Machine Learning HW2

Wenhan Yang

Due: March 10 2021

Question 1

1.

$\because \epsilon \sim N(0, \sigma^2)$, i.i.d. and independent of X
$\therefore \mathbb{E}(\epsilon|X) = \mathbb{E}(\epsilon) = 0, Var(\epsilon|X) = Var(\epsilon) = \sigma^2$

$$\begin{aligned}
\mathbb{E}_{\epsilon|X}[\hat{\omega}] &= \mathbb{E}_{\epsilon|X}[(X^TX)^{-1}X^Ty] \\
&= \mathbb{E}_{\epsilon|X}((X^TX)^{-1})\mathbb{E}_{\epsilon|X}(X^T)\mathbb{E}_{\epsilon|X}(y) \\
&= (X^TX)^{-1}X^T\mathbb{E}_{\epsilon|X}(X\omega + \epsilon) \\
&= (X^TX)^{-1}X^T\Big(\mathbb{E}_{\epsilon|X}(X\omega) + \mathbb{E}_{\epsilon|X}(\epsilon)\Big) = (X^TX)^{-1}X^T(X\omega + \mathbb{E}(\epsilon|X)) \\
&= (X^TX)^{-1}X^T(X\omega + 0) = (X^TX)^{-1}X^TX\omega = \omega
\end{aligned}$$

$\Longrightarrow \mathbb{E}_{\epsilon|X}[\hat{\omega}] - \omega = 0$, it is an unbiased estimator of ω.

2.

Note: X^TX is symmetric matrix, then its transposition is itself. Variance of a constant value is 0.

$$\begin{aligned}
Var[\hat{\omega}] &= Var[(X^TX)^{-1}X^Ty] \\
&= ((X^TX)^{-1}X^T)Var(y)((X^TX)^{-1}X^T)^T \\
&= (X^TX)^{-1}X^TVar(X\omega + \epsilon)X(X^TX)^T \\
&= (X^TX)^{-1}X^TVar(\epsilon)X(X^TX) \\
&= \sigma^2(X^TX)^{-1}X^TX(X^TX) = \sigma^2(X^TX)^{-1}
\end{aligned}$$

3.

Note: X^TX is symmetric matrix, then it is positive definite. $(X^TX + \lambda I)^{-1}$ is also positive definite because $X^TX + \lambda I$ is symmetric and thus positive definite; the inverse of a positive definite matrix is also positive definite (and also symmetric). Given $\hat{\omega}_r(\lambda) = (X^TX + \lambda I)^{-1}X^Ty$:

$$\begin{aligned}
Var[\hat{\omega}_r(\lambda)] &= Var[(X^TX + \lambda I)^{-1}X^Ty] \\
&= ((X^TX + \lambda I)^{-1}X^T)Var(y)((X^TX + \lambda I)^{-1}X^T)^T \\
&= (X^TX + \lambda I)^{-1}X^TVar(X\omega + \epsilon)X((X^TX + \lambda I)^{-1})^T \\
&= (X^TX + \lambda I)^{-1}X^TVar(\epsilon)X(X^TX + \lambda I)^{-1} \\
&= \sigma^2(X^TX + \lambda I)^{-1}(X^TX)(X^TX + \lambda I)^{-1} \\
Var[\hat{\omega}] &= \sigma^2(X^TX)^{-1}
\end{aligned}$$

Let $W = (X^TX)(X^TX + \lambda I)^{-1}$. Then:

$W^T = ((X^TX)(X^TX + \lambda I)^{-1})^T = (X^TX + \lambda I)^{-1}(X^TX)$
$W^{-1} = ((X^TX)(X^TX + \lambda I)^{-1}) = (X^TX + \lambda I)(X^TX)^{-1} = I + \lambda(X^TX)^{-1}$
$(W^T)^{-1} = ((X^TX + \lambda I)^{-1}(X^TX))^{-1} = (X^TX)^{-1}(X^TX + \lambda I) = I + \lambda(X^TX)^{-1}$

