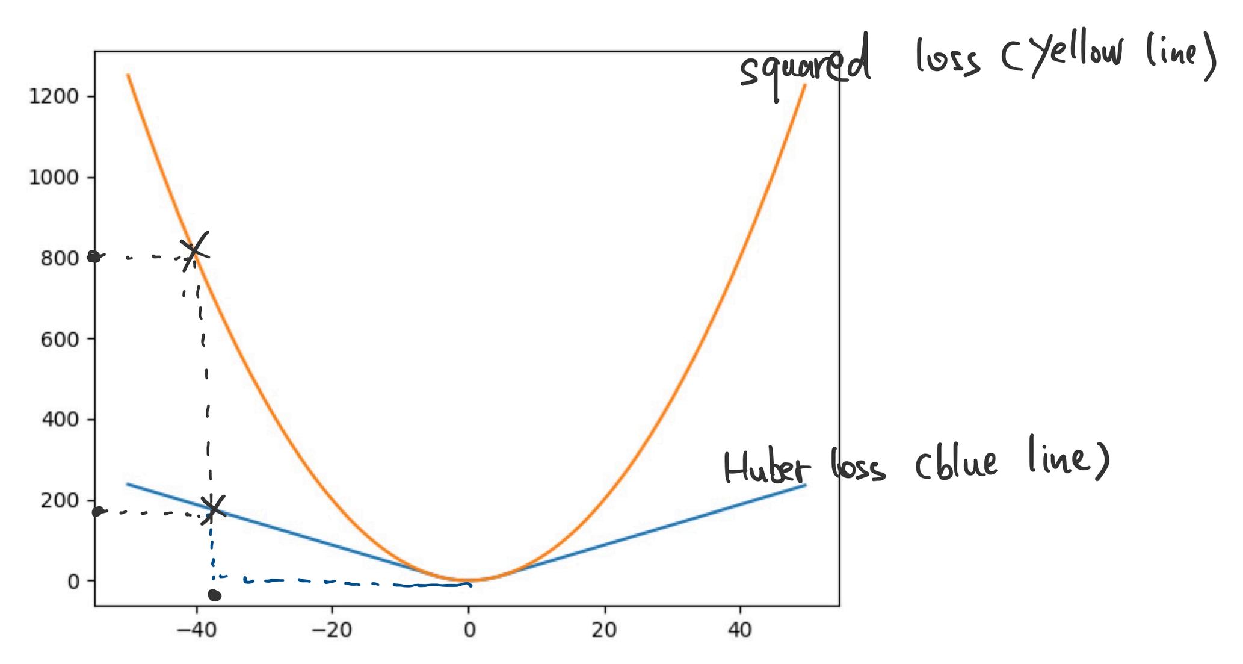
2018年10月12日 星期五　　下午7:51

Q1 (a)



Ans: within Huber loss, we have a delta value to classify a, when $|a| > \delta$. It's affect is linear. Compare with squared loss, Whether it is correct value or outliers. It will has exponential growth rate. As we can see on the graph., assume we have outliers $= -40$. The y-value of huber loss is significant smaller than the y-value of squared loss.

Q1 (b)

$$\because H'_\delta(a) = \begin{cases} a & \text{if } |a| \le \delta \\ \delta & \text{if } a > \delta \\ -\delta & \text{if } a < \delta \end{cases}$$

$$\because H'_\delta(y-t) = \begin{cases} y-t & \text{if } |y-t| \le \delta \\ \delta & \text{if } y-t > \delta \\ -\delta & \text{if } y-t < -\delta \end{cases}$$

$$\frac{\partial L_\delta}{\partial w} \begin{cases} ① \dfrac{\partial L_\delta}{\partial w} = \dfrac{dL_\delta}{dy}\dfrac{\partial y}{\partial w} = \dfrac{d}{dy}\left[\frac{1}{2}(y-t)^2\right] \cdot x_i \\ \qquad = (y-t)\,x_i \quad \text{if } |y-t| \le \delta \\ ② \dfrac{\partial L_\delta}{\partial w} = \dfrac{dL_\delta}{dy}\dfrac{\partial y}{\partial w} = \dfrac{d}{dy}\left(\delta\left((y-t) - \frac{1}{2}\delta^2\right)\right) x_i \\ \qquad = \delta x_i \quad \text{if } y-t > \delta \\ ③ \dfrac{\partial L_\delta}{\partial w} = \dfrac{dL_\delta}{dy}\dfrac{\partial y}{\partial w} = \dfrac{d}{dy}\left(-\delta\left((y-t) - \frac{1}{2}\delta^2\right)\right) x_i \\ \qquad = -\delta x_i \quad \text{if } y-t < \delta \end{cases}$$

