

# **Backpropagation – Mathematical Background**

Erhard Hinrichs

Seminar für Sprachwissenschaft  
Eberhard-Karls Universität Tübingen

# Product and Quotient Rule

Product Rule

$$\frac{d}{dx}(f(x) * g(x)) = \frac{d}{dx}(f(x)) * g(x) + f(x) * \frac{d}{dx}(g(x)) \quad (1)$$

Quotient Rule

$$\frac{d}{dx} \frac{f(x)}{g(x)} = \frac{\frac{d}{dx}(f(x)) * g(x) - f(x) * \frac{d}{dx}(g(x))}{[g(x)]^2} \quad (2)$$

# Chain Rule

$$\frac{d}{dx} [f(g(x))] = f' [g(x)] * g'(x) \quad (3)$$

Remarks:

- ▶  $f(u)$  is called *the outside function*
- ▶  $g(x)$  is called *the inside function*
- ▶ The chain rule can be stated in words as follows:  $\frac{df}{dx}$  equals the product of the derivatives of the outside function  $\frac{df}{du}$  and of the inside function  $\frac{du}{dx}$ .
- ▶ The chain rule is often helpful when taking derivatives of functions such as:  $\frac{d}{dx}(5x - 4)^6$ ,  $\frac{d}{dx}\sqrt{x^2 - 1}$ ,  $\frac{d}{dx}\frac{1}{x^2 - 4x + 5}$

## Chain Rule – Example

$$\frac{d}{dx} [f(g(x))] = f' [g(x)] * g'(x) \quad (4)$$

Find the derivative of  $f(x) = (x^2 + 1)^3$

$$\begin{aligned} f'(x) &= 3(x^2 + 1)^{3-1} * 2x^{2-1} \\ &= 3((x^2 + 1)^2(2x)) \\ &= 6x((x^2 + 1)^2) \end{aligned} \quad (5)$$

# Derivative of Sigmoid Function

$$\begin{aligned}\frac{d}{dx} \sigma(x) &= \frac{d}{dx} \left[ \frac{1}{1 + e^{-x}} \right] \\ &= \frac{(0)(1 + e^{-x}) - (-e^{-x})(1)}{(1 + e^{-x})^2} \\ &= \frac{e^{-x}}{(1 + e^{-x})^2} \\ &= \frac{1}{1 + e^{-x}} \frac{e^{-x}}{1 + e^{-x}}\end{aligned}\tag{6}$$

## Derivative of Sigmoid Function (continued)

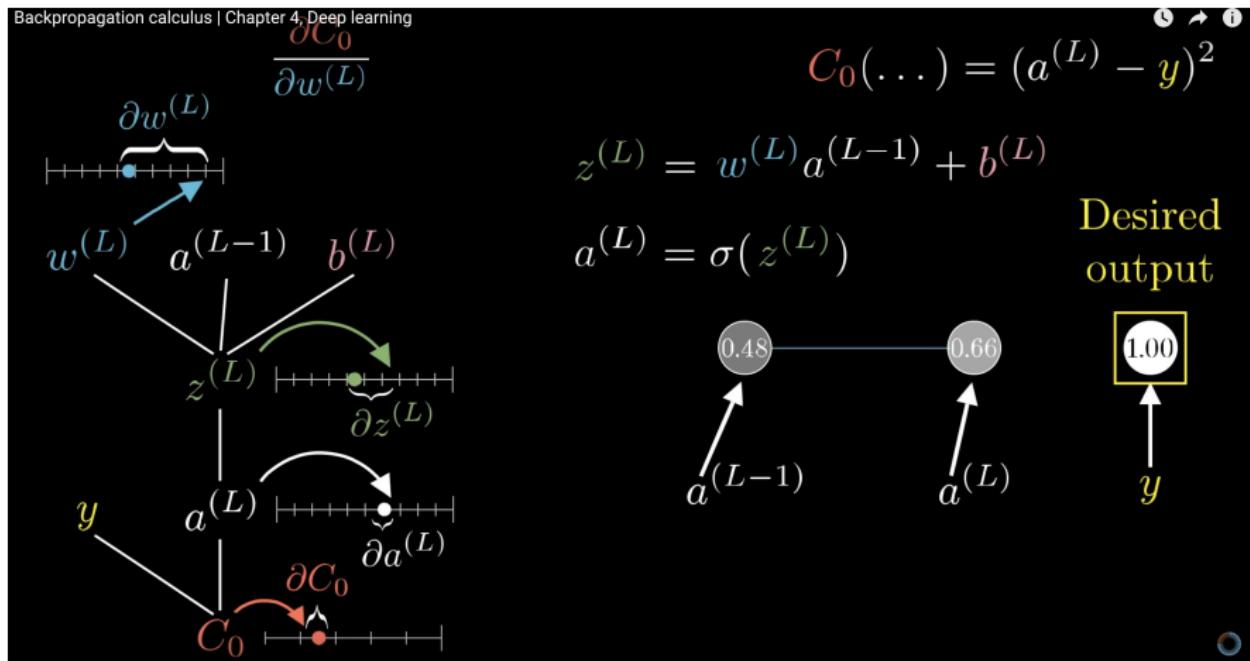
$$\begin{aligned} &= \frac{1}{1 + e^{-x}} \frac{e^{-x} + (1 - 1)}{1 + e^{-x}} \\ &= \frac{1}{1 + e^{-x}} \frac{1 + e^{-x} - 1}{1 + e^{-x}} \\ &= \frac{1}{1 + e^{-x}} \left[ \frac{1 + e^{-x}}{1 + e^{-x}} - \frac{1}{1 + e^{-x}} \right] \\ &= \frac{1}{1 + e^{-x}} \left[ 1 - \frac{1}{1 + e^{-x}} \right] \\ &= \sigma(x)(1 - \sigma(x)) \end{aligned} \tag{7}$$

