EECS545(Section001): Homework #1

Due on Jan.25, 2022 at 11:59pm

Instructor:Honglak Lee

Tiejin Chen **tiejin@umich.edu**

Problem 1

To finish running the code, please cancel the plot windows when plt.show() is running. **(a)**

(i)

for batch gradient descent, we will use the whole dataset as a batch. And we use learning rate is 0.001, the initial of b and w is 0, And $\epsilon = 0.2$ to get the parameter:

1. Batch gradient descent
   intercept:1.8803399934089935,slope:-2.689632966538344
2. Stochastic gradient descent
   intercept:1.8798794643605279,slope:-2.6898976027584025

(ii)

We use the same hyperparameter in part(i)(learning rate is 0.001, the initial of b and w is 0, And $\epsilon = 0.2$), and we gets the result:

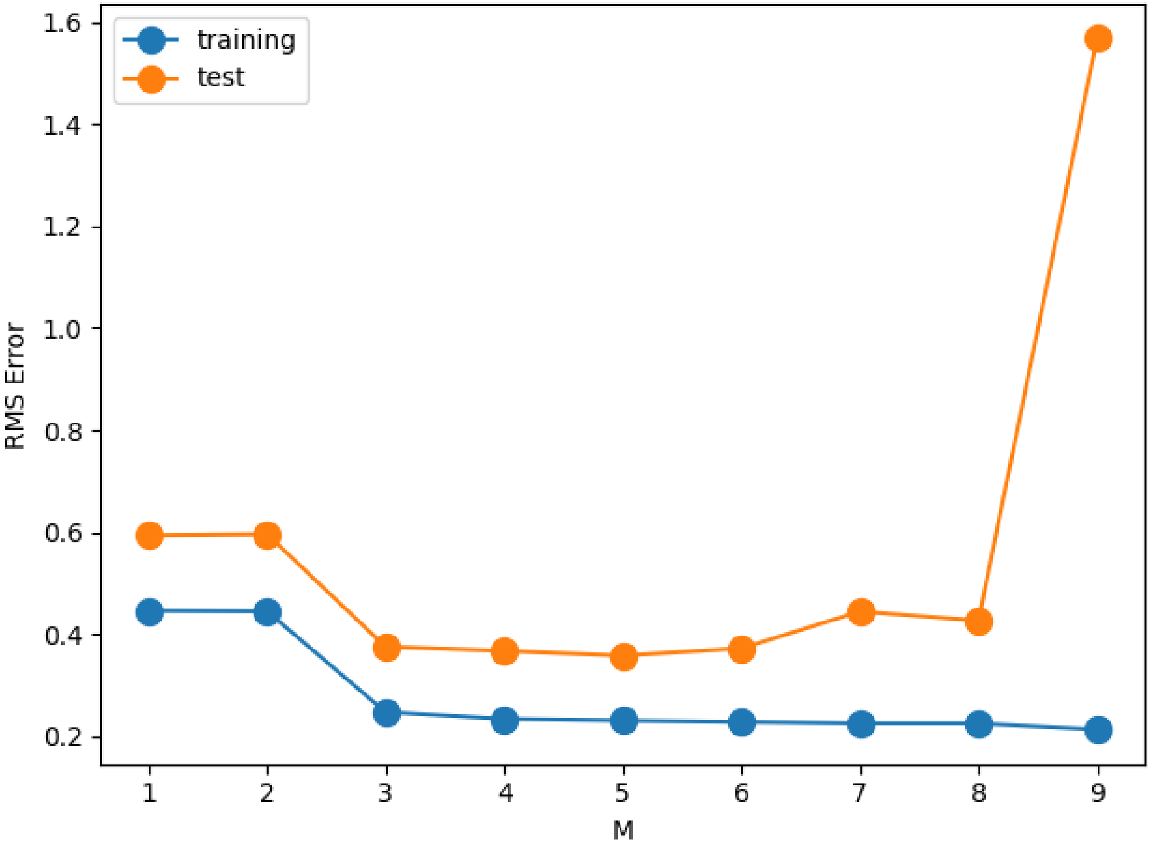
1. Batch gradient descent
   after 2824 epoches, it converges when eps sets to 0.2
2. Stochastic gradient descent
   after 2822 epoches, it converges when eps sets to 0.2

Stochastic gradient descent converges faster.

