

# 1) Derivation and Proof

(a)  $h(x) = w_1 x + w_0 = [w_0 \ w_1] \begin{bmatrix} 1 \\ x \end{bmatrix}$

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

$N$  pairs of data samples  $(x^{(i)}, y^{(i)})$

1D  $\Rightarrow$  one feature only

so, using the notations from the lecture notes,

$$L(w) = \frac{1}{2} \sum_{i=1}^N \left( \sum_{j=0}^M w_j \phi_j(x^{(i)}) - y^{(i)} \right)^2$$

here,  $M=2$ ,  $\phi_0 = 1$ ,  $\phi_1(x^{(i)}) = x^{(i)}$

using Matrix algebra,

$$L(w) = \frac{1}{2} \sum_{i=1}^N \left( w^T \phi(x^{(i)}) - y^{(i)} \right)^2 \quad \text{--- (1)}$$

where,  $w = [w_0 \ w_1]^T$

Expanding eq (1),

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

$$L(w) = \frac{1}{2} w^T \Phi^T \Phi w - w^T \Phi^T Y + \frac{1}{2} Y^T Y \quad \text{--- (2)}$$

$$h(x^{(i)}) = \sum_{j=0}^M w_j \phi_j(x^{(i)})$$

$M=2$

where,  $\mathbf{Y} = [y^1, y^2, \dots, y^n]^T$

$$\Phi = \begin{bmatrix} \phi_0(x^{(1)}) & \phi_1(x^{(1)}) \\ \vdots & \vdots \\ \phi_0(x^{(n)}) & \phi_1(x^{(n)}) \end{bmatrix}_{N \times 2} = \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix}_{N \times 2}$$

$$x = \begin{bmatrix} x^1 \\ \vdots \\ x^n \end{bmatrix}_{N \times 2}$$

Taking gradient of Eq ② w.r.t  $w$ ,

$$\nabla_w L(w) = \nabla_w \left( \frac{1}{2} w^T \Phi^T \Phi w - w^T \Phi^T Y + \frac{1}{2} Y^T Y \right)$$

$$\nabla_w L(w) = \Phi^T \Phi w - \Phi^T Y = 0$$

Taking pseudo inverse to solve above equation  
for  $w$ ,

$$w_m = (\Phi^T \Phi)^{-1} \Phi^T Y \quad \rightarrow \textcircled{3}$$

$$\Phi^T \Phi = \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix}_{2 \times N} \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix}_{N \times 2} = \begin{bmatrix} N & \dots & N \bar{x} \\ \vdots & \ddots & \vdots \\ N \bar{x} & \dots & N \bar{x} \end{bmatrix}_{2 \times 2}$$

$$= \begin{bmatrix} N & \dots & N \bar{x} \\ \vdots & \ddots & \vdots \\ N \bar{x} & \dots & N \bar{x} \end{bmatrix}_{2 \times 2}$$

$$= \begin{bmatrix} N & \dots & N \bar{x} \\ \vdots & \ddots & \vdots \\ N \bar{x} & \dots & N \bar{x} \end{bmatrix}_{2 \times 2}$$

$$= \begin{bmatrix} N & \dots & N \bar{x} \\ \vdots & \ddots & \vdots \\ N \bar{x} & \dots & N \bar{x} \end{bmatrix}_{2 \times 2}$$

where,  $\bar{x} = \frac{1}{N} \sum_{i=1}^N x^{(i)}$  } is the mean.

$$\Phi^T \Phi = \begin{bmatrix} N & N \bar{x} \\ N \bar{x} & N \bar{x} + \sum_{i=1}^N (x^{(i)})^2 \end{bmatrix} \quad \textcircled{4}$$

$$\begin{bmatrix} \bar{Y} \\ \bar{X} \end{bmatrix} = \begin{bmatrix} - & - & - & - & - & - \\ x_1 & - & - & - & - & - \\ x_2 & - & - & - & - & - \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_N & - & - & - & - & - \end{bmatrix}_{2 \times 2} \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_N \end{bmatrix}_{N \times 1} = \begin{bmatrix} \bar{Y} \\ \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_N \end{bmatrix}_{2 \times 1}$$

