

Machine Learning

Computational Linguistics

Emory University

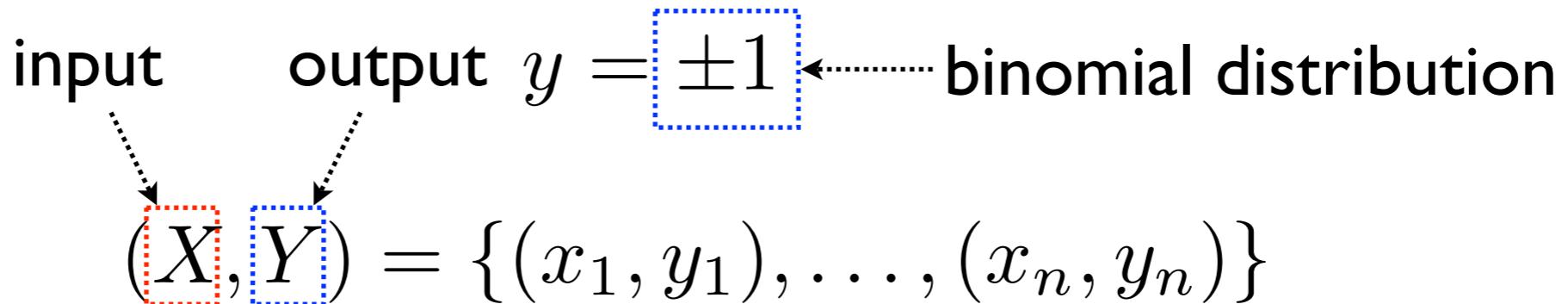
Jinho D. Choi



EMORY
UNIVERSITY



Supervised Learning



prediction $\hat{y} = \boxed{f(x)}$ predicts the output of x

Expected risk $E(f) = \int \boxed{\ell(\hat{y}; y)} \cdot \boxed{P(x, y)}$

loss function joint distribution **unknown!**

Empirical risk $\hat{E}(f) = \frac{1}{n} \sum_{i=1}^n \ell(\hat{y}_i; y_i)$ minimize!



Linear Prediction

$$\hat{E}(f) = \frac{1}{n} \sum_{i=1}^n \ell(\hat{y}_i; y_i)$$

$$\ell(\hat{y}; y) = \frac{1}{2} (\hat{y} - y)^2$$

$$\hat{y} = f(x) = \mathbf{w}^T \Phi(x) = \mathbf{w}^T \mathbf{x}$$

↑
feature vector

$$\ell(\mathbf{w}, \mathbf{x}; y) = \frac{1}{2} (\mathbf{w}^T \mathbf{x} - y)^2$$

Find a **weight vector** that minimizes the **loss**.

