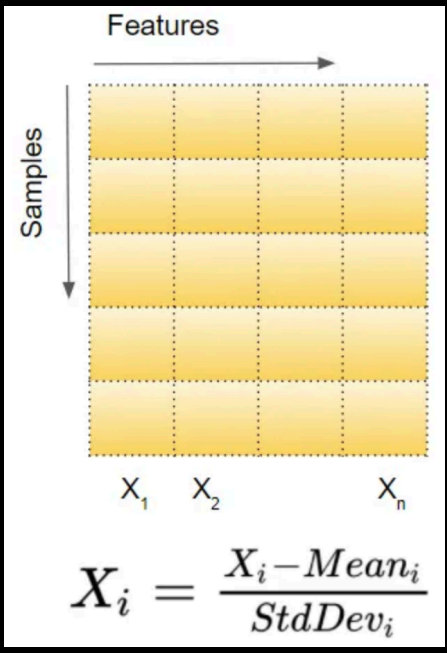
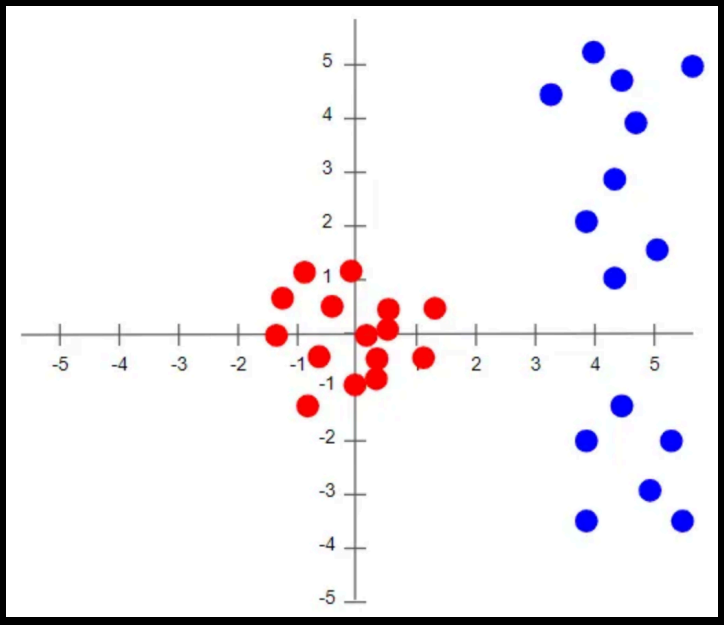
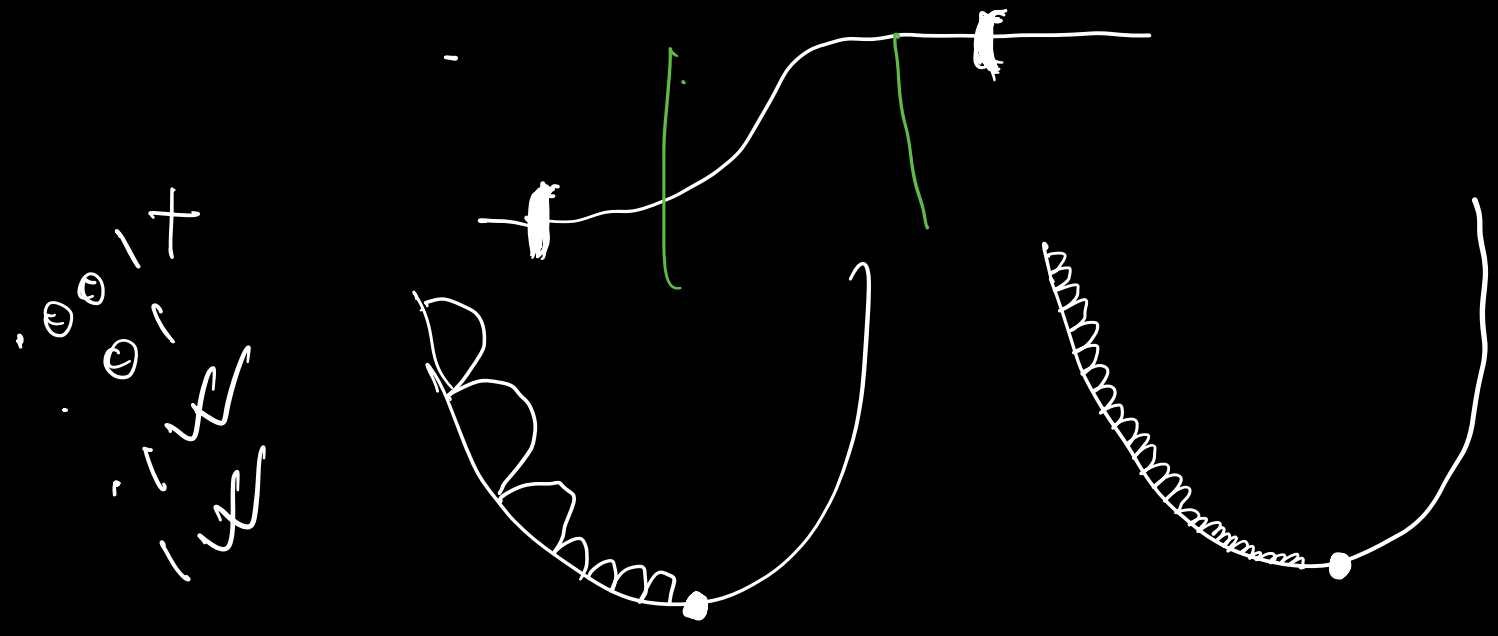
Today's Agenda