## Derivatives of ReLU and tanh

$$\frac{dReLU(z)}{dz} = \begin{cases} 0 & \text{for } z < 0 \\ 1 & \text{for } z \geq 0 \end{cases} \quad (8)$$

$$\frac{dtanh(z)}{dz} = 1 - \tanh^2(z) \quad (9)$$

# Backprop in a Simple FFN



# Applying the Chain Rule

$$\frac{\partial C_0}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C_0}{\partial a^{(L)}}$$

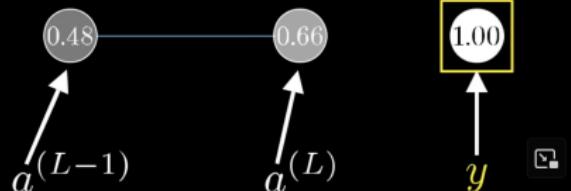
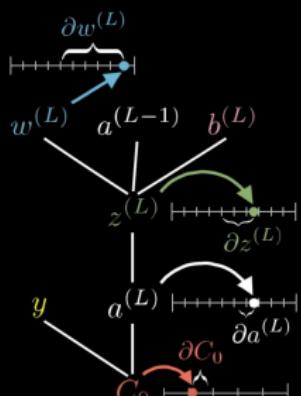
Chain rule

$$C_0(\dots) = (a^{(L)} - y)^2$$

$$z^{(L)} = w^{(L)}a^{(L-1)} + b^{(L)}$$

$$a^{(L)} = \sigma(z^{(L)})$$

Desired output



# Computing Relevant Derivatives

$$\frac{\partial C_0}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C_0}{\partial a^{(L)}}$$

$$C_0 = (a^{(L)} - y)^2$$

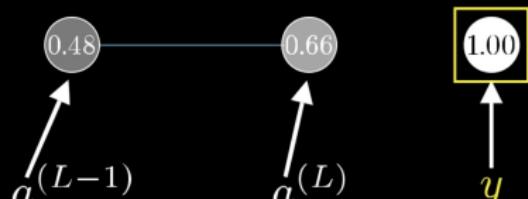
$$\frac{\partial C_0}{\partial a^{(L)}} = 2(a^{(L)} - y)$$

$$a^{(L)} = \sigma(z^{(L)})$$

$$\frac{\partial a^{(L)}}{\partial z^{(L)}} = \sigma'(z^{(L)})$$

$$\frac{\partial z^{(L)}}{\partial w^{(L)}} = a^{(L-1)}$$

$$z^{(L)} = w^{(L)}a^{(L-1)} + b^{(L)}$$



# Average of All Training Examples

$$\frac{\partial C_0}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C_0}{\partial a^{(L)}} = a^{(L-1)} \sigma'(z^{(L)}) 2(a^{(L)} - y)$$

