Imperial College
London

COURSEWORK 3

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Mathematics for Machine Learning

Author:
Arnold Cheung (CID: 01184493)

Date: November 7, 2019

1 Question 1

a) The likelihood $p(\boldsymbol{y}|\boldsymbol{X})$ can be factorised:

$$p(\boldsymbol{y}|\boldsymbol{X}) = \prod_{i=1}^{N} p(y_i|x_i) \tag{1}$$

where:

$$y_i \sim \mathcal{N}(\boldsymbol{w}^T \boldsymbol{\phi}(x_i), \sigma^2) \tag{2}$$

The maximum likelihood solution for the parameters σ^2 and w can be found as follows:
Begin by applying the *argmax* function to the likelihood, and take the log of the likelihood to turn the product into a sum, this can be done as the log function doesn't change the position of the maximum.:

$$\begin{aligned} argmax_w p(\boldsymbol{y}|\boldsymbol{X}) &= argmax_w \prod_{i=1}^{N} p(y_i|\boldsymbol{\phi}(x_i)) \\ &= argmax_w \sum_{i=1}^{N} \log p(y_i|\boldsymbol{\phi}(x_i)) \\ &= argmax_w \sum_{i=1}^{N} \log \mathcal{N}(y_i|\boldsymbol{w}^T \boldsymbol{\phi}(x_i), \sigma^2) \end{aligned} \tag{3}$$

The function with respect to w can be expanded using the Gaussian distribution:

$$\begin{aligned} f(\boldsymbol{w}, \sigma^2) &= \sum_{i=1}^{N} \log(\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2}\frac{(y_i - \boldsymbol{w}^T \boldsymbol{\phi}(x_i))^2}{\sigma^2})) \\ &= \sum_{i=1}^{N} \log(\frac{1}{\sqrt{2\pi\sigma^2}}) - \frac{1}{2\sigma^2}(y_i - \boldsymbol{w}^T \boldsymbol{\phi}(x_i))^2 \\ &= -N \log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \boldsymbol{w}^T \boldsymbol{\phi}(x_i))^2 \end{aligned} \tag{4}$$

Now take the partial derivative of f with respect to w and set it to 0 to find the maximum:

$$\begin{aligned} \frac{\partial f}{\partial \boldsymbol{w}} &= 0 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{w}} (y_i - \boldsymbol{w}^T \boldsymbol{\phi}(x_i))^2 \\ &= -\frac{1}{2\sigma^2} \frac{\partial}{\partial \boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{\Phi} \boldsymbol{w}||^2 \\ &=> -\frac{1}{2\sigma^2} \frac{\partial}{\partial \boldsymbol{w}} ||\boldsymbol{y} - \boldsymbol{\Phi} \boldsymbol{w}||^2 = 0 \end{aligned} \tag{5}$$

Solve for w:

$$\begin{aligned}
\frac{\partial}{\partial w}||\boldsymbol{y}-\boldsymbol{\Phi w}||^2 &= 0\\
\frac{\partial}{\partial w}(\boldsymbol{y}-\boldsymbol{\Phi w})^T(\boldsymbol{y}-\boldsymbol{\Phi w}) &= 0\\
\frac{\partial}{\partial w}(\boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{y}^T\boldsymbol{\Phi w}-\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{y}+\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi w}) &= 0\\
-\boldsymbol{y}^T\boldsymbol{\Phi}-\boldsymbol{y}^T\boldsymbol{\Phi}+2\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi} &= 0\\
2(\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}-\boldsymbol{y}^T\boldsymbol{\Phi}) &= 0\\
\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi} &= \boldsymbol{y}^T\boldsymbol{\Phi}\\
\boldsymbol{w}^T &= \boldsymbol{y}^T\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1} \qquad (6)
\end{aligned}$$

