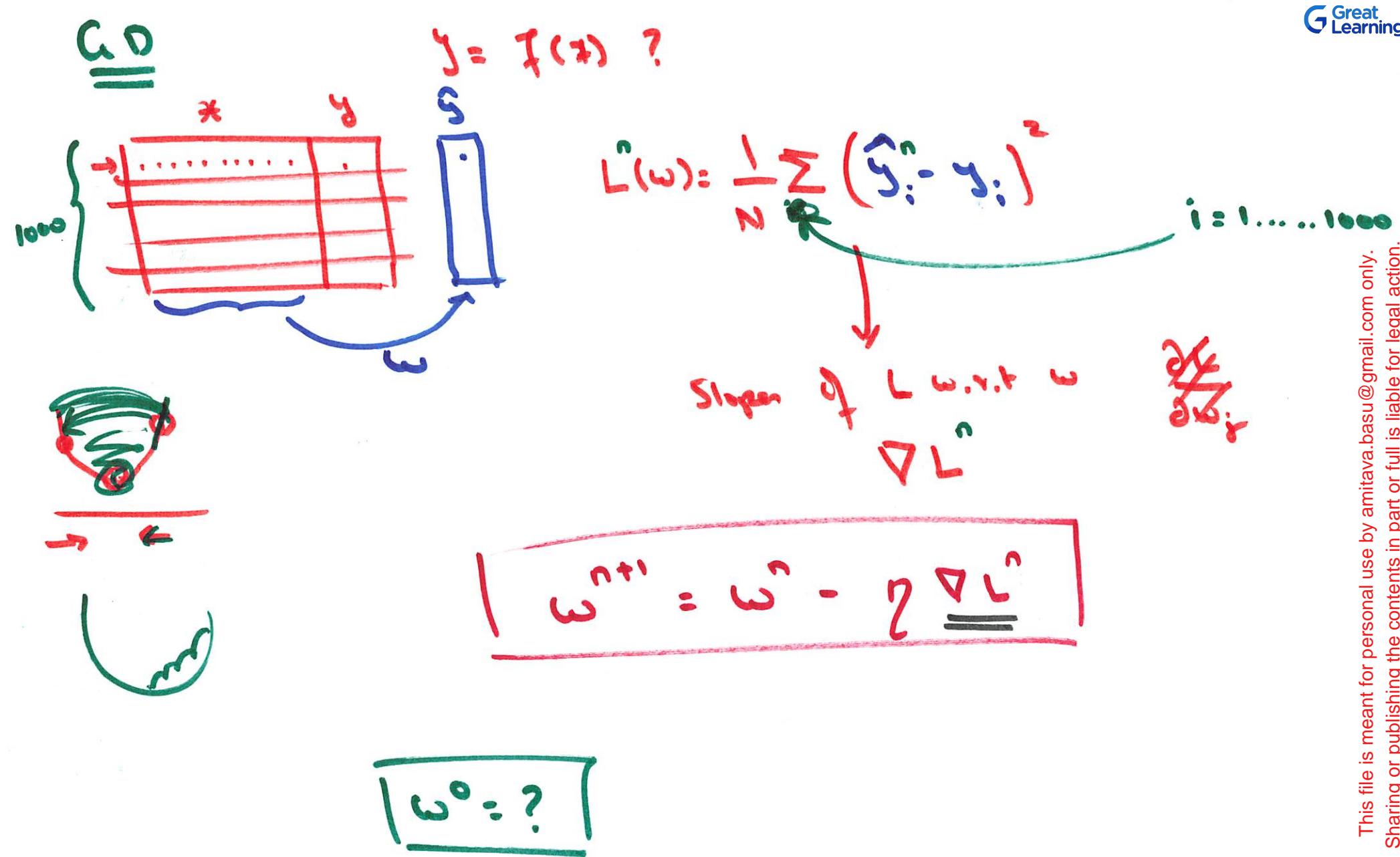


Agenda

- The Challenges: Over fitting & local optima
- The Training
 - Epochs, Batch Size, Iterations
 - Gradient Descent (GD) Vs Stochastic GD (SGD) Vs Mini-Batch GD
 - SGD with momentum
 - Learning rates and adaptive learning rates
 - Weight Initialization
 - Batch Normalization
- Guarding against over-fitting
 - L1/L2 Regularization
 - Data Augmentation
 - Drop outs
- Neural Network Architectures

weigh decay



This file is meant for personal use by amitava.basu@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

