Lesson 4: Feedforward Neural Network (FFNN)

Definition: a) A FFNN consists of layers of computational units (i.e., neurons), usually interconnected in a feedforward way.

b) A FFNN is an artificial Neural Network (ANN) wherein connections between the nodes (neurons) do not form a cycle.

Context :

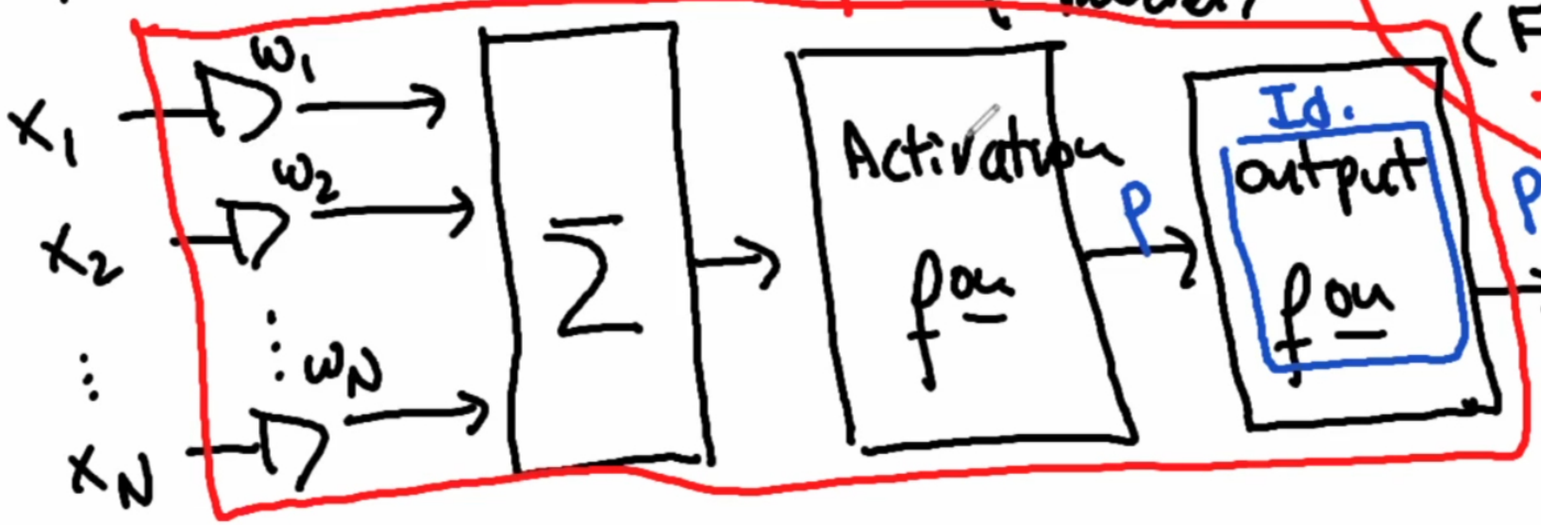
Biological NN (BNN) | Artificial NN (ANN)

Artificial Neural Network

i) FFNN
(Feedforward NN)

ii) FBNN / RNN
(Feedback NN / Recurrent NN)
with Memory

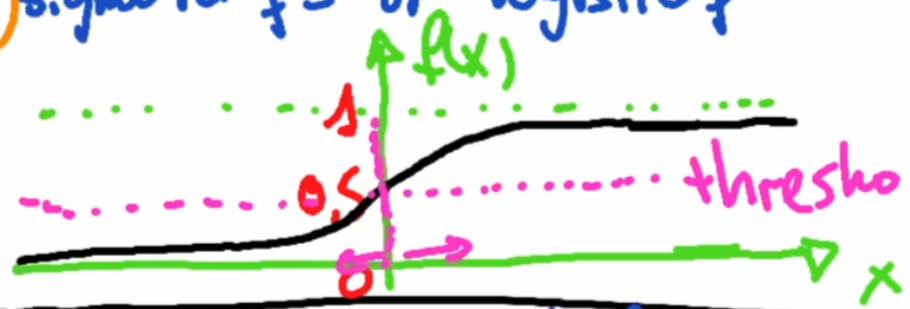
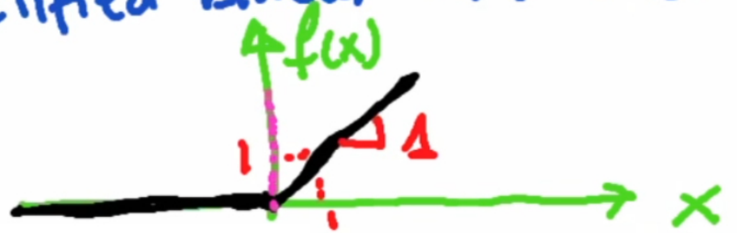
Def : Artificial Neuron (with a model)



In this course, output fon is the identity fon

Types of activation fcns (Basic)

For this course!

