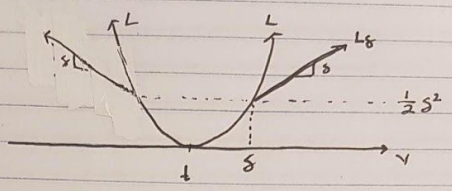
CSC411 Homework 3

CSC411 HW3 Feb 5, 2019

1. a) $L_\delta(y,t) = \begin{cases} \frac{1}{2}(y-t)^2 & |y-t| \le \delta \\ \delta(|y-t| - \frac{1}{2}\delta) & |y-t| > \delta \end{cases}$



After $(y-t) > \delta$, the loss only increases linearly, and thus it influences the classifier less.

b) $\rightarrow \frac{\partial L_\delta}{\partial w} = \frac{\partial H_\delta}{\partial y} \cdot \frac{\partial y}{\partial w}$

$y = w_1 x_1 + \ldots + w_n x_n$

$\frac{\partial y}{\partial w} = [x_1 \; x_2 \; \cdots \; x_n]$

$\frac{\partial y}{\partial w} = x^T$

Letting $a = y - t$

$\frac{\partial H_\delta}{\partial y} = \frac{\partial H_\delta}{\partial a} \frac{\partial a}{\partial y}^T$

$$\frac{\partial H_\delta}{\partial a} = \begin{cases} a & |a| \le \delta \\ \delta & |a| > \delta, \; a > \delta \\ -\delta & |a| > \delta, \; a < -\delta \end{cases}$$

$$\therefore \frac{\partial L_\delta}{\partial w} = \begin{cases} (y-t)x^T & |y-t| \le \delta \\ \delta x^T & |y-t| > \delta, \; |y-t| > \delta \\ -\delta x^T & |y-t| > \delta, \; |y-t| < -\delta \end{cases}$$

$\frac{\partial L_\delta}{\partial b} = \frac{\partial H_\delta(a)}{\partial a} \frac{\partial a}{\partial y}^T \frac{\partial y}{\partial b}^T$

$$= \begin{cases} y-t & |y-t| < \delta \\ \delta & |y-t| > \delta \quad y-t > \delta \\ -\delta & |y-t| > \delta \quad y-t < -\delta \end{cases}$$

2. a) $W^* = \underbrace{\text{argmin}\ \frac{1}{2}\sum_{i=1}^{n} a^{(i)}(y^{(i)} - W^T x^{(i)})^2 + \frac{\lambda}{2}\|W\|^2}_{C}$

$$Y = \begin{bmatrix} y^1 \\ \vdots \\ y^N \end{bmatrix} \qquad XW = \begin{bmatrix} W^T x^1 \\ \vdots \\ W^T x^N \end{bmatrix}$$

$$\left(\begin{bmatrix} a^1 & & \\ & \ddots & \\ & & a^N \end{bmatrix} \begin{bmatrix} y_1 - W^T x^1 \\ \vdots \\ y^N - W^T x^N \end{bmatrix} \right)^T \begin{bmatrix} y_1 - W^T x^1 \\ \vdots \\ y^N - W^T x^N \end{bmatrix} = [A(y - XW)]^T (y - XW) = (y - XW)^T A^T (y - XW)$$
$$= (y - XW)^T A (y - XW)$$

$$\therefore C = \frac{1}{2}(y - XW)^T A (y - XW) + \frac{\lambda}{2} W^T W$$

Let $B = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix}$

$$B^T A B = \begin{bmatrix} b_1 & \cdots & b_N \end{bmatrix} \begin{bmatrix} a^1 & & \\ & \ddots & \\ & & a^N \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix} = a^1 b_1^2 + \ldots + a^N b_N^2$$

$$\frac{d(B^T A B)}{dB} = \begin{bmatrix} 2a^1 b_1 & \cdots & 2a^N b_N \end{bmatrix}$$
$$= 2B^T A$$

