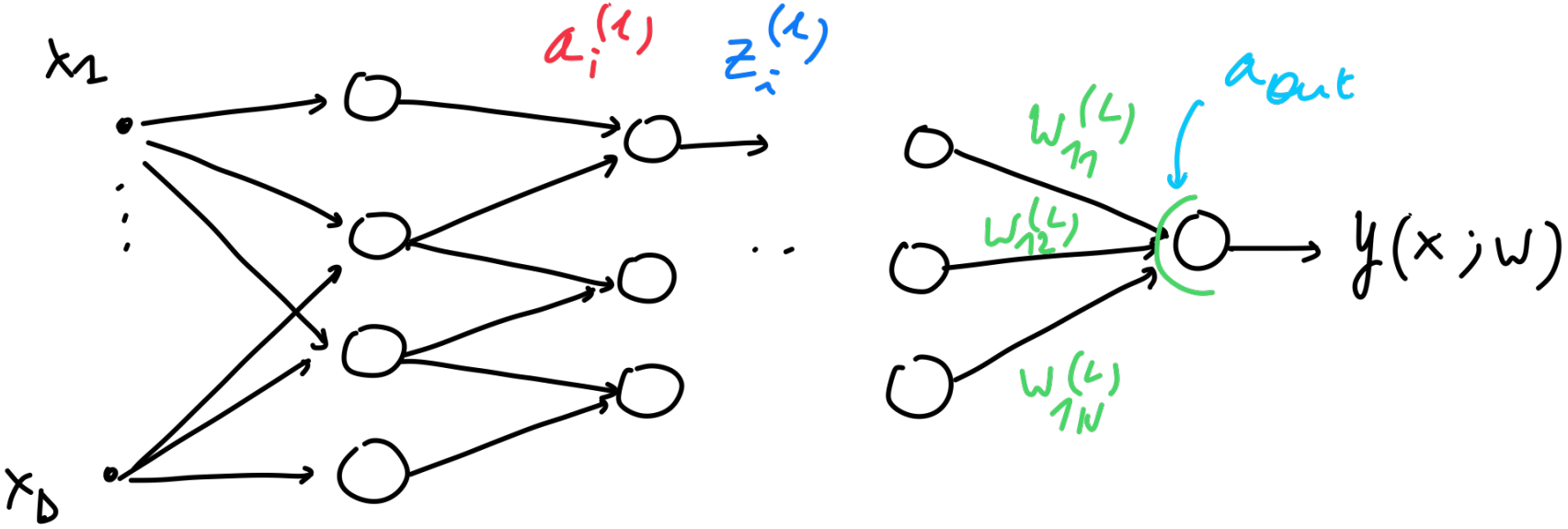


$a_i^{(l)}$ = preactivation of unit i in layer l

$$a_i^{(l)} = \sum_{j=1}^{N_{l-1}} w_{ij}^{(l)} z_j^{(l-1)} + w_{i0}^{(l)} \quad \longleftarrow \quad \frac{\partial a_i^{(l)}}{\partial w_{ij}^{(l)}} = z_j^{(l-1)}$$

$z_i^{(l)}$ = postactivation for unit i in layer l

$$z_i^{(l)} = \sigma(a_i^{(l)}) = \sigma\left(\sum_{j=1}^{N_{l-1}} w_{ij}^{(l)} z_j^{(l-1)} + w_{i0}^{(l)} \right)$$

$$p(t=0|x) = y(x;w)$$

$$p(t=1|x) = 1 - y(x;w)$$

$$p(\{t(x^{(i)}) = t^{(i)}\}_{i=1}^{N} | x) = \prod_{i=1}^{N} p(t(x^{(i)}) = t^{(i)} | x)$$

$$= \prod_{i=1}^{N} y(x^{(i)};w)^{1-t^{(i)}} \left(1 - y(x^{(i)};w)\right)^{t^{(i)}}$$

→ taking the log

$$L(w) = -(1-t^{(i)}) \log y(x^{(i)};w) - t^{(i)} \log\left(1 - y(x^{(i)};w)\right)$$

1] We want $\frac{\partial L}{\partial w_{ij}^{(l)}} = \frac{\partial L}{\partial a_i^{(l)}} \cdot \frac{\partial a_i^{(l)}}{\partial w_{ij}^{(l)}}$