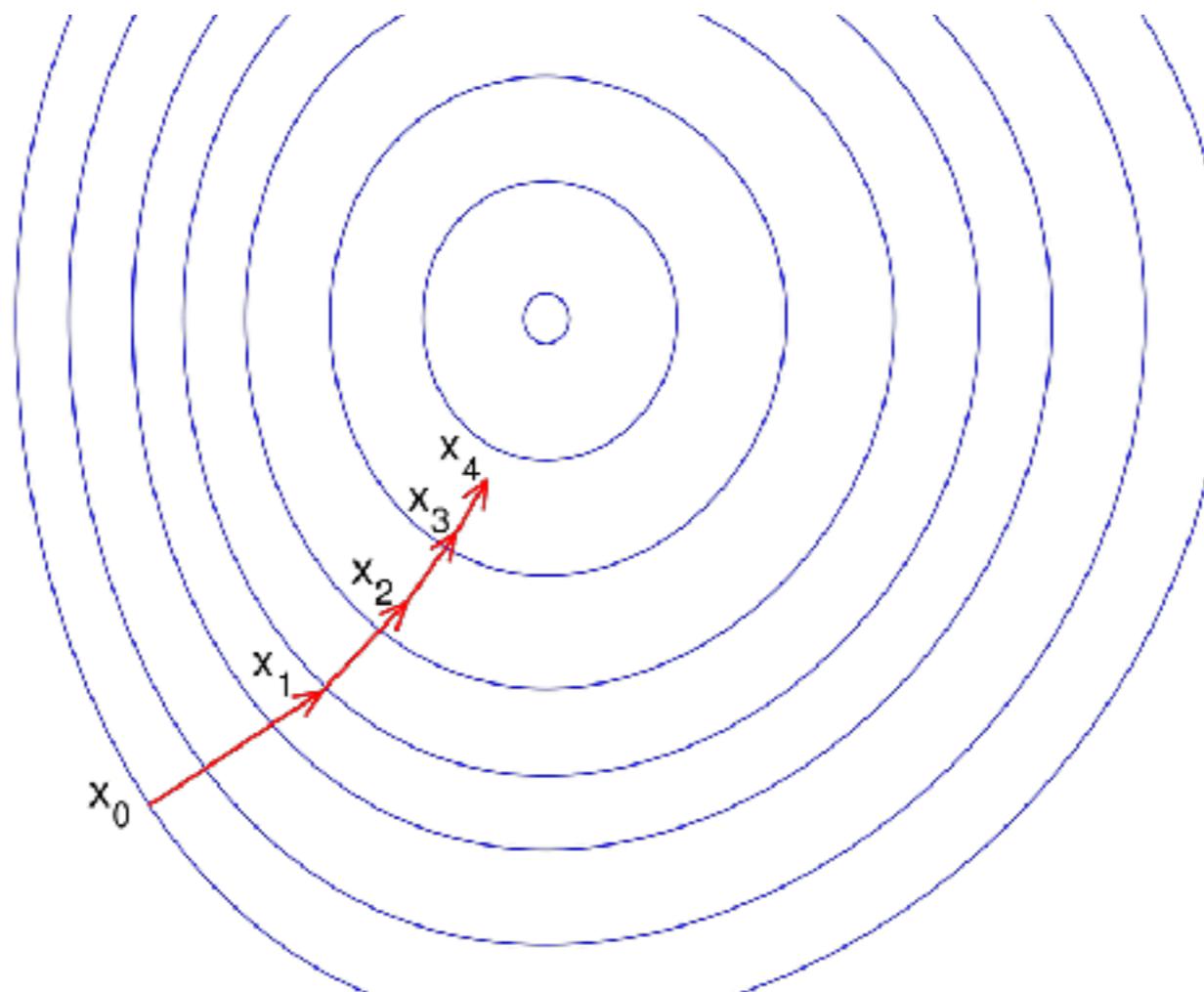


Gradient Descent

$$w_{t+1} \leftarrow w_t - \eta_t \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial w} \ell(w_t, x_i; y_i)$$

learning rate

derivative of the loss



Minimize loss

Derivative $\rightarrow 0$

Convex optimization

Global optimum?



Gradient Descent

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \mathbf{w}} \ell(\mathbf{w}_t, \mathbf{x}_i; y_i)$$

$$\ell(\mathbf{w}, \mathbf{x}; y) = \frac{1}{2} (\mathbf{w}^T \mathbf{x} - y)^2$$

$$\frac{\partial}{\partial \mathbf{w}} \ell(\mathbf{w}, \mathbf{x}; y) = \frac{\partial}{\partial \mathbf{w}} \frac{1}{2} (\mathbf{w}^T \mathbf{x} - y)^2 = \boxed{(\mathbf{w}^T \mathbf{x} - y) \mathbf{x}}$$

$$\boxed{\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t} - \eta_t \frac{1}{n} \sum_{i=1}^n (\mathbf{w}^T \mathbf{x}_i - y_i) \mathbf{x}_i$$

How often is the **weight vector** updated?



Stochastic Gradient Descent

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \boxed{\frac{1}{n} \sum_{i=1}^n (\mathbf{w}^T \mathbf{x}_i - y_i) \mathbf{x}_i}$$

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t \boxed{- \eta_t (\mathbf{w}_t^T \mathbf{x}_i - y_i) \mathbf{x}_i}$$

updated for **every** instance

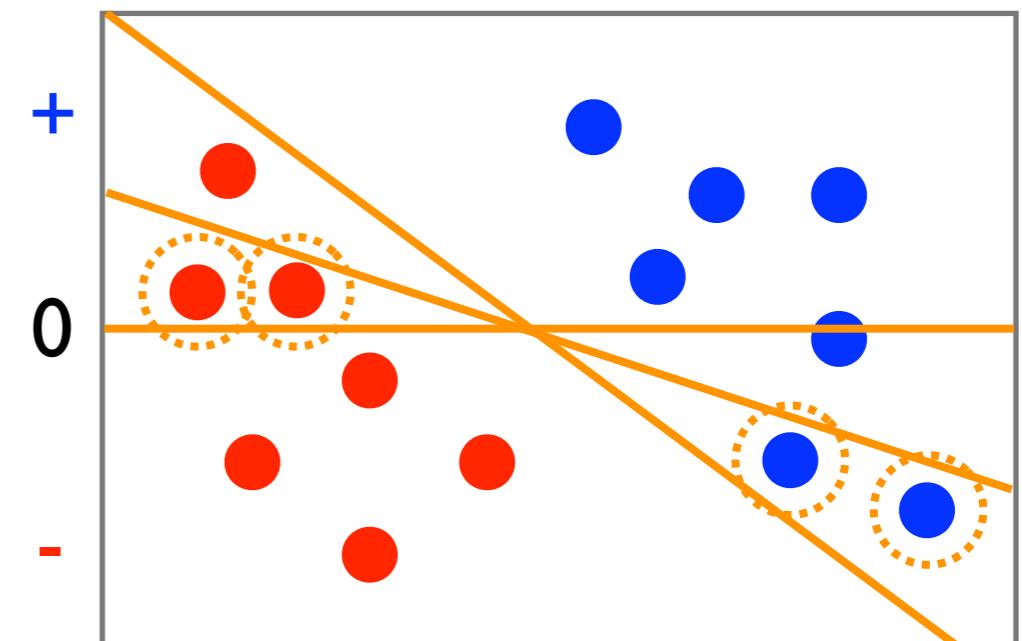
$$\mathbf{w}_0^T \mathbf{x}_1 > 0 \quad \mathbf{w}_1 \leftarrow \mathbf{w}_0 - \eta (\oplus + 1) \mathbf{x}_1$$

$$\mathbf{w}_1^T \mathbf{x}_2 < 0 \quad \mathbf{w}_2 \leftarrow \mathbf{w}_1 - \eta (\ominus + 1) \mathbf{x}_2$$

$$\mathbf{w}_2^T \mathbf{x}_3 < 0 \quad \mathbf{w}_3 \leftarrow \mathbf{w}_2 - \eta (\ominus - 1) \mathbf{x}_3$$

$$\mathbf{w}_3^T \mathbf{x}_4 > 0 \quad \mathbf{w}_4 \leftarrow \mathbf{w}_3 - \eta (\oplus - 1) \mathbf{x}_4$$

