Backpropagation (Chain Rule on Graphs)

$\Uparrow$

Chain Rule: $h(x) = f(g(x))$

$$\frac{dh}{dx} = \boxed{\frac{df}{\cancel{dg}} \cdot \frac{\cancel{dg}}{dx}} = f'(g(x))g'(x)$$

Ex: $h(x) = e^{x^2}$

$f(x) = e^x \quad f'(x) = e^x$

$g(x) = x^2 \quad g'(x) = 2x$

$h'(x) = f'(g(x))\,h'(x)$

$= e^{g(x)} \cdot g'(x) = e^{x^2} \cdot 2x$

$h(x) = f(\underbrace{g(t(x))})$

$$\frac{dh}{dx} = h'(x) = f'(g(t(x))) \cdot g'(t(x)) \cdot t'(x)$$

$h(x) = e^{(2x+3)^3}$

$f(x) = e^x \quad f'(x) = e^x$

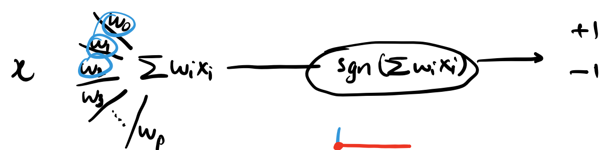
$g(x) = x^3 \quad g'(x) = 3x^2$

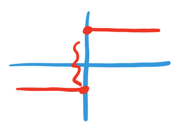
$t(x) = 2x+3 \quad t'(x) = 2$

$h'(x) = e^{(2x+3)^3} \cdot 3(2x+3)^2 \cdot 2$

$= 6(2x+3)^2 e^{(2x+3)^3}$

Perceptrons:





Gradient Descent: $\min_x f(x)$

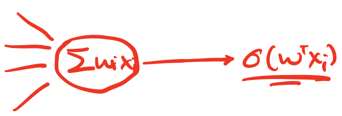
$$x^{(t+1)} = x^{(t)} - \eta f'(x^{(t)})$$

(logistic)