$\delta_i^{(l)}$ $z_j^{(l-1)}$

$L \leftarrow a_i^{(l)}, \ldots$; $a_i^{(l)} \leftarrow w_{ij}^{(l)}$

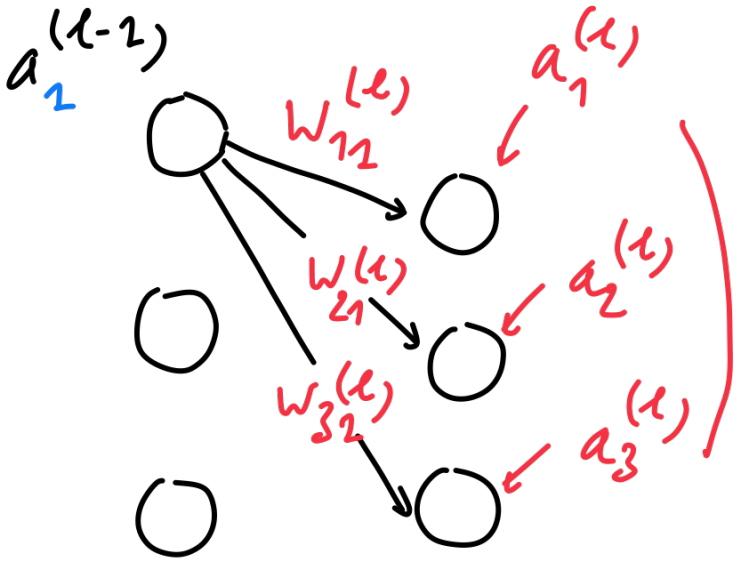
2] $\delta_{out} = \frac{\partial L}{\partial a_{out}} \rightarrow \frac{\partial \left(-(1-t^{(i)}) \log \sigma(a_{out}) - t^{(i)} \log(1-\sigma(a_{out}))\right)}{\partial a_{out}}$

$= -(1-t^{(i)}) \frac{\sigma'}{\sigma(a_{out})} + t^{(i)} \frac{\sigma'}{1-\sigma(a_{out})}$

$\sigma'(a) = \sigma(a)(1-\sigma(a))$

$= -(1-t^{(i)})(1-\sigma(a_{out})) + t^{(i)} \sigma(a_{out})$

$\delta_{out} = \sigma(a_{out}) - (1-t^{(i)})$



$$L\left(a_1^{(l)}\ a_2^{(l)}\ a_3^{(l)}\right)$$

$$\frac{\partial L}{\partial a_1^{(l-2)}} \leftarrow L\left(a_1^{(l)}(a_1^{(l-2)})\ a_2^{(l)}(a_1^{(l-2)})\right)$$

$$\frac{\partial L}{\partial a_i^{(l-2)}} = \sum_{j=1}^{N_L} \frac{\partial L}{\partial a_j^{(l)}} \cdot \frac{\partial a_j^{(l)}}{\partial a_i^{(l-2)}}$$

$$\delta_i^{(l-2)} = \sum_{j=1}^{N_L} \delta_j^{(l)}$$

$$\frac{\partial a_j^{(l)}}{\partial a_i^{(l-2)}} \rightarrow a_j^{(l)} = \sum_{i=1}^{N_{L-2}} w_{ji}^{(l)} \cdot \sigma\left(a_i^{(l-2)}\right) + w_{j0}^{(l)}$$

$$\frac{\partial a_j^{(l)}}{\partial a_i^{(l-2)}} = w_{ji}^{(l)}\, \sigma'\left(a_i^{(l-2)}\right)$$

Backpropagation

1) → Forward propagate $x^{(i)}$ through the network, and get all $a_i^{(l)}$, $z_i^{(l)}$

2) you get $\delta_{out} = \frac{\partial L}{\partial a_{out}} = \sigma(a_{out}) - (1 - t^{(i)})$

3) Backpropagate the δ_{out}

$$\delta_i^{(l-2)} = \sum_{j=1}^{N_L} \delta_j^{(l)} W_{ji}^{(l)} \cdot \sigma'(a_i^{(l-2)})$$

4) Get the gradient as

$$\frac{\partial L}{\partial w_{ij}^{(l)}} = \delta_i^{(l)} \cdot z_j^{(l-1)}$$

$$\frac{\partial L}{\partial w_{i0}^{(l)}} = \delta_i^{(l)}$$