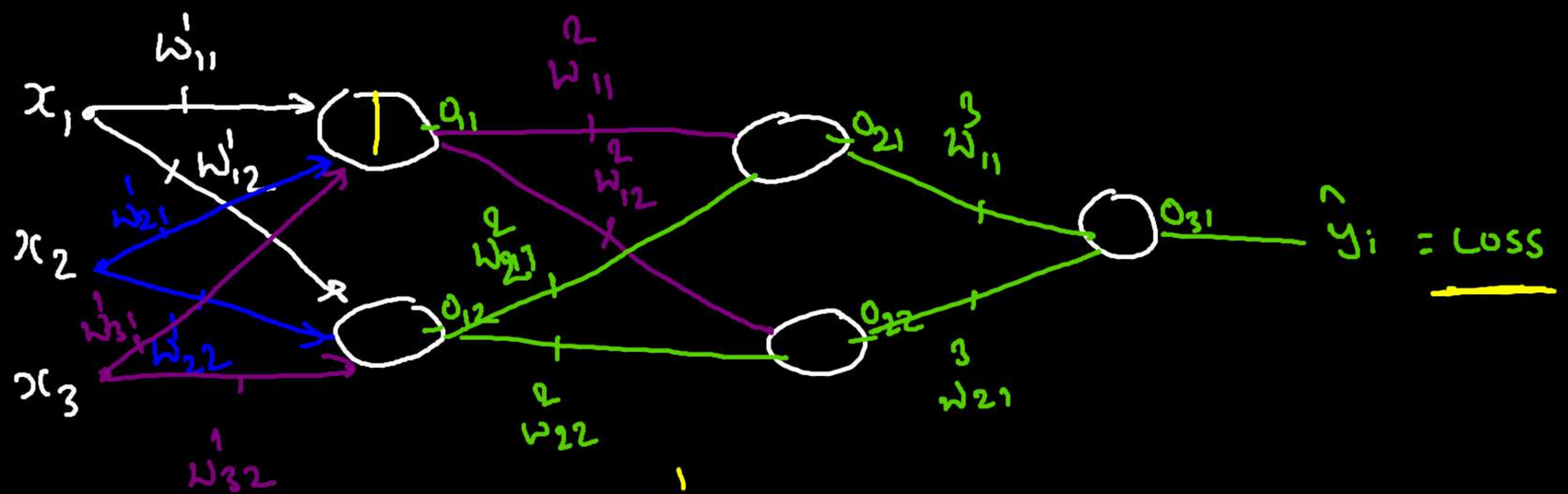
Weight Initialization - Deep MLP



Weight :-

$$I = \begin{bmatrix} w^1_{11} \\ w^1_{12} \\ w^1_{21} \\ w^1_{22} \\ w^1_{31} \\ w^1_{32} \end{bmatrix}_{6} \begin{matrix} \leftarrow 1 \\ \leftarrow 0.1 \\ \leftarrow 0.3 \\ \leftarrow 0.5 \\ \leftarrow 0 \\ \leftarrow 1 \end{matrix} \quad II = \begin{bmatrix} w^2_{11} \\ w^2_{12} \\ w^2_{21} \\ w^2_{22} \end{bmatrix}_{4} \leftarrow 1 \quad III = \begin{bmatrix} w^3_{11} \\ w^3_{21} \end{bmatrix}_{2} \leftarrow 1$$

$= 12$

Key insight

* Things not to do :-

(1) zero initialization → This is very very bad idea

$$\text{init } W_{ij}^{k} = 0 \quad \forall\, i\, j\, k$$

- → No training
- → All neurons compute the same thing
- → Same gradient updates

} Symmetric

Ensemble :- more different base models are, the better is the output of ensembling

(2) Non-zero constant initialization → linear

(3) Random initialization with small weight

✓ 0.0006) 0.0005 0.000000001) — —

convergence

$50 * 0.00001 = \frac{\partial L}{\partial O_{31}} * \frac{\partial O_{31}}{\partial O_{21}} * \frac{\partial O_{21}}{\partial O_{11}} * \frac{\partial O_{11}}{\partial w'_{11}}$

→ Vanishing Gradient Problem

(4) Random initialization with large weight

10 100 500 2525 — —

→ Exploding Gradient Problem

$W_0 = 50$, $W_1 = 50 - 10 * 100 = -950$

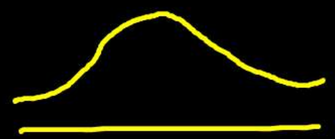
What can be done

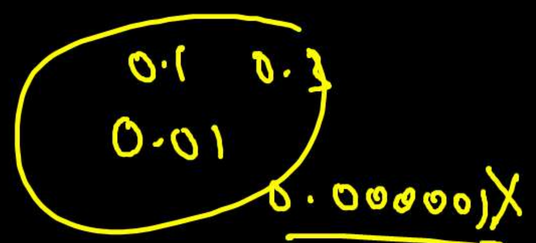
Heuristic Approach

Weight init

Uniform

Normal | Gaussian distribution





idea 1 :- Gaussian / Normal init

(a) Weight should be small (but not very small)

