Assignment 1

Ashutosh Mittal (amittal@kth.se)
DD2434 Machine Learning, Advanced Course

December 30, 2016

1 The Prior $p(\mathbf{X})$, $p(\mathbf{W})$, $p(f)$

Task 2.1: Theory

Question 1

Gaussian form of the likelihood function is a sensible choice because:

- Central Limit Theorem: The exact source of noise is seldom known, and it could be produced due to a number of random processes. CLT tells us that, subject to certain mild conditions, the sum of a set of random variables, which is of course itself a random variable, has a distribution that becomes increasingly Gaussian as the number of terms in the sum increases.
- Mathematical convenience: Gaussians are easy to work with. It is convenient to maximize the log likelihood function for joint probability of the training data. Conjugate prior for a Gaussian is a Gaussian as well, which helps in sequential estimation. Also, maximum likelihood estimator on gaussian distribution leads to highly intuitive minimum mean square error estimator.

The spherical co-variance matrix represents that the $\boldsymbol{q}$ dimensions of the input $\mathbf{Y}$ are independent of each other (bear no correlation with each other).

Question 2

If the data points are **not** assumed to be independent then we cannot write join probability of the N points as simple product of individual probabilities. The joint distribution would then be represented as:

$$\boldsymbol{P}(\mathrm{Y}|\boldsymbol{f},\mathrm{X}) = \prod_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{p}(\mathrm{y_i}|\mathrm{y}_{\boldsymbol{i-1}}, ..., \mathrm{y_1}, \boldsymbol{f}, \mathrm{X}) \tag{1}$$

Question 3

$$\mathbf{y_i} = \mathbf{W}\mathbf{x_i} + \boldsymbol{\epsilon} \tag{2}$$

$$\text{where } \boldsymbol{\epsilon} \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I})$$

Using eq 2 and assuming all the observation points have independent distribution:

$$\boldsymbol{P}(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^{N} \boldsymbol{\mathcal{N}}(\mathbf{y_i}|\mathbf{W}\mathbf{x_i}, \boldsymbol{\sigma}^2\mathbf{I}) \tag{3}$$

Question 4

In Bayesian probability theory, if the posterior distributions $\boldsymbol{p(\theta, x)}$ are in the same family as the prior probability distribution $\boldsymbol{p(\theta)}$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $\boldsymbol{p(x|\theta)}$.

In our case, chosing prior to be a Normal distribution makes sense because for a normal distributed likelihood function, a normal distributed prior would result in a posterior which is normal distributed as well. This will help us in sequential estimation.

Question 5

A generalized error function looks like:

$$\frac{1}{2}\Sigma_{n=1}^{N}\left[\boldsymbol{t_n} - \mathrm{w}^{\boldsymbol{T}}\boldsymbol{\phi}(\mathrm{x}_{\boldsymbol{n}})\right]^2 + \frac{\boldsymbol{\lambda}}{2}\Sigma_{j=1}^{M}|\boldsymbol{w_j}|^{\boldsymbol{q}}$$

Here q represents the order of the norm. The case of L1 norm (q=1) is known as lasso. Figure 3.3 in the course book (Bishop 2006) shows that for L1 norm we get a diamond around origin as the contour for the regularization term where as L2 norm has a circle. Using fig 3.4 in the course book (Bishop 2006) we can see that the meeting point for contours of two terms generalized error function for L1 and L2 norm. We can see that in L1 norm meeting point can correspond to some weights being zero however for L2 norm some weights might be very low valued but highly unlikely to be zero. This leads to a sparse model in which some basis functions play no role.

Question 6

$$\boldsymbol{P}(\mathrm{W}|\mathrm{x}, \mathrm{y}) = \frac{\mathbf{1}}{\boldsymbol{Z}}\boldsymbol{P}(\mathrm{y}|\mathbf{W}, \mathrm{x})\boldsymbol{P}(\mathrm{W})$$

Considering the simplification suggested in the question:
Note: Here $\mathbf{W}$ is $\boldsymbol{D}$x$\boldsymbol{q}$, $\mathbf{x_i}$ is $\boldsymbol{q}$x$\mathbf{1}$, $\mathbf{y_i}$ is $\boldsymbol{D}$x$\mathbf{1}$

$$\begin{aligned}
&P(\mathrm{W}|\mathrm{x_i}, \mathrm{y_i}) = \frac{1}{Z} P(\mathrm{y_i}|\mathrm{W}, \mathrm{x_i}) P(\mathrm{W}) \\
\Rightarrow\ &P(\mathrm{W}|\mathrm{x_i}, \mathrm{y_i}) \propto \mathcal{N}(\mathrm{W x_i}, \sigma^2 \mathrm{I}) \mathcal{N}(\mathrm{W_0}, \tau^2 \mathrm{I}) \\
\Rightarrow\ &P(\mathrm{W}|\mathrm{x_i}, \mathrm{y_i}) \propto \left[\frac{1}{(2\pi)^{1/2} \sigma^2} e^{-\frac{1}{2\sigma^2} (\mathrm{y_i} - \mathrm{W x_i})^T (\mathrm{y_i} - \mathrm{W x_i})} \right] \left[\frac{1}{(2\pi)^{q/2} \tau^2} e^{-\frac{1}{2\tau^2} (\mathrm{W} - \mathrm{W_0})^T (\mathrm{W} - \mathrm{W_0})} \right] \\
\Rightarrow\ &P(\mathrm{W}|\mathrm{x_i}, \mathrm{y_i}) \propto \left[e^{-\frac{1}{2\sigma^2} (\mathrm{y_i} - \mathrm{W x_i})^T (\mathrm{y_i} - \mathrm{W x_i})} \right] \left[e^{-\frac{1}{2\tau^2} (\mathrm{W} - \mathrm{W_0})^T (\mathrm{W} - \mathrm{W_0})} \right] \\
\Rightarrow\ &P(\mathrm{W}|\mathrm{x_i}, \mathrm{y_i}) \propto e^{-\frac{1}{2} \mathrm{W}^T \left[(\frac{1}{\sigma^2} \mathrm{x_i}^T \mathrm{x_i}) + (\frac{1}{\tau^2} I) \right] \mathrm{W}} e^{\mathrm{W}^T \left[\frac{1}{\tau^2} \mathrm{W_0} + \frac{1}{\sigma^2} (\mathrm{x_i}^T \mathrm{y_i}) \right]} e^{-\frac{1}{2\sigma^2} (\mathrm{y_i}^T \mathrm{y_i}) - \frac{1}{2\tau^2} (\mathrm{W_0^T W_0})}
\end{aligned} \tag{4}$$

