1) Derivation and Proof

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

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

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

1D $\Rightarrow$ one feature only

$$h(x^{(i)}) = \sum_{j=0}^{M-1} w_j \phi_j(x^{(i)}), \quad M = 2$$

so, using the notations from the lecture notes,

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

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

using Matrix algebra,

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

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

Expanding eq (1),

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

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

where, $Y = [y^1, y^2, \cdots y^N]^T$

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

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 the above equation for w,

$$w_{ML} = (\Phi^T \Phi)^{-1} \Phi^T Y \quad \text{— (3)}$$

$$\Phi^T \Phi = \begin{bmatrix} 1 & \cdots & 1 \\ x^1 & \cdots & x^N \end{bmatrix}_{2 \times N} \begin{bmatrix} 1 & x^1 \\ \vdots & \vdots \\ 1 & x^N \end{bmatrix}_{N \times 2} = \begin{bmatrix} N & N\bar{x} \\ N\bar{x} & \sum_{i=1}^{N} (x^{(i)})^2 \end{bmatrix}_{2 \times 2}$$

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

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

$$[\Phi]^T Y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x^1 & x^2 & \cdots & x^n \end{bmatrix}_{2\times n} \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^n \end{bmatrix}_{n\times 1} = \begin{bmatrix} \sum_{i=1}^{n} y^{(i)} \\ \sum_{i=1}^{n} x^{(i)} y^{(i)} \end{bmatrix}_{2\times 1} \quad (5)$$

substituting (5) and (6) in (3)

$$\theta_{ML} = \begin{bmatrix} n & \sum_{i=1}^{n} x^{(i)} \\ \sum_{i=1}^{n} x^{(i)} & \sum_{i=1}^{n} (x^{(i)})^2 \end{bmatrix}^{-1}_{2\times 2} \begin{bmatrix} \sum_{i=1}^{n} y^{(i)} \\ \sum_{i=1}^{n} x^{(i)} y^{(i)} \end{bmatrix}_{2\times 1}$$

$$\theta_{ML} = \frac{1}{\sum_{i=1}^{n} (x^{(i)})^2 - n\bar{x}^2} \begin{bmatrix} \sum_{i=1}^{n} (x^{(i)})^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} n\bar{y} \\ \sum_{i=1}^{n} x^{(i)} y^{(i)} \end{bmatrix}$$

$$\theta_{ML} = \frac{1}{\sum_{i=1}^{n} (x^{(i)})^2 - n\bar{x}^2} \begin{bmatrix} \left(\sum_{i=1}^{n} (x^{(i)})^2\right)\bar{y} - \bar{x}\sum_{i=1}^{n} x^{(i)} y^{(i)} \\ -n\bar{x}\bar{y} + \sum_{i=1}^{n} x^{(i)} y^{(i)} \end{bmatrix}$$

$$\theta_0 = \frac{\left(\sum_{i=1}^{n} (x^{(i)})^2\right)\bar{y} - \bar{x}\sum_{i=1}^{n} x^{(i)} y^{(i)}}{\sum_{i=1}^{n} (x^{(i)})^2 - n\bar{x}^2}$$

$$\theta_1 = \frac{\sum_{i=1}^{n} x^{(i)} y^{(i)} - n\bar{x}\bar{y}}{\sum_{i=1}^{n} (x^{(i)})^2 - n\bar{x}^2}$$

can be rewritten as,

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

$$w_0 = \frac{\left(\sum_{i=1}^{N} (x^{(i)})^2\right)\bar{y} - \bar{x}\sum_{i=1}^{N} x^{(i)} y^{(i)} + N\bar{x}^2\bar{y} - N\bar{x}\bar{y}^2}{\sum_{i=1}^{N} (x^{(i)})^2 - N\bar{x}^2}$$

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

$$w_0 = \bar{y} - w_1\bar{x}$$

D

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

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

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

using Eigenvalue decomposition, $A = U \Lambda U^T$

$U U^T = U^T U = I$, $\Lambda = \text{diag}(\lambda_1 \dots \lambda_d)$

$Au_i = \lambda_i u_i$ — (1) u_i denote i^{th} column of U.

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

we can replace z with u_i $\forall i$

$\Rightarrow$ so $u_i^T A u_i > 0$

$\Rightarrow$ substituting from eq (1)

$\Rightarrow u_i^T \lambda_i u_i > 0$, λ_i is scaler

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

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

So,

$\Rightarrow$ $\lambda_i > 0$ $\forall i$ Hence Proved.

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

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

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

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

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

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

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

$z^T A z > 0$ Hence Proved

(ii) Let $B \triangleq \Phi^T \Phi$, and λ_i and u_i are eigen value and eigen vector of B.

