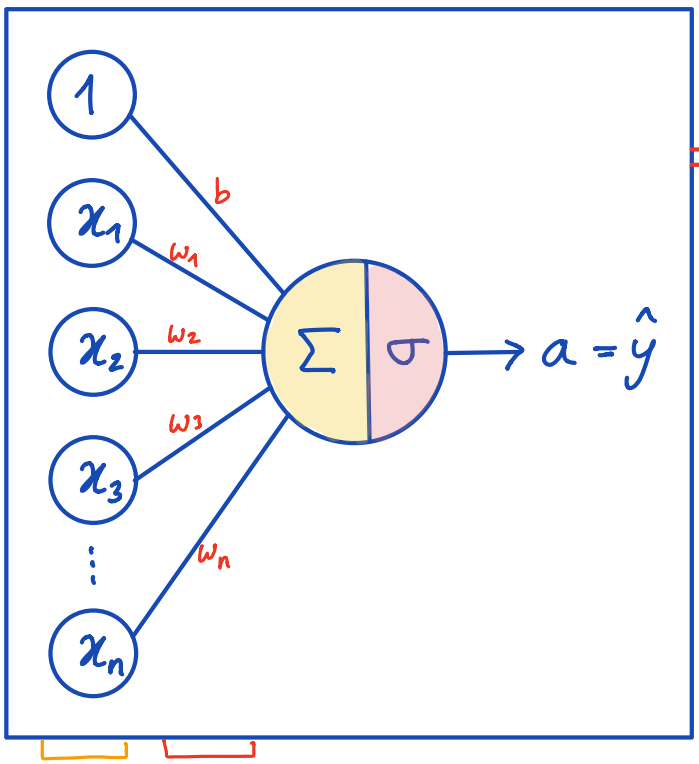
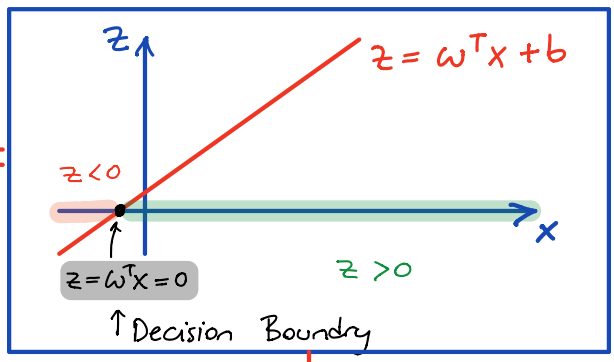
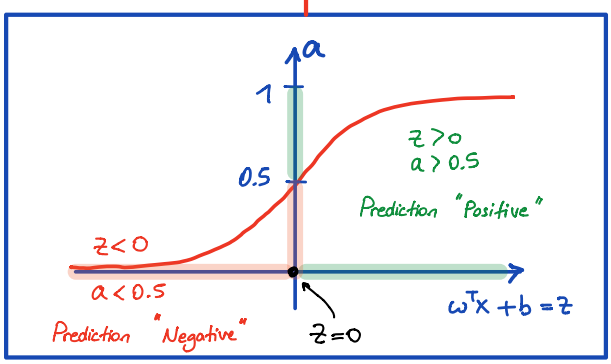
BINARY CLASSIFIKATION WITH NN







