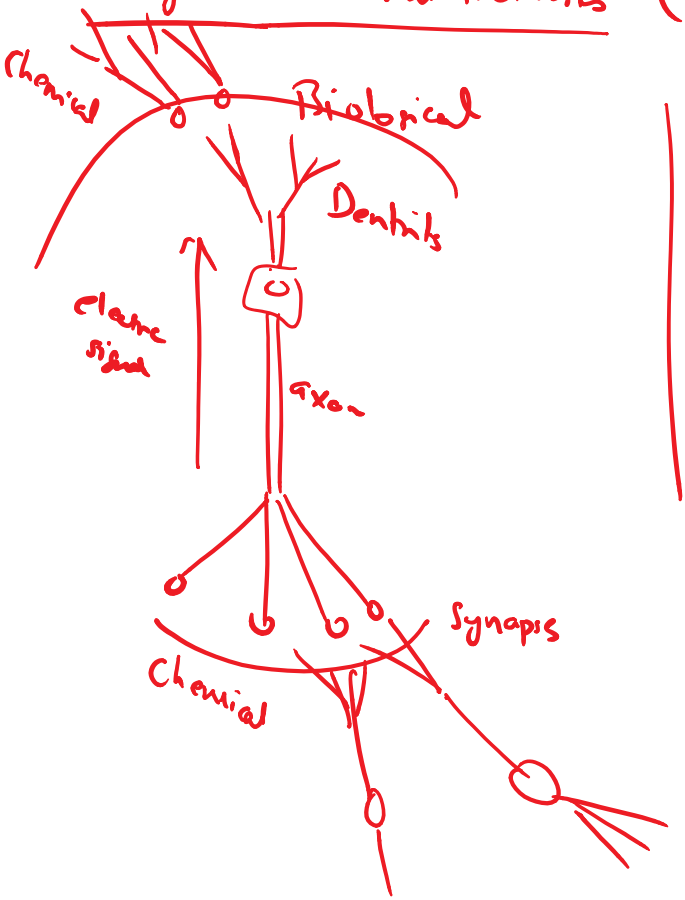
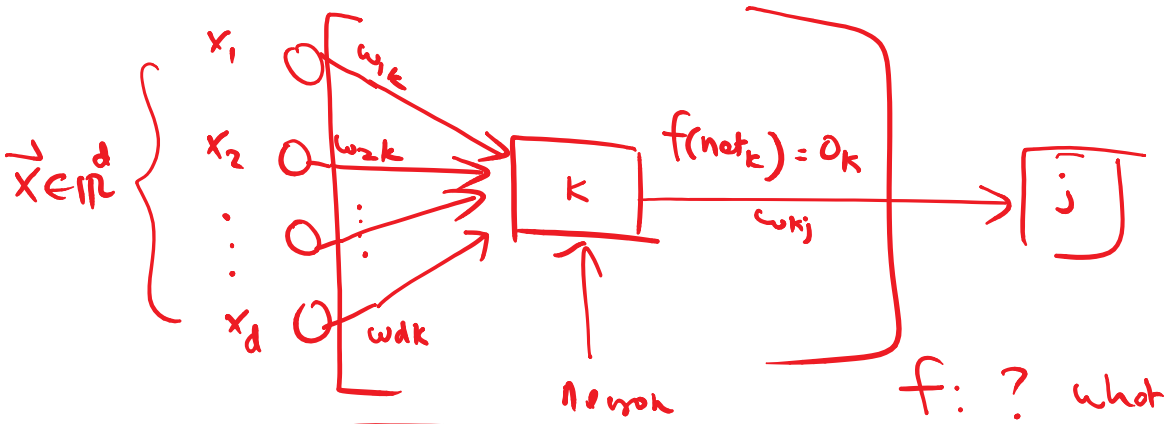
Artificial Neural Networks (ANN)



① aggregation of inputs from synapses
② thresholding
③ spiking ←
④ propagate to next set of connected neurons via dendrites