So Need to find eigen value and eigen vector of $A \triangleq B + \beta I$
$= \Phi^T \Phi + \beta I$

①

$$B u_i = \lambda_i u_i$$

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

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

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

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

(2)

First, need to show $\Phi^T\Phi$ is P.S.D

we know $\Phi^T\Phi$ is invertible, as it is the solution to the least-square problem, i.e, any eigenvalue of $\Phi^T\Phi$ is not zero,

so, $\Phi^T\Phi x = \lambda x$, then

$$\Rightarrow \quad x^T\Phi^T\Phi x = x^T\lambda x = \lambda x^T x$$

$$\Rightarrow \quad \frac{\|\Phi x\|^2}{\|x\|^2} = \lambda > 0$$, Hence $\Phi^T\Phi$ is P.D

so, $A \triangleq \Phi^T\Phi + \beta I$ eigenvalue is $\lambda_i+\beta$, eigenvector is u_i

we know A is a symmetric matrix, then,

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

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

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

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

$$z = \left[\frac{c_1 u_1^T}{} + \cdots + c_d u_d^T\right]$$

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

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

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

we know, $\lambda_i > 0$, so, for any $\beta \geq 0$,

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

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

Hence proved.

(G) $y \in \{-1, +1\}$

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

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

using sigmoid function

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

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

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

we can rewrite eq (1) using (2) and (3)

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

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

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

we can combine these terms as

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

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

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

$$\text{Now } \underbrace{\sum_{n=1}^{N} \log P\left(y^{(n)} \middle| x^{(n)}\right)}_{\text{Maximize}} = \underbrace{\sum_{n=1}^{N} \left(-w^T\phi(x^{(n)})\right) - \sum_{n=1}^{N} \left[\log\left(1+\exp\left(-y^{(n)} w^T \phi(x^{(n)})\right)\right)\right]}_{\text{Maximize.}}$$

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

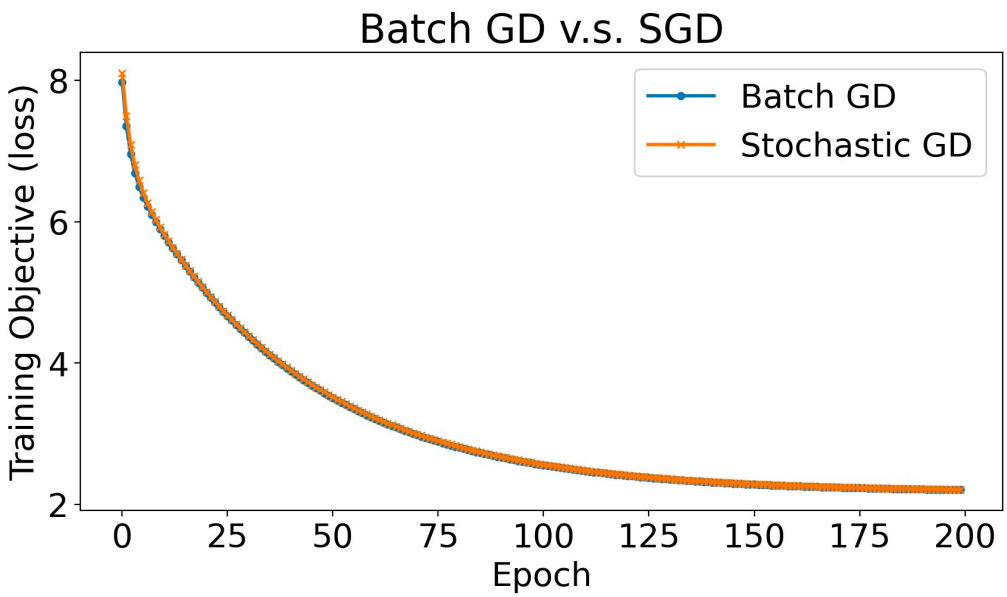
minimize $\sum_{n=1}^{N} \log\left(1+\exp\left(-y^{(n)} \cdot w^T \phi(x^{(n)})\right)\right)$

Hence Proved!

2 — 2.1 — GD and SGD

a) Auto grader

b)



