Imperial College
London

Coursework 2

Imperial College London

Department of Computing

CO496

Author:
Chengyi Zhang (CID: 01604256)

Date: November 5, 2018

1

1.1 Question 1a

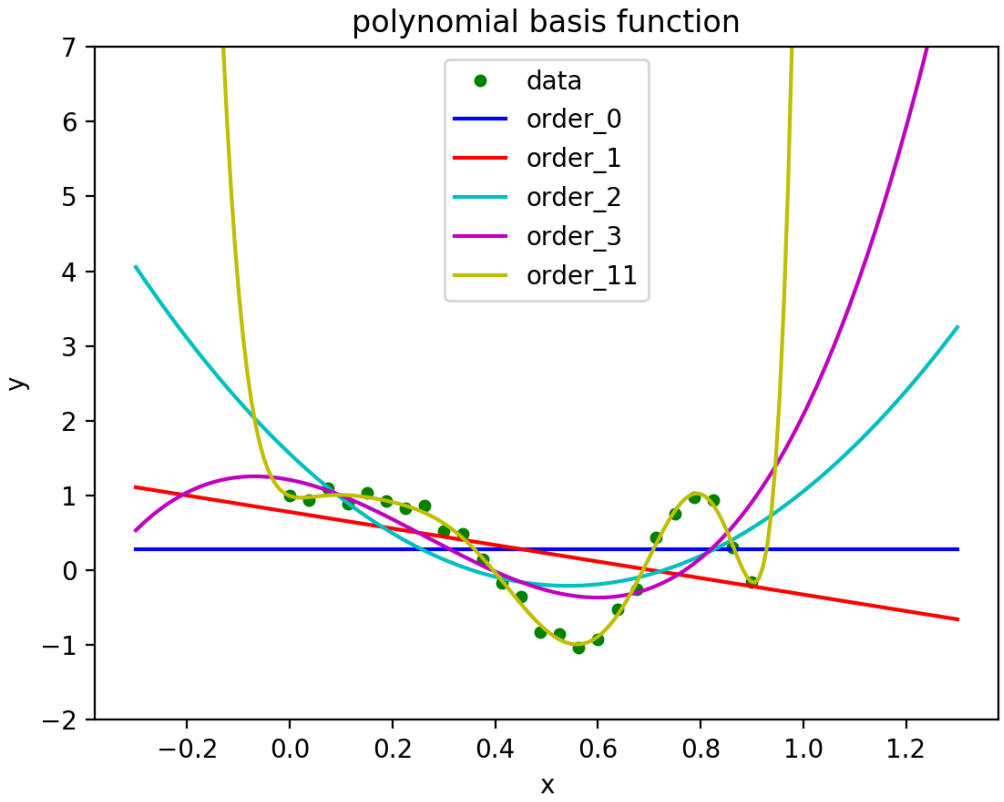

Figure 1: Fit the data using polynomial basis functions with various orders

1.2 Question 1b

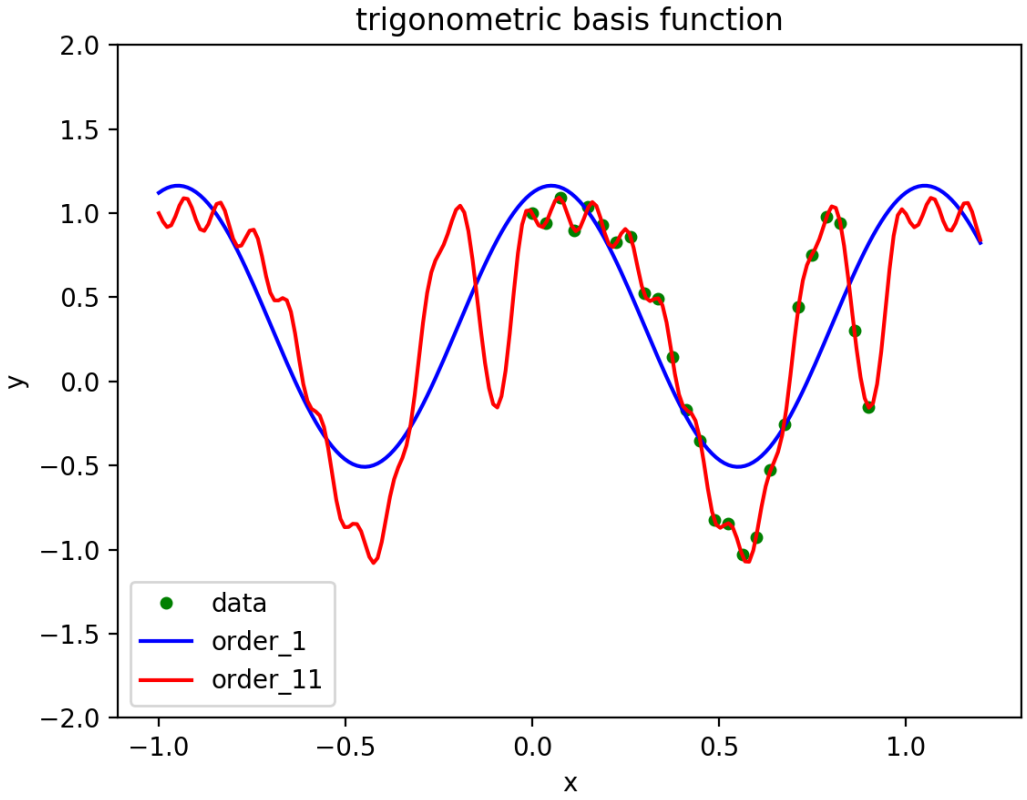

Figure 2: Fit the data using trigonometric basis functions with various orders

1.3 Question 1c

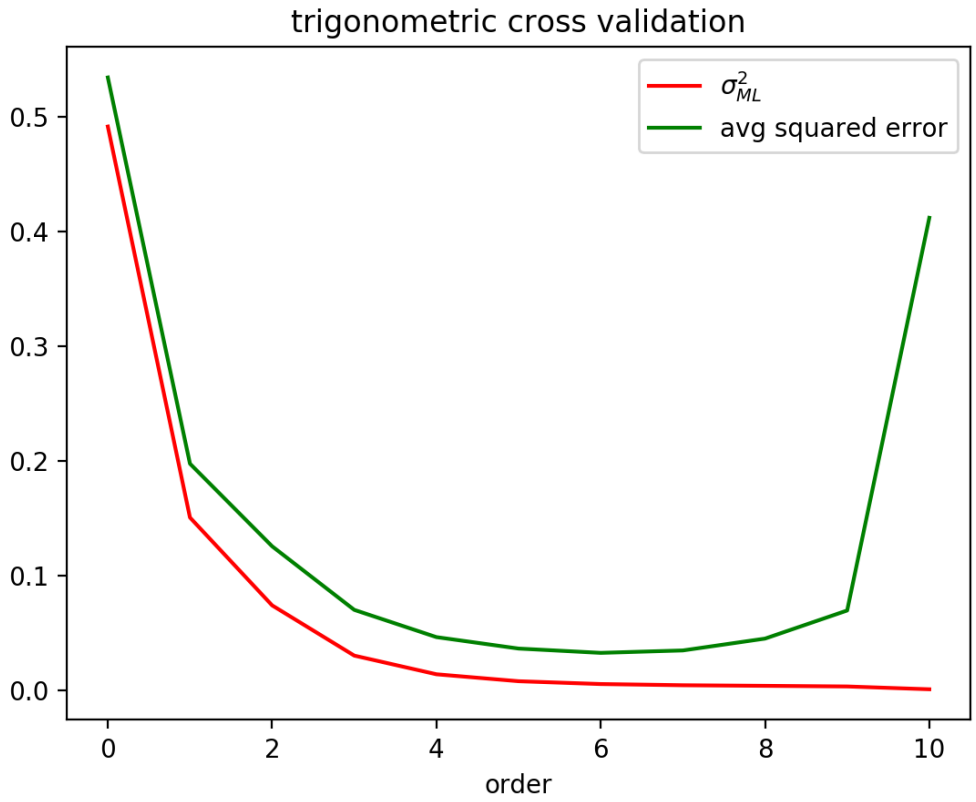

Figure 3: σ^2_{ML} and average squared error of order 0 to 10

1.4 Question 1d

When the model increases the order of parameters to fit the training data set better, σ^2_{ML} decreases. Meanwhile, however, the influence of noise is also getting stronger, causing the model to perform worse on testing data. When the model has more parameters than the data can express, the model starts to over-fit the training data set. Therefore, the average squared error on test data decreases at first, and then increases with the order.