Human Brain: 10 trillion connections



$$net_k = \sum_{i=1}^{d} w_{ik} x_i$$

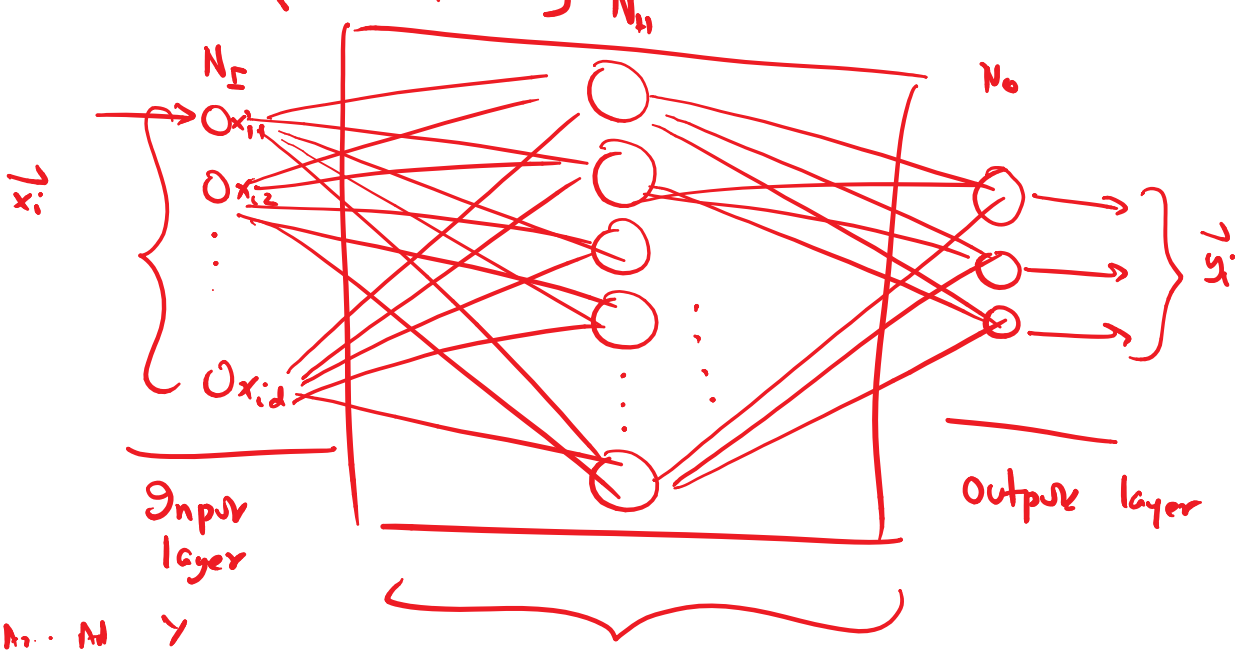
Linear Combination

f: ? what kind of functions should we allow.

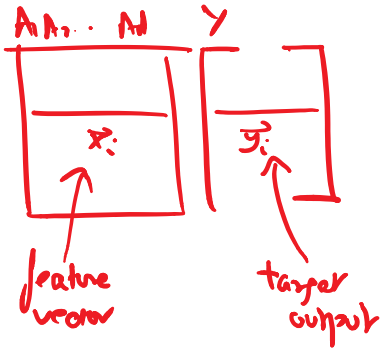
Activation function

topology of the network

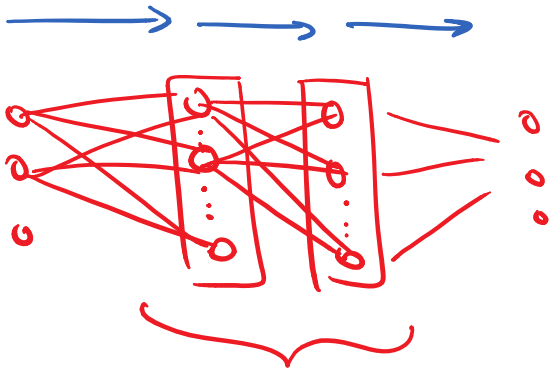
topology of the network
↘ set of connections
] architecture



hidden layer(s)



