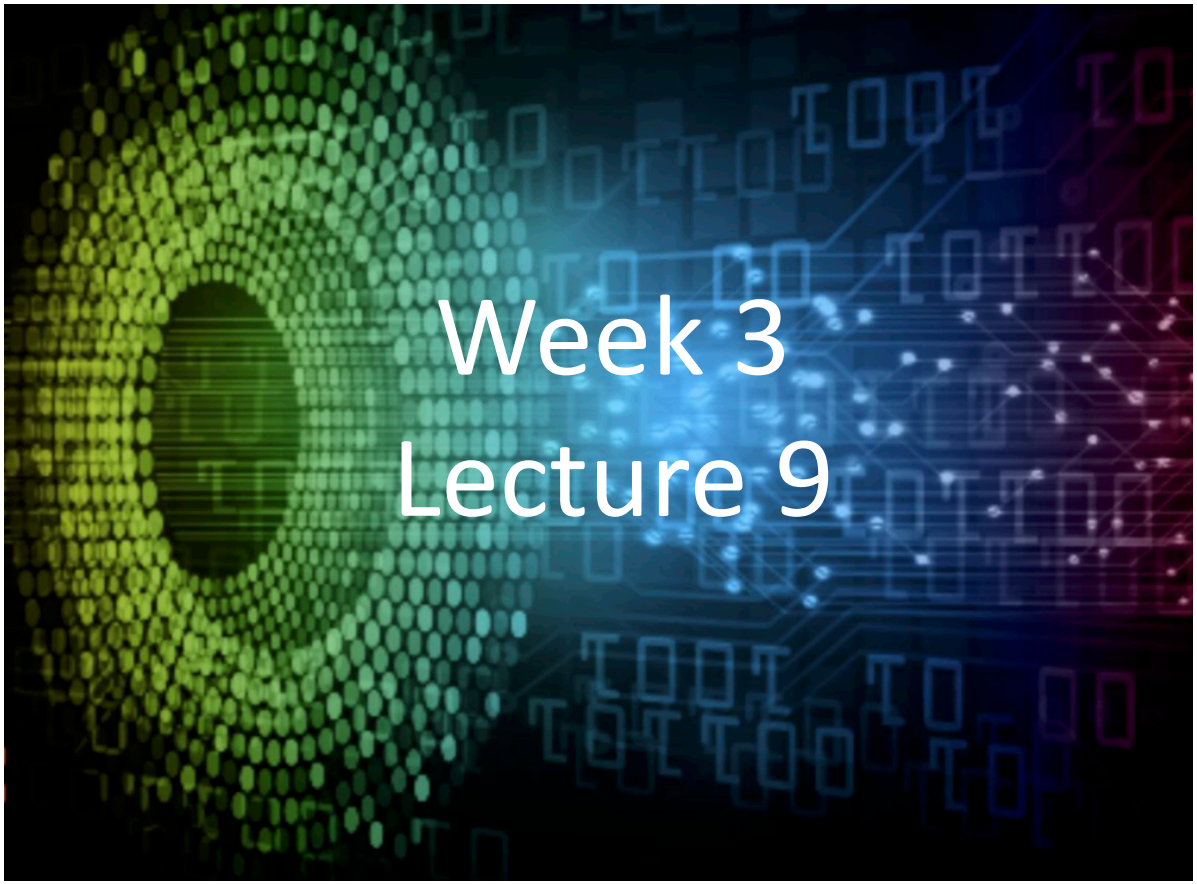
# Introduction to Deep Learning
# Applications and Theory




ECE 596 / AMATH 563

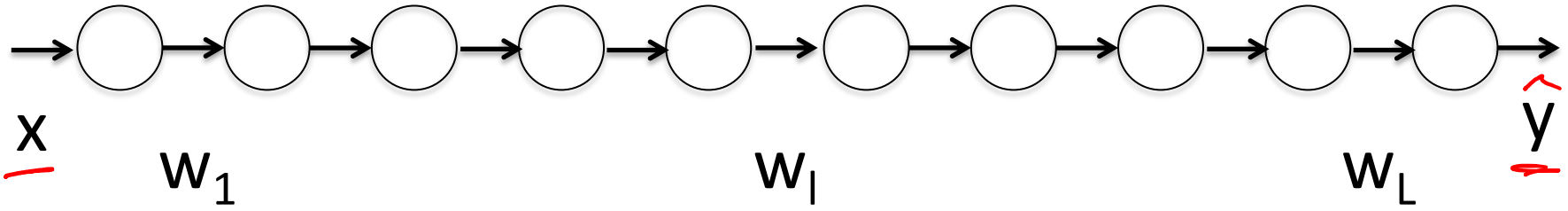
# Deep Learning Practices

1. Previous Lecture: Regularization
2. Normalization
3. Batch Normalization
4. Initialization
5. Hyperparameters tuning
6. eScience

