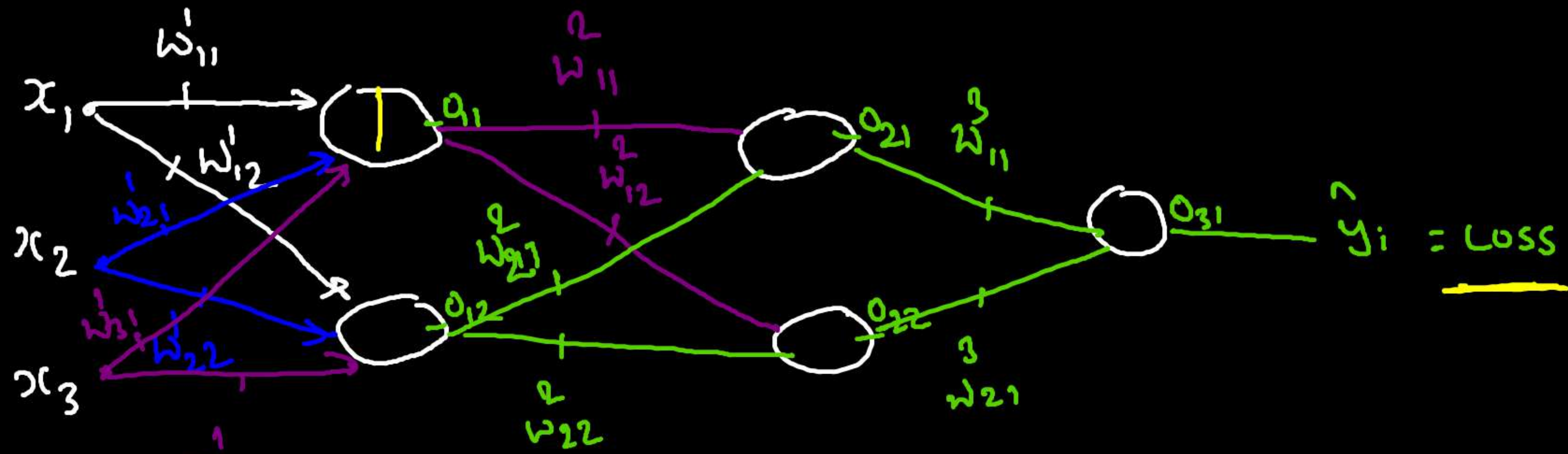


Weight Initialization - Deep MLP



Weight :-

$$I = \begin{bmatrix} w_{11}^1 \\ w_{12}^1 \\ w_{21}^1 \\ w_{22}^1 \\ w_{31}^1 \\ w_{32}^1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.1 \\ 0.3 \\ 0.5 \\ 0 \end{bmatrix} \quad \therefore \quad \begin{bmatrix} w_{11}^2 \\ w_{12}^2 \\ w_{21}^2 \\ w_{22}^2 \end{bmatrix} = \begin{bmatrix} w_{11}^3 \\ w_{21}^3 \end{bmatrix}$$

6 4 2

$= 12$

Webinar Chat

Joel Ratnam to Hosts and panelists

JR yes sir

Ravinder to Hosts and panelists

R Clear Sir.

rashmi to Hosts and panelists

R very much clear sir

Abhishek to Hosts and panelists

A please revise on vanishing and exploding gradient again

Vikram Rawlo to Hosts and panelists

VR yes

Santoshkumar Pandit to Hosts and panelists

SP yes

Sourav K to Hosts and panelists

SK yes sir

rashmi to Hosts and panelists

R yes

PRAMOD K. to Hosts and panelists

PK yes

Who can see your messages? Recording

To: Hosts and panelists

Type message here...

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

Talking: Learnvista Private Lim...

zm Webinar Chat

Usha Kumari to Hosts and panelists

UK 0

rashmi to Hosts and panelists

R bad

Priyanka to Hosts and panelists

P yes

PRAMOD K. to Hosts and panelists

PK yes

Usha Kumari to Hosts and panelists

UK yes sir

Sourav K to Hosts and panelists

SK Yes

Sahas Swamy to Hosts and panelists

SS yes sir

rashmi to Hosts and panelists

R clear sir

Priyanka to Hosts and panelists

P yes

Who can see your messages? Record

To: Hosts and panelists

Type message here...