As the likelihood function and prior as Normal distributed, we can expect the posterior to be normal distributed as well.

$$\begin{aligned}
&P(\mathrm{w_i}|x, \mathrm{y_i}) \sim \mathcal{N}(\mathrm{W}_p, \Sigma) \\
\Rightarrow\ &P(\mathrm{W}|\mathrm{x_i}, \mathrm{y_i}) \propto e^{-\frac{1}{2} (\mathrm{W} - \mathrm{W}_p)^T \Sigma^{-1} (\mathrm{W} - \mathrm{W}_p)} \\
\Rightarrow\ &P(\mathrm{W}|\mathrm{x_i}, \mathrm{y_i}) \propto e^{-\frac{1}{2} \mathrm{W}^T \Sigma^{-1} \mathrm{W}} e^{\mathrm{W}^T \Sigma^{-1} \mathrm{W}_p} e^{-\frac{1}{2} \mathrm{W}_p^T \Sigma^{-1} \mathrm{W}_p}
\end{aligned} \tag{5}$$

Comparing equations 4 and 5, we can get the respective parameters of the posterior:

$$\Sigma^{-1} = \left[(\frac{1}{\sigma^2} \mathrm{x_i}^T \mathrm{x_i}) + (\frac{1}{\tau^2} \mathrm{I}) \right] \tag{6}$$

$$\mathrm{W}_p = \Sigma \left[\frac{1}{\tau^2} \mathrm{W_0} + \frac{1}{\sigma^2} (\mathrm{x_i}^T \mathrm{y_i}) \right]$$

$$\Rightarrow \mathrm{W}_p = \left[(\frac{1}{\sigma^2} \mathrm{x_i}^T \mathrm{x_i}) + (\frac{1}{\tau^2} \mathrm{I}) \right] \left[\frac{1}{\tau^2} \mathrm{W_0} + \frac{1}{\sigma^2} (\mathrm{x_i}^T \mathrm{y_i}) \right] \tag{7}$$

No, $\boldsymbol{Z}$ has no role in determining the posterior's parameters. However if we have multiple data sets then we need to consider Z (probability of the respective dataset).In this case it will play a role in the posterior.

Question 7

Non-parametric model is one in which the predictor does not take a predetermined form but is constructed according to information derived from the data. Nonparametric regression requires larger sample sizes.

Given the parameters, future predictions in parametric models are independent of the observed data. However, in non-parametric models amount of information that parameters can capture about the data can grow as the amount of data grows. This makes them more `representative`. In contrast to this, the inability to control the complexity of the model in non-parametric methods often gives rise to problems with `interpretation`. Parametric methods score higher in interpretability.

Question 8

Using the prior stated in eq 11 of the assignment and the expression for $\mathbf{y}$ stated in eq 11 of the assignment we can compute the marginal for $\mathbf{y}$. Using the derivations in section 6.4.2 in the course book (Bishop 2006) we can write the marginal as follows:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{f})\, p(\boldsymbol{f})\, d\boldsymbol{f} = \mathcal{N}(\mathbf{0}, \mathbf{C}) \tag{8}$$

where elements of covariance matrix $\mathbf{C}$ are given by

$$C(i,j) = k(\mathbf{x}_i, \mathbf{x}_j) + \sigma^2 \delta_{ij} \tag{9}$$

We can clearly see that elements of $\boldsymbol{C}$ has two components: one is due to the noise $\boldsymbol{\epsilon}$, while the other term is due to the variance of the prior we have assumed. Kernel $\boldsymbol{K}(\mathbf{x_1}, \mathbf{x_2})$ would have high value if $\mathbf{x_1}, \mathbf{x_2}$ are apart and less if closer.

Using equation 6.66 and 6.67 from the course book we can write the mean and variance of a predictive term $\boldsymbol{x}_{N+1}$.

$$\begin{aligned} m(\mathbf{x}_{N+1}) &= \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{y} \\ \sigma^2(\mathbf{x}_{N+1}) &= c - \mathbf{k}^T \mathbf{C}_N^{-1} \mathbf{k} \end{aligned} \tag{10}$$

So the prediction has high variance if the kernel function has high value. Essentially, points near the known $(\boldsymbol{x}, \boldsymbol{y})$ points would have low variance while those away from it would have high variance. This can be seen in fig 1 taken from the course book.

Hence we can say Prior provides a starting point and an informed guess about what could be the distribution.

