1.1.3}$$

1.2 Using Bayes rule,

$$p(\lambda \mid \mathbf{X}, m) = \frac{p(\mathbf{X} \mid \lambda, m)\, p(\lambda \mid m)}{p(\mathbf{X} \mid m)} \tag{1.2.1}$$

Now, the marginal likelihood $p(\mathbf{X} \mid m)$ can be obtained by marginalising the marginal likelihood function $p(\mathbf{X} \mid \lambda, m)$ using the prior $p(\lambda \mid m)$.

$$p(\mathbf{X} \mid m) = \int_\lambda p(\mathbf{X} \mid \lambda, m)\, p(\lambda \mid m)\, d\lambda \tag{1.2.2}$$

Using equations 1.1.2, 1.2.1 and 1.2.2,

$$\boxed{p(\lambda \mid \mathbf{X}, m) = \frac{p(\mathbf{X} \mid \lambda, m)\, p(\lambda \mid m)}{\int_\lambda p(\mathbf{X} \mid \lambda, m)\, p(\lambda \mid m)\, d\lambda}} \tag{1.2.3}$$

1.3 Using Bayes rule,

$$p(m \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid m)\, p(m)}{p(\mathbf{X})} \tag{1.3.1}$$

Now, the marginal likelihood $p\left(\mathbf{X}\right)$ can be obtained by marginalising the marginal likelihood function $p\left(\mathbf{X} \,|\, m\right)$ using the prior $p\left(m\right)$.

$$p\left(\mathbf{X}\right) = \sum_{m} p\left(\mathbf{X} \,|\, m\right) p\left(m\right) \tag{1.3.2}$$

Using equations 1.2.2, 1.3.1 and 1.3.2.

$$\boxed{p\left(m \,|\, \mathbf{X}\right) = \frac{p\left(\mathbf{X} \,|\, m\right) p\left(m\right)}{\sum_{m} p\left(\mathbf{X} \,|\, m\right) p\left(m\right)}} \tag{1.3.3}$$

The difficulty in computing the mentioned quantities using the expressions obtained in parts **1.1**, **1.2** and **1.3** depends on the number of marginalisations that we need to perform while computing the aforementioned expressions. The more the number of marginalisations required, the more difficult it is to compute the quantity.

In the computation of $p\left(\theta \,|\, \mathbf{X}, \lambda, m\right)$ (in part **1.1**), we need to perform one marginalisation as described by equation 1.1.2.

In the computation of $p\left(\lambda \,|\, \mathbf{X}, m\right)$ (in part **1.2**), we need to perform two marginalisations as described by equations 1.1.2 and 1.2.2.

In the computation of $p\left(m \,|\, \mathbf{X}\right)$ (in part **1.3**), we need to perform three marginalisations as described by equations 1.1.2, 1.2.2 and 1.3.2.

Hence the ranking of the three quantities in terms of the difficulty of computing them (easiest to hardest) will be

$$\boxed{p\left(\theta \,|\, \mathbf{X}, \lambda, m\right) < p\left(\lambda \,|\, \mathbf{X}, m\right) < p\left(m \,|\, \mathbf{X}\right)}$$

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 2

Student Name: Atreya Goswami
Roll Number: 190201
Date: February 7, 2023

Given the following expressions for likelihood, prior, posterior and posterior predictive for a Bayesian linear regression setup, we need to determine what happens to the posterior predictive variance as we increase the size of the training set N.

Prior:	$p(\boldsymbol{w}) = \mathcal{N}\left(\boldsymbol{w} \mid \mathbf{0}, \lambda^{-1}\mathbf{I}\right)$
Likelihood:	$p(y \mid \boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}\left(y \mid \boldsymbol{w}^\top \boldsymbol{x}, \beta^{-1}\right)$
Posterior:	$p(\boldsymbol{w} \mid \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_n, y_1, y_2, \cdots y_n) = \mathcal{N}\left(\boldsymbol{w} \mid \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N\right)$
Posterior Predictive:	$p(y_* \mid \boldsymbol{x}_*) = \mathcal{N}\left(y_* \mid \boldsymbol{\mu}_N^\top \boldsymbol{x}_*, \sigma_N^2(\boldsymbol{x}_*)\right)$

where $\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N \left(\beta \sum_{n=1}^N y_n \boldsymbol{x}_n\right)$, $\boldsymbol{\Sigma}_N = \left(\beta \sum_{n=1}^N \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I}\right)^{-1}$ and $\sigma_N^2(\boldsymbol{x}_*) = \beta^{-1} + \boldsymbol{x}_*^\top \boldsymbol{\Sigma}_N \boldsymbol{x}_*$

Claim 2.1. The variance of the posterior predictive $\sigma_N^2(\boldsymbol{x}_*)$ increases with N.

Proof. To establish our claim, it is sufficient to prove that for any $K \in \mathbb{N}$, $\sigma_{K+1}^2(\boldsymbol{x}_*) > \sigma_K^2(\boldsymbol{x}_*)$.

$$\begin{aligned}
\boldsymbol{\Sigma}_{K+1} &= \left(\beta \sum_{n=1}^{K+1} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I}\right)^{-1} \\
&= \left(\beta \boldsymbol{x}_{K+1} \boldsymbol{x}_{K+1}^\top + \beta \sum_{n=1}^{K} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I}\right)^{-1} \\
&= \left(\beta \boldsymbol{x}_{K+1} \boldsymbol{x}_{K+1}^\top + \boldsymbol{\Sigma}_K^{-1}\right)^{-1} \\
&= \left[\boldsymbol{\Sigma}_K^{-1} + \left(\beta^{0.5} \boldsymbol{x}_{K+1}\right)\left(\beta^{0.5} \boldsymbol{x}_{K+1}\right)^\top\right]^{-1} \qquad (2.1.1)
\end{aligned}$$

Identity 2.2.

$$\left(\mathbf{M} + \boldsymbol{v}\boldsymbol{v}^\top\right)^{-1} = \mathbf{M}^{-1} - \frac{\left(\mathbf{M}^{-1}\boldsymbol{v}\right)\left(\boldsymbol{v}^\top \mathbf{M}^{-1}\right)}{1 + \boldsymbol{v}^\top \mathbf{M}^{-1} \boldsymbol{v}}$$