1) Batch Normalization

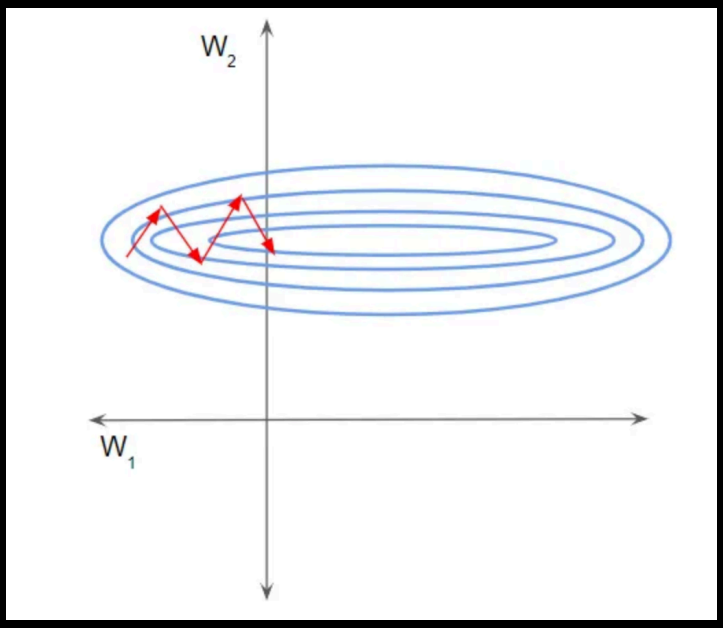
2) Weight Initialization



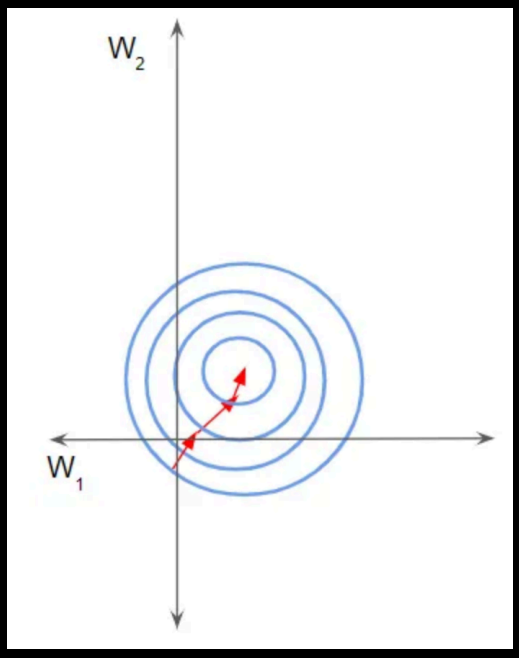


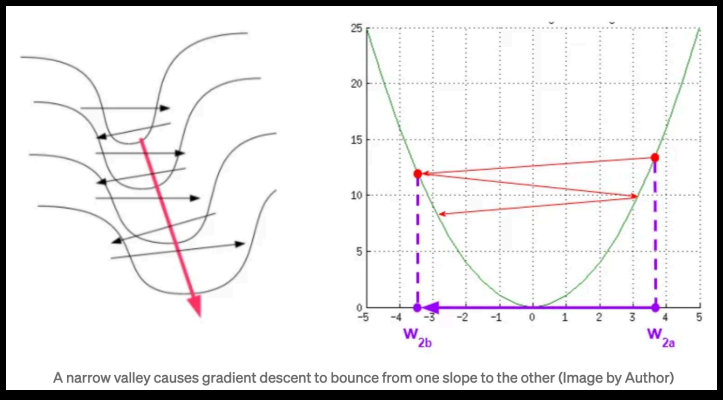


Without BN C.F Curve



B.N C.F Curve





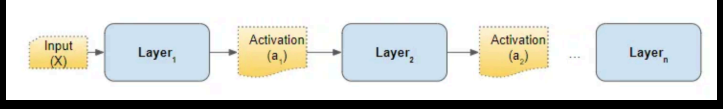
A narrow valley causes gradient descent to bounce from one slope to the other (Image by Author)

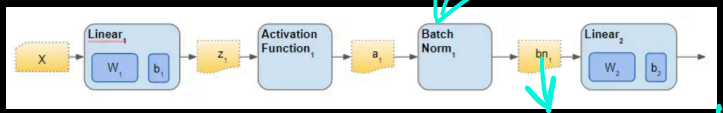
w/o BN N1 More Epochs less Time

N2 w BN less Epochs More Time

100 98%.

40 Epochs





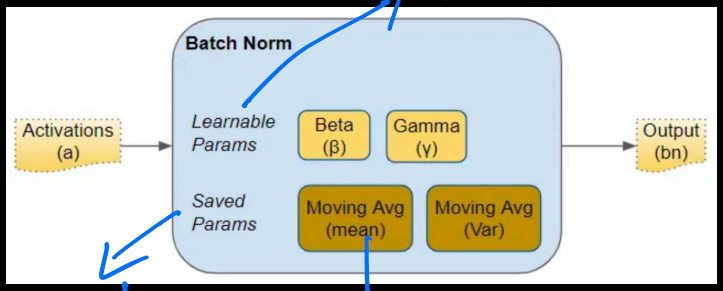
Normalized Output

act(wx+b) = a''

Normal (a'') -> BN(a'')

Trainable

=> act(wx+b)