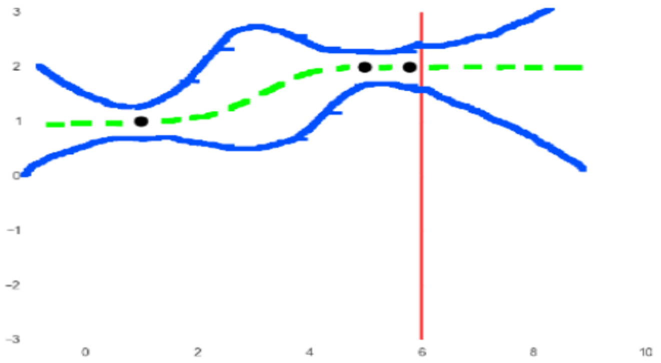


Figure 1: Variance and mean of predictive points

Question 9

The joint likelihood function is given by

$$\begin{aligned} p(\mathbf{Y}, \mathbf{X}, \boldsymbol{f}, \boldsymbol{\theta}) &= p(\mathbf{Y}|\boldsymbol{f}, \boldsymbol{\theta}_1)\, p(\boldsymbol{f}|\mathbf{X}, \boldsymbol{\theta}_2) \\ &= \Big[\prod_{i=1}^{N} \mathcal{N}(\boldsymbol{f}, \theta_1^2 \mathbf{I})\Big]\, \mathcal{N}(\mathbf{0}, k(\mathbf{X}, \mathbf{X})) \end{aligned} \tag{11}$$

where $\boldsymbol{\theta_1}$ is noise variance and $\boldsymbol{\theta_2}$ is parameter for defining kernel of the prior. Both of them are design hyper-parameters. The graphical model for the same is shown in fig 2.

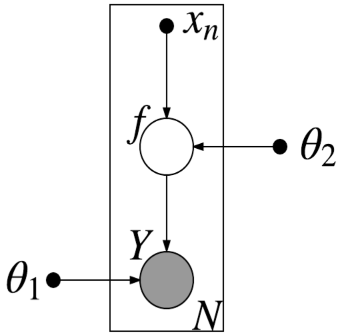


Figure 2: Graphical Model

Question 10

Until now $\mathbf{Y}$ was epressed in terms of $\boldsymbol{f}$ which was in turn connected with $\mathbf{X}$. However, we wish to directly predict $\mathbf{Y}$ from $\mathbf{X}$ thus we can perform the following marginalization over possible functions $\boldsymbol{f}$:

$$\boldsymbol{p(\mathrm{Y}|\mathrm{X},\theta) = \int p(\mathrm{Y}|f)p(f|\mathrm{X},\theta)df} \tag{12}$$

Here $\boldsymbol{\theta}$ is a hyperparameter which represents our belief about the system.
Using expressions for mean and variance of predicted points from eq 10 we can see that kernel defined for the prior affects $\boldsymbol{P(\mathrm{Y}|\mathrm{X},\theta)}$. There is high variance for points near known points where as high as we move away from them. This uncertainty is translated to the output through the kernel function.

Task 2.2: Practical

Question 11

- **Prior distribution**

$$\boldsymbol{\mathrm{W} = \mathcal{N}(\mathrm{W_0}, \tau^2\mathrm{I})}$$
$$\textit{where,}\, \boldsymbol{\mathrm{W_0} = [-1.3, 0.5]}$$
$$\boldsymbol{Var = 2 * \mathrm{I}}$$

  Figure 3 shows the distribution of this prior over the weights.

- **Posterior distribution** From the data $\mathbf{x}$, a point was sampled at random and the distribution of the posterior over the $\mathbf{W}$ looked as shown in figure 4a

- **Functions y vs. x** Sampling from the posterior over $\mathbf{W}$ figure 4b shows the functions.

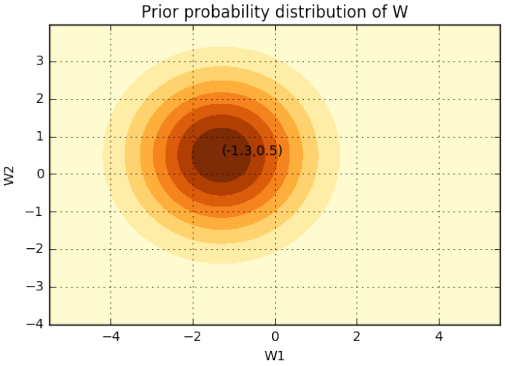


Figure 3: Distribution of the prior over the weights

- **Iterations** Sequential estimation was used thenafter. Plots of the posterior distribution and the corresponding **y** vs **x** plots are shown in figure 4.

We can see that as we iterating, the variance of the distribution over **W** is getting smaller and the functions **y** vs **x** are aligning to the same slope. Thus, we can conclude that we are zeroing down on the appropriate weights **W**. This is definitely desirable as can now determine the apt weights for the system.

