WHAT YOU WILL STUDY IN TODAY VIDEO ?

- What is Vanishing Gradient Problem?
- What is Exploding Gradient Problem?
- Dying ReLu Problem

① Vanishing Gradient

MLP

loss grad → too small → weight updation difficult → Slow training

Causes

main

① Activation fn → Sigmoid, Tan h

② Too much hidden layers

N-1

1 2 3

small

multiply

N-2

1 2 3 4 5 6 7

too small

multiply

③ Weight initialization

③ Weight initialization
↳ Starting weights too small

fix:

① Better act fn → ReLu
→ Leaky ReLu

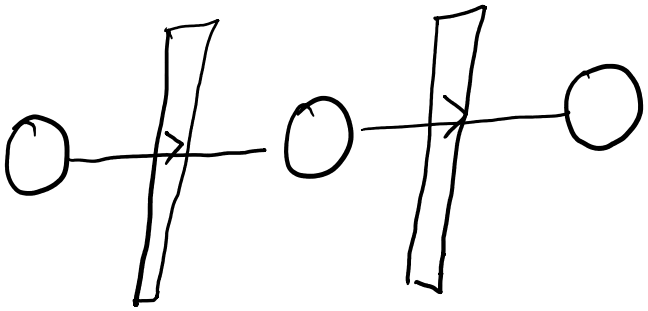
② Use Weight initialization techniques

Xavier/Glorot

He initialization

③ Batch normalization layer

Gradient don't shrink much



Mathematical

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

Derivative of sigmoid

$$\Rightarrow \sigma'(z) = \sigma(z)\left(1-\sigma(z)\right)$$

3 layers → $w = 0.5$

$x = 1$

1st layers → $z_1 = w \times x = 0.5 \times 1 = 0.5$

$a_1 = \frac{1}{1+e^{-0.5}} = 0.62$

2nd layer → $z_2 = 0.5 \times 0.62 = 0.31$

$a_2 = \frac{1}{1+e^{-0.31}} = 0.58$

3rd layer → $z_3 = 0.5 \times 0.58 = 0.29$

$a_3 = 0.57 \rightarrow$ O/P

Backpropagation

Compute gradient

3rd layer → $\sigma'(z_3) = 0.57 \times (1-0.57)$

$= 0.25$

2nd layer → $\sigma'(z_2) = 0.58 \times (1-0.58)$

$= 0.24$

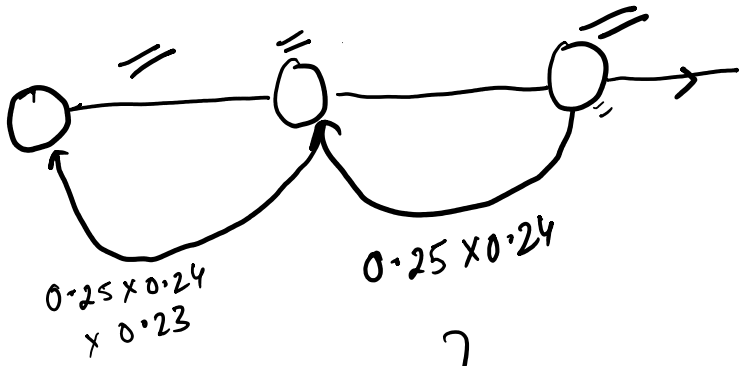
$\sigma'(z_1) = 0.62 \times (1-0.62)$

1st layer → $\sigma'(z_1) = 0.62 \times (1-0.62)$

$= 0.23$

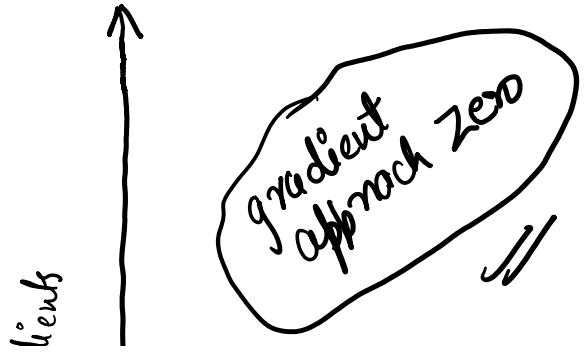
Overall gradient → $0.25 \times 0.24 \times 0.23$

