1 Refreshing Mathematics

Let $w \in \mathbb{R}^n$ is an n-dimensional column vector, and $f(w) \in \mathbb{R}$ is a function of w. In Lecture 2, we have defined the gradient $\nabla f(w) \in \mathbb{R}^n$ and Hessian matrix $H \in \mathbb{R}^{n\times n}$ of f with respect to w.

1. Let $f(w) = w^T Xb$, where $X \in \mathbb{R}^{n\times p}$ is a n × p matrix, and b is a p-dimensional column vector. Compute $\nabla f(w)$ using the definition of gradient.

$$\nabla f(w) = \begin{bmatrix} \frac{\partial f(w)}{w_1} & \cdots & \frac{\partial f(w)}{w_n} \end{bmatrix}^T \tag{1}$$

$$\begin{aligned}
\frac{\partial f(w)}{w_k} &= \frac{\partial}{\partial w_k}(\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}^T \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix}) \\
&= \frac{\partial}{\partial w_k}(b_1 \sum_{i=1}^{n} w_i x_{i1} + b_2 \sum_{i=1}^{n} w_i x_{i2} + \cdots + b_p \sum_{i=1}^{n} x_{ip}) \\
&= \frac{\partial}{\partial w_k} \sum_{j=1}^{p} \sum_{i=1}^{n} w_i x_j b_j \\
&= \sum_{i=1}^{p} b_i x_{ki}
\end{aligned} \tag{2}$$

Hence we have:

$$\begin{aligned}
\nabla f(w) &= \begin{bmatrix} \frac{\partial f(w)}{w_1} & \cdots & \frac{\partial f(w)}{w_n} \end{bmatrix}^T \\
&= \begin{bmatrix} \sum_{i=1}^{p} b_i x_{1i} & \cdots & \sum_{i=1}^{p} b_n x_{ni} \end{bmatrix}^T \\
&= Xb
\end{aligned} \tag{3}$$

2. Let $f(w) = tr(Bww^T A)$, where $A, B \in \mathbb{R}^{n\times n}$ are squared matrices of size $n \times n$, and $tr(A)$ is the trace of the squared matrix A. Using the definition of gradient, compute $\nabla f(w)$.

Again we have:

$$\nabla f(w) = \begin{bmatrix} \frac{\partial f(w)}{w_1} & \cdots & \frac{\partial f(w)}{w_n} \end{bmatrix}^T \tag{4}$$

$$\begin{aligned}
\frac{\partial f(w)}{w_k} &= \frac{\partial}{\partial w_k} tr(\begin{bmatrix} b_{11} & \cdots & b_{1n} \\ b_{21} & \cdots & b_{2n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nn} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}^T \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}) \\
&= \frac{\partial}{\partial w_k} tr(\begin{bmatrix} w_1 \sum_{i=1}^{n} b_{1i} w_i & \cdots & w_n \sum_{i=1}^{n} b_{1i} w_i \\ w_1 \sum_{i=1}^{n} b_{2i} w_i & \cdots & w_n \sum_{i=1}^{n} b_{2i} w_i \\ \vdots & \ddots & \vdots \\ w_1 \sum_{i=1}^{n} b_{ni} w_i & \cdots & w_n \sum_{i=1}^{n} b_{ni} w_i \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}) \\
&= \frac{\partial}{\partial w_k} (\sum_{p=1}^{n} \sum_{j=1}^{n} \sum_{i=1}^{n} b_{pi} w_i a_{jp} wj) \\
&= \sum_{p=1}^{n} \sum_{j=1}^{n} b_{pk} a_{jp} w_j + \sum_{p=1}^{n} \sum_{i=1}^{n} b_{pi} w_i a_{kp}
\end{aligned} \tag{5}$$