# Vanishing/Exploding Gradients

- **Very deep** neural network

L = #layers




$$\hat{y} = w_L \cdot \ldots w_l \cdot \ldots w_2 \cdot w_1 \cdot x$$

$$w_l = w > 1; \qquad \hat{y} = w^L x \to \infty$$

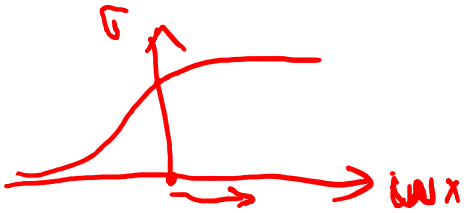
$$w_l = w < 1; \qquad \hat{y} = w^L x \to 0$$

# Vanishing/Exploding Gradients

- With **activation**:

$$\ldots w_3 \sigma_3(w_2 \sigma_2(\sigma_1'(w_1 x)))$$

- For **gradients**:




$$\frac{\partial J}{\partial w_1} = \sigma_3'(z_3) w_3 \sigma_2'(z_2) w_2 \sigma_1'(z_1) x$$

# Normalization of the datasets

- **Zero mean:**

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$x^{(i)\mu} = x^{(i)} - \mu$$

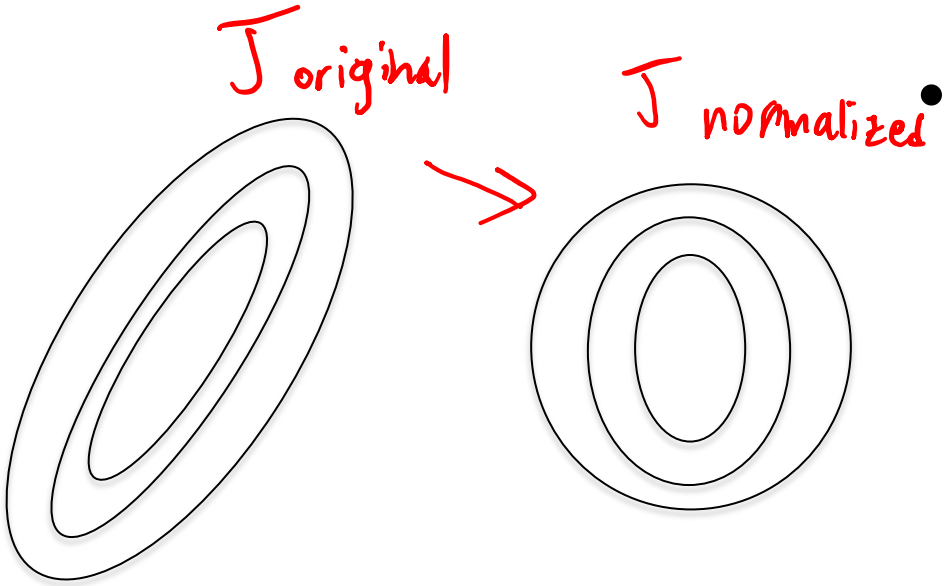
- **Normalized Variance:**

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m} x^{(i)^2}$$

$$x^{(i)\mu,\sigma^2} = x^{(i)\mu}./\sigma^2$$

# Intuition

- If **inputs have different scales**, the **cost function** will also have to include different scales

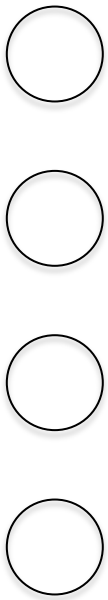



- Remember to **normalize all sets**: training, validation, testing

# Batch Normalization

- Normalize the outputs of each layer

$$z^{[l]} = w^{[l]} a^{[l-1]} + b^{[l]}$$



$$\mu = \frac{1}{m} \sum_{i=1}^{m} z^{(i)}$$

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^{m} (z^{(i)} - \mu)^2$$

$$z_{norm}^{(i)} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

# Learning batch normalization

- To make sure that the layers outputs will not be forced to be zero mean and variance 1

$$\tilde{z}^{[l](i)} = \gamma z^{[l](i)}_{norm} + \beta^{[l]}$$

- Forward propagation

$$a^{[l-1](i)} \rightarrow z^{[l](i)} \rightarrow_{BN} \tilde{z}^{[l](i)} \rightarrow a^{[l](i)}$$