(b)

(i)

We get the plot:



(ii)

We would say 5 degrees best fits the data because it has least test RMS error. And the charts shows some trend of under/over-fitting. When $M \leq 2$, we can see that both training and test RMS error is very high which is the performance of under-fitting. And also, when M 8 to 9, the trianing RMS error decrease while test RMS error increase a lot. This is the performance of over-fitting.

(c)

To make problem simple, we can change the objective function a little bit. And we can get:

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

We call the last form $L(w)$. Now we differentiate $L(w)$, and we find the differentiation of the first term is the result of normal linear regression, and we can get :

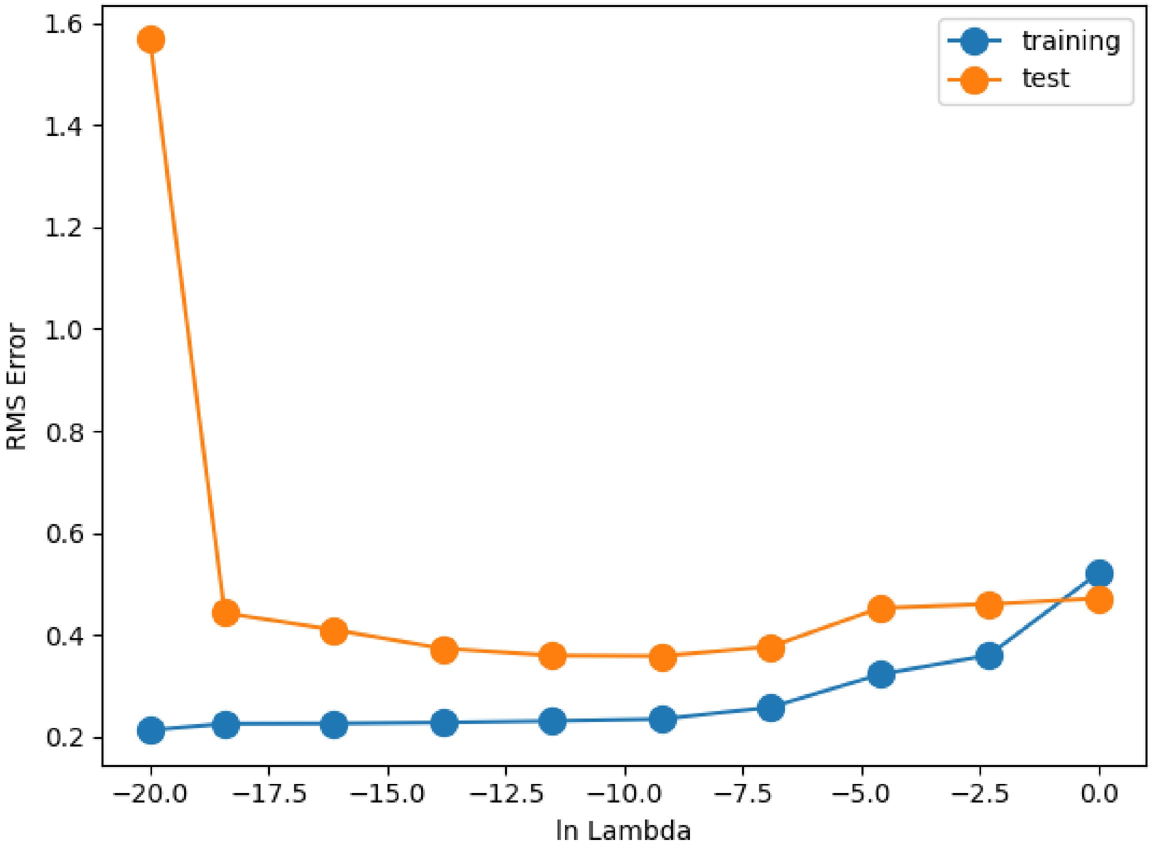
$$\frac{\partial L(w)}{\partial w} = 2\phi(x)^T\phi(x)w - 2\phi(x)^Ty + 2\lambda w$$

we set this to 0, we get the closed form of the ridge regression is:

$$w = (\phi(x)^T\phi(x) + \lambda I)^{-1}\phi(x)^Ty$$

Where I is the identity matrix.

Since ln fucntion does not have definition on 0, we will use -20 to present the x-axis of $\lambda = 0$ in the plot to make plot beautiful. And we can get the plot:



(ii)

From the plot, we can see when $\lambda = 10^{-4}$, the rms error for test set minimize. Hence we think $\lambda = 10^{-5}$ work the best.