Substituting ④ and ⑤ in ③

$$w_{ML} = \begin{bmatrix} 2 & \bar{Y} \\ 2 & \bar{x} \end{bmatrix} \begin{bmatrix} \bar{x} \\ \frac{1}{2} \sum_{i=1}^{2N} (x^{(i)})^2 \end{bmatrix}_{2 \times 2}^{-1} \begin{bmatrix} 2 & \bar{Y} \\ 2 & \bar{x} \\ \vdots & \bar{x} \\ 2 & y^{(i)} \end{bmatrix}_{2 \times 1}$$

$$w_{ML} = \frac{1}{\sum_{i=1}^{2N} (x^{(i)})^2 - N \bar{x}^2} \begin{bmatrix} \frac{\sum_{i=1}^{2N} (x^{(i)})^2}{2} & \bar{x} \\ \bar{x} & 1 \end{bmatrix} \begin{bmatrix} \bar{Y} \\ \bar{x} \\ \vdots \\ \bar{x} \\ y^{(i)} \end{bmatrix}_{2 \times 1}$$

$$w_{ML} = \frac{1}{\sum_{i=1}^{2N} (x^{(i)})^2 - N \bar{x}^2} \begin{bmatrix} (\sum_{i=1}^{2N} (x^{(i)})^2) \bar{Y} - \bar{x} \sum_{i=1}^{2N} x^{(i)} y^{(i)} \\ - \bar{x} \bar{x} \\ - \bar{x} \bar{x} \\ - \bar{x} \bar{x} \\ \sum_{i=1}^{2N} x^{(i)} y^{(i)} \end{bmatrix}$$

$$e_0 = \frac{(\sum_{i=1}^{2N} (x^{(i)})^2) \bar{Y} - \bar{x} \sum_{i=1}^{2N} x^{(i)} y^{(i)}}{\sum_{i=1}^{2N} (x^{(i)})^2 - N \bar{x}^2}$$

$$e_1 = \frac{\sum_{i=1}^{2N} x^{(i)} y^{(i)} - N \bar{x} \bar{Y}}{\sum_{i=1}^{2N} (x^{(i)})^2 - N \bar{x}^2}$$

can be rewritten as,

$$\omega_1 = \frac{2 \sum_{i=1}^N x^{(i)} y^{(i)} - \bar{x} \bar{y}}{2 \sum_{i=1}^N (x^{(i)})^2 - \bar{x}^2}$$

$$\begin{aligned}\omega_0 &= \frac{\left( \sum_{i=1}^N (x^{(i)})^2 \right) \bar{y} - \bar{x} \sum_{i=1}^N x^{(i)} y^{(i)} + N \bar{x}^2 \bar{y} - N \bar{x} \bar{y}}{\sum_{i=1}^N (x^{(i)})^2 - N \bar{x}^2} \\ &= \frac{\left( \sum_{i=1}^N (x^{(i)})^2 - N \bar{x}^2 \right) \bar{y} - \bar{x} \sum_{i=1}^N x^{(i)} y^{(i)} + N \bar{x}^2 \bar{y}}{\sum_{i=1}^N (x^{(i)})^2 - N \bar{x}^2}\end{aligned}$$

$$\omega_0 = \bar{y} - \omega_1 \bar{x}$$



(i) To Prove  $\Rightarrow A$  is PD if and only if  $\lambda_i > 0 \forall i$ .

$\Leftrightarrow$  If  $A$  is P.D, then prove  $\lambda_i > 0 \forall i$ .

