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

Student Name: Nidhi Hegde
Roll Number: 180472
Date: February 27, 2021

QUESTION 1

$$\textit{Given}: \qquad p(x|\eta) = \frac{1}{\sqrt{2\pi\eta}} e^{-\frac{x^2}{2\eta}}, \qquad p(\eta|\gamma) = \frac{\gamma^2}{2} e^{-\frac{\gamma^2}{2}\eta}$$

$$\implies p(x|\gamma) = \int p(x|\eta).p(\eta|\gamma) d\eta = \int \frac{\gamma^2}{2\sqrt{2\pi\eta}} e^{-\left(\frac{x^2}{2\eta}+\frac{\gamma^2}{2}\eta\right)} d\eta$$

The above integral computes the marginal distribution of x over the range of η, i.e. 0 to ∞. Since the above integral is hard to compute, we instead compute the **Moment Generating Function** (MGF) of this marginal distribution and then compare it with the MGF of some standard distributions, which will help us to find the distribution of $p(x|\gamma)$.

Moment generating function of $p(x|\gamma)$:

$$M_x(t) = E_{p(x|\gamma)}[e^{tx}] = \int_{-\infty}^{\infty} \left(\int_0^{\infty} \frac{\gamma^2}{2\sqrt{2\pi\eta}}.e^{-\left(\frac{x^2}{2\eta}+\frac{\gamma^2}{2}\eta\right)} d\eta \right) e^{tx} dx$$

We use the following steps to compute this integral:

- After rearranging the terms, we can integrate out x, using completing the squares trick.
- Using the result $\int_{-\infty}^{\infty} e^{-ax^2} = \sqrt{\frac{\pi}{a}}$, the inner integral reduces to $\sqrt{2\pi\eta}$.
- Finally, the outer integral takes the form $\int_0^{\infty} e^{b\eta}$. Solving this proper integral under the assumption that $|t| < \gamma$, we get a closed form solution as:

$$\int_0^{\infty} \frac{\gamma^2}{2\sqrt{2\pi\eta}} \left(\int_{-\infty}^{\infty} e^{\frac{(x-t\eta)^2}{2\eta}} dx \right) e^{\frac{(t^2-\gamma^2)\eta}{2}} d\eta$$

$$= \int_0^{\infty} \frac{\gamma^2}{2\cancel{\sqrt{2\pi\eta}}} \left(\cancel{\sqrt{2\pi\eta}} \right) e^{\frac{(t^2-\gamma^2)\eta}{2}} d\eta = \int_0^{\infty} \frac{\gamma^2}{2} e^{\frac{(t^2-\gamma^2)\eta}{2}} d\eta = \frac{\gamma^2}{(\gamma^2 - t^2)} = \frac{1}{(1-\frac{t^2}{\gamma^2})}$$

The form above matches the MGF of Laplace distribution $L(\mu, b)$ with parameters $\mu = 0$ and $b = \frac{1}{\gamma}$.
Thus,

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

The marginal distribution of a gaussian likelihood, with an exponential prior thus, has the form of a Laplace distribution centered at 0, with a scale parameter $\frac{1}{\gamma}$.
The marginal distribution shows the marginal probability of data obtained after marginalizing over all possible parameters η. In essence, it represents the probability of data after accumulating all the possibilities of η.

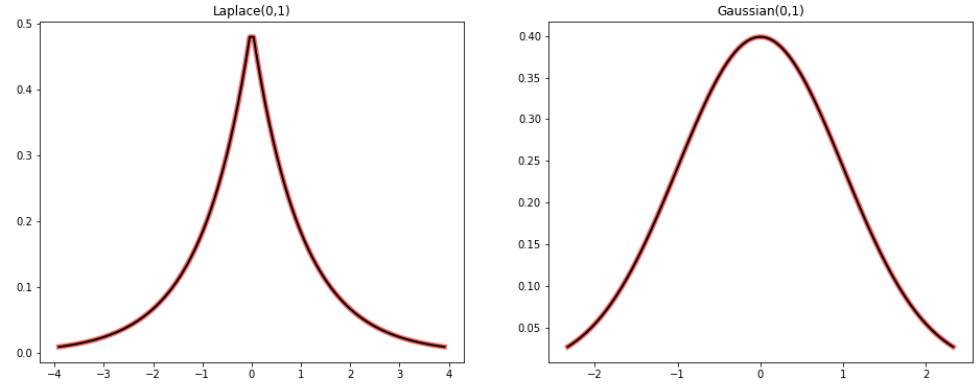


Figure 1: Plots of distributions $p(x|\gamma)\ and\ p(x|\eta)$ with $\gamma = \eta = 1$