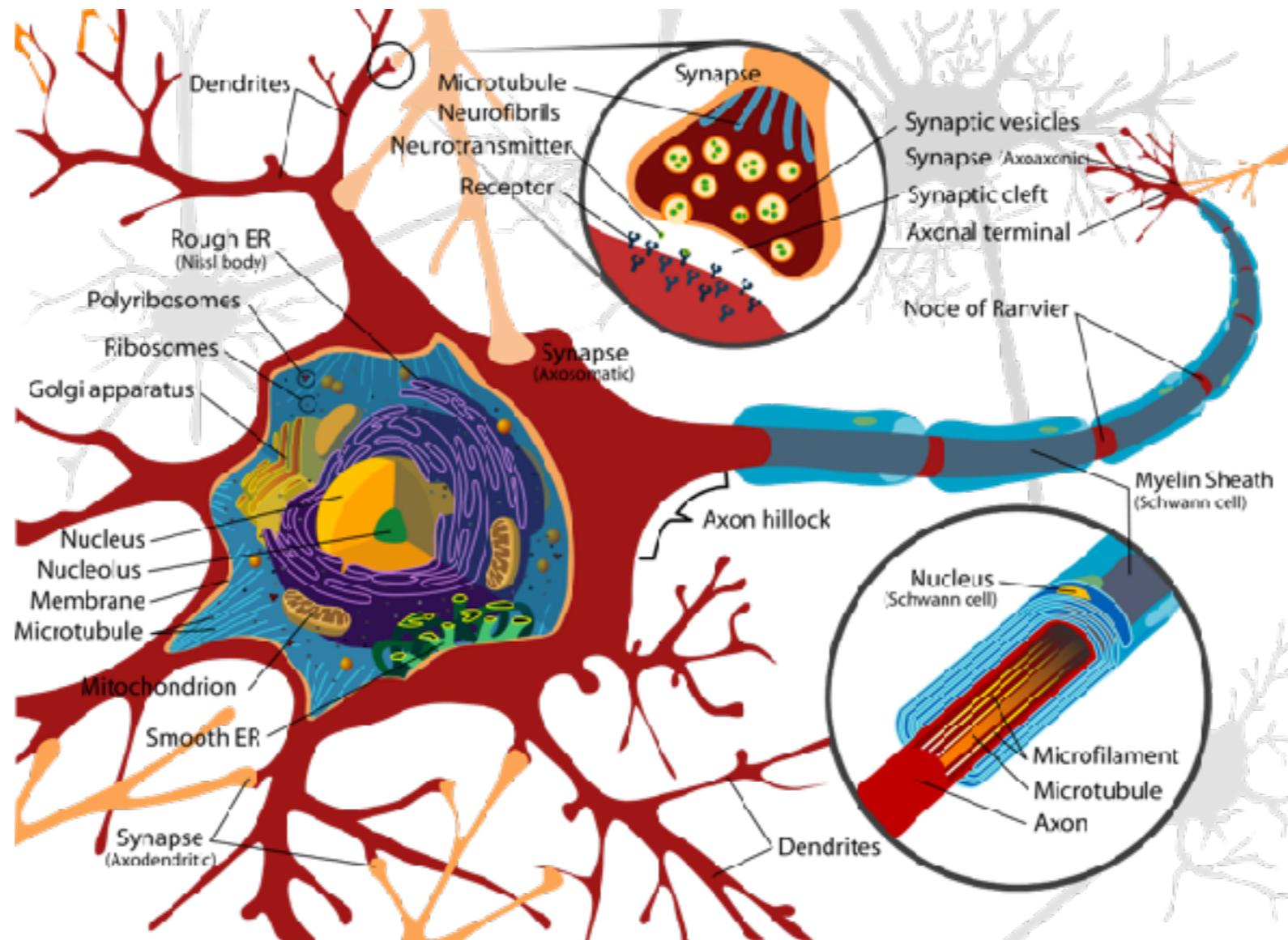
$$\mathbf{w}_0 \leftarrow 0$$



Human Brain

Learn
Adapt

Robust
Fault
Tolerant



How many **neurons**?

How many **synapsis** per **neuron**?