We can determine the content above in the figures. In Fig. 1, the curves with higher order are obviously fitting the training data better. However, the curve with order 3 increases rapidly when $x > 0$, and the curve with order 11, which almost perfectly fits the training data, changes extremely fast when x is not in the domain of training data. Similarly, in Fig. 2, the curve with order 11 is strongly influenced by the noise and thus it fits the data perfectly.

2

2.1 Question 2a

Set $\boldsymbol{Y} = \begin{bmatrix} y_1 \\ ... \\ y_N \end{bmatrix}_{N\times 1}$, $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(x_1) \\ ... \\ \boldsymbol{\phi}(x_N) \end{bmatrix}_{N\times m}$

Prior:

$$p(\boldsymbol{w}) \sim N(0, b^2 \boldsymbol{I}_{m\times m}),\ b^2 = \frac{\sigma^2}{\lambda} \tag{1}$$

(Negative) log likelihood:

$$\begin{aligned}
\mathbb{L}(y_i|x_i,\boldsymbol{w}) &= -\log(\prod_{i=1}^{N} p(y_i|x_i,\boldsymbol{w})) && (2)\\
&= -\log(\prod_{i=1}^{N} p(y_i|\phi(x_i),\boldsymbol{w})) && (3)\\
&= -log(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y_i-\boldsymbol{w}^\top\boldsymbol{\phi}(x_i))^2}{2\sigma^2}\} && (4)\\
&= \frac{N}{2}\log 2\pi + \frac{N}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^\top\boldsymbol{\phi}(x_i))^2 && (5)\\