On plotting both the curves, we see that the shape of the Laplacian distribution is sharply peaked at its mean, and goes down on either sides at a faster rate, while the gaussian distribution has a smoother peak and gradually goes down as we move away from the mean.

This is because the prior distribution, being an exponential distribution favours η to be near 0. As a result, gaussians with lesser variance contribute more to the integral, resulting in a sharper peak at the mean in the resulting distribution marginalized over η.

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

Student Name: Nidhi Hegde
Roll Number: 180472
Date: February 27, 2021

QUESTION 2

The variance of predictive posterior is given by:

$$\sigma_n^2(x_*) = \beta^{-1} + x_*^T \Sigma_N x_* \quad where \qquad \Sigma_N = (\beta \Sigma_{n=1}^N x_n x_n^T + \lambda I)^{-1}$$

Thus,

$$\sigma_n^2(x_*) = \beta^{-1} + x_*^T (\beta \Sigma_{n=1}^N x_n x_n^T + \lambda I)^{-1} x_* = \beta^{-1} + \beta^{-1} \left(x_*^T (\Sigma_{n=1}^N x_n x_n^T + \frac{\lambda}{\beta} I)^{-1} x_* \right)$$

Similarly,

$$\sigma_{n+1}^2(x_*) = \beta^{-1} + x_*^T (\beta \Sigma_{n=1}^{N+1} x_n x_n^T + \lambda I)^{-1} x_* = \beta^{-1} + \beta^{-1} \left(x_*^T (\Sigma_{n=1}^{N+1} x_n x_n^T + \frac{\lambda}{\beta} I)^{-1} x_* \right)$$

For simplicity, let's assume $\frac{\lambda}{\beta} = k$.
Taking the difference of both the above equations, we get:

$$\sigma_n^2(x_*) - \sigma_{n+1}^2(x_*) = \beta^{-1} x_*^T \left((kI + \Sigma_{n=1}^N x_n x_n^T)^{-1} - ((kI + \Sigma_{n=1}^N x_n x_n^T) + x_{N+1} x_{N+1}^T)^{-1} \right) x_* \quad (2)$$

Let us consider the matrix formula given in the question,
with $M = (kI + \Sigma_{n=1}^N x_n x_n^T)$ and $v = x_{N+1}$.
Thus, [2] reduces to

$$\sigma_n^2(x_*) - \sigma_{n+1}^2(x_*) = \beta^{-1} x_*^T \left(\cancel{M^{-1}} - \left[\cancel{M^{-1}} - \frac{(M^{-1} x_{N+1})(x_{N+1}^T M^{-1})}{1 + x_{N+1}^T M^{-1} x_{N+1}} \right] \right) x_*$$

$$\implies \sigma_n^2(x_*) - \sigma_{n+1}^2(x_*) = \beta^{-1} x_*^T \left(\frac{(M^{-1} x_{N+1})(x_{N+1}^T M^{-1})}{1 + x_{N+1}^T M^{-1} x_{N+1}} \right) x_* \quad (3)$$

M is a symmetric PSD matrix, since

$$\begin{aligned} M^T &= (kI + \Sigma_{n=1}^N x_n x_n^T)^T = (kI^T + \Sigma_{n=1}^N (x_n x_n^T)^T) = (kI + \Sigma_{n=1}^N x_n x_n^T) = M \\ and \quad a^T M a &= k a^T I a + \Sigma_{n=1}^N a^T x_n x_n^T a = k a^T a + \Sigma_{n=1}^N (x_n^T a)^T (x_n^T a) \geq 0 \qquad \forall a \end{aligned}$$

