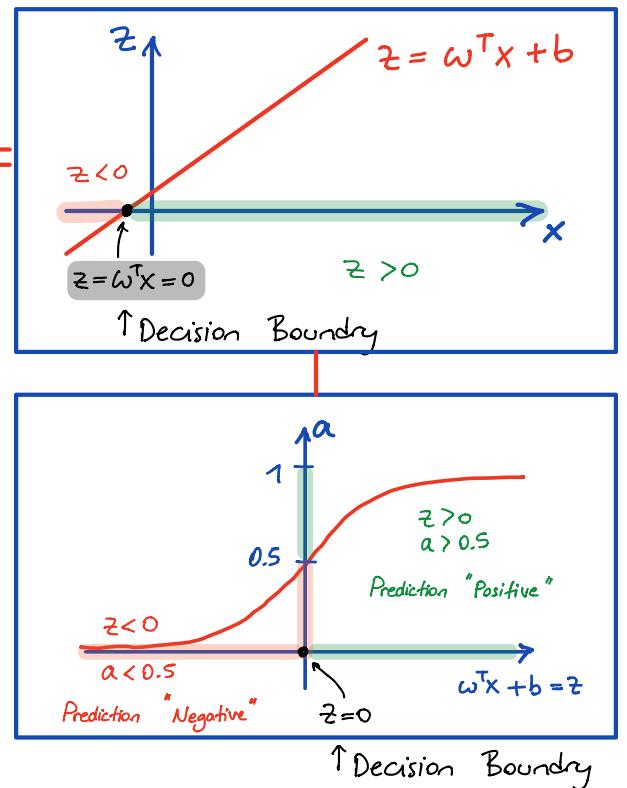
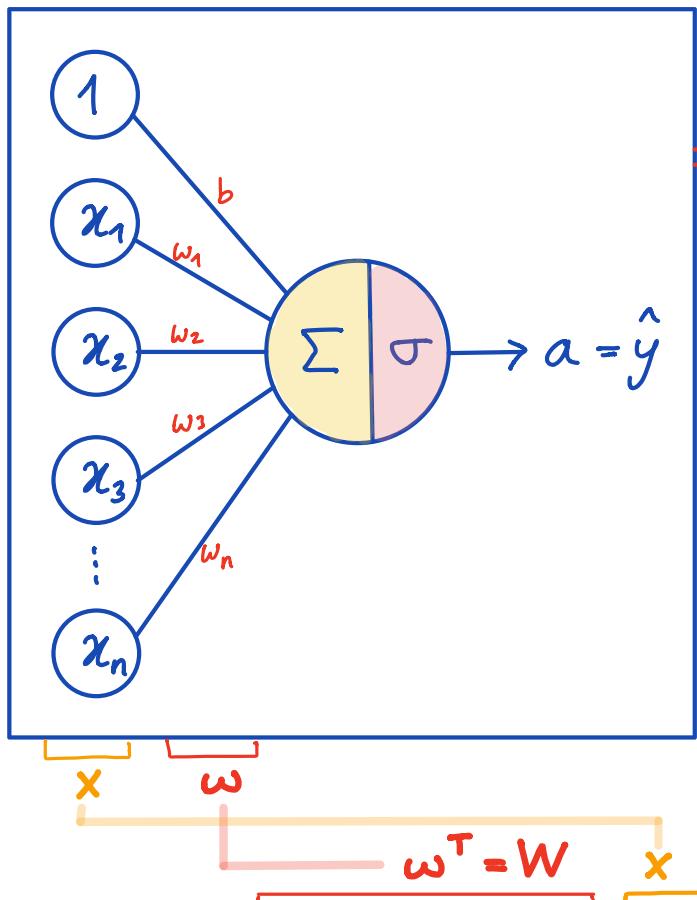