Problem 2

(a)

Since X is a matrix whose i-th row is $x^{(i)}$We can know that:

$$(Xw - y)^T = [\begin{bmatrix} w^T x^{(1)} \\ w^T x^{(2)} \\ ... \\ w^T x^{(N)} \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ ... \\ y^{(N)} \end{bmatrix}]^T = \begin{bmatrix} w^T x^{(1)} - y^{(1)} & ... & w^T x^{(N)} - y^{(N)} \end{bmatrix}$$

Let $R = diag(0.5r^{(1)}, 0.5r^{(2)}, ..., 0.5r^{(N)})$, Then we know:

$$(Xw - y)^T R = \begin{bmatrix} 0.5r^{(1)}(w^T x^{(1)} - y^{(1)}) & 0.5r^{(2)}(w^T x^{(2)} - y^{(2)}) & ... & 0.5r^{(N)}(w^T x^{(N)} - y^{(N)}) \end{bmatrix}$$

Thus, we have:

$$\begin{aligned} (Xw - y)^T R(Xw - y) &= \begin{bmatrix} 0.5r^{(1)}(w^T x^{(1)} - y^{(1)})... & 0.5r^{(N)}(w^T x^{(N)} - y^{(N)}) \end{bmatrix} \begin{bmatrix} w^T x^{(1)} - y^{(1)} \\ ... \\ w^T x^{(N)} - y^{(N)} \end{bmatrix} \\ &= \sum_{i=1}^{N} 0.5r^{(i)}(w^T x^{(i)} - y^{(i)})^2 \\ &= \frac{1}{2} \sum_{i=1}^{N} r^{(i)}(w^T x^{(i)} - y^{(i)})^2 \\ &= E_D(w) \end{aligned}$$

Thus we prove that $E_D(w)$ can be written as $(Xw-y)^T R(Xw-y)$. And R is a diagonal matrix with element $(0.5r^{(1)}, 0.5r^{(2)}, ..., 0.5r^{(N)})$.

(b)

We know:

$$\frac{\partial u^T A u}{\partial u} = 2Au$$

We have:

$$\frac{\partial E_D(w)}{\partial w} = (\frac{\partial(Xw - y))}{\partial w})^T \frac{\partial E_D(w)}{\partial(Xw - y)} = X^T R(Xw - y)$$

Let it to be 0, And we can get:

$$w = (X^T RX)^{-1} X^T Ry$$

(c)

We only consider kernel of pdf. And in the following process of proof, when we write $p(y^{(i)}|x^{(i)}; w)$, we mean $exp(-\frac{(y^{(i)} - w^T x^{(i)})^2}{2(\sigma^{(i)})^2})$,and we do not care about the constant front of this term. For likelihood function, we have:

$$L(w) = \prod_{i=1}^{N} p(y^{(i)}|x^{(i)}; w) = exp(-\sum_{i=1}^{N} \frac{(y^{(i)} - w^T x^{(i)})^2}{2(\sigma^{(i)})^2})$$

We take log-lokelihood function to get:

$$lnL(w) = -\sum_{i=1}^{N} \frac{1}{2} \frac{1}{(\sigma)^2}(w^T x^{(i)} - y^{(i)})^2$$

To maximize the likelihood function is equivalence to maximize log-likelihood function. And if we let $r^{(i)} = \frac{1}{(\sigma^{(i)})^2}$, then we can have:

$$max\ lnL(w) \Leftrightarrow max - \sum_{i=1}^{N}\frac{1}{2}r^{(i)}(w^Tx^{(i)} - y^{(i)})^2 \Leftrightarrow min\sum_{i=1}^{N}\frac{1}{2}r^{(i)}(w^Tx^{(i)} - y^{(i)})^2 \Leftrightarrow minE_D(w)$$

Thus, when $r^{(i)} = \frac{1}{(\sigma^{(i)})^2}$, this kind of MLE is equivalence to locally weighted linear regression.

