EECS 545 HW1

1. (a)

i. For Batch gradient descent and Stochastic gradient descent, we use the same hyperparameters, i.e., initial weight: (0, 0), learning rate: 0.001, and MSE≤0.2 as termination iteration condition.

The coefficients generated by Batch gradient descent is (1.88033999, -2.68963297).

The coefficients generated by Stochastic gradient descent is (1.88029847, -2.68956969).

ii. Use MSE≤0.2 as termination iteration condition and hyperparameters s are the same as in question i.

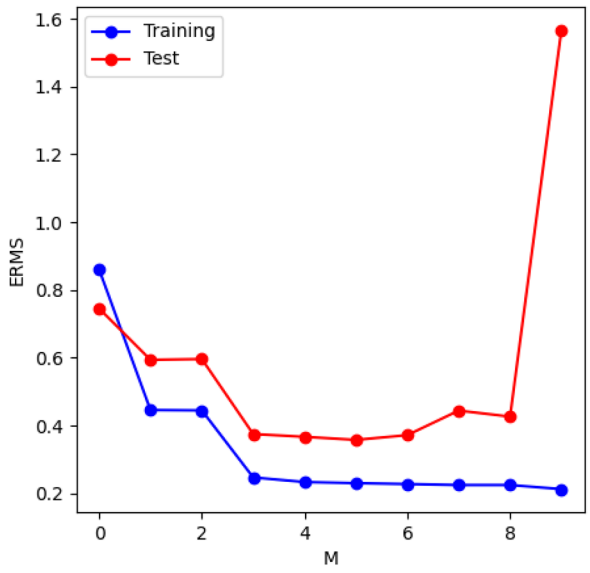
For Batch gradient descent, the training took 2824 iterations to converge.

For Stochastic gradient descent, the training took 2820 iterations to converge.

So, it can be said that the convergence speed of Stochastic gradient descent is slightly faster than that of Batch gradient descent. And as the size of the dataset increases, the convergence speed of Stochastic gradient descent will be much faster than that of Batch gradient descent.

1. (b)

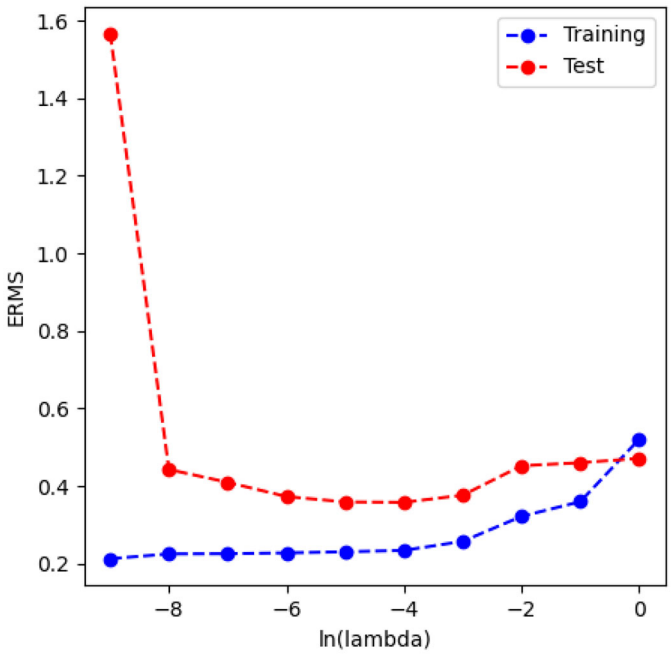
i.



ii. Based on the chart above, it can be said that a polynomial of degree 5 is the best fit for the data. When M=0, 1, 2, there is a certain underfitting because the RMSE of the training data is large. When M = 9, there is overfitting because the RMSE of the test data is large.

1. (c)

i. Use the logarithm of lambda to represent the value on the x-axis to better see the trend in the graph.



ii. From the graph above, λ = 1e - 5 is the best model because it has a lower RMSE.

2. (a)

Let $z = Xw - y$, then $z_i = w^T x_i - y_i$

$$E_D(w) = \frac{1}{2}\sum_{i=1}^{N} r_i (w^T x_i - y_i)^2$$

$$= \sum_{i=1}^{N} \frac{1}{2} r_i z_i^2$$

$$= z^T R z$$

$$= (Xw - y)^T R (Xw - y)$$

where $R_{ii} = \frac{1}{2} r_i$, $R_{ij} = 0$, $i \neq j$

2. (b)

$$\nabla_w E_D(w) = \nabla_w \left(\frac{1}{2} w^T X^T R X w - w^T X^T R y + \frac{1}{2} y^T R y\right)$$

$$= X^T R X w - X^T R y$$

$$= 0$$

Solve this equation, we can get $w^* = (X^T R X)^{-1} X^T R y$

2. (c)