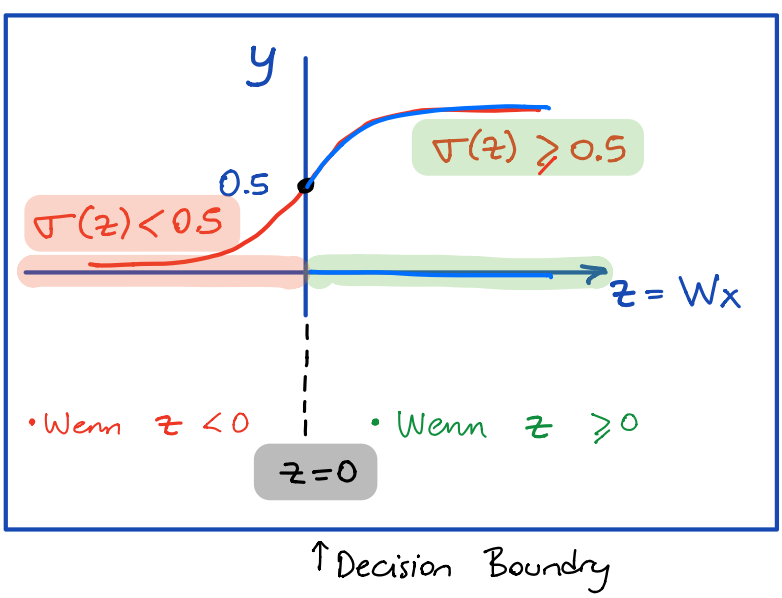
x w

$w^T = W$ x

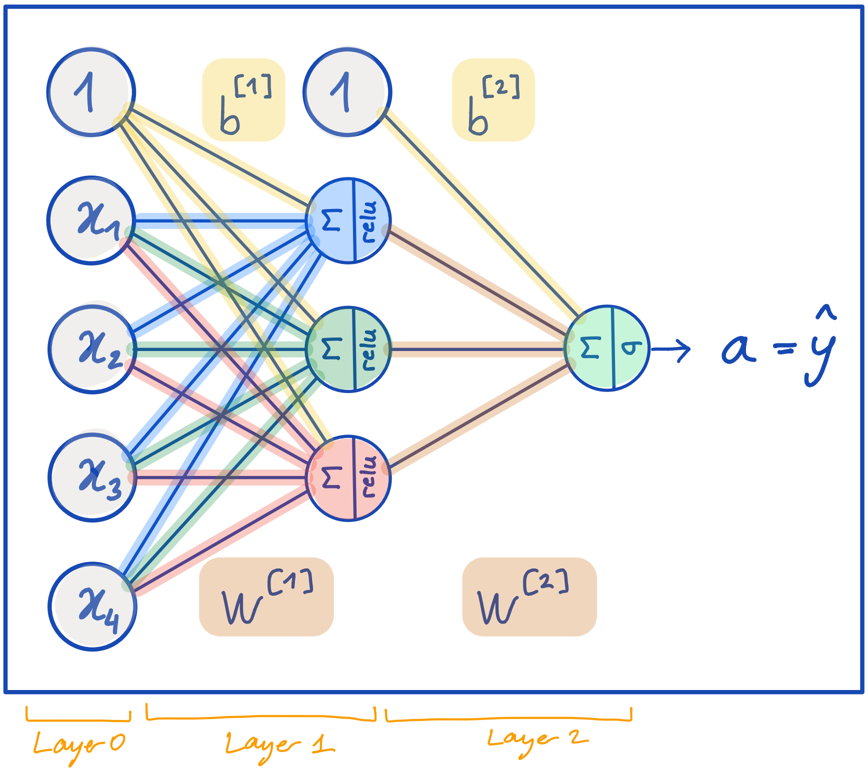
$$a = \sigma\left(\begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + b\right) = \sigma(Wx + b) = \sigma(z)$$

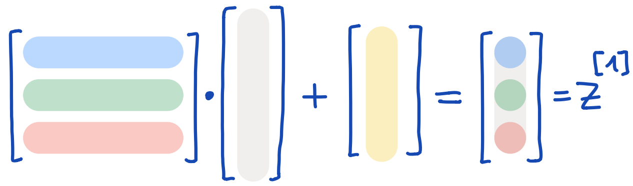
- The Sigmoid Function σ

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

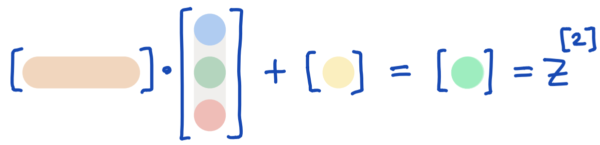


BINARY CLASSIFIKATION WITH DNN (Ng notation)