EMORY
UNIVERSITY



Hebbian Learning

Hebb's Learning Rule

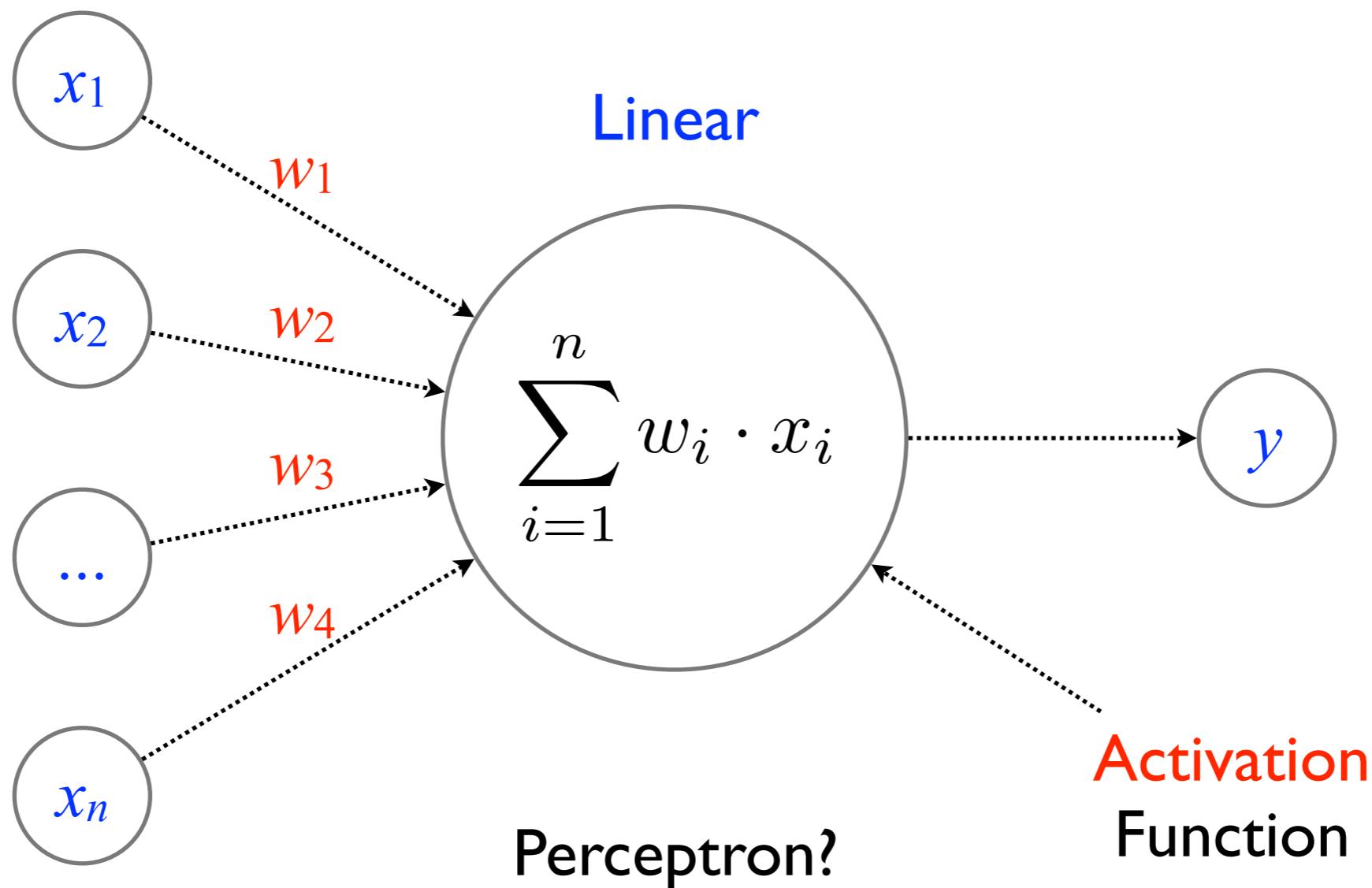
The **connections** between two **neurons** is strengthened if the **neurons** fire **simultaneously**.

Synchronous activation
increases the synaptic strength.

Asynchronous activation
decreases the synaptic strength.

