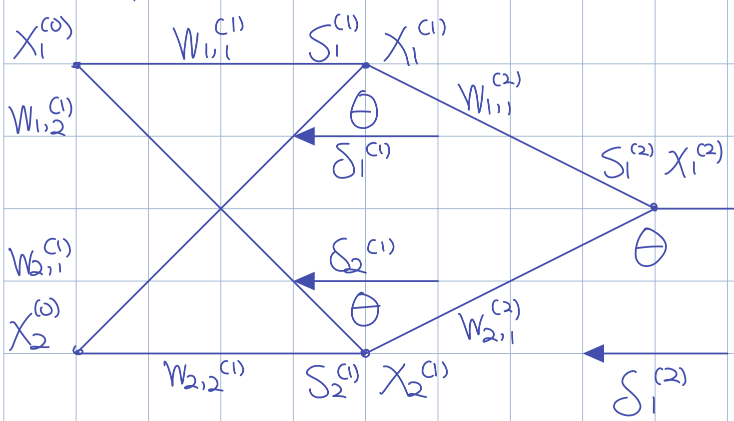
Example



$e(\ )$

$$e(\Omega) = (x_1^{(2)} - y)^2$$

Input Layer

Notations

$x_i^{(l)}$

$$s_i^{(l)} = \sum_{j=0}^{N} \left(x_j \, w_{j,i}^{(l-1)}\right)$$

$w_{i,j}^{(l)}$

$$\delta_i^{(l)} = \partial e / \partial s_i^{(l)}$$

Weight Connecting Node i In Layer l-1 To Node j In Layer l

Goal (Calculate Partial Derivative Of Error With Respect To Weights)

$$\theta'(s_1^{(2)})$$

$$\delta_1^{(2)} = \frac{\partial e}{\partial s_1^{(2)}} = \boxed{\frac{\partial e}{\partial x_1^{(2)}}} \times \boxed{\frac{\partial x_1^{(2)}}{\partial s_1^{(2)}}} = 2(x_1^{(2)} - y)\,\theta'(s_1^{(2)})$$

$$\frac{\partial}{\partial x_1^{(2)}}(x_1^{(2)}-y)^2 = 2(x_1^{(2)}-y)$$

$$\delta_1^{(1)} = \frac{\partial e}{\partial s_1^{(1)}} = \delta_1^{(2)} w_{1,1}^{(2)} \theta'(s_1^{(1)})$$

$$= \frac{\partial e}{\partial x_1^{(1)}} \times \underbrace{\frac{\partial x_1^{(1)}}{\partial s_1^{(1)}}}_{\theta'(s_1^{(1)})}$$

$$\frac{\partial e}{\partial x_1^{(1)}} = \underbrace{\frac{\partial e}{\partial s_1^{(2)}}}_{\delta_1^{(2)}} \times \underbrace{\frac{\partial s_1^{(2)}}{\partial x_1^{(1)}}}_{w_{1,1}^{(2)}}$$

$$\delta_2^{(1)} = \delta_1^{(2)} w_{2,1}^{(2)} \theta'(s_2^{(2)})$$

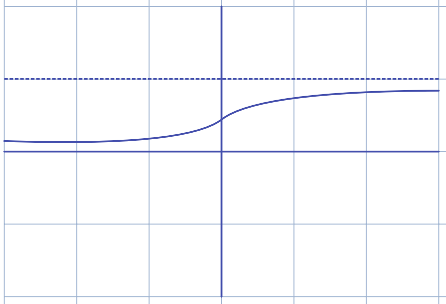
$$\frac{\partial e}{\partial w_{1,1}^{(1)}} = \underbrace{\frac{\partial e}{\partial s_1^{(1)}}}_{\delta_1^{(1)}} \times \underbrace{\frac{\partial s_1^{(1)}}{\partial w_{1,1}^{(1)}}}_{x_1^{(0)}}$$

Implementation of NN

θ (Activation Function) (Brings Non Linearity Into NN)

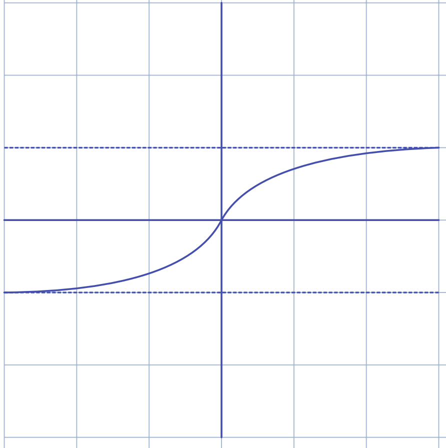
Sigmoid

$$\theta(s) = \frac{e^s}{1 + e^s} = \frac{1}{1 + e^{-s}}$$



tanh

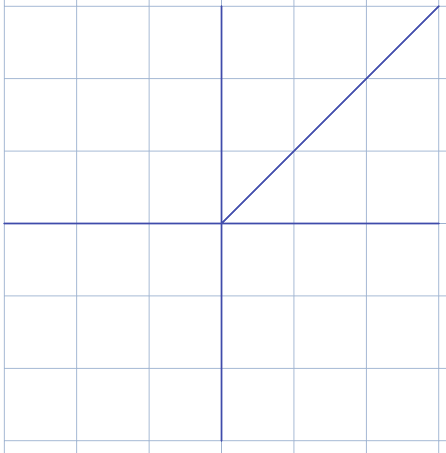
$$\theta(s) = \frac{e^s - e^{-s}}{e^s + e^{-s}}$$



Vanishing Gradient (Slow Learning)

ReLU

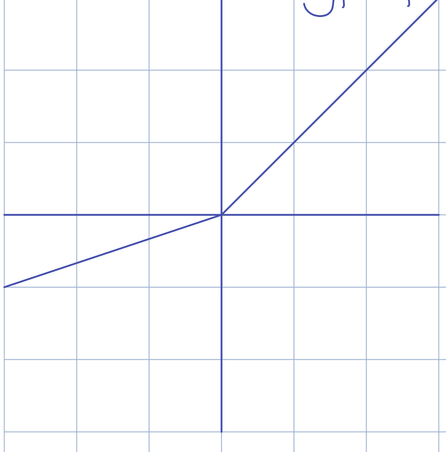
$$\theta(s) = \max(0, s)$$



Leaky ReLU

$$\theta(s) = \begin{cases} s, & s \geq 0 \\ \alpha s, & s < 0 \end{cases}$$

$\alpha \approx 0.1$ (Hyperparameter)



Data Augmentation

(Generate More Data)

Given $(\underline{x}, y)$ also include $(\underline{F(x)}, y)$

Weights Initialization (Section 8.4 DL Textbook)

$W_{i,j}^{(l)} \longrightarrow N(0, 10^{-3})$ N: Normal Distribution

Xavier Initialization

$$S_j^{(l)} = \sum_{i=1}^{n} x_i^{(l-1)} W_{i,j}^{(l)} \qquad W_{i,j}^{(l)} \sim N\left(0, \frac{1}{n}\right)$$

$$Var(S_j^{(l)}) \approx 1$$

If using ReLU $\longrightarrow W_{i,j}^{(l)} \sim N\left(0, \frac{2}{n}\right)$

Input Preprocessing

$D = \{(\underline{x_1}, y_1), (\underline{x_2}, y_2), \ldots, (\underline{x_n}, y_n)\}$

Sample Mean

$$\underline{M} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Sample Variance

$$\sigma(j) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i(j) - y(j))}$$

$j = 1, 2, 3, \ldots, d$

$$X_i'(j) = \frac{X_i(j) - M(j)}{\sigma(j)}$$ Zero Mean Unit Vector