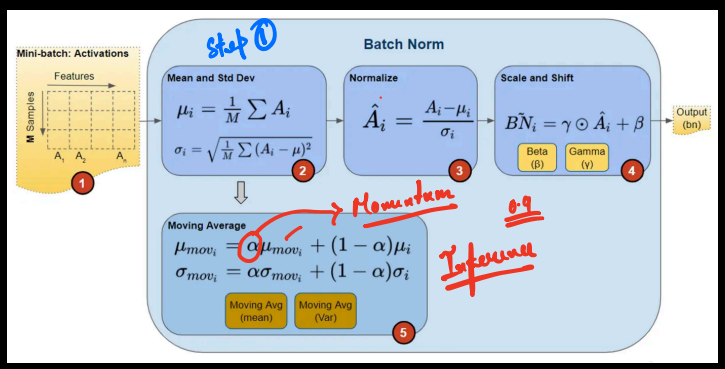
Non Learnable Parameters

Inference

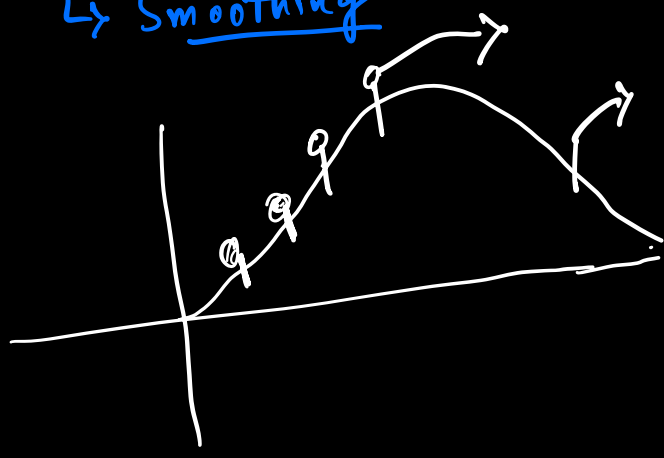
β & γ

Exponential Weighted Moving Avg (EWMA)

In a layer for every node activation is calculated and then BN is applied

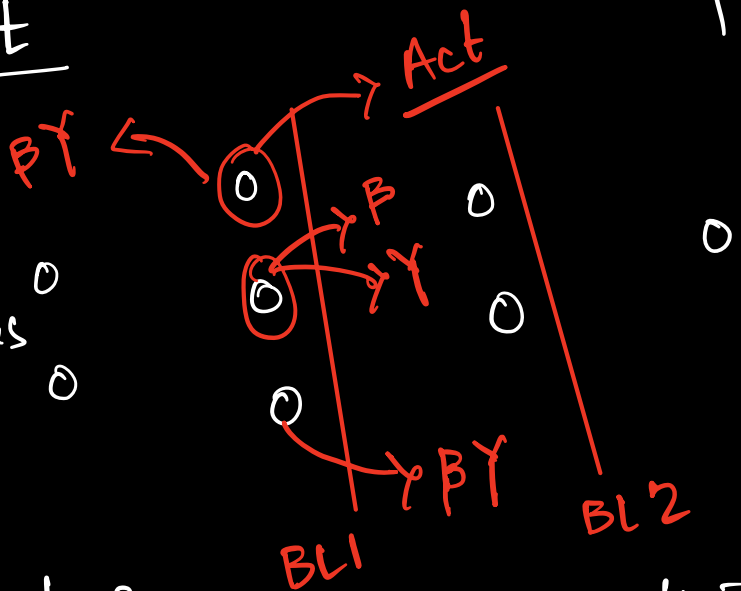


→ Smoothing



γ → Scaling factor

β → shift



Total Parameters

20

Learnable Parameters

↳ 10

5 × 4 = 20

Activations → Mean → Std-Dev → Normalization

↓

B·N (Scale & Shift)