Question 12

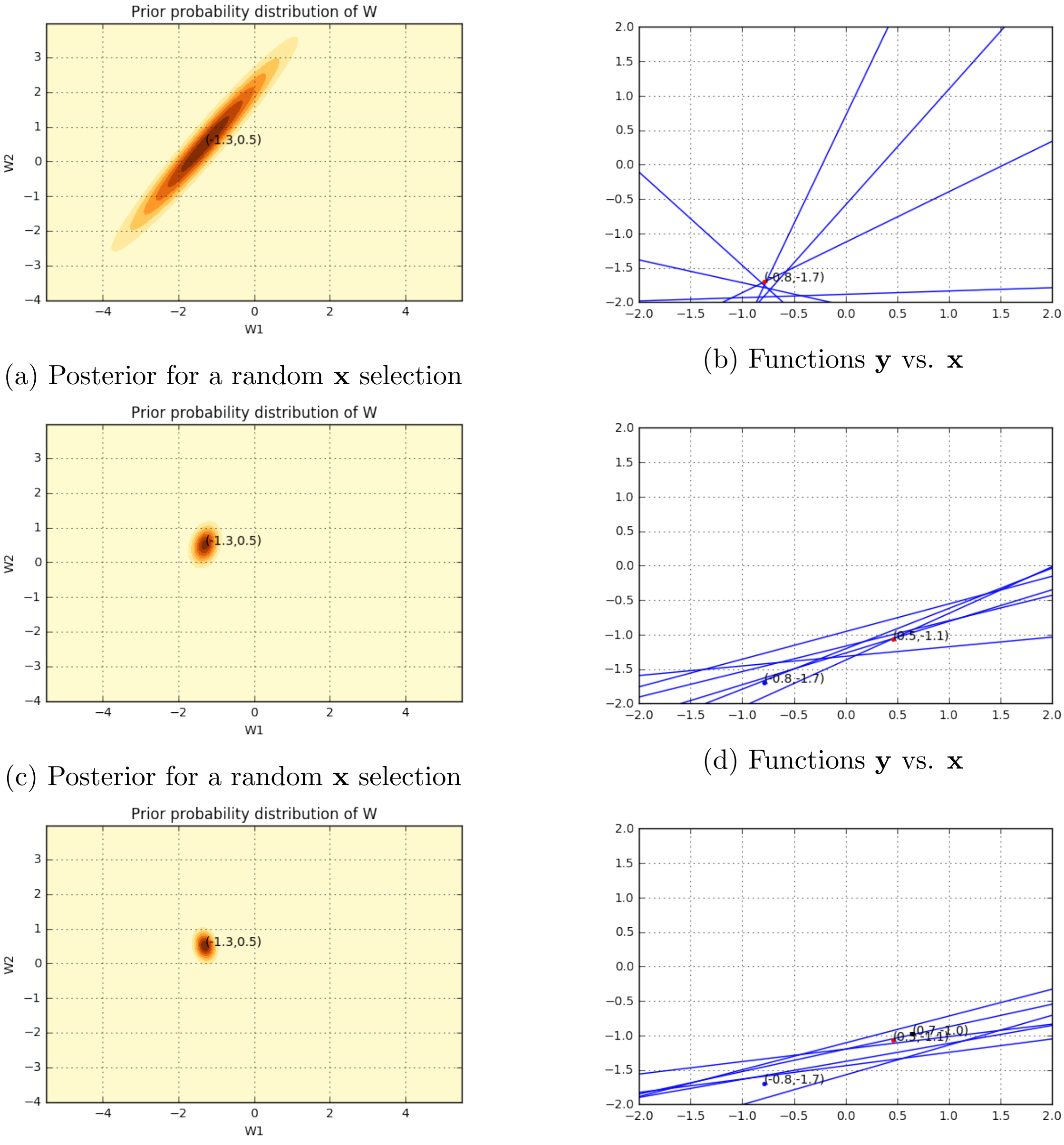
Here I used the squared variance kernel as specified in equation 17 of the assignment. Using this, $\boldsymbol{\mathcal{GP}}$ prior was generated and sampled as shown in figure 5 for varying values of the length-scale ($\boldsymbol{l}$) of the covariance function. It was observed that as $\boldsymbol{l}$ was increased, the sampled plots became smoother. Higher $\boldsymbol{l}$ causes the nearby points to be more related to each other. For $\boldsymbol{l} = \mathbf{10}$, effect of the nearby points was so pronounced that almost straight lines were observed. For $\boldsymbol{l} = \mathbf{0.01}$ each point behaves like an independent gaussian distributed point.

Question 13

Posterior is basically prior manipulated by the observed data with the premise of our beliefs about the distribution. If no data is provided, we cannot make the likelihood function and all we have is our belief about it. Thus, prior and posterior would essentially become the same.

Question 14

Figure 6 shows the the posterior distribution. We can clearly observe that the variance increases as we move away from the known points. Posterior is basically prior put together with the data. This is evident here. The observed data points reshape the randomly distributed samples of the prior to pass through/near them. This is desirable because as we keep on adding more data points we can get a robust and more accurate posterior.



(a) Posterior for a random **x** selection

(b) Functions **y** vs. **x**

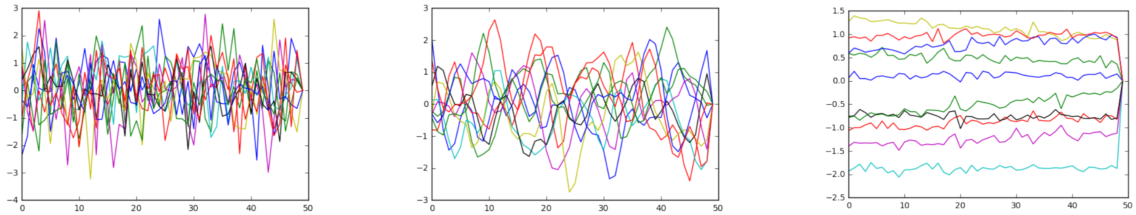
(c) Posterior for a random **x** selection