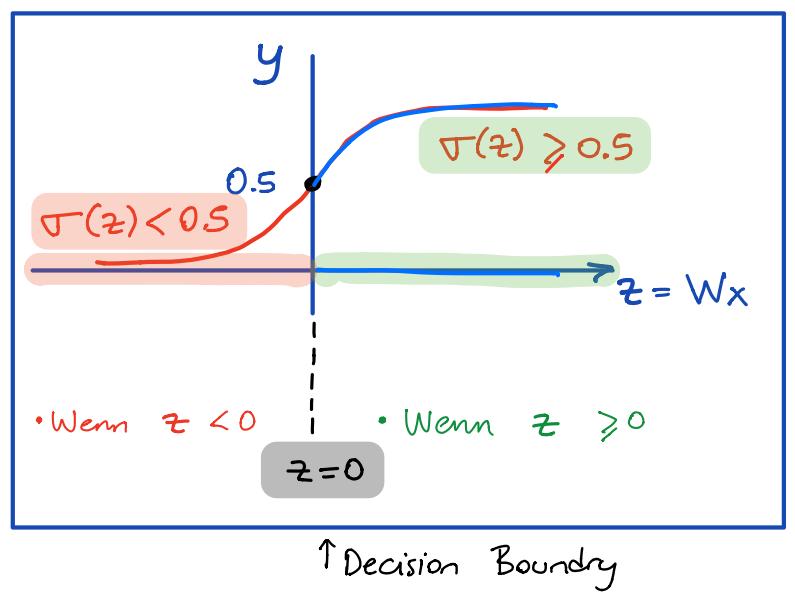
BINARY CLASSIFICATION WITH NN



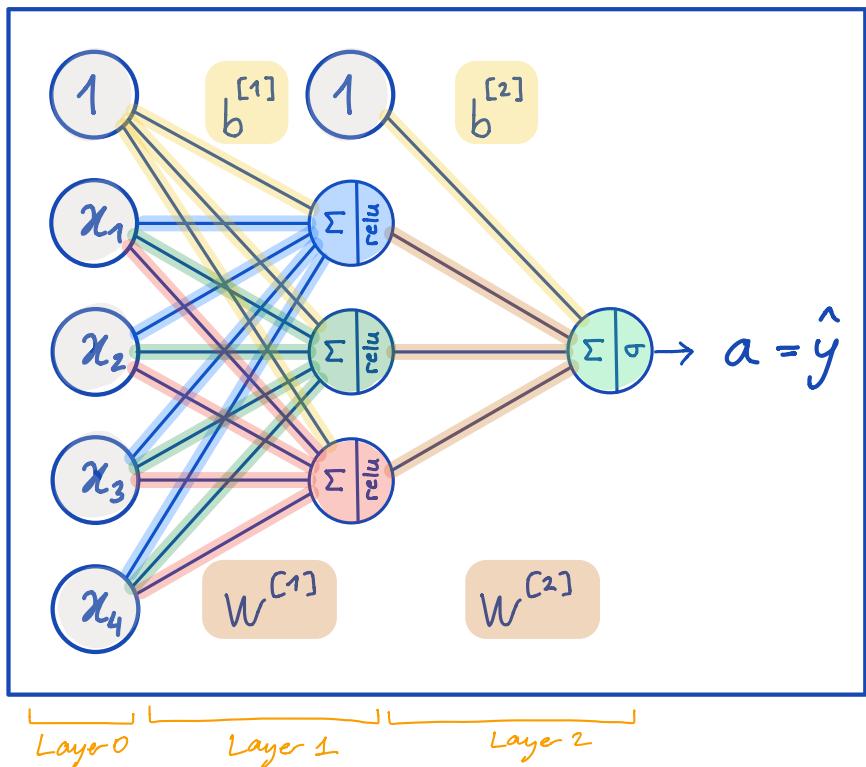
$$a = \sigma \left([w_1 \ w_2 \ \dots \ w_n] \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + b \right) = \sigma(Wx + b) = \sigma(z)$$

- The Sigmoid Function σ

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



BINARY CLASSIFICATION WITH DNN (Ng notation)



$$\begin{bmatrix} \text{blue bar} \\ \text{green bar} \\ \text{red bar} \end{bmatrix} \cdot \begin{bmatrix} \text{grey bar} \end{bmatrix} + \begin{bmatrix} \text{yellow bar} \end{bmatrix} = \begin{bmatrix} \text{blue bar} \\ \text{green bar} \\ \text{red bar} \end{bmatrix} = \underline{z}^{[1]}$$

$$\text{relu}\left(W^{[1]} \cdot x + b^{[1]}\right) = a^{[1]}$$

$$\begin{bmatrix} \text{orange bar} \end{bmatrix} \cdot \begin{bmatrix} \text{grey bar} \end{bmatrix} + \begin{bmatrix} \text{yellow bar} \end{bmatrix} = \begin{bmatrix} \text{green bar} \end{bmatrix} = \underline{z}^{[2]}$$

$$\sigma\left(W^{[2]} \cdot a^{[1]} + b^{[2]}\right) = a^{[2]}$$

Layer 0 Layer 1 Layer 2

$$\text{relu}\left(\begin{bmatrix} \text{blue bar} \\ \text{green bar} \\ \text{red bar} \end{bmatrix} \cdot \begin{bmatrix} \text{grey bar} \end{bmatrix}\right) + \begin{bmatrix} \text{yellow bar} \end{bmatrix} \xrightarrow{\text{Broadcasting}} \begin{bmatrix} \text{blue bar} \\ \text{green bar} \\ \text{red bar} \\ \text{blue bar} \\ \text{green bar} \\ \text{red bar} \\ \text{blue bar} \\ \text{green bar} \\ \text{red bar} \end{bmatrix}$$

$$\text{relu}\left(W^{[1]} \cdot x + b^{[1]}\right) = A^{[1]}$$

$$\sigma\left(\begin{bmatrix} \text{orange bar} \end{bmatrix} \cdot \begin{bmatrix} \text{grey bar} \end{bmatrix} + \begin{bmatrix} \text{yellow bar} \end{bmatrix}\right) \xrightarrow{\text{Broadcasting}} \begin{bmatrix} \text{green bar} \\ \text{green bar} \end{bmatrix}$$

$$\sigma\left(W^{[2]} \cdot A^{[1]} + b^{[2]}\right) = A^{[2]} = \hat{y}$$

CROSS-ENTROPY LOSS

$$A^{[2]} = \hat{y} = [\textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet}]$$

$$Y = [\textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet}]$$

$$\text{LOSS} = \text{DISTANCE}(\hat{y}, Y) = [\textcolor{brown}{\bullet} \textcolor{brown}{\bullet} \textcolor{brown}{\bullet} \textcolor{brown}{\bullet} \textcolor{brown}{\bullet} \textcolor{brown}{\bullet} \textcolor{brown}{\bullet}]$$

$$\bullet \quad \mathcal{L}(\bullet, \bullet) = \mathcal{L}(\hat{y}, y) = -y \cdot \log(\hat{y}) - (1-y) \cdot \log(1-\hat{y}) = \bullet$$

$$\bullet \quad \mathcal{L}([\textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet}], [\textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet}])$$

$$= \mathcal{L}(\hat{y}, Y) = -Y * \log(\hat{y}) - (1-Y) * \log(1-\hat{y})$$

$$= [\textcolor{brown}{\bullet} \textcolor{brown}{\bullet} \textcolor{brown}{\bullet} \textcolor{brown}{\bullet} \textcolor{brown}{\bullet} \textcolor{brown}{\bullet}]$$

elementwise

$$\bullet \quad Y * \log(\hat{y}) = [\textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet}] * \log([\textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet} \textcolor{lightgreen}{\bullet}])$$

COST FUNCTION

(average loss)

$$J(\theta) = \frac{1}{m} \cdot \text{np.sum}(\mathcal{L}(\hat{Y}, Y))$$

$$= \frac{1}{7} \cdot \underbrace{\text{np.sum}([\bullet \bullet \bullet \bullet \bullet \bullet \bullet])}_{\text{Sum over all Data points}}$$

(MINI BATCH GRADIENT DESCENT)

THE GRADIENT

$$\theta = [\text{blue bar} \quad \text{green bar} \quad \text{red bar} \quad \text{yellow bar} \quad \text{orange bar} \quad \bullet]^\top \quad (\text{Parameter})$$

$$\frac{\partial J}{\partial \theta} = d\theta = [\text{blue bar} \quad \text{green bar} \quad \text{red bar} \quad \text{yellow bar} \quad \text{orange bar} \quad \bullet]^\top \quad (\text{evaluate for current } \omega)$$

