$$p(y_1, y_2 \cdots y_m | x_1 \cdots x_m) = \prod_{i=1}^{m} P(y_i / x_i)$$

$$\max \log \left(\prod_{i=1}^{m} P(y_i / x_i) \right)$$

$$\cancel{\min -\log (} \quad \min \sum_{i=1}^{m} -p \log P(y_i / x_i)$$

$$y(\theta) = -\log p(y/x)$$

$\downarrow$

$w_1 \, w_2 \cdots$

$$= -\log \sigma((2y-1)z)$$

$$= \zeta(-(2y-1)z)$$

softplus

not activation

cost funct

$$= \zeta((1-2y)z)$$

$\sigma(z) = \frac{1}{1+e^{-z}}$

$= \frac{e^z}{1+e^z}$

$\frac{d}{dz}\sigma(z) = \sigma(1-\sigma)$

$\frac{\partial y(\theta)}{\partial z} =$

$\sigma((1-2y)z)(1-2y)$

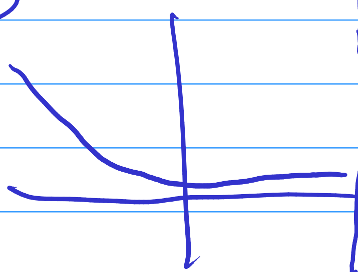
@ $y=1$ $\quad \zeta((1-2y)z) = \zeta(-z)$

when $z >> 0$

$\zeta(-z) \to 0$

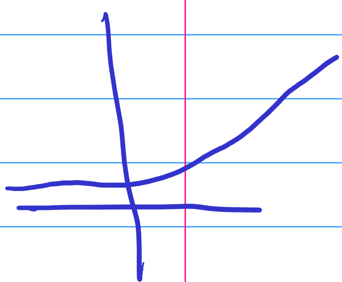
when $z << 0$

$\zeta(-z) \to$ large



@ $y=0$ $\quad z >> 0$

$\zeta(z) \Rightarrow$ large $\quad$ ∇ also large ($w^{(k+1)}$ changes quite alot)



$z << 0$

$\zeta(z) \to 0$ $\quad$ w's wont change

(Take 1 example, treat as SGD)

Why is this cost function any better than MSE

P(

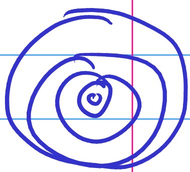
$$\max \log P(y_1, y_2 \dots y_m \mid x_1, x_2 \dots x_m)$$

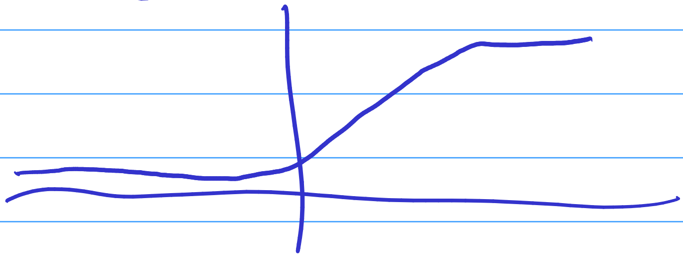
$$= \max \log \prod_{i=1}^{m} P(y_i \mid x_i)$$

$$= \max \log \prod \left[\sigma(z_i)^{y_i} (1-\sigma(z_i))^{1-y_i} \right]$$

$$= \max \sum \left(y_i \log \sigma(z_i) + (1-y_i) \log(1-\sigma(z_i)) \right)$$

$$= \min \left[\sum_i -\left(y_i \log \sigma(z_i) + (1-y_i) \log(1-\sigma(z_i)) \right) \right]$$





$$= \min -\left[y \log \sigma(z) + (1-y) \log(1-\sigma(z)) \right]$$