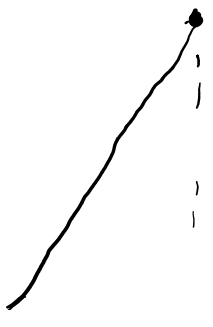
$= 0.0138$ → Small value

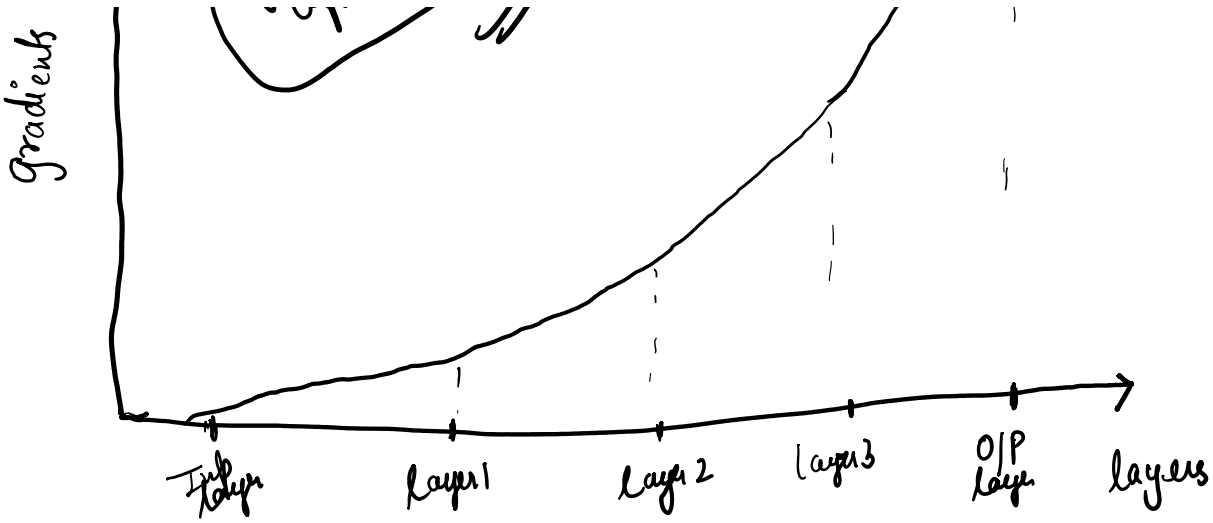


{No updation weight}

Graph





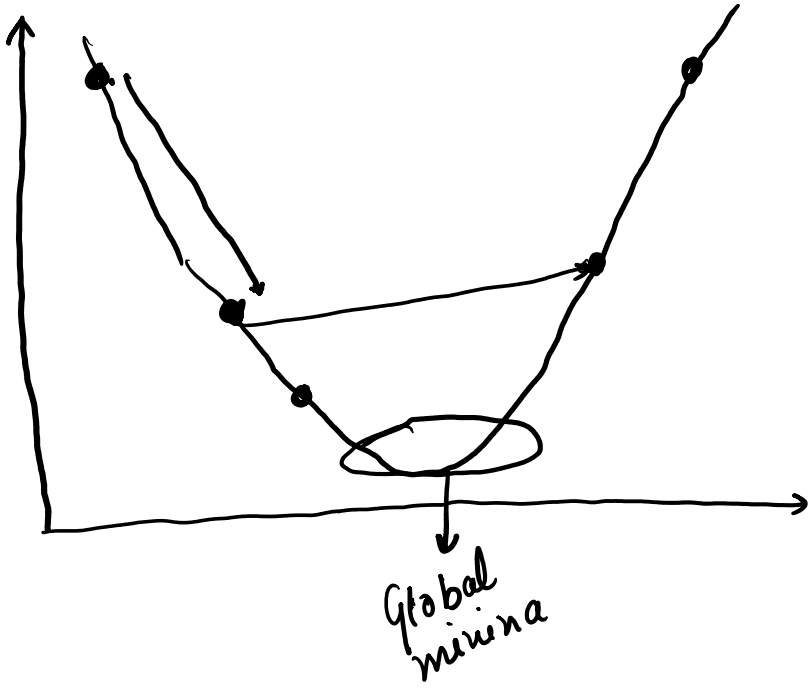

Gradients
Layer1
Layer2
Layer3
O/P
layers

② Exploding Gradient

big loss gradient → drastic change in weight updation ⇒

↓

Convergence fails.