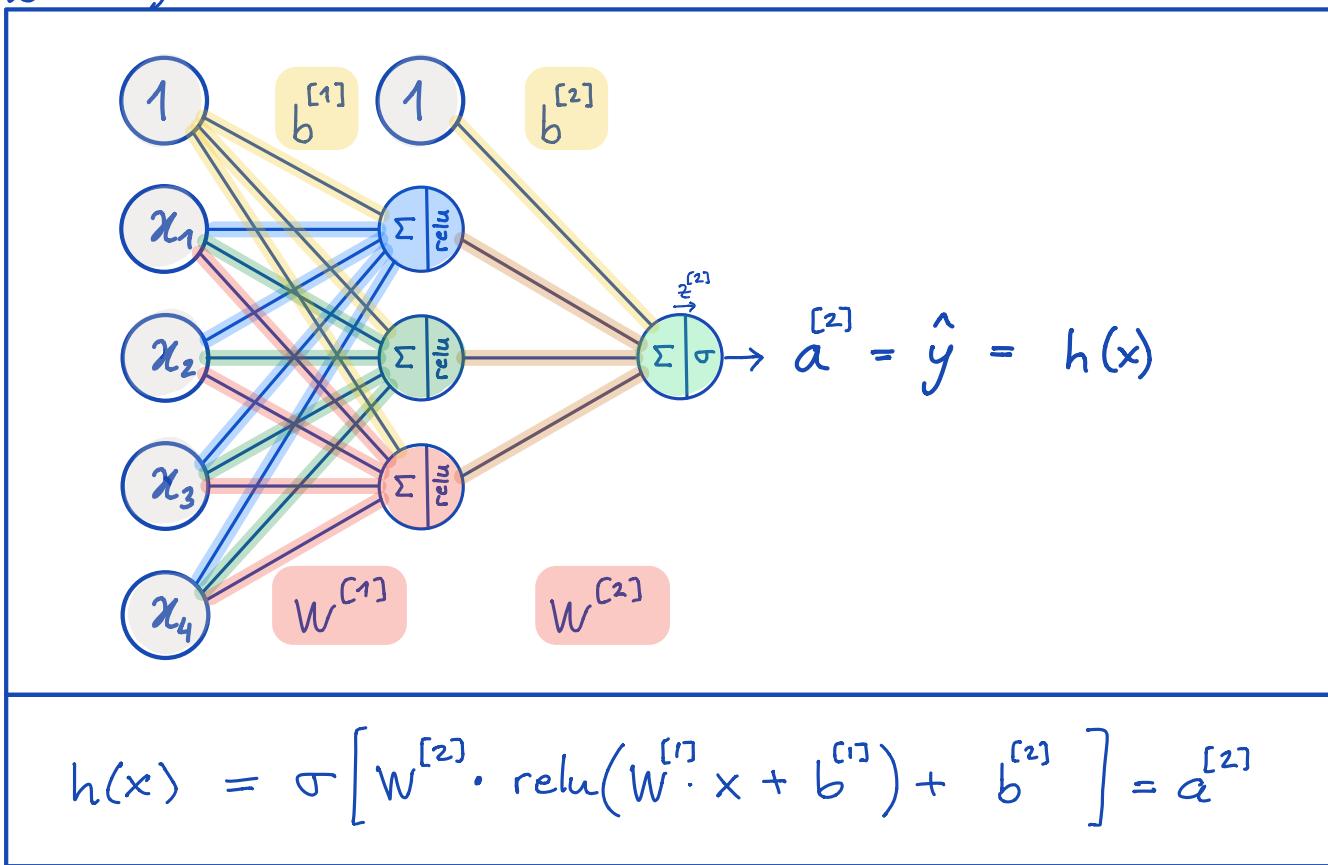
$$\nabla J = \left[\frac{\partial J}{\partial w_{11}^{[1]}} \dots \frac{\partial J}{\partial w_{34}^{[1]}} \frac{\partial J}{\partial b_1^{[1]}} \dots \frac{\partial J}{\partial b_3^{[1]}} \frac{\partial J}{\partial w_{11}^{[2]}} \dots \frac{\partial J}{\partial w_{13}^{[2]}} \frac{\partial J}{\partial b_{11}^{[2]}} \right]^\top \quad (\text{current } \omega)$$

Direction of steepest Ascent !

(Go in opposite direction) !

• Backpropagation - Multiple Layers

2-Layer Network (Graph Representation)



(Output as a function of x)

$$J = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(a^{(i)}, y^{(i)})$$

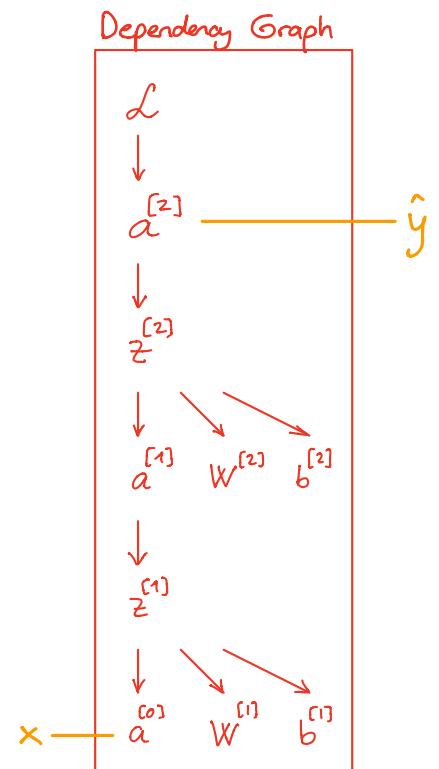
$$\mathcal{L} = - \left(y \cdot \log(a) + (1-y) \cdot \log(1-a) \right)$$