Same can be done for σ^2, by taking the partial derivative of f with respect to σ^2:

$$\begin{aligned}
\frac{\partial f}{\partial\sigma^2} &= -\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{N}(y_i-\boldsymbol{w}^T\boldsymbol{\phi}(x_i))^2\\
&=> -\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{N}(y_i-\boldsymbol{w}^T\boldsymbol{\phi}(x_i))^2 = 0 \qquad (7)
\end{aligned}$$

Solve for σ^2:

$$\begin{aligned}
-\frac{N}{2}\sigma^2+\frac{1}{2}\sum_{i=1}^{N}(y_i-\boldsymbol{w}^T\boldsymbol{\phi}(x_i))^2 &= 0\\
\sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(y_i-\boldsymbol{w}^T\boldsymbol{\phi}(x_i))^2\\
\sigma^2 &= \frac{1}{N}||\boldsymbol{y}-\boldsymbol{\Phi w}||^2 \qquad (8)
\end{aligned}$$

Substituting w to be in terms of Φ:

$$\sigma^2 = \frac{1}{N}||\boldsymbol{y}-\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}||^2 \qquad (9)$$

The plot below shows the maximum likelihood estimation using polynomial basis functions, where:

$$\phi_0(x) = 1$$
$$\phi_j(x) = x^j$$
$$for j = 1, 2, ...K$$

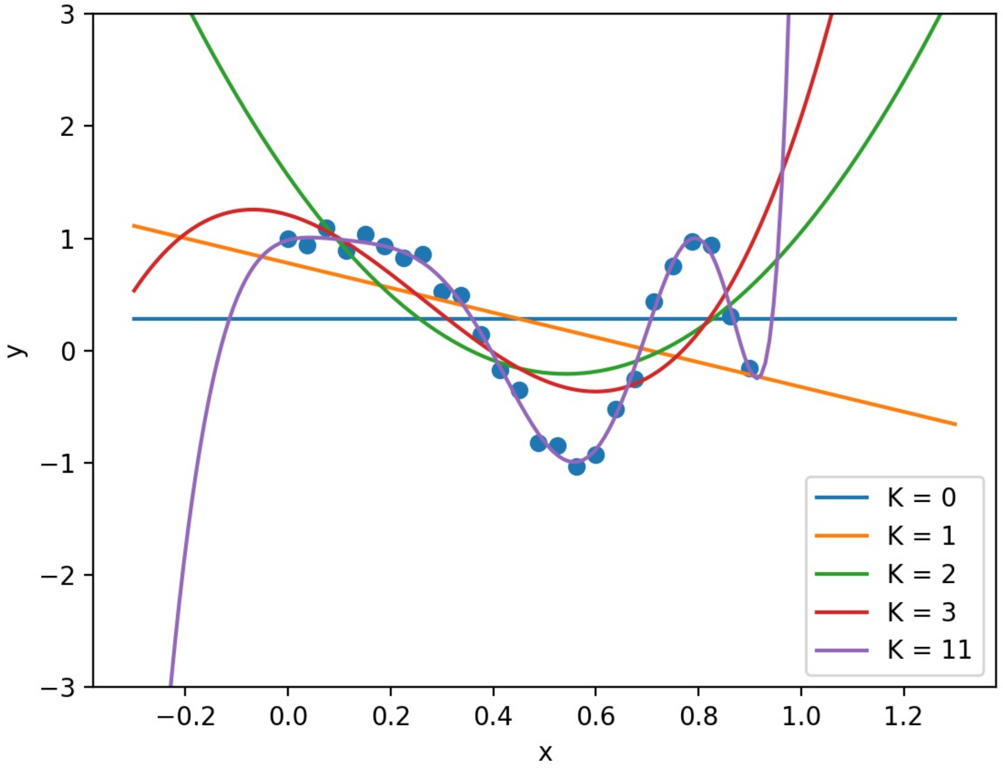

Figure 1: The plot showing the maximum likelihood estimation using polynomial basis functions of order K = 0, 1, 2, 3 and 11 along with the data

b) The plot below shows the maximum likelihood estimation using trigonometric basis functions, where:

$$\phi_0(x) = 1$$
$$\phi_{2j-1}(x) = sin(2\pi jx)$$
$$\phi_{2j}(x) = cos(2\pi jx)$$
$$for j = 1, 2, ...K$$