Hence,

$$\begin{aligned}
\nabla f(w) &= \begin{bmatrix} \frac{\partial f(w)}{w_1} & \cdots & \frac{\partial f(w)}{w_n} \end{bmatrix}^T \\
&= \begin{bmatrix} \sum_{j=1}^{n} \sum_{i=1}^{n} b_{i1} a_{ji} w_j \\ \vdots \\ \sum_{j=1}^{n} \sum_{i=1}^{n} b_{in} a_{ji} w_j \end{bmatrix} + \begin{bmatrix} \sum_{j=1}^{n} \sum_{i=1}^{n} b_{ij} a_{1i} w_j \\ \vdots \\ \sum_{j=1}^{n} \sum_{i=1}^{n} b_{ij} a_{ni} w_j \end{bmatrix} \\
&= B^T A^T w + ABw
\end{aligned} \tag{6}$$

3. Let $f(w) = tr(Bww^TA)$. Compute the Hessian matrix H of f with respect to w using the definition

$$H(w) = \begin{bmatrix} \frac{f}{\partial w_1^2} & \frac{f}{\partial w_1 w_2} & \cdots & \frac{f}{\partial w_1 w_n} \\ \frac{f}{\partial w_2 w_1} & \frac{f}{\partial w_2 w_2} & \cdots & \frac{f}{\partial w_2 w_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{f}{\partial w_n w_1} & \frac{f}{\partial w_n w_2} & \cdots & \frac{f}{\partial w_n w_n} \end{bmatrix} \tag{7}$$

From equation (5) we have:

$$\frac{\partial f}{\partial w_p} = \sum_{j=1}^{n}\sum_{i=1}^{n} b_{jp}a_{ij}w_i + \sum_{j=1}^{n}\sum_{i=1}^{n} b_{ji}a_{pj}w_i \tag{8}$$

Therefore:

$$\frac{f}{\partial w_p w_q} = \sum_{i=1}^{n} b_{ip}a_{qi} + \sum_{i=1}^{n} b_{iq}a_{pi} \tag{9}$$

Hence we can write $H(w)$ as:

$$\begin{aligned} H(w) &= \begin{bmatrix} \sum_{i=1}^n b_{i1}a_{1i} & \sum_{i=1}^n b_{i1}a_{2i} & \cdots & \sum_{i=1}^n b_{i1}a_{ni} \\ \sum_{i=1}^n b_{i2}a_{1i} & \sum_{i=1}^n b_{i2}a_{2i} & \cdots & \sum_{i=1}^n b_{i2}a_{ni} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n b_{in}a_{1i} & \sum_{i=1}^n b_{in}a_{2i} & \cdots & \sum_{i=1}^n b_{in}a_{ni} \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^n b_{i1}a_{1i} & \sum_{i=1}^n b_{i2}a_{1i} & \cdots & \sum_{i=1}^n b_{in}a_{1i} \\ \sum_{i=1}^n b_{i1}a_{2i} & \sum_{i=1}^n b_{i2}a_{2i} & \cdots & \sum_{i=1}^n b_{in}a_{2i} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n b_{i1}a_{ni} & \sum_{i=1}^n b_{i2}a_{ni} & \cdots & \sum_{i=1}^n b_{in}a_{ni} \end{bmatrix} \\ &= B^TA^T + AB \end{aligned} \tag{10}$$

4. If $A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, B = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$, Is $f(w)$ a convex function?:

If we want to prove $f(w)$ is a convex funciton, we need to prove $H(w)$ is positive-definite. From equation (10), we have:

$$H(w) = B^TA^T + AB = \begin{bmatrix} 2 & 4 \\ 4 & -6 \end{bmatrix} \tag{11}$$

Then for any $x = [x_1, x_2]^T$, we have:

$$\begin{aligned} x^TH(w)x &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 4 & -6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= 2x_1^2 + 8x_1x_2 - 6x_2^2 \end{aligned} \tag{12}$$

And we can find a counter example that when $x = [0, 1]$, $x^TH(w)x < 0$, hence $H(x)$ is not positive-definite and f(x) is not a convex function.

5. In Lecture 5, we have define the sigmoid function: $\sigma(a) = \frac{1}{1+e^{-a}}$, let $f(w) = \log(\sigma(w^Tx))$, where log is the natural logarithmic function. Compute $\nabla f(w)$ using the definition of gradient.