Find difference between $Var[\hat{\omega}]$ and $Var[\hat{\omega}_r(\lambda)]$:

$$\begin{aligned}
Var[\hat{\omega}] - Var[\hat{\omega}_r(\lambda)] &= \sigma^2(X^TX)^{-1} - \sigma^2(X^TX + \lambda I)^{-1}(X^TX)(X^TX + \lambda I)^{-1} \\
&= \sigma^2\Big((X^TX)^{-1} - (X^TX + \lambda I)^{-1}(X^TX)I(X^TX + \lambda I)^{-1}\Big) \\
&= \sigma^2\Big((X^TX)^{-1} - (X^TX + \lambda I)^{-1}(X^TX)(X^TX)^{-1}(X^TX)(X^TX + \lambda I)^{-1}\Big) \\
&= \sigma^2\Big((X^TX)^{-1} - \underline{\underline{((X^TX)(X^TX + \lambda I)^{-1})^T}}(X^TX)^{-1}\underline{\underline{((X^TX)(X^TX + \lambda I)^{-1})}}\Big) \\
&= \sigma^2\Big((X^TX)^{-1} - W^T(X^TX)^{-1}W\Big) \\
&= \sigma^2\Big(W^T(W^T)^{-1}(X^TX)^{-1}W^{-1}W - W^T(X^TX)^{-1}W\Big) \\
&= \sigma^2W^T\Big((W^T)^{-1}(X^TX)^{-1}W^{-1} - (X^TX)^{-1}\Big)W \\
&= \sigma^2W^T\Big((I + \lambda(X^TX)^{-1})(X^TX)^{-1}(I + \lambda(X^TX)^{-1}) - (X^TX)^{-1}\Big)W \\
&= \sigma^2W^T\Big((X^TX)^{-1} + \lambda(X^TX)^{-2} + \lambda(X^TX)^{-2} + \lambda^2(X^TX)^{-3} - (X^TX)^{-1}\Big)W \\
&= \sigma^2W^T\Big(2\lambda(X^TX)^{-2} + \lambda^2(X^TX)^{-3}\Big)W \\
&= \sigma^2(X^TX + \lambda I)^{-1}(X^TX)\Big(2\lambda(X^TX)^{-2} + \lambda^2(X^TX)^{-3}\Big)(X^TX)(X^TX + \lambda I)^{-1} \\
&= \sigma^2(X^TX + \lambda I)^{-1}\Big(2\lambda I + \lambda^2(X^TX)^{-1}\Big)(X^TX + \lambda I)^{-1}
\end{aligned}$$

Here, $(X^TX + \lambda I)^{-1}$ and $2\lambda I + \lambda^2(X^TX)^{-1}$ are both positive definite matrix. So, the product $(X^TX + \lambda I)^{-1}\Big(2\lambda I + \lambda^2(X^TX)^{-1}\Big)(X^TX + \lambda I)^{-1}$ is positive definite. For $\sigma \geq 0$, $Var[\hat{\omega}] - Var[\hat{\omega}_r(\lambda)] \geq 0$. Therefore, $Var[\hat{\omega}] \geq Var[\hat{\omega}_r(\lambda)]$

Question 2

1.

$y = X\theta + \epsilon$
Lasso Regression: $J(\theta) = ||X\theta - y||_2^2 + \lambda||\theta||_1$

$$\begin{aligned} J(\hat{\theta}) &= ||X\hat{\theta} - \hat{y}||_2^2 + \lambda||\hat{\theta}||_1 = ||X\hat{\theta} - (X\hat{\theta} + \epsilon)||_2^2 + \lambda||\hat{\theta}||_1 \\ &= ||\epsilon||_2^2 + \lambda||\hat{\theta}||_1 \\ &= \sum_{i=1}^{n} \epsilon_i^2 + \lambda|a| + \lambda|b| + \lambda||r||_1 \\ &= \sum_{i=1}^{n} \epsilon_i^2 + \lambda(|a| + |b|) + \lambda||r||_1 \end{aligned}$$

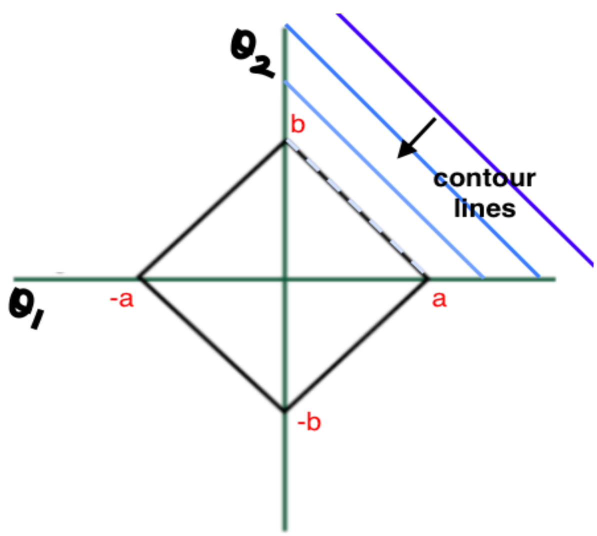
Note that the term $|a| + |b|$ satisfies $|a| + |b| \geq |a + b|$. Equality in triangular inequality holds if and only if numbers have the same sign or one is zero. So on the right hand side, when at least one of $|a|$ or $|b|$ is 0, the left hand side reaches its minimum $|a|$ or $|b|$, and J(θ) is also minimized. Take a=0, then:

$$J(\hat{\theta})_1 = \sum_{i=1}^{n} \epsilon_i^2 + \lambda|b| + \lambda||r||_1$$

where J(θ) is minimized, also the same when b=0:

$$J(\hat{\theta})_2 = \sum_{i=1}^{n} \epsilon_i^2 + \lambda|a| + \lambda||r||_1$$

$$\therefore J(\hat{\theta})_1 = J(\hat{\theta})_2 \longrightarrow |a| = |b| \longrightarrow |a + b| \leq 2|a| = 2|b|$$



In this special case, the contour lines are parallel to the $||\theta||_1$ edge, and here a and b have the same sign and the loss function is minimized on the line segment $a + b = 2|a| = 2|b|$.

If $(c, d, r^T)^T$ is another minimizer, then c+d = a+b, a≠c, b≠d, c and d have the same sign. But if a=b=0, then the norm edge shrink to a point(origin) and the solution is unique.

2.

Ridge Regression: $J(\theta) = ||X\theta - \hat{y}||_2^2 + \lambda||\theta||_2^2 = (y - X\theta)^T(y - X\theta) + \lambda\theta^T\theta$

Take derivative: $\nabla_\theta J(\theta) = -2X^T(y - X\theta) + 2\lambda\theta = 0$

get $\hat{\theta} = (X^TX + \lambda I)^{-1}X^Ty$

$$\text{which is } \begin{pmatrix} a \\ b \\ r \end{pmatrix} = \left[\begin{pmatrix} x_1 \\ x_2 \\ X_r \end{pmatrix}\begin{pmatrix} x_1 & x_2 & X_r \end{pmatrix} + \lambda I\right]^{-1}\begin{pmatrix} x_1 \\ x_2 \\ X_r \end{pmatrix}\left[\epsilon + \begin{pmatrix} \theta_1 & \theta_2 & \theta_r \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ X_r \end{pmatrix}\right]$$

$$= \begin{bmatrix} x_1^2 + \lambda & x_1x_2 & x_1X_r \\ x_1x_2 & x_2^2 + \lambda & x_2X_r \\ x_1X_r & x_2X_r & X_r^2 + \lambda \end{bmatrix}^{-1}\begin{pmatrix} x_1 \\ x_2 \\ X_r \end{pmatrix}\left[\epsilon + \theta_1x_1 + \theta_2x_2 + \theta_rX_r\right]$$

$$= A^{-1}\begin{pmatrix} x_1 \\ x_2 \\ X_r \end{pmatrix}\left[\epsilon + \theta_1x_1 + \theta_2x_2 + \theta_rX_r\right]$$

Note that A is symmetric. $\because x_1 = x_2 \therefore x_1^2 + \lambda = x_2^2 + \lambda$ and $x_1x_2 = x_2x_1$

$$\therefore A = \begin{bmatrix} a_{11} & a_{12} & ... & ... \\ a_{12} & a_{11} & ... & ... \\ : & : & & \\ : & : & S & \\ : & : & & \end{bmatrix}, \text{S is a square matrix}$$

$$\text{Let } B = A^{-1} = \frac{1}{|A|}\begin{bmatrix} a_{11}^* & a_{12}^* & ... & ... \\ a_{12}^* & a_{11}^* & ... & ... \\ : & : & & \\ : & : & S^* & \\ : & : & & \end{bmatrix} := \begin{bmatrix} b_{11} & b_{12} & ... & ... \\ b_{12} & b_{11} & ... & ... \\ : & : & & \\ : & : & S_b & \\ : & : & & \end{bmatrix}$$