GD version took 0.00 seconds

GD Test objective = 2.7017

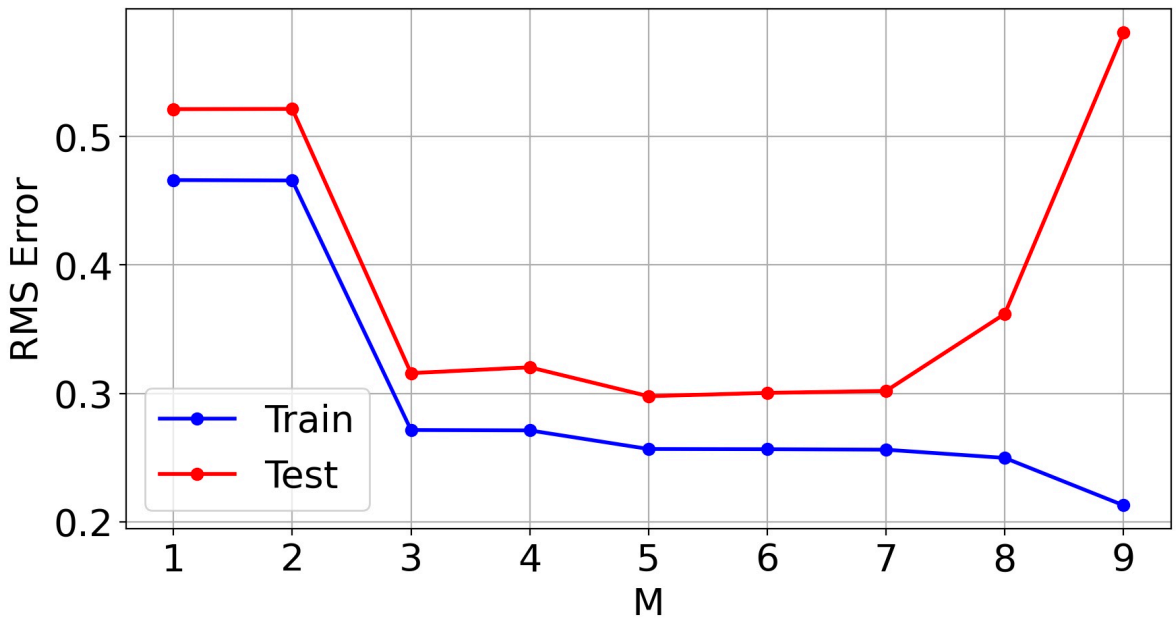
SGD version took 0.02 seconds

SGD Test objective = 2.6796

2.2 Over-fitting Study

(a) Autograder

(b)



(c) Degree of 5 polynomial will best fit the data. Yes, there was an evidence for both over/under-fitting.

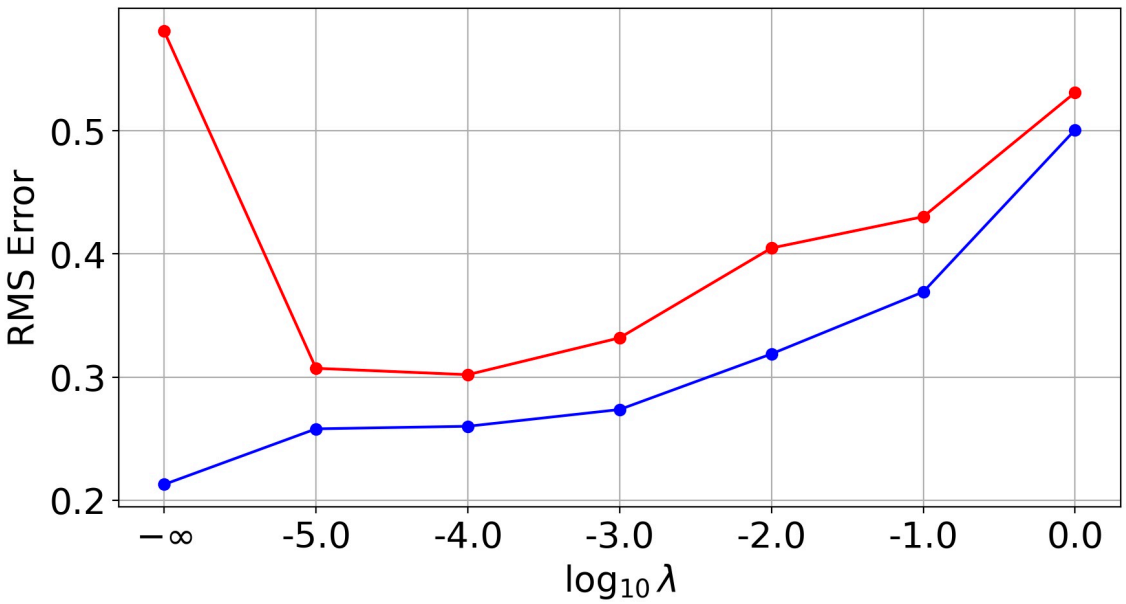
For $M > 7$, its an evidence of over-fitting

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

2.3 Regularization

(a) Autograder.