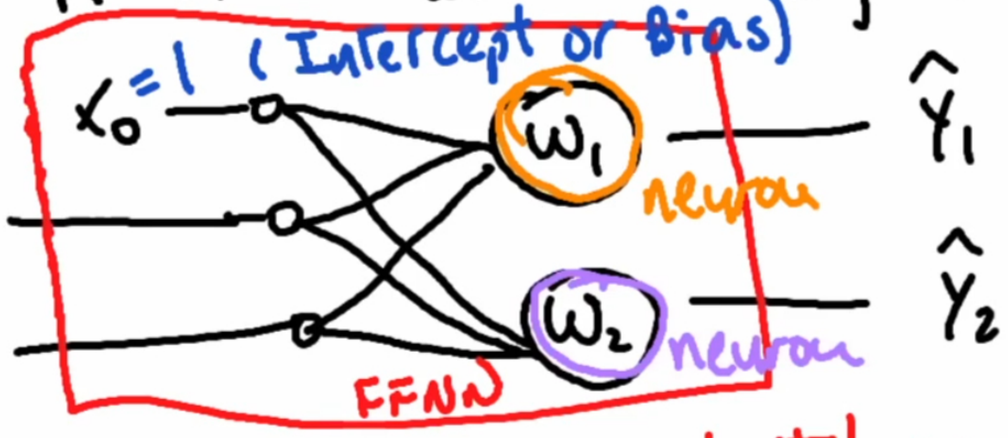
Graph	Equation	Derivative
1) Sigmoid fcn or logistic fcn 	$f(x) = \frac{1}{1+e^{-x}}$	$f'(x) = f(x)(1-f(x))$
2) Hyperbolic tangent fcn 	$f(x) = \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$	$f'(x) = 1 - f^{2}(x)$
3) Rectified Linear unit (ReLU) 	$f(x) = \begin{cases} 0, & \text{if } x \le 0 \\ x, & \text{if } x > 0 \end{cases} = \max\{0, x\}$ $= x \cdot \mathbb{I}\{x > 0\}$ → Indicator fcn: $\mathbb{I}\{x>0\} = \begin{cases} 1, & x>0 \\ 0, & x \le 0 \end{cases}$	$f'(x) = \begin{cases} 0, & x \le 0 \\ 1, & x > 0 \end{cases}$ $\begin{cases} 1, & x>0 \\ 0, & x \le 0 \end{cases}$

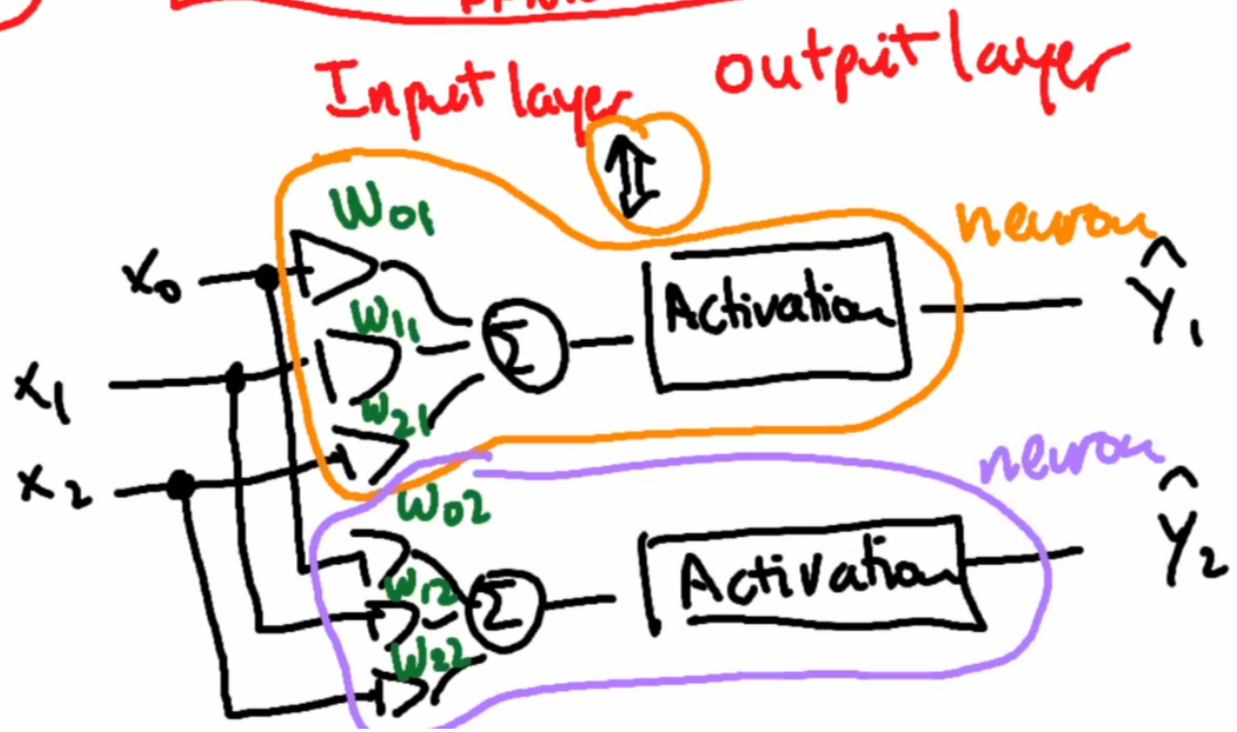
Example: A FFNN with two layers with two features.

