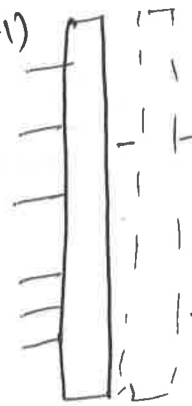
①

COL865
Deep Learning
Backpropagation
August 3rd, 2017

Hidden Unit (k)

$\hat{x}^{(k-1)}$

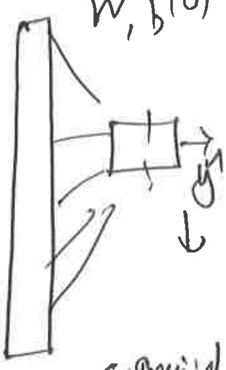


$\hat{x}^{(k)}$

$\rightarrow g^{(k)}(W^{(k)T}\hat{x}^{(k-1)} + b^{(k)})$

$W^{(k)}, b^{(k)}$ ⇓ Activation Unit

$W^{(0)}, b^{(0)}$



y'

sigmoid

$y = g^{(0)}(W^{(0)T}\hat{x}^{(k)} + b^{(0)})$

log likelihood loss

$$\boxed{\hat{J}(\theta) = \cancel{\tfrac{1}{m}}\sum_{i=1}^{m} L(y^{(i)}, \hat{y}^{(i)}; \theta)}$$

$\downarrow$

$\nabla \hat{J}(\theta)$

Gradient of $\hat{J}(\theta)$ w.r.t ~~to~~ network parameters

Chain Rule of ~~probability~~ Calculus:-

~~$y = g(x)$~~

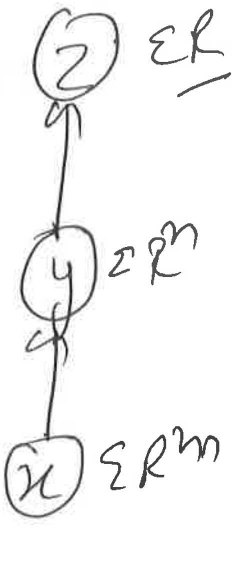
$x \in \mathbb{R}^m \rightarrow \boxed{g(x)} \rightarrow \vec{y} \in \mathbb{R}^n \rightarrow \boxed{f(y)} \rightarrow z \rightarrow \cancel{R}$

Imp

Jacobian Gradient Product

$$\boxed{\nabla_x z = ? \left(\frac{\partial y}{\partial x}\right)^T \nabla_y z}$$

$$\frac{\partial z}{\partial x_i} = \sum_j \left[\frac{\partial z}{\partial y_j} \cdot \left(\frac{\partial y_j}{\partial x_i}\right)\right]$$



Jacobian Matrix

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_m} \\ \vdots & & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_m} \end{bmatrix}$$

$$\frac{\partial z}{\partial y} = \begin{bmatrix} \frac{\partial z}{\partial y_1} \\ \vdots \\ \frac{\partial z}{\partial y_n} \end{bmatrix}$$

Forward_Computation (x, θ) {

For $k = 1$ to K

$\cancel{\hat{x}^{(k)} = \hat{A}(W^{(k)T} \hat{x}^{(k-1)})}$

$$\left.\begin{aligned} z^{(k)} &= W^{(k)T} \hat{x}^{(k-1)} + b^{(k)} \\ \hat{x}^{(k)} &= g^{(k)}(z^{(k)}); \end{aligned}\right\} \text{ kth hidden Layer}$$

$g^{(k)} \to$ Activation function

end for

$z^{(0)} = W^{(0)T} \hat{x}^{(K)} + b^{(0)}$

$\hat{y} = g^{(0)}(z^{(0)});$

$g^{(0)} \to$ sigmoid (output function)

}

②

Regularizer

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} L(\hat{y}, y^{(i)}; \theta) \cancel{+ \lambda \Omega(\theta)}$$

$\Downarrow$

$\cancel{\sum_{j=1}^{N} \theta_j^2}$

$\cancel{\theta_j}$

$\|\theta\|_2^2$

For clarity, one example:-

$$\boxed{J(\theta) = L(\hat{y}, y, \theta) \cancel{+ \lambda \Omega(\theta)}}$$

Backpropagation:-

Output Layer:-

$\cancel{d^{(o)} = \frac{\partial J(\theta)}{\partial \hat{y}}}$

$d_{z} \quad \nabla_{z^{(o)}} J$

$\cancel{d_{z^{(o)}}} = \cancel{\nabla_{\hat{y}} J(\theta)} \quad \nabla_{\hat{y}} L(y, \hat{y}, \theta) \odot g^{(o)\prime}(z^{(o)})$

$\cancel{d_{W^{(o)}}} =$

Hidden Layers:-

$\cancel{\nabla_z}$

$\nabla_{W^{(o)}} J$

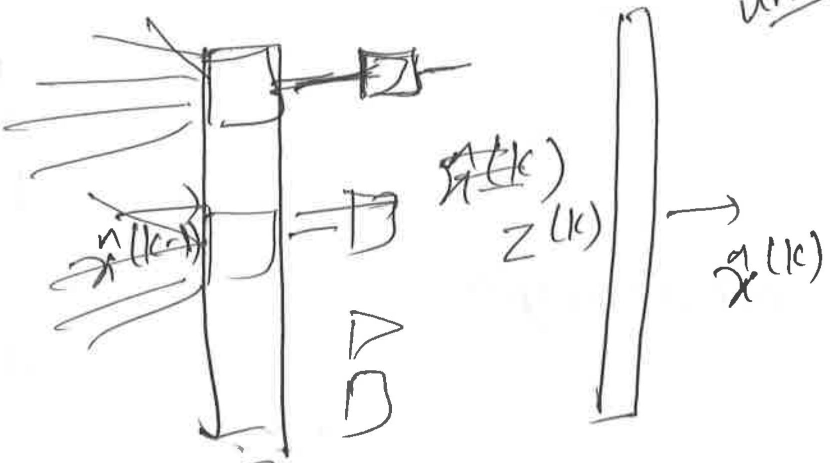
$= \quad \nabla_{z^{(o)}} J \cdot \hat{x}^{(k)T}$

$\nabla_{b^{(o)}} J = \nabla_{z^{(o)}}$

Hidden Layers:-

Activation units

$x - y - 2$



Available:-

∂J

$\nabla_{\hat{x}^{(k)}} J$

$$\boxed{\hat{x}^{(k)} = g^{(k)}(z^{(k)})}$$

$$z^{(k)} = W^T \hat{x}^{(k)} + b^{(k)}$$

$$z_i^{(k)} = W_{i:}^T \hat{x}^{(k)} + b_i^{(k)}$$

$$\hat{x}^{(k)} \cdot \frac{\partial J}{\partial z_i^{(k)}}$$

$W^{(k)}$

$$\nabla_{z^{(k)}} J = \cancel{g^k} \; \nabla_{\hat{x}^{(k)}} J \odot g^{(k)\prime}(z^{(k)})$$

$$\nabla_{W^{(k)}} J = \nabla_{z^{(k)}} J \; \cancel{z^{(k)T}} \; g(z^{(k)})^T$$

$$\cancel{\hat{x}^{(k)} \nabla_{z^{(k)}} J^T} \quad = (\hat{x}^{(k)})^T$$

$$\nabla_{b^{(k)}} J = \nabla_{z^{(k)}} J$$

$$\nabla_{\hat{x}^{(k-1)}} J = W^{(k)T} \cdot \cancel{\nabla_{z^{(k)}} J} \; \nabla_{z^{(k)}} J$$