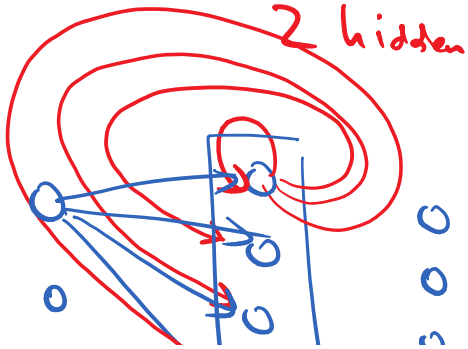
2 layer network



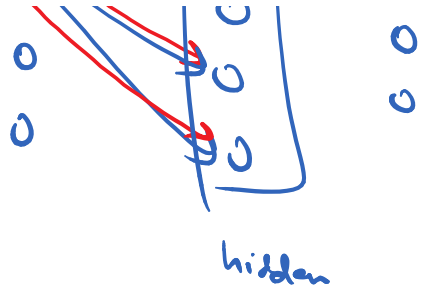
Feed Forward Networks

2 hidden layers

MLP: Multilayered Perceptrons



backward flow of info



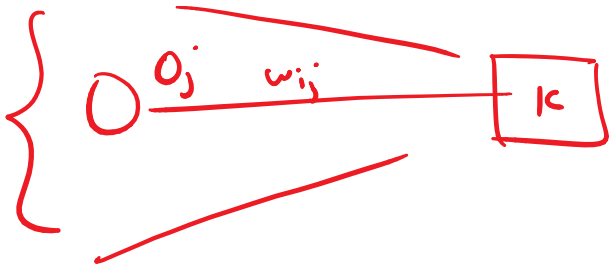
info

feed-back

Recurrent NN

whenever "history" is important

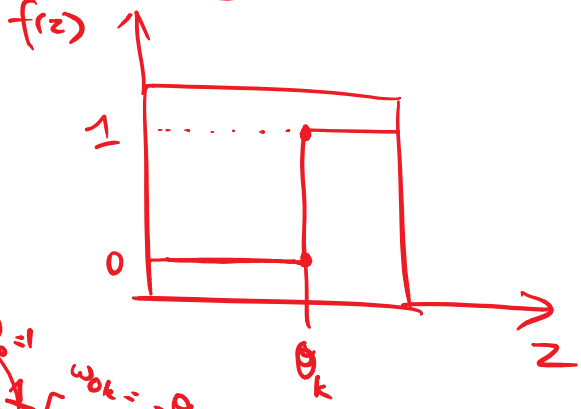
$$net_k = \sum_{i=1}^{\text{inputs}} w_{ik} o_i = \vec{w}^T \vec{o}$$



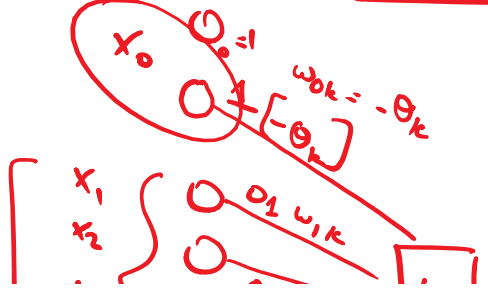
$f(net_k)$? ← different activation functions

$z = net_k$

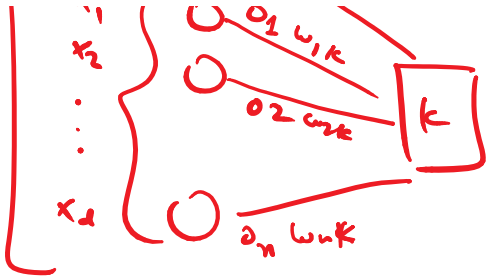
$$f(z) = \begin{cases} 1 & \text{if } z \geq \theta \\ 0 & \text{otherwise} \end{cases}$$ ← threshold