Again we have:

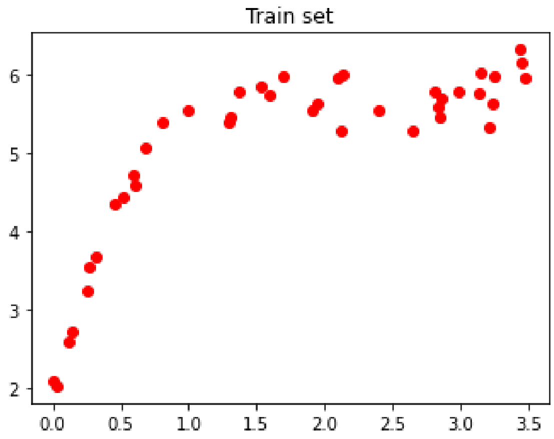
$$\begin{aligned} \nabla f(w) &= \begin{bmatrix} \frac{\partial f(w)}{w_1} & \cdots & \frac{\partial f(w)}{w_n} \end{bmatrix}^T \\ \frac{\partial f(w)}{w_k} &= \frac{\partial \log(\sigma(w^Tx))}{\partial w_k} \\ &= \frac{\partial \log(\sigma(w^Tx))}{\partial \sigma(w^Tx)} \frac{\partial \sigma(w^Tx))}{\partial w^Tx} \frac{\partial w^Tx}{\partial w_k} \qquad \text{(By chain rule)} \\ &= \sigma(w^Tx)^{-1} \frac{e^{-w^Tx}}{(1+e^{-w^Tx})^2} x_k \\ &= \frac{e^{-w^Tx}}{1+e^{-w^Tx}} x_k \\ &= (1 - \sigma(w^Tx))x_k \end{aligned} \tag{13}$$

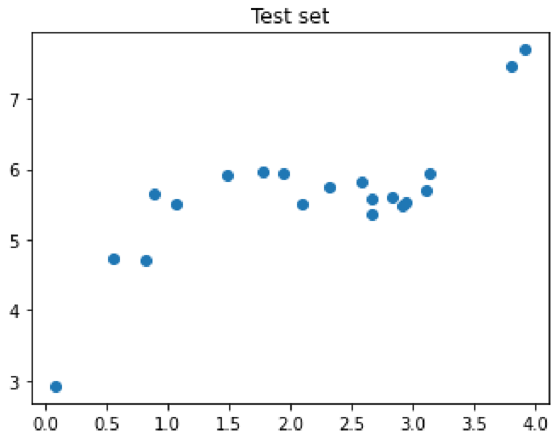
Therefore, we can write $\nabla f(w)$ as:

$$\begin{aligned}\nabla f(w) &= \begin{bmatrix} \frac{\partial f(w)}{w_1} & \cdots & \frac{\partial f(w)}{w_n} \end{bmatrix}^T \\ &= (1 - \sigma(w^T x)) \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \\ &= (1 - \sigma(w^T x))x \end{aligned} \tag{14}$$

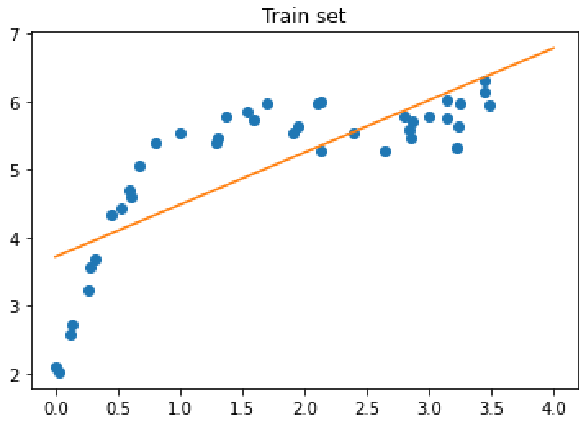
2 Linear and Polynomial Regression

1. Load the training data hw1xtr.dat and hw1ytr.dat into the memory and plot it on one graph. Load the test data hw1xte.dat and hw1yte.dat into the memory and plot it on another graph.



