Question 1

$$p(\theta|\mathbf{X},\lambda,m) = \frac{p(\theta|\lambda,m)p(\mathbf{X}|\theta,\lambda,m)}{p(\mathbf{X}|\lambda,m)} \quad (1)$$

$$p(\lambda|\mathbf{X},m) = \frac{p(\lambda|m)p(\mathbf{X}|\lambda,m)}{p(\mathbf{X}|m)} \quad (2)$$

$$p(m|\mathbf{X}) = \frac{p(m)p(\mathbf{X}|m)}{p(\mathbf{X})} \quad (3)$$

Ranking: (3) > (2) > (1). Note that each of the above contains a quantity which is an integral/summation of a quantity computed by the previous one, i.e., $p(\mathbf{X}|\lambda,m)$ requires integrating out θ from $p(\mathbf{X}|\theta,\lambda,m)$; $p(\mathbf{X}|m)$ requires integrating out λ from $p(\mathbf{X}|\lambda,m)$, and $p(\mathbf{X})$ requires integraing out m from $p(\mathbf{X}|m)$.

The posterior predictive distribution is given by **Question 2**

$$p(y_*|\boldsymbol{x}_*) = \mathcal{N}\left(\boldsymbol{\mu}_N^\intercal \boldsymbol{x}_*, \beta^{-1} + \boldsymbol{x}_*^\intercal \boldsymbol{\Sigma}_N \boldsymbol{x}_*\right), \boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N\left(\sum_{n=1}^{N} y_n \boldsymbol{x}_n\right), \boldsymbol{\Sigma}_N = \left(\beta \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\intercal + \lambda \mathbf{I}\right)^{-1}$$

Consider the difference in the variances when we increase the no. of observations by 1. We have

$$\delta_{N+1,N} = \left(\beta^{-1} + \boldsymbol{x}_*^\intercal \boldsymbol{\Sigma}_{N+1} \boldsymbol{x}_*\right) - \left(\beta^{-1} + \boldsymbol{x}_*^\intercal \boldsymbol{\Sigma}_N \boldsymbol{x}_*\right) = \boldsymbol{x}_*^\intercal (\boldsymbol{\Sigma}_{N+1} - \boldsymbol{\Sigma}_N) \boldsymbol{x}_*$$

Also notice that we can express $\boldsymbol{\Sigma}_{N+1} = (\boldsymbol{\Sigma}_N^{-1} + \beta \boldsymbol{x}_{N+1} \boldsymbol{x}_{N+1}^\intercal)^{-1}$ We use the following identity(called as Shermon Morrison Formula) with $\boldsymbol{v} = \sqrt{\beta}\boldsymbol{x}_{N+1}, \mathbf{M} = \boldsymbol{\Sigma}_N^{-1}$-

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

$$\implies \boldsymbol{\Sigma}_{N+1} = \boldsymbol{\Sigma}_N - \frac{\beta \boldsymbol{\Sigma}_N \boldsymbol{x}_{N+1} \boldsymbol{x}_{N+1}^\intercal \boldsymbol{\Sigma}_N}{1 + \beta \boldsymbol{x}_{N+1}^\intercal \boldsymbol{\Sigma}_N \boldsymbol{x}_{N+1}}$$

$$\implies \delta_{N+1,N} = -\frac{\beta (\boldsymbol{x}_*^\intercal \boldsymbol{\Sigma}_N \boldsymbol{x}_{N+1}) \boldsymbol{x}_{N+1}^\intercal \boldsymbol{\Sigma}_N \boldsymbol{x}_*}{1 + \beta \boldsymbol{x}_{N+1}^\intercal \boldsymbol{\Sigma}_N \boldsymbol{x}_{N+1}}$$

$$\implies \delta_{N+1,N} = -\frac{\beta \phi^\intercal \phi}{1 + \beta \boldsymbol{x}_{N+1}^\intercal \boldsymbol{\Sigma}_N \boldsymbol{x}_{N+1}} \qquad \because \text{define } \phi = \boldsymbol{x}_{N+1}^\intercal \boldsymbol{\Sigma}_N \boldsymbol{x}_*$$