SGD

$$(GD) \quad L(\omega) = \frac{1}{N} \sum_{i=1 \dots 1000} (\hat{y}_i - y_i)^2$$

$$(SGD) \quad L(\omega) = (\hat{y}_i - y_i)^2$$

randomly chosen

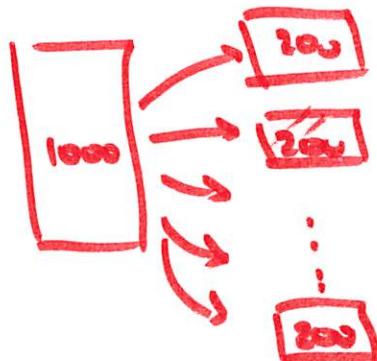
Mini Batch
SGD

$$N = 1000$$

$$N_b = 200$$

$$\text{iter} = 5$$

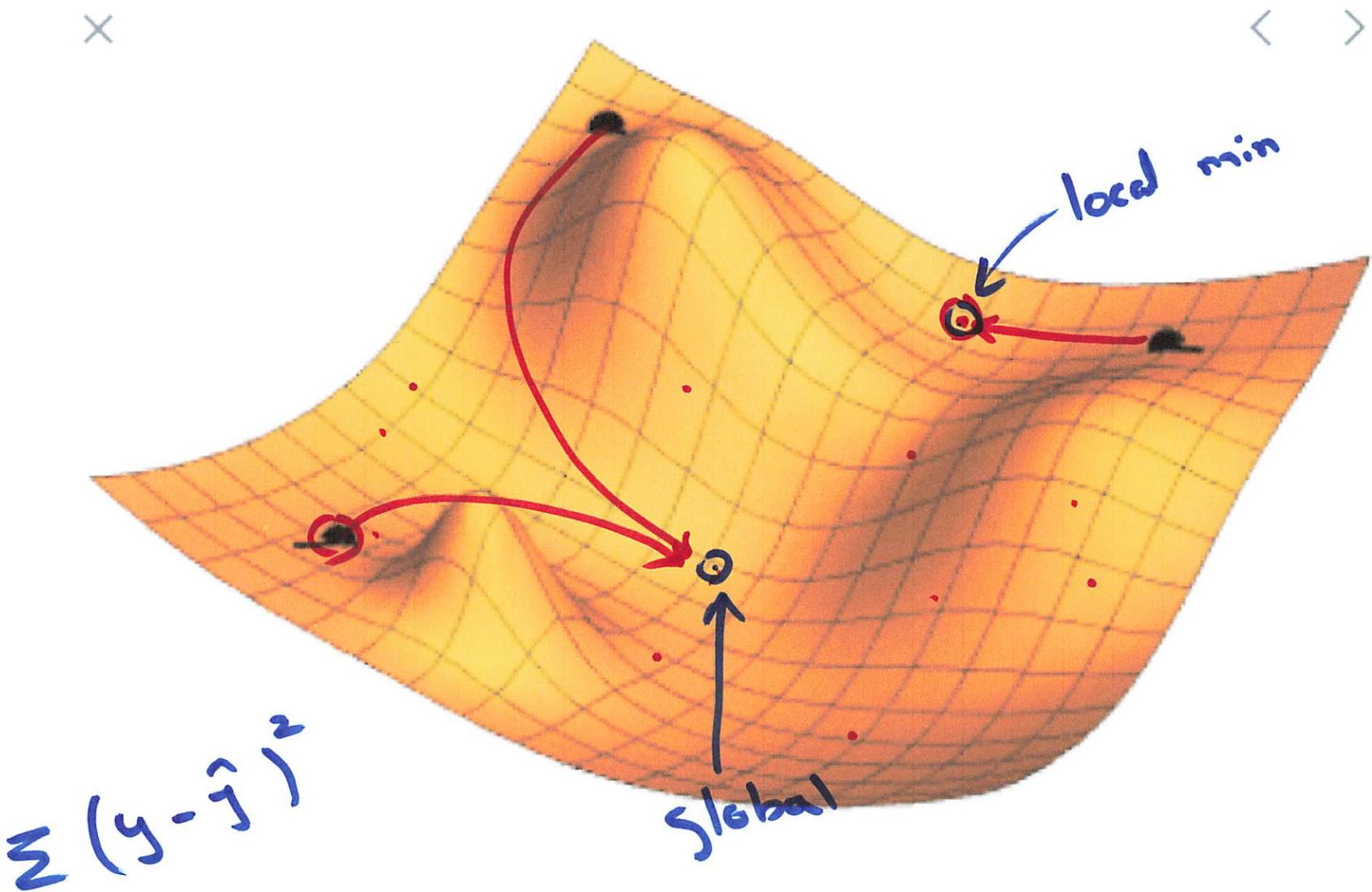
epoch



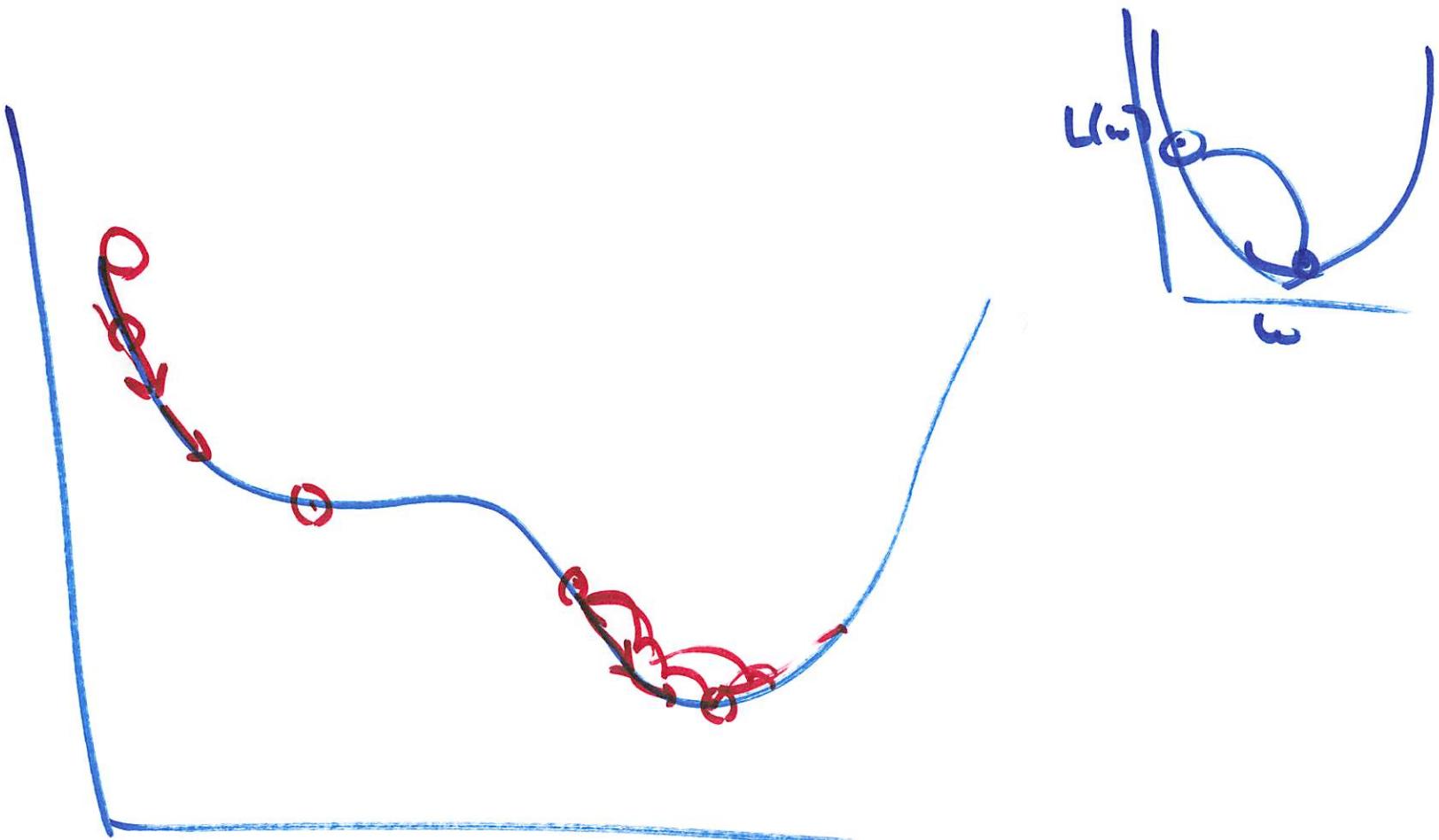
$$L(\omega) = \frac{1}{N_b} \sum_{i \in B} (\hat{y}_i - y_i)^2$$

batch size = N_b

$$\text{iterations} = \frac{N}{N_b}$$



This file is meant for personal use by amitava.basu@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.



$$\hat{J}'' = J' - \eta \nabla L'$$

SGD with momentum \Rightarrow

$$\hat{J}'' = J' - \eta (\alpha \nabla L' + (1-\alpha) \nabla L'')$$

Learning rate

- Choosing the Learning rate (η)
 - Too small, we will need too many iterations for convergence
 - Too large, we may skip the optimal solution
- Adaptive Learning Rate :
 - start with high learning rate and
 - gradually reduce the learning rate with each iteration.
 - Moreover, having different learning rates for different weight updates will help: Adagrad, RMS Prop