$\{z(wx+b)\}$

Case 1

Activation → Batch Normalization

Case 2

$Wx+b$) → Batch Normalization → Activation

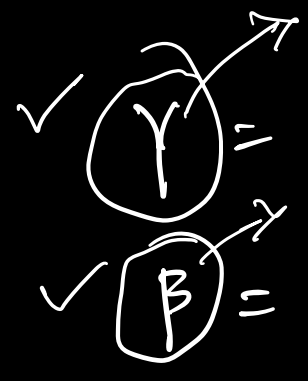
MBGD

Feed Forward

$Wx+b$

Act $(Wx+b)$

Normalization



$\frac{\partial L}{\partial \gamma}$

$\frac{\partial L}{\partial \beta}$

Batch Normalization $\frac{100}{4} = 25$

Back prop

Weights, Bias, β, γ

Moving Average Mean / Variance

→ Saved during Training

100 data points Batch size → 4

Iterations → 25

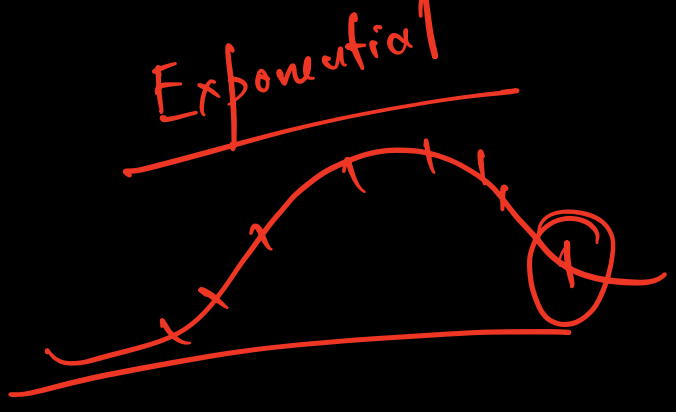
x_1 Moving Average & Mean

x_1 Moving Avg Std. Dev

⋮

x_{25} MAM

x_{25} MAV



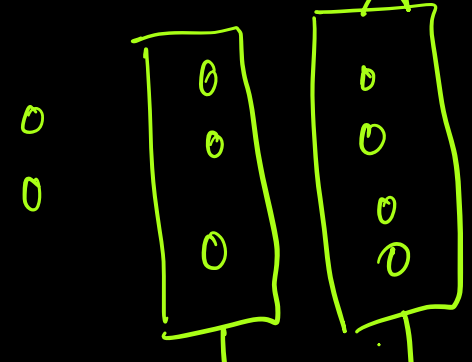
Pros :