(d) Functions **y** vs. **x**

(e) Posterior for a random **x** selection

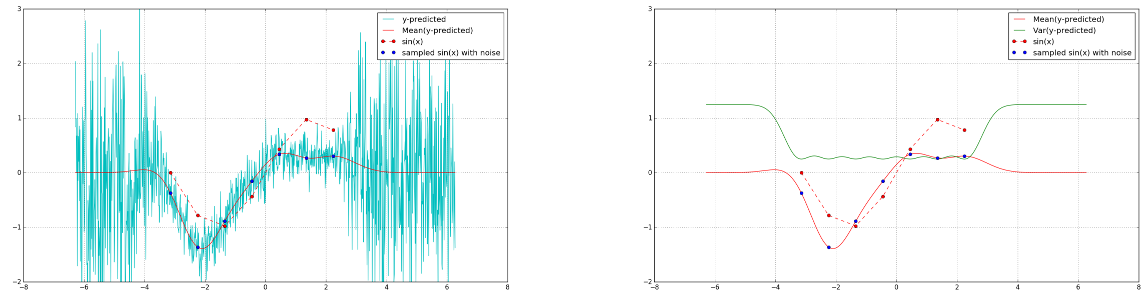
(f) Functions **y** vs. **x**

Figure 4: Plots of sequential estimation



(a) l = 0.01 (b) l = 1 (c) l = 10

Figure 5: Plots of samples from the $\mathcal{GP}$ prior with different length-scale of the covariance function



(a) With sampling from the posterior distribution (b) Mean and variance of posterior distribution

Figure 6: Plots of posterior distribution values predicted using squared variance kernel

Adding diagonal covariance matrix to squared exponential (fig 7): Doing this has enabled the kernel to somehow take account of the inherent noise in the observed data points. Before adding this diagonal covariance matrix, mean of the posterior was strictly passing through the data points but after adding it, it skips some of them if they seem to be affected by noise. Adding this matrix has improved the estimation process.

2 The Posterior $p(\mathrm{X}|\mathrm{Y})$

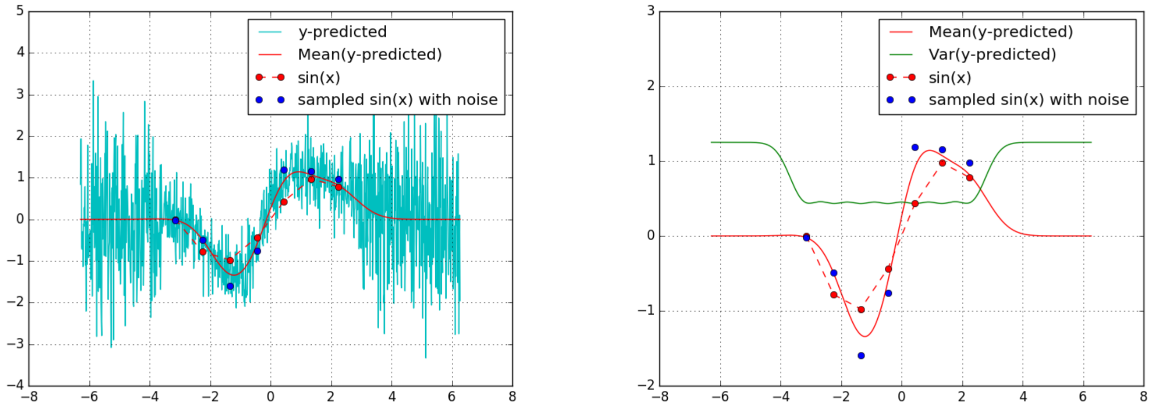
Task 2.3: Theory

Question 15

Prior can be seen a as a preference as it sets the constrains for the distribution of the prior. Any new data/information is seen in the light of this assumption about the prior. It gives a starting point and direction to which posterior should proceed. In eq 10 we have seen that even after marginalizing how uncertainty filters through.

Question 16

In this prior, different dimensions of **X** are independent and variance in each of these dimensions is unity. This prior encodes our preference that we wish to have **X** which has no



(a) With sampling from the posterior distribution (b) Mean and variance of posterior distribution

Figure 7: Plots of posterior distribution values predicted using squared variance kernel added with diagonal covariance matrix

correlation among its vector dimensions. In a way we are trying to reduce **Y** to a minimum dimensional space through this preference. Mean being zero represents that we has translated **Y** have zero mean as well.

Question 17

$$
\begin{aligned}
P(\mathbf{y_i}|\mathbf{W}) &= \int_{\mathbf{x_i}} P(\mathbf{y_i}|\mathbf{W},\mathbf{x_i})P(\mathbf{x_i})d\mathbf{x_i} \\
P(\mathbf{y_i}|\mathbf{W}) &= \int_{\mathbf{x_i}} \mathcal{N}(\mathbf{W}\mathbf{x_i},\sigma^2)\mathcal{N}(0,\mathbf{I})d\mathbf{x_i} \\
\Rightarrow P(\mathbf{y_i}|\mathbf{W}) &= \int_{\mathbf{x_i}} \left[\frac{1}{(2\pi)^{1/2}\sigma}e^{-\frac{1}{2\sigma^2}(\mathbf{y_i}-\mathbf{W}\mathbf{x_i})^T(\mathbf{y_i}-\mathbf{W}\mathbf{x_i})}\right]\left[\frac{1}{(2\pi)^{q/2}}e^{-\frac{1}{2}\mathbf{x_i}^T\mathbf{x_i}}\right]d\mathbf{x_i} \\
\Rightarrow P(\mathbf{y_i}|\mathbf{W}) &= \int_{\mathbf{x_i}} \frac{1}{(2\pi)^{(q+1)/2}\sigma}e^{-\frac{1}{2}\mathbf{x_i}^T\left[(\frac{1}{\sigma^2}\mathbf{W}^T\mathbf{W})+\mathbf{I}\right]\mathbf{x_i}}e^{\mathbf{x_i}^T\left[\frac{1}{\sigma^2}(\mathbf{W}^T\mathbf{y_i})\right]}e^{-\frac{1}{2\sigma^2}(\mathbf{y_i}^T\mathbf{y_i})}d\mathbf{x_i}
\end{aligned}
\tag{13}
$$