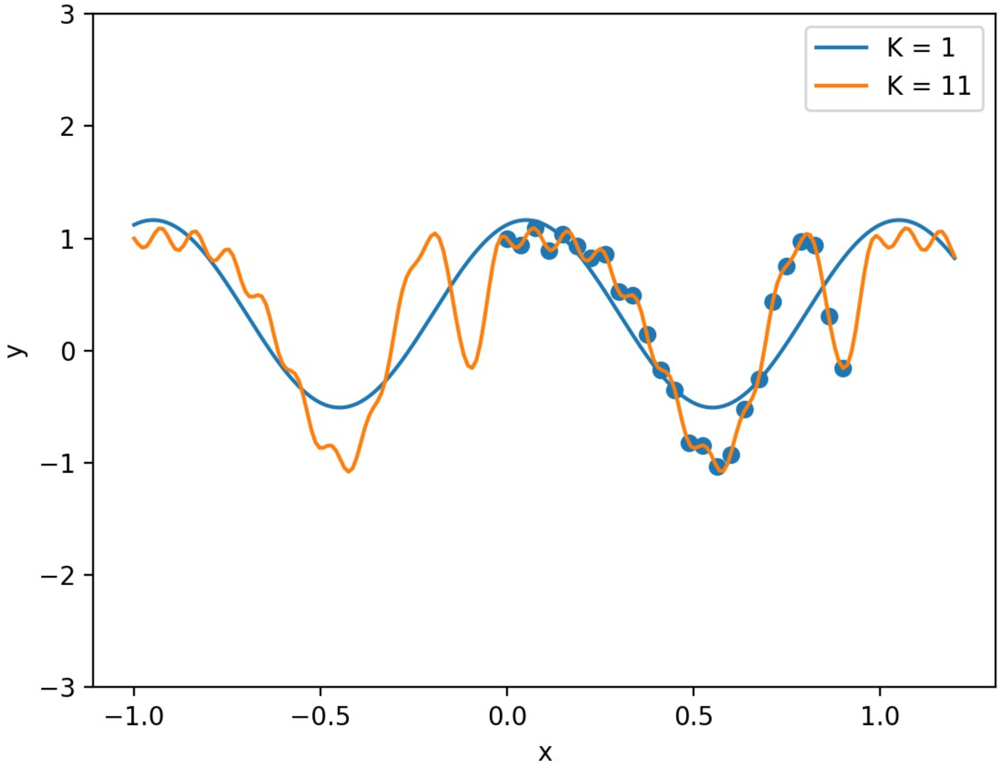

Figure 2: The plot showing the maximum likelihood estimation using trigonometric basis functions of order K = 1 and 11 along with the data

c) Average squared error and the maximum likelihood value of σ^2:

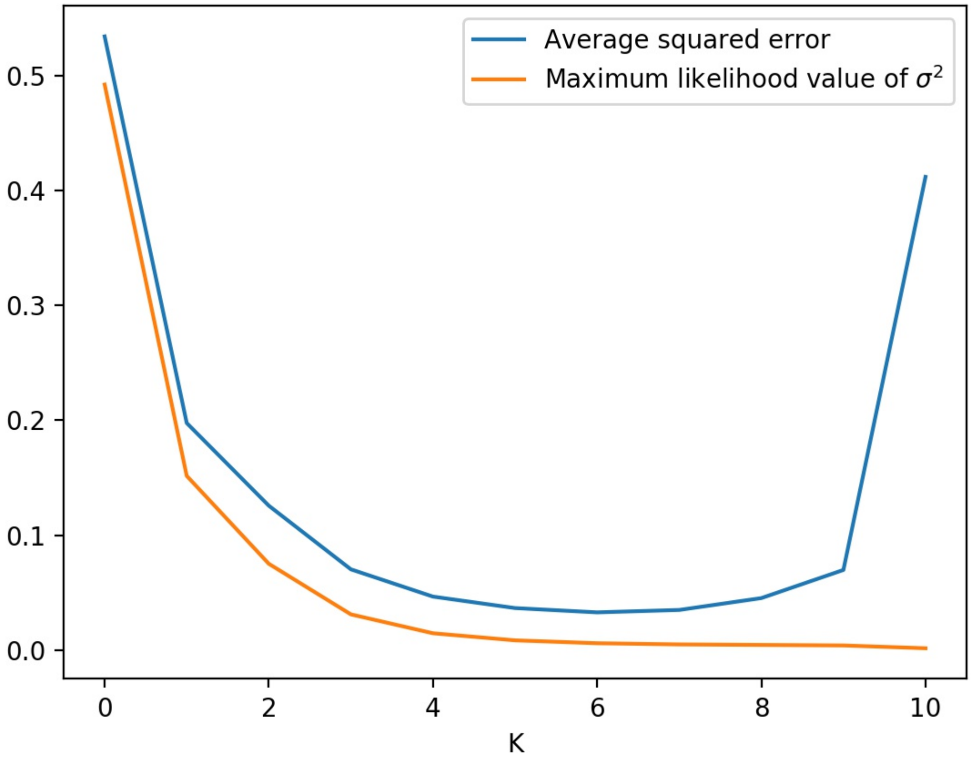

Figure 3: The average square error and the maximum likelihood value of σ^2 of the maximum likelihood estimation using trigonometric basis functions for K = 0 to K = 10, determined using leave-one-out cross validation on the 25 data points.

d) Over-fitting is a situation where the estimation model models to its training data too well, modelling even for the noise in the training data using high order basis functions. The result of over-fitting is that the model will have a very low error when tested against its original training data, however the error of the estimation of a new input will be high. Figure 3 shows the average square error of the cross validation and the maximum likelihood value of σ^2, which can be understood as the mean squared error of the estimation when tested against the training data. It can be seen that both lines rapidly decrease from K = 0 to K = 6, and as K further increases the average squared error of the cross validation starts to increase, while σ^2_{ML} continues to decrease as the error between the model and the training data continues to drop due to the order of the maximum likelihood estimator increases. Using Figure 2 as an example, when K=1, the model is under-fitting, as it doesn't represent the shape of the data good enough, and the model is over-fitting when K=11, the curve closely goes through most of the data points, the overall shape of the curve is however jagged as it is composed with high basis functions that models even to the noise present in the training data.

2 Question 2

a) The log-marginal likelihood is defined as:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\Phi},\alpha,\beta) &= \int p(\boldsymbol{y},\boldsymbol{w}|\boldsymbol{\Phi},\alpha,\beta)d\boldsymbol{w} \\ &= \int \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w},\beta\boldsymbol{I})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0},\alpha\boldsymbol{I})d\boldsymbol{w} \end{aligned} \tag{10}$$

Since both the likelihood and the prior are Gaussian, the marginal likelihood is therefore also Gaussian, it can be defined by finding the mean and the covariance of the Gaussian distribution:

$$p(\boldsymbol{y}|\boldsymbol{\Phi},\alpha,\beta) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{m_N},\boldsymbol{S_N}) \tag{11}$$

$$\boldsymbol{m_N} = \mathbb{E}[\boldsymbol{y}|\boldsymbol{\Phi},\alpha,\beta] = \mathbb{E}[\boldsymbol{\Phi}\boldsymbol{w}+\epsilon] = \boldsymbol{\Phi}\mathbb{E}[\boldsymbol{w}]+\boldsymbol{0} = \boldsymbol{0} \tag{12}$$