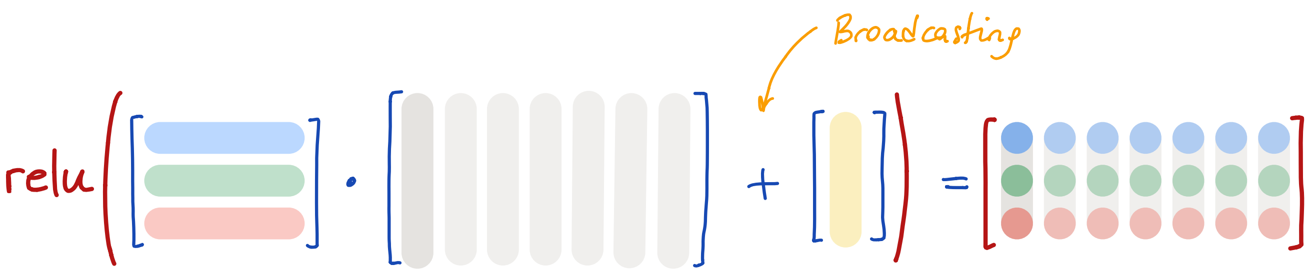




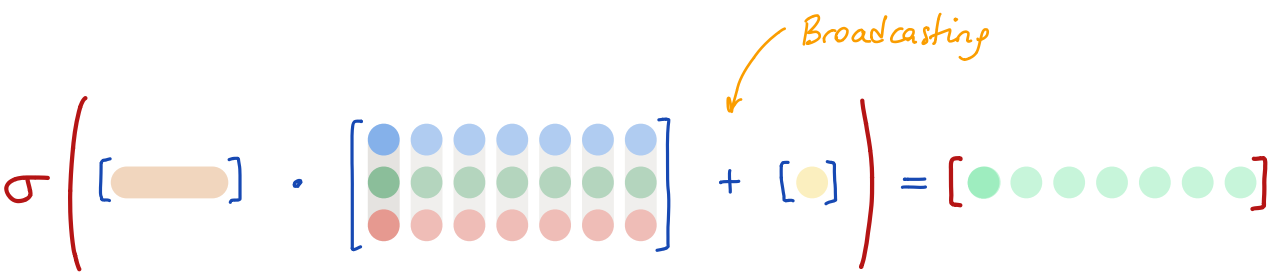
$$\mathrm{relu}\left(W^{[1]} \cdot x + b^{[1]}\right) = a^{[1]}$$



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



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



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

CROSS-ENTROPY LOSS

$$A^{[2]} = \hat{Y} = [\bullet\bullet\bullet\bullet\bullet\bullet\bullet]$$

$$Y = [\bullet\bullet\bullet\bullet\bullet\bullet\bullet]$$

$$\text{LOSS} = \text{DISTANCE}(\hat{Y}, Y) = [\overset{(1)\ (2)\ (3)\ \cdots\ (6)\ (7)}{\bullet\bullet\bullet\bullet\bullet\bullet\bullet}]$$

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

- $\mathcal{L}\left(\overset{(1)\ (2)\ (3)\ \cdots\ (6)\ (7)}{\bullet\bullet\bullet\bullet\bullet\bullet\bullet}, \overset{(1)\ (2)\ (3)\ \cdots\ (6)\ (7)}{\bullet\bullet\bullet\bullet\bullet\bullet\bullet}\right)$

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

$$= [\overset{(1)\ (2)\ (3)\ \cdots\ (6)\ (7)}{\bullet\bullet\bullet\bullet\bullet\bullet\bullet}]$$

- elementwise

$$Y * \log(\hat{Y}) = [\bullet\bullet\bullet\bullet\bullet\bullet\bullet] * \log([\bullet\bullet\bullet\bullet\bullet\bullet\bullet])$$

COST FUNCTION (average loss)

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

$$= \frac{1}{7} \cdot np.sum\left([\bullet\bullet\bullet\bullet\bullet\bullet\bullet]\right)$$

Sum over all Data points

(MINI BATCH GRADIENT DESCENT)

THE GRADIENT

$$\theta = [\quad]^T \quad \text{(Parameter)}$$

$$\frac{\partial J}{\partial \theta} =: d\theta = [\quad]^T \quad \text{(evaluate for current } \omega\text{)}$$