Sigmoid unit / perceptron

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

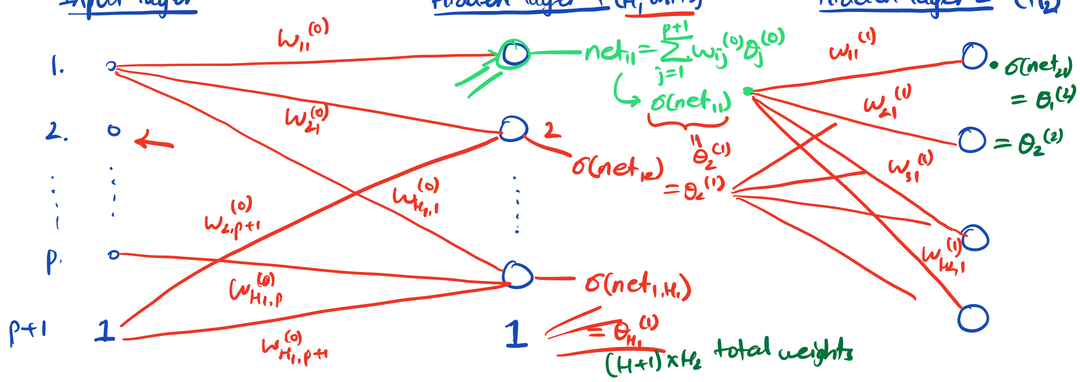
$$\sigma'(x) = \sigma(x)(1-\sigma(x))$$



$\to$ ReLU / tanh

$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$

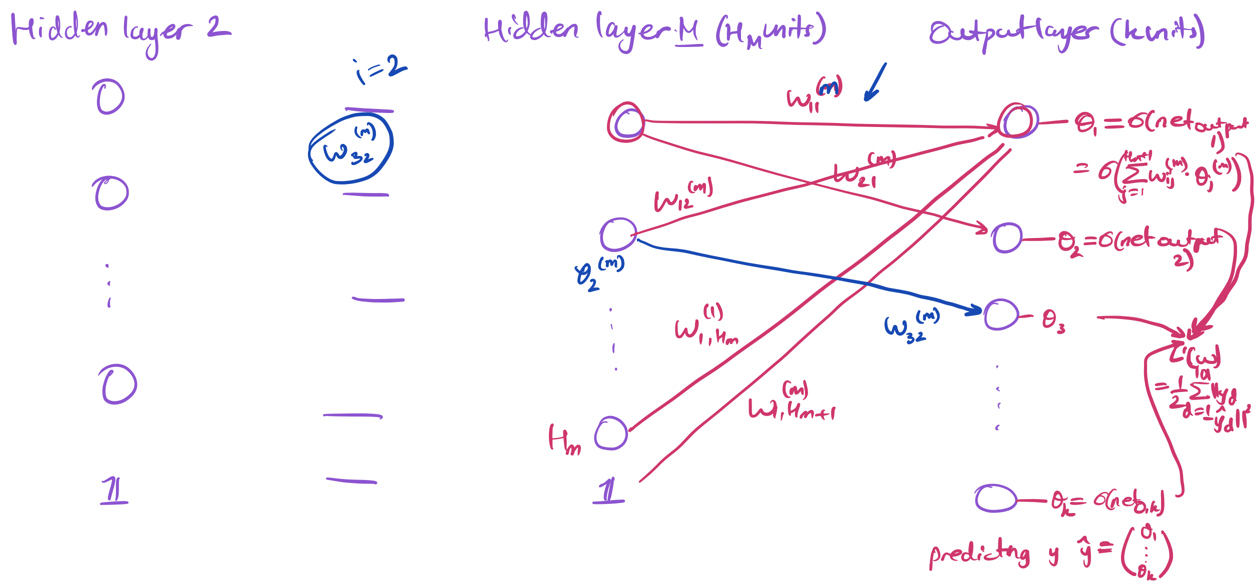
MLP / Neural Network: $D = \{(x, y) : x_i \in \mathbb{R}^p, y_i \in \mathbb{R}^K\}$



$w_{ji}^{(k)}$ = connecting ith input to jth node in kth layer

$x_i = \theta^{(0)} = \begin{bmatrix} \theta_1^{(0)} \\ \vdots \\ \theta_p^{(0)} \end{bmatrix}$

$(p+1) \times H_1$ total weights



$$L(w) = \frac{1}{2}\sum_{d=1}^{|D|} \|y_d - \hat{y}_d\|^2$$

$$= \underbrace{\frac{1}{2}\sum_{d=1}^{|D|}\sum_{k=1}^{K}}(y_{dk} - \theta_{dk})^2$$

w = vector of all weights in network:

$$\begin{bmatrix} w_{11}^{(0)} \\ w_{12}^{(0)} \\ \vdots \\ w_{H,p+1}^{(0)} \\ w_{11}^{(1)} \\ \vdots \\ w_{11}^{(m)} \\ \vdots \end{bmatrix}$$

$$\hat{\theta}^{(2)} = \sigma(W^{(1)} \cdot \theta^{(1)})$$

$$\theta^{(2)} = \begin{pmatrix} \hat{\theta}^{(2)} \\ 1 \end{pmatrix} \in \mathbb{R}^{(H_2+1)\times 1}$$

Aside:

$$\underline{W^{(0)}} = \begin{bmatrix} w_{11}^{(0)} & w_{12}^{(0)} & \cdots & w_{1,p+1}^{(0)} \\ w_{21}^{(0)} & w_{22}^{(0)} & \cdots & w_{2,p+1}^{(0)} \\ \vdots & & & \\ w_{H_1,1}^{(0)} & \cdots & \cdots & w_{H,p+1}^{(0)} \end{bmatrix}$$

$$\underline{\theta_1^{(1)}} = \sigma\left(\sum_{j=1}^{p+1} w_{\underline{1}j}^{(0)} \cdot \theta_{\underline{j}}^{(0)}\right)$$

$$= \sigma(\text{1st element of } W^{(0)} \cdot \theta^{(0)})$$

$$\theta_2^{(1)} = \sigma\left(\sum_{j=1}^{p+1} w_{2j}^{(0)} \cdot \theta_j^{(0)}\right)$$

$$= \sigma(\text{2nd el. } W^{(0)} \cdot \theta^{(0)})$$

$$\hat{\theta}^{(1)} = \sigma(\underbrace{W^{(0)}}_{\mathbb{R}^{H_1\times(p+1)}} \cdot \underbrace{\theta^{(0)}}_{\mathbb{R}^{(p+1)\times 1}}) = \mathbb{R}^{H_1 \times 1}$$

$$\theta^{(1)} = \begin{bmatrix} \hat{\theta}^{(1)} \\ 1 \end{bmatrix} \in \mathbb{R}^{(H_1+1)\times 1}$$

Input | H1 | H2 | H3 | ....

$$x = \theta^{(0)} \qquad \theta^{(1)} = \begin{bmatrix} \sigma(W^{(0)}\theta^{(0)}) \\ 1 \end{bmatrix} \rightarrow \theta^{(2)} \begin{bmatrix} \sigma(W^{(1)} \cdot \theta^{(1)}) \\ 1 \end{bmatrix} \quad \ldots\ldots$$

$$\theta = \sigma(W^{(m)} \cdot \theta^{(m)})$$

$$= \sigma\left(W^{(m)} \cdot \begin{bmatrix} \hat{\theta}^{(m)} \\ 1 \end{bmatrix}\right)$$

$$= \sigma\left(W^{(m)}\begin{pmatrix} \sigma\left(W^{(m-1)}\begin{pmatrix} \sigma(W^{(m-2)}\,\theta^{(m-2)}) \\ 1 \end{pmatrix}\right) \\ 1 \end{pmatrix}\right)$$

$$= \sigma\left(w^{(m)} \cdot \begin{bmatrix} \sigma(w^{(m-1)} \cdot \theta^{(m-1)}) \\ 1 \end{bmatrix}\right)$$