Log likehood:

$$\log p(y^{(1)} \sim y^{(N)} | X, w) = \log \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}\sigma^{(i)}} \exp\left(-\frac{(y^{(i)} - w^T X^{(i)})^2}{2(\sigma^{(i)})^2}\right)\right)$$

$$= \sum_{i=1}^{N} \left(-\frac{1}{2}\log 2\pi - \log \sigma^{(i)} - \frac{(y^{(i)} - w^T X^{(i)})^2}{2(\sigma^{(i)})^2}\right)$$

$$= -\frac{N}{2}\log 2\pi - \sum_{i=1}^{N} \log \sigma^{(i)} - \sum_{i=1}^{N} \frac{(y^{(i)} - w^T X^{(i)})^2}{2(\sigma^{(i)})^2}$$

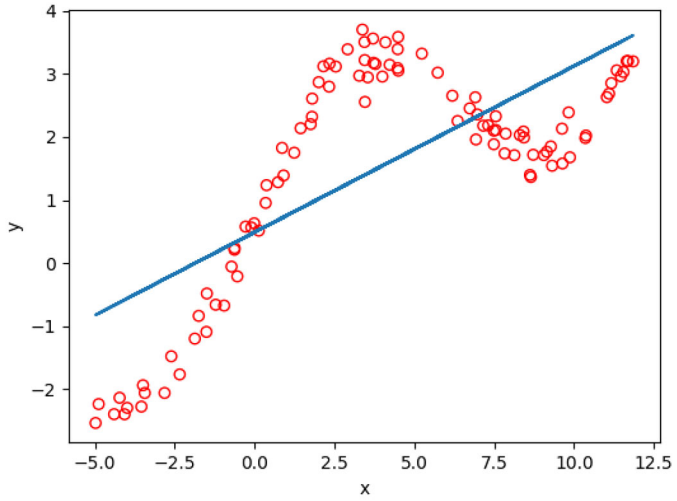
Let $\nabla_w \log p(y^{(1)} \sim y^{(N)} | X, w) = 0$:

$$\nabla_w \log p(y^{(1)} \sim y^{(N)} | X, w) = -\sum_{i=1}^{N} \frac{1}{\sigma^{(i)2}} (y^{(i)} - w^T X^{(i)}) X^{(i)} = 0$$

Notice that if we replace $\frac{1}{\sigma^{(i)2}}$ with $r^{(i)}$, we can get the same equation as problem b)., so we can find the maximum likehood estimate of w by solving $w^* = (X^T R X)^{-1} X^T R y$, And $r^{(i)}$ can be set as $\frac{1}{\sigma^{(i)2}}$.

2. (d)

i. The coefficients generated by normal equation is (0.49073707, 0.26333931).



ii.

$$E(w) = \frac{1}{2}\sum_{i=1}^{N}(w^T\phi(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2}\|w\|^2$$

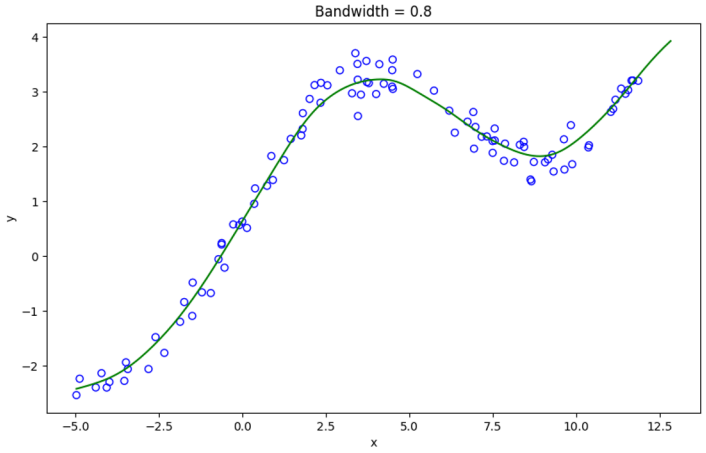
$$= \frac{1}{2}w^T\phi^T\phi w - w^T\phi^T y + \frac{1}{2}y^T y + \frac{\lambda}{2}w^T w$$

$$\nabla_w E(w) = \phi^T\phi w - \phi^T y + \lambda w$$

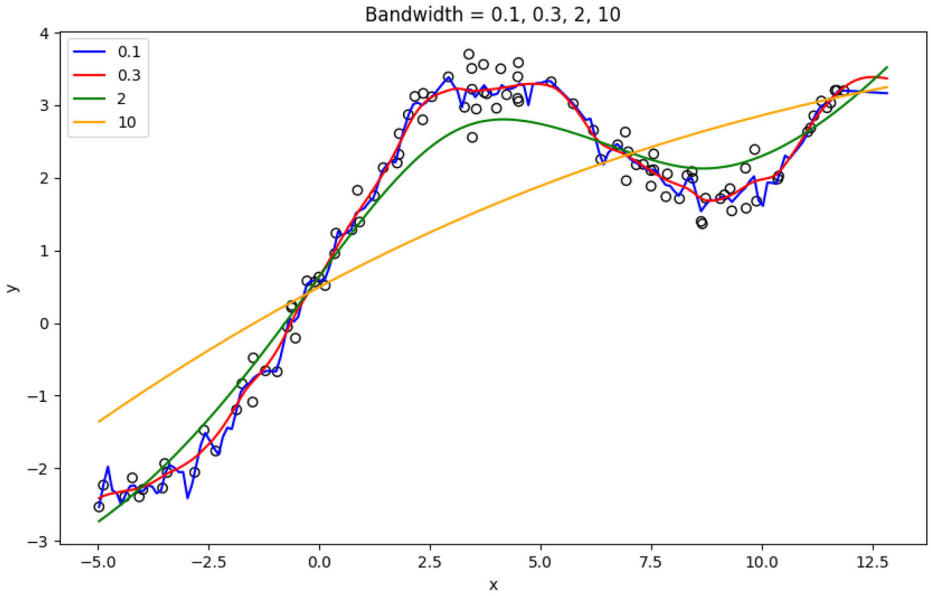
$$= (\lambda I + \phi^T\phi)w - \phi^T y$$