$$\therefore \begin{pmatrix} a \\ b \\ r \end{pmatrix} = \begin{bmatrix} b_{11} & b_{12} & -b_{1n}- \\ b_{12} & b_{11} & -b_{2n}- \\ : & : & \\ : & : & S_b \\ : & : & \end{bmatrix}\begin{pmatrix} x_1 \\ x_2 \\ X_r \end{pmatrix}\left[\epsilon + \theta_1x_1 + \theta_2x_2 + \theta_rX_r\right]$$

$$\therefore a = (b_{11}x_1 + b_{12}x_2 + (-b_{1n}-)X_r)(\epsilon + \theta_1x_1 + \theta_2x_2 + \theta_rX_r)$$
$$b = (b_{11}x_1 + b_{12}x_2 + (-b_{2n}-)X_r)(\epsilon + \theta_1x_1 + \theta_2x_2 + \theta_rX_r)$$

$\because x_1 = x_2$

Note that $-b_{1n}- = (x_1X_r)^* = (x_2X_r)^* = -b_{2n}-$

$\therefore a = b$

Question 3

1.

The standardization parameter are saved in $mean_std.pk$:

```
with open('mean_std.pk','rb') as read_file:
    df = pickle.load(read_file)
```

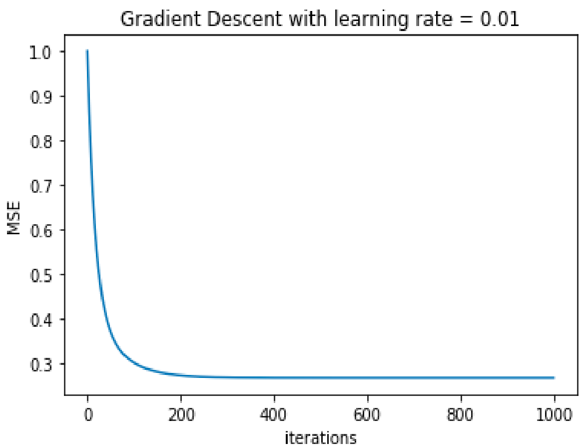
```
df
```

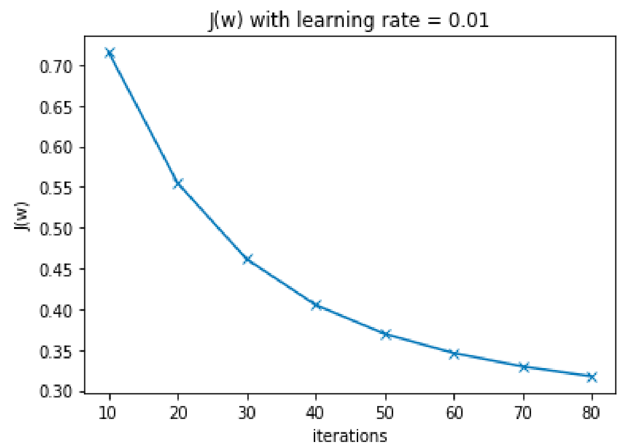
```
{'area': {'mean': 2000.6808510638298, 'std': 786.2026187430467},
 'n_bedroom': {'mean': 3.1702127659574466, 'std': 0.7528428090618782},
 'price': {'mean': 340412.6595744681, 'std': 123702.53600614739}}
```

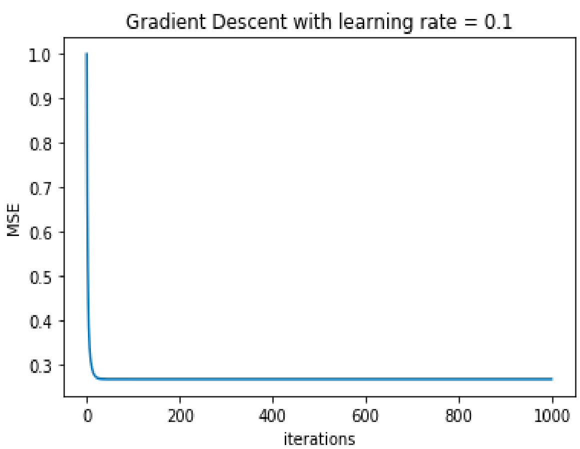
2.