$$\therefore \frac{\partial L_\delta}{\partial b} \begin{cases} ④ \dfrac{\partial L_\delta}{\partial b} = \dfrac{dL_\delta}{dy}\dfrac{\partial y}{\partial b} = \dfrac{d}{dy}\left[\frac{1}{2}(y-t)^2\right] \cdot \\ \qquad = (y-t) \quad \text{if } |y-t| \le \delta \\ ⑤ \dfrac{\partial L_\delta}{\partial b} = \dfrac{dL_\delta}{dy}\dfrac{\partial y}{\partial b} = \dfrac{d}{dy}\left(\delta\left((y-t) - \frac{1}{2}\delta^2\right)\right) \\ \qquad = \delta \quad \text{if } y-t > \delta \\ ⑥ \dfrac{\partial L_\delta}{\partial b} = \dfrac{dL_\delta}{dy}\dfrac{\partial y}{\partial b} = \dfrac{d}{dy}\left(-\delta\left((y-t) - \frac{1}{2}\delta^2\right)\right) \\ \qquad = -\delta \quad \text{if } y-t < \delta \end{cases}$$

$$\therefore \text{ we have } \frac{\partial L_\delta}{\partial w} = \begin{cases} (y-t)\,x_i & \text{if } |y-t| \le \delta \\ \delta x_i & \text{if } y-t > \delta \\ -\delta x_i & \text{if } y-t < \delta \end{cases}$$

$$\text{And } \frac{\partial L_\delta}{\partial b} = \begin{cases} (y-t) & \text{if } |y-t| \le \delta \\ \delta & \text{if } y-t > \delta \\ -\delta & \text{if } y-t < \delta \end{cases}$$

Q2 (a)

$$\because w^* = \underbrace{\arg\min \frac{1}{2}\sum_{i=1}^{N} a^{(i)}\left(y^{(i)} - w^T x^{(i)}\right)^2}_{①} + \underbrace{\frac{\lambda}{2}\|w\|^2}_{②}$$

$\therefore$ define a vector $v = y - Xw$, therefore the first term can be rewrite as $\frac{1}{2}\sum_{j=1}^{N} v_j^2 a^j$. Because we also know that A is a diagonal matrix where $A_{ii} = a^{(i)}$. Then $[Av]_j = a^{(j)} v_j$, and inner product $\langle v, Av \rangle = v^T A v = \sum_{j=1}^{N} v_j^2 a^{(j)}$.

$\therefore$ Moreover, we can rewrite the loss function as form of vector:

$$\begin{aligned} \therefore L(w) &= \frac{1}{2}(y - Xw)^T A (y - Xw) + \lambda \frac{1}{2}\|w\|^2 \\ &= \frac{1}{2}\left[y^T - (Xw)^T\right] A (y - Xw) + \lambda \frac{1}{2}\|w\|^2 \\ &= \frac{1}{2}\left[y^T - w^T X^T\right] A (y - Xw) + \lambda \frac{1}{2}\|w\|^2 \\ &= \frac{1}{2}\left[(y^T A - w^T X^T A)(y - Xw) + \lambda w^T w\right] \\ &= \frac{1}{2}\left[y^T A y - y^T A X w - w^T X^T A y + w^T X^T A X w + \lambda w^T w\right] \\ &= \frac{1}{2}\left[y^T A y - (y^T A X w)^T - w^T X^T A y + w^T X^T A X w + \lambda w^T w\right] \\ &= \frac{1}{2}\left[y^T A y - w^T X^T A y - w^T X^T A y + w^T X^T A X w + \lambda w^T w\right] \\ &= \frac{1}{2} y^T A y - w^T X^T A y + \frac{1}{2} w^T X^T A X w + \frac{1}{2}\lambda w^T w \end{aligned}$$

$\because \nabla_w w^T w = 2w$

$\because \nabla_w w^T A w = 2Aw$, we know that $X^T A X$ is symmetric.

since $X^T_{1\times n} A_{n\times n} X_{n\times 1}$ = matrix of 1×1.

