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

Student Name: Suryateja BV
Roll Number: 160729
Date: February 8, 2019

QUESTION 1

KL divergence is defined as, $KL\left(p||q\right) = \int_{-\infty}^{\infty} p\left(x\right) \log \frac{p(x)}{q(x)} dx$. Let us try with $p\left(x\right) = p_{data}\left(x\right)$ and $q\left(x\right) = p\left(x|\theta\right)$. Then,

$$KL\left(p||q\right) = \int_{-\infty}^{\infty} p_{data}\left(x\right) \log p_{data}\left(x\right) dx - \int_{-\infty}^{\infty} p_{data}\left(x\right) \log p\left(x|\theta\right) dx$$

Minimizing the KL divergence with respect to θ, and ignoring the first term since it doesn't depend on θ, we get,

$$\arg\min_{\theta} KL\left(p||q\right) = \arg\min_{\theta} - \int_{-\infty}^{\infty} p_{data}\left(x\right) \log p\left(x|\theta\right) dx$$

Note that we can approximate the integral (expectation wrt $p_{data}\left(x\right)$) on RHS using Monte Carlo approximation. Let us use the N observations $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$ given. As $N \to \infty$, Monte-Carlo approximation gets almost equal to the integral.

$$\arg\min_{\theta} KL\left(p||q\right) = \arg\max_{\theta} \sum_{n=1}^{N} \log p\left(x_n|\theta\right)$$

The RHS now is precisely the maximum log-likelihood estimate (same as MLE since log is increasing function). Hence, we have shown that minimising KL divergence is equivalent to maximising the MLE.

Note that, had we taken $q\left(x\right) = p_{data}\left(x\right)$ and $p\left(x\right) = p\left(x|\theta\right)$, then $KL\left(p||q\right)$ wouldn't have given us the solution.

$$KL\left(p||q\right) = \int_{-\infty}^{\infty} p\left(x|\theta\right) \log p\left(x|\theta\right) dx - \int_{-\infty}^{\infty} p\left(x|\theta\right) \log p_{data}\left(x\right) dx$$

This would require us to compute the expectation wrt $p\left(x|\theta\right)$ which is actually an unknown distribution. We can't make use of Monte carlo approximation here.

Student Name: Suryateja BV
Roll Number: 160729
Date: February 8, 2019

QUESTION 2

Let $\{X_i\}_{i=1}^{N}$ be N i.i.d Gaussian random variables drawn from $\mathcal{N}\left(\mu, \sigma^2\right)$. Let us take an $N \times 1$ matrix $A = \left[\frac{1}{N}\frac{1}{N}\cdots\frac{1}{N}\right]$ and let X be a random vector $X = \left[X_1 X_2 \cdots X_N\right]^T$. Let $\bar{X} = AX$. Note that $\bar{X}$ is a Gaussian random variable too. Then,

$$\begin{aligned}
\mathbb{E}\left[\bar{X}\right] = A\mathbb{E}\left[X\right] &= \left[\frac{1}{N}\frac{1}{N}\cdots\frac{1}{N}\right] \times \left[\mathbb{E}\left[X_1\right]\mathbb{E}\left[X_2\right]\cdots\mathbb{E}\left[X_N\right]\right]^T \\
&= \left[\frac{1}{N}\frac{1}{N}\cdots\frac{1}{N}\right] \times \left[\mu\mu\cdots\mu\right]^T = \mu
\end{aligned}$$

$$\begin{aligned}
Cov\left(\bar{X}\right) &= ACov\left(X\right)A^T \\
&= \left[\frac{1}{N}\frac{1}{N}\cdots\frac{1}{N}\right] diag\left(\sigma^2, \sigma^2, \cdots, \sigma^2\right) \left[\frac{1}{N}\frac{1}{N}\cdots\frac{1}{N}\right]^T \\
&= \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}
\end{aligned}$$

Note that we get a diagonal matrix because X_i's are independent, ie, $cov\left(X_i, X_i\right) = \sigma^2$ and $cov\left(X_i, X_j\right) = 0$ for i $\neq$ j. Thus, $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$

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

QUESTION 3

Student Name: Suryateja BV
Roll Number: 160729
Date: February 8, 2019

Proof of a simple trick being used in this problem : replacement of set of observations of Gaussian distribution by its empirical mean, assuming variance is known.