Let  $A$  is P.D, then  $\sum A z > 0 \forall z \in \mathbb{R}^n$

Using Eigenvalue decomposition,  $A = U \Lambda U^T$   
 $U^T = G U = I$ ,  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$

$$A u_i = \lambda_i u_i \quad | \quad u_i \text{ denote } i^{\text{th}} \text{ column of } U.$$

so,  $A$  is P.D., Then  $z^T A z > 0$

we can replace  $z$  with  $u_i + v_i$

$$\Rightarrow \text{so, } u_i^T A u_i > 0$$

$\Rightarrow$  substituting from Eq ①

$$\Rightarrow u_i^T \lambda_i u_i > 0, \quad \lambda_i \text{ is scalar}$$

$$\Rightarrow \lambda_i u_i^T u_i > 0$$

we know  $u_i^T u_i > 0$  from inner product property.

so,

$$\Rightarrow \lambda_i > 0 \quad \forall i \quad \underline{\text{Hence Proved.}}$$

Now, let  $\lambda_i > 0 \quad \forall i \in \{1, \dots, d\}$ , then

Prove  $A$  is PD

$\Rightarrow$  Let  $z = c_1 u_1^T + c_2 u_2^T + \dots + c_d u_d^T$ ,  
 $c_i \in \mathbb{R}$  is a constant and  $u_i$  are eigenvectors of  $A$   
 $u_i \in \mathbb{R}^{d \times 1}$   
 Then

$$z^T A z = [c_1 u_1^T + \dots + c_d u_d^T]^T A [c_1 u_1^T + \dots + c_d u_d^T]$$

$$A = U \Lambda U^T, \text{ so, } \Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix}, U = \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix}$$

$$\begin{aligned} Z^T A Z &= \left[ c_1 u_1^T + \cdots + c_d u_d^T \right]^T U \Lambda U^T \left[ c_1 u_1^T + \cdots + c_d u_d^T \right] \\ &= [c_1 \|u_1\|_2^2 + \cdots + c_d \|u_d\|_2^2]^T \Lambda \left[ c_1 \|u_1\|_2^2 + \cdots + c_d \|u_d\|_2^2 \right] \end{aligned}$$

$$Z^T A Z = \lambda_1 c_1^2 + \lambda_2 c_2^2 + \cdots + \lambda_d c_d^2 > 0$$

as we start with assumption that  
 $\lambda_i > 0 \quad \forall i \in \{1, \dots, d\}$

$Z^T A Z > 0$

Hence Proved

(ii) Let  $B \triangleq \Phi^T \Phi$ , and  $\lambda_i$  and  $c_i$  are eigenvalue and eigen vector of  $B$ .  
 so Need to find eigenvalue and eigen vector of  $A \triangleq B + \beta I$

$$\begin{aligned} &= \Phi^T \Phi + \beta I \end{aligned}$$

①

$$B u_i = \lambda_i u_i$$

$$\begin{aligned} \Rightarrow (B + \beta I) u_i &= B u_i + \beta u_i = \lambda_i u_i + \beta u_i \\ &= (\lambda_i + \beta) u_i \end{aligned}$$

$$\Rightarrow (B + \beta I) u_i = (\lambda_i + \beta) u_i$$

so Eigen value of  $B + \beta I$  i.e.  $\Phi^T \Phi + \beta I$  is  $\lambda_i + \beta$

and Eigen vector of  $B + \beta I$  i.e  $\Phi^T \Phi + \beta I$  is  $u_i$

(2)

$$\text{so, } A \triangleq \Phi^T \Phi + \beta I \quad \text{eigenvalue is } \lambda_i + \beta \\ \text{eigen vector is } u_i$$

we know  $A$  is a symmetric matrix, Then,

$$A = U \Lambda U^T, \text{ where } U = [u_1 \dots u_d]^T$$

$$\Lambda = \text{diag}(\lambda_1 + \beta, \dots, \lambda_d + \beta)$$

here  $\lambda_i > 0 \forall i$  as  $\Phi^T \Phi$  is P.D

Now similarly to what we did in (1) Problem,

$$z = [c_1 u_1^T + \dots + c_d u_d^T]$$

$$z^T A z = [c_1 u_1^T + \dots + c_d u_d^T]^T U \Lambda U^T [c_1 u_1^T + \dots + c_d u_d^T]$$