Train NN:

- GD: $w_{ji}^{(t+1)} = w_{ji}^{(t)} - \eta \boxed{\frac{\partial L(w)}{\partial w_{ji}^{(t)}}}$

Linear Reg: $\hat{y} = w^T x$

$$L(w) = \frac{1}{2} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

$$\frac{\partial L(w)}{\partial w} = 0$$

- Stochastic GD:

$$w_{ji}^{(t+1)} = w_{ji}^{(t)} - \eta \boxed{\frac{\partial L_d(w)}{\partial w_{ji}^{(t)}}}$$

$d = \{d_i\}$ $\quad w = (X^T X)^{-1} X^T y$

- Batch GD

$$L_d(w) = \frac{1}{2} \sum_{d_i \in \underline{d}} \sum_{k=1}^{d} (y_{d_i k} - \theta_{d_i, k})^2$$

$d = \{d_1, d_2, d_3\}$

Case 1: $w_{ji} = w_{ji}^{(m)} \to$ output layer

Goal: $\frac{\partial L_d(w)}{\partial w_{ji}}$

$$\frac{\partial L_d(w)}{\partial w_{ji}^{(m)}} = \frac{\partial}{\partial w_{ji}^{(m)}} \frac{1}{2} \sum_{k=1}^{k} (y_k - \underline{\theta_k})^2$$

$$= \frac{\partial}{\partial w_{ji}^{(m)}} \frac{1}{2} \sum_{k=1}^{k} (y_k - \sigma(net_{output,k}))^2$$

$$= \frac{\partial}{\partial w_{ji}^{(m)}} \frac{1}{2} \sum_{k=1}^{k} \left(y_{\underline{k}} - \sigma\left(\sum_{\ell=1}^{H_m+1} w_{k\ell}^{(m)} \theta_\ell^{(m)}\right)\right)^2$$

$$= \frac{\partial}{\partial w_{ji}^{(m)}} \frac{1}{2} \Big(y_j - \sigma\underbrace{\Big(\sum_{\ell=1}^{H_m+1} w_{j\ell}^{(m)} - \theta_\ell^{(m)}\Big)}_{\sigma(net\ output,j)}\Big)^2$$

$$\frac{\partial L_d(w)}{\partial \sigma(net_{o,j})} = \frac{\partial}{\partial \sigma(net_{o,j})} \frac{1}{2}(y_i - \sigma(net_{o,j}))^2$$

$$\frac{\partial}{\partial x} \frac{1}{2}(y-x)^2 = -(y-x)$$

$$\frac{\partial \sigma(x)}{\partial x} = \sigma'(x) = \sigma(x)(1-\sigma(x))^2$$

$$net_{o,j} = \sum_{\ell=1}^{H_m+1} w_{j\ell}^{(m)} \theta_\ell^{(m)}$$

$$= w_{j1}^{(m)} \theta_1^{(m)} + w_{j2}^{(m)} \theta_2^{(m)} + \dots + w_{jH_m+1}^{(m)} \theta_{H_m+1}^{(m)}$$

$$\frac{\partial L_d(w)}{\partial w_{ji}^{(m)}} = \boxed{\left[\frac{\partial L_d(w)}{\partial \sigma(net_{o,j})}\right]\left[\frac{\partial \sigma(net_{o,j})}{\partial (net_{o,j})}\right]} \cdot \frac{\partial (net_{o,j})}{\partial \underline{w_{ji}^{(m)}}}$$

$$= -(y_j - \sigma(net_{o,j}))\left[\sigma(net_{o,j})(1-\sigma(net_{o,j}))\right]\left[\theta_i^{(m)}\right]$$

$$\frac{\partial L_d(w)}{\partial net_{o,j}} = \frac{\partial L_d}{\partial \sigma(net_{o,j})} \cdot \frac{\partial \sigma(net_{o,j})}{\partial net_{o,j}}$$