Features

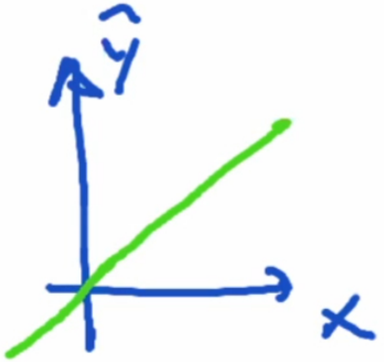
x_1

x_2





Without Bias (~~x_0~~)



With Bias (x_0)

$$W = \begin{bmatrix} 1 & 1 \\ w_1 & w_2 \\ 1 & 1 \end{bmatrix}$$

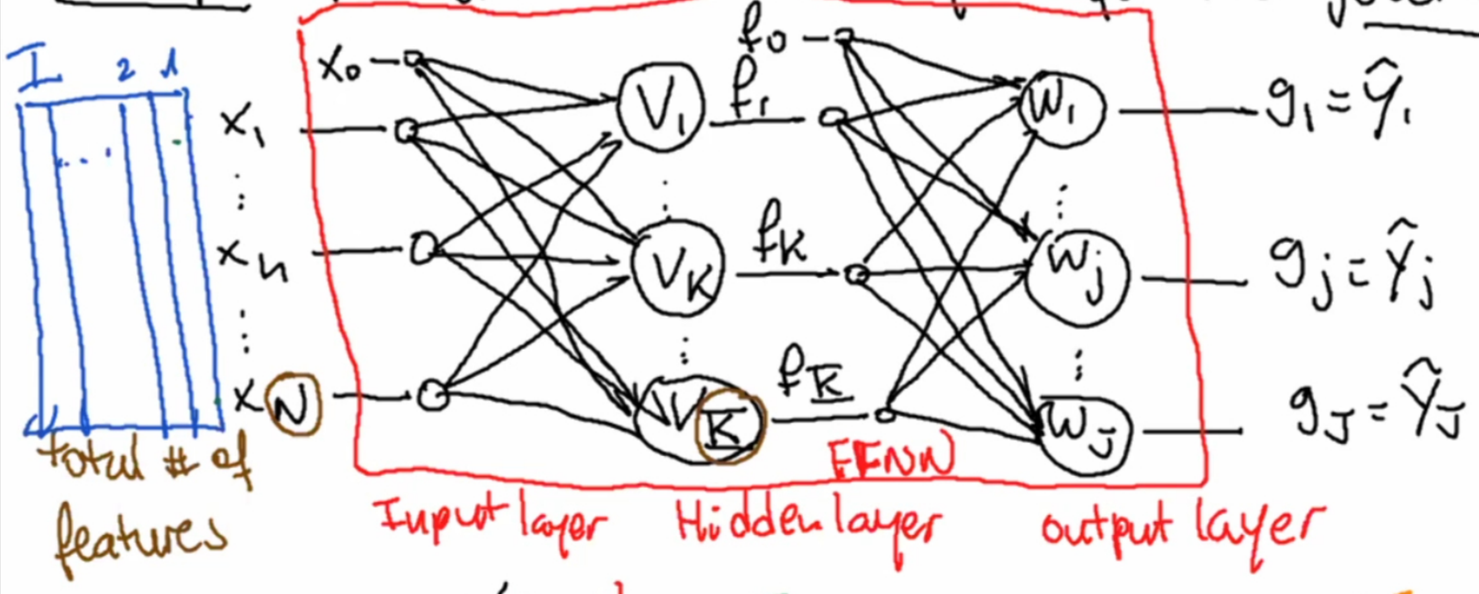
total # of features: N+1

3 × 2

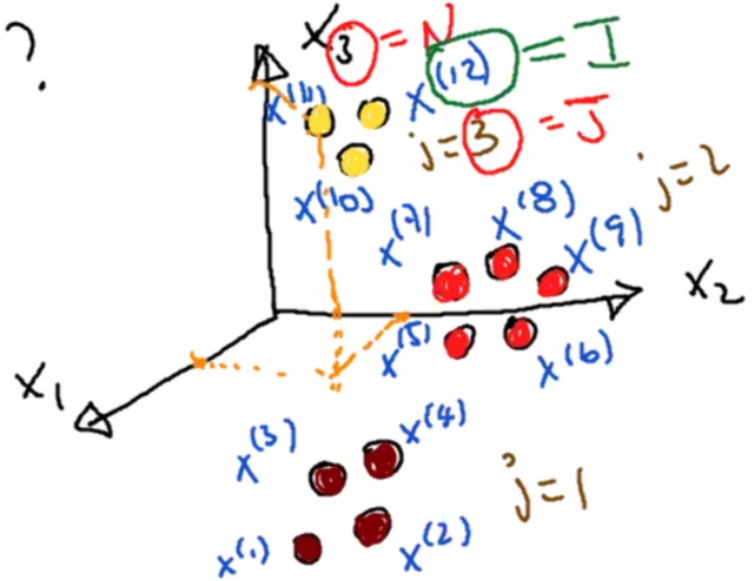
$$W = \begin{bmatrix} w_{01} & w_{02} \\ w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}$$

output-variable dimension

Example: A FFNN with three layers for the general case:

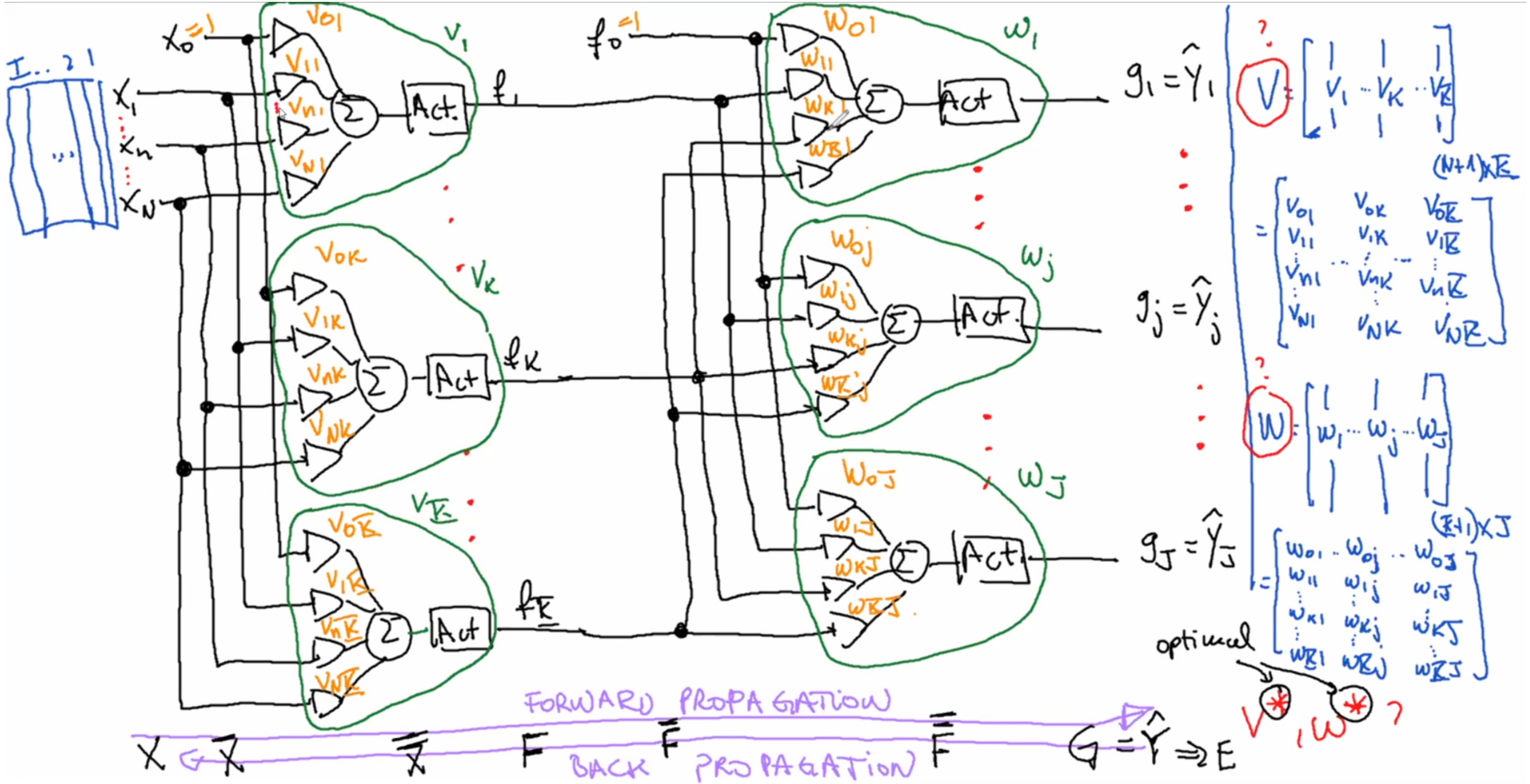


- ✓ $N \triangleq$ total # of features
- ✓ $K \triangleq$ total # of hidden neurons
- ✓ $J \triangleq$ output-variable dimension
- ✓ $I \triangleq$ total # of training examples
  - ↑ By definition



$$x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ x_3^{(i)} \end{bmatrix}$$

$$i \in \{1, 2, \cdots \underbrace{I}_{=12}\}$$


I .. 2 1
$x_0 = 1$
x_1
x_n
x_N
v_{01}
v_{11}
v_{n1}
v_{N1}
v_1
Act.
f_1
v_{0K}
v_{1K}
v_{nK}
v_{NK}
v_K
f_K
$v_{0\bar{K}}$
$v_{1\bar{K}}$
$v_{n\bar{K}}$
$v_{N\bar{K}}$
$v_{\bar{K}}$
$f_{\bar{K}}$
$f_0 = 1$
w_{01}
w_{11}
w_{K1}
$w_{\bar{K}1}$
w_1
$g_1 = \hat{y}_1$
w_{0j}
w_{ij}
w_{Kj}
$w_{\bar{K}j}$
w_j
$g_j = \hat{y}_j$
w_{0J}
w_{1J}
w_{KJ}
$w_{\bar{K}J}$
w_J
$g_J = \hat{y}_J$
FORWARD PROPAGATION
BACK PROPAGATION
X
Z
Z
F
F
F
$G = \hat{Y} \Rightarrow E$