Adam

AdaDelta

Adaptive Grad

$$\hat{g}^t = g^t - \frac{\eta}{\sqrt{\delta^t + \epsilon}} \nabla L^t$$

$$\delta^t = \sum_i (\nabla L^t)^2$$

RMS Prop

$$\hat{g}^t = g^t - \frac{\eta}{\sqrt{\delta^t + \epsilon}} \nabla L^t$$

$$\delta^t = \alpha \delta^{t-1} + (1-\alpha) \nabla L^t$$

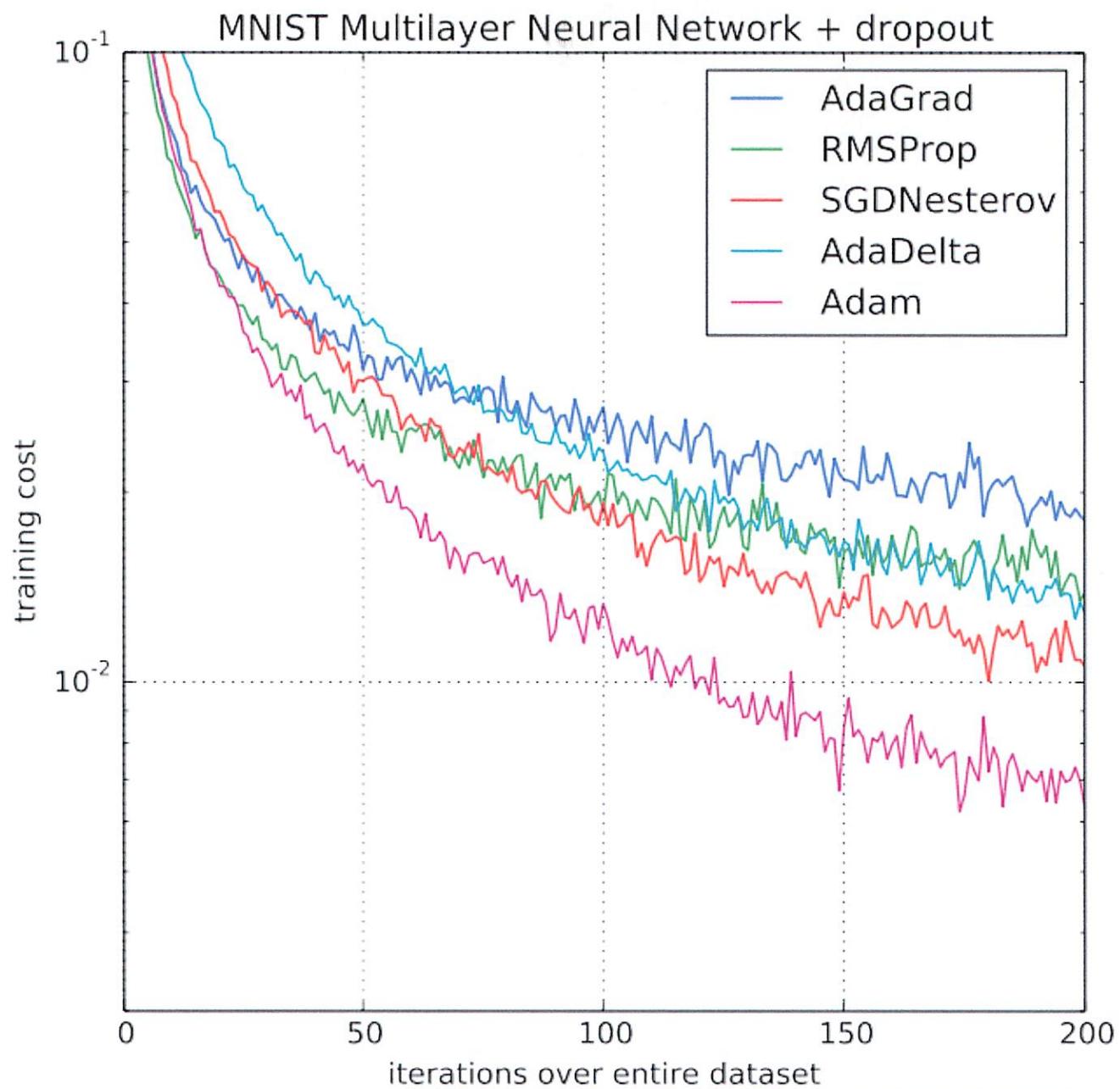
Adam

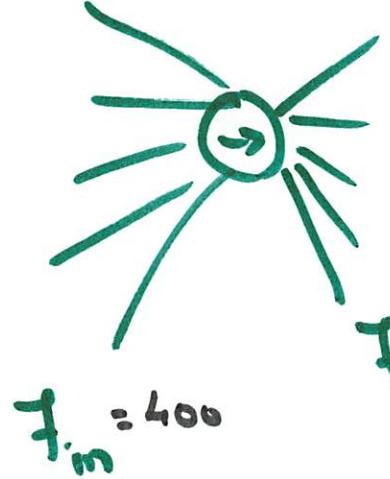


RMS Prop

+
SGD

with Momentum

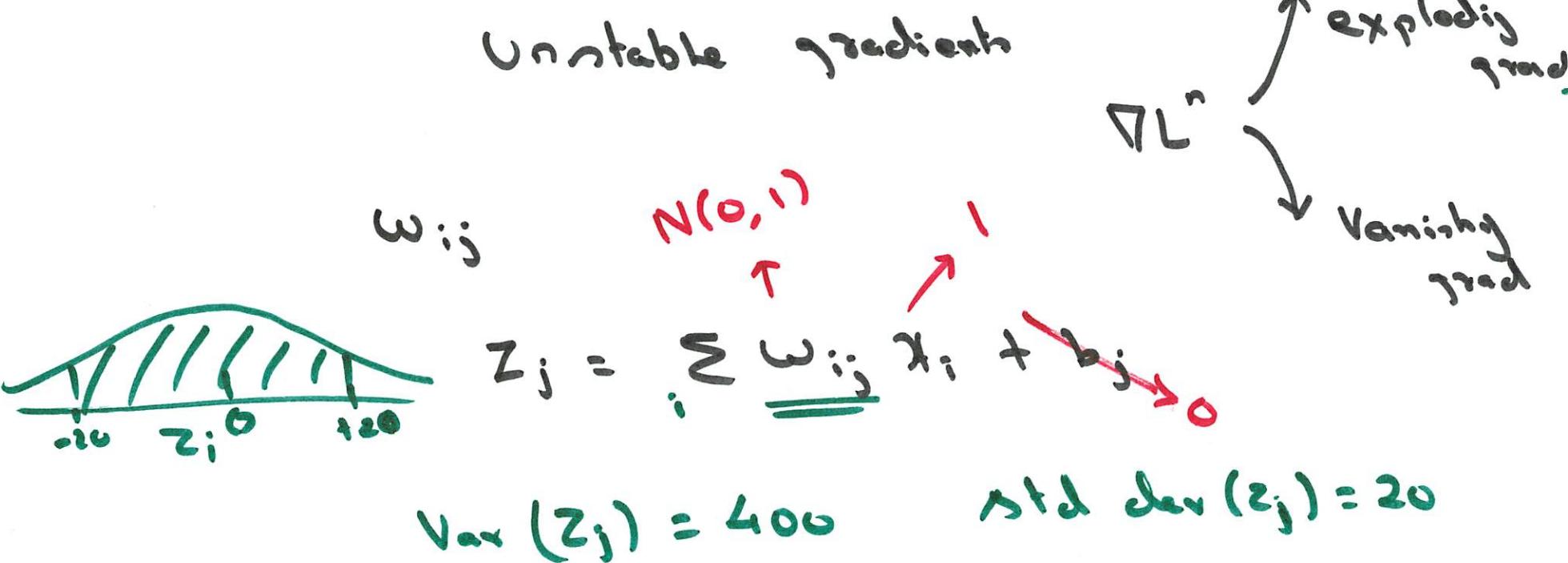


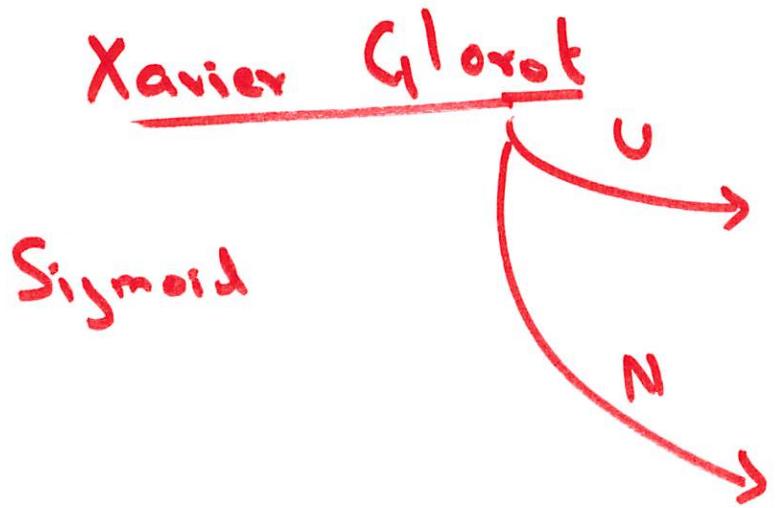


$$a_j = \sigma \left(\sum w_{ij} x_i + b_j \right)$$

$w_{ij} = 0$

$w_{ij} = \text{Normal}(0, 1)$



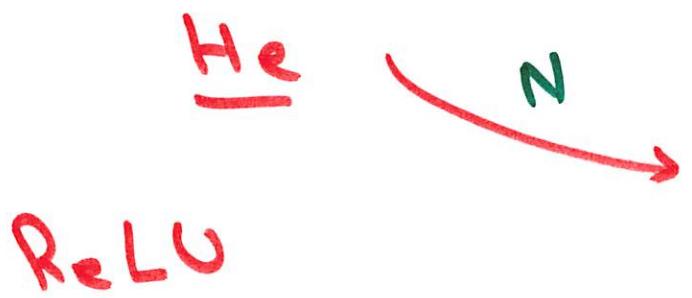


Uniform

$$\left[-\sqrt{\frac{6}{I_{in} + I_{out}}} , +\sqrt{\frac{6}{I_{in} + I_{out}}} \right]$$