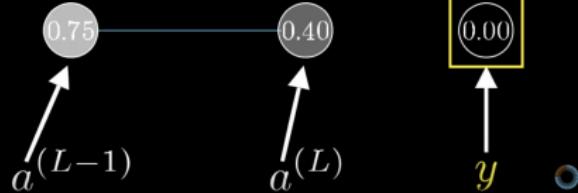
Average of all  
training examples

$$C_0 = (a^{(L)} - y)^2$$

$$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$$

$$\underbrace{\frac{\partial C}{\partial w^{(L)}}}_{\text{Derivative of full cost function}} = \overbrace{\frac{1}{n} \sum_{k=0}^{n-1} \frac{\partial C_k}{\partial w^{(L)}}}^{\sigma(z^{(L)})}$$

Derivative of  
full cost function



# Gradient

$$\frac{\partial C_0}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial C_0}{\partial a^{(L)}} = a^{(L-1)} \sigma'(z^{(L)}) 2(a^{(L)} - y)$$

$$\nabla C = \begin{bmatrix} \frac{\partial C}{\partial w^{(1)}} \\ \frac{\partial C}{\partial b^{(1)}} \\ \vdots \\ \frac{\partial C}{\partial w^{(L)}} \\ \frac{\partial C}{\partial b^{(L)}} \end{bmatrix}$$