$$\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2} = \frac{\sum (x_i-\bar{x}+\bar{x}-\mu)^2}{2\sigma^2} = \frac{N(\bar{x}-\mu)^2}{2\sigma^2} + \frac{S^2}{2\sigma^2}$$

where $\bar{x}$ and S^2 are empirical mean and empirical variance respectively. In the given problem, for the posterior distribution of μ_m, we need to only consider the school m, and we can replace the observations by empirical mean of school m as shown below–

$$\begin{aligned}
p\left(\mu_m|x^{(m)},\mu_0\sigma^2,\sigma_0^2\right) &= \frac{p\left(x^{(m)}|\mu_m,\sigma^2\right)p\left(\mu_m|\mu_0,\sigma_0^2\right)}{\int p\left(x^{(m)}|\mu_m,\sigma^2\right)p\left(\mu_m|\mu_0,\sigma_0^2\right)d\mu_m} \\
&\propto \exp\left(-\frac{\sum_{i=1}^{N_m}\left(x_i^{(m)}-\mu_m\right)^2}{2\sigma^2}\right)\exp\left(-\frac{(\mu_m-\mu_0)^2}{2\sigma_0^2}\right) \\
&\propto \exp\left(-\frac{N\left(\bar{x}^{(m)}-\mu\right)^2}{2\sigma^2}\right)\exp\left(-\frac{(\mu_m-\mu_0)^2}{2\sigma_0^2}\right)
\end{aligned}$$

Using completing the squares trick, we readily get the posterior mean μ_{mp} and variance σ_{mp}^2 as

$$\begin{aligned}
\mu_{mp} &= \frac{\sigma_0}{N_m\sigma_0^2+\sigma^2}\mu_0 + \frac{N_m\sigma_0^2}{N_m\sigma_0^2+\sigma^2}\bar{x}^{(m)} \\
\frac{1}{\sigma_{mp}^2} &= \frac{1}{\sigma_0^2} + \frac{N_m}{\sigma^2}
\end{aligned}$$

Thus, the posterior distribution is $\mu_m \sim \mathcal{N}\left(\mu_{mp},\sigma_{mp}^2\right)$ The marginal likelihood is given by,

$$\begin{aligned}
p\left(x|\mu_0,\sigma^2,\sigma_0^2\right) &= \prod_{m=1}^{M}\int p\left(x^{(m)}|\mu_m,\mu_0,\sigma^2,\sigma_0^2\right)p\left(\mu_m|\mu_0,\sigma_0^2\right)d\mu_m \\
&= \prod_{m=1}^{M}\frac{p\left(x^{(m)}|\mu_m,\sigma^2\right)p\left(\mu_m|\mu_0,\sigma_0^2\right)}{p\left(\mu_m|x^{(m)},\mu_{mp},\sigma_{mp}^2\right)}
\end{aligned}$$

The above is the exact value. We can use the same trick of replacing with $\bar{x}^{(m)}$ to get,

$$p\left(x|\mu_0,\sigma^2,\sigma_0^2\right) = \prod_{m=1}^{M}\mathcal{N}\left(\bar{x}^{(m)},\mu_m,\frac{\sigma^2}{N_m}\right)\mathcal{N}\left(\mu_m|\mu_0,\sigma_0^2\right) = \prod_{m=1}^{M}\mathcal{N}\left(\bar{x}^{(m)}|\mu_0,\sigma_0^2+\sigma^2/N_m\right)$$

We have used $\bar{x}^{(m)} = \mu_m+\epsilon$, and took expectation and variance. $\mathbb{E}\left[\bar{x}^{(m)}\right] = \mu_0$ and $\text{Var}\left[\bar{x}^{(m)}\right] = \sigma_0^2+\sigma^2/N_m$ Computation of MLE-II involves taking maximum wrt μ_0 of the marginal log likelihood. Doing so with the empirical mean replaced version of marginal likelihood, and taking derivative wrt μ_0 we get,

$$\sum_{m=1}^{M} \frac{\bar{x}^{(m)} - \mu_0}{\sigma_0^2 + \sigma^2/N_m} = 0$$
$$\implies \mu_0 = \frac{\sum_{m=1}^{M} \frac{\bar{x}^{(m)}}{\sigma_0^2+\sigma^2/N_m}}{\sum_{m=1}^{M} \frac{1}{\sigma_0^2+\sigma^2/N_m}}$$

The above is the MLE-II estimate of μ_0.