$$a^{[2]} = \sigma(z^{[2]})$$

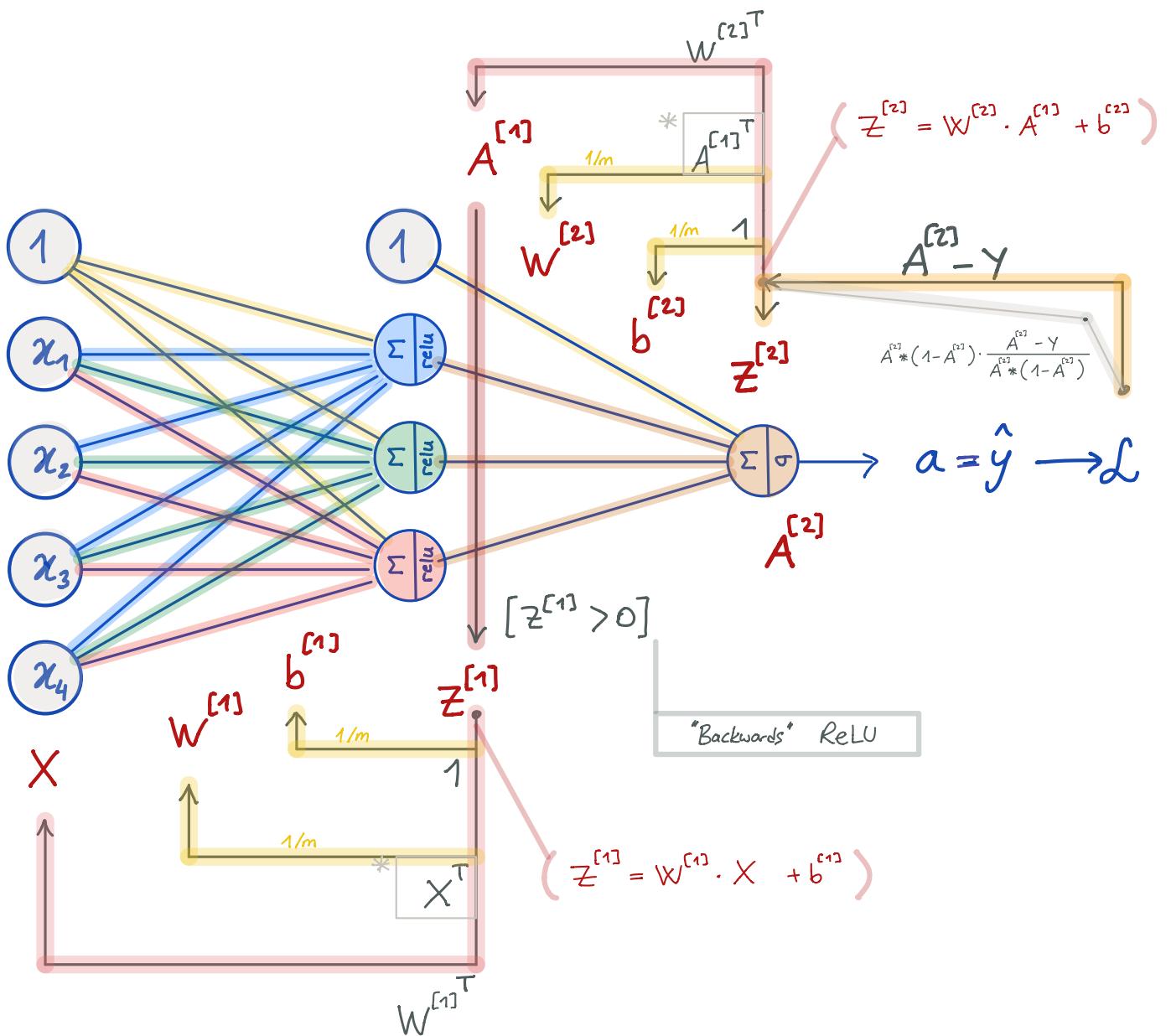
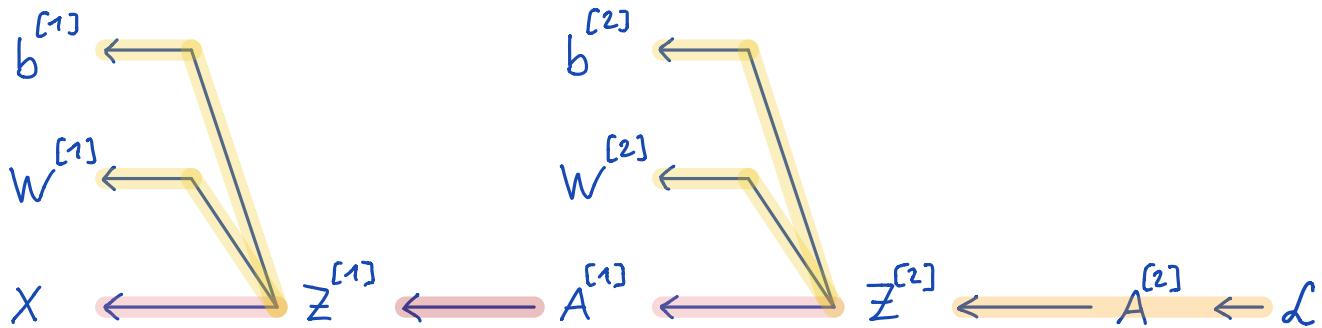
$$z^{[2]} = W^{[2]} \cdot a^{[1]} + b^{[2]}$$

$$a^{[1]} = \text{ReLU}(z^{[1]})$$

$$z^{[1]} = W^{[1]} \cdot a^{[0]} + b^{[1]}$$



BACKWARD PASS



* : The right term to be multiplied with $\frac{\partial \mathcal{L}}{\partial Z^{[2]}}$, but not $\frac{\partial \mathcal{L}}{\partial W}$! See notes below!