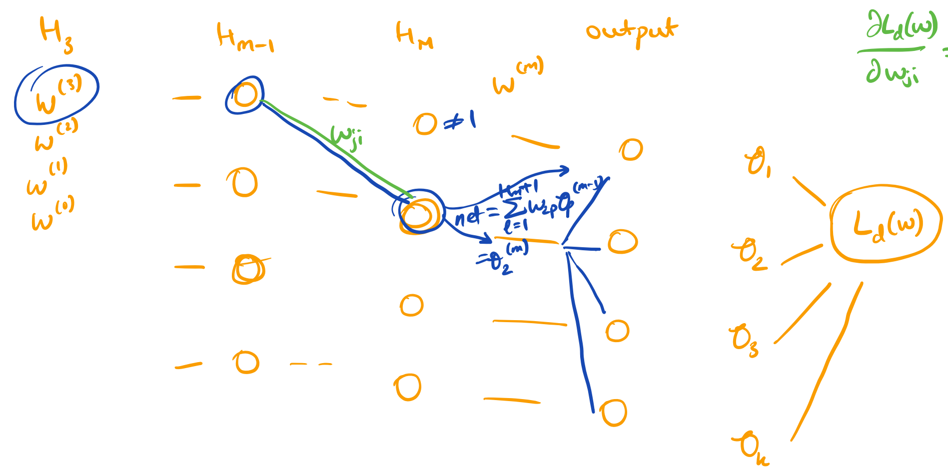
$$\frac{\partial L_d(w)}{\partial w_{ji}^{(m)}} = -\left[(y_j - \theta_j)\theta_j(1-\theta_j)\right]\theta_i^{(m)}$$

$$= -\delta_{output,j}\, \theta_i^{(m)}$$

$$\boxed{\delta_{output,j} = \frac{-\partial L_d(w)}{\partial net_{output,j}}}$$

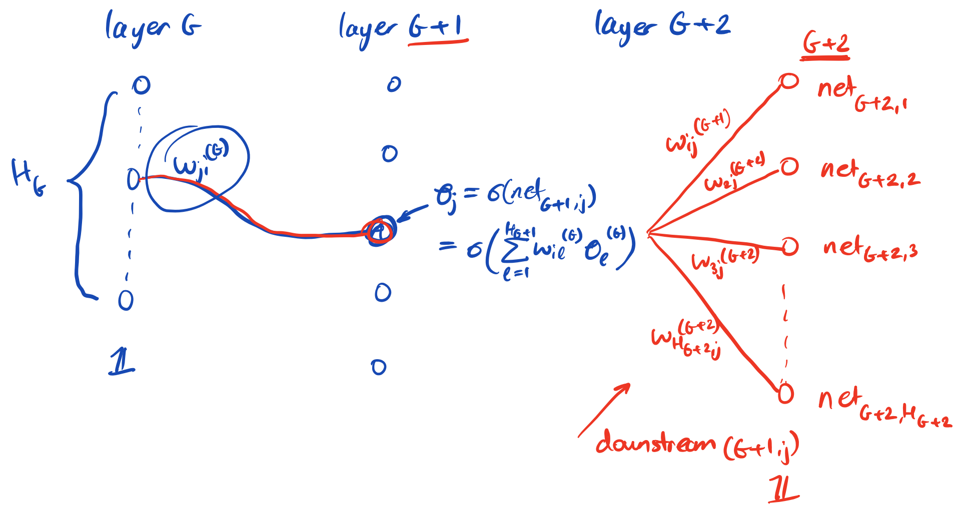
SGD:

$$(w_{ji}^{(m)})^{(t+1)} = (w_{ji}^{(m)})^{t} + \eta\, \delta_{o,j}\, \theta_i^{(m)}$$



$\frac{\partial L_d(w)}{\partial w_{ji}}$ = How does changing w_{ji} by small amount affect $L(w)$

Case 2: $w_{ji} = w_{ji}^{(G)}$



$$\frac{\partial L_d(w)}{\partial w_{ji}^{(G)}} = \frac{\partial L_d(w)}{\partial net_{G+1,j}} \times \frac{\partial net_{G+1,j}}{\partial w_{ji}^{(G)}}$$

$$= \left[\sum_{z \in downstream(G+1,j)} \frac{\partial L_d(w)}{\partial net_{G+2,z}} \cdot \frac{\partial net_{G+2,z}}{\partial net_{G+1,j}}\right] \frac{\partial net_{G+1,j}}{\partial w_{j,i}^{(G)}}$$

$$= \left[\left(\sum_{z} -\delta_{G+2,z}\right)\left(w_{zj}^{(G+1)} \theta_j^{(G+1)}(1-\theta_j^{(G+1)})\right)\right] \theta_i^{(G)}$$

SGD: $$(w_{ji}^{(G)})^{(t+1)} = (w_{ji}^{(G)})^{(t)} + \eta\, \theta_i^{(G)} \left[\theta_j^{(G+1)}(1-\theta_j^{(G+1)}) \sum_z \delta_{G+2,z} w_{zj}^{(G+1)}\right]$$

$$= (w_{ji}^{(G)})^{(t)} + \eta\, \theta_i^{(G)}\, \delta_{j,G+1}$$

$\delta_{j,G}$ = effect of net_{Gj} on loss