Thus, M^{-1} is also symmetric (since $M^{(-1)^T} = M^{T^{(-1)}}$) [**Property 1**], as well as P.S.D (since all the eigenvalues are reciprocal of that of M, and therefore positive as well). [**Property 2**]

$$\begin{aligned} M^{-1} \; is \; P.S.D. \quad &\implies \quad x^T M^{-1} x \geq 0 \;\; \forall x \;\in\; \{\mathbb{R} - 0\} \\ &\therefore \;\; x_{N+1}^T M^{-1} x_{N+1} \geq 0 \\ &\implies \;\; 1 + x_{N+1}^T M^{-1} x_{N+1} \geq 1 \end{aligned}$$

We have now estabilished that the denominator of [3] is positive.
Thus, in order to check the sign of [3] , we just need to infer the sign of the numerator, i.e. $\beta^{-1}x_*^T\left((M^{-1}x_{N+1})(x_{N+1}^T M^{-1})\right)x_*$
Ignoring β^{-1} since it is always positive, we can re-write the remaining numerator using [**Property 1**] as:

$$\begin{gathered} x_*^T(M^{-1}x_{N+1})(x_{N+1}^T M^{-1})x_* \\ = (x_{N+1}^T M^{-1}x_*)^T(x_{N+1}^T M^{-1}x_*) \\ = \mathbf{p}^T\mathbf{p} \;\; where \;\; \mathbf{p} = (x_{N+1}^T M^{-1}x_*) \\ \mathbf{p}^T\mathbf{p} > 0 \; \forall \mathbf{p} \in \mathbb{R} - 0 \\ \implies \sigma_n^2(x_*) - \sigma_{n+1}^2(x_*) > 0 \end{gathered}$$

We know that an inner product of a vector with itself is always positive for $\forall \mathbf{p} \in \mathbb{R} - 0$
Therefore, we can say that the variance $\sigma_n^2(x_*)$ is greater than $\sigma_{n+1}^2(x_*)$ (for any training example x_*) or in other words, as the number of training examples increases, the variance of the predictive posterior decreases.

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

Student Name: Nidhi Hegde
Roll Number: 180472
Date: February 27, 2021

QUESTION 3

The empirical mean, expressed as a linear transformation of a $\mathbb{R}^D$ Gaussian variable $\mathbf{x}$ can be written as:
($[1]^N$ denotes a N dimensional vector with all 1's)

$$\bar{x} = a^T\mathbf{x} + b \quad where \;\; \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \;\; a = \frac{1}{N}[1]^N \; and \; b = 0$$

Here x is sampled from a Gaussian distribution with mean $\boldsymbol{\mu} = \mu[1]^N$ and covariance matrix $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$.
Using the results discussed in the lectures, we know that the linear transformation of a gaussian random variable is another gaussian distribution mean and variance given as:

$$E[\bar{x}] = a^T\boldsymbol{\mu} = \frac{\mu}{N}\;([1]^N)^T[1]^N = \mu$$
$$var(\bar{x}) = a^T\boldsymbol{\Sigma}a = \frac{\sigma^2}{N^2}\;([1]^N)^T[1]^N = \frac{\sigma^2}{N}$$
$$\therefore \;\; p(\bar{x}) = \mathcal{N}(\bar{x} \mid \mu, \frac{\sigma^2}{N})$$

We know that the resulting distribution should have the same mean as the mean of the distribution of x, since the expected value of a random variable is the same as the expectation of its mean. Intuitively, if we sample N values from $p(x)$, their mean can be expressed as $\mu + \epsilon$, where ϵ is the gaussian noise. Thus, the resulting distribution will be centered at μ, which is consistent with the obtained result. Further, the variance of the empirical mean distribution reduces by a factor of N, when compared to the variance of $p(x)$. We can intuitively say that sampling N values from $p(x)$ is equivalent to sampling 1 value in $p(\bar{x})$, and thus as we draw more and more samples from $p(x)$, the empirical mean of the drawn samples moves towards the mean of the distrbution $p(\bar{x})$, i.e. as $n \to \infty$, variance of $p(\bar{x})$ tends to 0.