- Can do per minibatch

# Pseudo-code

For each mini-batch

compute F-prop:

in each layer replace by

$$z^{[l](i)} \rightarrow \tilde{z}^{[l](i)}$$

$\gamma, \beta$

EMA

Do B-prop to compute

$$\nabla_{w^{[l]}}, \nabla_{\beta^{[l]}}, \nabla_{\gamma^{[l]}}$$

update

Testing: X

# Initialization

- Zero → Problematic
- Random Normal (0,1) -> Problematic
- Xavier (tanh):

$$Var(w^{[l]}) : 1/n^{[l-1]}$$

$$w^{[l]} = N(0,1) \cdot \sqrt{\frac{1}{n^{[l-1]}}}$$

# Initialization

- He (ReLU):

$$Var(w^{[l]}) : 2/n^{[l-1]}$$

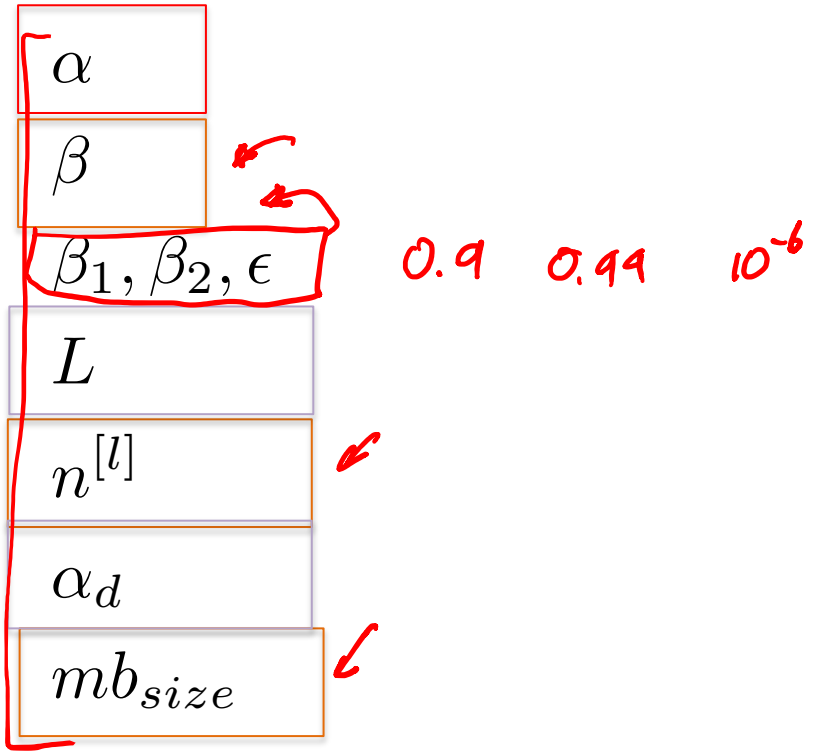
$$w^{[l]} = N(0,1) \cdot \sqrt{\frac{2}{n^{[l-1]}}}$$

- Other:

$$Var(w^{[l]}) : \frac{2}{n^{[l-1]} + n^{[l]}}$$

# Hyper Parameters

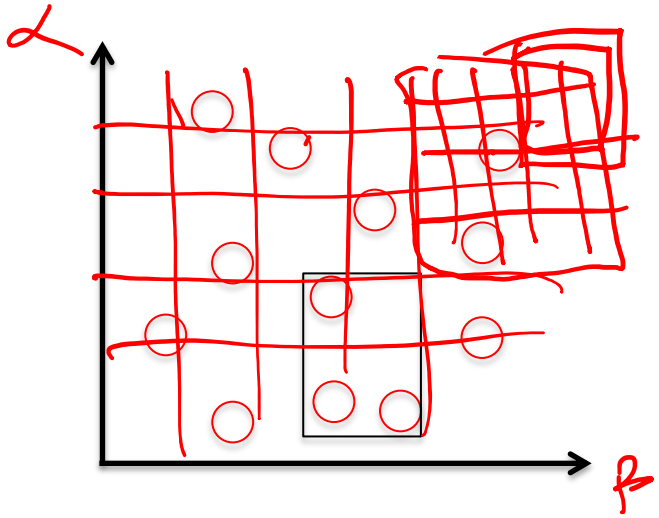
Params:



# Search

- Grid Search, Random Search
- Coarse to fine
- Linear for : L, $n^{[l]}$
- $\text{Log}_{10}$ scale for

$$\alpha, 1-\beta$$



# Data Science

- **UW eScience** lecture

**Sarah Stone**
eScience Executive Director

Weds
11:30