$$\boldsymbol{S_N} = Cov[\boldsymbol{y}|\boldsymbol{\Phi},\alpha,\beta] = var[\boldsymbol{\Phi}\boldsymbol{w}+\epsilon] = \boldsymbol{\Phi}var[\boldsymbol{w}]\boldsymbol{\Phi}^T+\beta\boldsymbol{I} = \boldsymbol{\Phi}\alpha\boldsymbol{I}\boldsymbol{\Phi}^T+\beta\boldsymbol{I} \tag{13}$$

Therefore the marginal likelihood is:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\Phi},\alpha,\beta) &= \mathcal{N}(\boldsymbol{y}|\boldsymbol{0},\boldsymbol{\Phi}\alpha\boldsymbol{I}\boldsymbol{\Phi}^T+\beta\boldsymbol{I}) \\ &= \mathcal{N}(\boldsymbol{y}|\boldsymbol{0},\alpha\boldsymbol{\Phi}\boldsymbol{\Phi}^T+\beta\boldsymbol{I}) \end{aligned} \tag{14}$$

And the log-marginal likelihood (lml) is:

$$f(\alpha,\beta) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\alpha\boldsymbol{\Phi}\boldsymbol{\Phi}^T+\beta\boldsymbol{I}| - \frac{1}{2}\boldsymbol{y}^T(\alpha\boldsymbol{\Phi}\boldsymbol{\Phi}^T+\beta\boldsymbol{I})^{-1}\boldsymbol{y} \tag{15}$$

The gradient with respect to α and β can be found by finding the derivative of the log-marginal likelihood with respect to $[\alpha,\beta]$:

$$\frac{\partial f}{\partial \alpha} = -\frac{1}{2}(tr(\Sigma^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T) - \boldsymbol{y}^T\Sigma^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T\Sigma^{-1}\boldsymbol{y}) \tag{16}$$

$$\frac{\partial f}{\partial \beta} = -\frac{1}{2}(tr(\Sigma^{-1}) - \boldsymbol{y}^T\Sigma^{-1}\Sigma^{-1}\boldsymbol{y}) \tag{17}$$

Where:

$$\Sigma = \alpha\boldsymbol{\Phi}\boldsymbol{\Phi}^T+\beta\boldsymbol{I}$$

b) The maximum log-marginal likelihood with the given dataset and the linear basis functions can be found using gradient descent. The following plot shows the gradient descent(ascent) with starting position $[\alpha = 0.5, \beta = 0.5]$, step size $= 0.005$ and 100 iterations:

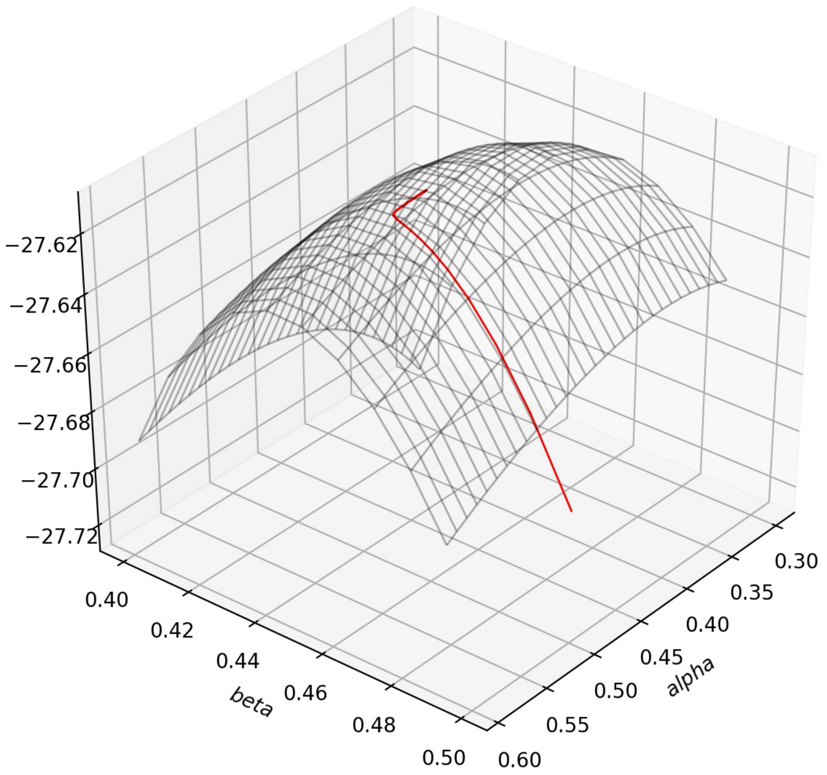

Figure 4: The log-marginal likelihood plot varying with α and β. The red line represents the gradient ascent steps with with starting position $[\alpha = 0.5, \beta = 0.5]$, step size $= 0.005$ and 100 iterations. The maximum is found at $[\alpha = 0.460, \beta = 0.448]$ with maximum log-marginal likelihood $= -27.6$.

c) Below shows a plot of maximum log-marginal likelihoods found by gradient ascent against the order of trigonometric basis functions. Comparing to Figure 3, the maximum marginal likelihood method finds the maximum at around K=4, while the maximum likelihood method method shows a decreasing error (variance) as K increases, which results in overfitting. The marginal likelihood method has the advantage naturally trading off between model complexity and data-fit, however the maximum likelihood method doesn't require a prediction of a prior, which, if the prior is incorrect can lead to wrong prediction using the maximum marginal likelihood method.