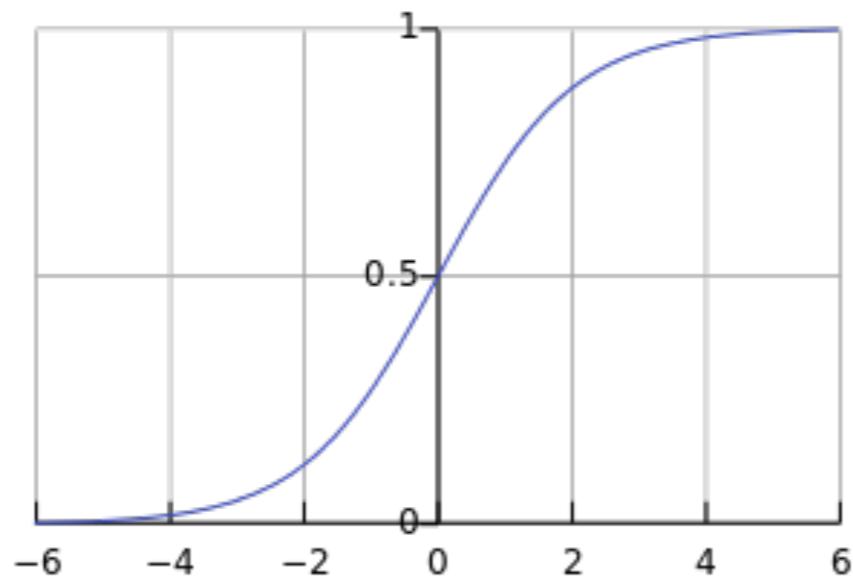


Single-layer Neural Network

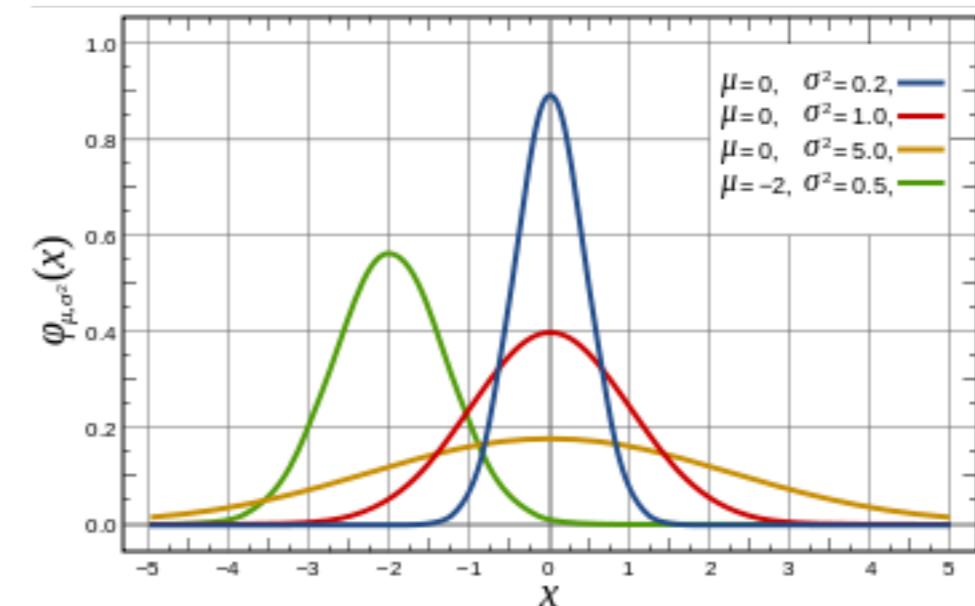


Non-Linear Function

Sigmoid



Gaussian



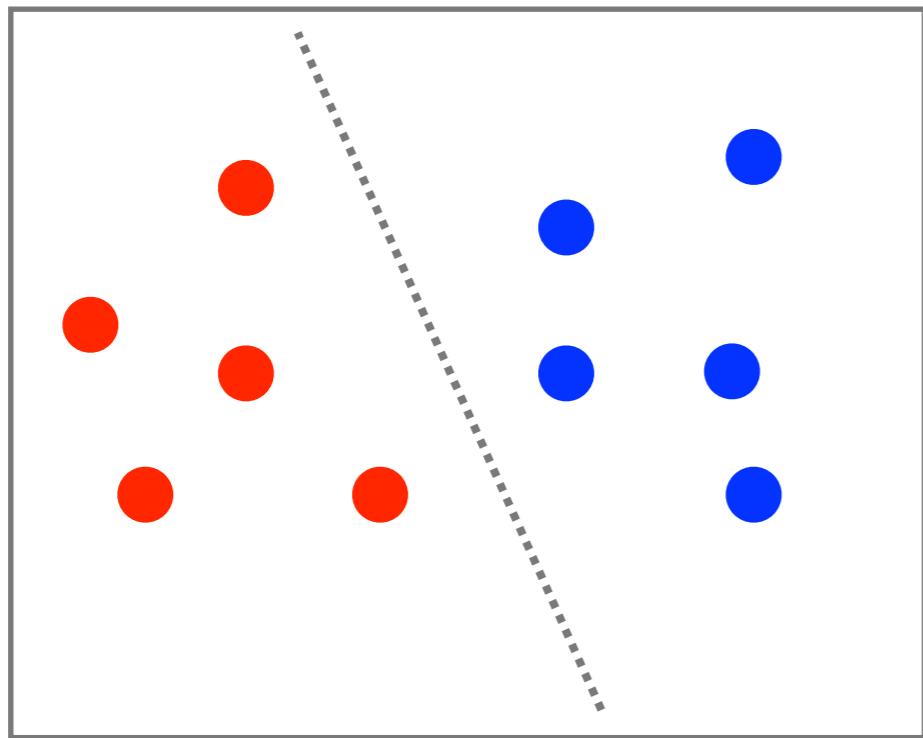
$$\frac{1}{1 + e^{-w \cdot x}}$$

$$e^{-\frac{\|w-x\|^2}{2a^2}}$$

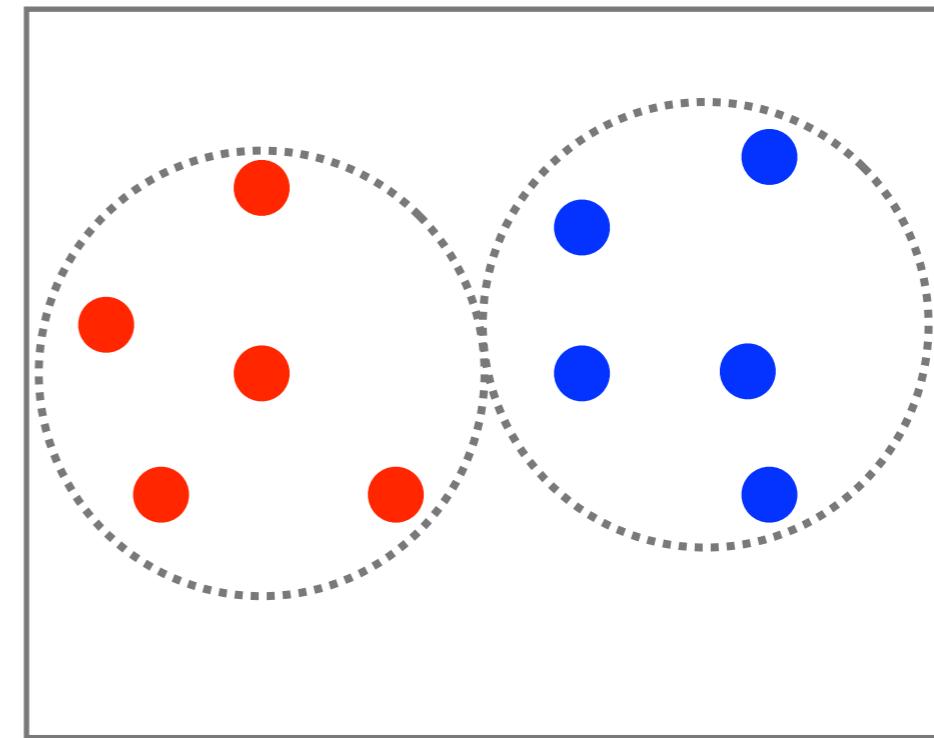


Sigmoid vs. Gaussian

Sigmoid



Gaussian



Find **hyperplanes**