binary neuron with threshold θ



$z = net_k = \sum w_{ik} o_i > \theta_k$



$$z = net_k = \sum w_{ik} o_i \geq \theta_k$$

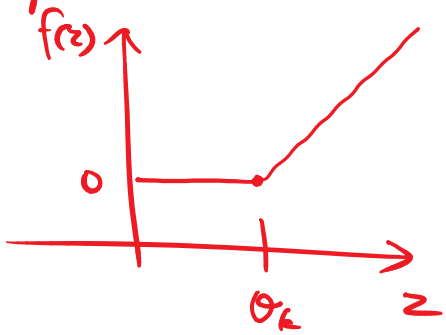
$$z_k \quad \sum_{i=1}^{n} w_{ik} o_i - \theta_k \geq 0$$

$$f(z_k) = \begin{cases} 1 & \text{if } z_k \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{i=0}^{n} w_{ik} o_i$$

---

(2) rectified linear neuron

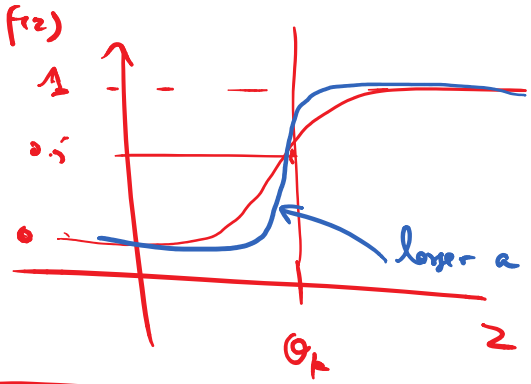


$$f(z_k) = \begin{cases} z_k & \text{if } z \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

assuming that θ_k is a bias neuron

$$net_k = z_k = \sum_{i=0}^{n} \_\_\_\_$$

(3) Sigmoid (logistic)