THE DERIVATIVES

① $\frac{\partial \mathcal{L}}{\partial a}$ for

$$\mathcal{L}(a, y) = -y \cdot \log_e(a) - (1-y) \cdot \log(1-a)$$

- $\frac{\partial \mathcal{L}}{\partial a} = \frac{-y}{a} - \frac{(1-y)}{(1-a)} \cdot (-1) = \frac{-y(1-a) + a(1-y)}{a(1-a)} =$
- $= \frac{-y + ya + a - ay}{a(1-a)} = \frac{a - y}{a(1-a)}$

② $\frac{da}{dz}$ for

$$a = \sigma(z) = \frac{1}{1+e^{-z}} \Rightarrow \frac{1}{a} = 1+e^{-z}$$

- $\frac{da}{dz} = \frac{d}{dz} (1+e^{-z})^{-1} = -1 \cdot (1-e^{-z})^{-2} \cdot (-e^{-z})$

$$= -1 \cdot a^2 \cdot \left(1 - \frac{1}{a}\right)$$

$$= -a^2 \cdot \left(\frac{a-1}{a}\right) = a(1-a)$$

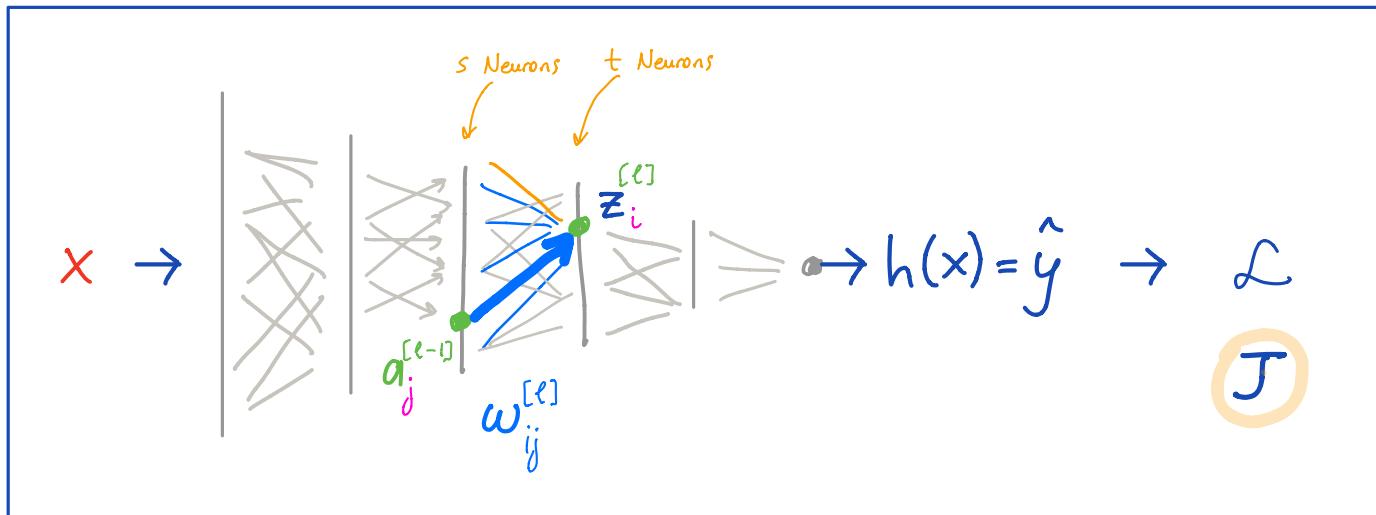
- $\frac{\partial \mathcal{L}}{\partial z} = \frac{\partial \mathcal{L}}{\partial a} \cdot \frac{\partial a}{\partial z} = \frac{a-y}{a(1-a)} \cdot \cancel{a(1-a)} = a-y$

- $\frac{\partial \mathcal{L}}{\partial z^{[2]}} = A^{[2]} - Y$

This holds, as long as
 $A^{[2]} = \sigma(z^{[2]})$

$$= [\text{green dots}] - [\text{green dots}] = [\text{yellow dots}]$$

③ $\frac{\partial z}{\partial a}$ for $z = Wa + b$



$$i \rightarrow \begin{bmatrix} w_{in} & \dots & w_{ij} & \dots & w_{is} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{in} & \dots & w_{ij} & \dots & w_{is} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{ti} & \dots & w_{tj} & \dots & w_{ts} \end{bmatrix} \cdot \begin{bmatrix} a_1^{(l-1)} \\ a_j^{(l-1)} \\ a_s^{(l-1)} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_i \\ b_t \end{bmatrix} = \begin{bmatrix} z_1^{(l)} \\ z_i^{(l)} \\ z_t^{(l)} \end{bmatrix}$$

$$W^{(l)} \cdot a^{(l-1)} + b^{(l)} = z^{(l)}$$

$$\frac{\partial z_i^{(l)}}{\partial a_j^{(l-1)}} = w_{ij}^{(l)}$$

$$\begin{aligned} z_1^{(l)} &= w_{in}^{(l)} \cdot a_1^{(l-1)} + \dots + w_{ij}^{(l)} \cdot a_j^{(l-1)} + \dots + w_{ts}^{(l)} \cdot a_s^{(l-1)} + b_1^{(l)} \\ &\vdots \\ z_i^{(l)} &= w_{in}^{(l)} \cdot a_1^{(l-1)} + \dots + w_{ij}^{(l)} \cdot a_j^{(l-1)} + \dots + w_{ts}^{(l)} \cdot a_s^{(l-1)} + b_i^{(l)} \\ &\vdots \\ z_t^{(l)} &= w_{in}^{(l)} \cdot a_1^{(l-1)} + \dots + w_{ij}^{(l)} \cdot a_j^{(l-1)} + \dots + w_{ts}^{(l)} \cdot a_s^{(l-1)} + b_t^{(l)} \end{aligned}$$

- Two Layout Options for $\frac{\partial \bar{z}}{\partial \bar{a}}$:

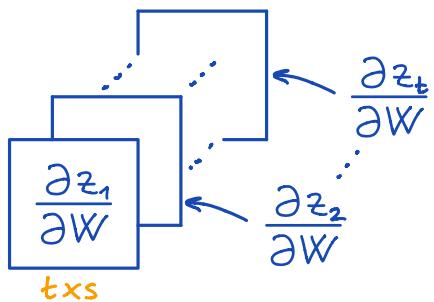
- Numerator Layout

- Denominator Layout

$$\circ \frac{\partial z}{\partial a} = \begin{bmatrix} \frac{\partial z_1}{\partial a_1} & \frac{\partial z_2}{\partial a_1} & \frac{\partial z_t}{\partial a_1} \\ \frac{\partial z_1}{\partial a_2} & \frac{\partial z_2}{\partial a_2} & \frac{\partial z_t}{\partial a_2} \\ \vdots & \vdots & \vdots \\ \frac{\partial z_1}{\partial a_s} & \frac{\partial z_2}{\partial a_s} & \frac{\partial z_t}{\partial a_s} \end{bmatrix}_{(s \times t)} = W^{(c)T}$$