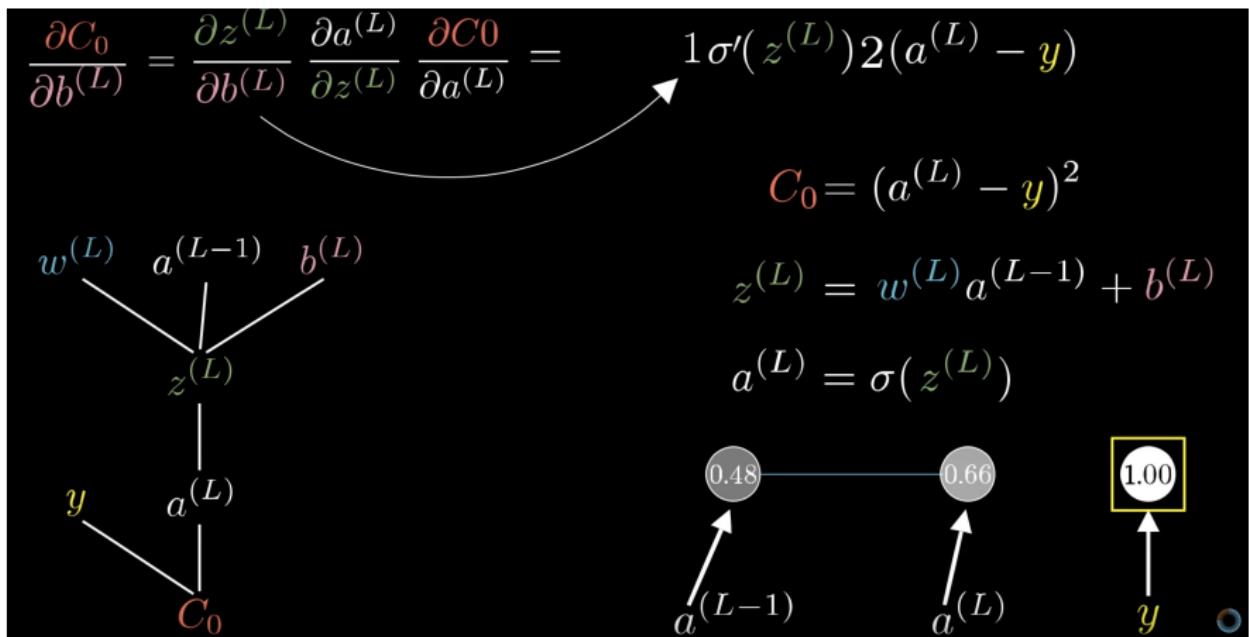
$C_0 = (a^{(L)} - y)^2$

$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$

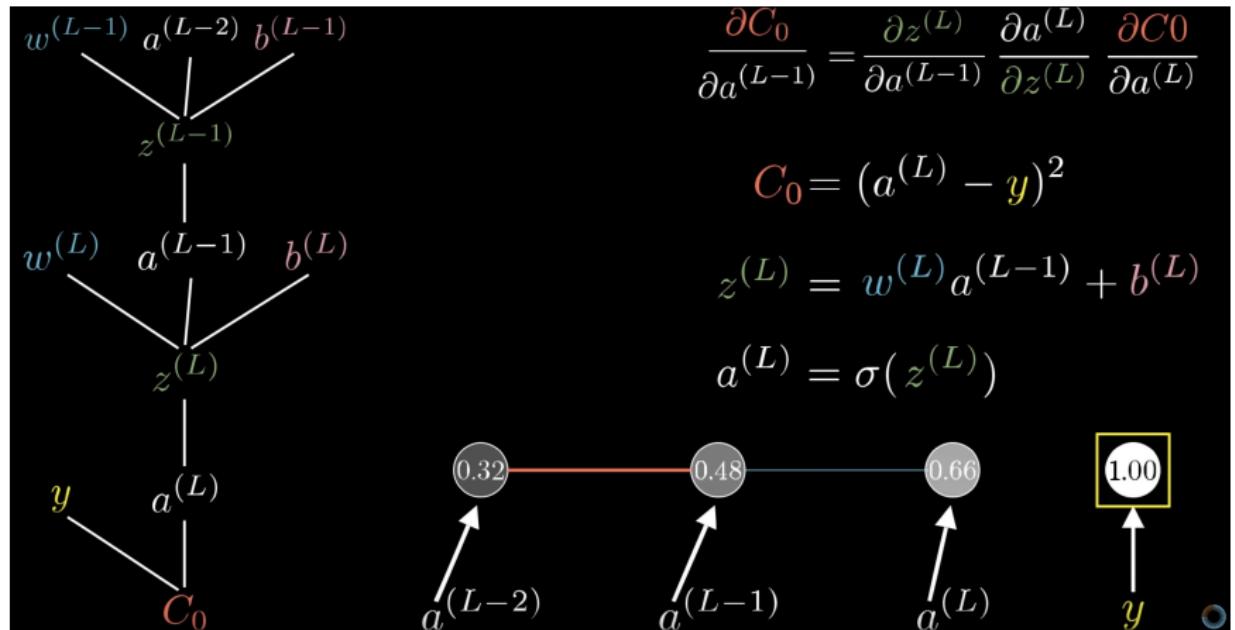
$a^{(L)} = \sigma(z^{(L)})$

The diagram illustrates the forward pass of a neural network layer. It shows three nodes in a row. The first node contains the value 0.48 and is labeled  $a^{(L-1)}$ . An arrow points from this node to the second node, which contains the value 0.66 and is labeled  $a^{(L)}$ . A third node contains the value 1.00 and is labeled  $y$ . This final node is highlighted with a yellow border. Below the nodes, there is a small blue circle.