Using the matrix identity 2.2, we can rewrite the expression in 2.1.1 as

$$\boldsymbol{\Sigma}_{K+1} = \boldsymbol{\Sigma}_K - \frac{\beta \boldsymbol{\Sigma}_K \boldsymbol{x}_{K+1} \boldsymbol{x}_{K+1}^\top \boldsymbol{\Sigma}_K}{1 + \beta \boldsymbol{x}_{K+1}^\top \boldsymbol{\Sigma}_K \boldsymbol{x}_{K+1}} \qquad (2.1.2)$$

Now, we substitute for $\mathbf{\Sigma}_{K+1}$ in the expression of $\sigma^2_{K+1}\left(\boldsymbol{x}_*\right)$ using equation 2.1.2.

$$\begin{aligned}
\sigma^2_{K+1}\left(\boldsymbol{x}_*\right) &= \beta^{-1} + \boldsymbol{x}_*^\top \mathbf{\Sigma}_{K+1} \boldsymbol{x}_* \\
&= \beta^{-1} + \boldsymbol{x}_*^\top \mathbf{\Sigma}_K \boldsymbol{x}_* - \boldsymbol{x}_*^\top \frac{\beta \mathbf{\Sigma}_K \boldsymbol{x}_{K+1} \boldsymbol{x}_{K+1}^\top \mathbf{\Sigma}_K}{1 + \beta \boldsymbol{x}_{K+1}^\top \mathbf{\Sigma}_K \boldsymbol{x}_{K+1}} \boldsymbol{x}_* \\
&= \sigma^2_K\left(\boldsymbol{x}_*\right) - \beta \boldsymbol{x}_*^\top \frac{\mathbf{\Sigma}_K \boldsymbol{x}_{K+1} \boldsymbol{x}_{K+1}^\top \mathbf{\Sigma}_K}{1 + \beta \boldsymbol{x}_{K+1}^\top \mathbf{\Sigma}_K \boldsymbol{x}_{K+1}} \boldsymbol{x}_* \qquad (2.1.3)
\end{aligned}$$

Since the covariance matrix $\mathbf{\Sigma}_K$ is a symmetric, positive semi-definite matrix by definition [Mur22, p. 77],

$$\begin{aligned}
\boldsymbol{x}_{K+1}^\top\, \mathbf{\Sigma}_K\, \boldsymbol{x}_{K+1} &\geq 0 \\
\text{or,} \quad 1 + \beta \boldsymbol{x}_{K+1}^\top\, \mathbf{\Sigma}_K\, \boldsymbol{x}_{K+1} &> 0 \qquad (2.1.4)
\end{aligned}$$

Also,

$$\begin{aligned}
\mathbf{\Sigma}_K \boldsymbol{x}_{K+1} \boldsymbol{x}_{K+1}^\top \mathbf{\Sigma}_K &= \mathbf{\Sigma}_K^\top \boldsymbol{x}_{K+1} \boldsymbol{x}_{K+1}^\top \mathbf{\Sigma}_K \\
&= \left(\boldsymbol{x}_{K+1}^\top \mathbf{\Sigma}_K\right)^\top \left(\boldsymbol{x}_{K+1}^\top \mathbf{\Sigma}_K\right) \quad = \quad \mathbf{A} \quad (\text{let})
\end{aligned}$$

Since $\mathbf{A}$ is a gram matrix by definition, so it is a positive semi-definite matrix [Mur22, p. 237]. Hence we can write

$$\begin{aligned}
\boldsymbol{x}_*^\top \mathbf{A} \boldsymbol{x}_* &\geq 0 \\
\text{or,} \quad \beta \boldsymbol{x}_*^\top \mathbf{\Sigma}_K \boldsymbol{x}_{K+1} \boldsymbol{x}_{K+1}^\top \mathbf{\Sigma}_K \boldsymbol{x}_* &\geq 0 \qquad (2.1.5)
\end{aligned}$$

Now, using equations 2.1.4 and 2.1.5, we can write

$$\beta \boldsymbol{x}_*^\top \frac{\mathbf{\Sigma}_K \boldsymbol{x}_{K+1} \boldsymbol{x}_{K+1}^\top \mathbf{\Sigma}_K}{1 + \beta \boldsymbol{x}_{K+1}^\top \mathbf{\Sigma}_K \boldsymbol{x}_{K+1}} \boldsymbol{x}_* \quad > \quad 0 \qquad (2.1.6)$$

Using equations 2.1.3 and 2.1.6,

$$\sigma^2_{K+1}\left(\boldsymbol{x}_*\right) > \sigma^2_K\left(\boldsymbol{x}_*\right) \qquad (2.1.7)$$

Thus, we have proved our claim that $\sigma^2_N\left(\boldsymbol{x}_*\right)$ increases with N.

■

As the size of the training set N increases, the posterior predictive variance $\sigma^2_N\left(\boldsymbol{x}_*\right)$ increases.

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Atreya Goswami
Roll Number: 190201
Date: February 7, 2023

QUESTION 3

We are given that x is a scalar drawn from the univariate Gaussian $p\left(x \mid \eta\right) = \mathcal{N}\left(x \mid 0, \eta\right)$ and the variance η is drawn from an exponential distribution $p\left(\eta \mid \gamma\right) = \text{Exp}\left(\eta \mid \frac{\gamma^2}{2}\right) = \frac{\gamma^2}{2}\exp\left(-\frac{\gamma^2}{2}\eta\right)$. We need to derive the marginal distribution $p\left(x \mid \gamma\right) = \int_\eta p\left(x \mid \eta\right) p\left(\eta \mid \gamma\right) d\eta$. Since the integral is hard to compute, we instead compute the MGF of the marginal distribution.