② Non-zero constant initialization → linear

③ Random initialization with small weight

✓ convergence

$$50 * 0.00001 = \frac{\partial L}{\partial w_{31}} * \frac{\partial w_{31}}{\partial w_{21}} * \frac{\partial w_{21}}{\partial w_{01}} * \frac{\partial w_{01}}{\partial w'_{11}}$$

→ Vanishing Gradient Problem

④ Random initialization with large weight

10 100 500 2525 — —

→ Exploding Gradient Problem

$$w_0 = 50, w_n = 50 - 10 * 100 = -950$$

Talking: Learnvista Private Lim...

zm Webinar Chat

Sourav K to Hosts and panelists

SK Yes sir ..

Priyanka to Hosts and panelists

P yes

PRAMOD K. to Hosts and panelists

PK yes

Santoshkumar Pa... to Hosts and panelists

SP yes sir

Nagarajan K to Hosts and panelists

NK yes sir

mansoor ali Niza... to Hosts and panelists

MA Clear sir

Joel Ratnam to Hosts and panelists

JR yes sir

Ravinder to Hosts and panelists

R Clear Sir.

rashmi to Hosts and panelists

R very much clear sir

Who can see your messages? Record

To: Hosts and panelists

Type message here...

What can be done

Heuristic Approach

Weight init

uniform



Normal / Gaussian distribution



0.1 0.3
0.01
0.000001X

idea 1 :- Gaussian / Normal init

- (a) weight should be small (but not very small)
- (b) All weight should not be zero

Talking: Learnvista Private Lim...

zm Webinar Chat

R yes

PRAMOD K. to Hosts and panelists

PK yes

Usha Kumari to Hosts and panelists

UK 2

PRAMOD K. to Hosts and panelists

PK both side fan in?

ok

Usha Kumari to Hosts and panelists

UK yes sir

PRAMOD K. to Hosts and panelists

PK yes

rashmi to Hosts and panelists

R yes

Ravinder to Hosts and panelists

R yes

Santoshkumar Pandit to Hosts and panelists

SP yes sir

Who can see your messages? Recording

To: Hosts and panelists

Type message here...

③ Good-variance

④ Weight should not be the same

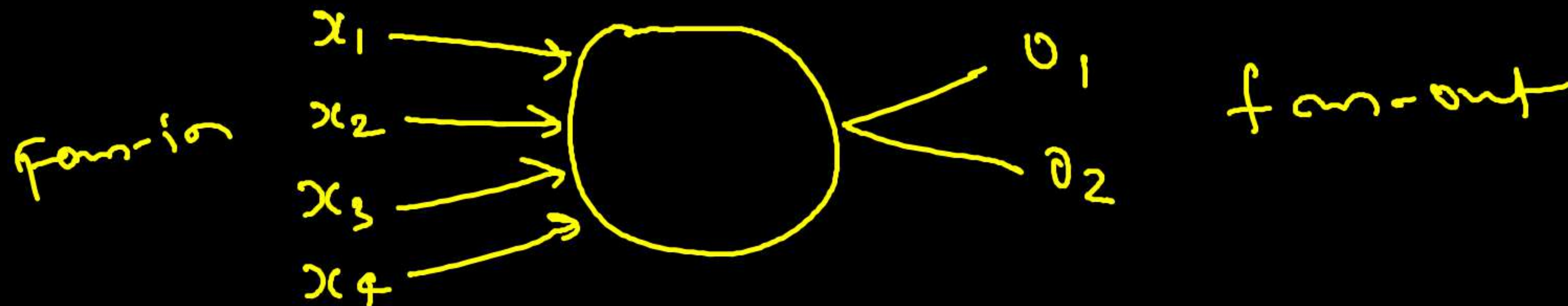
→ better init strategies (Heuristic Approach)

→ lots of experiment

→ Theory/math part is weak

fan-in

fan-out



Talking: Learnvista Private Lim...

zm Webinar Chat

rashmi to Hosts and panelists

R yes

Ravinder to Hosts and panelists

R yes

Santoshkumar Pandit to Hosts and panelists

SP yes sir

Vikram Rawlo to Hosts and panelists

VR yes

Usha Kumari to Hosts and panelists

UK ye s sir

Sourav K to Hosts and panelists

SK Now yes...

Ravinder to Hosts and panelists

R yes

PRAMOD K. to Hosts and panelists

PK now it is

Abhishek to Hosts and panelists

A but in hidden layer we use ReLU, why sigmoid?

Who can see your messages? Recording

To: Hosts and panelists

Type message here...

$W_{ij}^k \rightarrow \text{unif} \left[\frac{-1}{\sqrt{\text{fan-in}}}, \frac{1}{\sqrt{\text{fan-in}}} \right]$ Formula

$\checkmark \text{ fan-in} = 4$
 $\text{fan-out} = 2$

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

$W_{ij}^k = (-0.5, +0.5)$ — 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 | |

zm Webinar Chat

PRAMOD K. to Hosts and panelists

PK yes

rashmi to Hosts and panelists

R yes

Ravinder to Hosts and panelists

R yes

Santoshkumar Pandit to Hosts and panelists

SP yes sir

Vikram Rawlo to Hosts and panelists

VR yes

Usha Kumari to Hosts and panelists

UK ye s sir

Sourav K to Hosts and panelists

SK Now yes...