$$= [c_1 \|u_1\|_2^2 + \dots + c_d \|u_d\|_2^2]^T \Lambda [c_1 \|u_1\|_2^2 + \dots + c_d \|u_d\|_2^2]$$

$$z^T A z = (\lambda_1 + \beta) c_1^2 + \dots + (\lambda_d + \beta) c_d^2$$

we know,  $\lambda_i > 0$ , so, for any.

$$\beta > 0,$$

$$z^T A z > 0, \text{ means } A \text{ is P.D}$$

i.e.,  $\Phi^T \Phi + \beta I$  is PD

Hence Proved.

G

$$y \in \{-1, +3\}$$

$$\sum_{n=1}^N \log P(y^{(n)} | x_c^{(n)})$$

$$\sum_{n=1}^N \left[ \mathbb{I}(y^{(n)} = 1) \log P(y^{(n)} = 1 | x_c^{(n)}) + \mathbb{I}(y^{(n)} = -1) \log P(y^{(n)} = -1 | x_c^{(n)}) \right]$$

1

Using sigmoid function

$$P(y^{(n)} = 1 | x^{(n)}) = \sigma(\omega^T \phi(x^{(n)})) \quad \rightarrow \textcircled{2}$$

$$P(y^{(n)} = -1 | x^{(n)}) = 1 - \sigma(\omega^T \phi(x^{(n)})) \quad \rightarrow \textcircled{3}$$

$$\text{where, } \sigma(a) = \frac{1}{1 + e^{-a}}$$

we can rewrite Eq \textcircled{1} using \textcircled{2} and \textcircled{3}

$$\sum_{n=1}^N \left[ \mathbb{I}(y^{(n)} = 1) \log \frac{1}{1 + e^{-\omega^T \phi(x^{(n)})}} + \mathbb{I}(y^{(n)} = -1) \log \frac{e^{-\omega^T \phi(x^{(n)})}}{1 + e^{-\omega^T \phi(x^{(n)})}} \right]$$

$$\sum_{n=1}^N \left[ \mathbb{I}(y^{(n)} = 1) - \log (1 + e^{\omega^T \phi(x^{(n)})}) + \mathbb{I}(y^{(n)} = -1) \left( -\omega^T \phi(x^{(n)}) - \log (1 + e^{\omega^T \phi(x^{(n)})}) \right) \right]$$

$$\sum_{n=1}^N \left( -\omega^T \phi(x^{(n)}) \right) - \sum_{n=1}^N \left[ \mathbb{I}(y^{(n)}=1) \log \left( 1 + e^{-\omega^T \phi(x^{(n)})} \right) + \mathbb{I}(y^{(n)}=-1) \log \left( 1 + e^{-\omega^T \phi(x^{(n)})} \right) \right]$$

we can combine these terms as

where  $\mathbb{I}(y^{(n)}=1) = 1$  when  $y^{(n)}=1$  and 0 otherwise

similarly  $\mathbb{I}(y^{(n)}=-1) = 1$  when  $y^{(n)}=-1$  and 0 otherwise.

$$\sum_{n=1}^N \left( -\omega^T \phi(x^{(n)}) \right) - \sum_{n=1}^N \left( \log \left( 1 + e^{-y^{(n)} \omega^T \phi(x^{(n)})} \right) \right)$$

$$\Rightarrow \sum_{n=1}^N \log P(y^{(n)}|x^{(n)}) = \sum_{n=1}^N \left( -\omega^T \phi(x^{(n)}) \right) - \sum_{n=1}^N \left( \log \left( 1 + \exp(-y^{(n)} \omega^T \phi(x^{(n)})) \right) \right)$$

Maximize

Maximize.