By definition, the moment generating function of any distribution with pdf $p\left(x\right)$ is given by

$$\mathcal{M}_X\left(t\right) = \int_{-\infty}^{\infty} e^{tx} p\left(x\right) dx$$

Hence, MGF of the marginal likelihood $p\left(x \mid \gamma\right)$ is given by

$$\begin{aligned}
& \int_x e^{tx} \int_\eta p\left(x \mid \eta\right) p\left(\eta \mid \gamma\right) d\eta dx \\
= \quad & \int_x \int_\eta \frac{\gamma^2}{2} e^{tx} \mathcal{N}\left(x \mid 0, \eta\right) \exp\left(-\frac{\gamma^2}{2}\eta\right) d\eta dx \\
= \quad & \frac{\gamma^2}{2} \int_\eta \int_x e^{tx} \frac{1}{\sqrt{2\pi\eta}} \exp\left(-\frac{x^2}{2\eta}\right) \exp\left(-\frac{\gamma^2}{2}\eta\right) dx d\eta \\
= \quad & \frac{\gamma^2}{2} \int_\eta \int_x \frac{1}{\sqrt{2\pi\eta}} \exp\left(tx - \frac{x^2}{2\eta} - \frac{\gamma^2}{2}\eta\right) dx d\eta \\
= \quad & \frac{\gamma^2}{2} \int_\eta \int_x \frac{1}{\sqrt{2\pi\eta}} \exp\left[-\frac{\left(x - \eta t\right)^2}{2\eta} - \frac{\gamma^2 - t^2}{2}\eta\right] dx d\eta \\
= \quad & \frac{\gamma^2}{2} \int_\eta \exp\left[-\frac{\gamma^2 - t^2}{2}\eta\right] \int_x \frac{1}{\sqrt{2\pi\eta}} \exp\left[-\frac{\left(x - \eta t\right)^2}{2\eta}\right] dx d\eta \\
= \quad & \frac{\gamma^2}{2} \int_\eta \exp\left[-\frac{\gamma^2 - t^2}{2}\eta\right] \left[\int_{-\infty}^{\infty} \mathcal{N}\left(x \mid \eta t, \eta\right) dx\right] d\eta \\
= \quad & \frac{\gamma^2}{2} \int_0^{\infty} \exp\left[-\frac{\gamma^2 - t^2}{2}\eta\right] d\eta \qquad \left(\because \int_{-\infty}^{\infty} \mathcal{N}\left(x \mid \eta t, \eta\right) dx = 1\right) \\
= \quad & \frac{\gamma^2}{2} \left. \frac{\exp\left(-\frac{\gamma^2 - t^2}{2}\eta\right)}{-\frac{\gamma^2 - t^2}{2}} \right|_0^{\infty} \\
= \quad & \frac{\gamma^2}{\gamma^2 - t^2} \quad = \quad \frac{1}{1 - \frac{t^2}{\gamma^2}}
\end{aligned}$$

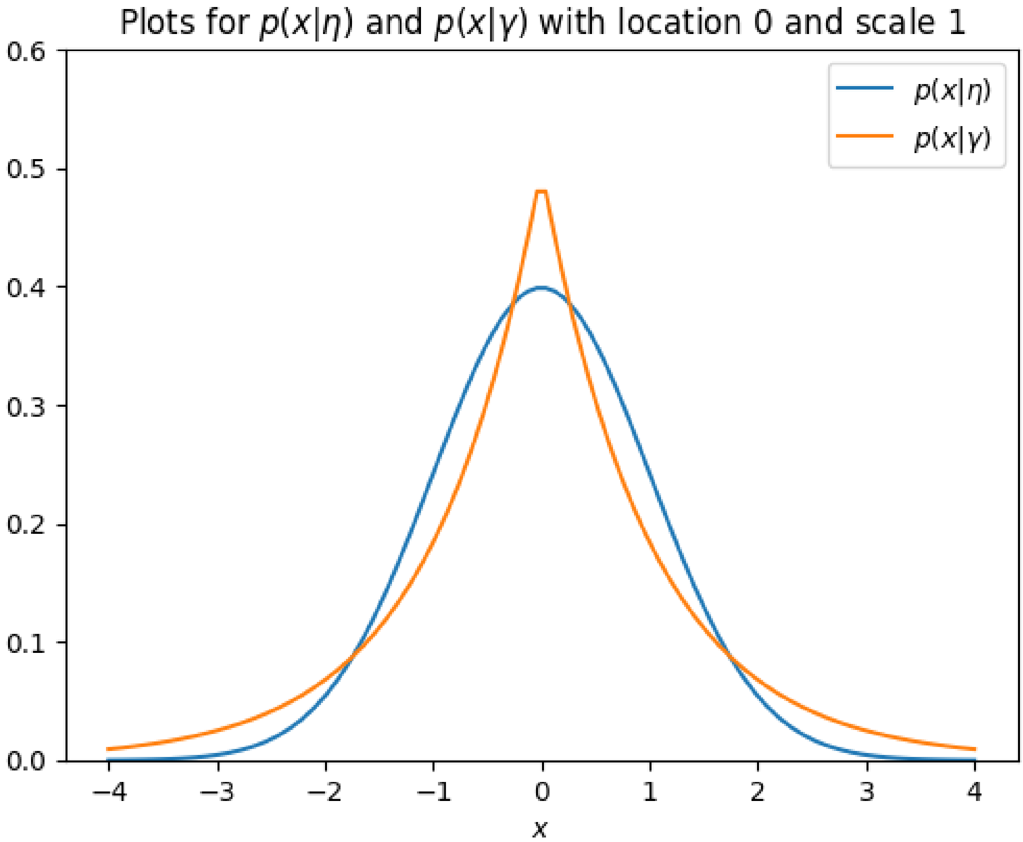
From the table of MGFs of various pdfs on the Wikipedia page https://en.wikipedia.org/wiki/Moment-generating_function, we know the MGF of the Laplace distribution $\mathcal{L}(\mu, b)$ is given by

$$\frac{e^{t\mu}}{1 - b^2\, t^2}, \qquad |\, t\,| < \frac{1}{b}$$

By comparing the obtained MGF with that of the Laplace distribution and choosing $|\, t\,| \ < \gamma$, we can conclude that the marginal distribution $p\,(x\,|\,\gamma)$ is a Laplace distribution with parameters 0 and $\frac{1}{\gamma}$.

$$p\,(x\,|\,\gamma) = \mathcal{L}\left(x\,|\,0, \frac{1}{\gamma}\right)$$

The marginal distribution $p\,(x\,|\,\gamma)$ is a distribution of the scalar x conditioned only on the prior parameters γ. This is derived by marginalising the distribution $p\,(x\,|\,\eta)$ over η using the prior $p\,(\eta\,|\,\gamma)$.



Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Atreya Goswami
Roll Number: 190201
Date: February 7, 2023

QUESTION 4

We are given student data from M schools, N_m is the number of students in school m. For student n in school m, we are given a response variable $y_n^{(m)} \in \mathbb{R}$ and a feature vector $\boldsymbol{x}_n^{(m)} \in \mathbb{R}^D$. We assume a linear regression model to model the response variable as

$$p\left(\boldsymbol{y}^{(m)} \,|\, \mathbf{X}^{(m)}, \boldsymbol{w}_m\right) = \mathcal{N}\left(\boldsymbol{y}^{(m)} \,|\, \mathbf{X}^{(m)}\boldsymbol{w}_m,\ \beta^{-1}\mathbf{I}_{N_m}\right)$$

where $\boldsymbol{w}^{(m)} \in \mathbb{R}^D$ is the regression weight vector for school m, $\mathbf{X}^{(m)}$ and $\boldsymbol{y}^{(m)}$ are $\boldsymbol{x}_n^{(m)\top}$ and $y_n^{(m)}$ respectively stacked vertically for all the students of school m and have dimensions $N_m \times D$ and $N_m \times 1$ respectively. We assume the prior on the weight vector $\boldsymbol{w}_m$ for school m to be $p\left(\boldsymbol{w}_m \,|\, \boldsymbol{w}_0\right) = \mathcal{N}\left(\boldsymbol{w}_m \,|\, \boldsymbol{w}_0, \lambda^{-1}\mathbf{I}_D\right)$. λ and β are assumed to be known, $\boldsymbol{w}_0$ is unknown.

For school m, given the likelihood $p\left(\boldsymbol{y}^{(m)} \,|\, \mathbf{X}^{(m)}, \boldsymbol{w}_m\right)$ and prior $p\left(\boldsymbol{w}_m \,|\, \boldsymbol{w}_0\right)$, we can derive the marginal likelihood $p\left(\boldsymbol{y}^{(m)} \,|\, \mathbf{X}^{(m)}\right)$ as

$$\begin{aligned} p\left(\boldsymbol{y}^{(m)} \,|\, \mathbf{X}^{(m)}, \boldsymbol{w}_0\right) &= \int_{\boldsymbol{w}_m} p\left(\boldsymbol{y}^{(m)} \,|\, \mathbf{X}^{(m)}, \boldsymbol{w}_m\right) p\left(\boldsymbol{w}_m \,|\, \boldsymbol{w}_0\right) d\boldsymbol{w}_m \\ &= \int_{\boldsymbol{w}_m} \mathcal{N}\left(\boldsymbol{y}^{(m)} \,|\, \mathbf{X}^{(m)}\boldsymbol{w}_m,\ \beta^{-1}\mathbf{I}_{N_m}\right) \mathcal{N}\left(\boldsymbol{w}_m \,|\, \boldsymbol{w}_0,\ \lambda^{-1}\mathbf{I}_D\right) d\boldsymbol{w}_m \end{aligned}$$

Using the results for Linear Gaussian Models presented on Slide 11 in Lecture 5 [Bis06, 2.115],

$$p\left(\boldsymbol{y}^{(m)} \,|\, \mathbf{X}^{(m)}, \boldsymbol{w}_0\right) = \mathcal{N}\left(\boldsymbol{y}^{(m)} \,|\, \mathbf{X}^{(m)}\boldsymbol{w}_0,\ \beta^{-1}\mathbf{I}_{N_m} + \lambda^{-1}\mathbf{X}^{(m)}\mathbf{X}^{(m)\top}\right) \tag{4.1}$$

Assuming the data from all the schools to be i.i.d., we can write the marginal likelihood $p\left(\boldsymbol{y} \,|\, \mathbf{X}, \boldsymbol{w}_0\right)$ as

$$p\left(\boldsymbol{y} \,|\, \mathbf{X}, \boldsymbol{w}_0\right) = \prod_{m=1}^{M} p\left(\boldsymbol{y}^{(m)} \,|\, \mathbf{X}^{(m)}, \boldsymbol{w}_0\right) \tag{4.2}$$

Now, to get the MLE estimate of $\boldsymbol{w}_0$, we need to need to maximise the log of marginal likelihood function $p\left(\boldsymbol{y} \,|\, \mathbf{X}, \boldsymbol{w}_0\right)$ w.r.t. $\boldsymbol{w}_0$.

$$\hat{\boldsymbol{w}}_{0,\,\text{MLE}} = \arg\max_{\boldsymbol{w}_0} f(\boldsymbol{w}_0)$$

where,

$$\begin{aligned} f(\boldsymbol{w}_0) &= \log p\left(\boldsymbol{y} \,|\, \mathbf{X}, \boldsymbol{w}_0\right) \\ &= \log \prod_{m=1}^{M} p\left(\boldsymbol{y}^{(m)} \,|\, \mathbf{X}^{(m)}, \boldsymbol{w}_0\right) \qquad \text{(Using equation 4.2)} \\ &= \sum_{m=1}^{M} \log \mathcal{N}\left(\boldsymbol{y}^{(m)} \,|\, \mathbf{X}^{(m)}\boldsymbol{w}_0,\ \beta^{-1}\mathbf{I}_{N_m} + \lambda^{-1}\mathbf{X}^{(m)}\mathbf{X}^{(m)\top}\right) \qquad \text{(Using equation 4.1)} \\ &= C - \sum_{m=1}^{M} \left(\boldsymbol{y}^{(m)} - \mathbf{X}^{(m)}\boldsymbol{w}_0\right)^{\top} \left(\beta^{-1}\mathbf{I}_{N_m} + \lambda^{-1}\mathbf{X}^{(m)}\mathbf{X}^{(m)\top}\right)^{-1} \left(\boldsymbol{y}^{(m)} - \mathbf{X}^{(m)}\boldsymbol{w}_0\right) \end{aligned}$$