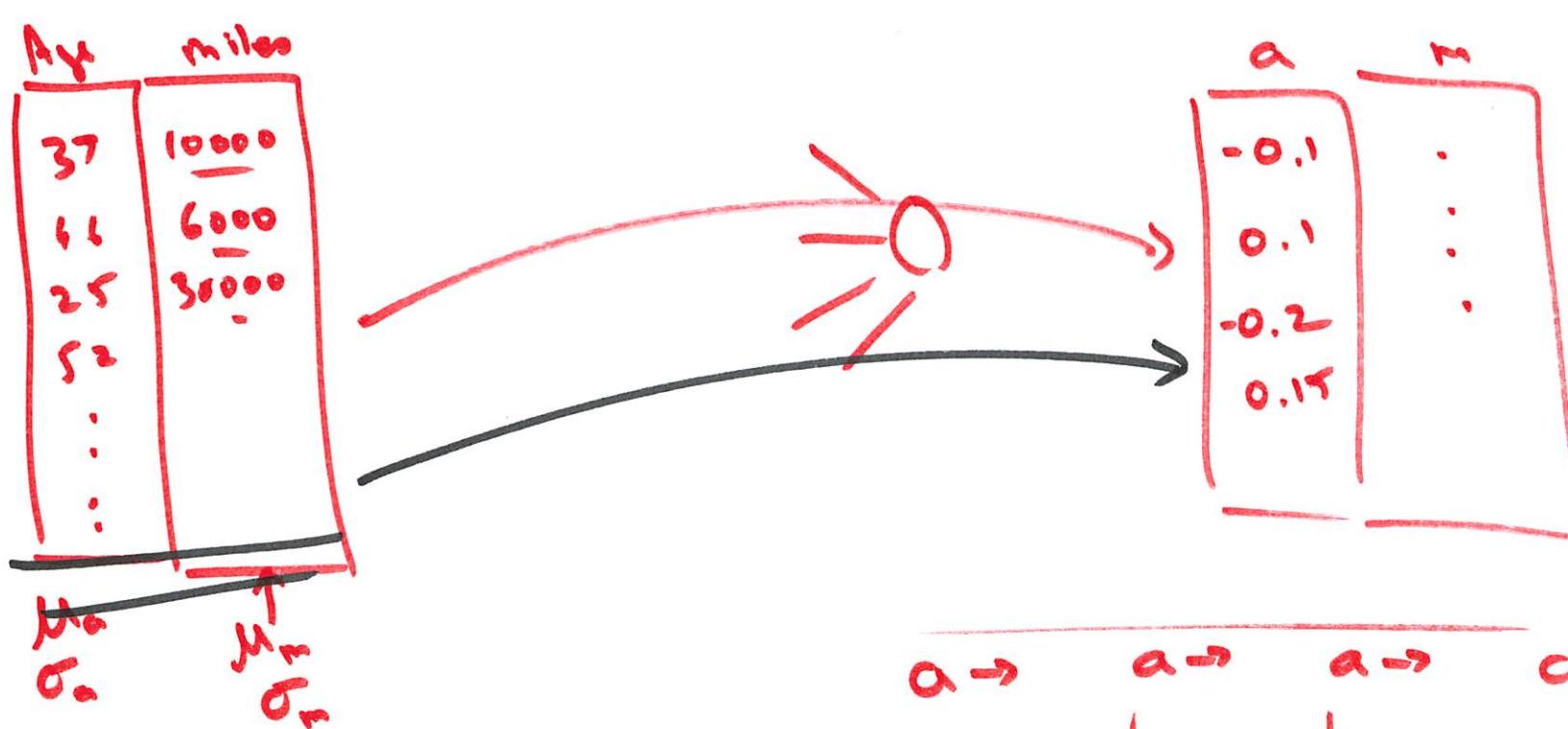
Normal

$$\left(0, \sqrt{\frac{2}{I_{in} + I_{out}}} \right)$$



Normal

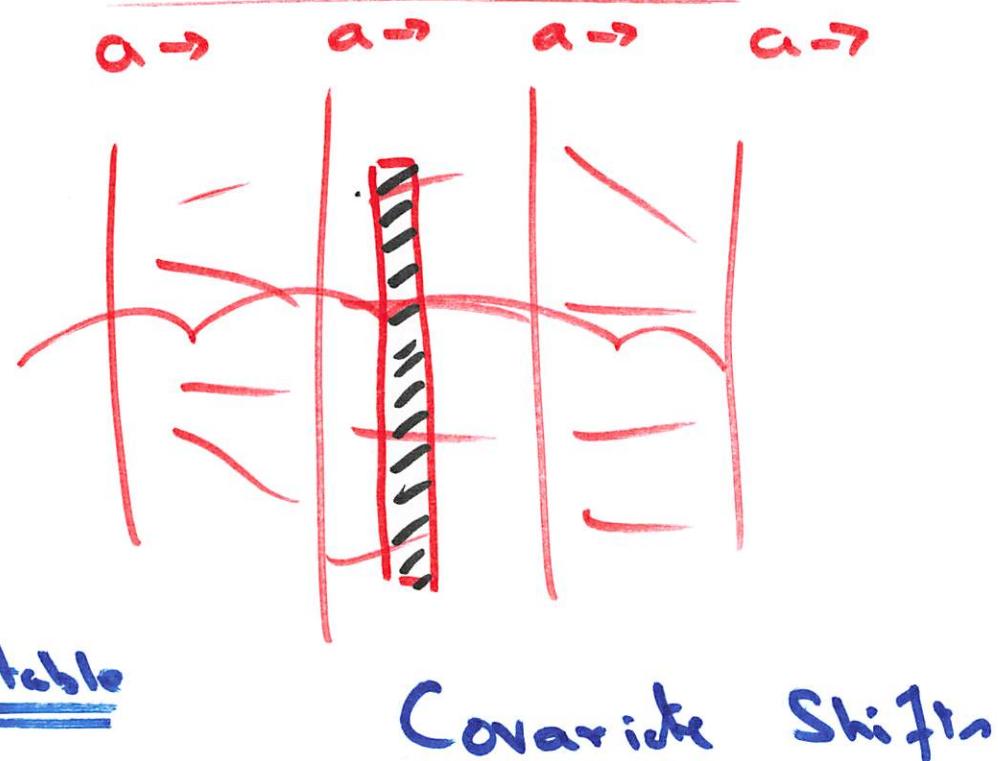
$$\left(0, \sqrt{\frac{2}{I_{in}}} \right)$$

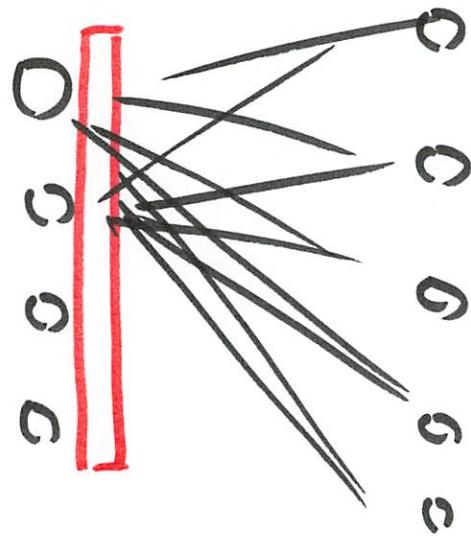


$$\hat{x}_a = \frac{x_a - Ma}{\sigma_a}$$

$$\hat{x}_m = \frac{x_m - Mm}{\sigma_m}$$

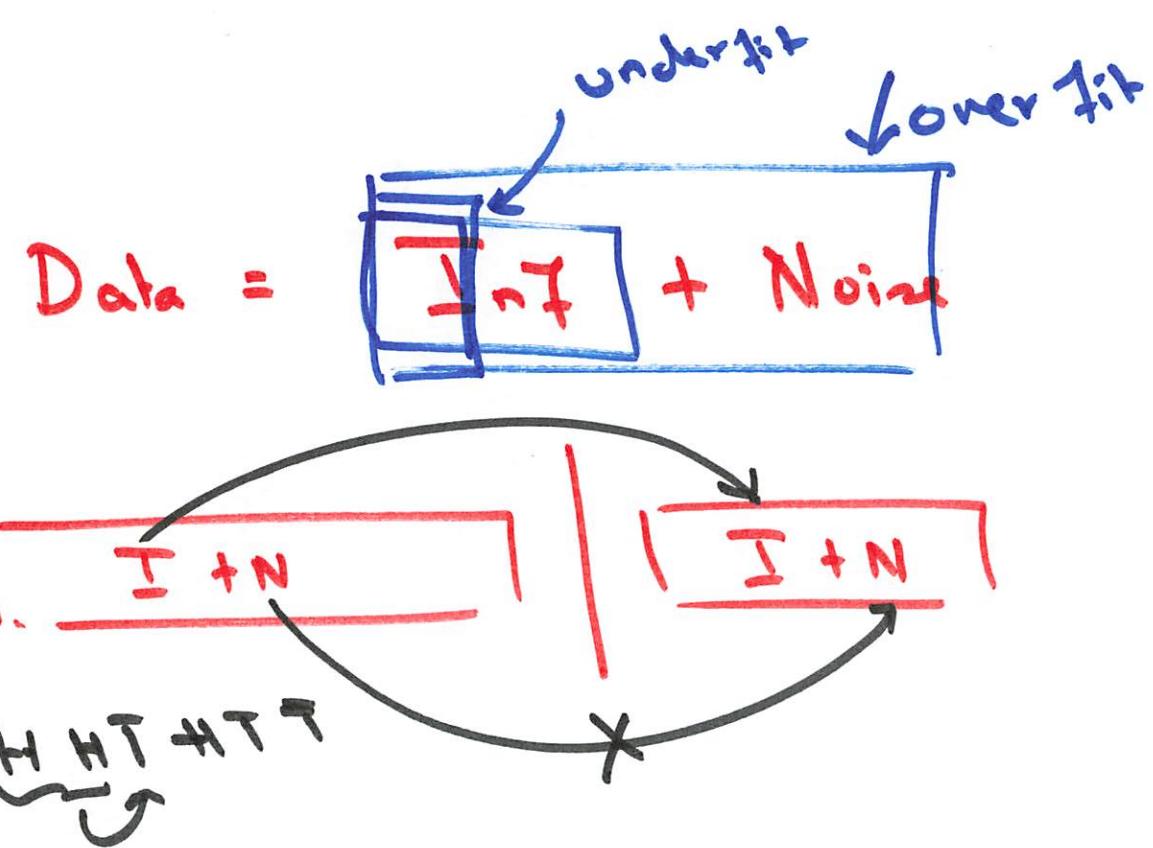
long → Forward → Stable





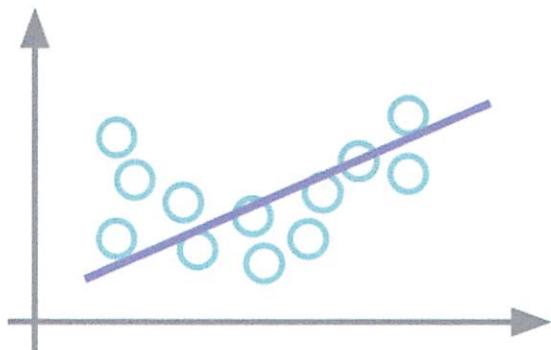
$$\hat{a}_i = \frac{a_i - \mu}{\sigma}$$
$$\hat{a}_i = \gamma \hat{a}_i + \beta$$