Substituting the obtained MLE-II estimate of μ_0 in μ_{mp}, we get

$$\mu_{mp} = \frac{\sigma_0}{N_m\sigma_0^2 + \sigma^2} \frac{\sum_{m=1}^{M} \frac{\bar{x}^{(m)}}{\sigma_0^2+\sigma^2/N_m}}{\sum_{m=1}^{M} \frac{1}{\sigma_0^2+\sigma^2/N_m}} + \frac{N_m\sigma_0^2}{N_m\sigma_0^2 + \sigma^2}\bar{x}^{(m)}$$

This increases the probability of marginal likelihood of X given the hyperparameters. There is no change in the form of the solution. The posterior is still a normal distribution with a different mean. Instead of taking any random μ_0, we took a specific μ_0.

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

QUESTION 4

Student Name: Suryateja BV
Roll Number: 160729
Date: February 8, 2019

We have $p\left(Z|\alpha\right) = \int p\left(Z|\pi,\alpha\right)p\left(\pi|\alpha\right)d\pi$. Since the entries are generated independently, we have

$$\begin{aligned}
p\left(Z|\alpha\right) &= \int\int\cdots\int_K \prod_{n=1}^{N}\left(\prod_{k=1}^{K} p\left(Z_{nk}|\pi_k,\alpha\right)\right)\prod_{k'=1}^{K} p\left(\pi_{k'}|\alpha\right)d\pi_1 d\pi_2\cdots d\pi_K \\
&= \prod_{k=1}^{K}\int\prod_{n=1}^{N}\left(\pi_k\right)^{z_{nk}}\left(1-\pi_k\right)^{1-z_{nk}}p\left(\pi_k|\alpha\right)d\pi_k \\
&= \prod_{k=1}^{K}\int\left(\pi_k\right)^{\sum z_{nk}}\left(1-\pi_k\right)^{N-\sum z_{nk}}\frac{\pi_k^{\frac{\alpha}{K}-1}}{B\left(\frac{\alpha}{K},1\right)} \\
&= \prod_{k=1}^{K}\frac{B\left(\sum_{n=1}^{N} z_{nk}+\frac{\alpha}{K}, N+1-\sum_{n=1}^{N} z_{nk}\right)}{B\left(\frac{\alpha}{K},1\right)}
\end{aligned}$$

where $B\left(a,b\right) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. So, we see that it can be written in the form of product of ratios of beta functions.

For the next part, note that $p\left(Z_{nk}=1|Z_{-nk}\right) = \int_0^1 \pi_k p\left(\pi_k|Z_{-nk}\right)d\pi_k = \mathbb{E}\left[\pi_k\right]$ wrt the posterior distribution of $p\left(\pi_k|Z_{-nk}\right)$

$$p\left(\pi_k|Z_{-nk}\right) = \frac{p\left(Z_{-nk}|\pi_k\right)p\left(\pi_k\right)}{\int_0^1 p\left(Z_{-nk}|\pi_k\right)p\left(\pi_k\right)d\pi_k}$$

The posterior distribution can be easily obtained by seeing the column as tossing a coin N-1 times, getting $\sum_{i=1,i\neq n}^{N} z_{ik} = t_{-nk}$ heads. We readily get the posterior distribution as,

$$\begin{aligned}
p\left(\pi_k|Z_{-nk}\right) &= \text{Beta}\left(\frac{\alpha}{K}+t_{-nk}, N-t_{-nk}\right) \\
\implies \mathbb{E}\left[\pi_k\right] &= \frac{\frac{\alpha}{K}+t_{-nk}}{\frac{\alpha}{K}+N}
\end{aligned}$$