$$
\begin{aligned}
\Sigma =& \big[(\frac{1}{\sigma^2}\mathbf{W}^T\mathbf{W}) + \mathbf{I}\big]^{-1} \\
\mu =& \Sigma\big[\frac{1}{\sigma^2}(\mathbf{W}^T\mathbf{y_i})\big] \\
\Rightarrow P(\mathbf{y_i}|\mathbf{W}) =& \int_{\mathbf{x_i}} \frac{||\Sigma||}{(2\pi)^{1/2}\sigma}\mathcal{N}(\mu, \Sigma)e^{\frac{1}{2}\mu^T\Sigma^{-1}\mu}e^{-\frac{1}{2\sigma^2}(\mathbf{y_i}^T\mathbf{y_i})}d\mathbf{x_i} \\
\Rightarrow P(\mathbf{y_i}|\mathbf{W}) =& \frac{||\Sigma||}{(2\pi)^{1/2}\sigma}e^{\frac{1}{2}\mu^T\Sigma^{-1}\mu}e^{-\frac{1}{2\sigma^2}(\mathbf{y_i}^T\mathbf{y_i})} \\
\Rightarrow P(\mathbf{y_i}|\mathbf{W}) =& \frac{||\Sigma||}{(2\pi)^{1/2}\sigma}e^{-\frac{1}{2}\mathbf{y_i}^T\left[\frac{1}{\sigma^2}\mathbf{I}-\frac{1}{\sigma^4}\mathbf{W}\Sigma^T\mathbf{W}^T\right]\mathbf{y_i}}
\end{aligned} \tag{14}
$$

$$
\begin{aligned}
\Sigma_2^{-1} =& \big[\frac{1}{\sigma^2}\mathbf{I} - \frac{1}{\sigma^4}\mathbf{W}\Sigma^T\mathbf{W}^T\big] \\
\Rightarrow \Sigma_2 =& \sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^T \\
\Rightarrow P(\mathbf{y_i}|\mathbf{W}) =& \mathcal{N}(0, \Sigma_2)
\end{aligned} \tag{15}
$$

$$
\begin{aligned}
P(\mathbf{Y}|\mathbf{W}) =& \prod_{1}^{N} P(\mathbf{y_i}|\mathbf{W}) \\
\Rightarrow P(\mathbf{Y}|\mathbf{W}) =& \mathcal{N}(0, \Sigma_2)
\end{aligned} \tag{16}
$$

Question 18

The first one is maximum likelihood (ML), second one MAP and the last one is Type-II Maximum-Likelihood.
MAP is same as ML except for the regularization term which is introduced due to the prior distribution. Type-II Maximum-Likelihood is same as MAP except for the fact that it is marginalized over **X**.
As we get more data, MAP can be used to adjust the parameters as per the new data however, however ML does not. ML needs complete recalculation where MAP can be progressively computed as we get ore data.
The last two expressions of assignment's eq 25 are equal because the denominator term is independent of **W** as it will be integrated out. Hence it can be ignored.

Question 19

Using results from eq 16

$$
\begin{aligned}
\mathcal{L}(\mathbf{W}) =& -\ln(p(\mathbf{Y}|\mathbf{W})) \\
\Rightarrow \mathcal{L}(\mathbf{W}) =& \frac{ND}{2}\ln(2\pi) + \frac{N}{2}\ln|\Sigma_2| + \frac{1}{2}\Sigma_{i=1}^{N}\mathbf{y_i}^T\Sigma_2^{-1}\mathbf{y_i}
\end{aligned}
$$

Using results from the publication ”Probabilistic Principal Component Analysis” by Christopher M. Bishop and Michael E. Tipping in Journal of Royal Statistical Society (1999):

$$\begin{aligned}\frac{d\mathcal{L}}{d\mathbf{W}} &= \frac{N}{2|\Sigma_2|}\frac{d|\Sigma_2|}{d\mathbf{W}} + \frac{N}{2}\frac{d}{d\mathbf{W}} Tr(\Sigma_2^{-1}\mathbf{Y}^T\mathbf{Y}) \\ \Rightarrow \frac{d\mathcal{L}}{d\mathbf{W}} &= N(\Sigma_2^{-1}[\frac{1}{N}\mathbf{Y_i}^T\mathbf{Y}]\Sigma_2^{-1}\mathbf{W} \,-\, \Sigma_2^{-1}\mathbf{W})\end{aligned}$$

Task 2.4: Practical

Question 20

After constructing $\mathbf{Y}$ as per equations 28-34 from the assignment, attempt was made to estimate $\mathbf{x}$.

- First I tried to directly estimate $\mathbf{x}$, however the algorithm could only learn the linear relation f_{lin} and a curve similar to $\mathbf{x}sin(\mathbf{x})$ was obtained.
- Understanding the fact that the algorithm is able to learn only the linear part (during the theoretical derivation also we assumed input $\mathbf{x}$ to have a linear relation with $\mathbf{Y}$), now $\mathbf{x'}$ was estimated instead of $\mathbf{x}$. Figure 8 shows the estimated curve obtained.

The obtained results are not very accurate and seem to deviate from the actual curve. Also, I haven't been able to estimate the $f_{non-lin}$.