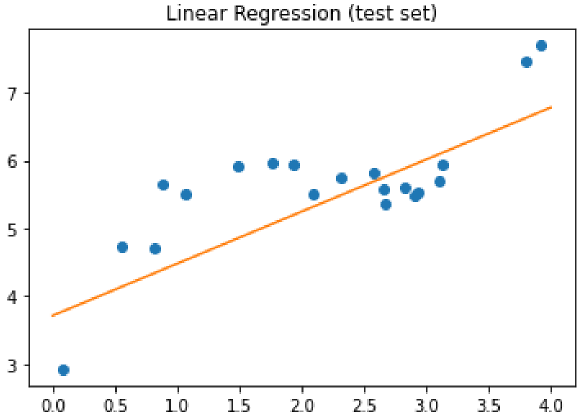


2. Add a column vector of 1's to the features, then use the linear regression formula discussed in Lecture 3 to obtain a 2-dimensional weight vector. Plot both the linear regression line and the training data on the same graph. Also report the average error on the training set using Eq. (1).



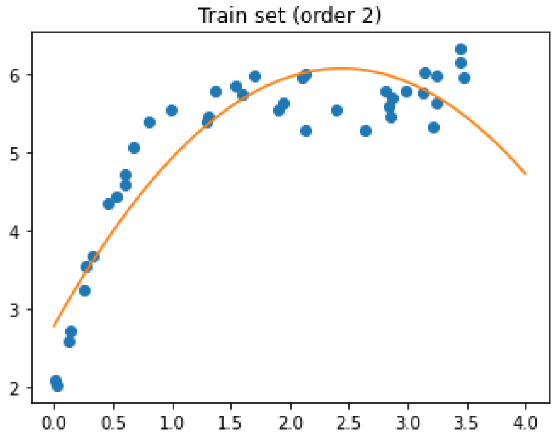
The average error on the train set is 0.508589.

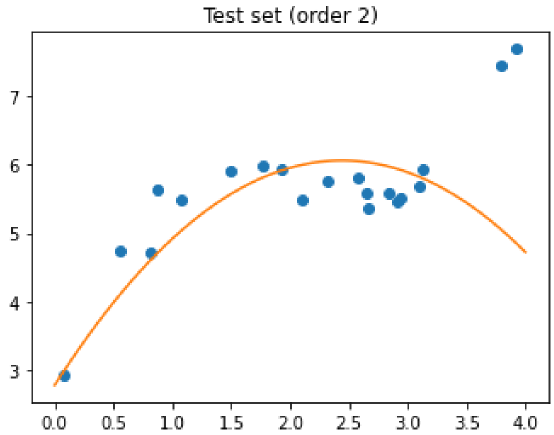
3. Plot both the regression line and the test data on the same graph. Also report the average error on the test set using Eq. (1).



The average error on the test set is 0.443912.

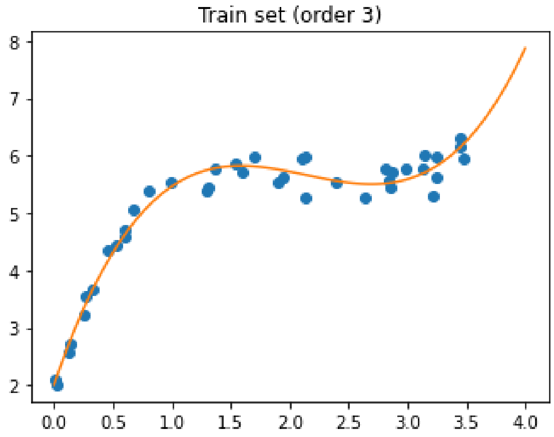
4. Implement the 2nd-order polynomial regression by adding new features x2 to the inputs. Repeat (b) and (c). Compare the training error and test error. Is it a better fit than linear regression?