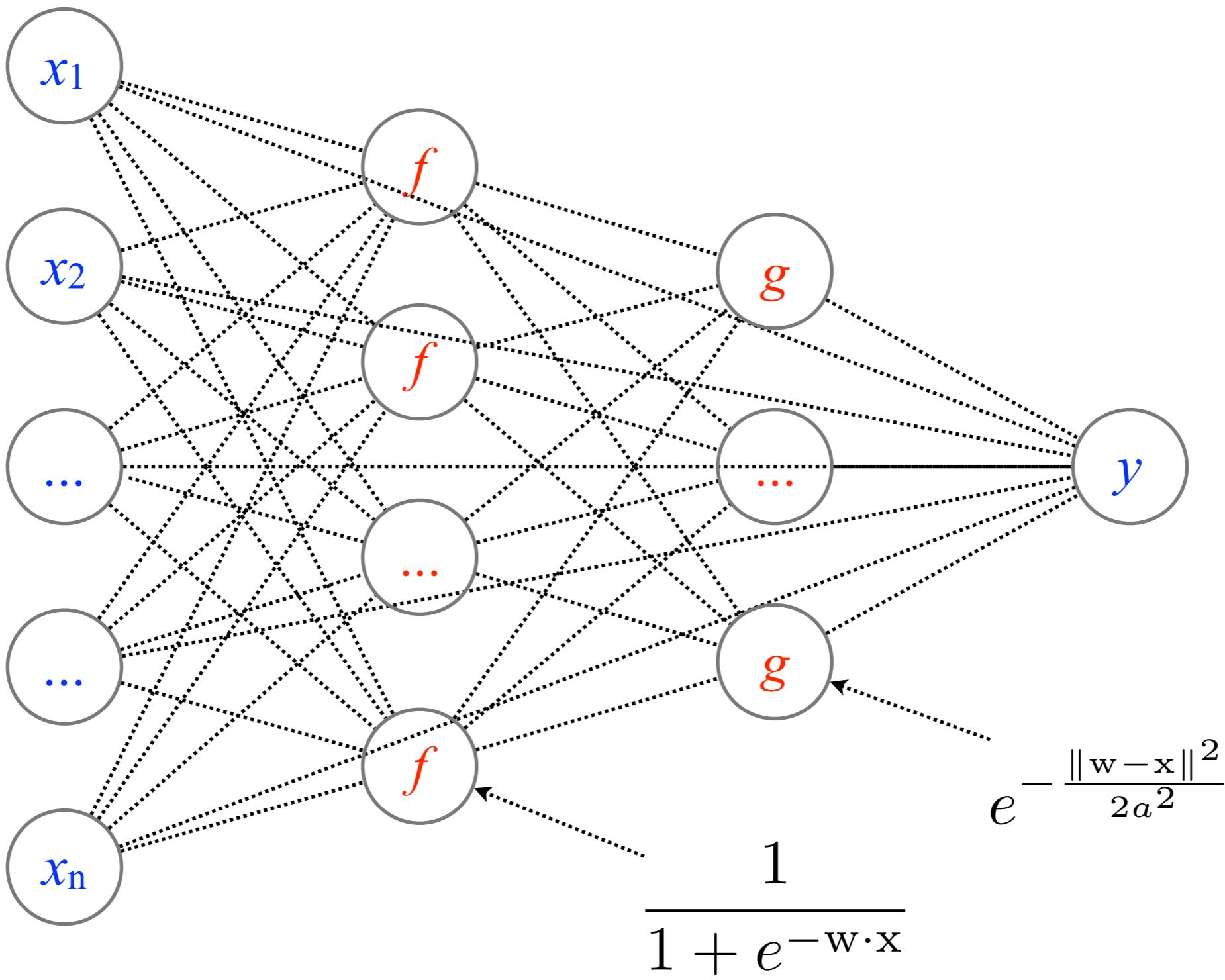
Find **hyperspheres**



EMORY
UNIVERSITY

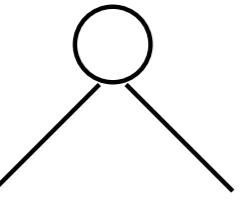
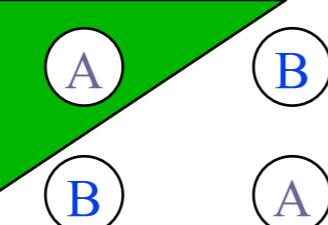
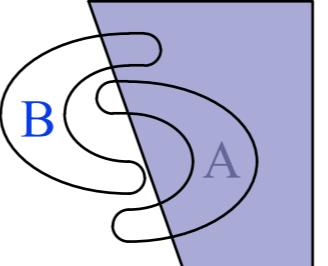
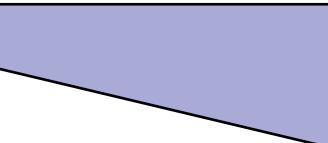
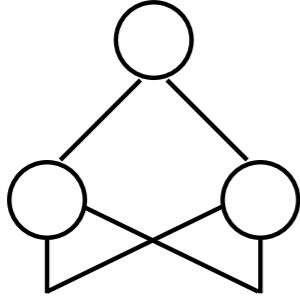
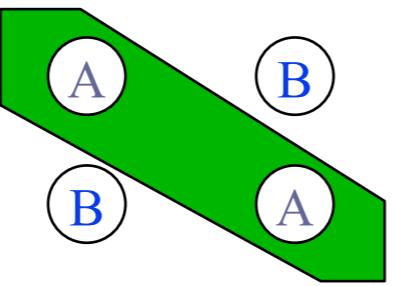
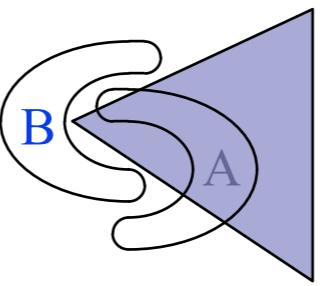
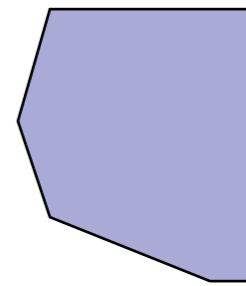