$$f(z) = \frac{1}{1+e^{-z}}$$

$$\frac{1}{1+e^{-az}}$$

$a > 0$

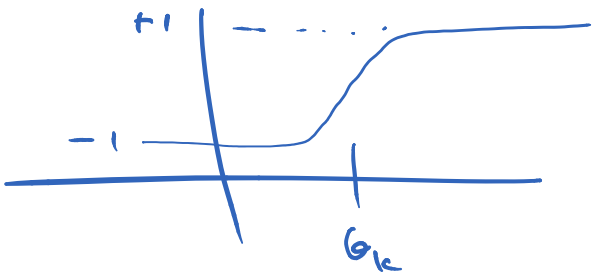
$1+e^{-z}$

default

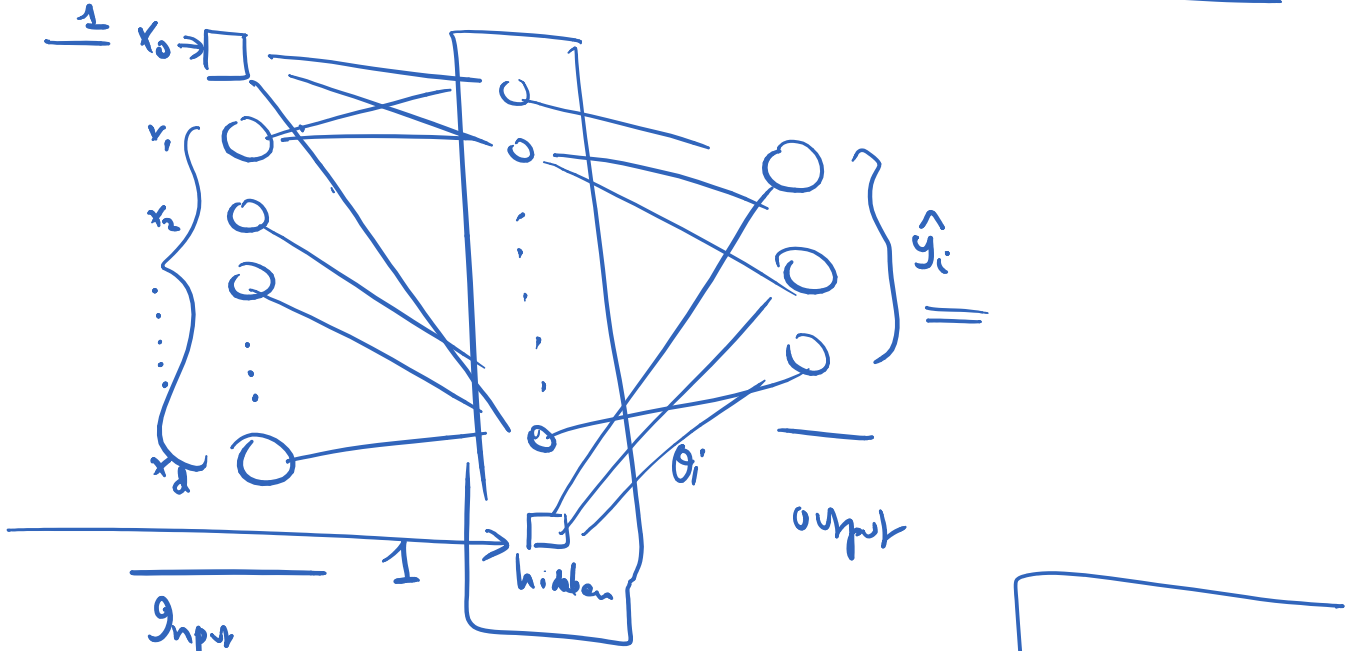
$a>0$

slope parameter

(4) Sigmoid (tanh)



$$f(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$



$\vec{x}_i \rightarrow$ feed forward through the network

$$|w_{ij}| = N_I \times N_H + N_H \times N_o$$

total # of weights/edges in ANN

$N_I = 100$

$N_H = 10\,000$

$$E = \sum^{n} \|\hat{y}_i - y_i\|^2$$

Initialize all the weights

$w_{ij} \neq 0$

small

-0.5 to $+0.5$

uniform

$N(0, \sigma^2)$

$N_H = 10{,}000$

$N_O = 10$

$$E = \underbrace{\sum_{i=1}^{n} \|\hat{y}_i - y_i\|^2}_{\text{Squared error function}}$$

$N(0, \sigma^2)$ — small

given training input $\vec{x}_i$, predict $\hat{y}_i$, compare to true target $\vec{y}_i$

$$\min_W (E) = \min_W \sum_{i=1}^{n} \|\hat{y}_i - y_i\|^2$$

all the weights

$W = \{w_{ij}\}$ all pairs

activation functions are non-linear, e.g. sigmoid

non-linear non-convex objective

:(bad news: no guarantee of the solution

Stochastic gradient descent "back propagation"

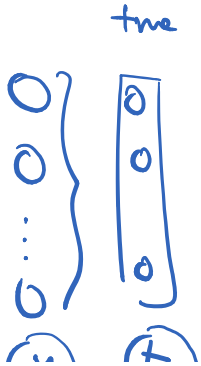
:)