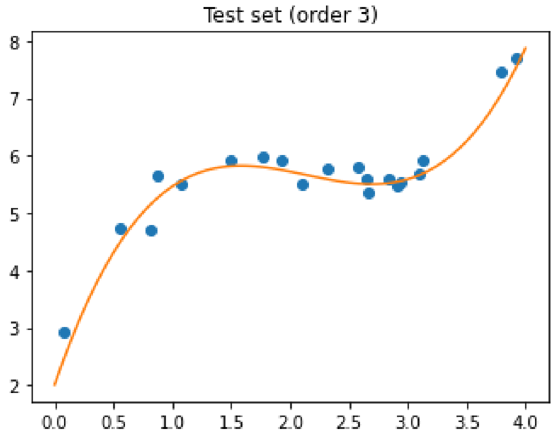




The average error on the train set is 0.200985, the average error on the test set is 0.853263. Compare to the linear regression, the 2nd-order polynomial regression fit better on training set but not on testing set. Hence it is not a better fit.

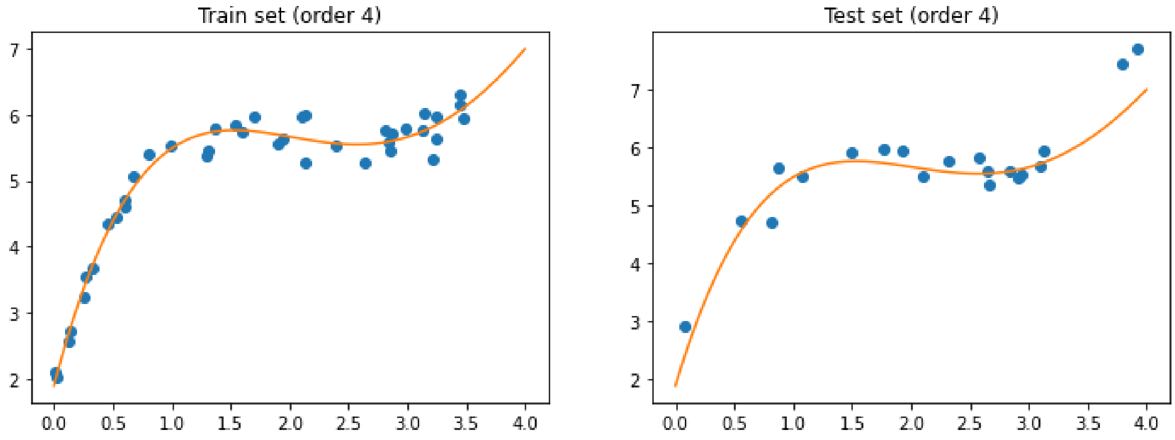
5. Implement the 3rd-order polynomial regression by adding new features x2,x3 to the inputs. Repeat (b) and (c). Compare the training error and test error. Is it a better fit than linear regression and 2nd-order polynomial regression?





The average error on the train set is 0.039229, the average error on the test set is 0.056418. Compare to the 2nd-order polynomial regression, the 3rd-order polynomial regression fit better both on training set and testing set, hence it is indeed a better fit.