(b)



(c) $\lambda = 10^{-4}$ seems to best work for 9^{th} degree polynomial as from plot we can see, RMS Error is min at $\log_{10}\lambda = -4.0$

(3) Locally weighted linear regression

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

$r^{(i)} \in \mathbb{R}$ is local weight for sample $(x^{(i)}, y^{(i)})$

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

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

and $Y \in \mathbb{R}^{N \times 1}$, $w \in \mathbb{R}^{D \times 1}$

So, $$E_D(w) = (w^T X - Y^T) R (w^T X - Y^T)^T$$

here R is a diagonal matrix

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

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

as R is a diagonal matrix

$$= \sum_{i=1}^{N} (w^T X - Y^T)_i^2 R_{ii} \quad \Big| \; R_{ii} = \frac{1}{2} r^{(i)}$$

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

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

Hence Proved

(b)

$$E_D(w) = (w^T X - Y^T) R (w^T X - Y^T)^T$$

$(N \times N)$ $(N \times 1)$

$$\nabla_w E_D(w) = (w^T X - Y^T) R \nabla_w (w^T X - Y^T)^T$$
$$= (w^T X - Y^T) R X^T$$

$X \in \mathbb{R}^{D \times N}$
$R \in \mathbb{R}^{N \times N}$
$Y \in \mathbb{R}^{N \times 1}$
$w \in \mathbb{R}^{D \times 1}$

$$\Rightarrow (w^T X - Y^T) R X^T = 0$$
$$\Rightarrow w^T X R X^T - Y^T R X^T = 0$$
$$\Rightarrow w^T X R X^T = Y^T R X^T$$
$$\Rightarrow (w^T X R X^T)^T = (Y^T R X^T)^T$$
$$\Rightarrow (X R X^T) w = (R X^T)^T Y$$
$$\Rightarrow \boxed{w^* = (X R X^T)^{-1} (R X^T)^T Y}$$

$X \in \mathbb{R}^{D \times N}$, $R \in \mathbb{R}^{N \times N}$, $Y \in \mathbb{R}^{N \times 1}$

$w \in \mathbb{R}^{D \times 1}$

(c)

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

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

Eq

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

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

Taking log.

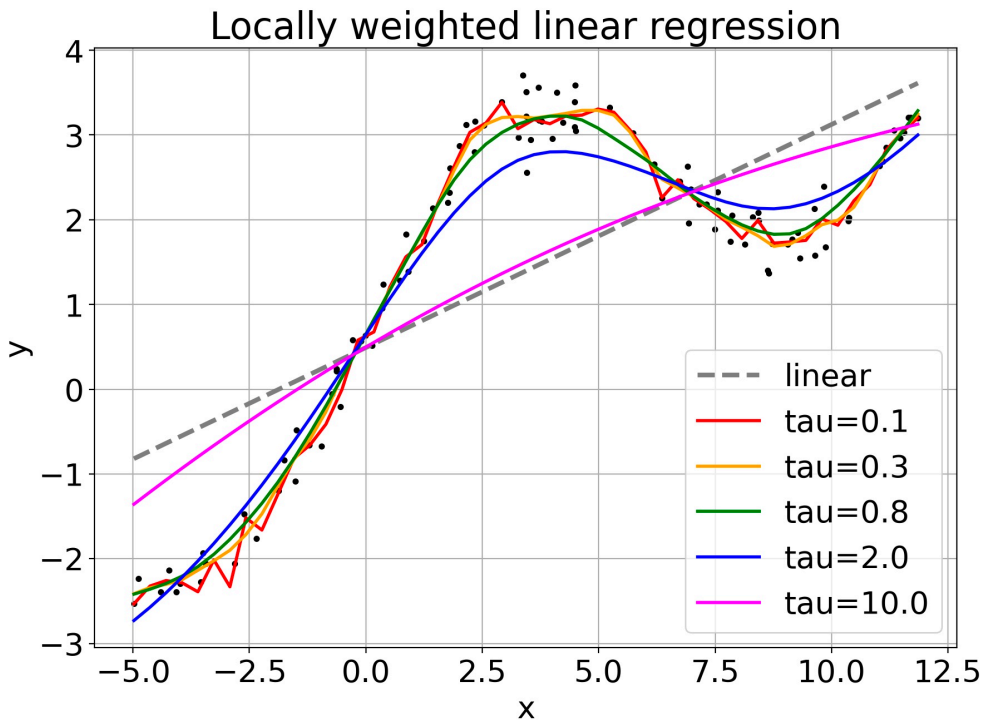
$$\log\left(p(Y \mid X, w, \sigma)\right) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\,\sigma^{(i)}}\right) - \sum_{i=1}^{n} \frac{\left(y^{(i)} - w^T x^{(i)}\right)^2}{2\left(\sigma^{(i)}\right)^2}$$

Hence problem is reduced to solving a weighted linear regression problem $E_D(w)$ with weight $w^{(i)} = \frac{1}{(\sigma^{(i)})^2}$

(d)

(i) Autograder.

(ii)



(iii) when τ is too small, fit is becoming over-fit and when τ is too big fit is becoming under-fit.

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

when τ is too small, it is excluding the point that increase confidence in the fit.