mem: μ_p, σ_p

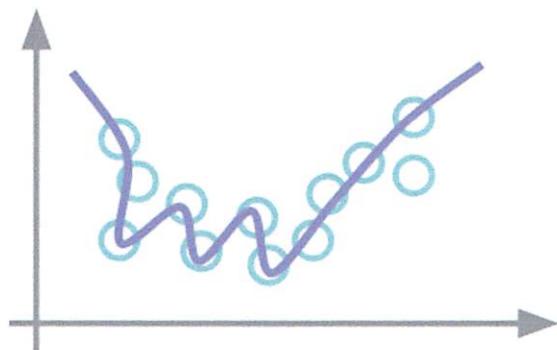
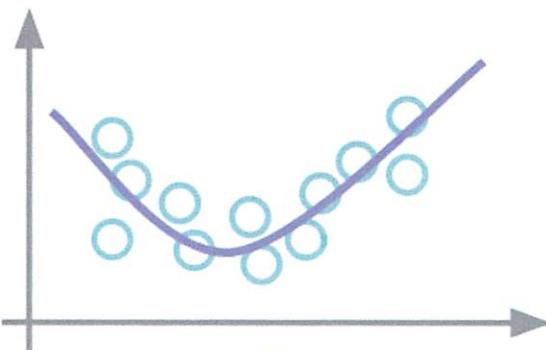


This file is meant for personal use by amitava.basu@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

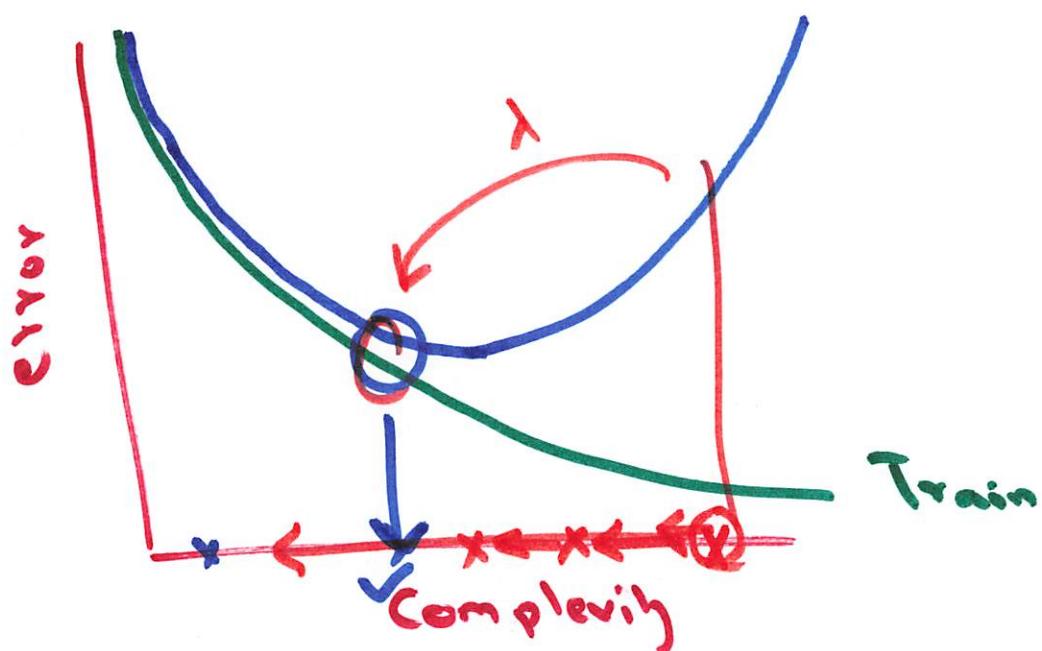
Under Vs Over-fitting



Underfit



overfit



$L_1 + L_2$ Reg.

∇

w_i :

