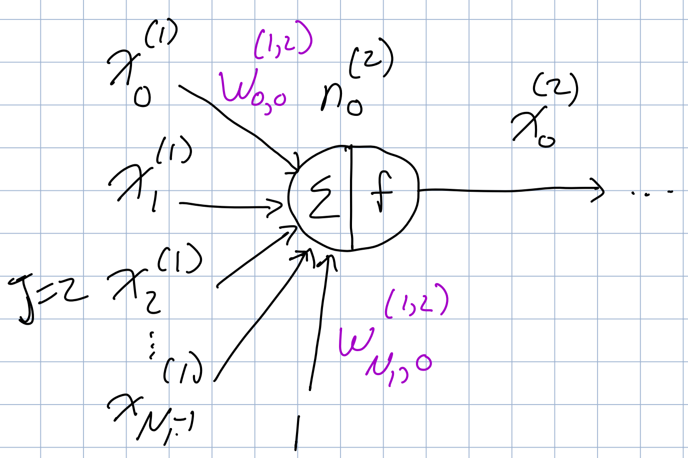
Layer 1 Layer 2



$$n_j^{(2)} = \sum_{i=0}^{N_1} w_{i,j}^{(1,2)} x_i^{(1)}$$

L L−1

$$= W^{(1,2)^T} X^{(1)}$$

$$\begin{bmatrix} w_{0,0}^{(1,2)} \\ w_{1,0}^{(1,2)} \\ w_{2,0}^{(1,2)} \\ \vdots \\ w_{N_1}^{(1,2)} \end{bmatrix} \qquad \begin{bmatrix} x_0^{(1)} \\ x_1^{(1)} \\ x_2^{(1)} \\ \vdots \\ x_{N-1}^{(1)} \\ 1 \end{bmatrix}$$

Training Set.

$X_p^{(1)}$ y_p

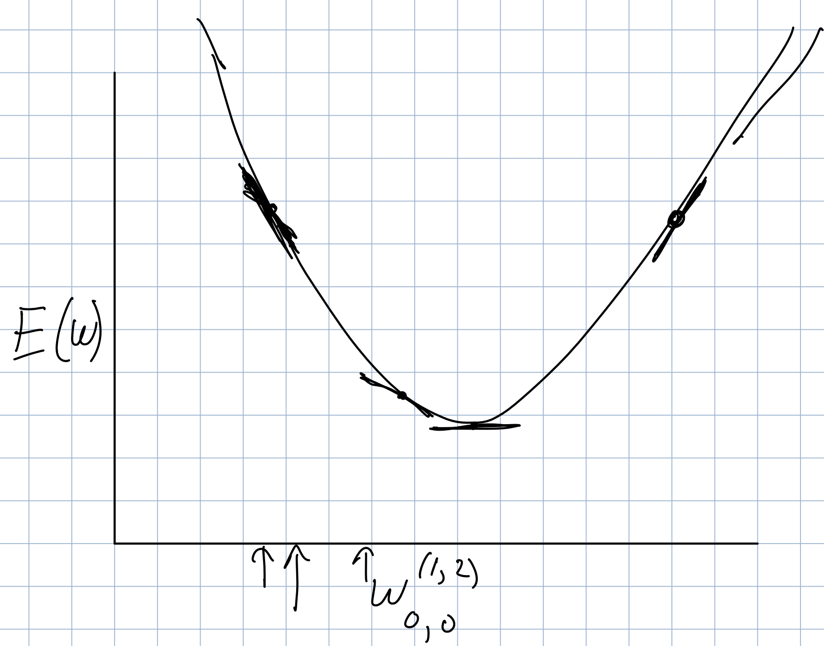
p = pattern

$$x_j^{(2)} = f(n_j^{(2)})$$

$$E_p = \frac{1}{2}\left(y_{0,p} - x_{0,p}^{(2)}\right)^2$$

SSE $E = \sum_{p \in P} E_p$

MSE $E = \frac{1}{|P|} \sum E_p$



$$\frac{dE}{dw_{0,0}^{(1,2)}}$$

Guess w

While not happy:

Compute $\frac{dE}{dw}$ for every w

$w \leftarrow w - \alpha \frac{dE}{dw}$

α : "learning rate"

$$\frac{dE}{dw_{J,0}} = \frac{d}{dw_{J,0}} E$$

Assume that f is linear

$$\underline{x_0^{(2)} = n_0^{(2)}}$$

$$= \frac{d}{dw_{J,0}} \sum_p E_p$$

$$= \sum_p \underbrace{\frac{d}{dw_{J,0}} E_p}$$

$$\frac{d}{dc}\frac{1}{2}(a-c)^2 = (a-c)\frac{d}{dc}(a-c)$$

$$\frac{d}{dw_{J,0}^{(1,2)}} E_p = \frac{d}{dw_{J,0}^{(1,2)}} \frac{1}{2}\left(\underline{y_{0,p} - x_{0,p}^{(2)}}\right)^2$$

$$= \left(y_{0,p} - x_{0,p}^{(2)}\right)\frac{d}{dw_{J,0}^{(1,2)}}\left(y_{0,p} - x_{0,p}^{(2)}\right)$$

$$= \quad '' \quad \left(\underbrace{\cancel{\frac{d}{dw_{J,0}^{(1,2)}} y_{0,p}}}_{\to 0} - \underbrace{\frac{d}{dw_{J,0}^{(1,2)}} x_{0,p}^{(2)}}\right)$$

$$= \left(y_{0,p} - x_{0,p}^{(2)}\right)(-1)\underbrace{\left(\frac{d}{dw_{J,0}^{(1,2)}} \sum_{j=0}^{N_1} w_{j,0}^{(1,2)} x_{j,p}\right)}$$

$$= \sum_{j=0}^{N_1} \underbrace{\frac{d}{dw_{J,0}^{(1,2)}} w_{j,0}^{(1,2)}} x_{j,p}$$

$$\frac{d}{dw_{J,0}} w_{0,0} x_{0,p} + \frac{d}{dw_{J,0}} w_{1,0} x_{1,p} + \dots + \frac{d}{dw_{J,0}} w_{J,0} x_{J,p} + \dots$$

$$= x_{j,p} + 0 + 0$$

$$= (y_{o,p} - x_{o,p}^{(2)})(-1)\, x_{j,p}^{(2)}$$

Assume $f()$ is interesting

$$\frac{d}{dc} f(a(c)) = f'(a(c)) \times \frac{d}{dc} a(c)$$

$$\frac{d}{dw_{j,o}} E_p = (y_{o,p} - x_{o,p}^{(2)}) \frac{d}{dw_{j,o}^{(1,2)}} \left(y_{o,p} - x_{o,p}^{(2)}\right)$$

$$(y_{o,p} - f(n_{o,p}^{(2)}))$$

$$= (\quad \ldots \quad) \left[\cancel{\frac{d}{dw_{j,o}^{(1,2)}} y_{o,p}} - \frac{d}{dw_{j,o}^{(1,2)}} f(n_{o,p}^{(2)}) \right]$$

0

$$f'(n_{o,p}^{(2)}) \underbrace{\frac{d}{dw_{j,o}^{(1,2)}} n_{o,p}^{(2)}}$$

$$f'(n_{o,p}^{(2)}) \frac{d}{dw} \Sigma$$

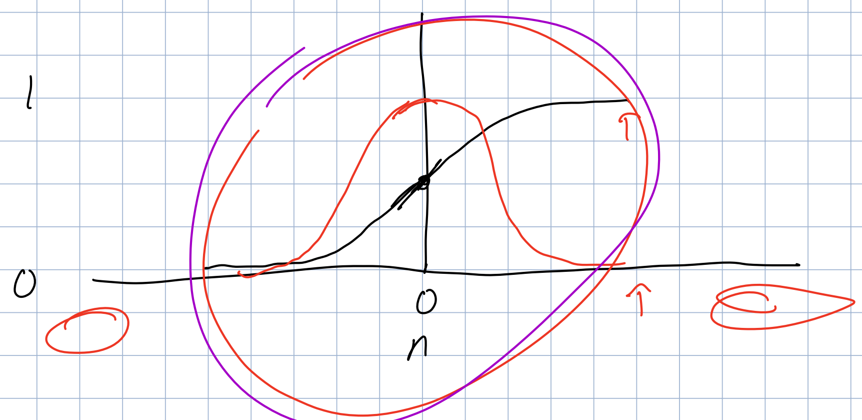
$$f'(n_{o,p}^{(2)}) \underbrace{\sum_{j=0}^{N_1} \frac{d}{dw_{j,o}^{(1,2)}} w_{j,o} x_{j,p}}$$

$$\boxed{\frac{dE_p}{dw_{j,o}}} = -(y_{o,p} - x_{o,p}^{(2)})\, f'(n_{o,p}^{(2)})\, x_{j,p}^{(0)}$$