And to get MLE estimator, we differentiate $lnL(w)$, and let it to 0 to get:

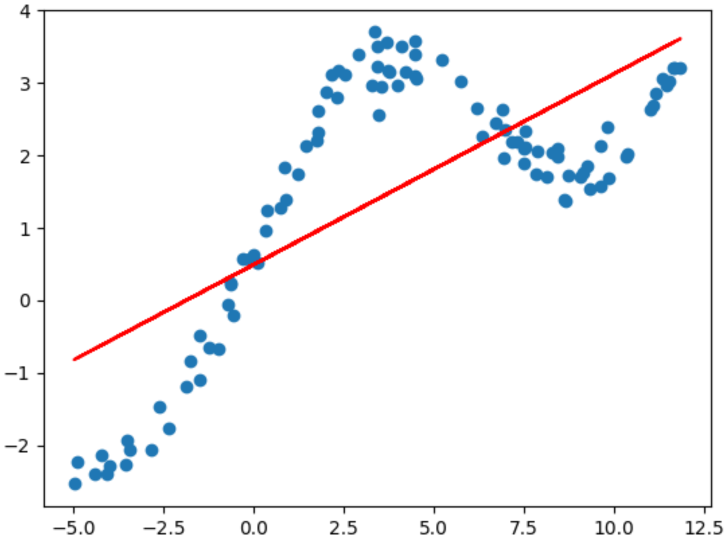
$$w^* = (X^T\Sigma X)^{-1}X^T\Sigma y$$

where Σ is a diagonal matrix with element $(\frac{1}{2(\sigma^{(1)})^2}, \frac{1}{2(\sigma^{(2)})^2}, ..., \frac{1}{2(\sigma^{(N)})^2})$.

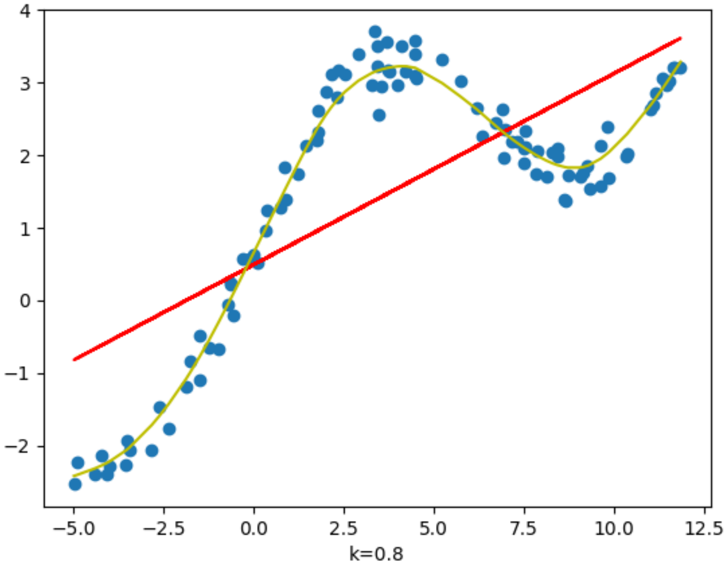
(d)

(i)

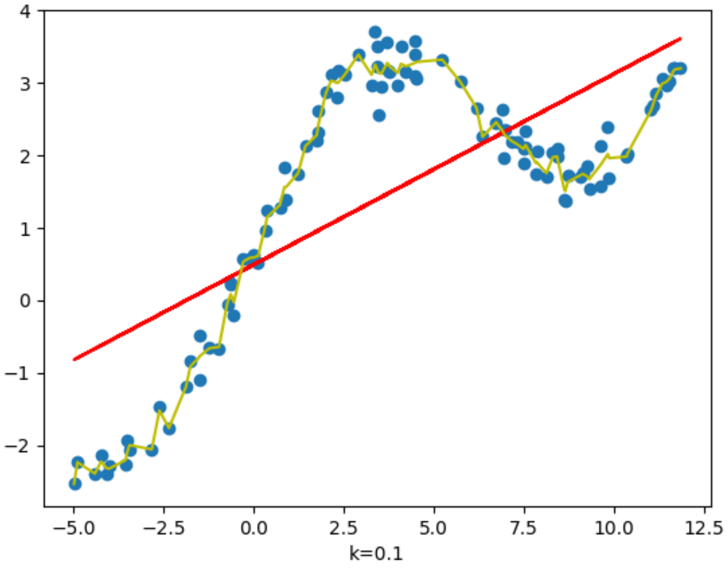
we can get:

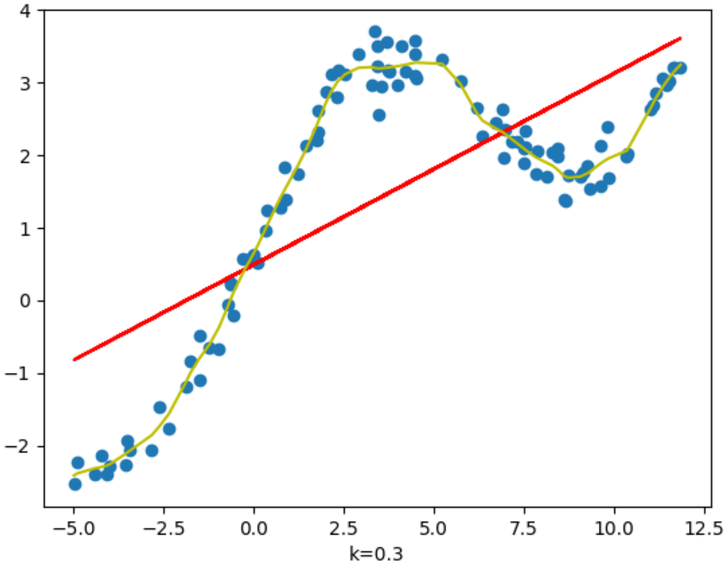


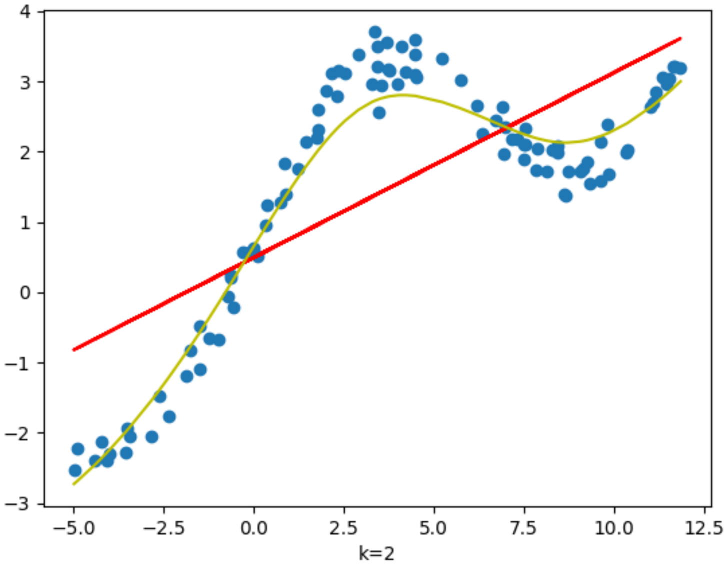
(ii)



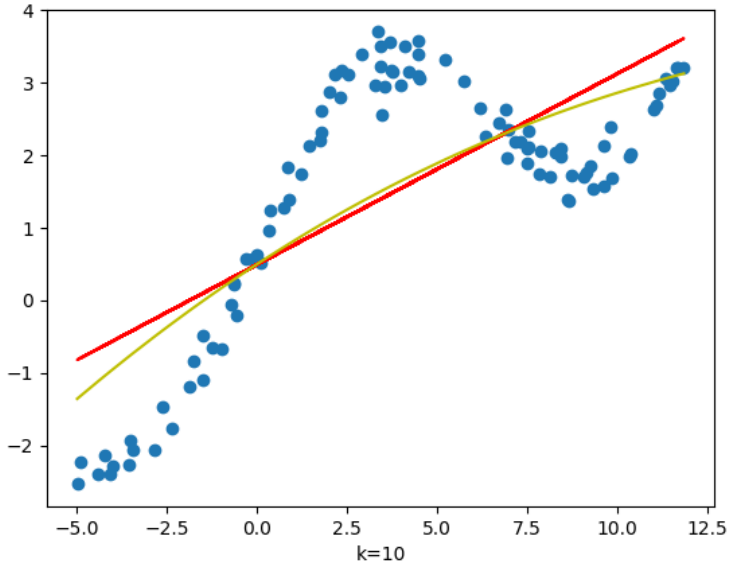
(iii) We have following 4 plots:







Problem 2 continued on next page. . .



And we can see from the image that when τ is too small(equal to 0.1), it will over-fitting. And when τ it too large(equal to 10), it will get closer to normal linear regression and seems like a straight line. This is kind of under-fitting.