&= \frac{N}{2}\log 2\pi + \frac{N}{2}\log\sigma^2 + \frac{1}{2\sigma^2}(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{w})^\top(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{w}) && (6)
\end{aligned}$$

(Negative) log prior:

$$\begin{aligned}
\mathbb{L}(\boldsymbol{w}) &= -\log p(\boldsymbol{w}) && (7)\\
&= -\log((2\pi b^2)^{-\frac{m}{2}})\exp\{-\frac{\boldsymbol{w}^\top\boldsymbol{w}}{2b^2}\} && (8)\\
&= \frac{m}{2}\log 2\pi + \frac{m}{2}\log b^2 + \frac{\boldsymbol{w}^\top\boldsymbol{w}}{2b^2} && (9)
\end{aligned}$$

Differentiate the product of both log and solve:

$$\frac{\partial \mathbb{L}(y_i|x_i,\boldsymbol{w})\mathbb{L}(\boldsymbol{w})}{\partial \boldsymbol{w}} = \frac{1}{2\sigma^2}(-2\boldsymbol{\Phi}^\top\boldsymbol{Y} + 2\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{w}) + \frac{\boldsymbol{w}}{b^2} = 0 \tag{10}$$

$$(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \frac{\sigma^2}{b^2}\boldsymbol{I})\boldsymbol{w} = \boldsymbol{\Phi}^\top\boldsymbol{Y} \tag{11}$$

$$\boldsymbol{w} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \frac{\sigma^2}{b^2}\boldsymbol{I})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{Y} \tag{12}$$

