Yinhong Qin
yq2021@nyu.edu

Homework 1

ECE 6143A-Machine Learning
September 21, 2021

Problem 1

1. The training set size I choose is 400. Thus the testing set size is 100. The partition method is random so the training set may vary between different training processes.

2. Please check the attached file "*ECE6143-HW1-Figures.pdf*" for the pictures of trained model overlaid on the testing set with different choices of orders.
   Note: The "order" here refers to the total number of terms in

$$f(x;\boldsymbol{\theta}) = \sum_{i=0}^{N} \theta_i x^i$$

   Therefore, the term with the highest power of x of a model with order d is

$$\theta_{d-1} x^{d-1}$$

3. Since the choice of both training and testing set may vary for each experiment, the best choice of d – the order of the polynomial model is not a constant value. Below are some results:

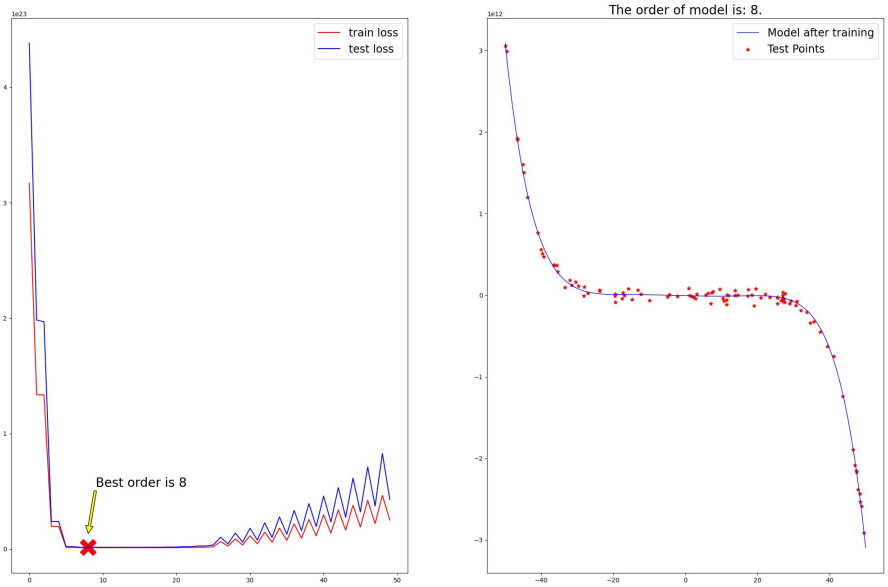

Figure 1.1 When testing loss reaches the minimum, order = 8.

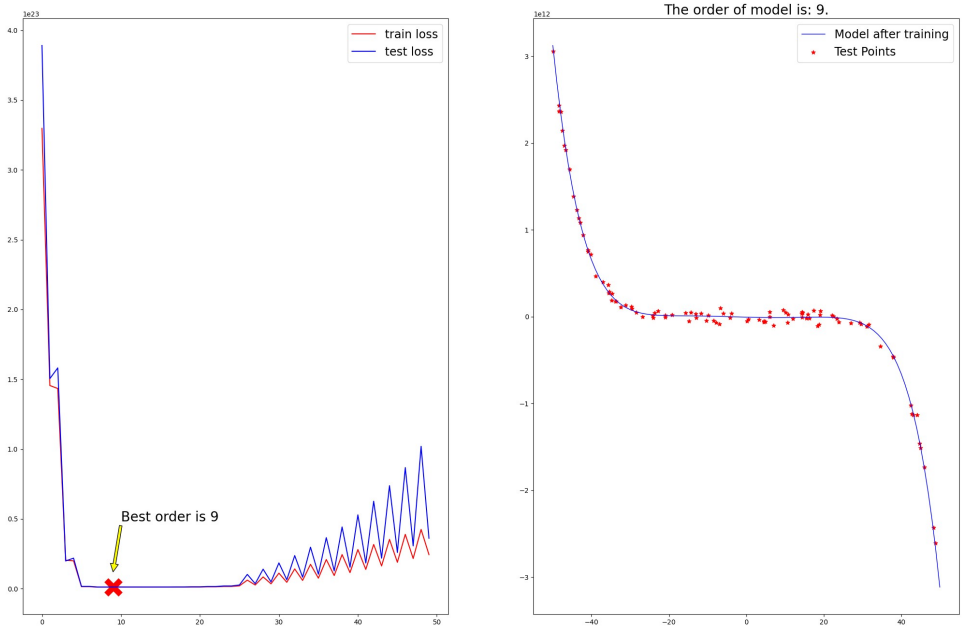

Figure 1.2 When testing loss reaches the minimum, order = 9.