Clearly $\beta \phi^\intercal \phi \geqslant 0$ and since Σ_N is a positive semidefinite matrix we have

$$\boldsymbol{v}^\intercal \boldsymbol{\Sigma}_{\boldsymbol{N}} \boldsymbol{v} \geqslant 0, \forall v \in \mathbb{R}^D \implies 1 + \beta \boldsymbol{x}_{N+1}^\intercal \boldsymbol{\Sigma}_N \boldsymbol{x}_{N+1} > 0$$

Hence we have $\delta_{N+1,N} \leqslant 0$. Thus the variance of the posterior predictive distribution decreases as the number of training examples increases. This is expected as the number of training instances increases our model becomes more certain of its predictions and the uncertainity in parameters decreases.

We use Fubini's theorem to find the moment generating function of the marginalised distribution and compare it with the MGF's of some of the common distributions. The marginalised distribution will turn out to be a **Laplace** distribution. The marginalised distribution

$$p(x|\gamma) = \int_0^\infty p(x|\eta)p(\eta|\gamma)d\eta = \int_0^\infty \frac{1}{\sqrt{2\pi\eta}}\frac{\gamma^2}{2}\exp\left(-\frac{x^2}{2\eta}\right)\exp\left(-\frac{\eta\gamma^2}{2}\right)d\eta$$

The moment generating function of $p(x|\gamma)$ is given by -

$$M(t) = \int_{-\infty}^\infty p(x|\gamma)e^{tx}dx = \frac{\gamma^2}{2\sqrt{2\pi}}\int_{\eta=0}^\infty\int_{x=-\infty}^\infty \frac{1}{\sqrt{\eta}}\exp\left(tx - \frac{x^2}{2\eta} - \frac{\eta\gamma^2}{2}\right)dxd\eta$$

The integral w.r.t x can be evaluated first as -

$$I_x = \exp\left(\frac{\eta t^2}{2}\right)\int_{-\infty}^\infty \exp\left(\frac{-(x-\eta t)^2}{2\eta}\right)dx = \sqrt{2\pi\eta}\exp\left(\frac{\eta t^2}{2}\right)\int_{-\infty}^\infty \mathcal{N}(x|\eta t, \eta)dx$$
$$= \sqrt{2\pi\eta}\exp\left(\frac{\eta t^2}{2}\right)$$

Question 3

Hence we have the following -

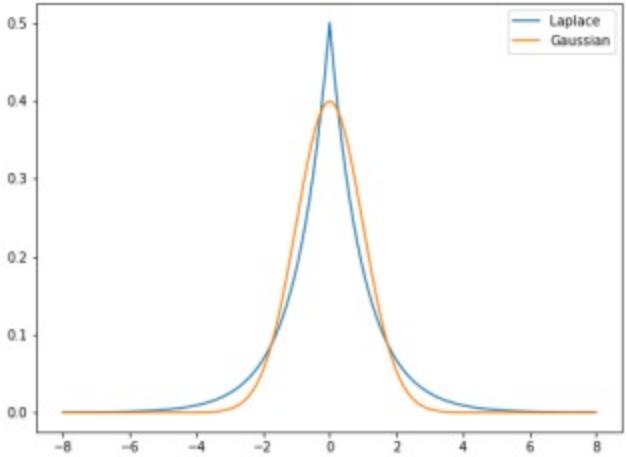
$$M(t) = \frac{\gamma^2}{2}\int_0^\infty \exp\left(\frac{\eta t^2 - \eta\gamma^2}{2}\right)d\eta = \begin{cases}\frac{\gamma^2}{\gamma^2-t^2} & t<\gamma \\ \infty & t\geqslant\gamma\end{cases}$$

Thus $M(t)$ is defined only for $(t<\gamma)$ and has the value given as above. Comparing this with the MGF's of known distributions we get this corresponds to the MGF of a Laplace distribution with $\mu = 0, b(\text{scale}) = \gamma$. The final marginalised expression is given by -

$$p(x|\gamma) = \text{Laplace}(x|0,\gamma) = \frac{1}{2\gamma}\exp\left(-\frac{|x|}{\gamma}\right)$$

The mean of the Laplace is 0 and its scale is γ. The marginal distribution $p(x|\gamma)$ is essentially the distribution obtained on integrating over the variance parameter of a Gaussian distribution. Intuitively this means summing up over several Gaussian distributions with varying variances. The uncertainty in the variance parameter is reflected in the resulting distribution. Note that the Laplace distribution is more peaked around its mean as compared to a Gaussian distribution(which is kind of smooth). The Laplace distribution is not differentiable only at $x = 0$ where as the Gaussian distribution is differentiable everywhere in its support.