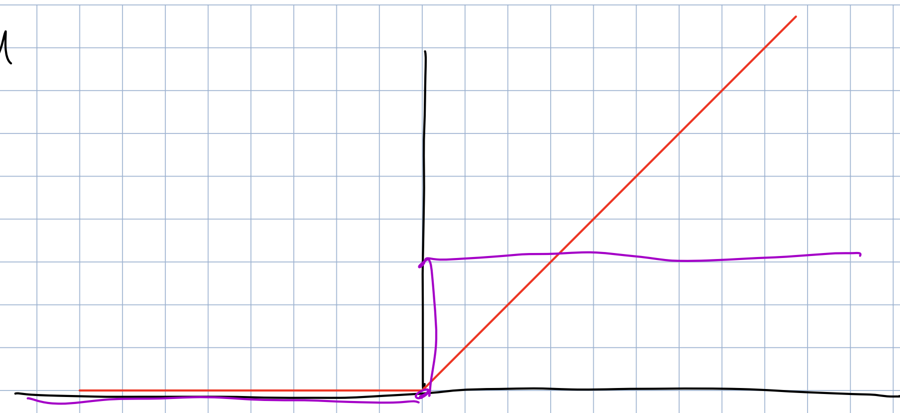
$$f(n) = \text{sigmoid}(n) = \frac{1}{1-e^{-n}}$$

$$\frac{dE}{dw_{j,0}}$$

$\uparrow$



RELU

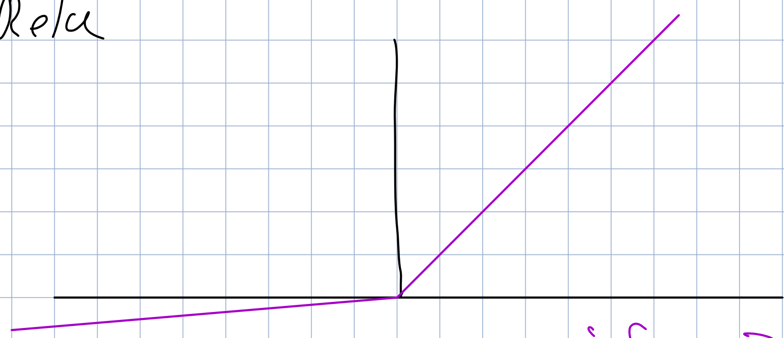


$n=0$

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

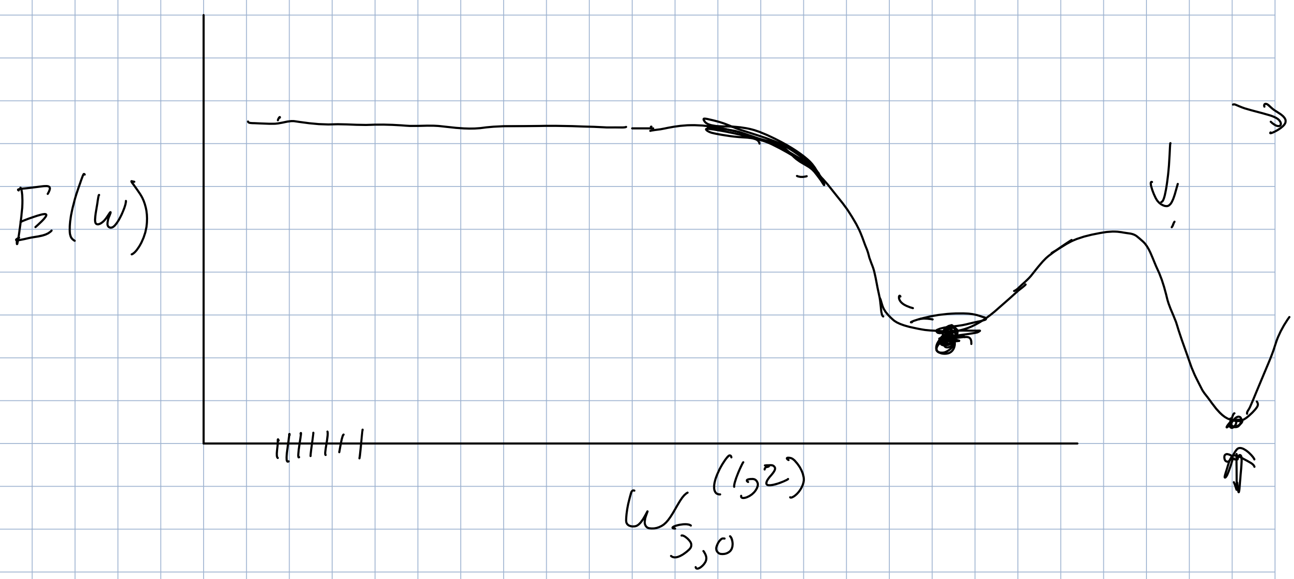
$$f'(n) = \begin{cases} 1 & \text{if } n > 0 \\ \cancel{1} & n = 0 \\ 0 & \text{if } n < 0 \end{cases}$$