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

Student Name: Nidhi Hegde
Roll Number: 180472
Date: February 27, 2021

QUESTION 4

1. If we treat each school's data as a single observation (i.e. equal to the empirical mean of observations from that school), then using Problem 3, we get the likelihood distribution of data from each school as:

$$p(\bar{x}^{(m)} \mid \mu_m) = \mathcal{N}(\bar{x}^{(m)} \mid \mu_m, \frac{\sigma^2}{N_m})$$

We also have the prior over μ_m as $\mathcal{N}(\mu_m \mid \mu_o, \sigma_o^2)$. This information can be expressed in form of a linear gaussian model as:

$$\bar{x}^{(m)} = \mu_m + \epsilon \quad where \quad \epsilon \sim \mathcal{N}(0, \frac{\sigma^2}{N_m})$$

We know that for a linear gaussian model:

$$x = Az + b + \epsilon,\ where\ \ x \sim \mathcal{N}(\mu, \Lambda^{-1}),\ \epsilon \sim \mathcal{N}(0, L^{-1})$$
$$\implies p(z|x) = \mathcal{N}(z \mid \Sigma[A^T L(x-b) + \Lambda\mu], \Sigma),\ where\ \ \Sigma = (\Lambda + A^T L A)^{-1}$$

Thus, the posterior for class m can be directly obtained using these formulas with $A = I,\ b = 0,\ L^{-1} = \frac{\sigma^2}{N_m},\ \mu = \mu_o,\ \Lambda^{-1} = \sigma_o^2$ as:

$$p(\mu_m \mid \bar{x}^{(m)}) = \mathcal{N}(\mu_m \mid \mu_N^{(m)}, {\sigma_N^2}^{(m)}) \quad where$$
$$\mu_N^{(m)} = \frac{N_m\sigma_o^2}{\sigma^2 + N_m\sigma_o^2}\bar{x}^{(m)} + \frac{\sigma^2}{\sigma^2 + N_m\sigma_o^2}\mu_o \quad , \quad \sigma_N^{2^{(m)}} = \left(\frac{1}{\sigma_o^2} + \frac{N_m}{\sigma^2}\right)^{-1}$$

This expression holds for $\forall m \in [1...M]$.
Now, we just need to show that conditioning on the mean of a class is equivalent to conditioning on all the individual data samples of the class.
Let the observed data for class m be $\mathbf{X}^{(m)} = \{x_1^{(m)}, x_2^{(m)} ... x_{N_m}^{(m)}\}$ and the empirical mean for the class be represented as $\bar{\mathbf{X}}^{(m)}$. Thus, we can simply marginalize $p(\mu_m \mid \bar{x}^{(m)})$ over those values of $\bar{x}^{(m)}$ which are consistent with the mean of observed data, i.e. $\bar{\mathbf{X}}^{(m)}$

$$p(\mu_m \mid \mathbf{X}^{(m)}) = \sum_{\bar{x}^{(m)} = \bar{\mathbf{X}}^{(m)}} p(\mu_m \mid \bar{x}^{(m)}) p(\bar{x}^{(m)} \mid \mathbf{X}^{(m)}) = p(\mu_m \mid \bar{x}^{(m)}) p(\bar{\mathbf{X}}^{(m)} \mid \mathbf{X}^{(m)})$$
$$\implies \quad p(\mu_m \mid \mathbf{X}^{(m)}) \ = \ p(\mu_m \mid \bar{x}^{(m)})$$

This further proves the assumption we make (treating all samples from a class as a single observation equal to its mean).
Thus, the posterior is expressed as:

$$\implies \quad p(\mu_m \mid {x_n^{(m)}}_{n=1}^{N_m}) = \mathcal{N}(\mu_m \mid \mu_N^{(m)}, {\sigma_N^2}^{(m)}) \quad \forall \ \ m$$
$$where \ \ \mu_N^{(m)} = \frac{N_m\sigma_o^2}{\sigma^2 + N_m\sigma_o^2}\bar{x}^{(m)} + \frac{\sigma^2}{\sigma^2 + N_m\sigma_o^2}\mu_o \quad , \quad {\sigma_N^2}^{(m)} = \left(\frac{1}{\sigma_o^2} + \frac{N_m}{\sigma^2}\right)^{-1}$$

2. The total marginal likeihood for the above model can be written as:

$$p(\mathbf{x} \mid \mu, \sigma_o^2, \sigma^2) = \prod_{m=1}^{M} p(x^{(m)} \mid \mu, \sigma_o^2, \sigma^2)$$