$$L(w) = \frac{1}{N} \sum (y - g)^2 + \frac{\lambda}{p} (\text{penalty})$$

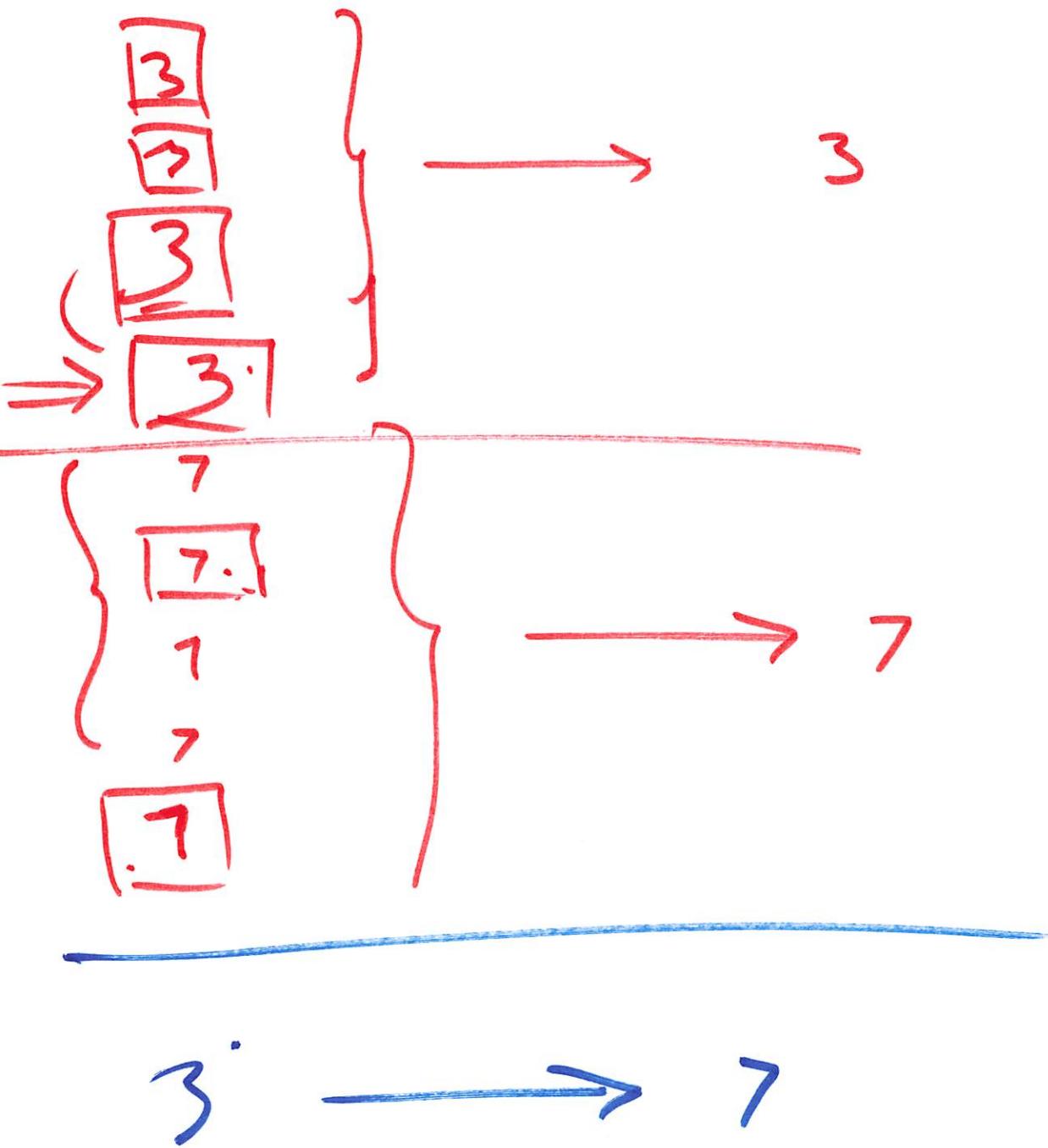
penalty

$$\rightarrow \sum |w_i|$$

$$\rightarrow \sum (w_i)^2$$

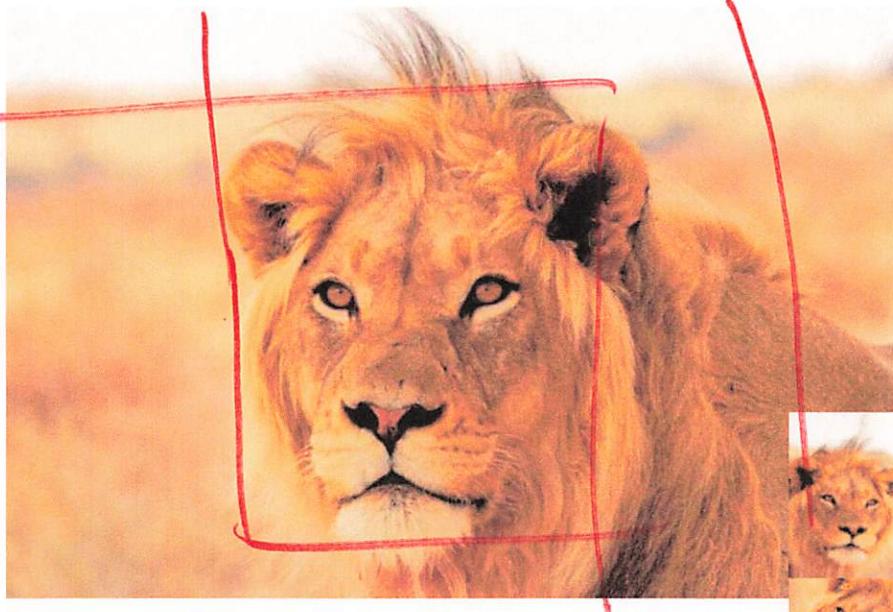
Lasso =

Ridge



This file is meant for personal use by amitava.basu@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Data Augmentation



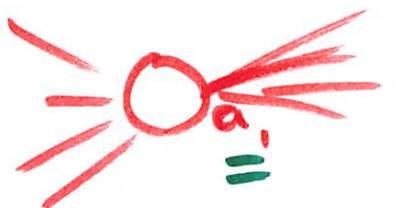
$\pm 15^\circ$
Noise
Shift
mirror
Stretch
Color
Crop
GAN

-source: towardsdatascience.com, MachineX: Image Data Augmentation Using Keras

THIS FILE IS MEANT FOR PERSONAL USE BY amitava.basu@gmail.com ONLY.

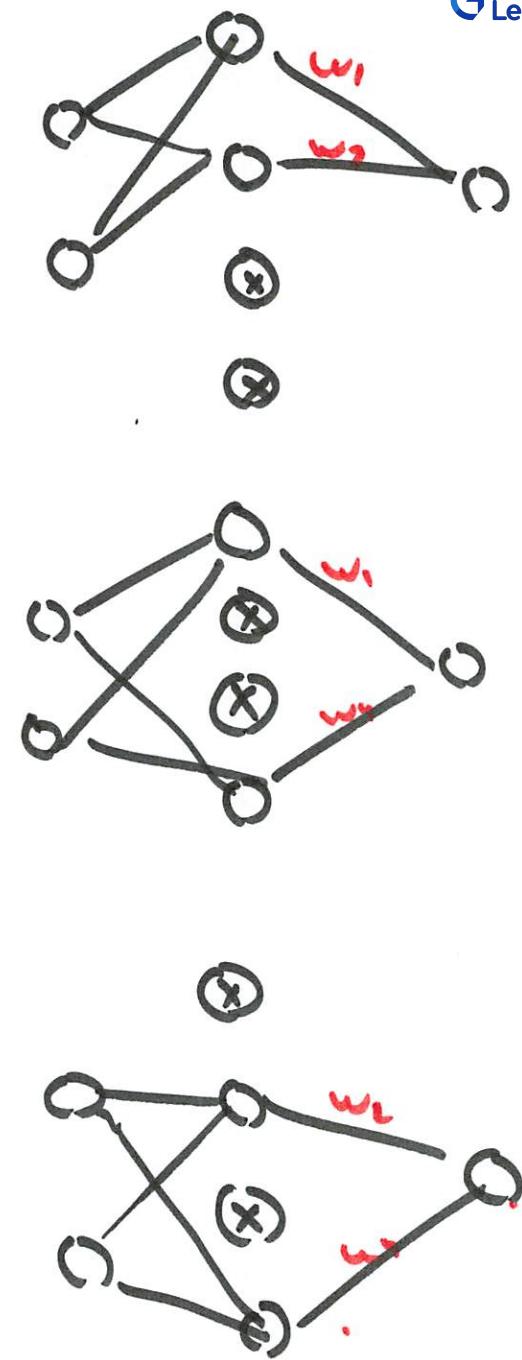
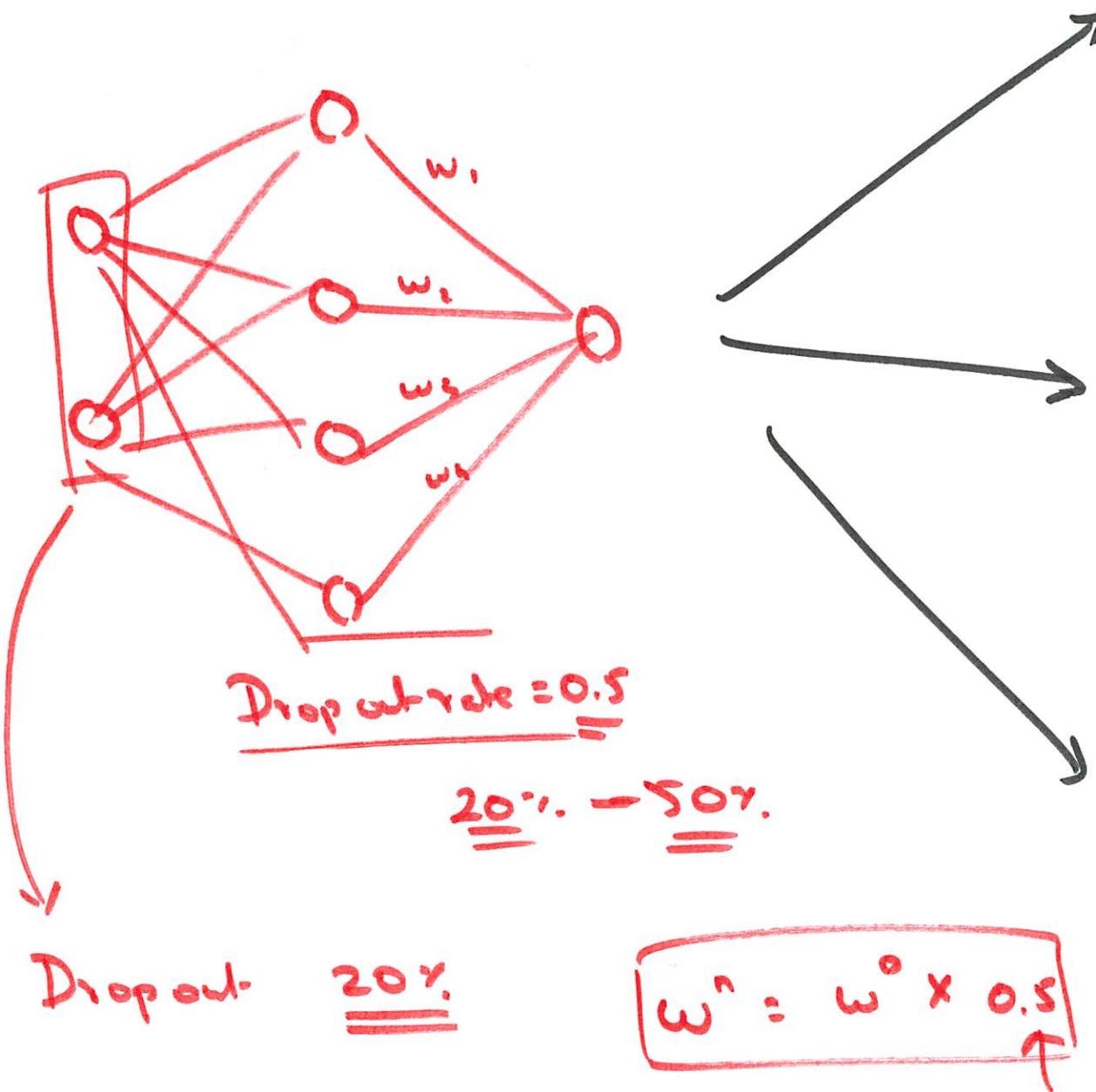
Sharing or publishing the contents in part or full is liable for legal action.