Plot of Gaussian and Laplace distribution



The Gaussian and Laplacian shown above are 0 meaned with scale of the Laplace and variance of the Gaussian distrbution being 1.

The marginal likelihood for each school's data will be $p(\boldsymbol{y}^{(m)}|\mathbf{X}^{(m)}, \boldsymbol{w}_0) = \int p(\boldsymbol{y}^{(m)}|\mathbf{X}^{(m)}, \boldsymbol{w}_m)p(\boldsymbol{w}_m|\boldsymbol{w}_0)d\boldsymbol{w}_m$. This is clearly a Gaussian since both the terms in the integrand are Gaussian (one is the likelihood and the other is the prior on $\boldsymbol{w}_m$. Integrating out $\boldsymbol{w}_m$ gives (using Gaussian marginal from conditional formula)

Question 4

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

The overall marginal likelihood (function of the shared parameter $\boldsymbol{w}_0$) will be

$$\mathcal{L}(\boldsymbol{w}_0) = \sum_{m=1}^{M} \log \mathcal{N}(\boldsymbol{y}^{(m)}|\mathbf{X}^{(m)}\boldsymbol{w}_0, \lambda^{-1}\mathbf{X}^{(m)}\mathbf{X}^{(m)^\top} + \beta^{-1}\mathbf{I}_{N_m})$$

Learning $\boldsymbol{w}_0$ by maximizing the marginal likelihood above (i.e., basically doing MLE-II) helps in pooling data from all the school in estimating the shared parameter $\boldsymbol{w}_0$ which in turn is used in the prior $p(\boldsymbol{w}_m) = \mathcal{N}(\boldsymbol{w}_m|\boldsymbol{w}_0, \lambda^{-1}\mathbf{I}_D)$ of each school's regression weights. So all the weights are shrunk (jointly regularized) towards this shared parameter $\boldsymbol{w}_0$ (which is learned by pooling the data from all the school). In contrast, if we fix $\boldsymbol{w}_0$ to some value then each school s $\boldsymbol{w}_m$ will be regularized separately and there won't be any sharing of information.

Question 5

The conditional distribution is given by $p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x},y)}{p(\boldsymbol{x})}$.

Note that $p(\boldsymbol{x}) = \int p(\boldsymbol{x}, y)dy = \frac{1}{N}\sum_{n=1}^{N} \mathcal{N}(\boldsymbol{x} - \boldsymbol{x}_n|\mathbf{0}, \sigma^2\mathbf{I}_D)$ using the Gaussian marginal property.

Also, $p(\boldsymbol{x}, y) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{N}([\boldsymbol{x} - \boldsymbol{x}_n, y - y_n]^\top|\mathbf{0}, \sigma^2\mathbf{I}_{D+1}) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{N}(\boldsymbol{x} - \boldsymbol{x}_n|\mathbf{0}, \sigma^2\mathbf{I}_D)\mathcal{N}(y - y_n|0, \sigma^2)$ since the covariance matrix of the Gaussian is diagonal.

