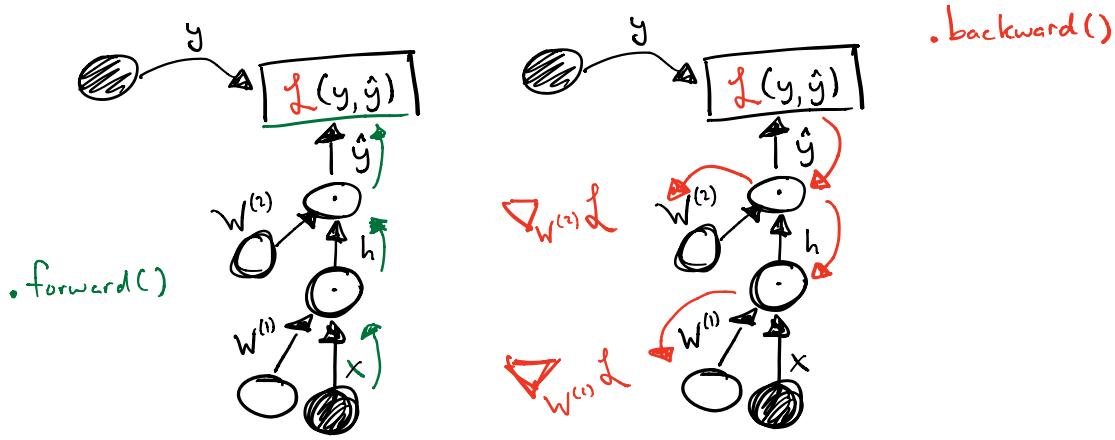
DS 4440 Backprop cont'd

Last time: Gradients on computation graphs via backpropagation.



To implement this, each node (or layer) in our graph needs to define how to run forwards and backwards.

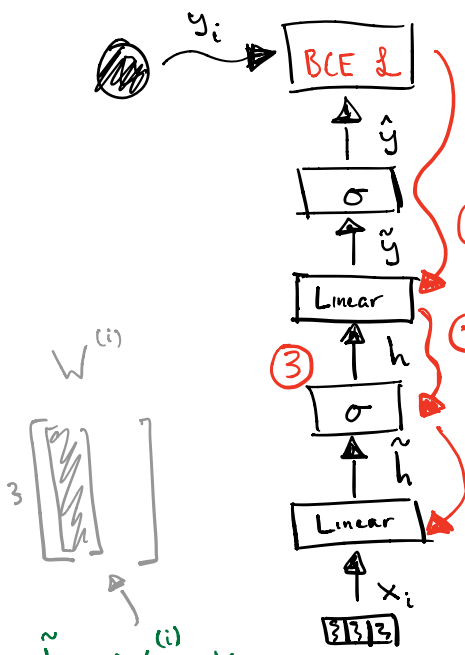
Let's try and generalize the algorithm, starting from a concrete example.

We'll consider binary classification with loss:

Probabilities

$$\text{BCELoss} = -y \lg \hat{y} - (1-y) \lg (1-\hat{y})$$

these are 0/1



$\hat{y} = \sigma(\tilde{y})$

$\tilde{y} = h \cdot W^{(0)}$

$W^{(0)} \in \mathbb{R}^{(2\times 1)}$

$W^{(i)} \in \mathbb{R}^{(3\times 2)}$

$x_i \in \mathbb{R}^{(1\times 3)}$

$\tilde{h}_0 = W^{(i)}_{:0} \cdot x_i$

① $$\frac{\partial \mathcal{L}}{\partial \tilde{y}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \tilde{y}} = (\hat{y} - y) = \delta^{(0)}$$

local error

② $$\frac{\partial \mathcal{L}}{\partial W^{(0)}} = \frac{\partial \tilde{y}}{\partial W^{(0)}} \delta^{(0)} = h \cdot (\hat{y} - y)$$

③ Local error @ h

$$\delta^{(h)} = \frac{\partial \mathcal{L}}{\partial h} = \frac{\partial \tilde{y}}{\partial h} \delta^{(0)} = \underbrace{W^{(0)} \cdot (\hat{y} - y)}_{\delta^{(h)} \in \mathbb{R}^{(2\times 1)}}$$