$$= 0$$

$$\Rightarrow w_{ML} = (\lambda I + \phi^T\phi)^{-1}\phi^T y$$



iii. It can be seen that when τ is too small, such as 0.1 or 0.3, the model tends to overfit. Conversely, when τ is too large, say 10, the model tends to underfit.



3. (a)

$$L = \frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - h(x^{(i)}))^2 = \frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - w_0 - w_1 x^{(i)})^2$$

$$\begin{cases} \frac{\partial L}{\partial w_0} = \sum_{i=1}^{N}(y^{(i)} - w_0 - w_1 x^{(i)})^2 = 0 & ① \\ \frac{\partial L}{\partial w_1} = \sum_{i=1}^{N}[(y^{(i)} - w_0 - w_1 x^{(i)})x^{(i)}] = 0 & ② \end{cases}$$

① : $\sum_{i=1}^{N}(y^{(i)} - w_1 x^{(i)}) = \sum_{i=1}^{N} w_0 = N w_0$

$\Rightarrow \underline{\underline{w_0 = \bar{Y} - w_1 \bar{X}}}$

② : $\sum_{i=1}^{N}(y^{(i)}x^{(i)} - w_0 x^{(i)} - w_1 x^{(i)2}) = 0$

$\Rightarrow \sum_{i=1}^{N}(y^{(i)}x^{(i)} - \bar{Y}x^{(i)} + w_1 \bar{X}x^{(i)} - w_1 x^{(i)2}) = 0$

$\Rightarrow \sum_{i=1}^{N}(w_1 x^{(i)2} - w_1 \bar{X}x^{(i)}) = \sum_{i=1}^{N}(y^{(i)}x^{(i)} - \bar{Y}x^{(i)})$

$\Rightarrow w_1\left(\frac{1}{N}\sum_{i=1}^{N}x^{(i)2} - \bar{X}^2\right) = \frac{1}{N}\sum_{i=1}^{N}y^{(i)}x^{(i)} - \bar{Y}\bar{X}$

$$\underline{\underline{w_1 = \frac{\frac{1}{N}\sum_{i=1}^{N}y^{(i)}x^{(i)} - \bar{Y}\bar{X}}{\frac{1}{N}\sum_{i=1}^{N}x^{(i)2} - \bar{X}^2}}}$$

3. (b)

i.

($\Leftarrow$) Since A can be expressed via the spectral decomposition.

$A = U\Lambda U^T$, with $UU^T = U^TU = I$ and $\Lambda = \text{diag}(\lambda_1, \sim\lambda_d)$

For any $z \neq 0$, $z^TAz = z^TU\Lambda U^Tz = (U^Tz)^T\Lambda(U^Tz)$

Let $y = U^Tz$, $z^TAz = y^T\Lambda y = \lambda_1 y_1^2 + \cdots \lambda_d y_d^2$

Since $\forall i, \lambda_i > 0$, $z^TAz > 0$.

So that A is PD.

($\Leftarrow$) Since A is PD, $\forall z \neq 0$, $z^T U \Lambda U z > 0$,

Let $y = U^T z$, $y^T \Lambda y = \lambda_1 y_1^2 + \cdots + \lambda_d y_d^2 > 0$.

Then $y^T \Lambda y > 0$ holds for any λ_i if and only if every λ is positive.

ii.

Suppose that $\phi^T \phi = U \Lambda U^T$, $U^T U = U U^T = I$

Then $\phi^T \phi + \beta I = U \Lambda U^T + U \beta I U^T = U(\Lambda + \beta I) U^T$

The eigenvalues of $\phi^T \phi + \beta I$ are diagonal elements $\lambda_i + \beta$.

According SVD, the eigenvalues are equal to the singular values,

So, the ridge regression has an effect of shifting all singular values by β.

Let $y = U^T z$. $z^T (\phi^T \phi + \beta I) z = (\lambda_1 + \beta) y_1^2 + \cdots + (\lambda_d + \beta) y_d^2$

Since $\forall z$, $z^T (\phi^T \phi) z = (\phi z)^T (\phi z) \geq 0$, the eigenvalues of $\phi^T \phi$ should be non-negative,

So for $\forall \beta > 0$, $\lambda_i + \beta$ should be positive. i.e. $\phi^T \phi + \beta I$ is PD. #