Ignoring terms constant w.r.t. $\boldsymbol{w}_0$, the final expression of the objective function for MLE-II is given by

$$\boxed{f(\boldsymbol{w}_0) \quad = \quad - \sum_{m=1}^{M} \left(\boldsymbol{y}^{(m)} - \mathbf{X}^{(m)}\boldsymbol{w}_0\right)^{\top} \left(\beta^{-1}\mathbf{I}_{N_m} + \lambda^{-1}\mathbf{X}^{(m)}\mathbf{X}^{(m)\top}\right)^{-1} \left(\boldsymbol{y}^{(m)} - \mathbf{X}^{(m)}\boldsymbol{w}_0\right)}$$

Estimating the mean of the prior for the weight vectors for each school using MLE-II will give us much better estimation for school-specific weight vectors, instead of simply using a fixed value of w_0 from some previous experiment or intuition. This eliminates any bias in our belief about the prior mean and instead learn the mean objectively from the data itself. This will help in better fitting of the model to estimate the school specific weight vectors.

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 5

Student Name: Atreya Goswami
Roll Number: 190201
Date: February 7, 2023

We are given a regression model with input $\boldsymbol{x} \in \mathbb{R}^D$ and output $y \in \mathbb{R}$. The joint probability distribution of the inputs and outputs is

$$p(\boldsymbol{x}, y) = \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{x} - \boldsymbol{x}_n, y - y_n) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{N}\left(\left[(\boldsymbol{x} - \boldsymbol{x}_n)^\top, y - y_n\right]^\top \,|\, \mathbf{0}, \sigma^2\mathbf{I}_{D+1}\right)$$

where $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ denote the training examples. We are required to derive the conditional distribution $p(y \,|\, \boldsymbol{x})$ and $\mathbb{E}[y \,|\, \boldsymbol{x}]$.

First, we observe that the joint distribution $p(\boldsymbol{x}, y)$ is a mixture of N normal joint probability distributions, each of which can be easily factorised (into the marginal distributions $p(\boldsymbol{x})$ and $p(y)$) as their covariance matrix $\sigma^2\mathbf{I}_{D+1}$ is diagonal. Thus we can write

$$\begin{aligned}
&\mathcal{N}\left(\left[(\boldsymbol{x} - \boldsymbol{x}_n)^\top, y - y_n\right]^\top \,|\, \mathbf{0}, \sigma^2\mathbf{I}_{D+1}\right) \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{D+1}{2}}}\exp\left[-\frac{1}{2\sigma^2}\left\{(\boldsymbol{x} - \boldsymbol{x}_n)^\top(\boldsymbol{x} - \boldsymbol{x}_n) + (y - y_n)^2\right\}\right] \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{D}{2}}}\exp\left[-\frac{1}{2\sigma^2}\left\{(\boldsymbol{x} - \boldsymbol{x}_n)^\top(\boldsymbol{x} - \boldsymbol{x}_n)\right\}\right]\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left[-\frac{1}{2\sigma^2}(y - y_n)^2\right] \\
&= \mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \,|\, \mathbf{0}, \sigma^2\mathbf{I}_D\right)\,\mathcal{N}\left(y - y_n \,|\, 0, \sigma^2\right) \qquad (5.1)
\end{aligned}$$

Now, we derive the marginal distribution $p(\boldsymbol{x})$ from the joint distribution $p(\boldsymbol{x}, y)$ by integrating out y.

$$\begin{aligned}
p(\boldsymbol{x}) &= \int_y p(\boldsymbol{x}, y)\,dy \\
&= \int_y \frac{1}{N}\sum_{n=1}^{N}\mathcal{N}\left(\left[(\boldsymbol{x} - \boldsymbol{x}_n)^\top, y - y_n\right]^\top \,|\, \mathbf{0}, \sigma^2\mathbf{I}_{D+1}\right)dy \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \,|\, \mathbf{0}, \sigma^2\mathbf{I}_D\right)\int_y \mathcal{N}\left(y - y_n \,|\, 0, \sigma^2\right)dy \qquad \text{(using equation 5.1)} \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \,|\, \mathbf{0}, \sigma^2\mathbf{I}_D\right) \qquad (5.2)
\end{aligned}$$

Using Bayes rule,

$$\begin{aligned}
p(y \,|\, \boldsymbol{x}) &= \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})} \\
&= \frac{\frac{1}{N}\sum_{n=1}^{N}\mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \,|\, \mathbf{0}, \sigma^2\mathbf{I}_D\right)\,\mathcal{N}\left(y - y_n \,|\, 0, \sigma^2\right)}{\frac{1}{N}\sum_{n=1}^{N}\mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \,|\, \mathbf{0}, \sigma^2\mathbf{I}_D\right)}
\end{aligned}$$

Thus,

$$\boxed{p\left(y \mid \boldsymbol{x}\right) \quad = \quad \sum_{n=1}^{N} \mathcal{N}\left(y - y_n \mid 0, \sigma^2\right) \left[\frac{\mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \mid \mathbf{0}, \sigma^2 \mathbf{I}_D\right)}{\sum_{n=1}^{N} \mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \mid \mathbf{0}, \sigma^2 \mathbf{I}_D\right)}\right]} \tag{5.3}$$