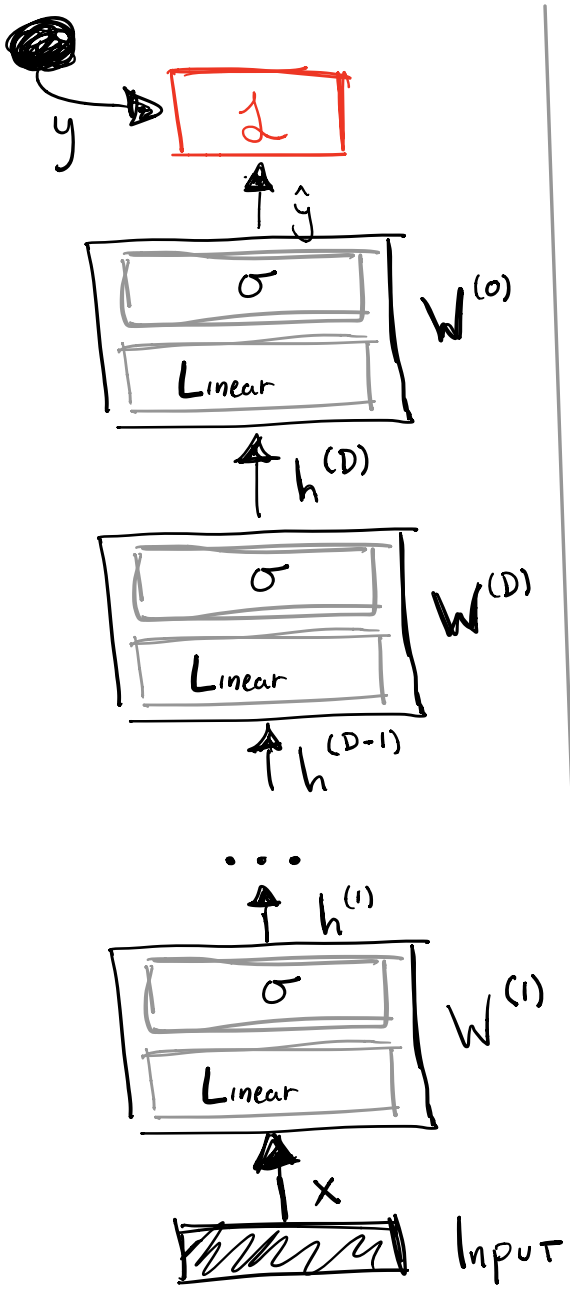
④ $$\underbrace{\frac{\partial \mathcal{L}}{\partial W^{(i)}_{:0}}}_{\mathbb{R}^{3\times 1}} = \boxed{\frac{\partial \tilde{h}_0}{\partial W^{(i)}_{:0}}} \boxed{\frac{\partial \mathcal{L}}{\partial \tilde{h}_0}}$$

$$\boxed{= \frac{\partial h}{\partial \tilde{h}_0} \delta^{(h)} = \left[\sigma(\tilde{h}_0)(1-\sigma(\tilde{h}_0)) \quad \emptyset\right] \delta^{(h)}}$$

$$\boxed{= x_i} = \underbrace{x_i}_{1\times 3} \cdot \underbrace{\sigma(\tilde{h}_0)(1-\sigma(\tilde{h}_0)) \cdot \delta_0^{(h)}}_{\text{Scalar}}$$

Let's try and generalize this.

Consider an arbitrarily deep network with D feed forward layers (Then an output layer).



$$\nabla_{W^{(k)}} \mathcal{L} = \nabla_{W^{(k)}} h^{(k)} \cdot \delta^{(h^k)}$$

Matrices!

$$\nabla_{h^{(k)}} \mathcal{L} = \nabla_{h^{(k)}} h^{(k+1)} \delta^{(h^{k+1})}$$

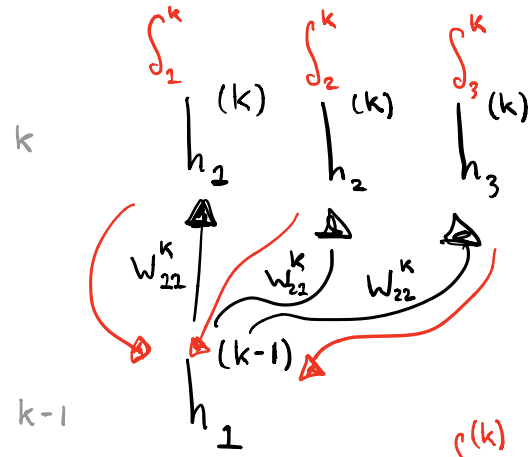
$\nabla_{h^{(k)}} h^{(k+1)}$ is a Jacobian