$V = \begin{bmatrix} | & | & | \\ v_1 & v_K & v_{\bar{K}} \\ | & | & | \end{bmatrix}$
$(N+1) \times \bar{K}$
$= \begin{bmatrix} v_{01} & v_{0K} & v_{0\bar{K}} \\ v_{11} & v_{1K} & v_{1\bar{K}} \\ v_{n1} & v_{nK} & v_{n\bar{K}} \\ v_{N1} & v_{NK} & v_{N\bar{K}} \end{bmatrix}$
$W = \begin{bmatrix} | & | & | \\ w_1 & w_j & w_J \\ | & | & | \end{bmatrix}$
$(\bar{K}+1) \times J$
$= \begin{bmatrix} w_{01} & w_{0j} & w_{0J} \\ w_{11} & w_{1j} & w_{1J} \\ w_{K1} & w_{Kj} & w_{KJ} \\ w_{\bar{K}1} & w_{\bar{K}j} & w_{\bar{K}J} \end{bmatrix}$
optimal
V^*, W^* ?

Goal of FFNN: Define $\theta = (V, W)$. Then,

$$\theta^* = (V^*, W^*) = \underset{\theta}{\operatorname{argmin}}\, E(\underline{x}, \underline{y}; \theta) \quad ?$$

$\hat{y} \triangleq$ Estimated output (FFNN)

$y \triangleq$ Actual output

~~MyE~~

Example:

$$z = \begin{array}{c|c} \text{index} & \text{value} \\ 1 & 3 \\ 2 & 2 \\ 3 & 1 \end{array}$$