(b) All weight should not be zero

(c) Good - variance

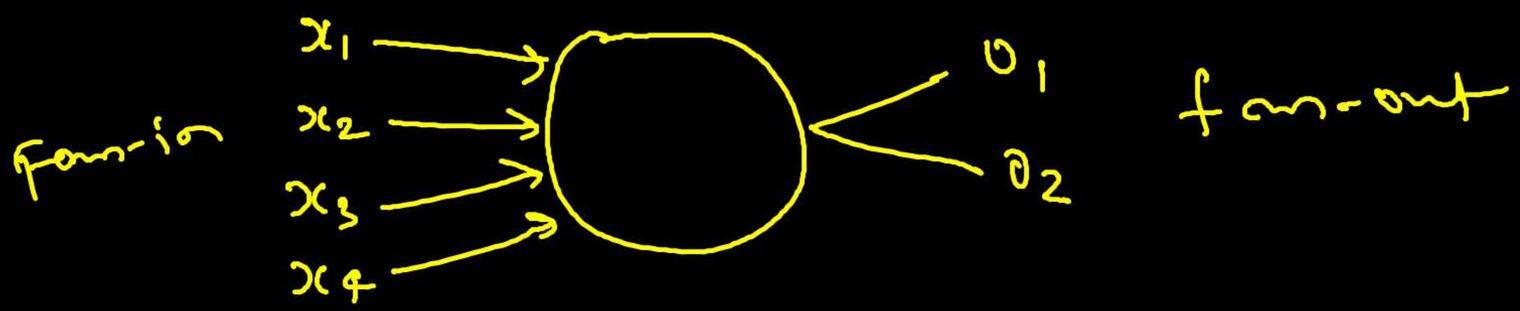
(d) weight should not be the same

→ better init strategies (Heuristic Approach)

→ lots of experiment

→ Theory / maths part is - Weak

fan - in fan - out



$$W_{ij}^{k} \longrightarrow \text{Unif}\left[\frac{-1}{\sqrt{fan\_in}}, \frac{1}{\sqrt{fan\_in}}\right]$$ Formula

$fan\_in = 4$

$fan\_out = 2$

$$\left[\frac{-1}{\sqrt{4}}, \frac{+1}{\sqrt{4}}\right] = \left[\frac{-1}{2}, \frac{1}{2}\right]$$

$W_{ij}^{k} = (-0.5, +0.5)$ — Range

W_{11}^{1}	W_{12}^{1}	W_{21}^{1}	W_{22}^{1}	W_{31}^{1}	W_{32}^{1}	W_{11}^{2}	W_{21}^{2}
0.4	-0.4	0.3	0.2	0.1	-0.1	-0.2	0.3

NOTE : There is no concrete agreement amongst all the researcher as which idea is the best one, because mathematics is fairly weak

Tanh & Sigmoid very popular

idea 2 :- 2010 - Xavier/Glorot init

(a) Xavier/Glorot Normal init

fan-in

fan-out

4

2

$$W_{ij}^{k} \sim N(0, \sigma_{ij}), \quad \sigma_{ij} = \pm\sqrt{\frac{2}{\text{fan-in} + \text{fan-out}}}$$

-0.5 to 0.5

$$\sqrt{\frac{2}{6}} = \sqrt{\frac{1}{3}} = \frac{1}{1.73}$$

(5) uniform Xavier / Glorot init

$$W_{ij}^{k} \sim U\left[\frac{-\sqrt{6}}{\sqrt{fan\text{-}in + fan\text{-}out}}, \frac{+\sqrt{6}}{\sqrt{fan\text{-}in + fan\text{-}out}}\right]$$

$fan\text{-}in = 4$

$fan\text{-}out = 2$

$$\frac{-\sqrt{6}}{\sqrt{4+2}}, \frac{+\sqrt{6}}{\sqrt{4+2}} = -1 \text{ to } +1$$

idea 3 :- He-init - 2015-2016 ⇒ Relu / Leaky

Activation - 2017 - Selu

uniform :-

(a) $w_{ij}^{k} \sim U\left[\frac{-\sqrt{6}}{\sqrt{fan\_in}}, \frac{+\sqrt{6}}{\sqrt{fan\_in}}\right]$