This can be directly computed using the result of linear gaussian model as:

$$\bar{x}^{(m)} = \mu_m + \epsilon \quad where \quad \epsilon \sim \mathcal{N}(0, \frac{\sigma^2}{N_m})$$

We know the results for a linear gaussian model:

$$x = Az + b + \epsilon, \; where \;\; x \sim \mathcal{N}(\mu, \Lambda^{-1}), \; \epsilon \sim \mathcal{N}(0, L^{-1})$$

Thus, the marginal distribution here can be directly obtained using these formulas with $A = I, \; b = 0, \; L = \frac{N_m}{\sigma^2}, \; \mu = \mu_o, \; \Lambda = \frac{1}{\sigma_o^2}$ as:

$$p(\mathbf{x} \mid \mu, \sigma_o^2, \sigma^2) = \prod_{m=1}^{M} \mathcal{N}\left(\bar{x}^{(m)} \mid \mu_o, \sigma_o^2 + \frac{\sigma^2}{N_m}\right)$$

Taking negative-log of this marginal likelihood and ignoring the constant terms, we get the form of the objective as:

$$\arg\min_{\mu_o} \sum_{m=1}^{M} (\sigma_o^2 + \frac{\sigma^2}{N_m})^{-1} (\bar{x}^{(m)} - \mu_o)^2$$

Solving this gives the solution as $\mu_o^{opt} = \frac{\sum_{m=1}^{M} \frac{N_m \bar{x}^{(m)}}{\sigma^2 + N_m \sigma_o^2}}{\sum_{m=1}^{M} \frac{N_m}{\sigma^2 + N_m \sigma_o^2}}$

3. Putting MLE-II solution of μ_o obtained from part 2 in the posterior obtained in part 1, we get the posterior distribution for class m as:

$$\implies \quad p(\mu_m \,| {x_n^{(m)}}_{m=1}^{N_m}) = \mathcal{N}(\mu_m \mid \mu_N^{(m)}, {\sigma_N^2}^{(m)}) \quad \forall \;\; m$$

$$where \;\; \mu_N^{(m)} = \frac{N_m \sigma_o^2}{\sigma^2 + N_m \sigma_o^2} \bar{x}^{(m)} + \frac{\sigma^2}{\sigma^2 + N_m \sigma_o^2} \frac{\sum_{m=1}^{M} \frac{N_m \bar{x}^{(m)}}{\sigma^2 + N_m \sigma_o^2}}{\sum_{m=1}^{M} \frac{N_m}{\sigma^2 + N_m \sigma_o^2}} \quad , \quad {\sigma_N^2}^{(m)} = \left(\frac{1}{\sigma_o^2} + \frac{N_m}{\sigma^2}\right)^{-1}$$

We observe that MLE-II estimate also captures the properties of data from other classes in the expression of mean of the posterior of any class m.
Another observation is that MLE-II pushes the mean of the posterior more towards the empirical mean for that class, since the contribution of $\bar{x}^{(m)}$ term increases in the final expression of the posterior's mean, as now we are not using any arbitrary μ_o, but a weighted combination of means of all classes.
This can be seen more clearly when the mean $\mu_N^{(m)}$ for class m can be re-arranged as:

$$\mu_N = \frac{N_m \sigma_o^2}{\sigma^2 + N_m \sigma_o^2} \bar{x}^{(m)} + \frac{N_m \sigma^2}{(\sigma^2 + N_m \sigma_o^2)^2} \frac{\bar{x}^{(m)}}{\left(\sum_{m=1}^{M} \frac{N_m}{\sigma^2 + N_m \sigma_o^2}\right)} + \frac{\sigma^2}{\sigma^2 + N_m \sigma_o^2} \frac{\sum_{m'=1, m' \neq m}^{M} \frac{N'_m \bar{x}^{(m')}}{\sigma^2 + N'_m \sigma_o^2}}{\sum_{m=1}^{M} \frac{N_m}{\sigma^2 + N_m \sigma_o^2}}$$

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

Student Name: Nidhi Hegde
Roll Number: 180472
Date: February 27, 2021

QUESTION 5

The expression for likelihood of the above model is given as:

$$p(y^{(m)}|w_m, \mathbf{X}^{(m)}) = \mathcal{N}(y^{(m)}|\mathbf{X}^{(m)}w_m^T, \beta^{-1}\mathbf{I}_n) \quad \forall m$$