To maximize  $\sum_{n=1}^N \log P(y^{(n)}|x^{(n)})$  we need to

minimize  $\sum_{n=1}^N \log \left( 1 + \exp(-y^{(n)} \cdot \omega^T \phi(x^{(n)})) \right)$

Hence Proved.

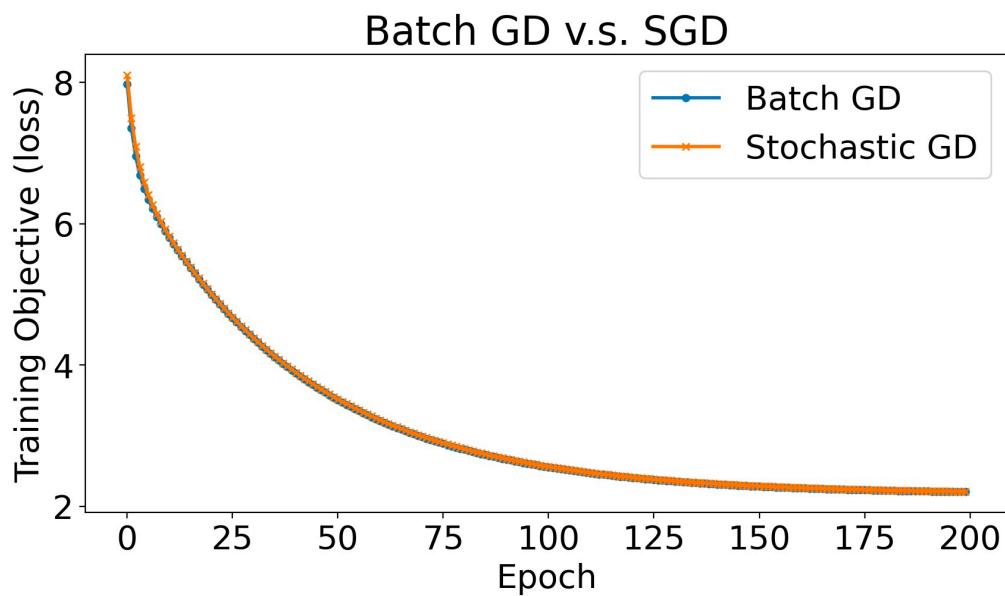
2

## 2.1 — GD and SGD

a)

Auto grader

b)