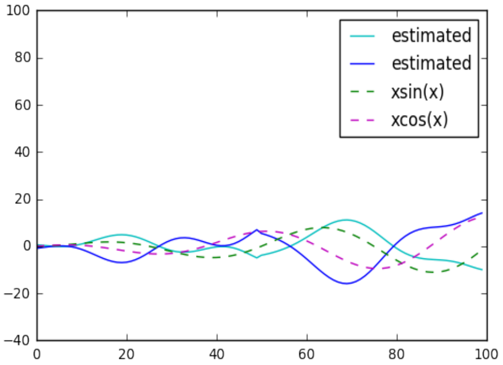


Figure 8: Predicted $\mathbf{x'}$ (x axis varies from 0 to 4π)

3 The Evidence $p(\mathbf{Y})$

Task 2.5: Theory

Question 21

Simplicity refers to number of model parameters. In M_0 there is no parameter thus it can be seen as the simplest model. This implies we don't need to estimate any parameter before applying this model. It performs equally good and bad on all data sets. However, the same simplicity can be a bane as it might capture the effect of parameter of $\boldsymbol{x}$ on the data set. Thus, other models capturing this would outperform $\boldsymbol{M_0}$ for some data sets.

Question 22

We decide upon a model which is based upon $\boldsymbol{(\mathrm{x}, \theta)}$ and try and predict the probability of the data sets ($\boldsymbol{\mathcal{D}}$). $\boldsymbol{M_0}$ has no parameters to accommodate behavior of the data thus not flexible at all. However $\boldsymbol{M_1}$ has $\boldsymbol{\theta_1}$ which can be set to fit the data, thus this model is more flexible. $\boldsymbol{M_1}$ has heavy probability mass for data sets that depend on $\boldsymbol{x_1^i}$ however fails to capture the effect of $\boldsymbol{x_2^i}$ or a mean offset ($\boldsymbol{\theta_0}$).

Question 23

As we are increasing the number of parameters in the model we are increasing the complexity of the system but as the same time we are increasingly restricting the data set they predict. For examle $\boldsymbol{M_0}$ has same probability of for all values of $\boldsymbol{x}$ however $\boldsymbol{M_3}$ has high probability only for some values of $\mathbf{x_1, x_2}$. $\boldsymbol{M_0}$ is suited for all data sets and is uncertain for all data sets equally. $\boldsymbol{M_3}$ is the standard logistic regression and least uncertain since it uses most parameters. It is good for data sets where there might be bias and both $\mathbf{x_{i1}, x_{i2}}$ are varying. $\boldsymbol{M_1}$ is good for data sets where only $\mathbf{x_{i1}}$ varies. $\boldsymbol{M_2}$ is good when $\mathbf{x_{i1}, x_{i2}}$ varies but the bias is not significant. $\boldsymbol{M_2}$ is more certain that $\boldsymbol{M_1}$. $\boldsymbol{H_0}$ is suited for data sets which don't depend on $\mathbf{x_1, x_1}$. $\boldsymbol{M_1}$ is good for data sets which vary with $\mathbf{x_1}$ but not with $\mathbf{x_2}$. $\boldsymbol{M_0}$ treats all data sets equally and is least restrictive and least flexible. $\boldsymbol{M_3}$ is most flexible and thus most restrictive as well. $\boldsymbol{M_2}$ is the same as $\boldsymbol{M_3}$ but without the bias weight. $\boldsymbol{M_1}$ is the same as $\boldsymbol{M_2}$ except it ignores the second dimension of x, thus less flexible and less restrictive as well.

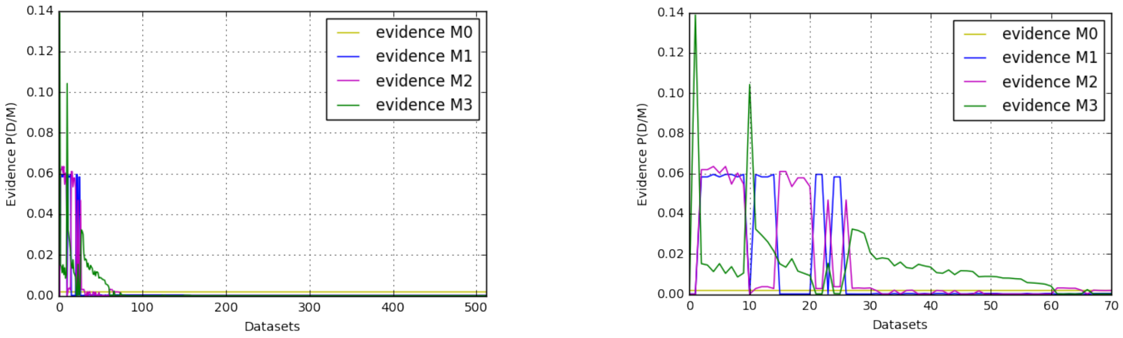
Question 24

The process of Marginalisation includes eliminating the dependence of the probability on a certain parameter using the probability distribution of this parameter. Taking into account uncertainty at each step is the spirit of Bayesian analysis. The implication would be that by applying equation (40) in the assignment, we can accommodate the distribution of parameter $\boldsymbol{\theta}$ in the probability distribution of data set and thus eliminate its dependence in $\boldsymbol{\mathcal{P}(\mathcal{D}/\mathcal{M}_\rangle)}$.

Question 25

The chosen prior signifies that the $\boldsymbol{\theta_i}$'s are not correlated. They vary independently with mean zero. The weights correspond to a sharp linear boundary in $\mathbf{x}$ space. High variance of weights signify they can vary over a large range of values which helps the model to have these sharp transitions otherwise it would become flat like $\boldsymbol{M_0}$.