Now by definition,

$$\begin{aligned}
\mathbb{E}\left[y \mid \boldsymbol{x}\right] &= \int_y y\, p\left(y \mid \boldsymbol{x}\right) dy \\
&= \int_y y \sum_{n=1}^{N} \mathcal{N}\left(y - y_n \mid 0, \sigma^2\right) \left[\frac{\mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \mid \mathbf{0}, \sigma^2 \mathbf{I}_D\right)}{\sum_{n=1}^{N} \mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \mid \mathbf{0}, \sigma^2 \mathbf{I}_D\right)}\right] dy \\
&= \sum_{n=1}^{N} \left[\frac{\mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \mid \mathbf{0}, \sigma^2 \mathbf{I}_D\right)}{\sum_{n=1}^{N} \mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \mid \mathbf{0}, \sigma^2 \mathbf{I}_D\right)}\right] \int_y \left(y - y_n + y_n\right) \mathcal{N}\left(y - y_n \mid 0, \sigma^2\right) dy \\
&= \sum_{n=1}^{N} \left[\frac{\mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \mid \mathbf{0}, \sigma^2 \mathbf{I}_D\right)}{\sum_{n=1}^{N} \mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \mid \mathbf{0}, \sigma^2 \mathbf{I}_D\right)}\right] \left\{\int_y \left(y - y_n\right) \mathcal{N}\left(y - y_n \mid 0, \sigma^2\right) dy + y_n\right\} \\
&= \sum_{n=1}^{N} \left[\frac{\mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \mid \mathbf{0}, \sigma^2 \mathbf{I}_D\right)}{\sum_{n=1}^{N} \mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \mid \mathbf{0}, \sigma^2 \mathbf{I}_D\right)}\right] \left(\mathbb{E}\left[y - y_n\right] + y_n\right)
\end{aligned}$$

Thus,

$$\boxed{\mathbb{E}\left[y \mid \boldsymbol{x}\right] \quad = \quad \sum_{n=1}^{N} y_n \left[\frac{\mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \mid \mathbf{0}, \sigma^2 \mathbf{I}_D\right)}{\sum_{n=1}^{N} \mathcal{N}\left(\boldsymbol{x} - \boldsymbol{x}_n \mid \mathbf{0}, \sigma^2 \mathbf{I}_D\right)}\right]} \tag{5.4}$$

We know that for a regression model, the conditional distribution $p\left(y \mid \boldsymbol{x}\right)$ captures how well the model fits the data. In the expression for $p\left(y \mid \boldsymbol{x}\right)$ obtained in equation 5.3, we observe that the conditional distribution behaves like a "weighted sum" of normally distributions in the response variable y centered around the training output points y_n. The "weights" are Gaussians in the feature vector $\boldsymbol{x}$ centered around the training input vectors $\boldsymbol{x}_n$, normalised by the sum of all the Gaussians. Thus, if $\boldsymbol{x}$ is close to $\boldsymbol{x}_n$ for any $n \in \{1, 2, 3, \cdots N\}$, then $p\left(y \mid \boldsymbol{x}\right)$ will be high if y is close to the corresponding y_n, which makes intuitive sense.

The same intuition works for $\mathbb{E}\left[y \mid \boldsymbol{x}\right]$ as the the expected value of y given $\boldsymbol{x}$ will be close to one of the y_n's if $\boldsymbol{x}$ is close to the corresponding $\boldsymbol{x}_n$.

Probabilistic Machine Learning (CS772A), Spring 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION

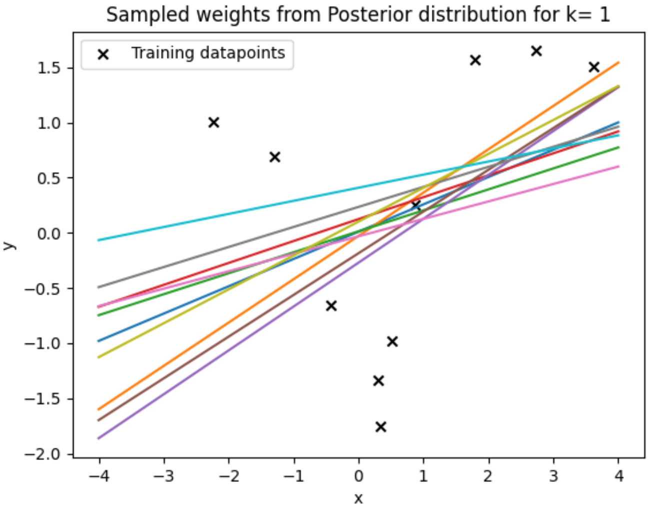
6

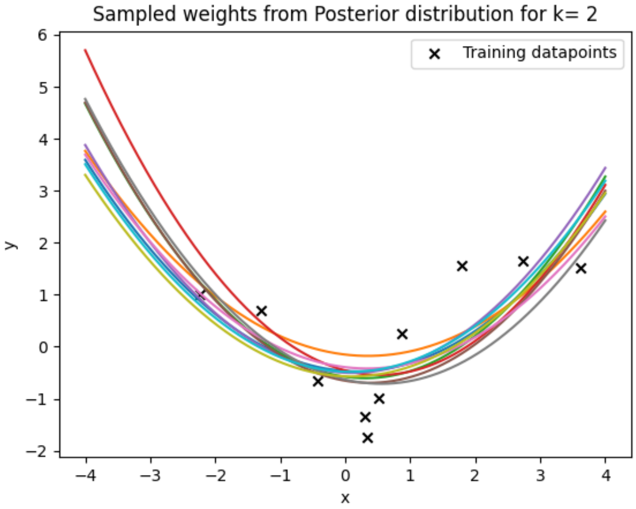
Student Name: Atreya Goswami
Roll Number: 190201
Date: February 7, 2023