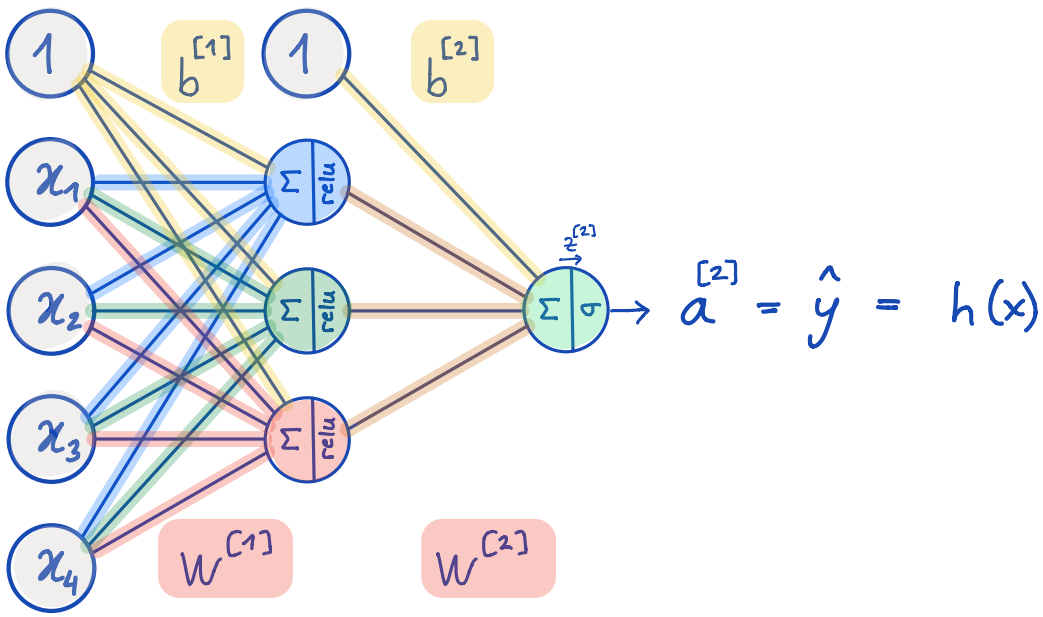
$$\nabla J = \left[\frac{\partial J}{\partial w_{11}^{[1]}} \cdots \frac{\partial J}{\partial w_{34}^{[1]}} \quad \frac{\partial J}{\partial b_1^{[1]}} \cdots \frac{\partial J}{\partial b_3^{[1]}} \quad \frac{\partial J}{\partial w_{11}^{[2]}} \cdots \frac{\partial J}{\partial w_{13}^{[2]}} \quad \frac{\partial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

Direction of steepest Ascent !

(Go in opposite direction) !

• Backpropagation – Multiple Layers

2 – Layer Network (Graph Representation)



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

(Output as a function of x)

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

$$\mathcal{L} = -\left(y \cdot \log(a) + (1-y) \cdot \log(1-a) \right) \qquad a = a^{[2]}$$