# Derivative of Bias

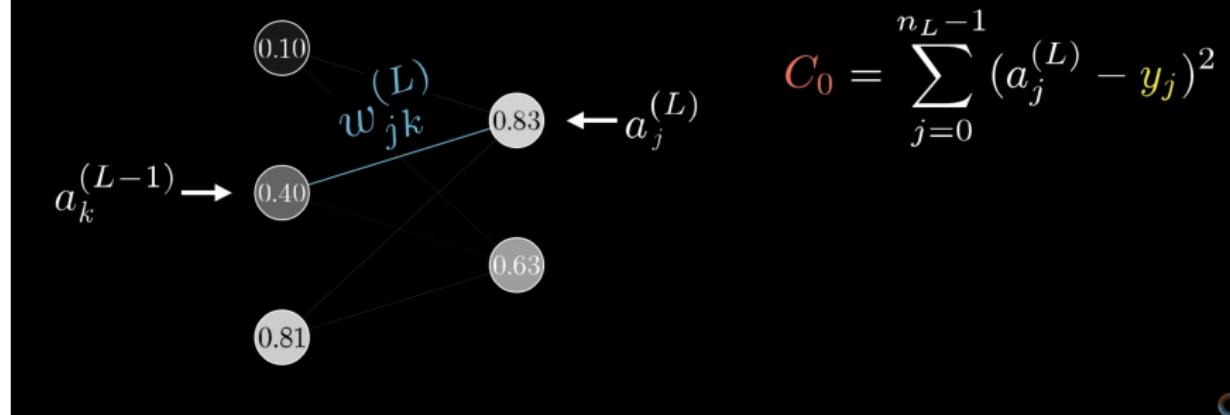


# Adding Layers



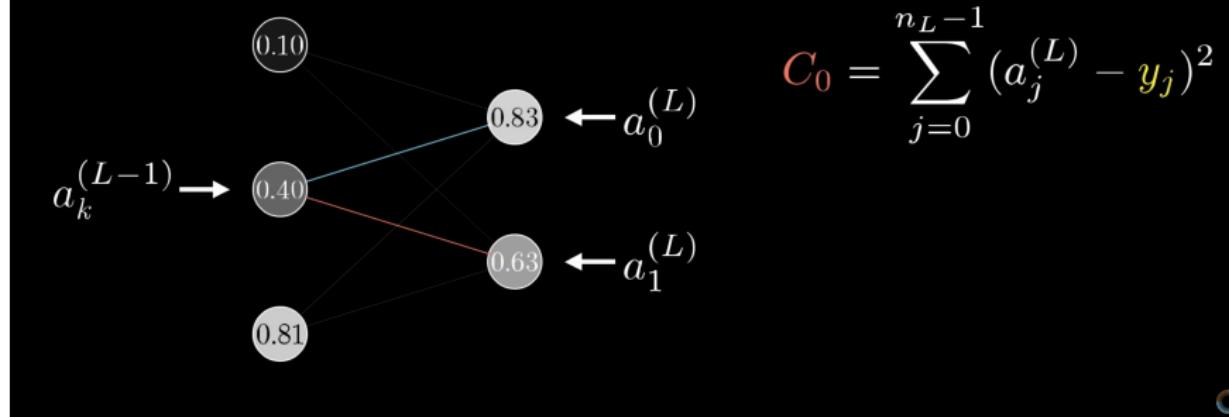
# Adding Nodes Per Layer

$$\frac{\partial C_0}{\partial w_{jk}^{(L)}} = \frac{\partial z_j^{(L)}}{\partial w_{jk}^{(L)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial C_0}{\partial a_j^{(L)}}$$
$$z_j^{(L)} = \dots + w_{jk}^{(L)} a_k^{(L-1)} + \dots$$
$$a_j^{(L)} = \sigma(z_j^{(L)})$$



# Summing Over Nodes Per Layer

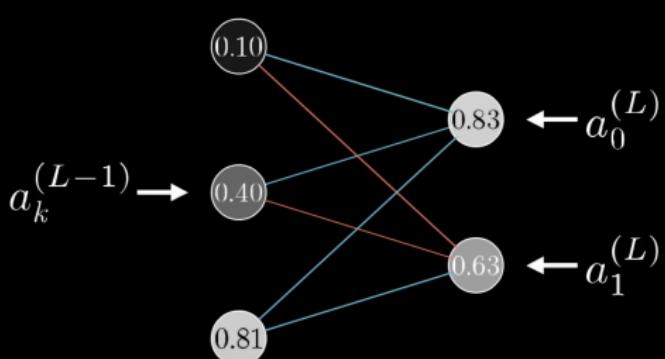
$$\frac{\partial C_0}{\partial a_k^{(L-1)}} = \underbrace{\sum_{j=0}^{n_L-1} \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial C_0}{\partial a_j^{(L)}}}_{\text{Sum over layer L}} \quad z_j^{(L)} = \dots + w_{jk}^{(L)} a_k^{(L-1)} + \dots$$
$$a_j^{(L)} = \sigma(z_j^{(L)})$$



# Summing Over Layers

aylist: Neural networks

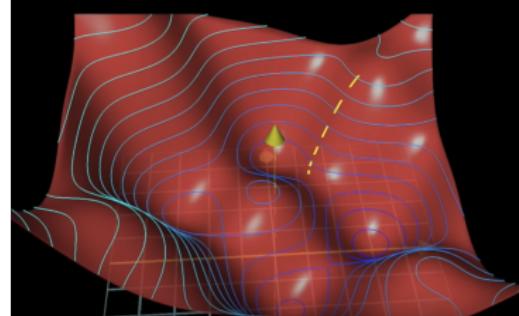
$$\frac{\partial a_k^{(L-1)}}{\partial a_k^{(L-1)}} = \underbrace{\sum_{j=0}^{n_L-1} \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial C_0}{\partial a_j^{(L)}}}_{\text{Sum over layer L}} \quad z_j^{(L)} = \cdots + w_{jk}^{(L)} a_k^{(L-1)} + \cdots$$
$$a_j^{(L)} = \sigma(z_j^{(L)})$$



$$C_0 = \sum_{j=0}^{n_L-1} (a_j^{(L)} - y_j)^2$$



# Summary


$$\nabla C \leftarrow \begin{cases} \frac{\partial C}{\partial w_{jk}^{(l)}} = a_k^{(l-1)} \sigma'(z_j^{(l)}) \boxed{\frac{\partial C}{\partial a_j^{(l)}}} \\ \sum_{j=0}^{n_{l+1}-1} w_{jk}^{(l+1)} \sigma'(z_j^{(l+1)}) \frac{\partial C}{\partial a_j^{(l+1)}} \\ \text{or} \\ 2(a_j^{(L)} - y_j) \end{cases}$$