Notice: if $A = I$

$$\frac{d(B^T B)}{dB} = 2B^T$$

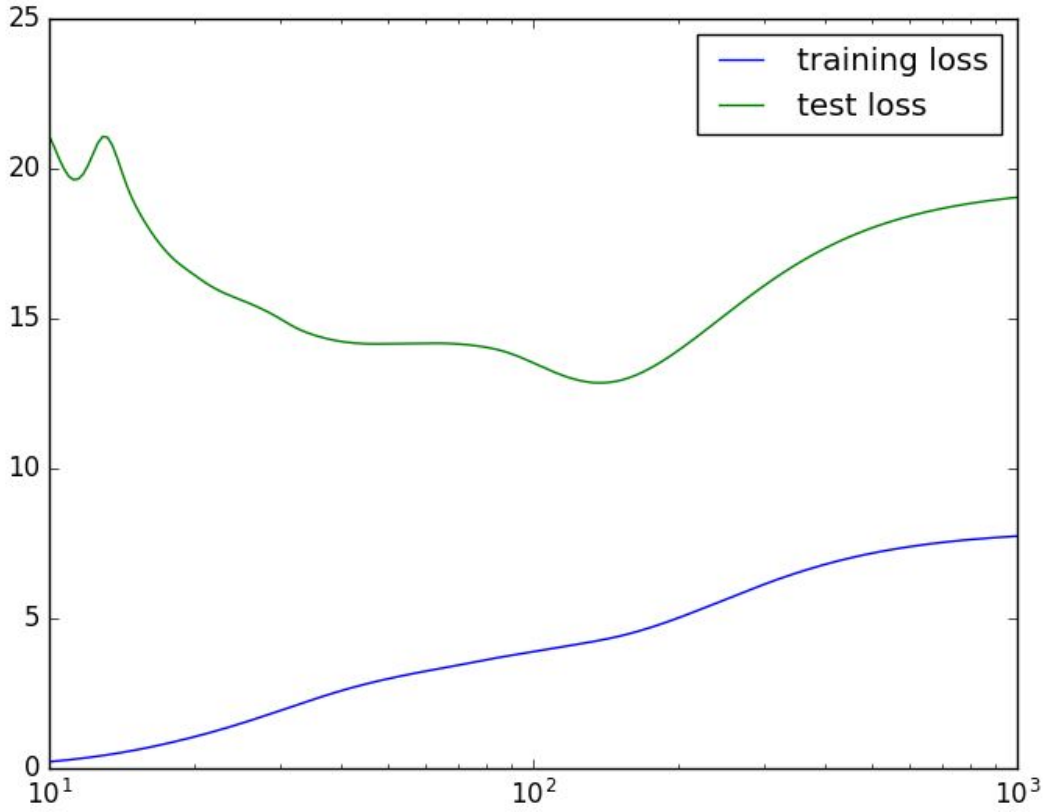
$$XW = \begin{bmatrix} W^T x^1 \\ \vdots \\ W^T x^N \end{bmatrix} \qquad \frac{d(XW)}{dW} = X$$

$$\frac{dC}{dW} = \frac{1}{2} \cdot \frac{d\left((y - XW)^T A (y - XW)\right)}{d(y - XW)} \cdot \frac{d(y - XW)}{dW} + \frac{\lambda}{2} \frac{d(W^T W)}{dW}$$

$$= \frac{1}{2} \cdot 2 \cdot (y - XW)^T A (-X) + \frac{\lambda}{2} \cdot 2 \cdot W^T$$
$$= -(y - XW)^T A X + \lambda W^T$$

Setting to 0

$$[0] = -(Y^T - W^T X^T) A X + \lambda W^T$$
$$[0] = -Y^T A X + W^T X^T A X + \lambda W^T$$
$$Y^T A X = W^T (X^T A X + \lambda I)$$
$$X^T A Y = (X^T A X + \lambda I) W$$
$$\boxed{W = (X^T A X + \lambda I)^{-1} X^T A Y}$$ as required