This form of result makes intuitive sense. Before observing Z_{-nk}, $p\left(Z_{nk}=1\right) = \mathbb{E}\left[\pi_k\right]$ wrt $p\left(\pi_k\right)$, the prior distribution, which is $\frac{\frac{\alpha}{K}}{\frac{\alpha}{K}+1}$. This is supported by $\frac{\alpha}{K}+1$ datapoints. Considering only the observations Z_{nk}, the value is $\frac{t_{-nk}}{N-1}$ which is supported by N-1 datapoints. So,

$$\mathbb{E}\left[\pi_k\right] = \frac{\left(\frac{\alpha}{K}+1\right)\left(\frac{\frac{\alpha}{K}}{\frac{\alpha}{K}+1}\right)+\left(N-1\right)\left(\frac{t_{-nk}}{N-1}\right)}{\frac{\alpha}{K}+1+N-1}$$

It is like a weighted average of our prior and posterior beliefs. Also note that, as $K\to\infty$, the value becomes $\frac{t_{-nk}}{N}$. This is like replacing a missing value in a column with the column average.

Finally, to get the expected number of 1s in a column and in the entire matrix,

$$\mathbb{E}\left[z_{nk}\right] = 1 \times p\left(z_{nk}=1|\alpha\right) + 0 \times p\left(z_{nk}=0|\alpha\right) = \int_0^1 p\left(z_{nk}=1|\pi_k,\alpha\right) p\left(\pi_k|\alpha\right) d\pi_k$$

$$= \mathbb{E}\left[\pi_k\right] = \frac{\frac{\alpha}{K}}{\frac{\alpha}{K}+1}$$

By linearity of expectations, (and given that z_{nk}'s independently generated), we get the expected number of ones in a column as $N\mathbb{E}\left[z_{nk}\right]$ and number of ones in the entire matrix as $NK\mathbb{E}\left[z_{nk}\right]$. Thus, number of 1s in a column is $\frac{\frac{N\alpha}{K}}{\frac{\alpha}{K}+1}$ and number of 1s in the entire matrix is $\frac{N\alpha}{\frac{\alpha}{K}+1}$.

The marginal prior on w, after integrating out b is given by,

$$
\begin{aligned}
p\left(w|\sigma^2_{spike}, \sigma^2_{slab}\right) &= p\left(w|b=1, \sigma^2_{spike}, \sigma^2_{slab}\right) \times p\left(b=1\right) + p\left(w|b=0, \sigma^2_{spike}, \sigma^2_{slab}\right) \times p\left(b=0\right) \\
&= \frac{1}{2}\mathcal{N}\left(w|0, \sigma^2_{slab}\right) + \frac{1}{2}\mathcal{N}\left(w|0, \sigma^2_{spike}\right)
\end{aligned}
$$

The plot of marginal prior with $\left(\sigma^2_{spike}, \sigma^2_{slab}\right) = (1, 100)$ is shown in Figure 1. Comparing with $\mathcal{N}(0,1)$ distribution, we see that the marginal prior is a bit fat tailed, and the probability that a sample will be close to mean 0 is also less. This marginal prior doesn't force w to take the value of 0 as aggresively as a standard Normal distribution.

