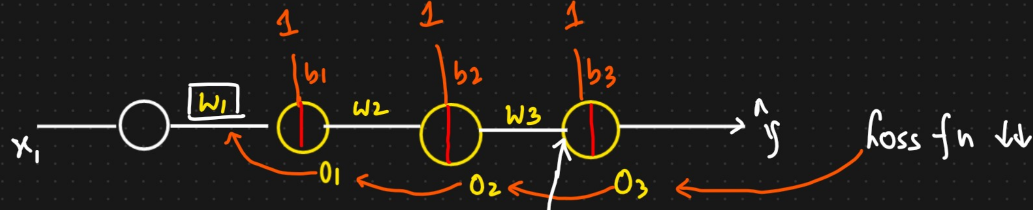
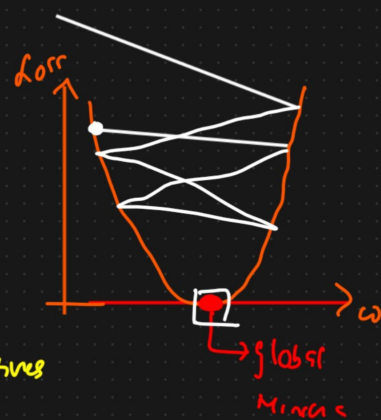
Exploding Gradient Problem ⇒ Weight Initialization Technique



$\rightarrow z = (O_2 * w_3 + b_3) \rightarrow \sigma(z)$

Optimizers



$$W_{1new} = W_{1old} - \boxed{\eta \boxed{\frac{\partial L}{\partial W_{1old}}}}$$

$\left\{ \begin{array}{l} W_{1new} >>>> W_{1old} \\ W_{1new} <<<< W_{1old} \end{array} \right\}$

⇗ Chain Rule of Derivatives

$$\frac{\partial L}{\partial W_{1old}} = \frac{\partial L}{\partial O_3} * \boxed{\frac{\partial O_3}{\partial O_2}} * \frac{\partial O_2}{\partial O_1} * \frac{\partial O_1}{\partial W_{1old}}$$

big * big * big * big ⇒ big Value

Weight initialization

⇓

Very high value

$$\frac{\partial O_3}{\partial O_2} = \boxed{\frac{\partial \sigma(z)}{\partial z}} * \frac{\partial z}{\partial O_2}$$

$$= [0 - 0.25] * \frac{\partial (O_2 * w_3 + b_3)}{\partial O_2}$$

$$= [0 - 0.25] * w_3 \Rightarrow 500 - 1000$$

Weight Initializing Techniques

① Uniform Distribution ✓

② Xavier / Glorot Initialization ✓

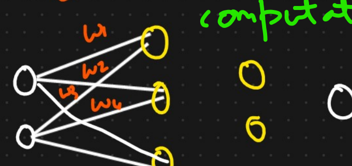
③ Kaiming He Initialization ✓

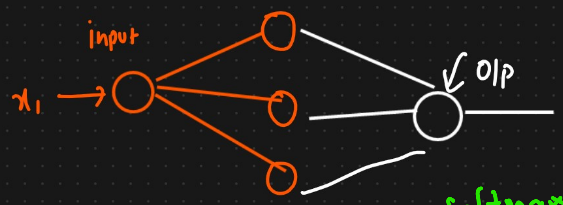
Key Points

① Weights should be small ✓ → avoid exploding gradient Problem

② Weights should not be same ✓ ← to avoid same computation

③ Weights should have good variance

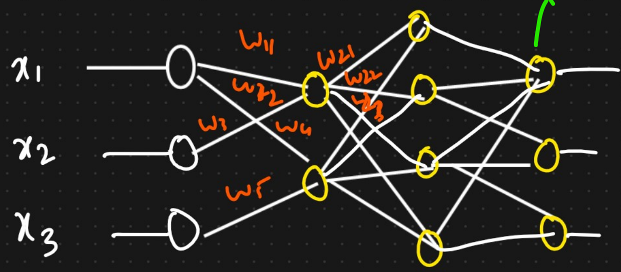




input = 1 (in count)

Output = 1 (in count)

softmax activation



input = 3 (in count)

O/P = 3 (in count)

① Uniform Distribution

$$W_{ij} \approx \text{Uniform Distribution} \in \left[\frac{-1}{\sqrt{\text{input}}}, \frac{1}{\sqrt{\text{input}}}\right]$$

$$\left[\frac{-1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right] \longrightarrow \text{for case above}$$

② Xavier/Glorot Initialization

Researcher ⟶ Xavier Glorot

① Xavier Normal Initialization

$$W_{ij} \approx \in N(0, \sigma)$$

$$\sigma = \sqrt{\frac{2}{(\text{input} + \text{output})}}$$

② Xavier Uniform

$$W_{ij} \approx \text{Uniform Distribution} \left[\frac{-\sqrt{6}}{\sqrt{\text{input} + \text{output}}}, \frac{\sqrt{6}}{\sqrt{\text{i/p} + \text{o/p}}}\right]$$

3 Kaiming He Initialization

1 He Normal

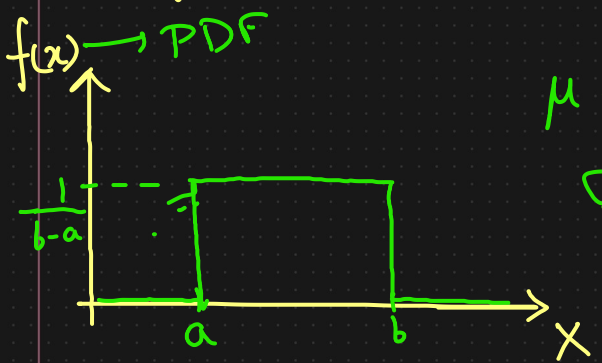
$$W_{ij} \sim N(0, \sigma)$$

$$\sigma = \sqrt{\frac{2}{input}}$$

2 He uniform

$$W_{ij} \sim \text{Uniform Distribution} \left[-\sqrt{\frac{6}{input}}, \sqrt{\frac{6}{input}} \right]$$

Uniform Distribution



$$\mu = \frac{a+b}{2}$$

$$\sigma = \sqrt{\frac{(b-a)^2}{12}}$$

area under curve

$$P(c \leq X \leq d) = \frac{d-c}{b-a}$$

For continuous random variables