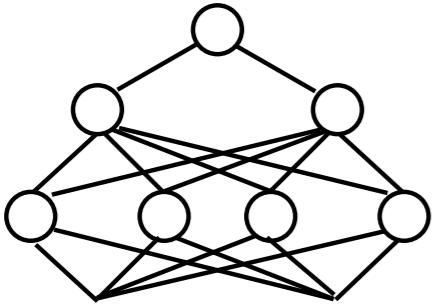
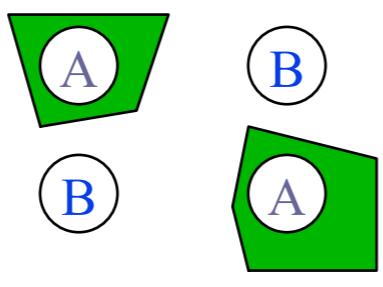
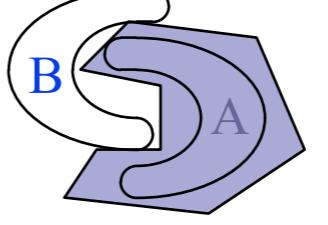
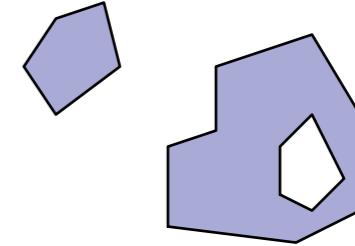


Multi-layer Neural Network



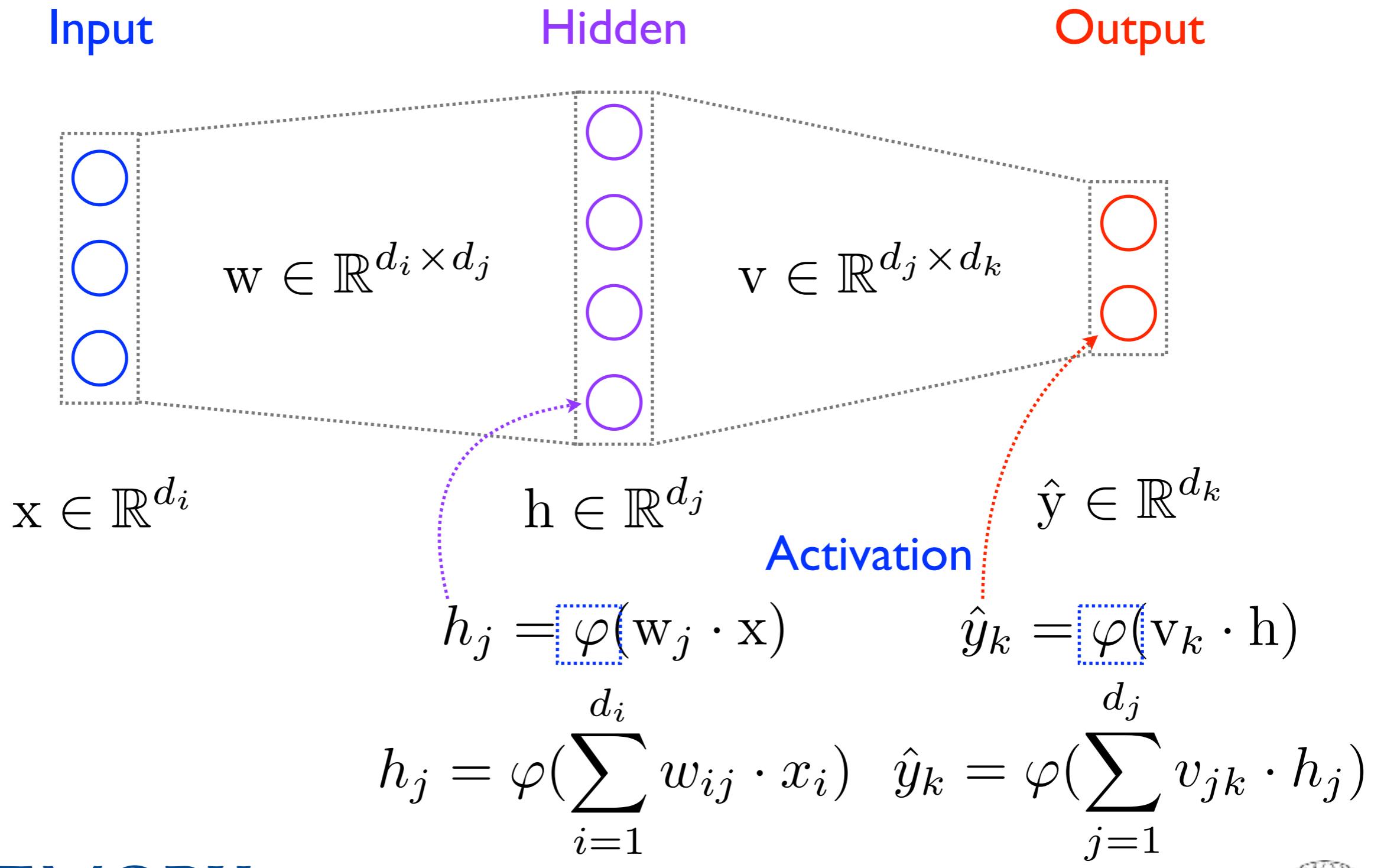
Multi-layer Neural Network

Comparison between different layers

# of Layers	Exclusive OR	Meshed Regions	General Regions
			
			
			