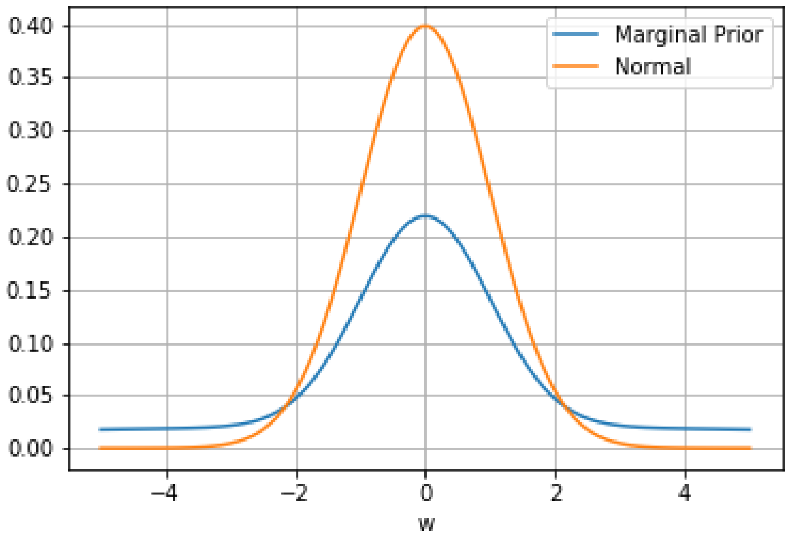


Figure 1: Marginal prior on w

$$
\begin{aligned}
p\left(b=1|w, \sigma^2_{spike}, \sigma^2_{slab}\right) &= \frac{p\left(w|b=1, \sigma^2_{spike}, \sigma^2_{slab}\right) \times p\left(b=1\right)}{p\left(w|b=1, \sigma^2_{spike}, \sigma^2_{slab}\right) \times p\left(b=1\right) + p\left(w|b=0, \sigma^2_{spike}, \sigma^2_{slab}\right) \times p\left(b=0\right)} \\
&= \frac{\mathcal{N}\left(w|0, \sigma^2_{slab}\right)}{\mathcal{N}\left(w|0, \sigma^2_{slab}\right) + \mathcal{N}\left(w|0, \sigma^2_{spike}\right)}
\end{aligned}
$$

$$
\begin{aligned}
p\left(b=1|x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2\right) &= \int p\left(b=1|w, x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2\right) p\left(w|x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2\right) dw \\
&= \int p\left(b=1|w, \sigma^2_{spike}, \sigma^2_{slab}\right) p\left(w|x, \sigma^2_{spike}, \sigma^2_{slab}, \rho^2\right) dw
\end{aligned}
$$

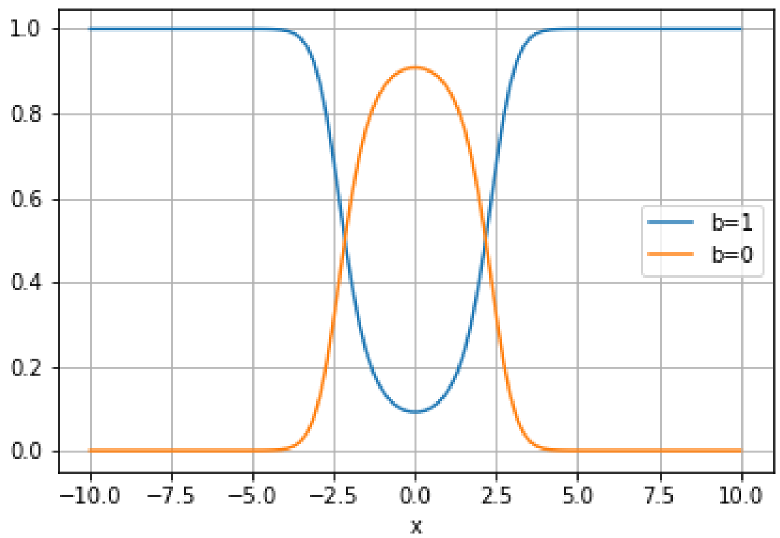
The above simplification stems from the fact that x in turn depends on w.

$$
\begin{aligned}
p\left(w|x,\sigma^2_{spike},\sigma^2_{slab},\rho^2\right) &= \frac{p\left(x|w,\sigma^2_{spike},\sigma^2_{slab},\rho^2\right)p\left(w|\sigma^2_{spike},\sigma^2_{slab},\rho^2\right)}{\int p\left(x|w,\sigma^2_{spike},\sigma^2_{slab},\rho^2\right)p\left(w|\sigma^2_{spike},\sigma^2_{slab},\rho^2\right)dw}\\
&= \frac{\mathcal{N}\left(x|w,\rho^2\right)\left(\mathcal{N}\left(w|0,\sigma^2_{slab}\right)+\mathcal{N}\left(w|0,\sigma^2_{spike}\right)\right)}{\int \mathcal{N}\left(x|w,\rho^2\right)\left(\mathcal{N}\left(w|0,\sigma^2_{slab}\right)+\mathcal{N}\left(w|0,\sigma^2_{spike}\right)\right)dw}\\
&= \frac{\mathcal{N}\left(x|w,\rho^2\right)\left(\mathcal{N}\left(w|0,\sigma^2_{slab}\right)+\mathcal{N}\left(w|0,\sigma^2_{spike}\right)\right)}{\mathcal{N}\left(x|0,\rho^2+\sigma^2_{slab}\right)+\mathcal{N}\left(x|0,\rho^2+\sigma^2_{spike}\right)}
\end{aligned}
$$