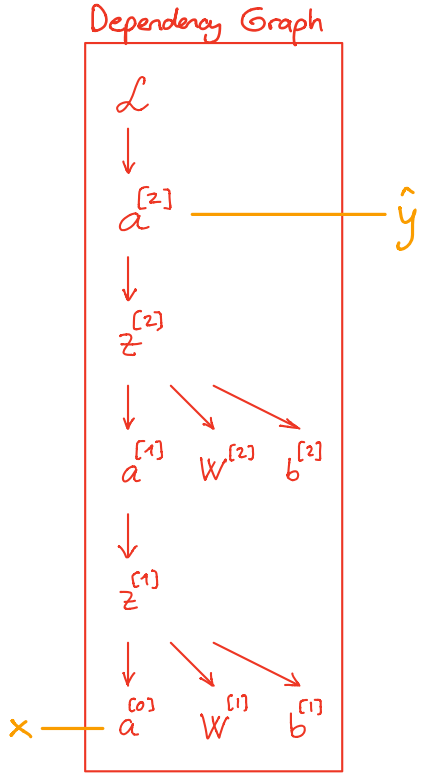
$$a^{[2]} = \sigma(z^{[2]})$$

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

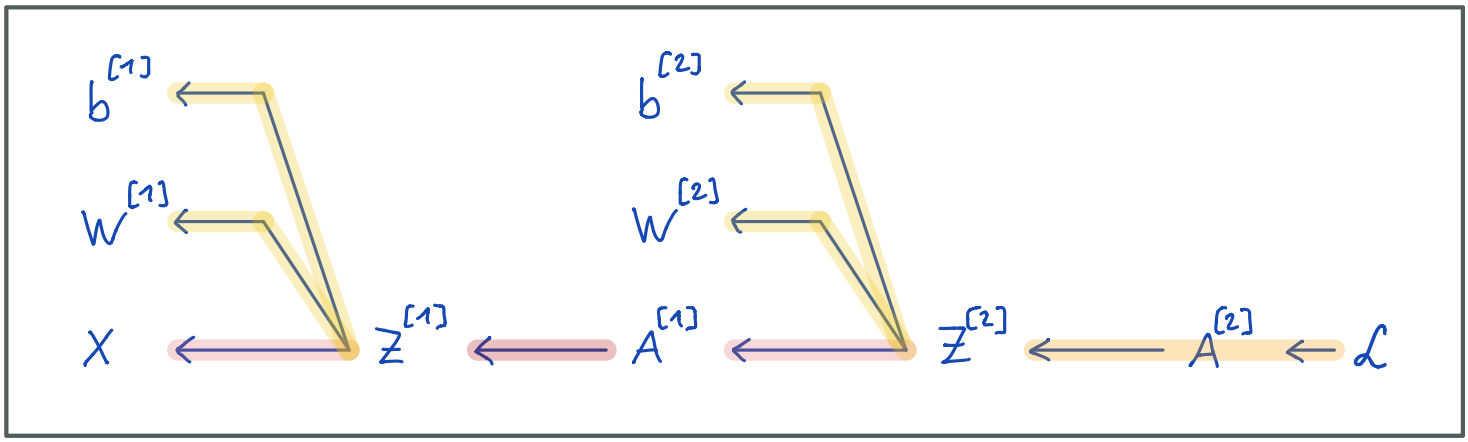
$$a^{[1]} = \text{ReLU}(z^{[1]})$$

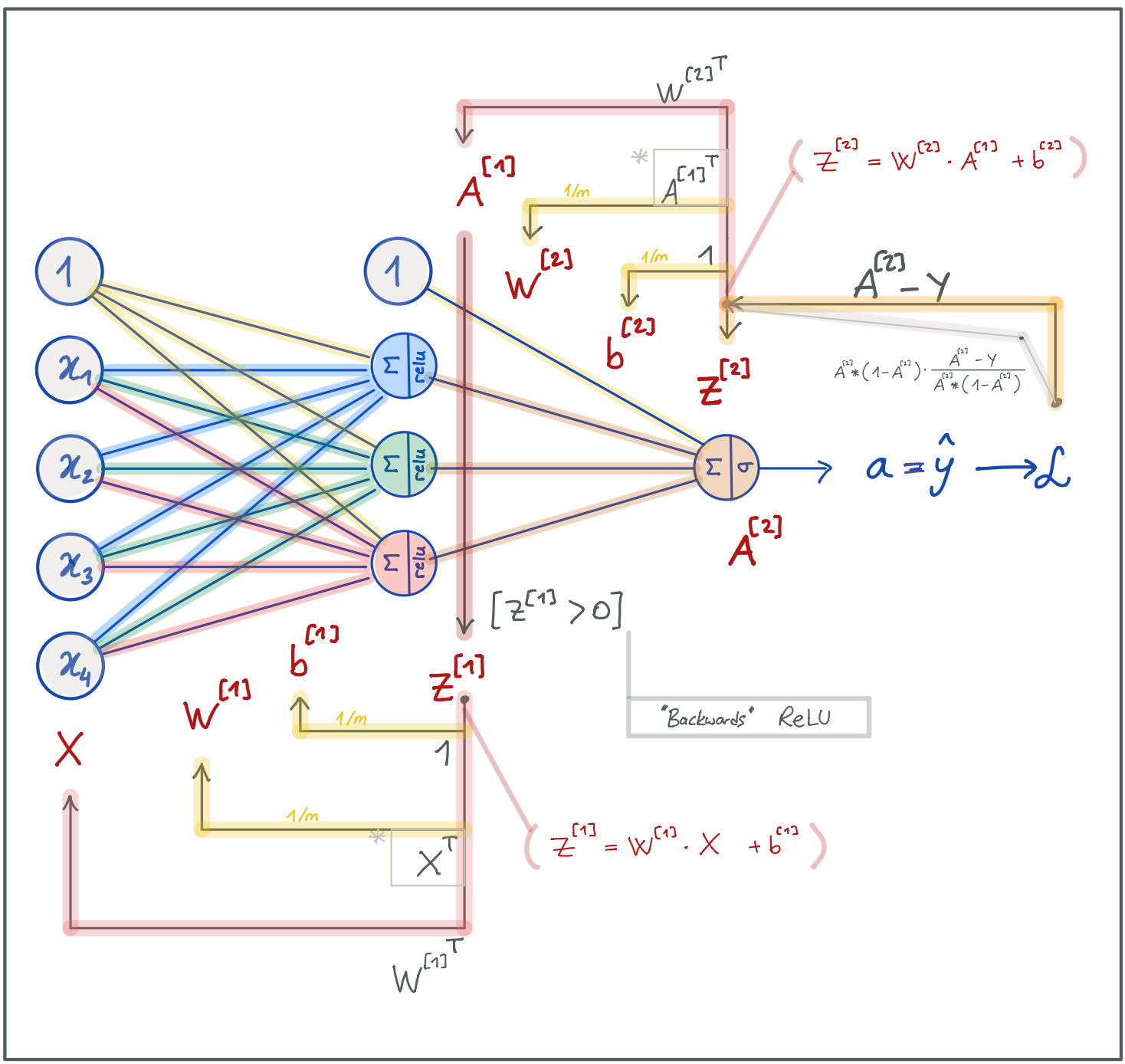
$$z^{[1]} = W^{[1]} \cdot a^{[0]} + b^{[1]}$$