Co Adaptation



$$-0.1a_1 + 0.1a_2$$

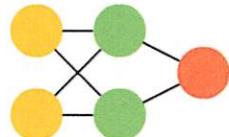
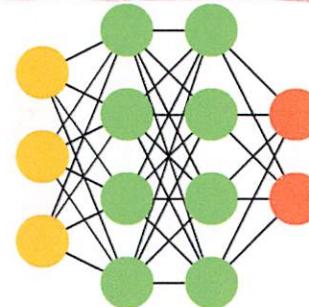
$$\underline{-0.3a_1 + 0.3a_2}$$



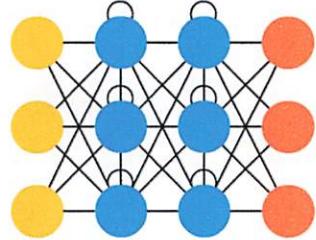
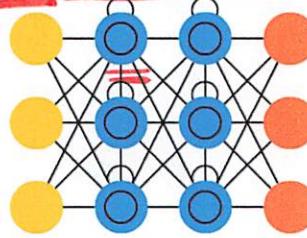
Types of NN

- Feed Forward
 - MLP
 - DNN
 - CNN
 - RNN
 - LSTM
 - . . .
 - Transf

Feed Forward (FF)

Deep Feed Forward (DFF)

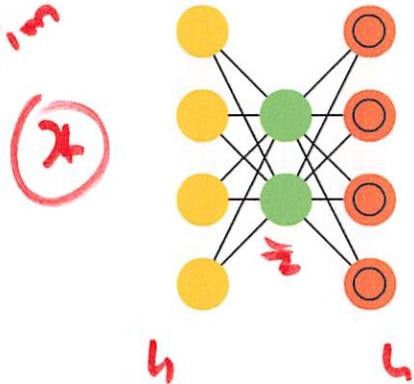
Recurrent Neural Network (RNN)

Long / Short Term Memory (LSTM)

17
1
17
17

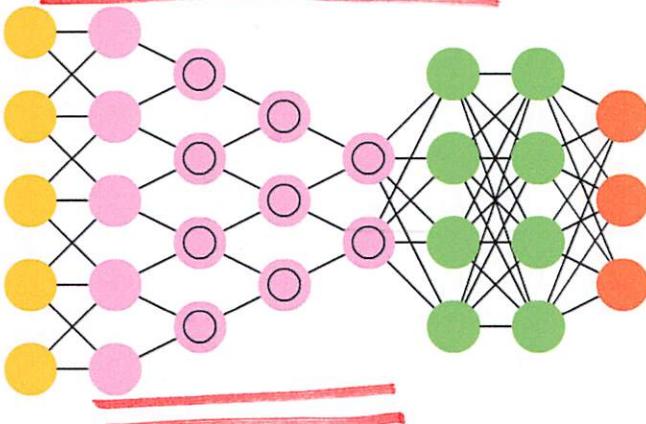
*Compressed
dim*

Auto Encoder (AE)



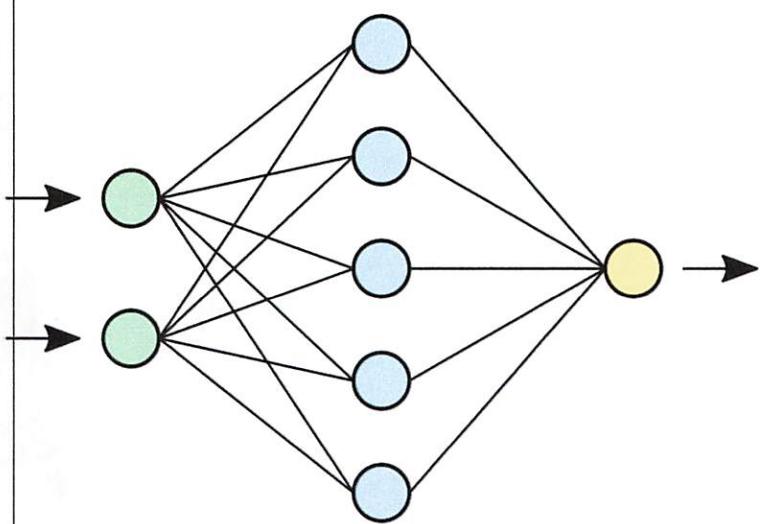
→ | 1 1 0 1 1 |
3 → | | | | |

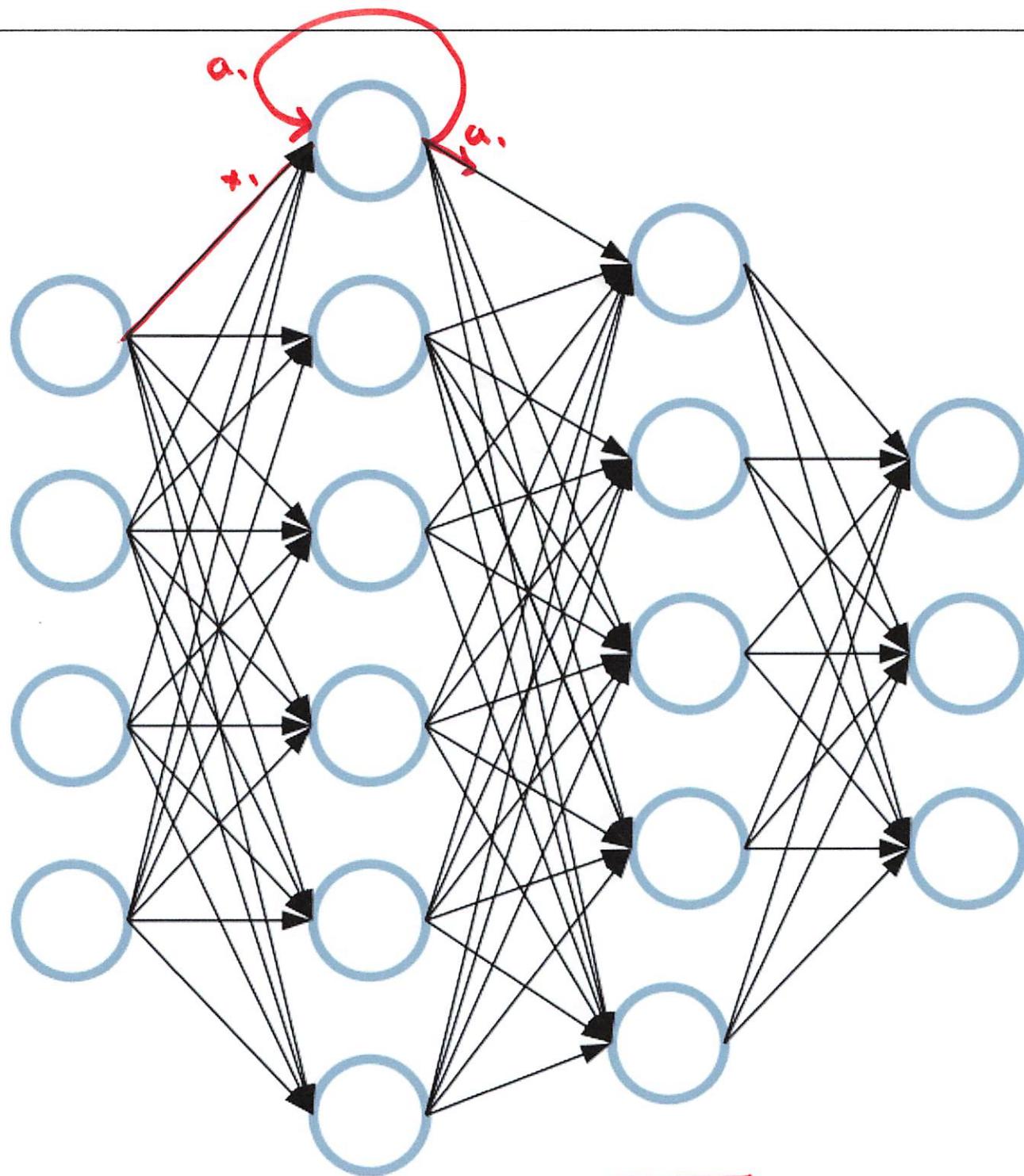
$| \hat{x} - x |$

Deep Convolutional Network (DCN)

17
17
17
17

Feed Forward Net.