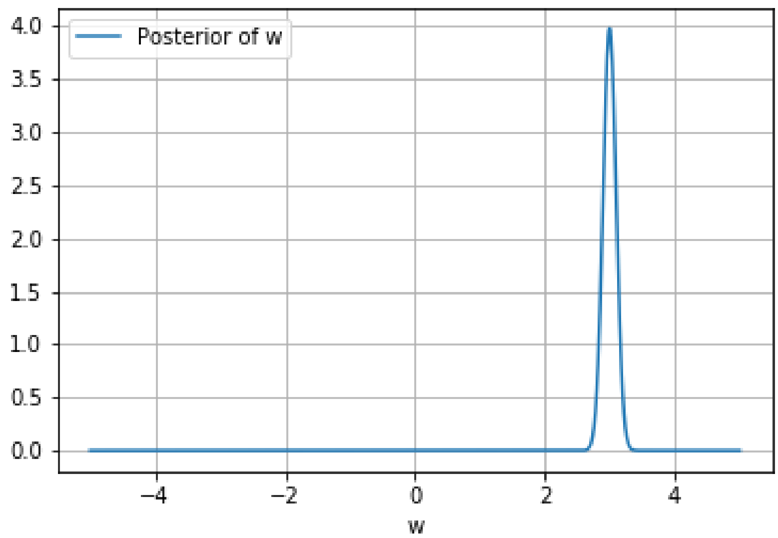
The denominator computation is easy, because $x = w + \epsilon \implies \mathbb{E}[x] = \mathbb{E}[w]$ and $\text{Var}[x] = \text{Var}[w] + \rho^2$. When we take $w \sim \mathcal{N}\left(w|0,\sigma^2_{slab}\right)$, we get $x \sim \mathcal{N}\left(x|0,\rho^2+\sigma^2_{slab}\right)$. Similarly, we get $x \sim \mathcal{N}\left(x|0,\rho^2+\sigma^2_{spike}\right)$ for $w \sim \mathcal{N}\left(w|0,\sigma^2_{spike}\right)$. Note that x is sum of two independent Gaussian random variables and hence Gaussian. So, we have

$$
\begin{aligned}
&p\left(b=1|x,\sigma^2_{spike},\sigma^2_{slab},\rho^2\right)\\
&= \int \frac{\mathcal{N}\left(w|0,\sigma^2_{slab}\right)}{\mathcal{N}\left(w|0,\sigma^2_{slab}\right)+\mathcal{N}\left(w|0,\sigma^2_{spike}\right)} \times \frac{\mathcal{N}\left(x|w,\rho^2\right)\left(\mathcal{N}\left(w|0,\sigma^2_{slab}\right)+\mathcal{N}\left(w|0,\sigma^2_{spike}\right)\right)}{\mathcal{N}\left(x|0,\rho^2+\sigma^2_{slab}\right)+\mathcal{N}\left(x|0,\rho^2+\sigma^2_{spike}\right)}dw\\
&= \frac{\mathcal{N}\left(x|0,\rho^2+\sigma^2_{slab}\right)}{\mathcal{N}\left(x|0,\rho^2+\sigma^2_{slab}\right)+\mathcal{N}\left(x|0,\rho^2+\sigma^2_{spike}\right)}
\end{aligned}
$$

The following plot shows $p\left(b=1|x,\sigma^2_{spike},\sigma^2_{slab},\rho^2\right)$ for $\left(\sigma^2_{spike},\sigma^2_{slab},\rho^2\right) = (1, 100, 0.01)$. Graph of $\left(b=0|x,\sigma^2_{spike},\sigma^2_{slab},\rho^2\right)$ is shown for comparison. If a value of x close to 0 is observed, this means that there is a very high chance of b being 0, ie, an irrelevant feature. If a value of x greater than 4 is observed, then it is almost certain that b is 1, ie, a relevant feature.



The plot of $p\left(w|x, \sigma_{spike}^2, \sigma_{slab}^2, \rho^2\right)$ is shown next. The value of x is taken to be 3, and other hyperparameters are same as before. As is evident from the figure, we see that there is a huge spike at w = 3. This is obvious because the noisy observation was at x=3, so w is also expected to be very close to 3.