Dependency Graph



BACKWARD PASS





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

canan.yildiz@tau.edu.tr

THE DERIVATIVES

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

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

$$= \frac{-y + \cancel{ya} + a - \cancel{ay}}{a(1-a)} = \frac{a-y}{a(1-a)}$$

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

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

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

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

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

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

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

$$= [\bullet\bullet\bullet\bullet\bullet\bullet] - [\bullet\bullet\bullet\bullet\bullet\bullet] = [\bullet\bullet\bullet\bullet\bullet\bullet]$$

canan.yildiz@tau.edu.tr

THE BACKPROP MAP

$\mathcal{L}$

$A^{[2]}$

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

(1×7)

element-wise division

$A^{[2]} * (1 - A^{[2]})$

$Z^{[2]}$

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

$W^{[2]^T}$ $\quad$ $A^{[1]^T}$ * $\quad$ 1

$Z^{[2]} = W^{[2]} \cdot A^{[1]} + b^{[2]}$

$A^{[1]}$

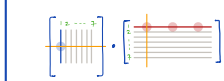
$$dA^{[1]} = W^{[2]^T} \cdot dZ^{[2]}$$

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

$W^{[2]}$

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

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



$b^{[2]}$

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

$(1 \times 1) =$ (1×7) axis = 1

$[Z^{[1]} > 0]$

$Z^{[1]}$

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

(3×7) $\quad$ (3×7)

$W^{[1]^T}$ $\quad$ X^T * $\quad$ 1

X

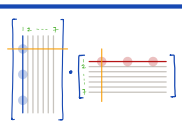
$$dX = W^{[1]^T} \cdot dZ^{[1]}$$

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

$W^{[1]}$

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

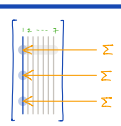
$(3 \times 4) = (3 \times 7) \cdot (7 \times 4)$