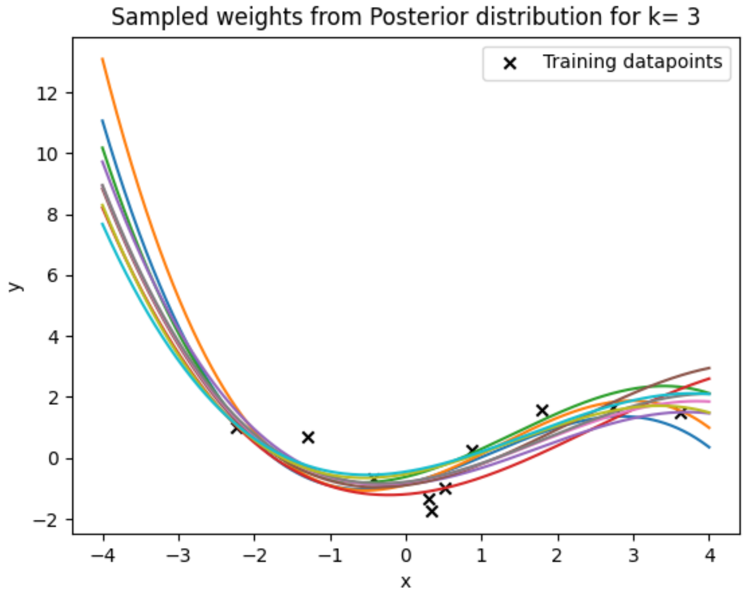
1. Posterior of w

$\mathbf{X}, \boldsymbol{y} = \{\phi(x_n)^T, y_n\}_{n=1}^N$

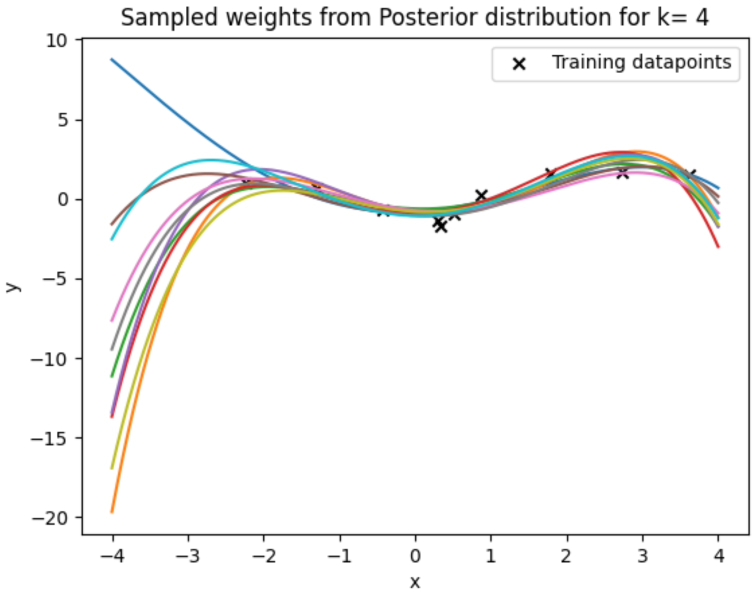
Likelihood: $p\left(\boldsymbol{y}|\mathbf{X}, \boldsymbol{w}\right) = \mathcal{N}\left(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}_N\right)$, Prior: $p\left(\boldsymbol{w}|\lambda\right) = \mathcal{N}\left(\boldsymbol{w}|\mathbf{0}, \lambda^{-1}\mathbf{I}_D\right)$
Posterior: $p\left(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \beta, \lambda\right) = \mathcal{N}\left(\boldsymbol{w}|\mu_N, \Sigma_N\right)$, where
$\Sigma_N = \left(\lambda\mathbf{I}_D + \beta\mathbf{X}^T\mathbf{X}\right)^{-1}$, $\mu_N = \beta\Sigma_N\mathbf{X}^T\boldsymbol{y}$







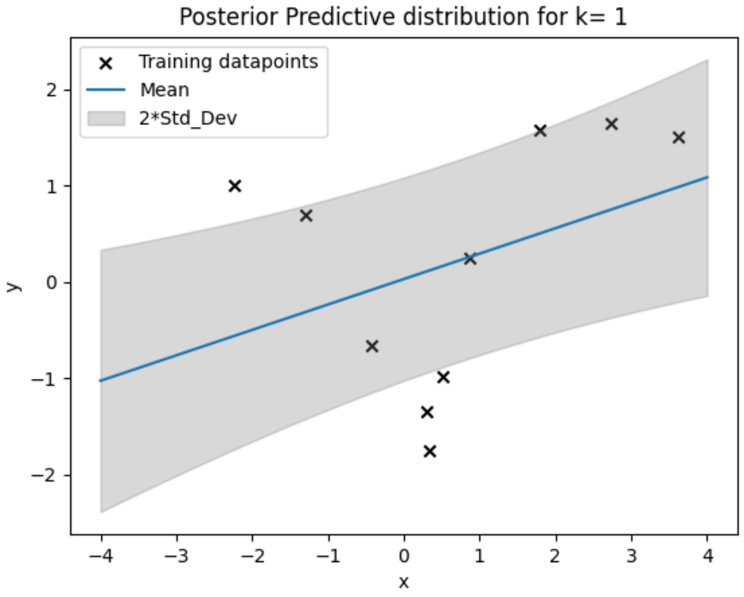
Sampled weights from Posterior distribution for k= 3