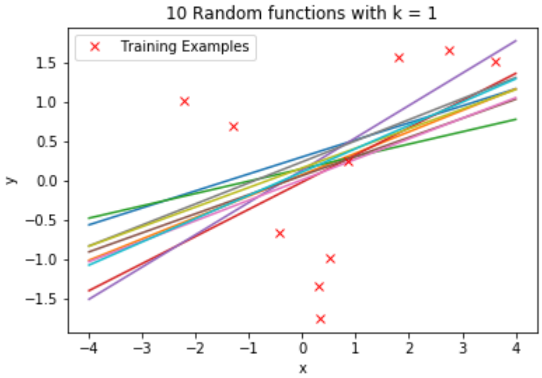


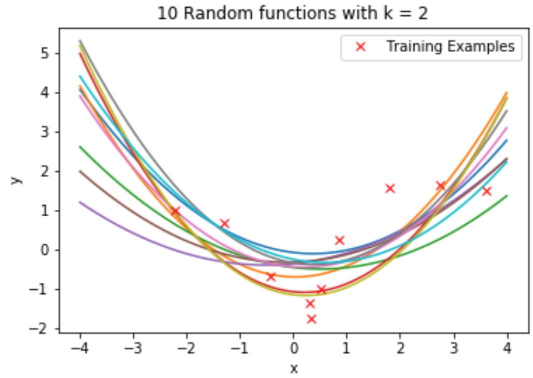
Topics in Probabilistic Modeling & Inference (CS698X), Spring 2019
Indian Institute of Technology Kanpur
Homework Assignment Number 1

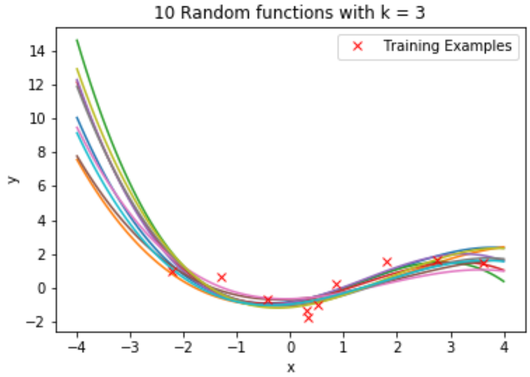
Student Name: Suryateja BV
Roll Number: 160729
Date: February 8, 2019

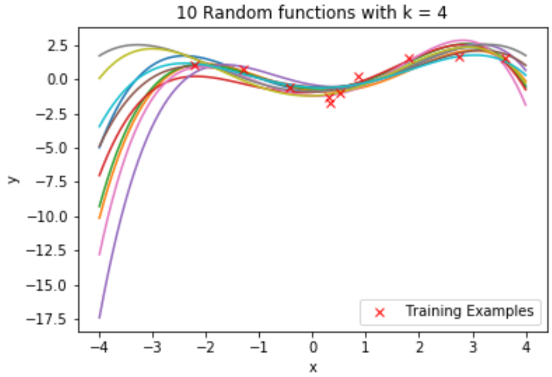
QUESTION 6

The plot of 10 random functions inferred from the posterior are shown below. k = 1 clearly doesn't fit the data that well. We can only judge which model fits well looking at the marginal log likelihood (model comparison). The model with k = 3 seems to explain data the best since it has the highest marginal log likelihood value.









k	Marginal Log likelihood
1	-32.35
2	-22.78
3	-22.08
4	-22.39

Table 1: Marginal log likelihood for various values of k

The log likelihood values taking w_{map} are shown below. From the table, model with k = 4 has the highest log likelihood value. This is different from the marginal log likelihood case where we found model with k = 3 to be the best. We should consider marginal log likelihood over log likelihood to be the model indicator because we are considering the uncertainty in w in marginal LL computation. We integrate over all possible values of w to get marginal LL. However, for w_{map} LL, we ignore the uncertainty in w by taking a specific value of w = w_{map}. We can't put our faith in a model by looking at just a single value of w.

k	w_{map} **Log likelihood**
1	-28.09
2	-15.36
3	-10.9
4	-7.23

Table 2: Marginal log likelihood for various values of k

The plots of posterior predictive mean along with +/- 2 times the standard deviation is shown next. Since k = 3 is our best model, from the figure it is obvious that we'd want an (x, y) pair in the region of [-4, -2.5] to improve our learned model, as the variance / standard deviation is pretty high in this region. More specifically, the standard deviation is maximum at x = -4, so I'd prefer to get the y value for x = -4.