$b^{[1]}$

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

$(3 \times 1) =$ (3×7) axis = 1

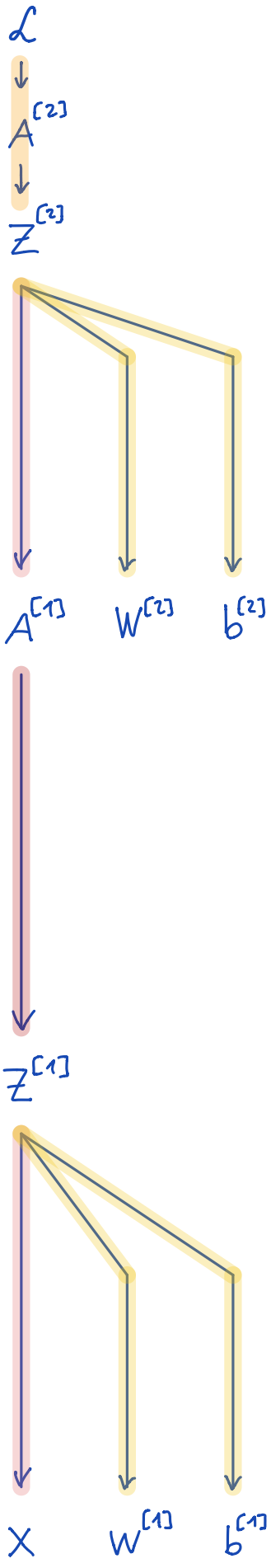


* The right term to be multiplied with $dZ^{[2]}$, but not $\partial Z / \partial W$! See notes below!

canan.yildiz@tau.edu.tr

THE BACKPROP MAP (v2)

Dependency Graph



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

$$\frac{\partial Z^{[2]}}{\partial A^{[1]}} = W^{[2]T} \qquad \frac{\partial Z^{[2]}}{\partial W^{[2]}} = A^{[1]T} \qquad \frac{\partial Z^{[2]}}{\partial b^{[2]}} = 1$$

$$Z^{[2]} = W^{[2]} \cdot a^{[1]} + b^{[2]} = [\quad] \cdot \begin{bmatrix} \bullet \\ \bullet \\ \bullet \end{bmatrix} + [\bullet] = [\bullet]$$

(ternary operator in Python)

$$\frac{\partial A^{[1]}}{\partial Z^{[1]}} = [Z^{[1]} > 0] = \begin{bmatrix} 1 \text{ if } \bullet > 0 \text{ else } 0 \\ 1 \text{ if } \bullet > 0 \text{ else } 0 \\ 1 \text{ if } \bullet > 0 \text{ else } 0 \end{bmatrix}$$

$$a^{[1]} = \text{relu}(Z^{[1]}) = \begin{bmatrix} z \text{ if } \bullet > 0 \text{ else } 0 \\ z \text{ if } \bullet > 0 \text{ else } 0 \\ z \text{ if } \bullet > 0 \text{ else } 0 \end{bmatrix} = \max(z, 0)$$

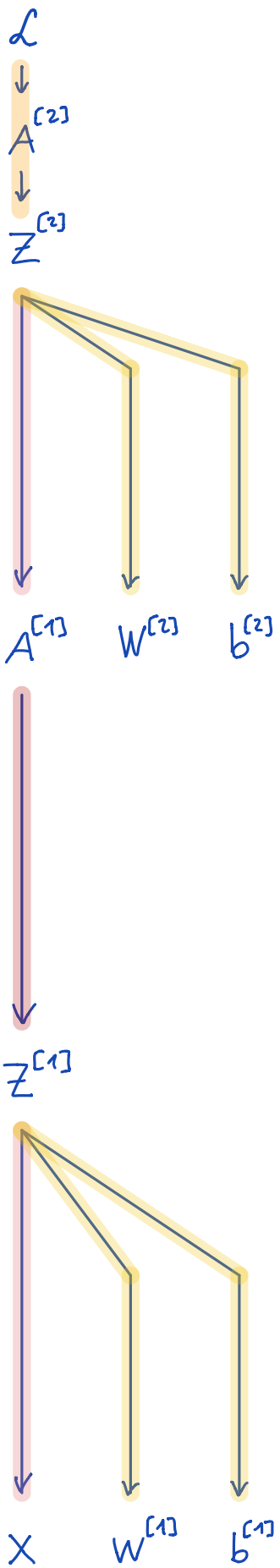
$$\frac{\partial Z^{[1]}}{\partial X} = W^{[1]T} \qquad \frac{\partial Z^{[1]}}{\partial W^{[1]}} = X^T \qquad \frac{\partial Z^{[1]}}{\partial b^{[1]}} = 1$$

$$Z^{[1]} = W^{[1]} \cdot X + b^{[1]} = \begin{bmatrix} \quad \\ \quad \\ \quad \end{bmatrix} \cdot \begin{bmatrix} \quad \\ \quad \end{bmatrix} + \begin{bmatrix} \quad \\ \quad \\ \quad \end{bmatrix} = \begin{bmatrix} \bullet \\ \bullet \\ \bullet \end{bmatrix}$$

canan.yildiz@tau.edu.tr

THE BACKPROP MAP (v3)