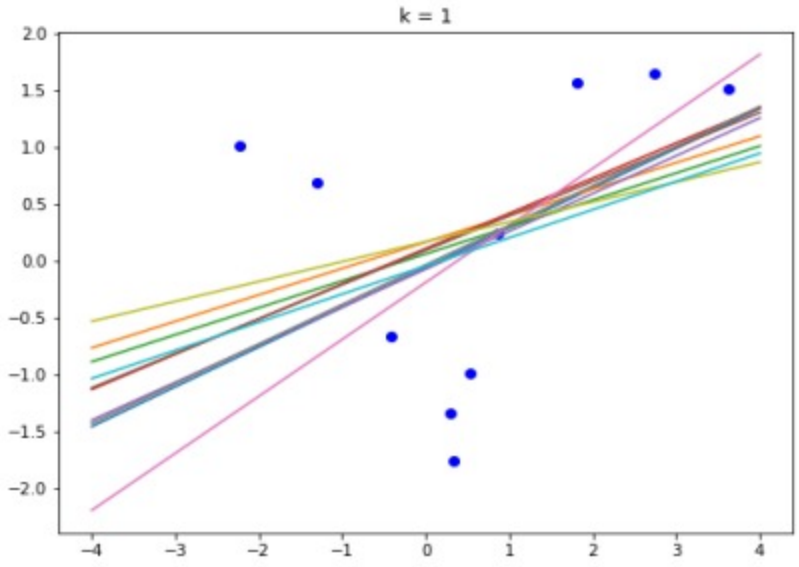
Plugging in the above results in the expression $p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x},y)}{p(\boldsymbol{x})}$, we get

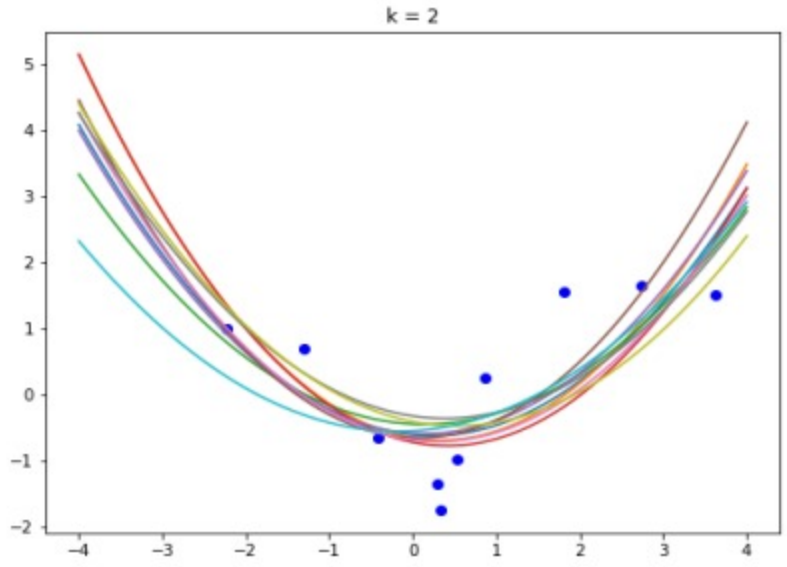
$$p(y|\boldsymbol{x}) = \frac{\frac{1}{N}\sum_{n=1}^{N} \mathcal{N}(\boldsymbol{x} - \boldsymbol{x}_n|\mathbf{0}, \sigma^2\mathbf{I}_D)\mathcal{N}(y - y_n|0, \sigma^2)}{\frac{1}{N}\sum_{m=1}^{N} \mathcal{N}(\boldsymbol{x} - \boldsymbol{x}_m|\mathbf{0}, \sigma^2\mathbf{I}_D)} = \sum_{n=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_n)\mathcal{N}(y|y_n, \sigma^2)$$

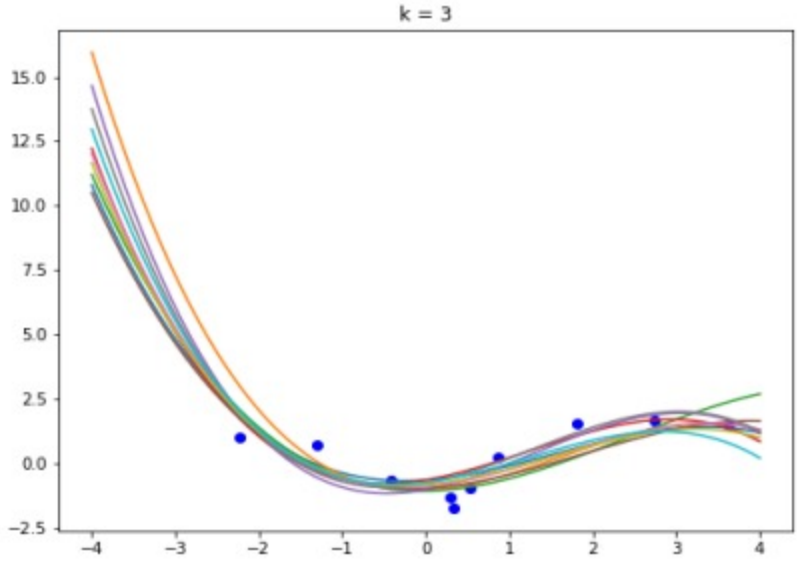
where $k(\boldsymbol{x}, \boldsymbol{x}_n) = \frac{\mathcal{N}(\boldsymbol{x}-\boldsymbol{x}_n|\mathbf{0},\sigma^2\mathbf{I}_D)}{\sum_{m=1}^{N}\mathcal{N}(\boldsymbol{x}-\boldsymbol{x}_m|\mathbf{0},\sigma^2\mathbf{I}_D)}$ Thus we see that the conditional distribution is a weighted sum of Gaussians centered at the training outputs and the weights are given by the similarity of $\boldsymbol{x}$ with the respective inputs. The expectation $\mathbb{E}[y|\boldsymbol{x}]$ is simply $\sum_{n=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_n)y_n$.