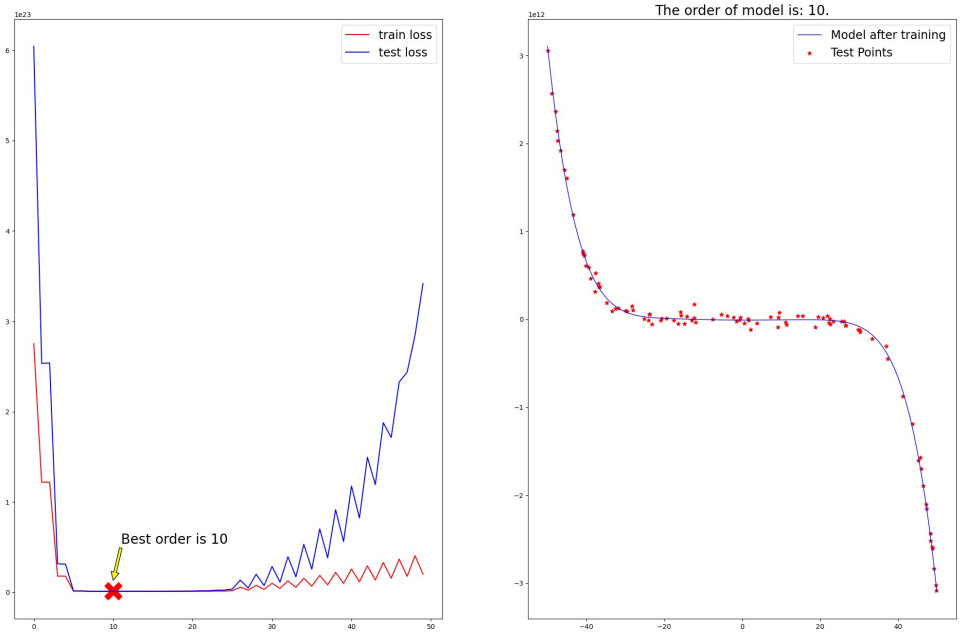

Figure 1.3 When testing loss reaches the minimum, order = 10.

Problem 2

1. The training set size I choose is 350. Therefore the size of testing set is 50. The partition method is random so the training set may vary between different training processes.

2. The training loss and testing loss curve is thus variable each time. The best choice of regularized factor is also different among different experiments and is depended on how the data set is partitioned. Below are some results:

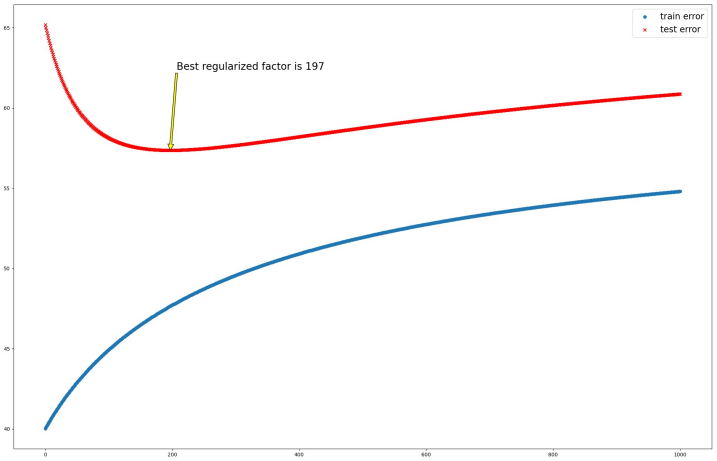

Figure 2.1 When testing loss reaches the minimum, $\lambda = 197$.

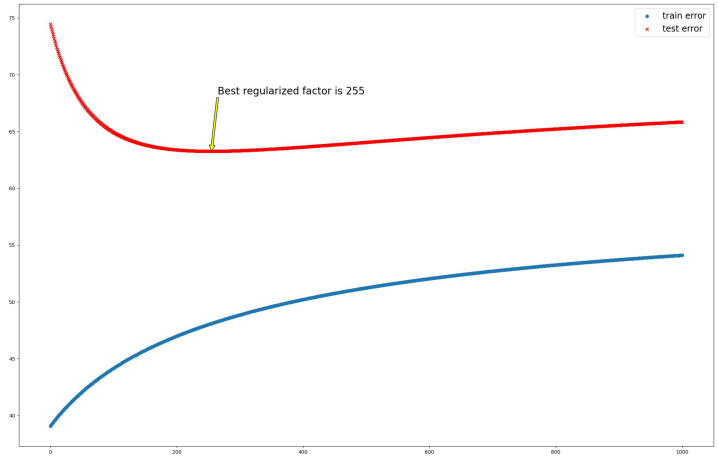

Figure 2.2 When testing loss reaches the minimum, $\lambda = 255$.

