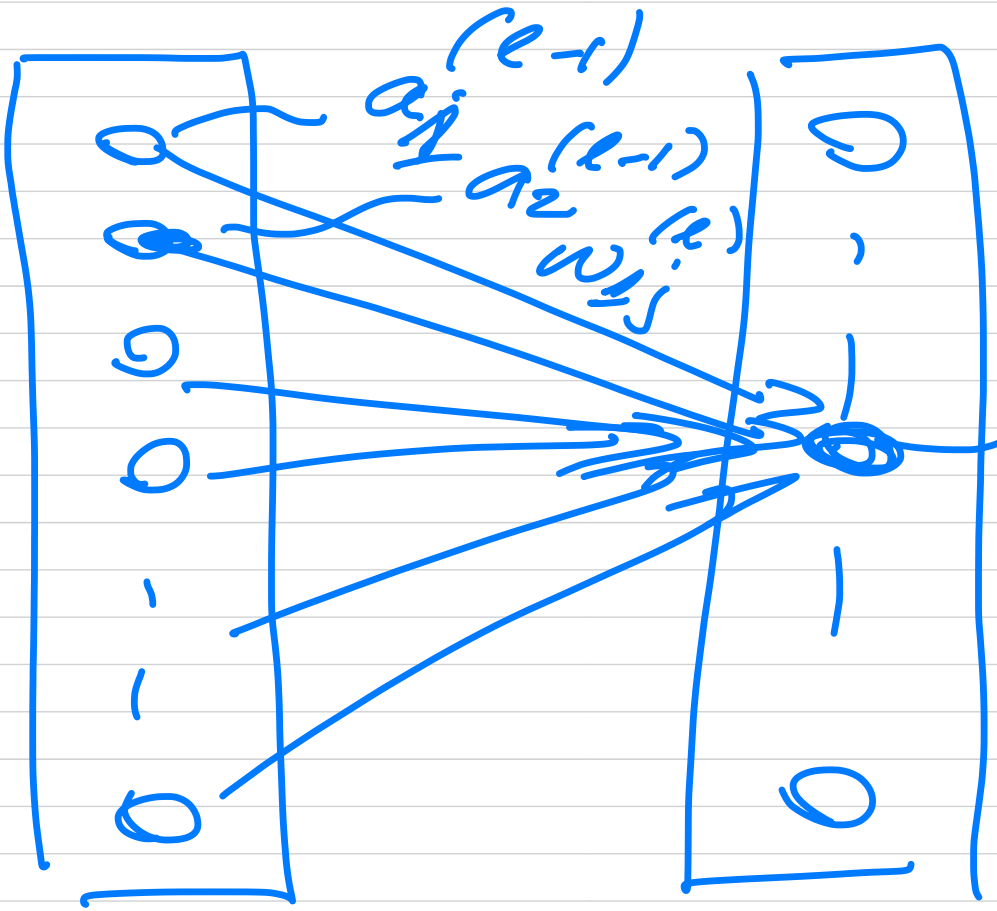
FYS-STK3155/4155, lecture October 13, 2025

FYS-STK3155/4155 October 13



$$z_j^{(l)} = \sum_{i=1}^{m_{l-1}} w_{ij}^{(l)} a_i^{(l-1)} + b_j^{(l)}$$

$$a_j^{(l)} = \sigma(z_j^{(l)})$$

$$l = L$$

$$a_j^{(L)} \rightsquigarrow C(a_j^{(L)}, y_j, \theta)$$

MSE

$$C(\theta) = \frac{1}{2} \sum_{i=1}^{n} (a_j^{(L)} - y_j)^2$$

$$\frac{\partial C}{\partial \theta} = 0$$

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

$$\frac{\partial z_j^{(\ell)}}{\partial a_i^{(\ell-1)}} = w_{ij}^{(\ell)}$$

$$\frac{\partial a_j^{(\ell)}}{\partial z_j^{(\ell)}} = a_j^{(\ell)}\left(1 - a_j^{(\ell)}\right)$$

↑ sigmoid function

start with L (final)

$$C(\theta) = \frac{1}{2} \sum_i \left(a_i^{(L)} - y_i\right)^2$$

$$\frac{\partial C}{\partial w_{jk}^{(L)}} = \underbrace{\left(a_j^{(L)} - y_j\right)}_{\text{From } C(\theta)} \leftarrow \frac{\partial C}{\partial a_j^{(L)}}$$

$$\times \underbrace{\left[a_j^{(L)}\left(1 - a_j^{(L)}\right)\right]}_{\sigma\left(z_j^{(L)}\right)} a_k^{(L-1)}$$

σ'

$$\delta_j^{(l)} = \frac{\partial C}{\partial a_j^{(l)}} \cdot \sigma'(z_j^{(l)}) \qquad \frac{\partial a_j^{(l)}}{\partial z_j^{(l)}}$$

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)} a_k^{(l-1)}$$

$$\frac{\partial C}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

$$w_{jk}^{(l)} \leftarrow w_{jk}^{(l)} - \eta \delta_j^{(l)} a_k^{(l-1)}$$

$$b_j^{(l)} \leftarrow b_j^{(l)} - \eta \delta_j^{(l)}$$

$$\Rightarrow \quad \delta_j^{(l)} = \frac{\partial C}{\partial z_j^{(l)}}$$

$$= \sum_k \left(\frac{\partial C}{\partial z_k^{(l+1)}} \right) \frac{\partial z_k^{(l+1)}}{\partial z_j^{(l)}}$$

$$\delta_k^{(l+1)}$$

$$z_j^{(l+1)} = \sum_{i=1}^{N_l} w_{ij}^{(l+1)} a_i^{(l)} + b_j^{(l+1)}$$

$$a_i^{(l)} = \sigma(z_i^{(l)})$$

$$\frac{\partial z_k^{(l+1)}}{\partial z_j^{(l)}} = w_{kj}^{(l+1)} \sigma'(z_j^{(l)})$$

$$\delta_j^{(l)} = \sum_k \delta_k^{(l+1)} w_{kj}^{(l+1)} \sigma'(z_j^{(l)})$$

updates of gradients

$$w_{jk}^{(\ell)} \leftarrow w_{jk}^{(\ell)} - \eta \, \delta_j^{(\ell)} a_k^{(\ell-1)}$$

$$b_j^{(\ell)} \leftarrow b_j^{(\ell)} - \eta \underbrace{\frac{\partial C}{\partial b_j^{(\ell)}}}_{\delta_j^{(\ell)}}$$

- algorithm
  - Define architecture (model)
    - # nodes
    - # hidden layers
    - activation functions
    - cost/loss function
  - initialize $\theta = \{W, b\}$
  - set up design matrix X

- learning rate, hyper-parameters
- gradient methods

(i) Perform first FF step and continue from hidden layer $l = 1$ to $l = L$ $a^{(L)}$

- compute $\delta_j^{(L)}$

- Then backpropagate

$l = L-1$, $l = L-2$, ...

$l = 1$

compute $\delta_j^{(l)}$

- Train gradients, update

$$w_{jk}^{(l)} \leftarrow w_{jk}^{(l)} - \eta \, \delta_j^{(l)} a_k^{(l-1)}$$

$$b_j^{(l)} \leftarrow b_j^{(l)} - \eta \, \delta_j^{(l)}$$

- continue till convergence