Differentiate the loss function and solve:

$$L(\boldsymbol{x}) = (\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{w})^\top(\boldsymbol{Y}-\boldsymbol{\Phi}\boldsymbol{w}) + \lambda\boldsymbol{w}^\top\boldsymbol{w} \tag{13}$$

$$\frac{\partial L}{\partial \boldsymbol{w}} = -2\boldsymbol{\Phi}^\top \boldsymbol{Y} + 2\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\boldsymbol{w} + 2\lambda \boldsymbol{w} = 0 \tag{14}$$

$$(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \lambda \boldsymbol{I})\boldsymbol{w} = \boldsymbol{\Phi}^\top \boldsymbol{Y} \tag{15}$$

$$\boldsymbol{w} = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Phi}^\top \boldsymbol{Y} \tag{16}$$

With $b^2 = \frac{\sigma^2}{\lambda}$, the solutions are the same. Note that the scaling coefficient, $\frac{1}{2\sigma^2}$, and other constants in the product of log likelihood and log prior do not affect the solution.

The loss function has two parts. First, we calculate the sum of the squared error between the actual value y_i and the expected value of the corresponding input data x_i and our parameters $\boldsymbol{w}$. However, if we have parameters with very high order, it is likely that the model can fit training data and result perfectly while it may have bad performance on test data, and such situation is called over-fitting. Therefore, we add the second part, named regularization, as a penalty to the order of parameters. Since the square of any entry of parameter, w_i^2, is equal to or larger than zero, then with $\lambda > 0$, the value of the second part is non-decreasing (increasing in most circumstances where $w_i \neq 0$) when the order is increasing.

2.2 Question 2b

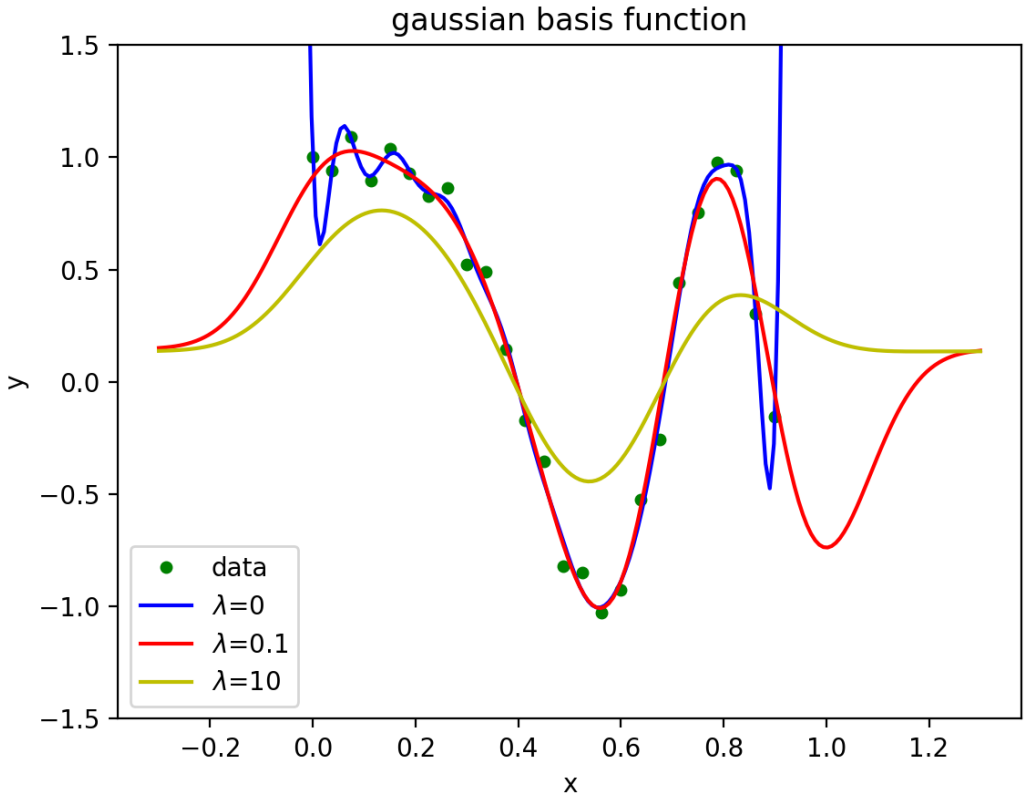

Figure 4: Ridge regression with different λ