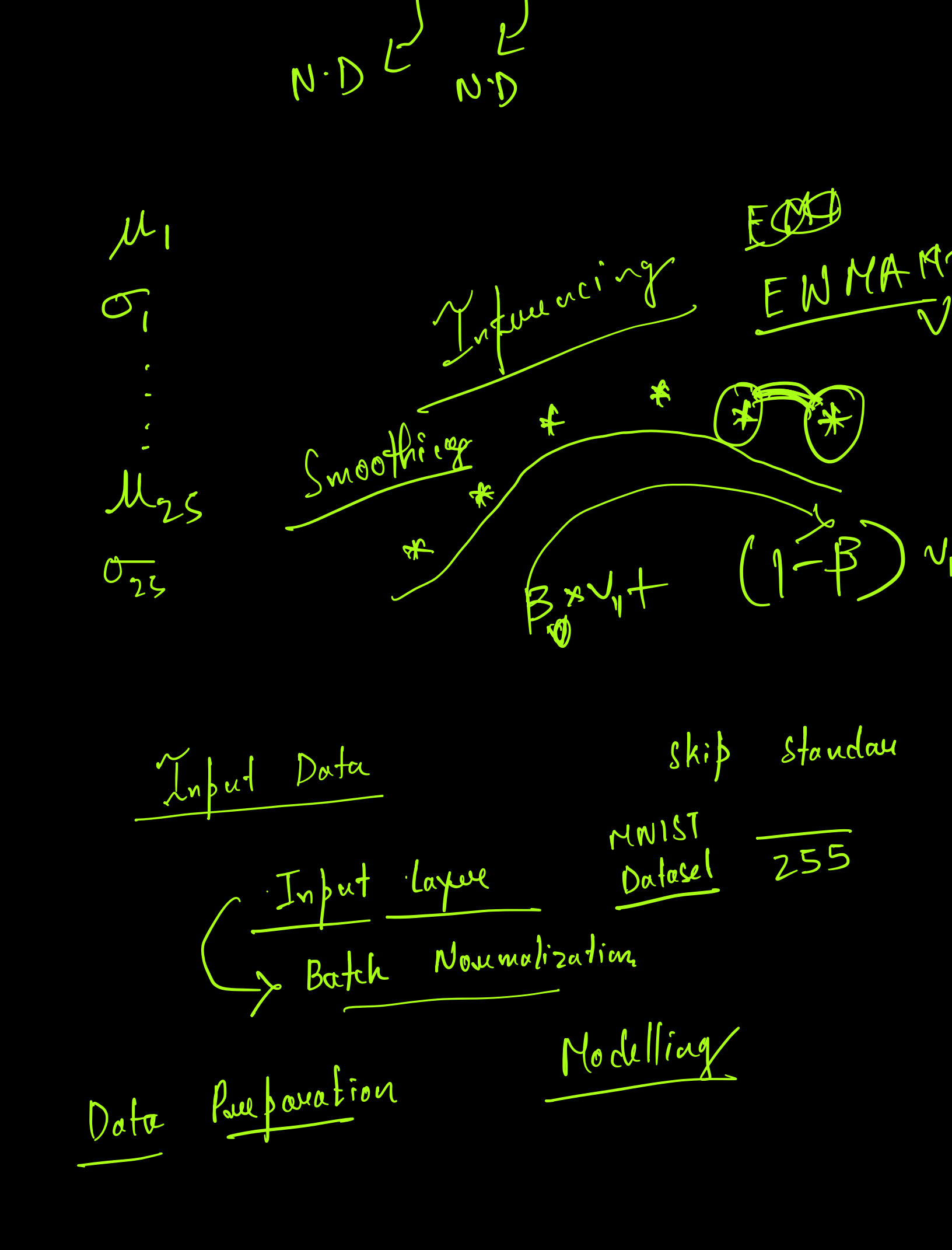
1) Faster Training Time

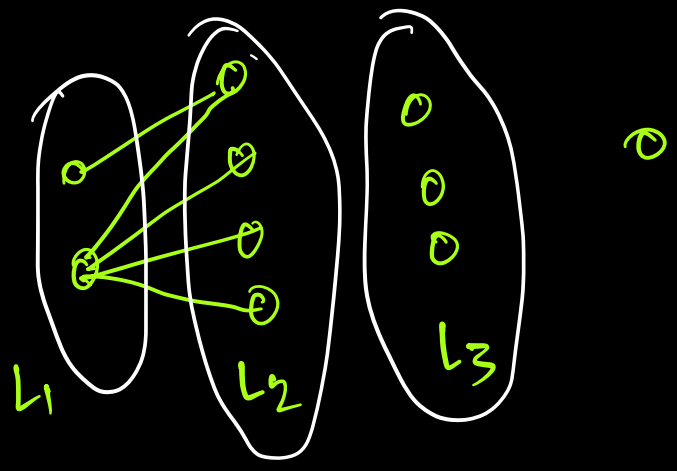
2) Regularizer

3) Not to worry about Weight intialization

4) Stable Training




N·D
N·D
μ_1
σ_1
μ_{25}
σ_{25}
Influencing
Smoothing
EWMA
$\beta \times v_{t-1}$
$(1-\beta)$
Input Data
skip
standar
Input layer
MNIST Dataset
255
Batch Normalization
Data Preparation
Modelling



Different Distributions

↳ keeps on changing

↳ Training slows / bit unstable

Weight Initialization

1) Zero Initialization ✗
2) Non - Zero Initialization ✗
3) Random Initialization

$$Wx + b$$

$$\Rightarrow 0 \cdot x + 0$$

$$\Rightarrow 0$$

Intution

$z_{11} = wx + b$ → $a_{11} = \max(0, z_{11})$

$w_{11}^{1} = w_{11} - \eta \frac{\partial L}{\partial w_{11}}$

$z_{12} = wx + b$ → $a_{12} = \max(0, z_{12})$

$a_{11} = a_{12}$ (No Training)

equal to 0

Random Initialization

(1) Small values → Vanishing Gradient

(2) Large values → Exploding Gradient

According to Requirement

1) Mid values (neither small / large)
2) All weights should be different

3) Range Defined (Limit)

Variance $= \sqrt{\frac{1}{n}}$ Exponentially

previous layer nodes, Fan In

$= \frac{1}{3} = \frac{1}{5}$

Hueristics

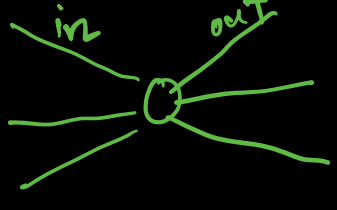
(tanh) Xavier Initialization (Normal Distri)

① Glorot

$$W_{ini} = \sqrt{\frac{1}{fan\_in}}$$

(Relu) ~~KQ~~ ② He Initialize (Normal Distribution)

$$W_{ini} = \sqrt{\frac{2}{fan\_in}}$$



(3) Xavier Initialization (Uniform Distribution)

[- limit, limit]

$$limit = \sqrt{\frac{6}{fan\_in + fan\_out}}$$

(4) He Initialization (Uniform Distribution)

[-limit, limit]

$$limit = \sqrt{\frac{6}{fan\_in}}$$