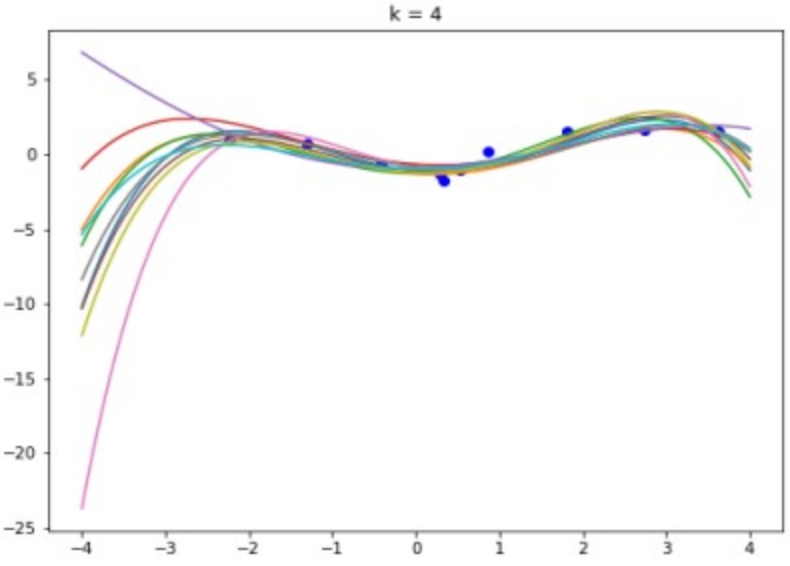
Plots for sampled weights from Posterior