GD version took 0.00 seconds

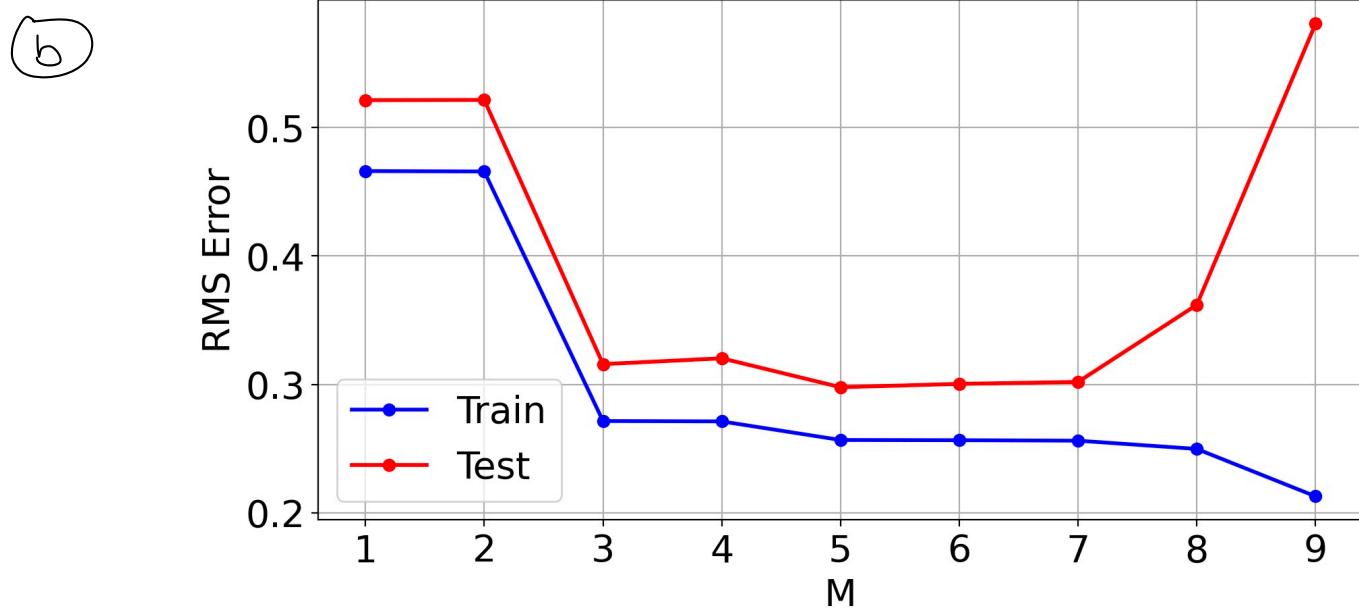
GD Test objective = 2.7017

SGD version took 0.02 seconds

SGD Test objective = 2.6796

## 2.2 Over-fitting Study

(a) Autograder



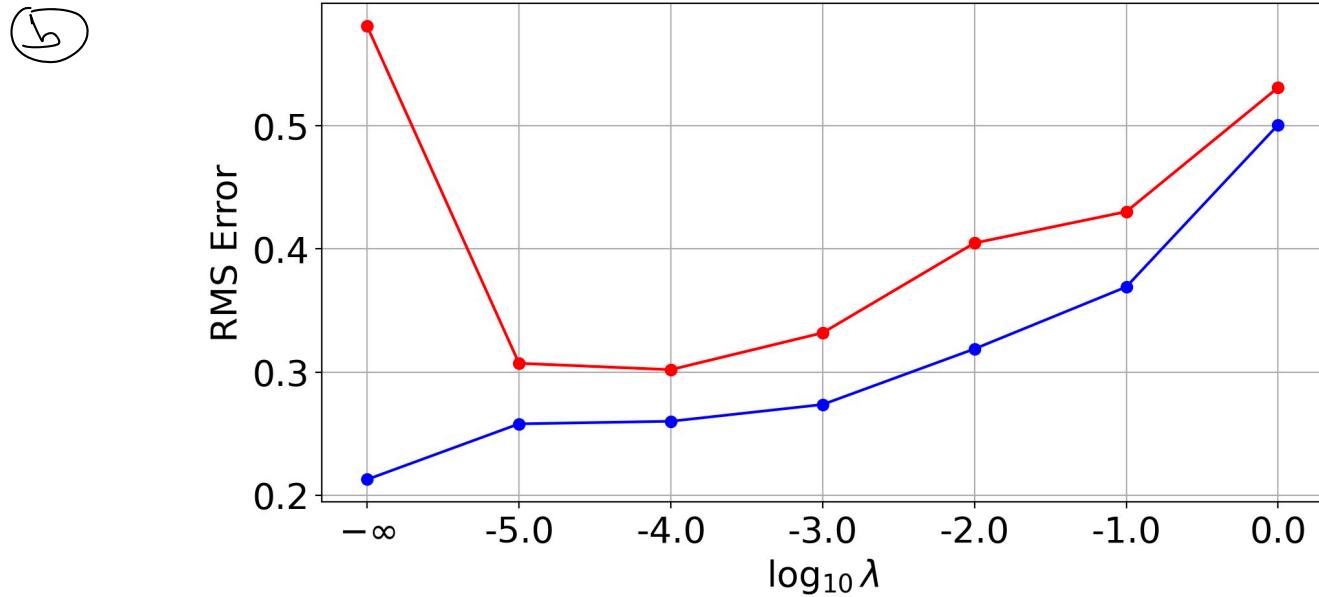
(c) Degree of S polynomial will best fit the data. Yes, there was an evidence for both over/under-fitting.