(Hessian Formulation
or
"Denominator Layout")

④ $\frac{\partial z}{\partial W}$ for $z = Wa + b$



Derivative of a vector
with respect to a matrix:
3D - Tensor $(t \times s \times t)$

	$= \frac{\partial z}{\partial W} = (- \quad \dots \quad -)$
\uparrow	$= (- \quad a^T \quad -)$

o $dW := \frac{\partial L}{\partial W}$ is what we actually need

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial W}$$

↑

[for weight update
 $W := W - \alpha \cdot dW$]

With a little trick we can calculate these derivatives more easily (without explicitly calculating $\partial z / \partial W$).

- IDEA : Calculate the gradients for W row-by-row.

Note : The first row of W affects only z_1 !

Question : What if layer l did contain only the cell z_1 ?

Let $w_{1.}$ denote the first row of W .

$$\frac{\partial \mathcal{L}}{\partial w_{1.}} = dW_{1.} = \frac{\partial \mathcal{L}}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_{1.}} = \frac{\partial \mathcal{L}}{\partial z_1} \cdot a^T \quad \frac{\partial z_2}{\partial w_{1.}} = 0$$

$$\frac{\partial \mathcal{L}}{\partial w_{2.}} = dW_{2.} = \frac{\partial \mathcal{L}}{\partial z_2} \cdot \frac{\partial z_2}{\partial w_{2.}} = \frac{\partial \mathcal{L}}{\partial z_2} \cdot a^T$$

:

$$\frac{\partial \mathcal{L}}{\partial w_t.} = dW_t. = \frac{\partial \mathcal{L}}{\partial z_t} \cdot \frac{\partial z_t}{\partial w_t.} = \frac{\partial \mathcal{L}}{\partial z_t} \cdot a^T$$

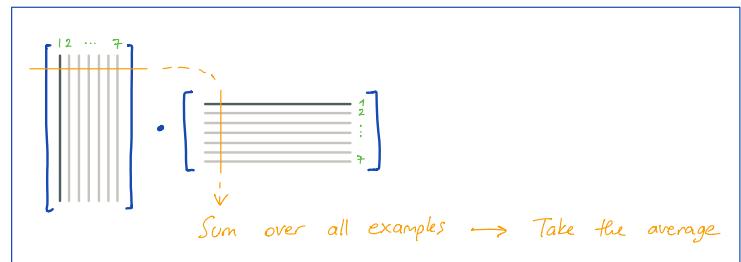
$$\Rightarrow dW = \begin{bmatrix} dz_1 \cdot a^T \\ dz_2 \cdot a^T \\ \vdots \\ dz_t \cdot a^T \end{bmatrix} = \begin{bmatrix} dz_1 \\ dz_2 \\ \vdots \\ dz_t \end{bmatrix} \cdot [\quad a^T \quad]$$

$$dW = dz \cdot a^T$$

- Question: What happens, if we have multiple examples ? ($m > 1$)

$$dW = \frac{1}{m} dZ \cdot A^T$$

[ℓ] [ℓ] [$\ell-1$]



$$\textcircled{4} \quad \frac{dA}{dZ} \quad \text{for } A = \text{relu}(Z)$$

$$\circ \frac{dA}{dZ} = [z > 0] = [\text{True if } (z > 0) \text{ else False}] \\ = [1 \text{ if } (z > 0) \text{ else } 0]$$

EXP.

$$\begin{array}{c}
 z \downarrow \\
 \boxed{z = \begin{bmatrix} 5 & -7 & 6 \\ 2 & 1 & -4 \end{bmatrix}} \quad \boxed{dZ = \begin{bmatrix} 0.5 & 0 & 0.7 \\ 0.1 & -0.6 & 0 \end{bmatrix}} \\
 \downarrow (\text{relu}) \quad * [z > 0] \uparrow (\text{backwards relu}) \\
 A = \begin{bmatrix} 5 & 0 & 6 \\ 2 & 1 & 0 \end{bmatrix} \quad dA = \begin{bmatrix} 0.5 & 0.05 & 0.7 \\ 0.1 & -0.6 & 0.8 \end{bmatrix}
 \end{array}$$

$$A \downarrow \quad dA = \frac{\partial L}{\partial A}$$

$$\Rightarrow \frac{dA}{dZ} = [z > 0] = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

L

THE BACKPROP MAP

$$dA^{[2]} := \frac{\partial L}{\partial A^{[2]}} = \frac{A^{[2]} - Y}{A^{[2]} * (1 - A^{[2]})} = -Y/A^{[2]} + (1-Y)/(1-A^{[2]})$$

↑ element-wise division

$$dZ^{[2]} := \frac{\partial L}{\partial Z^{[2]}} = dA^{[2]} * A^{[2]} * (1 - A^{[2]}) = A^{[2]} - Y$$

$$dA^{[1]} = W^{[2]T} \cdot dZ^{[2]}$$

$(3 \times 7) = (3 \times 1) \cdot (1 \times 7)$

$$dW^{[2]} = \frac{1}{m} \cdot dZ^{[2]} \cdot A^{[1]T}$$

$(1 \times 3) = (1 \times 7) \circ (7 \times 3)$

$$db^{[2]} = \frac{1}{m} \cdot \text{np.sum}(dZ^{[2]})$$

$(1 \times 1) = \boxed{\text{axis}=1} \quad (1 \times 7)$

$$dZ^{[1]} := \frac{\partial L}{\partial Z^{[1]}} = dA^{[1]} * [Z^{[1]} > 0] \quad (\text{element-wise})$$

(3×7)

$$dX = W^{[1]T} \cdot dZ^{[1]}$$

$(4 \times 7) = (4 \times 3) \cdot (3 \times 7)$

$$dW^{[1]} = \frac{1}{m} \cdot dZ^{[1]} \cdot X^T$$

$(3 \times 4) = (3 \times 7) \circ (7 \times 4)$

$$db^{[1]} = \frac{1}{m} \cdot \text{np.sum}(dZ^{[1]})$$

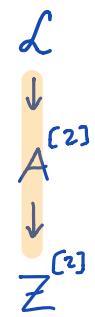
$(3 \times 1) = \boxed{\text{axis}=1} \quad (3 \times 7)$

* The right term to be multiplied with $dZ^{[2]}$,
but not $\partial Z / \partial W$! See notes below!