3. We can notice that when λ is increasing the testing loss will decrease and the training loss will increase.

Problem 3

1. Given the logistic squashing function:

$$g(z) = \frac{1}{1 + \exp(-z)}$$

Consider

$$\begin{aligned} g(-z) &= \frac{1}{1 + \exp(z)} \\ &= \frac{1 + \exp(z) - \exp(z)}{1 + \exp(z)} \\ &= 1 - \frac{\exp(z)}{1 + \exp(z)} \\ &= 1 - \frac{\exp(-z) \cdot \exp(z)}{\exp(-z) \cdot (1 + \exp(z))} \\ &= 1 - \frac{1}{1 + \exp(-z)} \\ &= 1 - g(z) \end{aligned}$$

2. Let

$$g(z) = \frac{1}{1 + \exp(-z)} = y$$

Then,

$$\begin{aligned} \exp(-z) &= \frac{1}{y} - 1 \\ z &= -\ln(\frac{1}{y} - 1) \\ z &= -\ln(\frac{1 - y}{y}) \\ z &= \ln(\frac{y}{1 - y}) \end{aligned}$$

Therefore,

$$g^{-1}(y) = \ln(\frac{y}{1 - y})$$

Problem 4

1. The derivative of logistic squashing function is:

$$\begin{aligned} g'(z) &= -\frac{1}{(1+\exp(-z))^2} * \frac{\mathrm{d}(1+\exp(-z))}{\mathrm{d}z} \\ &= -\frac{1}{(1+\exp(-z))^2} * (-\exp(-z)) \\ &= g(z)\frac{\exp(-z)}{1+\exp(-z)} \\ &= g(z)\frac{\exp(-z)+1-1}{1+\exp(-z)} \\ &= g(z)(1-g(z)) \end{aligned}$$

2. The partial derivative with respect to θ_i is:

$$\begin{aligned} \frac{\partial R_{emp}(\theta)}{\partial \theta_j} &= \frac{\partial \frac{1}{N}\sum_{i=1}^{N}(y_i-1)log(1-f(\mathbf{x}_i,\boldsymbol{\theta}))-y_i log(f(\mathbf{x}_i,\boldsymbol{\theta}))}{\partial \theta_j} \\ &= -\frac{1}{N}\sum_{i=1}^{N}[(y_i-1)\frac{1}{1-f(\mathbf{x}_i,\boldsymbol{\theta})}+y_i\frac{1}{f(\mathbf{x}_i,\boldsymbol{\theta})}]\frac{\partial f(\mathbf{x}_i,\boldsymbol{\theta})}{\partial \theta_j} \end{aligned}$$

Since

$$f(\mathbf{x}_i,\boldsymbol{\theta}) = \frac{1}{1+\exp(-\theta^T\mathbf{x})}$$

and

$$\begin{aligned} g'(z) &= \frac{\mathrm{d}\frac{1}{1+\exp(-z)}}{\mathrm{d}z} \\ &= g(z)(1-g(z)) \end{aligned}$$

Therefore,

$$\frac{\partial f(\mathbf{x}_i,\boldsymbol{\theta})}{\partial \theta_j} = f(\mathbf{x}_i,\boldsymbol{\theta})(1-f(\mathbf{x}_i,\boldsymbol{\theta}))\mathbf{x}_i^j$$

Then

$$\begin{aligned}\frac{\partial R_{emp}(\theta)}{\partial \theta_j} &= -\frac{1}{N}\sum_{i=1}^{N}[(y_i-1)\frac{1}{1-f(\mathbf{x}_i,\boldsymbol{\theta})}+y_i\frac{1}{f(\mathbf{x}_i,\boldsymbol{\theta})}](f(\mathbf{x}_i,\boldsymbol{\theta})(1-f(\mathbf{x}_i,\boldsymbol{\theta}))\mathbf{x}_i^j) \\ &= -\frac{1}{N}\sum_{i=1}^{N}[(y_i-1)f(\mathbf{x}_i,\boldsymbol{\theta})+y_i(1-f(\mathbf{x}_i,\boldsymbol{\theta}))]\mathbf{x}_i^j \\ &= \frac{1}{N}\sum_{i=1}^{N}(f(\mathbf{x}_i,\boldsymbol{\theta})-y_i)\mathbf{x}_i^j\end{aligned}$$

3. Here are the results of different choices of learning rate and tolerance:
(The results pictures are attached in "*ECE6143-HW1-Figures.pdf*". Please check that file.)

Learning rate	**Tolerance**	**Iterations**	**Binary classification error**
0.1	0.001	14434	6
1	0.001	40541	0
2	0.001	51016	0
10	0.001	75108	0
1	0.0001	751092	0
1	0.01	1458	6

It is shown that the value of learning rate and tolerance will influence the iterations until convergence and binary classification error. The iteration number will increase if the learning rate increases while with the same tolerance. It is because that a larger learning rate is more likely to "miss" the local minimum. When tolerance decreases, the iteration number will increase very quickly because it need more iterations to meet the convergence condition.