Forward Propagation



Backward Propagation

$$\hat{y}_k = \varphi\left(\sum_{j=1}^{d_j} v_{jk} \cdot h_j\right)$$

j'th hidden neuron
↓
k'th output neuron

$$\mathcal{E}(y_k, \hat{y}_k) = \frac{1}{2} (\hat{y}_k - y_k)^2$$

Least Square

$$z_k = \mathbf{v}_k \cdot \mathbf{h}$$

$$\mathcal{E} = \frac{1}{2} (\varphi(z_k) - y_k)^2$$

$$\frac{\partial \mathcal{E}}{\partial v_{jk}} = \frac{\partial \mathcal{E}}{\partial \hat{y}_k} \cdot \frac{\partial \hat{y}_k}{\partial z_k} \cdot \frac{\partial z_k}{\partial v_{jk}}$$



Backward Propagation

$$\frac{\partial \mathcal{E}}{\partial v_{jk}} = \boxed{\frac{\partial \mathcal{E}}{\partial \hat{y}_k}} \cdot \boxed{\frac{\partial \hat{y}_k}{\partial z_k}} \cdot \boxed{\frac{\partial z_k}{\partial v_{jk}}} \quad \boxed{\varphi(z_k) = \frac{1}{1 + e^{-z_k}}}$$

$$\frac{\partial}{\partial v_{jk}} \boxed{z_k} = \frac{\partial}{\partial v_{jk}} \left(\sum_{l=1}^{d_j} v_{lk} \cdot h_l \right) = h_j$$

$$\begin{aligned} \frac{\partial}{\partial z_k} \boxed{\hat{y}_k} &= \frac{\partial}{\partial z_k} \boxed{\varphi(z_k)} = \varphi(z_k)(1 - \varphi(z_k)) \\ &= \hat{y}_k(1 - \hat{y}_k) \end{aligned}$$

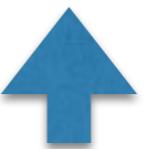
$$\frac{\partial}{\partial \hat{y}_k} \boxed{\mathcal{E}} = \frac{\partial}{\partial \hat{y}_k} \boxed{\frac{1}{2}(\hat{y}_k - y_k)^2} = \hat{y}_k - y_k$$

$$\Delta \mathcal{E}_{jk} = \boxed{(\hat{y}_k - y_k)} \cdot \boxed{\hat{y}_k(1 - \hat{y}_k)} \cdot \boxed{h_j}$$



Backward Propagation

$$\Delta \mathcal{E}_{jk} = (\hat{y}_k - y_k) \cdot \hat{y}_k(1 - \hat{y}_k) \cdot h_j$$



$$\hat{y}_k = \varphi\left(\sum_{j=1}^{d_j} v_{jk} \cdot h_j\right)$$

$$h_j = \varphi\left(\sum_{i=1}^{d_i} w_{ij} \cdot x_i\right)$$



$$\Delta \mathcal{E}_{ij} = (h_j - ?) \cdot h_j(1 - h_j) \cdot x_i$$



Backward Propagation

$$\Delta \mathcal{E}_{jk} = \boxed{(\hat{y}_k - y_k) \cdot \hat{y}_k(1 - \hat{y}_k)} \cdot h_j \\ \delta_k$$

$$\Delta \mathcal{E}_{ij} = \boxed{(h_j - ?)} \cdot h_j(1 - h_j) \cdot x_i \\ \frac{\partial \mathcal{E}}{\partial h_j} = \boxed{\sum_{l=1}^{d_k} (\delta_l \cdot v_{jl})}$$

Sum of all the gradients
made by the j 'th hidden unit

$$\Delta \mathcal{E}_{ij} = \sum_{l=1}^{d_k} (\delta_l \cdot v_{jl}) \cdot h_j(1 - h_j) \cdot x_i$$



Backward Propagation

Input → Hidden

$$h_j = \varphi\left(\sum_{i=1}^{d_i} w_{ij} \cdot x_i\right)$$

Hidden → Output

$$\hat{y}_k = \varphi\left(\sum_{j=1}^{d_j} v_{jk} \cdot h_j\right)$$

$$\delta_k = (\hat{y}_k - y_k) \cdot \hat{y}_k(1 - \hat{y}_k)$$

$$\Delta \mathcal{E}_{ij} = \sum_{l=1}^{d_k} (\delta_l \cdot v_{jl}) \cdot h_j(1 - h_j) \cdot x_i \quad \Delta \mathcal{E}_{jk} = \delta_k \cdot h_j$$

$$w_{ij}^{t+1} \leftarrow w_{ij}^t - \eta_t \Delta_{ij}$$

$$v_{jk}^{t+1} \leftarrow v_{jk}^t - \eta_t \Delta_{jk}$$



EMORY
UNIVERSITY