THE BACKPROP MAP (v2)

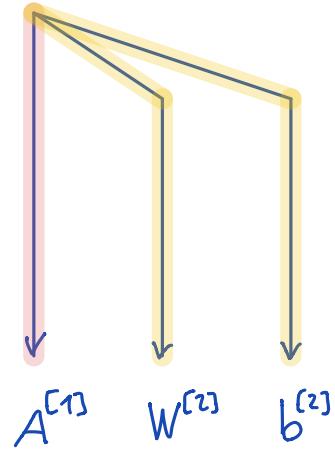
Dependency Graph

$$\frac{\partial \mathcal{L}}{\partial z^{[2]}} = A^{[2]} - Y = [●] - [●]$$



$$\begin{array}{c} \frac{\partial \mathcal{L}}{\partial z^{[2]}} = W^{[2]T} \\ \frac{\partial z^{[2]}}{\partial A^{[1]}} = W^{[2]} \\ \frac{\partial z^{[2]}}{\partial W^{[2]}} = A^{[1]T} \\ \frac{\partial z^{[2]}}{\partial b^{[2]}} = 1 \end{array}$$

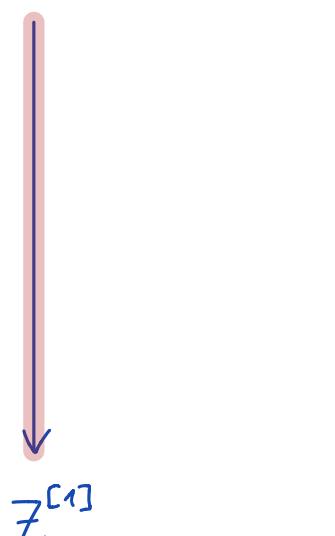
$z^{[2]} = W^{[2]T} \cdot a^{[1]} + b^{[2]} = [●] \cdot [●, ●, ●] + [●] = [●]$



$$\frac{\partial A^{[1]}}{\partial z^{[1]}} = [z^{[1]} > 0] = \begin{bmatrix} 1 & \text{if } \textcolor{blue}{●} > 0 & \text{else } 0 \\ 1 & \text{if } \textcolor{green}{●} > 0 & \text{else } 0 \\ 1 & \text{if } \textcolor{red}{●} > 0 & \text{else } 0 \end{bmatrix}$$

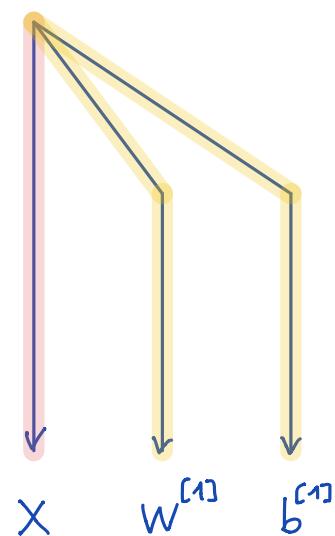
(ternary operator in Python)

$a^{[1]} = \text{relu}(z^{[1]}) = \begin{bmatrix} z & \text{if } \textcolor{blue}{●} > 0 & \text{else } 0 \\ z & \text{if } \textcolor{green}{●} > 0 & \text{else } 0 \\ z & \text{if } \textcolor{red}{●} > 0 & \text{else } 0 \end{bmatrix} = \max(z, 0)$



$$\begin{array}{c} \frac{\partial \mathcal{L}}{\partial z^{[1]}} = W^{[1]T} \\ \frac{\partial z^{[1]}}{\partial X} = X^T \\ \frac{\partial z^{[1]}}{\partial W^{[1]}} = 1 \\ \frac{\partial z^{[1]}}{\partial b^{[1]}} = 1 \end{array}$$

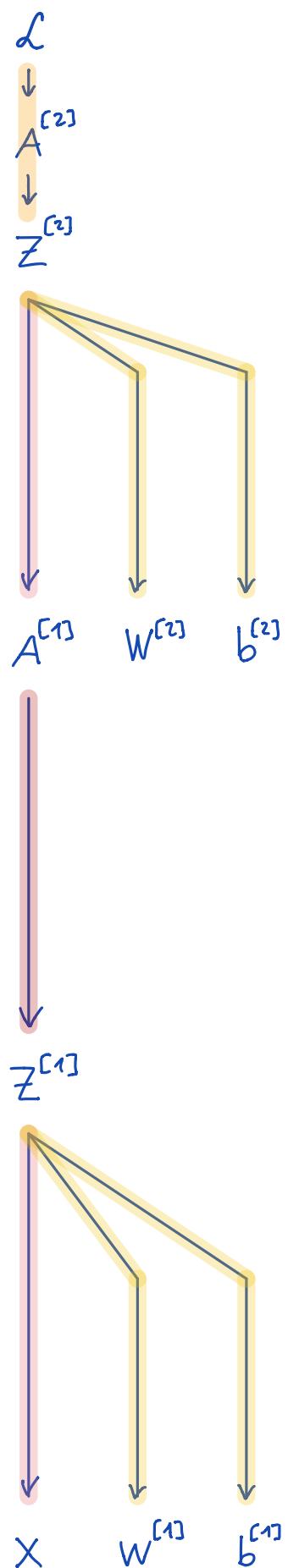
$z^{[1]} = W^{[1]T} \cdot X + b^{[1]} = [\textcolor{blue}{●}, \textcolor{green}{●}, \textcolor{red}{●}] \cdot [\textcolor{lightblue}{●}, \textcolor{lightgreen}{●}, \textcolor{lightred}{●}] + [\textcolor{yellow}{●}] = [●, ●, ●]$