$\min z = 1$

$\operatorname{argmin} z = 3$

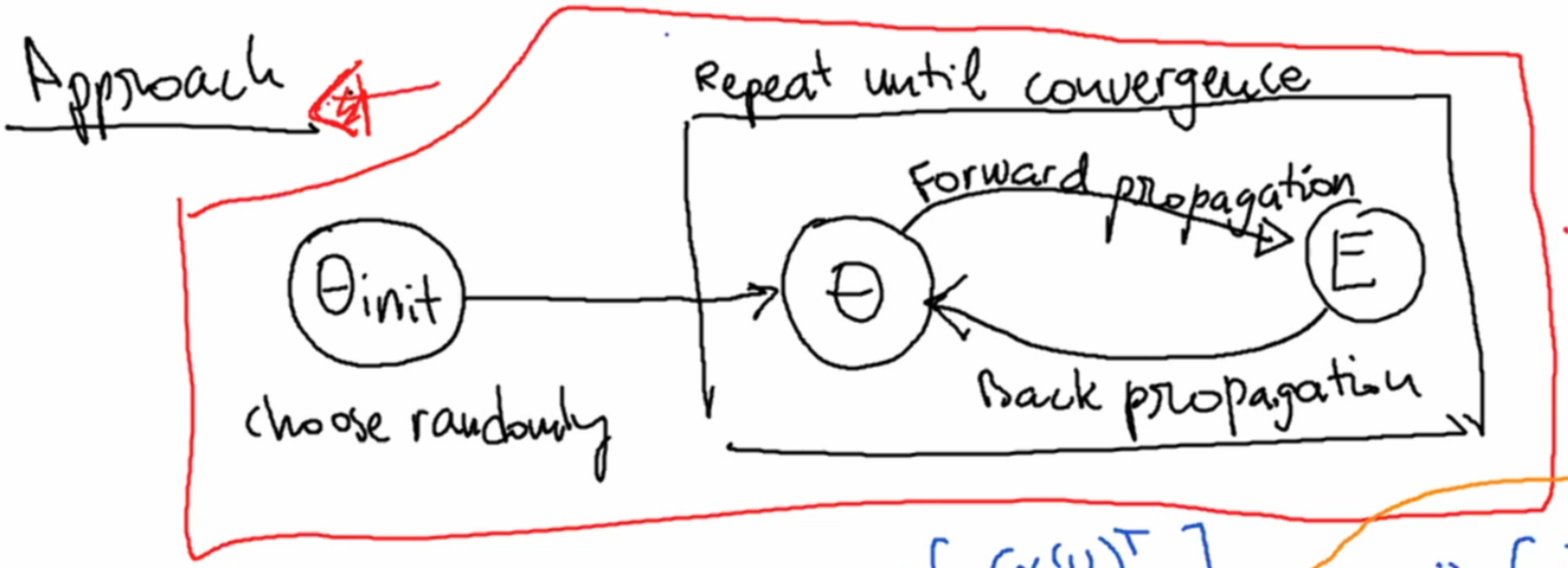
Def: Error

$$\underline{E} = \frac{1}{2} \sum_{i=1}^{I} \left(\| \hat{y}^{(i)} - y^{(i)} \|_2 \right)^2 \triangleq \text{Sum of squared errors (SSE)}$$

2-norm

Example: 2-norm $\quad \hat{y}^{(i)} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \; y^{(i)} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \; \| \hat{y}^{(i)} - y^{(i)} \|_2 = \sqrt{(1-3)^2 + (2-4)^2}$

$$= \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\hat{y}_j^{(i)} - y_j^{(i)} \right)^2$$



$\Rightarrow \Theta^* \Rightarrow \hat{Y}^*$ elementwise inversion

Ex. $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$; $A^{-1} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \end{bmatrix}$

Forward propagation :

$$X = \begin{bmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(i)})^T \\ \vdots \\ (x^{(I)})^T \end{bmatrix}_{I \times N}; \quad x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_N^{(i)} \end{bmatrix}; \quad (x^{(i)})^T = \begin{bmatrix} x_1^{(i)} & \cdots & x_N^{(i)} \end{bmatrix}$$

$$\overline{X} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} X^* = \begin{bmatrix} (\bar{x}^{(1)})^T \\ \vdots \\ (\bar{x}^{(I)})^T \end{bmatrix}_{I \times (N+1)}; \quad \overline{\overline{X}} = \underset{I \times (N+1)}{\overline{X}} \cdot \underset{(N+1) \times K}{V} = \begin{bmatrix} (\bar{x}^{(1)})^T V \\ \vdots \\ (\bar{x}^{(I)})^T V \end{bmatrix}_{I \times K}$$

sigmoid: $F = \frac{1}{1 + \exp(-\overline{\overline{X}})}$ ($I \times K$); $\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{I \times K}$