Here y_n denotes an $N_m \times 1$ vector, while w_m denotes a $D \times 1$ vector. The total likelihood can be written as:

$$p(y|\mathbf{X}, \mathbf{w}) = \prod_{m=1}^{m=M} \mathcal{N}(y^{(m)}|\mathbf{X}^{(m)}w_m^T, \beta^{-1}\mathbf{I}_n) \tag{4}$$

Here, $\mathbf{X} = \{X^{(1)}, X^{(2)}, ..., X^{(M)}\}$, $y = \{y^{(1)}, y^{(2)}, ..., y^{(M)}\}$ and $\mathbf{w} = \{w_1, w_2, ..., w_M\}$.
The prior is given as:

$$p(w_m) = \mathcal{N}(w_m|w_0^\top, \lambda^{-1}I_D) \ \ \forall m$$

The above model can be written in terms of a linear Gaussian model as:

$$y^{(m)} = \mathbf{X}^{(m)}w_m^\top + \epsilon \qquad s.t. \qquad w_m^T \sim \mathcal{N}(w_0^\top, \lambda^{-1}I_D), \epsilon \sim \mathcal{N}(0, \beta^{-1}I_{N_m}) \quad \forall m$$

Here, the respective parameters of the gaussian linear model, when compared with the standard form are $\mathbf{A} = \mathbf{X}^{(m)}$, $b = 0$, $\Lambda^{-1} = \lambda^{-1}I_D$ and $L^{-1} = \beta^{-1}I_{N_m}$.

We can obtain the marginal likelihood $p(y_m \ |w_o, \mathbf{X}^{(m)})$ for $m \in [1...M]$, directly using the formula as:

$$p(y_m \ |w_o, \mathbf{X}^{(m)}) = \mathcal{N}(y^{(m)}|\mathbf{X}^{(m)}w_0^\top, \lambda^{-1}\mathbf{X}^{(m)}\mathbf{X}^{(m)^\top} + \beta^{-1}\mathbf{I}_{N_m})$$

The total marginal likelihood can be written as:

$$p(Y|w_0, \mathbf{X}) \ = \ \prod_{m=1}^{M} \mathcal{N}(y^{(m)}|\mathbf{X}^{(m)}w_0^\top, \lambda^{-1}\mathbf{X}^{(m)}\mathbf{X}^{(m)^\top} + \beta^{-1}\mathbf{I}_{N_m}) \tag{5}$$

We can directly write [5] since the data of each of the M schools is i.i.d. with respect to other schools.

Taking log on both sides,

$$\log p(Y|w_0, \mathbf{X}) = \sum_{m=1}^{M} log(\mathcal{N}(y^{(m)}|\mathbf{X}^{(m)}w_0^\top, \lambda^{-1}\mathbf{X}^{(m)}\mathbf{X}^{(m)^\top} + \beta^{-1}I_{N_m}))$$

$$= -\sum_{m=1}^{M} \left(\frac{N_m log(2\pi)}{2} + \frac{log(|\Sigma_m|)}{2} + \frac{(y^{(m)} - \mathbf{X}^{(m)}w_0^\top)^\top \Sigma_m^{-1}(y^{(m)} - \mathbf{X}^{(m)}w_0^\top)}{2} \right)$$

$$where, \qquad \Sigma_m = \lambda^{-1}\mathbf{X}^{(m)}\mathbf{X}^{(m)^\top} + \beta^{-1}I_{N_m}$$

The solution obtained by maximizing the above log likelihood w.r.t. w_0 is the MLE-II estimate of w_0.
The final objective function after simplification, can be written as:

$$\arg\min_{w_0} \sum_{m=1}^{M} (y^{(m)} - \mathbf{X}^{(m)} w_0^\top)^\top \Sigma_m^{-1} (y^{(m)} - \mathbf{X}^{(m)} w_0^\top) \tag{6}$$

The benefit of using this estimate of w_0 over fixing it to some known value, is that it will allow us to capture the properties of the entire data in the distribution of posterior for any class m, since the solution for optimal w_0 will be in terms of $\left[\mathbf{X^{(m)}}, \mathbf{\Sigma_m}\right]_{m=1}^{M}$.