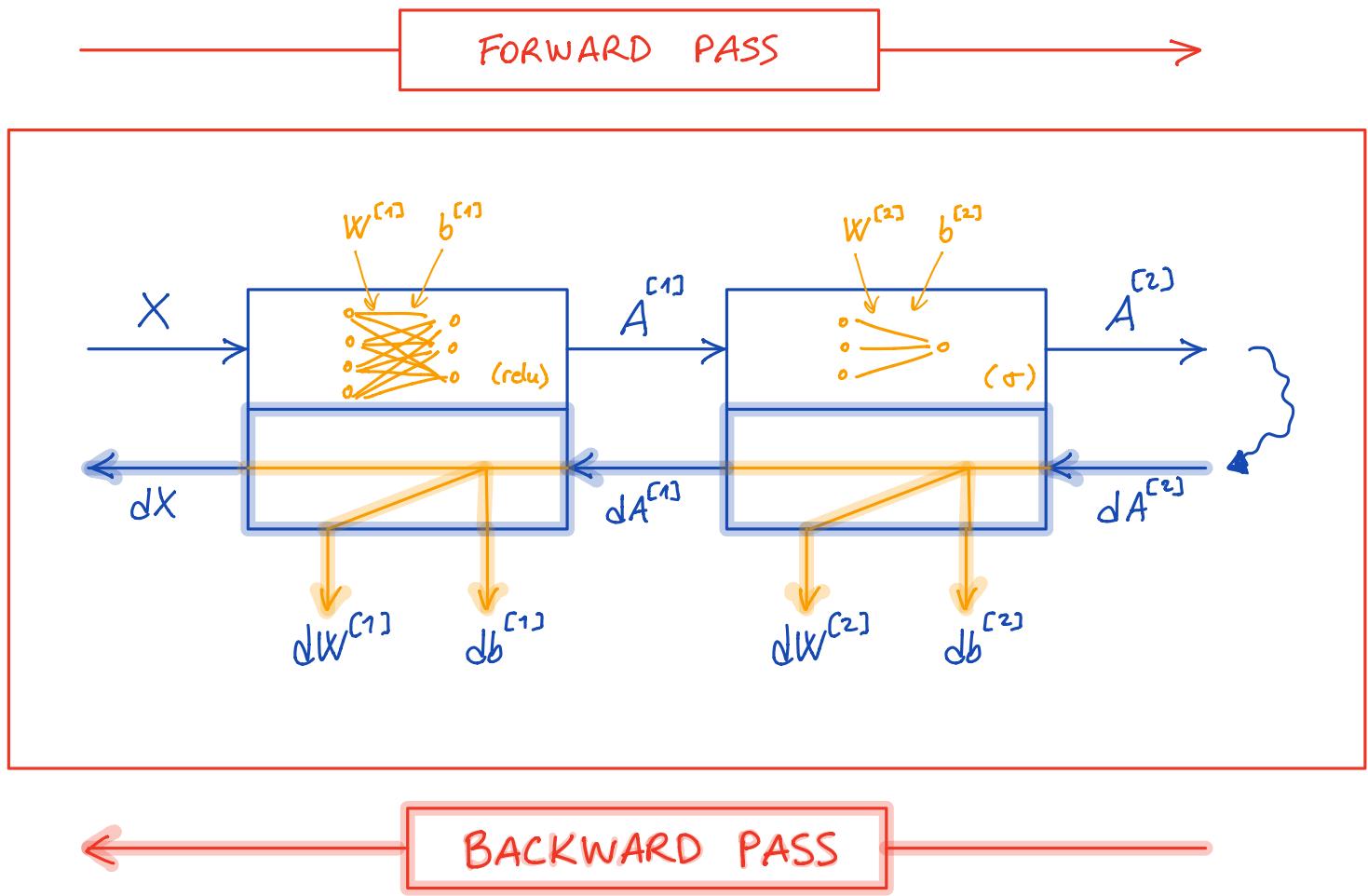
THE BACKPROP MAP (v3)

Dependency Graph

		$A^{[2]} - Y$
[] (1×7)		
	$A^{[1]T}$ (7×3)	1 (1×1)
$\frac{\partial L}{\partial A^{[1]}} = W^{[2]T} \cdot (A^{[2]} - Y)$ (3×7)	$\frac{\partial J}{\partial W^{[2]}} = \frac{1}{m} (A^{[2]} - Y) \cdot A^{[1]T}$ (1×3)	$\frac{\partial J}{\partial b^{[2]}} = \frac{1}{m} \cdot np \cdot \text{sum}(A^{[2]} - Y)$ (1×1)
$\frac{\partial A^{[1]}}{\partial Z^{[1]}} = [Z^{[1]} > 0] = \begin{bmatrix} 1/0 & \\ 1/0 & \\ 1/0 & \end{bmatrix}$ (3×7)		
$\frac{\partial L}{\partial Z^{[1]}} = W^{[2]T} \cdot (A^{[2]} - Y) * [Z^{[1]} > 0]$ (3×7)		(elementwise)
	X^T (7×4)	1 (1×1)
$\frac{\partial J}{\partial X} = W^{[1]T} \cdot \frac{\partial L}{\partial Z^{[1]}}$ (4×7)	$\frac{\partial J}{\partial W^{[1]}} = \frac{1}{m} \left(\frac{\partial L}{\partial Z^{[1]}} \cdot X^T \right)$ (3×4)	$\frac{\partial J}{\partial b^{[1]}} = \frac{1}{m} \cdot np \cdot \text{sum} \left(\frac{\partial L}{\partial Z^{[1]}} \right)$ (3×1)



• PROCESS OVERVIEW



WEIGHT UPDATES

$$\begin{bmatrix} W^{[1]} \\ b^{[1]} \\ W^{[2]} \\ b^{[2]} \end{bmatrix}_{\text{new}} := \begin{bmatrix} W^{[1]} \\ b^{[1]} \\ W^{[2]} \\ b^{[2]} \end{bmatrix}_{\text{old}} - \alpha \cdot \begin{bmatrix} dW^{[1]} \\ db^{[1]} \\ dW^{[2]} \\ db^{[2]} \end{bmatrix}$$