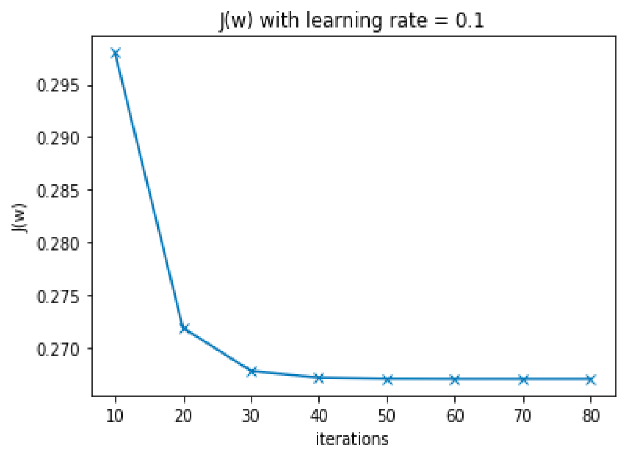
One example outcome of w is:

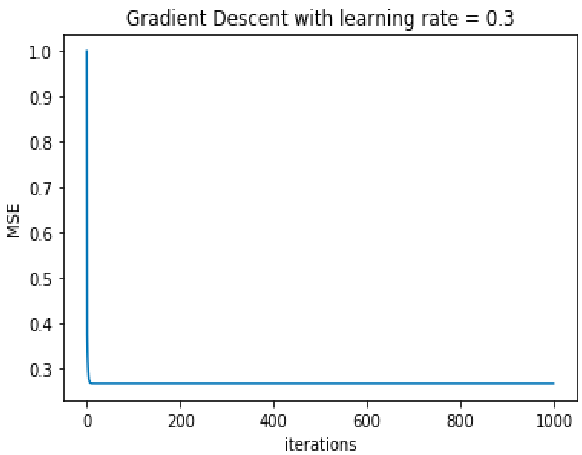
```
At alpha = 0.01, w0 = [-8.7459824e-17], w1 = [0.8846979], w2 = [-0.05311083].
At alpha = 0.1, w0 = [-1.006301e-16], w1 = [0.8847658], w2 = [-0.05317871].
At alpha = 0.3, w0 = [-9.1769005e-17], w1 = [0.8847659], w2 = [-0.05317878].
```

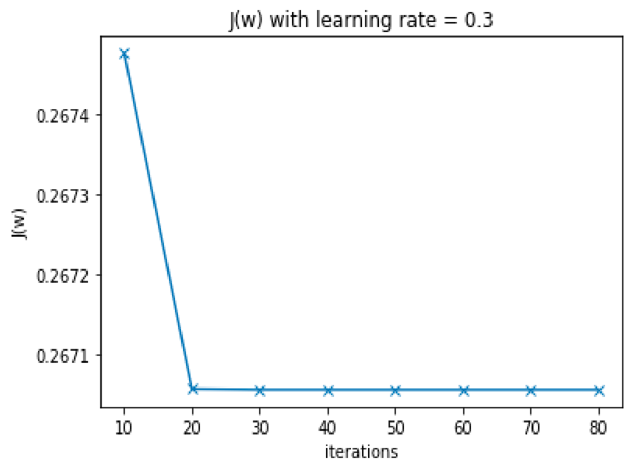












3.

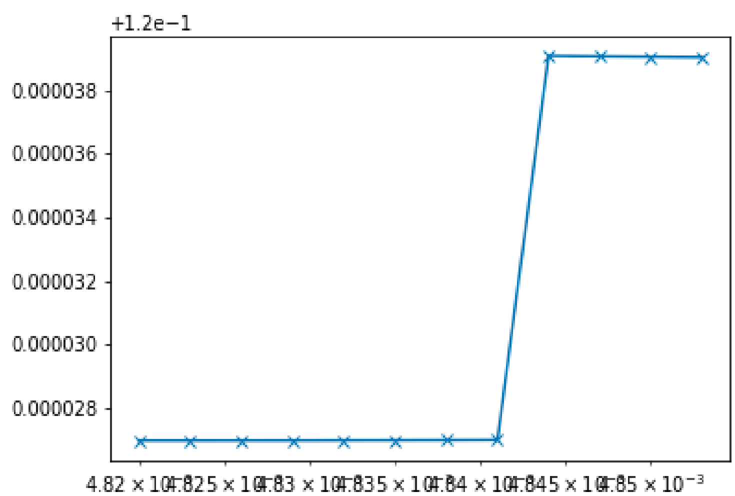
Using the W obtained from learning rate = 0.3:

```
The w is w0 = [-9.1769005e-17], w1 = [0.8847659], w2 = [-0.05317878], and pred_price is [493159.44]
```