$\because \nabla_w w^T x = x$

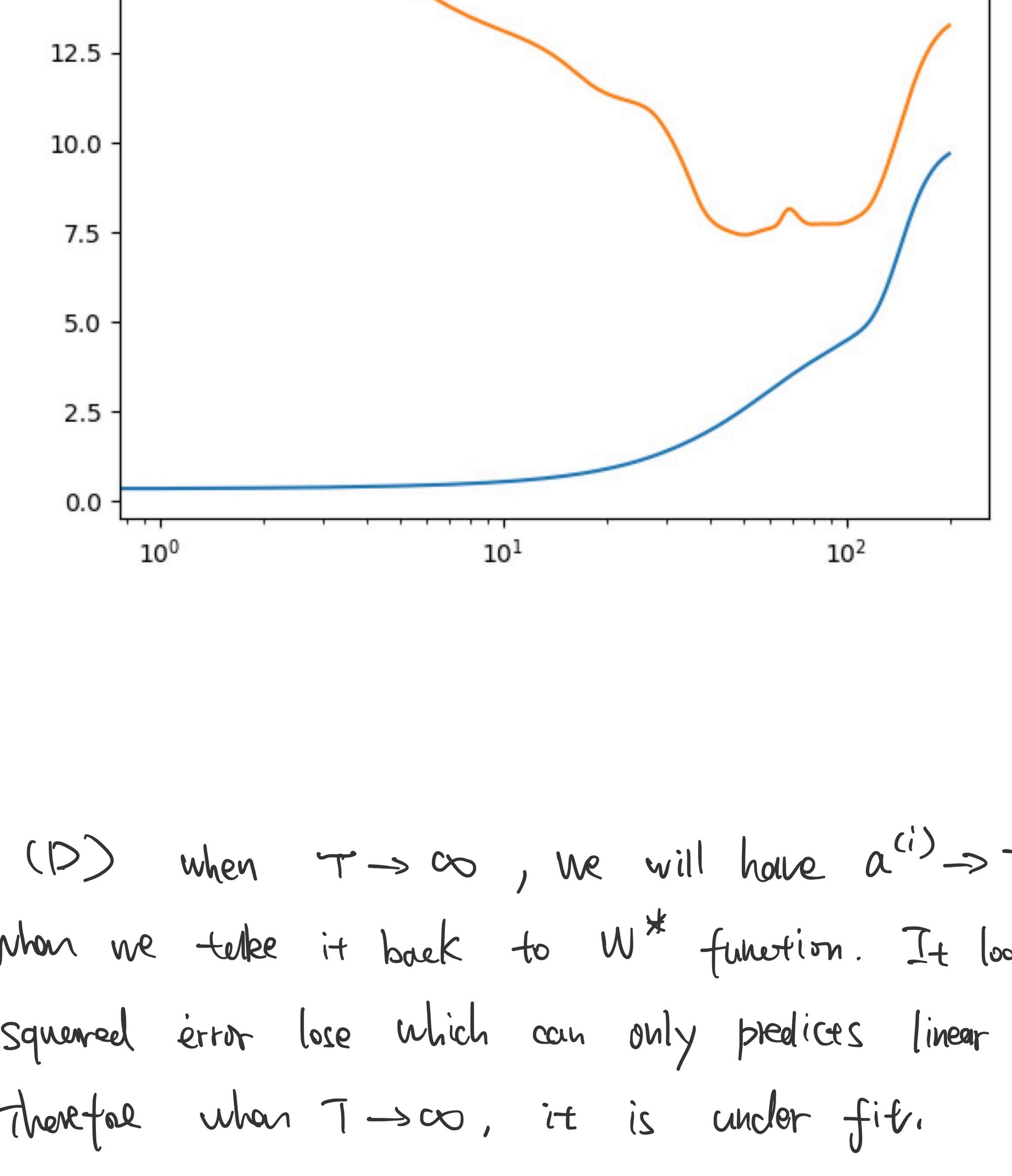
$$\begin{aligned} \therefore \text{Compute } \nabla_w L(w) &= \nabla_w \frac{1}{2} y^T A y - \nabla_w w^T X^T A y + \nabla_w \frac{1}{2} w^T X^T A X w + \nabla_w \frac{1}{2}\lambda w^T w \\ &= 0 - X^T A y + X^T A X w + \lambda w \\ &= -X^T A y + X^T A X w + \lambda w \end{aligned}$$

when we setting $\nabla_w L(w) = 0$, we have optimal w^*

$$\begin{aligned} -X^T A y + X^T A X w^* + \lambda w^* &= 0 \\ X^T A X w^* + \lambda w^* &= X^T A y \\ (X^T A X + I\lambda)\, w^* &= X^T A y \\ w^* &= (\,{}^T A X + I\lambda)^{-1} X^T A y \end{aligned}$$

proved!

Q2 (c):



Q2 (D) when $\tau \to \infty$, we will have $a^{(i)} \to \frac{1}{n}$, therefore when we take it back to w^* function. It looks like squared error lose which can only predices linear model. Therefore when $\tau \to \infty$, it is under fit.

when $\tau \to 0$, we will have $a^{(i)} \to$ large value. Therefore it will over-fit the training sample. Which means it will increases loss of validation set. So when $\tau \to 0$, it is over fit.