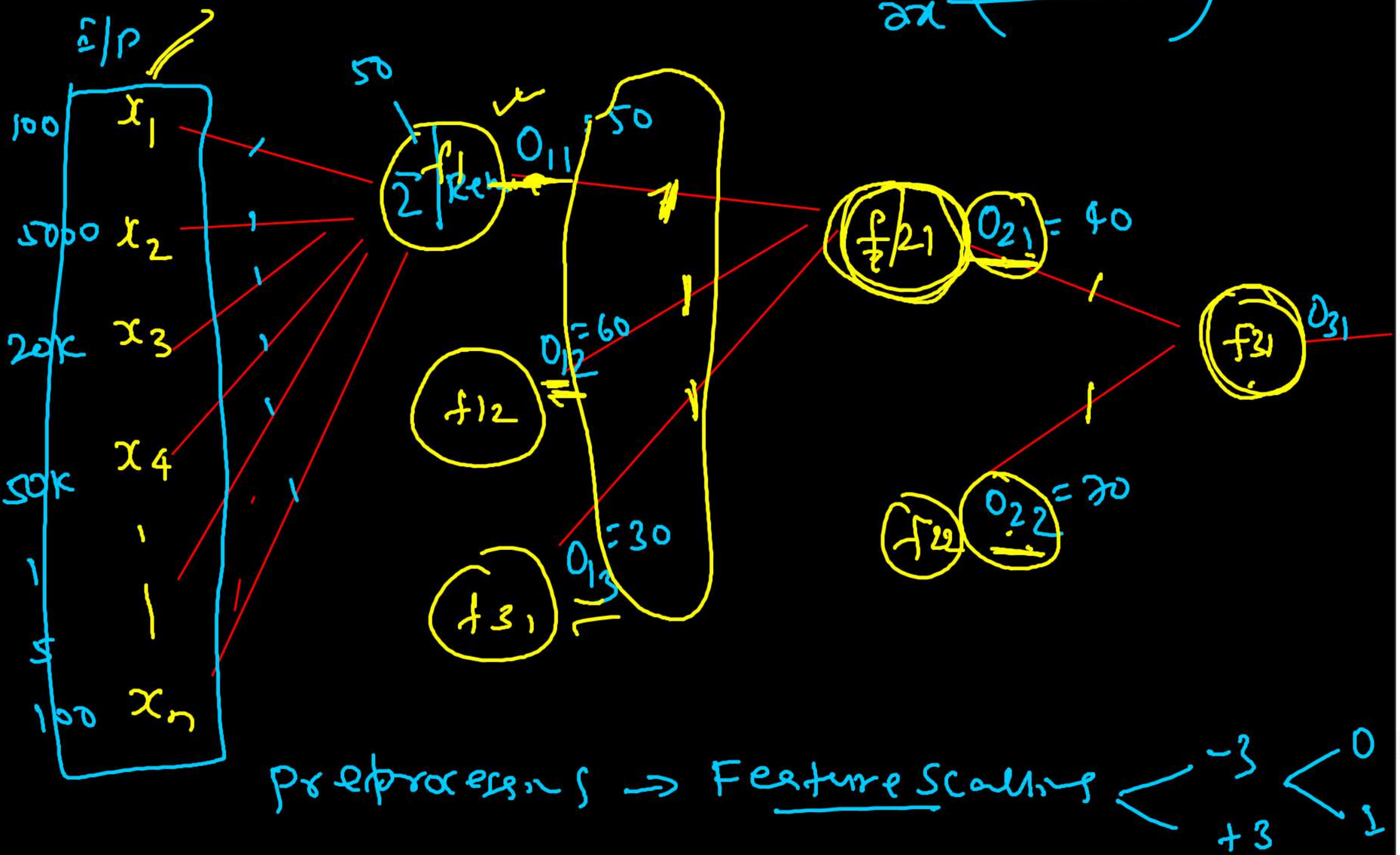
(b) Normal/Gaussian :: $w_{ij}^{k} \sim N(0, \sigma_{ij})$

where

$$\sigma = \sqrt{\frac{2}{fan\_in}}$$

Batch - Normalization

$$\frac{\partial}{\partial x}\left(x^{50} \quad x^{\frac{1}{2}}\right)$$



Preprocessing → Feature Scaling $\begin{matrix} -3 \\ +3 \end{matrix}$ $\begin{matrix} 0 \\ 1 \end{matrix}$

Webinar Chat

ok

Sourav K to Hosts and panelists
SK: Yes sir ,,

Usha Kumari to Hosts and panelists
UK: ye s

Priyanka to Hosts and panelists
P: yes

Santoshkumar Pandit to Hosts and panelists
SP: yes sir

PRAMOD K. to Hosts and panelists
PK: yes

Santoshkumar Pandit to Hosts and panelists
SP: Sir during batch normalisation does weight initialisation concepts work during their weight also ?

ok.....

yes

clear

Who can see your messages? Recording on

To: Hosts and panelists

Type message here...

Advantage - Batch Normalization

1. Faster convergence
2. handle Overfitting Problem

Please Note - non-trainable Parameter

↓

Feature Scaling on output