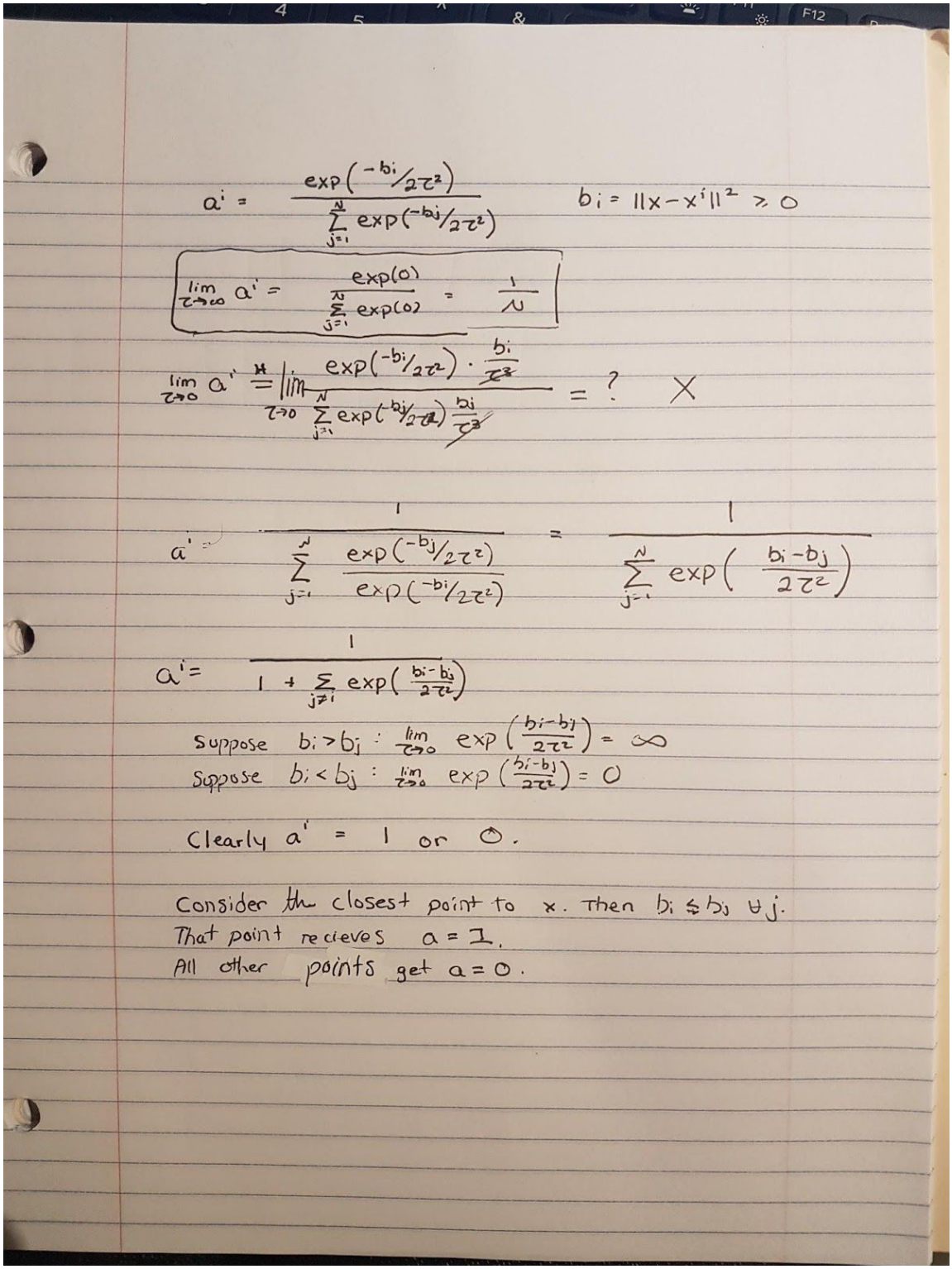


Training and test loss v.s. tau value

2d) As tau goes to infinity, all residuals get an equal weighting and the LRLS algorithm behaves like linear regression. Thus, the training loss increases because points further away influence the local decision. As tau goes to 0, only the closest points are considered. It is locally correct at the expense of points further away. Thus, the line is drawn better for each specific point and the training loss goes down. This is precisely the behaviour we see. A proof of behaviour as tau goes to 0 and infinity is on the next page.

Interestingly, the validation/test loss achieves a minimum somewhere in the middle. This makes sense. As tau goes to 0, the line is overfit and error increases. As tau goes to infinity, the line is locally underfit, and error increases.

Interpreting Tau



$$a^i = \frac{\exp(-b_i/2\tau^2)}{\sum_{j=1}^{N} \exp(-b_j/2\tau^2)} \qquad b_i = \|x - x^i\|^2 \geq 0$$

$$\lim_{\tau \to \infty} a^i = \frac{\exp(0)}{\sum_{j=1}^{N} \exp(0)} = \frac{1}{N}$$

$$\lim_{\tau \to 0} a^i \overset{H}{=} \lim_{\tau \to 0} \frac{\exp(-b_i/2\tau^2) \cdot \frac{b_i}{\tau^3}}{\sum_{j=1}^{N} \exp(-b_j/2\tau^2) \frac{b_i}{\tau^3}} = ? \quad \times$$

$$a^i = \frac{1}{\sum_{j=1}^{N} \frac{\exp(-b_j/2\tau^2)}{\exp(-b_i/2\tau^2)}} = \frac{1}{\sum_{j=1}^{N} \exp\left(\frac{b_i - b_j}{2\tau^2}\right)}$$

$$a^i = \frac{1}{1 + \sum_{j \neq i} \exp\left(\frac{b_i - b_j}{2\tau^2}\right)}$$

Suppose $b_i > b_j$: $\lim_{\tau \to 0} \exp\left(\frac{b_i - b_j}{2\tau^2}\right) = \infty$

Suppose $b_i < b_j$: $\lim_{\tau \to 0} \exp\left(\frac{b_i - b_j}{2\tau^2}\right) = 0$

Clearly $a^i = 1$ or 0.

Consider the closest point to x. Then $b_i \leq b_j \ \forall j$.
That point recieves $a = 1$.
All other points get $a = 0$.