For  $M > 7$ , its an evidence of over-fitting  
For  $M < 5$ , its an evidence of under-fitting.

2.3

## Regularization

(a) Autograder.



(c)  $\lambda = 10^{-4}$  seems to best work  
for  $q^{\text{th}}$  degree polynomial as from  
plot we can see, RMS error is min  
at  $\log_{10} \lambda = -4.0$

(3)

## Locally weighted linear regression

$$E_D(\omega) = \frac{1}{2} \sum_{i=1}^N \pi^{(i)} (\omega^T x^{(i)} - y^{(i)})^2 \quad \left| \begin{array}{l} \pi^{(i)} \in \mathbb{R} \text{ is} \\ \text{local weight for} \\ \text{samp} \end{array} \right. \quad \left. \begin{array}{l} \text{sample} \\ (x^{(i)}, y^{(i)}) \end{array} \right.$$

(a)  $E_D(\omega) = (\omega^T X - Y^T) R (\omega^T X - Y^T)^T$

where  $X \in \mathbb{R}^{D \times N}$  whose  $i^{\text{th}}$  column is  $x^{(i)} \in \mathbb{R}^{D \times 1}$   
 and  $Y \in \mathbb{R}^{N \times 1}$ ,  $\omega \in \mathbb{R}^{D \times 1}$

so,  $E_D(\omega) = (\omega^T X - Y^T) R (\omega^T X - Y^T)^T$

here  $R$  is a diagonal matrix

$$R = \frac{1}{2} \text{diag}(\pi^1, \pi^2, \dots, \pi^N)$$

$$E_D(\omega) = \sum_{i,j}^N (\omega^T x - Y^T)_i R_{ij} (\omega^T X - Y^T)_j$$

as  $R$  is a diagonal matrix

$$= \sum_{i=1}^N (\omega^T x - Y^T)_i^2 R_{ii} \quad \left. \right\} R_{ii} = \frac{1}{2} \pi^{(i)}$$

$$= \sum_{i=1}^N \frac{1}{2} \pi^{(i)} (\omega^T x^{(i)} - y^{(i)})^2$$

$$= \frac{1}{2} \sum_{i=1}^N \pi^{(i)} (\omega^T x^{(i)} - y^{(i)})^2$$

Hence Proved

(b)  $E_D(\omega) = \underset{N \times N}{(\omega^T X - \gamma^T)} R \underset{N}{(\omega^T X - \gamma^T)^T}$

$$\nabla_{\omega} E_D(\omega) = (\omega^T X - \gamma^T) R \nabla_{\omega} (\omega^T X - \gamma^T)^T$$

$$= (\omega^T X - \gamma^T) R X^T$$

$X \in \mathbb{R}^{D \times N}$   
 $R \in \mathbb{R}^{N \times N}$   
 $\gamma \in \mathbb{R}^{N \times 1}$   
 $\omega \in \mathbb{R}^{D \times 1}$

$$\Rightarrow (\omega^T X - \gamma^T) R X^T = 0$$

$$\Rightarrow \omega^T X R X^T - \gamma^T R X^T = 0$$

$$\Rightarrow \omega^T X R X^T = \gamma^T R X^T$$

$$\Rightarrow (\omega^T X R X^T)^T = (\gamma^T R X^T)^T$$

$$\Rightarrow (X R X^T) \omega = (R X^T)^T \gamma$$

$$\Rightarrow \boxed{\omega^* = (X R X^T)^{-1} (R X^T)^T \gamma}$$

$X \in \mathbb{R}^{D \times N}$   
 $R \in \mathbb{R}^{N \times N}, \quad \gamma \in \mathbb{R}^{N \times 1}$   
 $\omega \in \mathbb{R}^{D \times 1}$