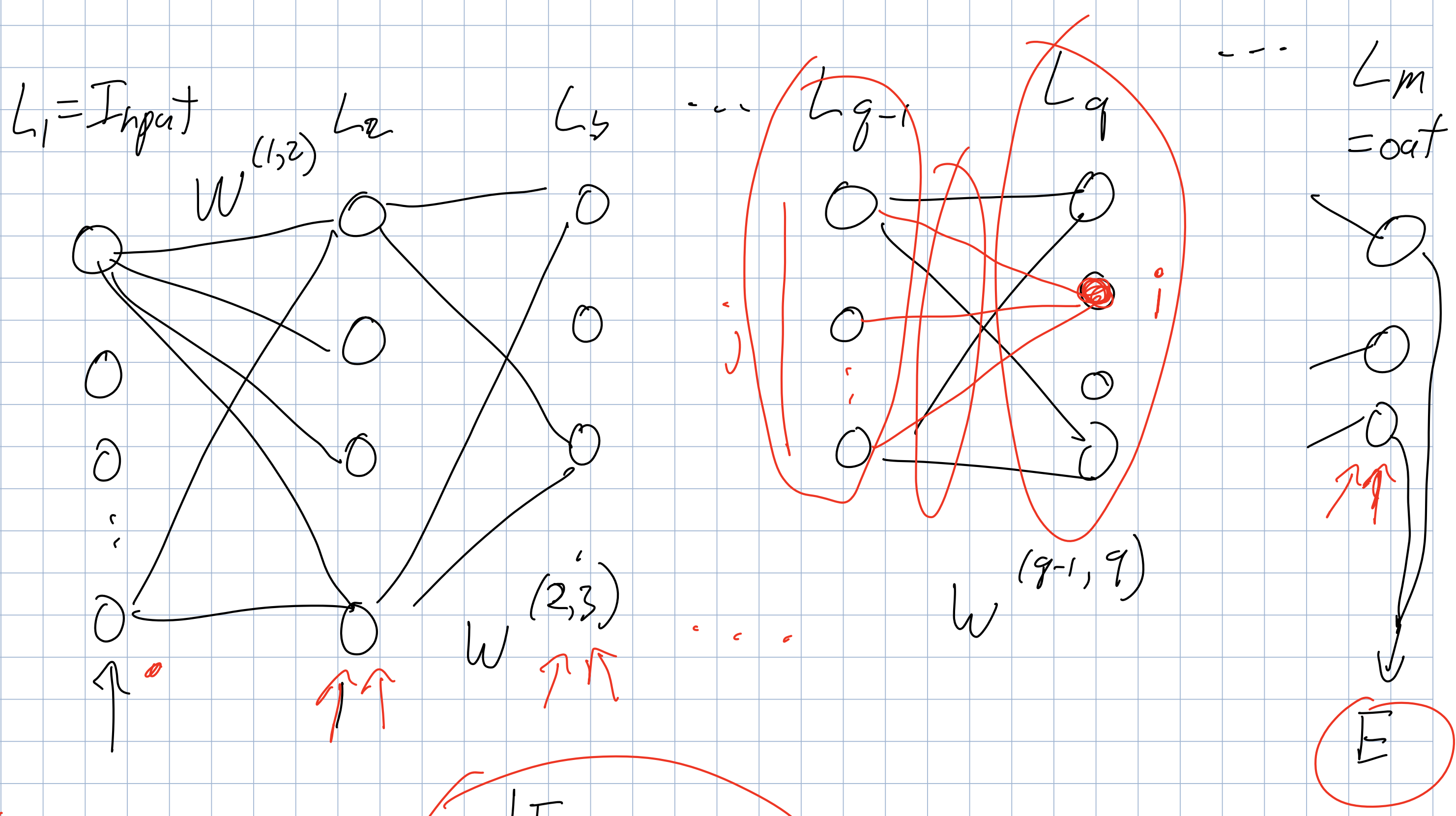
Leaky Relu



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

$$0 \leq \beta << 1$$





$$\frac{\partial E}{\partial x_{j,p}^{(q)}}$$

$$\frac{\partial E}{\partial w_{i,j}^{(q-1,q)}} \; ?$$

$$n_{i,p}^{(q)} = \sum_{j=0}^{N_{q-1}} w_{j,i}^{(q-1,q)} \times x_{j,p}^{(q-1)}$$

(net input for unit i, pattern p, layer q)

$$x_{i,p}^{(q)} = f\left(n_{i,p}^{(q)}\right)$$