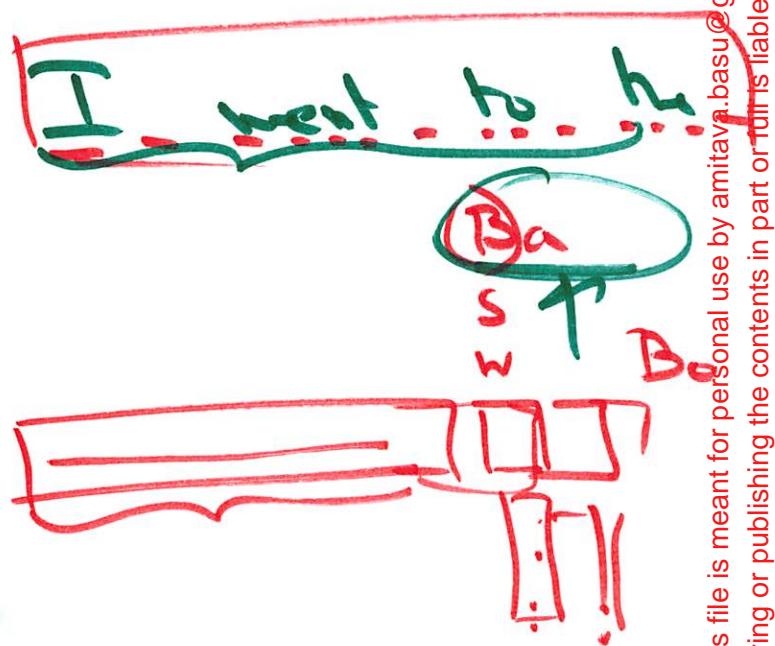
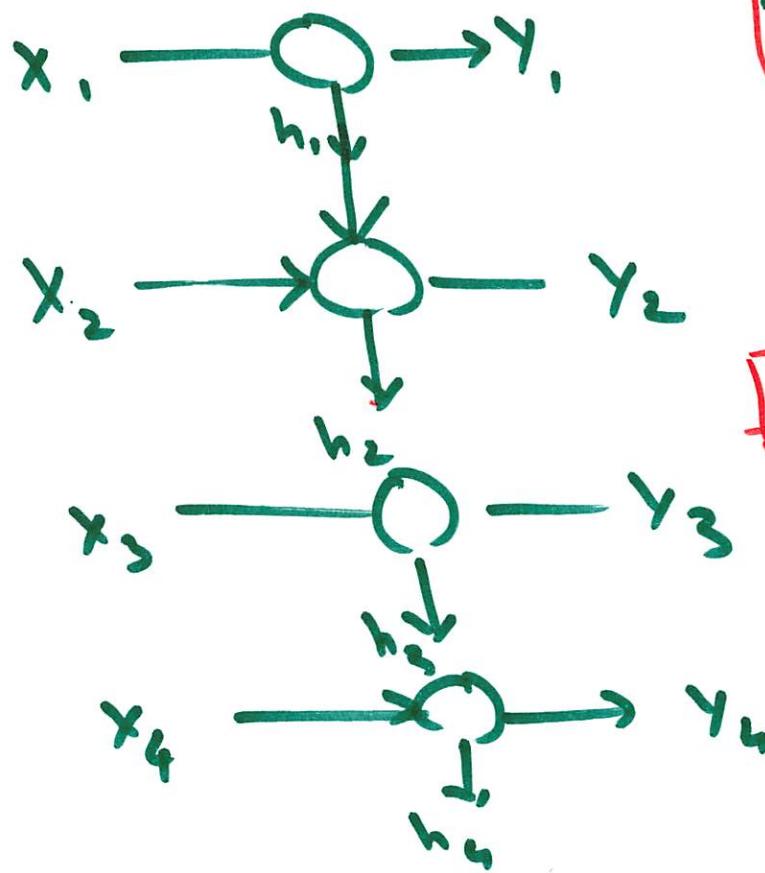




$$y = f(\omega x + b)$$

$$y = f(\omega x + \omega_0 b + b)$$

$$f = \sigma(\quad)$$

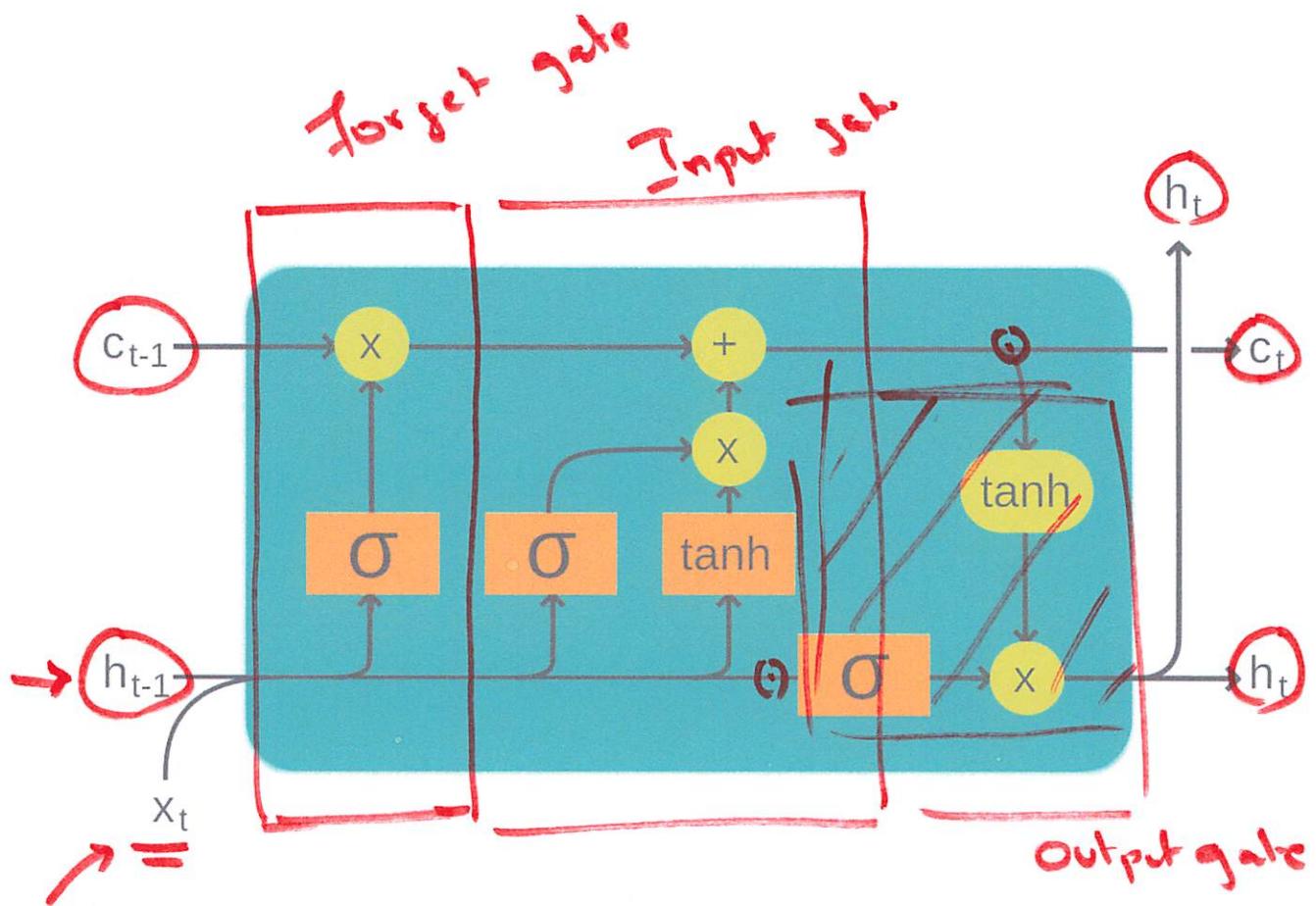


86

Text generation using an RNN

The meaning of life is ~~the tradition of the ancient human reproduction~~: it is less favorable to the good boy for when to remove her bigger. In the show's ~~agreement~~ unanimously resurfaced. The wild pastured with ~~consistent~~ street forests were incorporated by the 15th century BE. In 1996 the primary rapford undergoes an effort that the reserve conditioning, written into Jewish cities, sleepers to incorporate the .St Eurasia that activates the population. Mar??a Nationale, Kelli, Zedlat-Dukastoe, Florendon, Ptu's thought is. To adapt in most parts of North America, the dynamic fairy Dan please believes, the free speech are much related to the





Legend:

Layer	Componentwise	Copy	Concatenate