Matrix comprising all

Pairwise partials

$$J = \begin{array}{cc} & \\ d & \begin{bmatrix} \frac{\partial h_1^{k+1}}{\partial h_1^k} & \cdots & \frac{\partial h_d^{k+1}}{\partial h_1^k} \\ & \ddots & \\ & & \frac{\partial h_d^{k+1}}{\partial h_d^k} \end{bmatrix} \\ & d \end{array}$$

So if we think about a simple fully connected layer, the local error is intuitive.



$$\frac{\partial h_1^{(k)}}{\partial h_1^{(k-1)}} = \frac{\partial W_{11}^k \cdot h_1^{(k-1)}}{\partial h_1^{(k-1)}} = W_{11}^k$$

local error signal at j

$$\delta_1^{(k-1)} = \underbrace{\begin{bmatrix} \\ \end{bmatrix}}_{J} \begin{bmatrix} \\ \end{bmatrix}^{\delta^{(k)}} = \sum_{j=1}^{d} \boxed{\frac{\partial h_j^k}{\partial h_1^{k-1}}} \boxed{\delta_j^{(k)}}$$

Influence of h_1^{k-1} on h_j^k

$$= \sum_{j=1}^{d} W_{j1} \delta_j^{(k)}$$

<u>Exercise</u> Let's think about implementation of an arbitrary layer such that it supports backprop.

```
def forward(    )
    ...

def backward( )
    ...
```

- What information do we need for backprop?

(See Exercise)

(Bonus Derivation)

① $$\frac{\partial \mathcal{L}}{\partial \tilde{y}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \tilde{y}} = (\hat{y} - y)$$

$$\frac{\partial}{\partial \hat{y}} -y \lg \hat{y} - (1-y) \lg (1-\hat{y})$$

$$= \frac{\partial}{\partial \hat{y}} -y \lg \hat{y} - \boxed{\frac{\partial}{\partial \hat{y}} (1-y) \lg (1-\hat{y})}$$

$$= \cancel{\left(\frac{\partial}{\partial \hat{y}} -y\right)}^{\emptyset} \cdot \lg \hat{y} - y \frac{\partial}{\partial \hat{y}} \lg \hat{y}$$

$$= -y \frac{1}{\hat{y}} = \frac{-y}{\hat{y}}$$

$$\cancel{\frac{\partial}{\partial \hat{y}} (1-y)}^{\emptyset} \cdot \lg (1-\hat{y}) + (1-y) \frac{\partial}{\partial \hat{y}} \lg (1-\hat{y})$$

$$= (1-y) \frac{\partial}{\partial \hat{y}} (1-\hat{y}) \frac{1}{(1-\hat{y})}$$

$$= (1-y) \cdot -1 \cdot \frac{1}{(1-\hat{y})}$$

$$= \frac{-(1-y)}{(1-\hat{y})}$$

$$= \frac{-y}{\hat{y}} + \frac{(1-y)}{(1-\hat{y})}$$

$$= \frac{\partial \mathcal{L}}{\partial \hat{y}}$$

$$\frac{\partial \hat{y}}{\partial \tilde{y}} = \frac{\partial \sigma(\tilde{y})}{\partial \tilde{y}} = \hat{y} \cdot (1-\hat{y})$$

$= \sigma(\tilde{y})$

So... $$\frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \tilde{y}} = \left(\frac{-y}{\hat{y}} + \frac{(1-y)}{(1-\hat{y})}\right) \cdot \hat{y}(1-\hat{y})$$

$$= (\cancel{\hat{y}}(1-\hat{y})) \cdot \frac{-y}{\cancel{\hat{y}}} + (\hat{y}\cancel{(1-\hat{y})}) \cdot \frac{(1-y)}{\cancel{(1-\hat{y})}}$$

$$= (1-\hat{y}) \cdot -y + \hat{y}(1-y)$$

$$= -y(1-\hat{y}) + \hat{y}(1-y) = -y + \cancel{\hat{y}y} + \hat{y} - \cancel{\hat{y}y} = \hat{y} - y$$