Causes ⇒

① Weight initialization ⇒

↳ Starting weights are big ⇐

RNN

② Too many layers

③ High learning rate

$$w_{new} = w_{old} - \alpha \frac{dL}{dw}$$

fix

① Weight initialization

Xavier He

② Batch normalization layer

Mathematical

Assume, $w = 3$, $x = 1$

3 layered network

Act fn → Relu → $f(x) = \max(0, x)$

forward prop

1st layer → $z_1 = 3 \times 1 = 3$

$a_1 = \max(0, 3) = 3$

2nd layer $\longrightarrow$ $z_2 = 3 \times 3 = 9$

$a_2 = 9$

3rd layer $\longrightarrow$ $z_3 = 3 \times 9 = 27$

$a_3 = 27 \longrightarrow$ O/P

Back Prop.

Loss fn = MSE

$L = (y - \hat{y})^2$

3rd layer $\quad \frac{dL}{dz_3} = 1 \longrightarrow$ Loss grad at layer 3.

Loss grad w.r.t to weighted sum

2nd layer $\quad$ Apply chain rule $\longrightarrow \frac{dL}{dz_2} = w \times \frac{dL}{dz_3}$

$= 3 \times 1$

$= 3$

Loss grad for layer 2

1st $\quad \frac{dL}{dz_1} = w \times \frac{dL}{dz_2} = 3 \times 3$

1st layer

$$\frac{dL}{dz_1} = \cdots dz_2$$

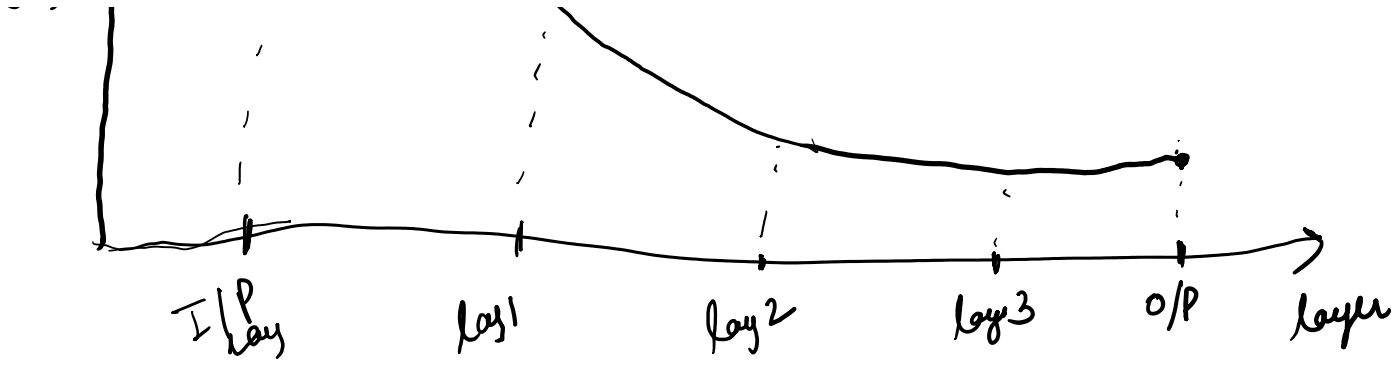
$$= 9.$$

& so on

27, - - - - -

(Drastic change) → unstable weight updation → model converge ✗

Graph


I/P
layer
layer1
layer2
layer3
O/P
layer

③ Dying Relu

Relu → max (0, z)

4 → 4
5 → 5
-1 → 0
-2 → 0
-4 → 0

Cause → Relu act fn

fix → leaky Relu, Parametric Relu, ELU