Question 6



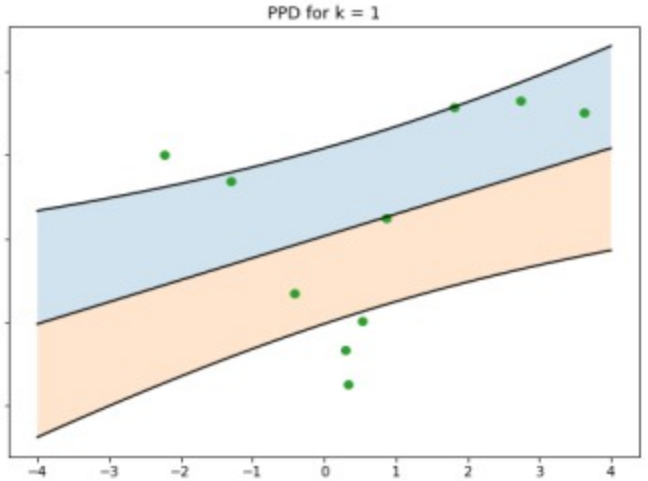


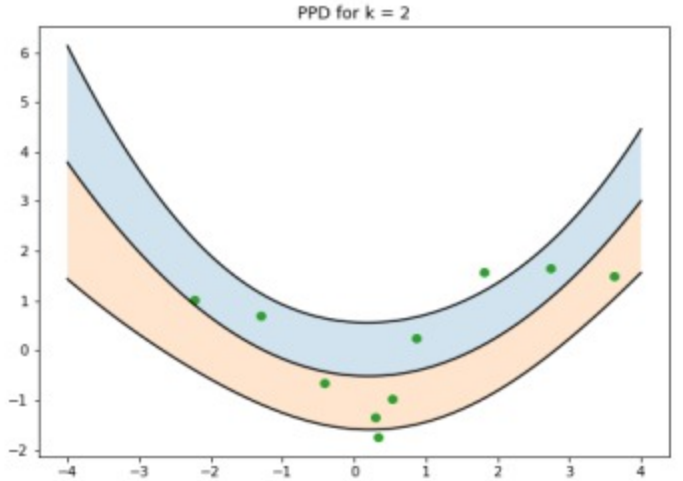


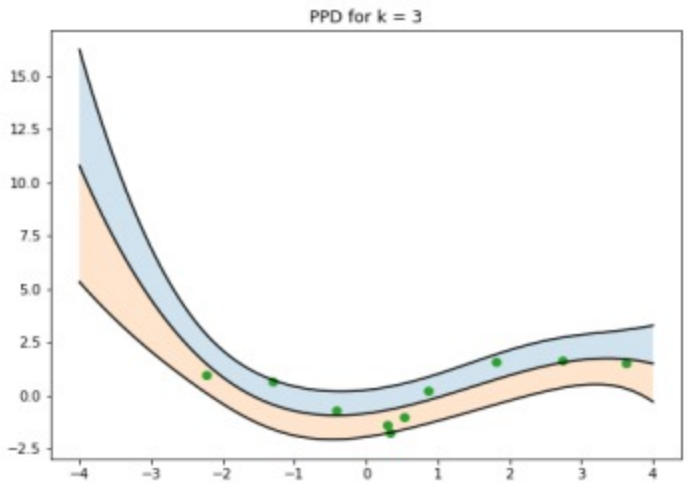


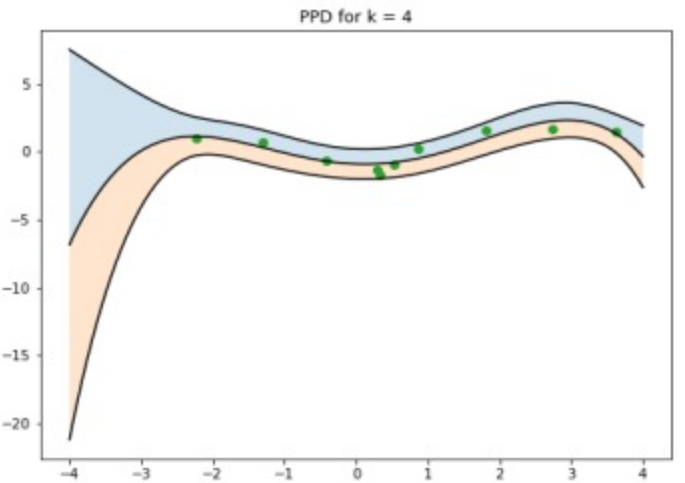
Each of the curves has a separate scale on the Y axis.

Posterior predictive distribution plots









Likelihood

k	Log Marginal Likelihood	Log Likelihood(using MAP estimate)
1	-32.3520	-28.0940
2	-22.7721	-15.3606
3	-22.0790	-10.9358
5	-22.3867	-7.2252

Observations

- Comparing the marginal likelihood corresponding to each of the 4 possible values of k we conclude that the model corresponding to $k = 3$ approximates the best.

- The model having maximum likelihood using the MAP estimate is the one corresponding to $k = 4$. Note that this model is different from the one which we chose on the basis of marginal likelihood(above). Log marginal likelihood is however a more reliable criterion to select the best model because, it considers the effects of all possible weight vectors in the prediction(posterior averaging over the possible weights), where as using the MAP estimate is essentially using a single best estimate of the weight. Posterior averaging gives robustness to the model.

- The best model corresponds to $k = 3$, hence choosing the inputs in the region where the model has more variance in its predictions would be a good idea. Choosing $x^{'}$ in the range [-4, -3] along with $y^{'}$ in the range [6, 12.5] would be a good idea.