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

Student Name: Nidhi Hegde
Roll Number: 180472
Date: February 27, 2021

QUESTION 6

1. The expression for the posterior, on using direct results of gaussian can be written as:

$$p(\mathbf{w}|\mathbf{\Phi_k}(\mathbf{X}), \mathbf{y}, \beta) = \mathcal{N}(\mathbf{w}|\mu_N, \Sigma_N) \quad where,$$

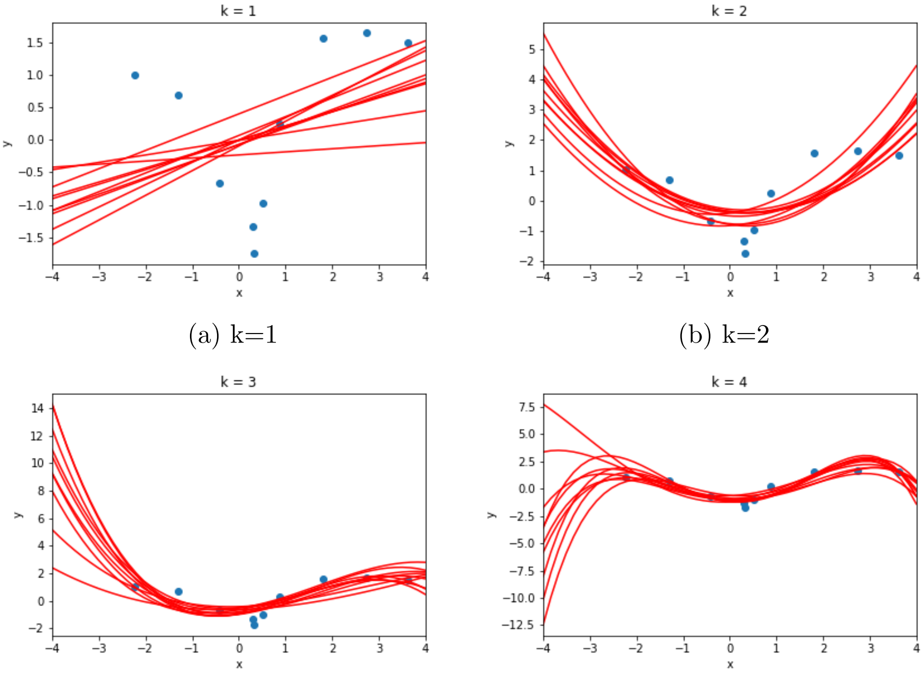
$$\mu_N = \left(\Phi_k(\mathbf{X})\Phi_k(\mathbf{X})^T + \frac{\mathbf{I}}{\beta}\right)^{-1} \Phi_k(\mathbf{X})^T\mathbf{y} \quad and \quad \Sigma_N = \left(\ \beta\Phi_k(\mathbf{X})\Phi_k(\mathbf{X})^T + \mathbf{I}\right)^{-1}$$

Computing μ_N and Σ_N for this case, for different values of k, we get:

$$\mu_N = \left(\mathbf{K} + \frac{\mathbf{I}}{\beta}\right)^{-1} \Phi_k(\mathbf{X})^T\mathbf{y} \quad and \quad \Sigma_N = (\ 4\mathbf{K} + \mathbf{I})^{-1} \quad \forall k$$

$$where, \quad \mathbf{K}_{ij} = \Phi_k(x_i)^T\Phi_k(x_j) \ \ \forall i,j \in [1..N] \ \ and \ \ k \in [1,2,3,4]$$

The plots with random sampling from the posterior distribution for different values of k can be visualized as:



(a) k=1

(b) k=2

(c) k=3

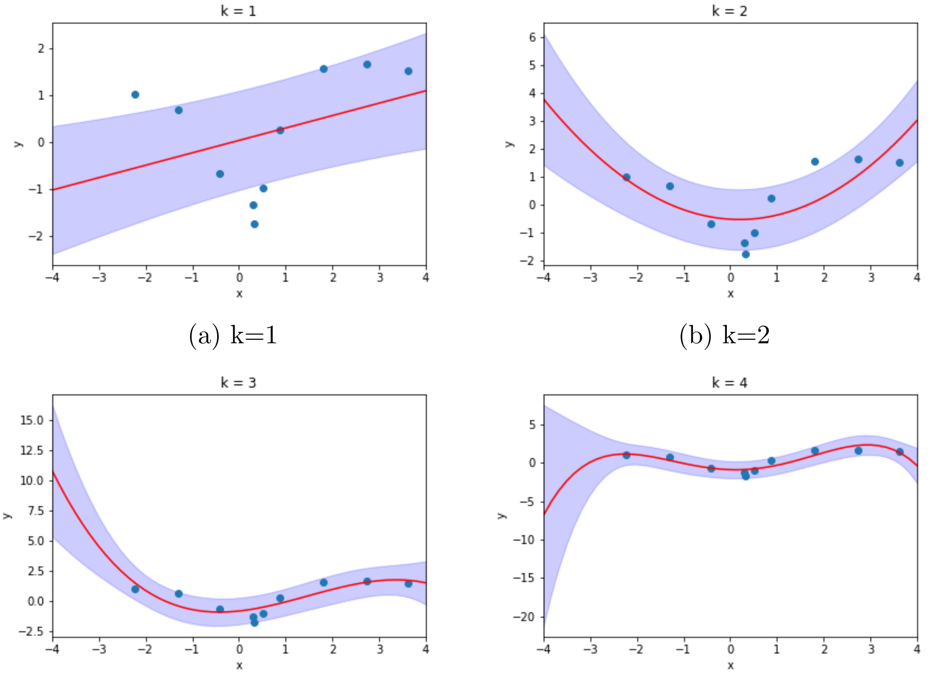
(d) k=4

Figure 2: Weights sampled from posterior(red), training points(blue)

2. The expression for PPD in this case is:

$$p(y_*|\mathbf{\Phi_k}(\mathbf{x_*}), \mathbf{\Phi_k}(\mathbf{X}), \mathbf{y}, \beta) = \mathcal{N}(y_* \mid \mu_N^T x_*, \ \frac{1}{\beta} + x_*^T\Sigma_N x_*)$$

Here μ_N and Σ_N refer to the mean and covariance matrix of the posterior distribution The plots in this case can be visualized as:



(a) k=1

(b) k=2

(c) k=3

(d) k=4

Figure 3: Mean of the posterior predictive(red), ±2 standard deviation (shaded)

3. The marginal likelihood for each k can be written as:

$$p(y|\mathbf{\Phi_k}(\mathbf{X}), \beta) = \mathcal{N}(y \mid 0, \Phi_k(\mathbf{X})\Phi_k(\mathbf{X})^T + \frac{I}{4})$$

Taking log, we get:

$$\log p(y|\mathbf{\Phi_k}(\mathbf{X}), \beta) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}y^T\Sigma^{-1}y$$

On comparing each value of k(individual values shown in code), we get the maximum value for marginal log likeihood for $k = 3$. Thus, model $k = 3$ seems to explain the data the best.

4. We know that w_{MAP} is same as the mean of posterior, i.e, μ_N, and the covariance matrix Σ is given by $\beta^{-1}I$. Thus, we compute the log-likelihood as:

$$\log p(y|\mathbf{\Phi_k}(\mathbf{X}), \mathbf{w_{MAP}}, \beta) = -\frac{N}{2}\log(2\pi) + \frac{N}{2}\log(\beta) - \frac{\beta}{2}(y - \mu_N)^T(y - \mu_N)$$

On comparing each value of k (individual values shown in code), we get the maximum value of log likelihood at MAP solution for $k = 4$.
No, the answer obtained in this case is different from part 3.
The criterion of marginal log likelihood is better for deciding the best model, since the best model will be the one which doesn't perform good on only one value of w, but rather the one that performs consistently over most of the w's. Marginal likelihood takes this into account by marginalizing over all values of w, and hence gives a better estimate.

5. We can see from Figure 3 that the predictive variance is large in the region $[-4, -3]$, since there is no training example in this region. Thus, the model is very uncertain there. Hence, choosing a new training example in the region $[-4, -3]$ may ”improve” the model, by incorporating the context of that region as well.