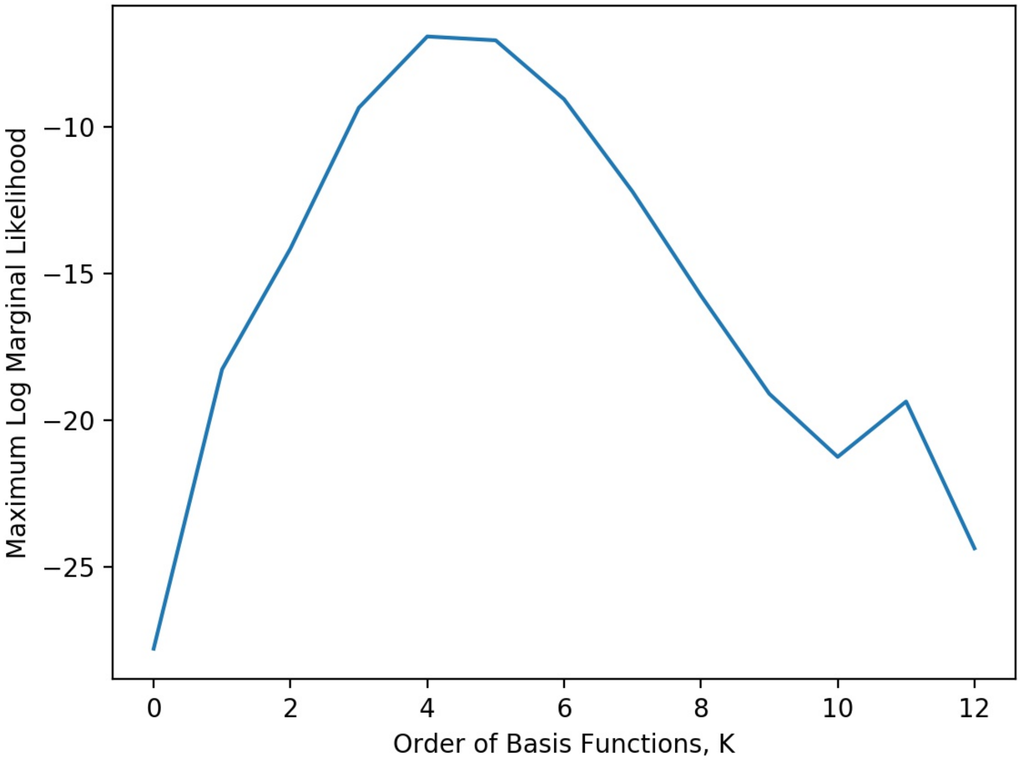

Figure 5: Maximum log-marginal likelihood plot against the order of trigonometric basis functions. Each maximum was found using gradient ascent with starting position $[\alpha = 0.25, \beta = 0.25]$, step size = 0.0001 and 10000 iterations.

d) The parameter posterior of a Gaussian likelihood and a Gaussian parameter prior is also Gaussian:

$$p(w|\boldsymbol{x},\boldsymbol{x}) = \mathcal{N}(w|\boldsymbol{m_N},\boldsymbol{S_N}) \tag{18}$$

Where:

$$\boldsymbol{S_N} = ((\alpha \boldsymbol{I})^{-1} + \frac{1}{\beta}\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1} \tag{19}$$

$$\boldsymbol{m_N} = \boldsymbol{S_N}(\frac{1}{\beta}\boldsymbol{\Phi}^T\boldsymbol{y}) \tag{20}$$

Using the parameter posterior distribution, the predictive mean and five sample sets pf parameters have been used to predict the $\boldsymbol{y}$ values with the give $\boldsymbol{x}$ values:

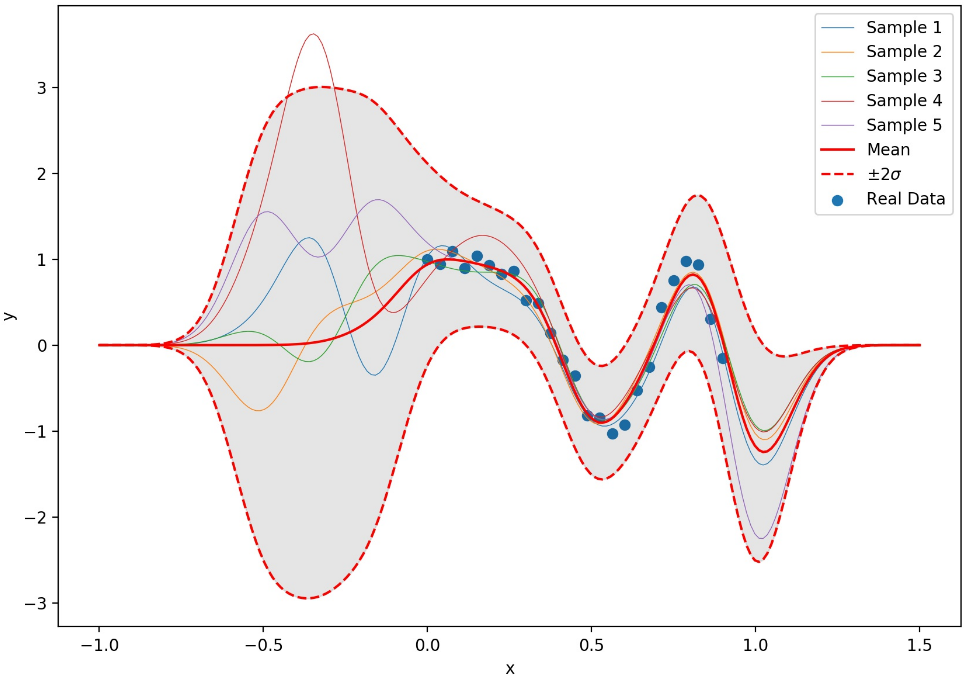

Figure 6: The plot of the real data, predictive mean (line in the centre in red), the region of ±2 standard deviation (shaded in grey), and five predictions using the sampled weights from the parameter posterior.