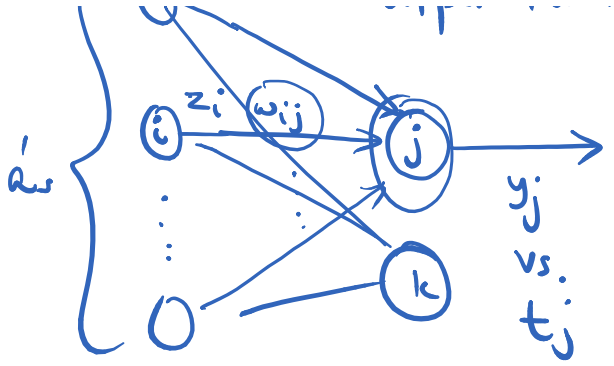
↓

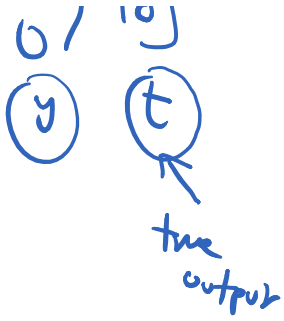
local minima

$$E = \frac{1}{2} \sum_i (t_i - y_i)^2$$

t_i: true; y_i: output of ANN







hidden

$$net_j = \sum_{a=1}^{N_{H1}} w_{aj} z_a$$

$$\nabla w_{ij} = \frac{\partial E}{\partial w_{ij}}$$

$$E = \frac{1}{2} \sum_i (t_i - y_i)^2$$

$$w_{ij} = w_{ij} - \eta \nabla w_{ij}$$

$$= \frac{\partial E}{\partial y_j} \cdot \frac{\partial y_j}{\partial net_j} \cdot \frac{\partial net_j}{\partial w_{ij}}$$

$$E = \frac{1}{2} \sum_i (t_i - f(net_j))$$

y_j

$$\frac{\partial}{\partial y_j} E = \frac{1}{2}(t_j - y_j)^2 + \cdot \text{Constr}$$

$$= \frac{1}{2} \cdot 2(t_j - y_j)(-1)$$

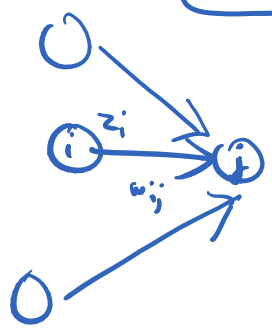
$$= -(t_j - y_j) f'(net_j)$$

$$\frac{\partial y_j}{\partial net_j} = \frac{\partial f(net_j)}{\partial net_j}$$

$$= f'(net_j)$$

z_i

$$\delta_j = -\frac{\partial E}{\partial net_j}$$



$$\frac{\partial net_j}{\partial w_{ij}}$$

$$net_j = \sum_{a=1}^{N_H} w_{aj} \cdot z_a$$

$$\frac{\partial net_j}{\partial w_{ij}} = \frac{\partial w_{ij} \cdot z_i + Const}{\partial w_{ij}}$$

Hidden to output

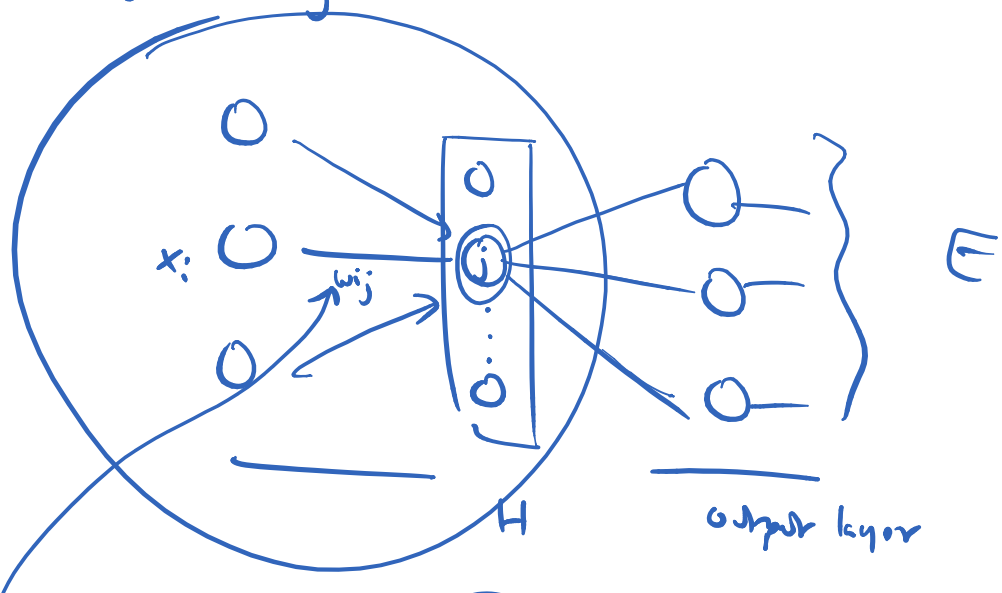
i j

$$\nabla w_{ij} = \frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial net_j} \cdot \frac{\partial net_j}{\partial w_{ij}} = \delta_j \cdot z_i$$

$$\delta_j = -(t_j - y_j) \cdot f'(net_j)$$

Input to hidden

i j



$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial net_j} \cdot \frac{\partial net_j}{\partial w_{ij}}$$