(c)  $\{(x^{(i)}, y^{(i)}) ; i=1, \dots, N\}$

$$P(y^{(i)} | x^{(i)}; \omega) = \frac{1}{\sqrt{2\pi} \sigma^{(i)}} \exp \left( - \frac{(y^{(i)} - \omega^T x^{(i)})^2}{2(\sigma^{(i)})^2} \right)$$

so

$$P(y^{(i)} | x^{(i)}, w, \sigma) = \mathcal{N}\left(y^{(i)} | w^T x^{(i)}, \sigma^2\right)$$

$$P(Y | X, w, \sigma) = \prod_{i=1}^n \mathcal{N}\left(y^{(i)} | w^T x^{(i)}, \sigma^2\right)$$

Taking log.

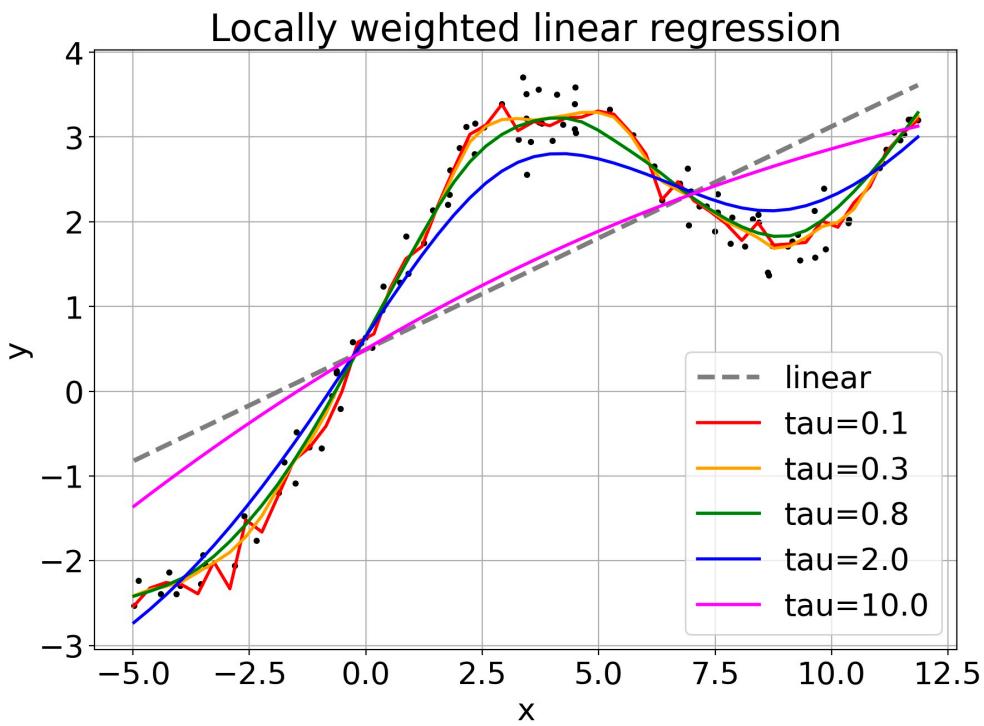
$$\log(P(Y | X, w, \sigma)) = \sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi} \sigma^{(i)}}\right) - \sum_{i=1}^n \frac{(y^{(i)} - w^T x^{(i)})^2}{2\sigma^{(i)2}}$$

Hence problem is reduced to solving a weighted linear regression problem  $E_D(w)$  with

weight

$$w^{(i)} = \frac{1}{2\sigma^{(i)2}}$$

(i) Autograder.



(ii) When  $\tau$  is too small, fit is becoming over-fit and when  $\tau$  is too big fit is becoming under-fit.

i.e., when  $\tau$  is too large, it is including the point which is degrading the fit which is underfitting.

when  $\tau$  is too small, it is excluding the point that increase confidence in the fit.