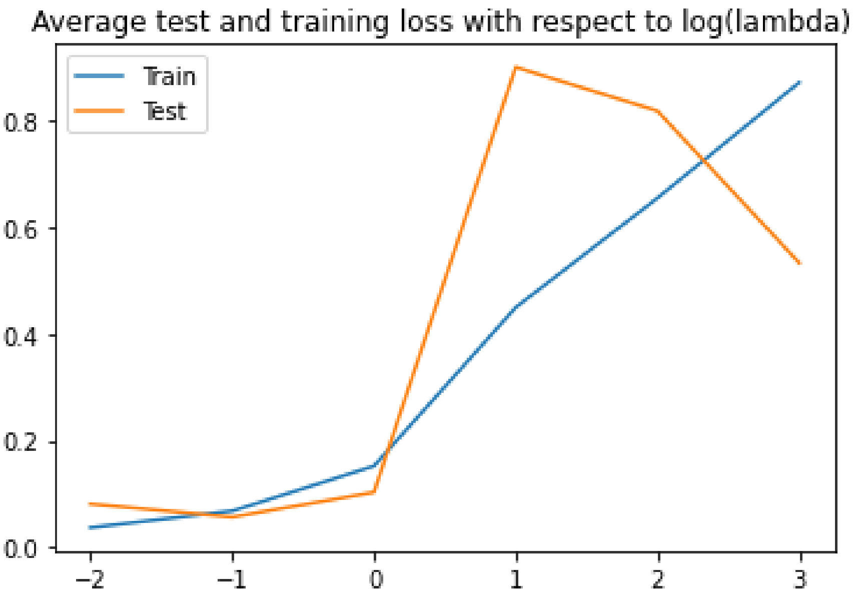
6. Implement the 4th-order polynomial regression by adding new features x2,x3,x4 to the inputs. Repeat (b) and (c). Compare the training error and test error. Compared with the previous results, which order is the best for fitting the data?



The average error on the train set is 0.035645, the average error on the test set is 0.127222. Compare to previous data, the 3rd-order polynomial regression fit best on test set and has a good performance on training set. Hence the 3rd-order polynomial regression is the best fit.

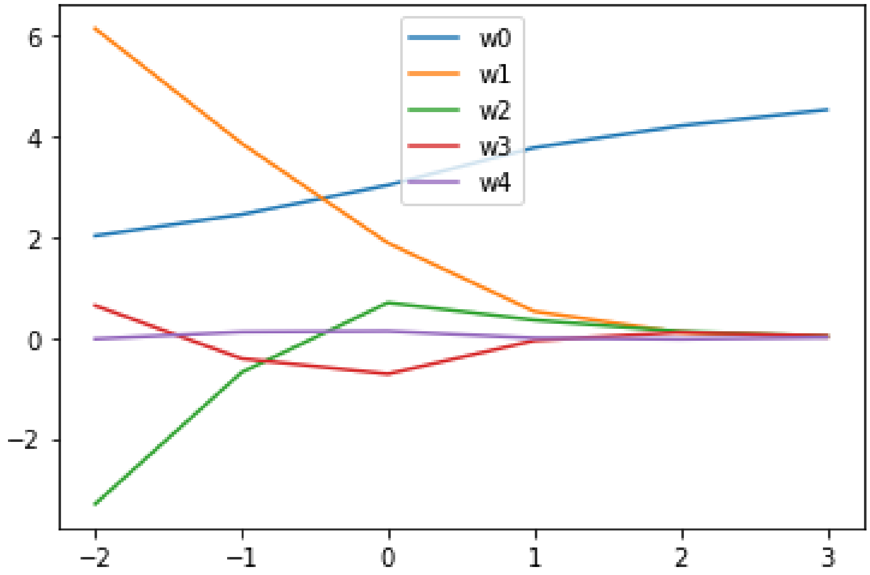
3 Regularization and Cross-Validation

1. Using the training data to implement l2-regularized for the 4th-order polynomial regression (page 12 of Lecture 4, note that we do not penalize the bias term w0), vary the regularization parameter $\lambda \in \{0.01, 0.1, 1, 10, 100, 1000\}$. Plot the training and test error (averaged over all in- stances) using Eq. (1) as a function of λ (you should use a log10 scale for λ). Which λ is the best for fitting the data?



From the above graph, we can see that when $\lambda = 0.1$, the 4-th order polynomial regression has the lowest test data. Hence $\lambda = 10$ is the best λ.

2. Plot the value of each weight parameter (including the bias term w0) as a function of λ.



3. Write a procedure that performs five-fold cross-validation on your training data (page 7 of Lecture 4). Use it to determine the best value for λ. Show the average error on the validation set as a function of λ. Is the same as the best λ in (a)? For the best fit, plot the test data and the l2-regularized 4th-order polynomial regression line obtained.