Ravinder to Hosts and panelists

R yes

PRAMOD K. to Hosts and panelists

PK now it is

Who can see your messages? Recording

To: Hosts and panelists

Type message here...

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

Temp Sigmoid very popular

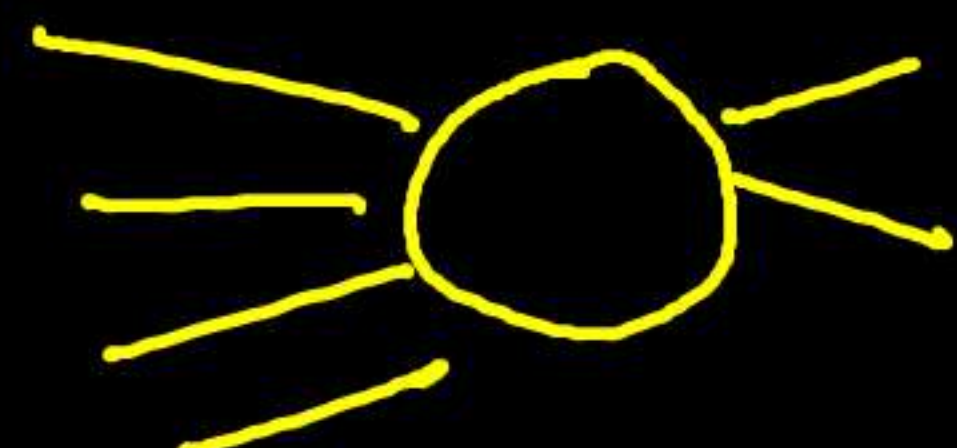
idea 2 :- 2010 - Xavier/Glorot init

(a) Xavier/Glorot Normal init

$$\sqrt{\frac{2}{6}} = \sqrt{\frac{1}{3}} = \frac{1}{1.73} \approx 0.577$$

fan-in

fan-out



4

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

Webinar Chat

when we have to apply sigmoid

in which case sigmoid would be applied?

Santoshkumar Pandit to Hosts and panelists

SP yes

rashmi to Hosts and panelists

R yes

Ravinder to Hosts and panelists

R yes

Usha Kumari to Hosts and panelists

UK relu

yes sir

Santoshkumar Pandit to Hosts and panelists

SP Yes little remember of Dalton theory

Usha Kumari to Hosts and panelists

UK centrifugal force acts here

Sourav K to Hosts and panelists

SK Yes

Who can see your messages? Record

To: Hosts and panelists

Type message here...

⑤ uniform Xavier/Glorot init

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

$$f_{in-in} = 4$$

$$f_{in-out} = 2$$

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

Webinar Chat

yes

Usha Kumari to Hosts and panelists

UK ye s sir

Sourav K to Hosts and panelists

SK Now yes...

Ravinder to Hosts and panelists

R yes

PRAMOD K. to Hosts and panelists

PK now it is

Abhishek to Hosts and panelists

A but in hidden layer we use ReL why sigmoid?

Vikram Rawlo to Hosts and panelists

VR yes

Abhishek to Hosts and panelists

A so is this relevant now?

when we have to apply sigmoid

in which case sigmoidwud b applied?

Who can see your messages? Record

To: Hosts and panelists

Type message here...

idea 3 :- He-inst - 2015-2016 \rightarrow ReLU / Leaky

Activation - 2017 - Selu

④ uniform :-

$$w_{ij}^k \sim u \left[\frac{-\sqrt{6}}{\sqrt{fan-in}}, \frac{+\sqrt{6}}{\sqrt{fan-in}} \right]$$