Dependency Graph



$A^{[2]} - Y$

(1x7)

$W^{[2]T}$ (3x1)	$A^{[1]T}$ (7x3)	1 (1x1)
$\frac{\partial \mathcal{L}}{\partial A^{[1]}} = W^{[2]T} \cdot (A^{[2]} - Y)$ (3x7)	$\frac{\partial J}{\partial W^{[2]}} = \frac{1}{m}(A^{[2]} - Y) \cdot A^{[1]T}$ (1x3)	$\frac{\partial J}{\partial b^{[2]}} = \frac{1}{m} \cdot np.sum(A^{[2]} - Y)$ (1x1)

$$\frac{\partial A^{[1]}}{\partial Z^{[1]}} = [Z^{[1]} > 0] = \begin{bmatrix} 1/0 \\ 1/0 \\ 1/0 \end{bmatrix}$$

(3x7)

$$\frac{\partial \mathcal{L}}{\partial Z^{[1]}} = W^{[2]T} \cdot (A^{[2]} - Y) * [Z^{[1]} > 0]$$

(elementwise)

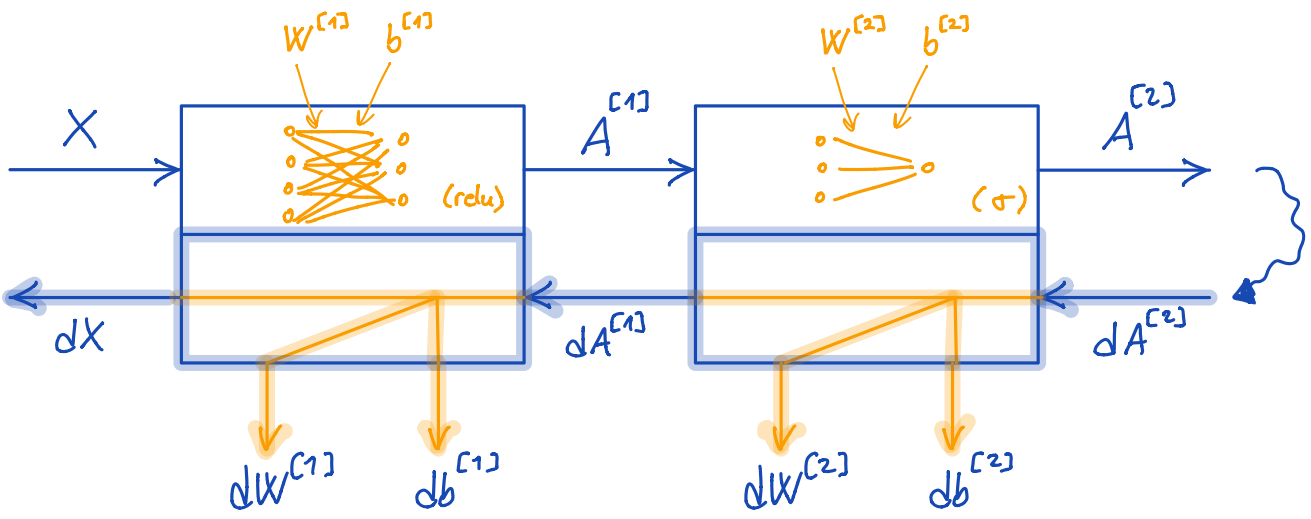
(3x7)

$W^{[1]T}$ (4x3)	X^T (7x4)	1 (1x1)
$\frac{\partial J}{\partial X} = W^{[1]T} \cdot \frac{\partial \mathcal{L}}{\partial Z^{[1]}}$ (4x7)	$\frac{\partial J}{\partial W^{[1]}} = \frac{1}{m}\left(\frac{\partial \mathcal{L}}{\partial Z^{[1]}} \cdot X^T\right)$ (3x4)	$\frac{\partial J}{\partial b^{[1]}} = \frac{1}{m} \cdot np.sum\left(\frac{\partial \mathcal{L}}{\partial Z^{[1]}}\right)$ (3x1)

canan.yildiz@tau.edu.tr

- PROCESS OVERVIEW

FORWARD PASS



BACKWARD PASS

WEIGHT UPDATES

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