Problem 3

(a)

We compute the partial derivatives of loss function to get:

$$\frac{\partial L}{\partial w_0} = -\sum_{i=1}^{n}(y^{(i)} - w_0 - w_1 x^{(i)})$$

$$\frac{\partial L}{\partial w_1} = -\sum_{i=1}^{n}(y^{(i)} - w_0 - w_1 x^{(i)})x^{(i)}$$

Let them to be 0, and we can get:

$$\sum_{i=1}^{n}(y^{(i)} - w_0 - w_1 x^{(i)}) = 0 \rightarrow n\bar{Y} - nw_0 - nw_1\bar{X} = 0 \rightarrow w_0 = \bar{Y} - w_1\bar{X}$$

$$\sum_{i=1}^{n}(y^{(i)} - w_0 - w_1 x^{(i)})x^{(i)} = 0 \rightarrow \sum_{i=1}^{n} y^{(i)}x^{(i)} - nw_0\bar{X} - w_1\sum_{i=1}^{n}(x^{(i)})^2 = 0$$

for second term, we plug in $w_0 = \bar{Y} - w_1\bar{X}$ to get:

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

Hence we prove what we need.

(b)

(i)

Problem 3 continued on next page. . .

Proof. ($\Rightarrow$) if A is PD. And for each i,we have:

$$u_i^T A u_i = u_i^T \lambda_i u_i = \lambda_i \left\| u_i \right\|^2$$

Since A is PD, we know, the LHS $u_i^T A u_i > 0$. Hence RHS must greater than 0. And we know $\left\| u_i \right\|^2 > 0$ when $u_i \neq 0$. To make RHS greater than 0, λ_i should greater than 0. Hence we can get for each i $\lambda_i > 0$

($\Leftarrow$) First we prove Λ is a PD matrix. For every non 0 $z = (z_1, ..., z_d)^T$. We have:

$$z^T \Lambda z = (z_1, ..., z_d) diag(\lambda_1, ..., \lambda_d)(z_1, ..., z_d)^T = \sum_{i=1}^{d} \lambda_i^2 z_i^2$$

For each i ,we have $\lambda_i > 0$. And for z_i, there must exists some i such that $z_i \neq 0$. We use z_{i_k} to present such element. And we know $\{i_k\} \subset \{1, .., d\}$. And the length of $\{i_k\}$ is n. Hence We have:

$$\sum_{i=1}^{d} \lambda_i^2 z_i^2 = \sum_{j=1}^{n} \lambda_j^2 z_{i_j}^2 > 0$$

Thus we prove that Λ is a PD matrix. Then, for every $z \in R^d$, we have:

$$z^T A z = z^T U \Lambda U^T z$$

We let $p = U^T z$, then we can find that p is also a non zero vector belongs to R^d. Thus, we have:

$$z^T A z = p^T \Lambda p > 0$$

Since Λ is pd. Thus we prove that A is a PD matrix.

□

(ii)

Proof. Assumed $\Phi^T \Phi$ have spectral decomposition so that $\Phi^T \Phi = U_\Phi \Lambda_\Phi U_\Phi^T$, which Λ_Φ has elements $(\lambda_{\Phi 1}, ..., \lambda_{\Phi n})$ are the eigenvalues of $\Phi^T \Phi$. Now we have:

$$U_\Phi (\Lambda_\Phi + \beta I) U_\Phi^T = U_\Phi \Lambda_\Phi U_\Phi^T + \beta U_\Phi U_\Phi^T = \Phi^T \Phi + \beta I$$

Hence we get the spectral decomposition of $\Phi^T \Phi + \beta I$ is:

$$\Phi^T \Phi + \beta I = U_\Phi (\Lambda_\Phi + \beta I) U_\Phi^T$$

And $(\Lambda_\Phi + \beta I)$ is a diagonal matrix with element $(\lambda_{\Phi 1} + \beta, ..., \lambda_{\Phi n} + \beta)$ which are the eigenvalues of $\Phi^T \Phi + \beta I$. Hence we can say that ridge regression has an effect of shifting all singular values by a constant β.

Now to prove for any $\beta > 0$, $\Phi^T \Phi + \beta I$ is pd, we only need to prove $\Phi^T \Phi$ is a PSD matrix. For any vector x that have size of (n,1),we have:

$$x^T (\Phi^T \Phi) x = (\Phi x)^T (\Phi x) = \left\| \Phi x \right\|^2 \geq 0$$

Hence $\Phi^T \Phi$ is a psd matrix. For every i, $\lambda_{\Phi i} \geq 0$. And thus after we give the eigenvalue a constant shifting which is greater than 0. we can have, for every i, $\lambda_{\Phi i} + \beta > 0$. Thus, using the conclusion in (i), we can get $\Phi^T \Phi + \beta I$ is a pd matrix. □