⑤ Normal/Gaussian :- $w_{ij}^k \sim N(0, \sigma_{ij})$

\propto

$$\sigma = \sqrt{\frac{2}{fan-in}}$$

Webinar Chat

A ok sir

rashmi to Hosts and panelists

R ok

Nagarajan K to Hosts and panelists

NK when to use uniform or normal ? which distribution are we referring to

rashmi to Hosts and panelists

R :D

yes

uniform

Nagarajan K to Hosts and panelists

NK ok sir

Usha Kumari to Hosts and panelists

UK yes sir

Santoshkumar Pandit to Hosts and panelists

SP yes

Usha Kumari to Hosts and panelists

UK interesting

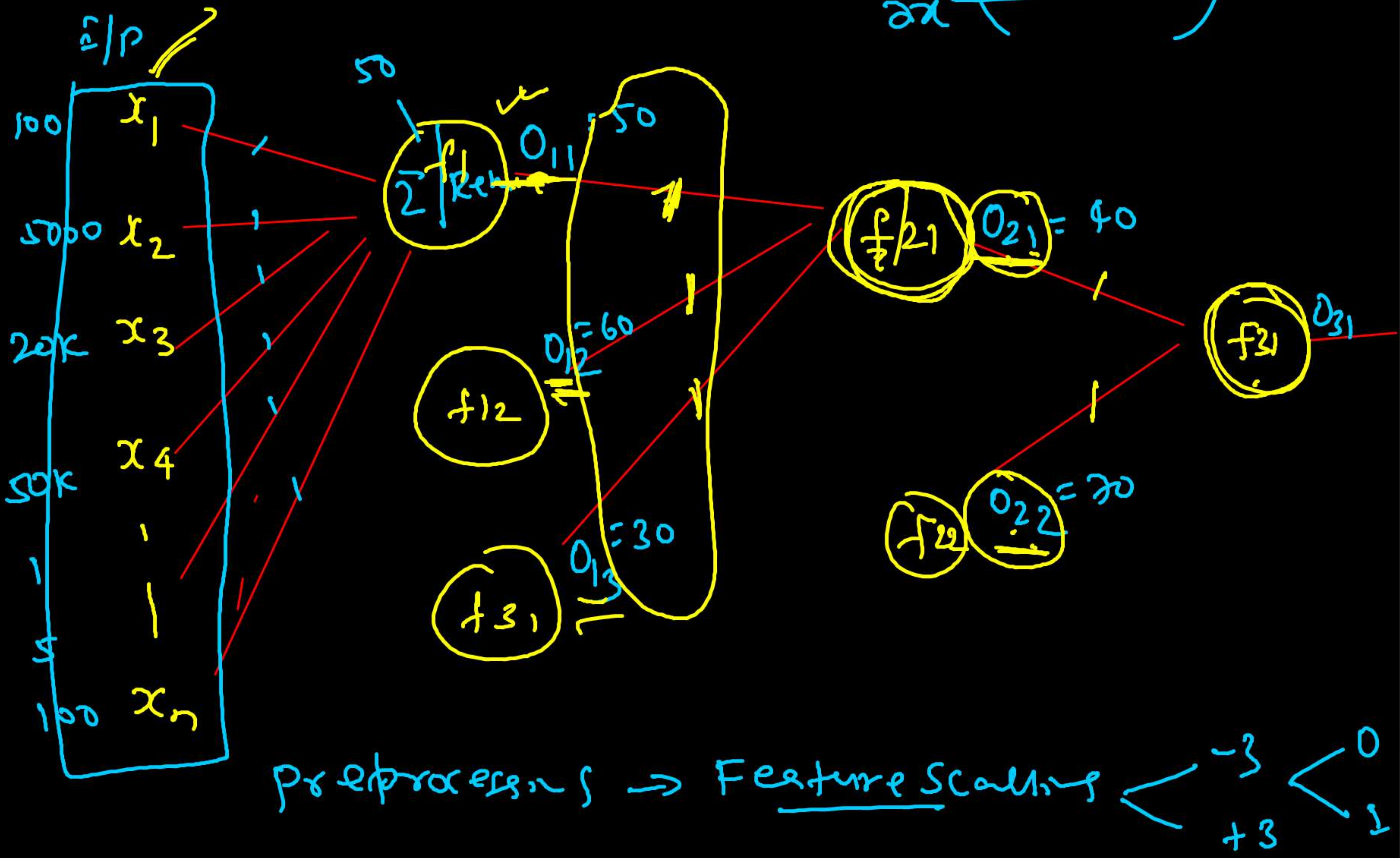
Who can see your messages? Recording on

To: Hosts and panelists

Type message here...

Batch - Normalization

$$\frac{d}{dx} \left(x^{50} x^2 \right)$$



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

rashmi to Hosts and panelists

R x1 x2

Priyanka to Hosts and panelists

P output values

rashmi to Hosts and panelists

R u asked sir f1

Sourav K to Hosts and panelists

SK Yes

Nagarajan K to Hosts and panelists

NK yes

Sourav K to Hosts and panelists

SK yes sir

Priyanka to Hosts and panelists

P yes

PRAMOD K. to Hosts and panelists

PK Bigger value means out of range values. correct?

ok

Who can see your messages? Recording on

To: Hosts and panelists

Type message here...