δ_j $\quad$ x_i

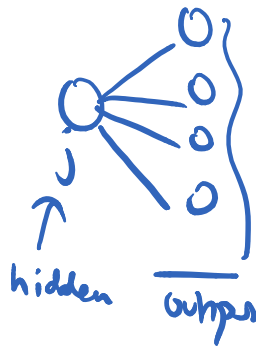
$$net_j = \sum_{a=1}^{N_I} x_{ai} w_{aj}$$

$$\frac{\partial net_j}{\partial w_{ij}} = x_i \cdot w_{ij} + Const$$

$$= x_i$$

$$\delta_j = -\frac{\partial E}{\partial net_j}$$

$$\sum_{k=1}^{N_O} \frac{\partial E}{\partial y_k} \cdot \frac{\partial y_k}{\partial net_k} \cdot \frac{\partial z_j}{\partial}$$

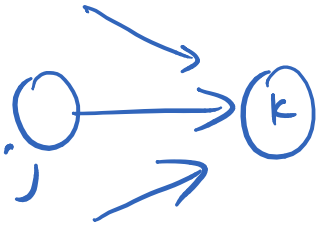


$$\frac{\partial E}{\partial net_j} = \sum_{k=1}^{N_o} \frac{\partial E}{\partial y_k} \cdot \frac{\partial y_k}{\partial net_k} \cdot \frac{\partial net_k}{\partial z_j} \cdot \frac{\partial z_j}{\partial net_j}$$

output k — net_k — output j — net_j

$$= \sum_{k=1}^{N_o} \frac{\partial E}{\partial net_k} \cdot w_{jk} \cdot f'(net_j)$$

$-\delta_k$



$$net_k = \sum_{a=1}^{N_h} w_{ak} \cdot z_a$$

$$\frac{\partial}{\partial z_j} = \frac{\partial}{\partial z_j} w_{jk} \cdot z_j = w_{jk}$$

$$\delta_j = -\sum_{k=1}^{N_o} \delta_k w_{jk} \cdot f'(net_j)$$

Summary $H \to O$

HO: $\nabla w_{ij} = \delta_j z_i = -(t_j - y_j) f'(net_j) z_i$

$I \to H$

IH: $\nabla w_{ij} = \delta_j x_i = -\left(\sum_{k=1}^{N_o} \delta_k w_{jk}\right) f'(net_j) x_i$

$f'(net_j)$ or $f'(z)$

If f is sigmoid (logistic)

$$f(z) = \frac{1}{1+e^{-z}}$$

$$f(z) = \frac{1}{1+e^{-z}}$$

$$\frac{d f(z)}{dz} = \frac{0 - (1)(e^{-z})(-1)}{(1+e^{-z})^2}$$

$$= \left(\frac{e^{-z}}{(1+e^{-z})}\right)\left(\frac{1}{1+e^{-z}}\right)$$

$1 - f(z)$ $\quad$ $f(z)$

$$f'(net_j) = f'(z) = f(z)(1-f(z))$$

SGD

epochs: 1 run through all points

for $k = 1$ to n in random order ← SGD

$\vec{x}_k$ ← Chosen point

y_k, $\hat{y}_k$

① feed forward

② back prop

for all i & j in (H, O)

$w_{ij} = w_{ij} - \eta \nabla w_{ij}$

for all i & j in (I, H)

$w_{ij} = w_{ij} - \eta \nabla w_{ij}$

η : step size