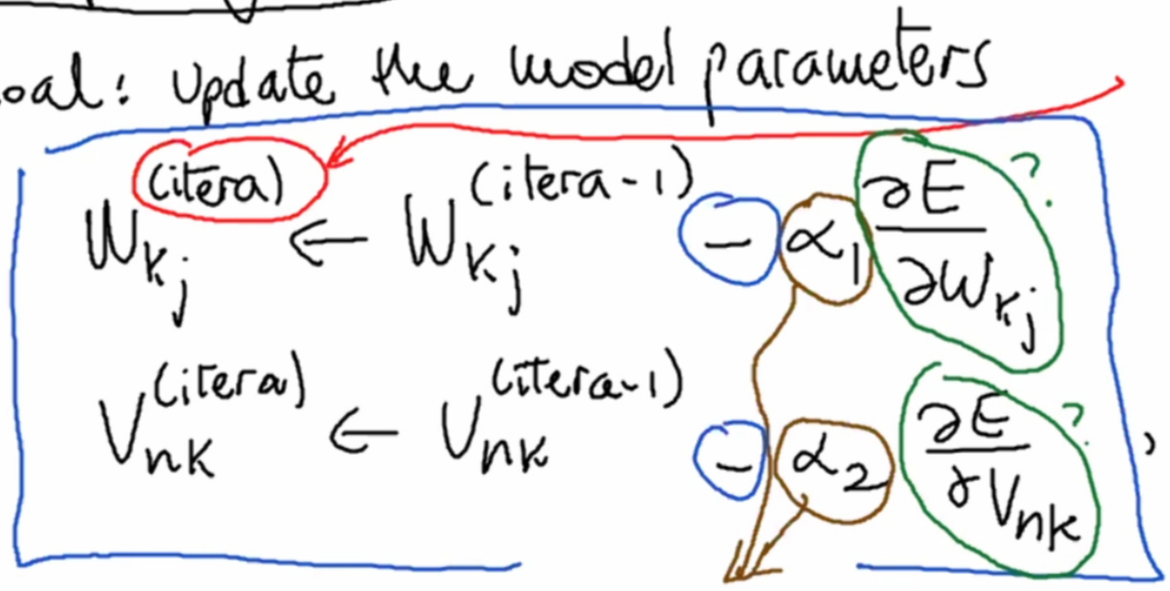
$$\overline{F} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} F \Big]_{I \times (K+1)} = \begin{bmatrix} (\bar{f}^{(1)})^T \\ \vdots \\ (\bar{f}^{(I)})^T \end{bmatrix}; \quad \overline{\overline{F}} = \underset{I \times (K+1)}{\overline{F}} \cdot \underset{(K+1) \times J}{W} = \begin{bmatrix} (\bar{f}^{(1)})^T W \\ \vdots \\ (\bar{f}^{(I)})^T W \end{bmatrix}_{I \times J}$$

$$G = \left(1 + \exp\left(-\overline{\overline{F}}\right)\right)^{-1}_{I \times J} = \hat{Y}$$

$$E = \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} (\hat{y}_j^{(i)} - y_j^{(i)})^2$$

Back propagation

Goal: Update the model parameters



$itera = \{1, 2, \cdots, Itera\_max\}$

$k = \{0, 1, \cdots K\}$, $j = \{1, 2, \cdots J\}$

$n = \{0, 1, \cdots, N\}$, $k = \{1, 2, \cdots K\}$

Batch Gradient Descent learning rate

"Group" (BGD) "0.01" or "0.001"

$$\frac{\partial E}{\partial w_{kj}} = \frac{\partial}{\partial w_{kj}} \frac{1}{2} \sum_{i=1}^{I} \sum_{j'=1}^{J} \left(\hat{y}_{j'}^{(i)} - y_{j'}^{(i)} \right)^2 = \frac{\partial}{\partial w_{kj}} \frac{1}{2} \sum_{i=1}^{I} \sum_{j'=1}^{J} \left(g_{j'}^{(i)} - y_{j'}^{(i)} \right)^2$$

$$= \sum_{i=1}^{I} \frac{1}{2} \frac{\partial}{\partial w_{kj}} \sum_{j'=1}^{J} \left(g_{j'}^{(i)} - y_{j'}^{(i)} \right)^2 = \sum_{i=1}^{I} \frac{1}{2} \cdot 2 \left(g_j^{(i)} - y_j^{(i)} \right) \frac{\partial g_j^{(i)}}{\partial w_{kj}} = \sum_{i=1}^{I} \left(g_j^{(i)} - y_j^{(i)} \right) g_j^{(i)} \left(1 - g_j^{(i)} \right) \bar{f}_k^{(i)}$$

$$\frac{\partial}{\partial w_{kj}} \left[\left(g_1^{(i)} - y_1^{(i)} \right)^2 + \cdots + \left(g_j^{(i)} - y_j^{(i)} \right)^2 + \cdots + \left(g_J^{(i)} - y_J^{(i)} \right)^2 \right]$$