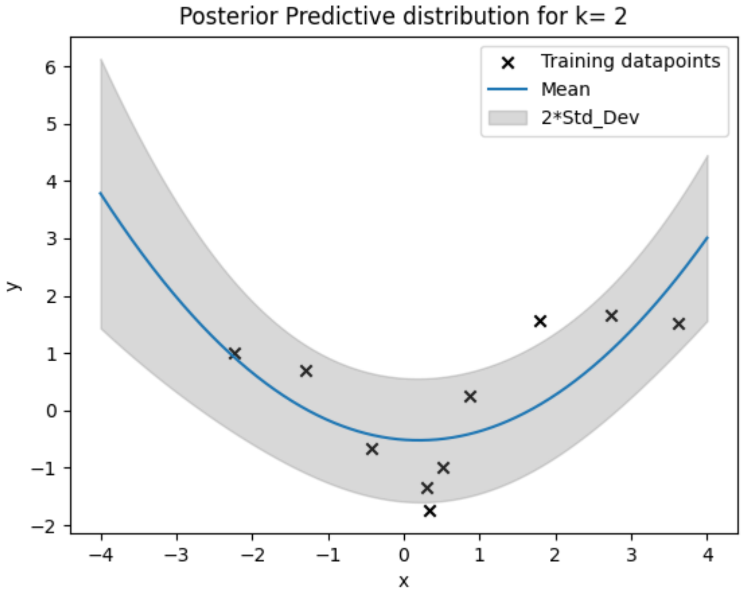


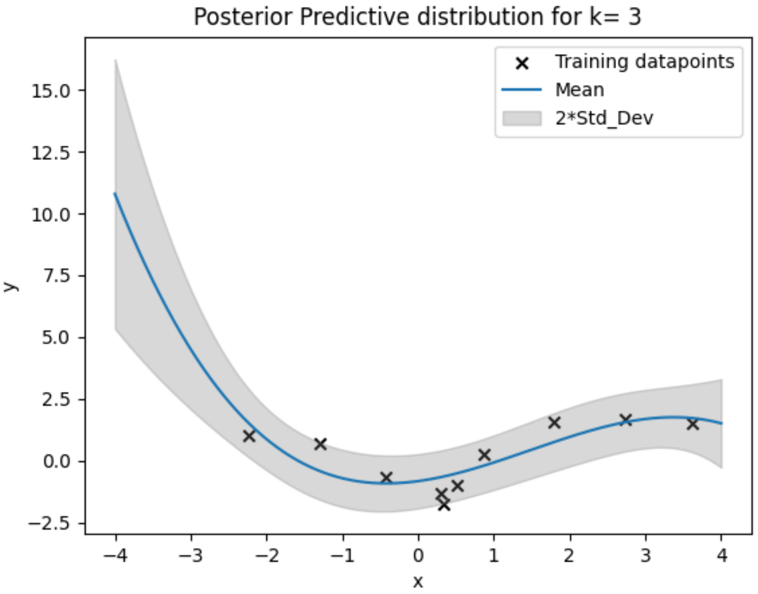
Sampled weights from Posterior distribution for k= 4

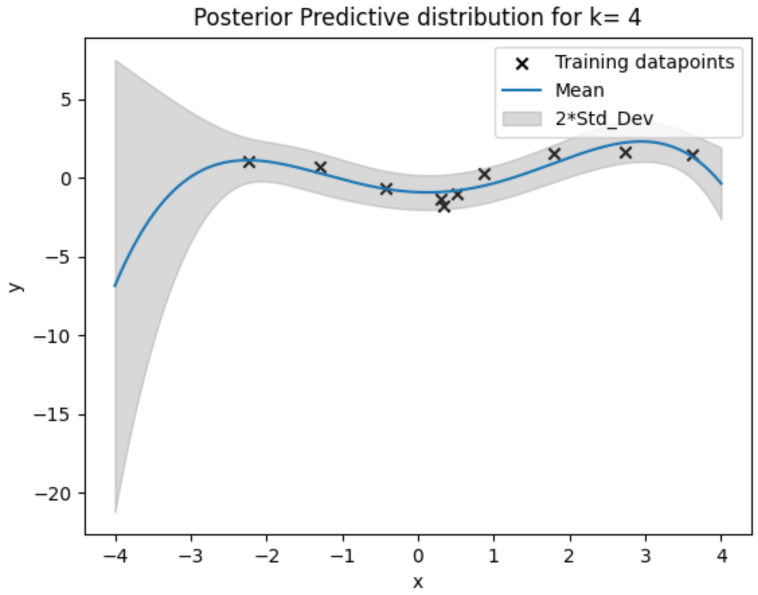
2. Posterior predictive of w

$p\left(y_*|\phi(\boldsymbol{x}_*), \mathbf{X}, \boldsymbol{y}\right) = \mathcal{N}\left(\mu_N^T\phi(\boldsymbol{x}_*), \beta^{-1} + \phi(\boldsymbol{x}_*)^T\Sigma_N\phi(\boldsymbol{x}_*)\right)$ on $x_* \in [-4, 4]$









3. Log Marginal Likelihood

Log Marginal Likelihood $\log p\left(\boldsymbol{y}|\mathbf{X},\beta,\lambda\right) = \log\mathcal{N}\left(\boldsymbol{y}|\mathbf{0},\beta^{-1}\mathbf{I}_N + \lambda^{-1}\mathbf{X}\mathbf{X}^T\right)$

Log Marginal Likelihood with $k = 1$: -32.352015280445244
Log Marginal Likelihood with $k = 2$: -22.77215317878222
Log Marginal Likelihood with $k = 3$: -22.07907064224274
Log Marginal Likelihood with $k = 4$: -22.386776180355803

Thus, the model with $k=3$ best fits the model by using log marginal likelihood comparison

4. Log Likelihood using w_{MAP}

Log likelihood using MAP estimate $\log p\left(\boldsymbol{y} \mid \boldsymbol{w}_{\mathrm{MAP}}, \mathbf{X}, \beta\right)=\log \mathcal{N}\left(\boldsymbol{y} \mid \mathbf{X} \boldsymbol{w}_{\mathrm{MAP}}, \beta^{-1} \mathbf{I}_{N}\right)$
$w_{\mathrm{MAP}}=\mu_{N}$ of the posterior distribution derived above

Log Likelihood with $k=1$: -28.094004379075553
Log Likelihood with $k=2$: -15.360663659052214
Log Likelihood with $k=3$: -10.935846883615739
Log Likelihood with $k=4$: -7.225291259028582

Thus, the model with $k=3$ best fits the model by using log likelihood comparison (by w_{MAP}).

Generally the marginal likelihood is a better way to compare model performance as it marginalises the likelihood over the weights as opposed to comparing likelihoods for a single estimate (MAP) for weights. But in our given case, the marginal likelihoods come out to be very close for k = 2,3,4 which makes it difficult to choose a best model (probably due to the small training data size). On the other hand, the log likelihood computed using MAP estimate of weights comes out to be different significantly for different values of k.

Thus in this case, log likelihood using MAP estimate of weights is more reasonable in my opinion.

5. From the plots, it is evident that there is no training data at all in [-4,-3], which is indicated by the posterior predictive variance in these regions. So more training data (x', y') should be chosen in [-4,-3] which will help in better model fitting by affecting the posterior, which in turn will help in better predictions in this region.

References

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[Mur22] Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022. URL: http://probml.github.io/book1.