Question 4

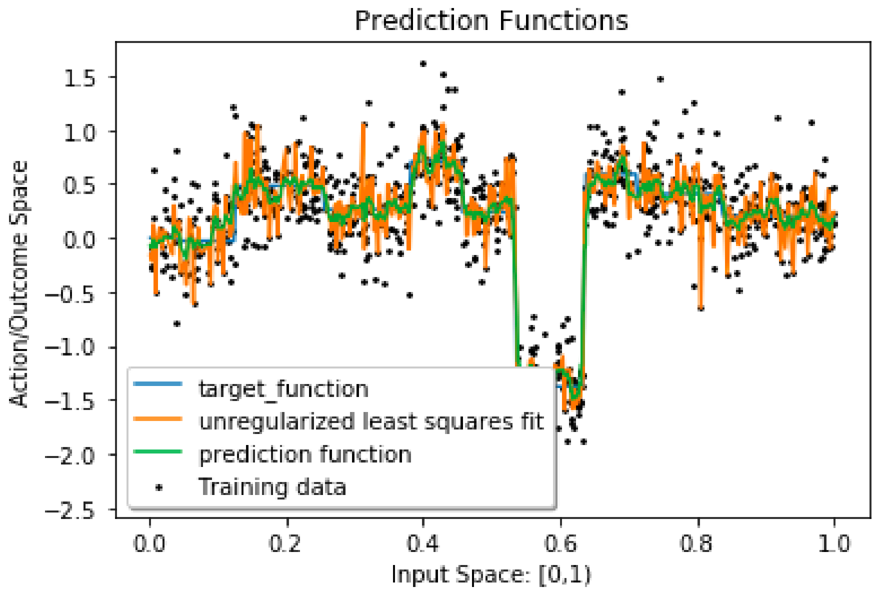
1.

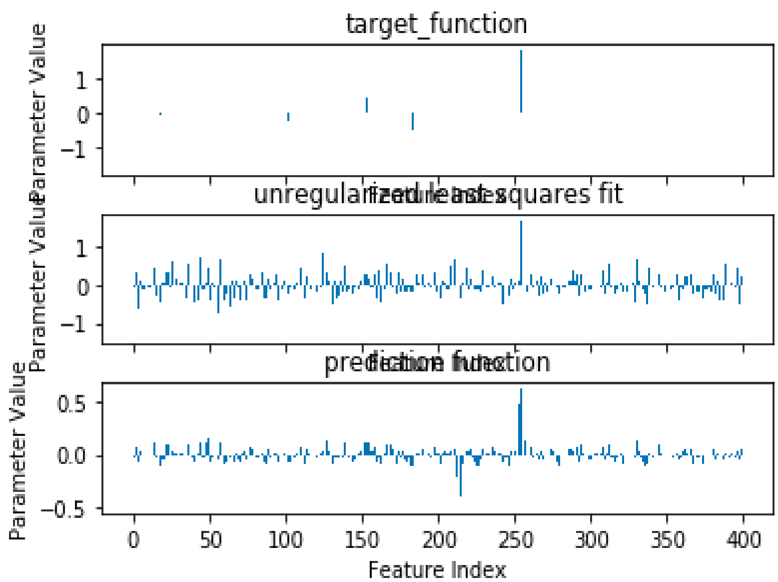
	param_l2reg	mean_test_score	mean_train_score
0	0.004820	0.120027	0.073094
1	0.004823	0.120027	0.073100
2	0.004826	0.120027	0.073106
3	0.004829	0.120027	0.073112
4	0.004832	0.120027	0.073117
5	0.004835	0.120027	0.073123
6	0.004838	0.120027	0.073129
7	0.004841	0.120027	0.073135
8	0.004844	0.120039	0.073141
9	0.004847	0.120039	0.073147
10	0.004850	0.120039	0.073153
11	0.004853	0.120039	0.073159



Take the best λ at 0.004841.

2.





The parameter value aligns with the piece-wise prediction functions: where the prediction functions decreases sharply the parameter value decreases relatively significantly. The scale of coefficients are pretty small, basically around 0 and approximate in the range (-1,1). A few coefficients have higher weight, and the coefficients have the most weight in all these 3 fits occur at 255th coefficient.