A yellow arrow points from the top equation to the bottom right term, indicating the update rule for the weight  $w_{jk}^{(l+1)}$ .



# Derivatives of log functions

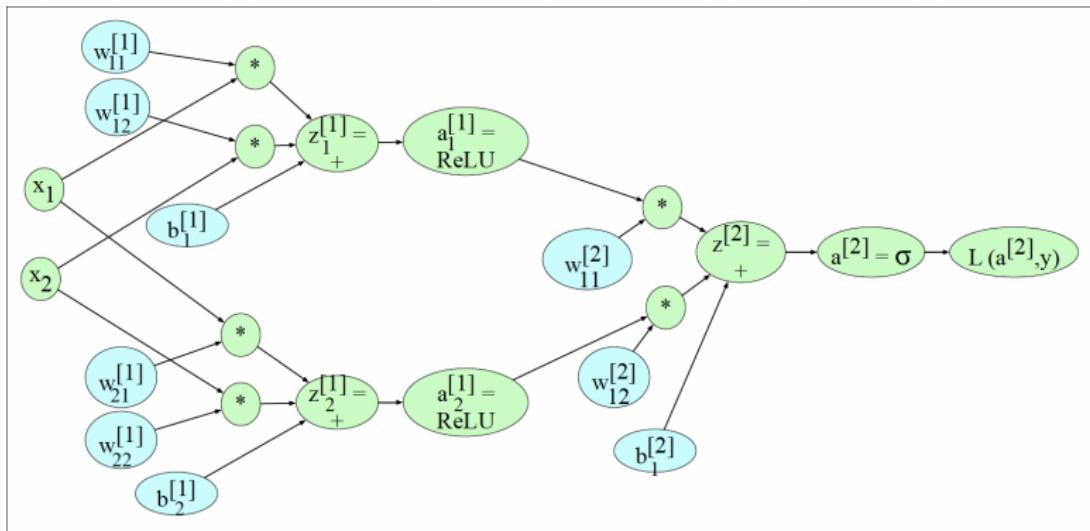
Derivative of Common Logarithm

$$\frac{d}{dx} \log_a x = \frac{1}{x * \ln(a)} \quad (10)$$

Derivative of Natural Logarithm

$$\frac{d}{dx} \ln x = \frac{1}{x} \quad (11)$$

# Sample Computation Graph for a simple 2-layer NN



## Derivative of the CE Loss Function with Respect to z

$$L_{CE}(a^{[2]}, y) = - \left[ y \log a^{[2]} + (1 - y) \log(1 - a^{[2]}) \right] \quad (12)$$

$$\begin{aligned} \frac{\partial L}{\partial a^{[2]}} &= - \left( \left( y \frac{\partial \log(a^{[2]})}{\partial a^{[2]}} \right) + (1 - y) \frac{\partial \log(1 - a^{[2]})}{\partial a^{[2]}} \right) \\ &= - \left( \left( y \frac{1}{a^{[2]}} \right) + (1 - y) \frac{1}{1 - a^{[2]}} (-1) \right) \\ &= - \left( \frac{y}{a^{[2]}} + \frac{y - 1}{1 - a^{[2]}} \right) \end{aligned} \quad (13)$$

## Derivative of the Sigmoid and the Chain Rule

$$\frac{\partial a^{[2]}}{\partial z} = a^{[2]}(1 - a^{[2]}) \quad (14)$$

$$\begin{aligned}\frac{\partial L}{\partial z} &= \frac{\partial L}{\partial a^{[2]}} \frac{\partial a^{[2]}}{\partial z} \\ &= - \left( \frac{y}{a^{[2]}} + \frac{y-1}{1-a^{[2]}} \right) a^{[2]}(1 - a^{[2]}) \\ &= a^{[2]} - y\end{aligned} \quad (15)$$