The bias weight would be captured by $\boldsymbol{\theta_0}$ thus we can assume the parameters to have zero mean.



(a) Distribution over 512 datasets

(b) Zoomed for first 70 data sets

Figure 9: Plots of distribution of evidence for different models

Task 2.6: Practical

Question 26

For each model, summing the evidence over all the data set gives 1 which is what is expected. Evidence is merely a distribution of probability weights over the different data sets. Sum of probabilities has to be 1 for any probability which is evident here as well.

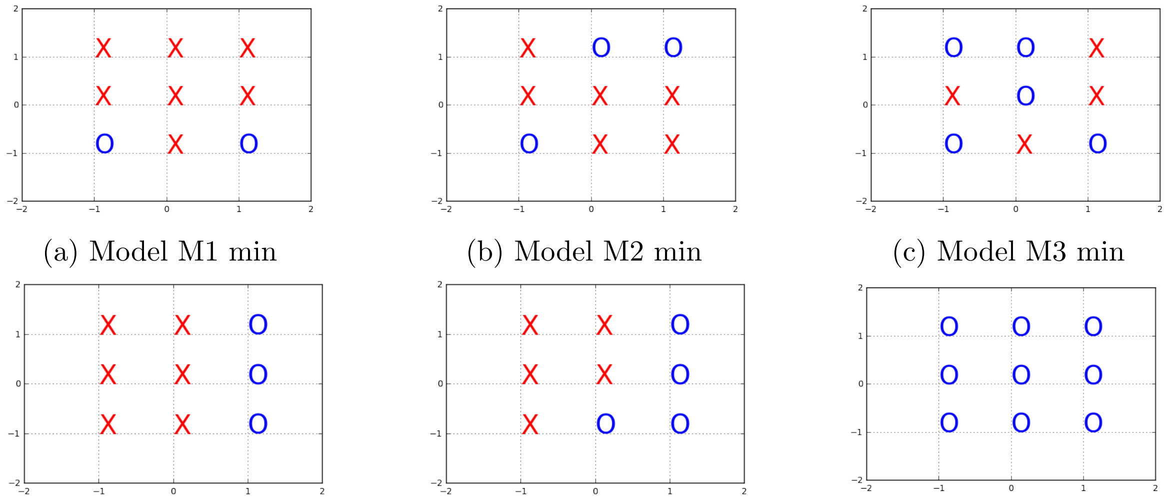
Question 27

Figure 9 shows the plot of evidence for the four models considered in the assignment.

As explained above in theory questions, we can clearly observe that as we decrease the number of parameters in the model, probability mass is distributed over more data sets. Model $\boldsymbol{M_4}$ is concentrated to a few data sets where as the simplest model $\boldsymbol{M_0}$ covers the almost the entire range of datasets. As suggested by Occum's razor, one should choose the simplest model that explains all of the data in consideration. This plot can help in chosing the model in this regard.

Question 28

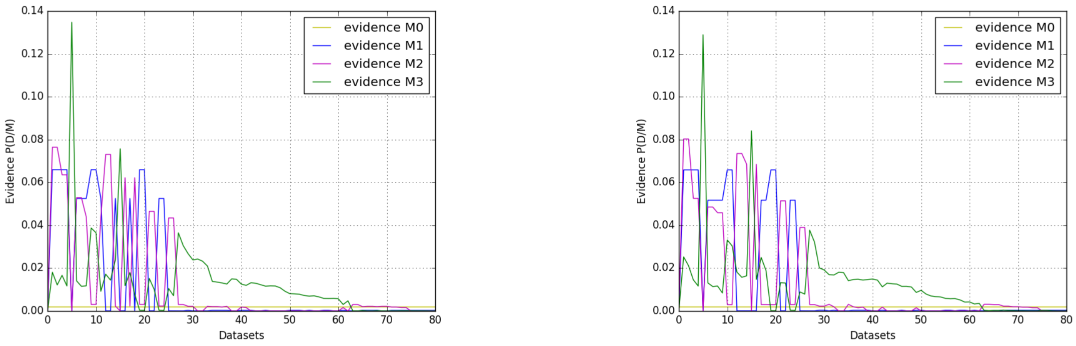
Plots for the data-sets which are given the highest and lowest evidence for each model are shown in figure 10. Model $\boldsymbol{M_1}$ has max probability for the data set shown in 10d and understandably so because the data set has only $\boldsymbol{x_1}$ is varying while if we move along $\boldsymbol{x_2}$ output is same. Minimum of model $\boldsymbol{M_1}$ has a representation in which both $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ vary where as model only captures the variation of $\boldsymbol{x_1}$. Model $\boldsymbol{M_2}$ does not include the bias in its model, thus we might be seeing the respective max and min. However, it is difficult to explicitly explain the figures with this reasoning. Model $\boldsymbol{M_3}$ has maximum for 10f because it is able to capture the bias which none of the other model captures.



(a) Model M1 min (b) Model M2 min (c) Model M3 min

(d) Model M1 max (e) Model M2 max (f) Model M3 max

Figure 10: Plots of data sets with highest and lowest evidence for different models



(a) Mean = [0,0,0] (b) Mean = [5,5,5]

Figure 11: Plots of evidence of data sets with different means of prior

Question 29

When mean was shifted to $[\mathbf{5, 5, 5}]$ some changes were observed. The effect of the parameters over the evidence of data sets became much more pronounced. Clearly, the difference between max and min evidence enlarged as shown in fig 11. The evidence plot became largely concentrated over the the data sets which follow the model.
However, when non- diagonal co-variace matrix was used, no observable difference was there in the results.