$$\frac{\partial g_j^{(i)}}{\partial w_{kj}} = \frac{\partial}{\partial w_{kj}} \left[\left(1 + e^{-\left(\bar{f}^{(i)} \right)^T w_j} \right)^{-1} \right] = \frac{-e^{-\left(\bar{f}^{(i)} \right)^T w_j} \cdot \frac{\partial}{\partial w_{kj}} \left(-\left(\bar{f}^{(i)} \right)^T w_j \right)}{\left(1 + e^{-\left(\bar{f}^{(i)} \right)^T w_j} \right)^2} \qquad -\bar{f}_k^{(i)}$$

$$\left(\frac{1}{\gamma} \right)' = \frac{-\gamma'}{\gamma^2}$$

$$= \frac{1}{1 + e^{-\left(\bar{f}^{(i)} \right)^T w_j}} \left(1 - \frac{1}{1 + e^{-\left(\bar{f}^{(i)} \right)^T w_j}} \right) \bar{f}_k^{(i)} = g_j^{(i)} \left(1 - g_j^{(i)} \right) \bar{f}_k^{(i)}$$

$$\frac{\partial E}{\partial v_{nk}} = \frac{\partial}{\partial v_{nk}}\left[\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(g_j^{(i)} - y_j^{(i)}\right)^2\right] = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{1}{2}\frac{\partial}{\partial v_{nk}}\left(g_j^{(i)} - y_j^{(i)}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(g_j^{(i)} - y_j^{(i)}\right)\frac{\partial g_j^{(i)}}{\partial v_{nk}} \; ?$$

$$\frac{\partial g_j^{(i)}}{\partial v_{nk}} = \frac{\partial g_j^{(i)}}{\partial \bar{f}_k^{(i)}} \cdot \frac{\partial \bar{f}_k^{(i)}}{\partial v_{nk}}$$

(★) (★★)

$$= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(g_j^{(i)} - y_j^{(i)}\right) g_j^{(i)}\left(1 - g_j^{(i)}\right) w_{kj}\, f_k^{(i)}\left(1 - f_k^{(i)}\right)\bar{x}_n^{(i)}$$

(★) $$\frac{\partial g_j^{(i)}}{\partial \bar{f}_k^{(i)}} = \frac{\partial}{\partial \bar{f}_k^{(i)}}\left[\left(1 + e^{-\left(\bar{f}^{(i)}\right)^T w_j}\right)^{-1}\right]$$

$$= \frac{-e^{-\left(\bar{f}^{(i)}\right)^T w_j}\left(-w_{kj}\right)}{\left(1 + e^{-\left(\bar{f}^{(i)}\right)^T w_j}\right)^2}$$

$$= \frac{1}{1 + e^{-\left(\bar{f}^{(i)}\right)^T w_j}}\left(1 - \frac{1}{1 + e^{-\left(\bar{f}^{(i)}\right)^T w_j}}\right) w_{kj} = g_j^{(i)}\left(1 - g_j^{(i)}\right) w_{kj}$$

(★★) $$\frac{\partial \bar{f}_k^{(i)}}{\partial v_{nk}} = \frac{\partial f_k^{(i)}}{\partial v_{nk}} = \frac{\partial}{\partial v_{nk}}\left[\left(1 + e^{-\left(\bar{x}^{(i)}\right)^T v_k}\right)^{-1}\right]$$

$$= \frac{-e^{-\left(\bar{x}^{(i)}\right)^T v_k}\left(-\bar{x}_n^{(i)}\right)}{\left(1 + e^{-\left(\bar{x}^{(i)}\right)^T v_k}\right)^2}$$

$$= \frac{1}{1 + e^{-\left(\bar{x}^{(i)}\right)^T v_k}}\left(1 - \frac{1}{1 + e^{-\left(\bar{x}^{(i)}\right)^T v_k}}\right)\bar{x}_n^{(i)} = f_k^{(i)}\left(1 - f_k^{(i)}\right)\bar{x}_n^{(i)}$$

Pseudo-code

$\alpha_1 = \alpha_2 = 0.001$

Initialize V, W randomly

for iters $= \{1, 2, \cdots, \text{Iters\_max}\}$:

FWP: $X, \overline{X}, \overline{\overline{X}}, F, \overline{F}, \overline{\overline{F}}, G, E$

BP:

$$W_{kj}^{(\text{iters}+1)} \leftarrow W_{kj}^{(\text{iters})} - \alpha_1 \frac{\partial E}{\partial W_{kj}} \qquad k = \{0, 1, \cdots, K\},\ j = \{1, 2, \cdots J\}$$

$$V_{nk}^{(\text{iters}+1)} \leftarrow V_{nk}^{(\text{iters})} - \alpha_2 \frac{\partial E}{\partial V_{nk}} \qquad n = \{0, 1, \cdots N\},\ k = \{1, 2, \cdots K\}$$

}

$\Rightarrow V^*, W^*$.