TUTOR LESSON 1 - DEEP LEARNING (BACKPROPAGATION)

Given the neural network described by the following equations:

$$h^{(1)} = W^{(1)} \cdot x + b^{(1)} \in \mathbb{R}^2$$
$$a^{(1)} = \mathrm{ReLU}(h^{(1)}) \in \mathbb{R}^2$$
$$h^{(2)} = (W^{(2)})^T \cdot a^{(1)} + b^{(2)} \in \mathbb{R}$$
$$y = \sigma(h^{(2)}) \in \mathbb{R}$$
$$J = \frac{1}{2}(t-y)^2 \in \mathbb{R}$$

With:

$$x = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}, \quad W^{(1)} = \begin{bmatrix} 0.5 & 0.75 \\ 1 & -1 \end{bmatrix}, \quad b^{(1)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

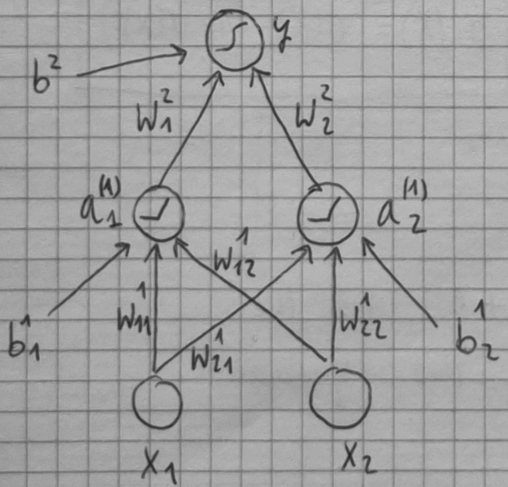
$$W^{(2)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad b^{(2)} = -1 \quad \text{and} \quad t = 1.$$

Compute the value of $\frac{\partial J}{\partial b_2^{(1)}}$, where $b_2^{(1)}$ is the second element of the bias vector $b^{(1)} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \end{bmatrix}$.

Hint: $\sigma(0) = 0.5$.

-NOTATION: h is the PRE-ACTIVATION, a is the POST-ACTIVATION, y is the OUTPUT, t is the TARGET, J is the LOSS FUNCTION, W are the WEIGHTS and b the BIASES.

① DRAW THE NETWORK



② FORWARD PASS

$$h^{(1)} = \begin{bmatrix} h_1^1 \\ h_2^1 \end{bmatrix} = \begin{bmatrix} W_{11}^1 & W_{12}^1 \\ W_{21}^1 & W_{22}^1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} b_1^1 \\ b_1^2 \end{bmatrix} = \begin{bmatrix} W_{11}^1 x_1 + W_{12}^1 x_2 + b_1^1 \\ W_{21}^1 x_1 + W_{22}^1 x_2 + b_2^1 \end{bmatrix}$$

$$= \begin{bmatrix} 1/4 + 3/4 + 0 \\ 1/2 - 1 + 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -1/2 \end{bmatrix}$$

$$a^{(1)} = ReLU(h^1) = \begin{bmatrix} a_1^1 \\ a_2^1 \end{bmatrix} = \begin{bmatrix} ReLU(h_1^1) \\ ReLU(h_2^1) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\boxed{ReLU(X) = \max(0, X)}$$

$$h^{(2)} = \begin{bmatrix} W_1^2 & W_2^2 \end{bmatrix} \begin{bmatrix} a_1^1 \\ a_2^1 \end{bmatrix} + b^2 = W_1^2 a_1^1 + W_2^2 a_2^1 + b^2 = 1 + 0 + (-1) = 0.$$

$$y = \sigma(h^{(2)}) = \sigma(0) = \frac{1}{2}.$$

$$J = \frac{1}{2}(t-y)^2 = \frac{1}{2}\left(1-\frac{1}{2}\right)^2 = \frac{1}{2}\cdot\left(\frac{1}{2}\right)^2 = \frac{1}{2}\cdot\frac{1}{4} = \frac{1}{8}.$$

③ BACKWARD PASS

Remember that we want to compute $\frac{\partial J}{\partial b_2^1}$.

• FIRST WAY:

$$y = \sigma(h^2)$$

$$\mathbb{R} \ni \frac{\partial J}{\partial b_2^1} = \frac{\partial\left[\frac{1}{2}(t-y)^2\right]}{\partial b_2^1} = -(t-y)\cdot\frac{\partial y}{\partial b_2^1} = \underbrace{-(t-y)\cdot y(1-y)}_{= C}\cdot\frac{\partial h^2}{\partial b_2^1}$$

$$= C\cdot\frac{\partial(W_2^2 a_2^1)}{\partial b_2^1} = C\cdot W_2^2\cdot\frac{\partial[ReLU(h_2^1)]}{\partial b_2^1}$$

$$\boxed{\sigma'(X) = \sigma(X)\cdot(1-\sigma(X))}$$

$$= C\cdot W_2^2\cdot\underbrace{ReLU'(h_2^1)}_{= 0}\cdot\frac{\partial h_2^1}{\partial b_2^1} = 0.$$

This comes from the fact that $h^2 = W_1^2 a_1^1 + W_2^2 a_2^1 + b^2$ and the fact that the only term that depends on b_2^1 inside this expression is $W_2^2 a_2^1$ (see the network drawing).

• SECOND WAY

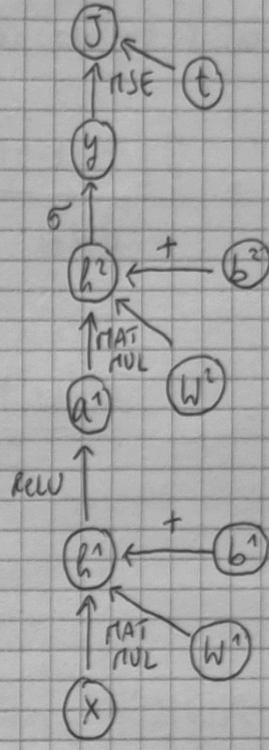
Now we will compute the partial derivative using the COMPUTATIONAL GRAPH, where each node is a variable (scalar, vector, matrix) and the edges encode the operations between nodes.

MATRIX FORM:

$$\mathbb{R}^2 \ni \frac{\partial J}{\partial b^1} = \overset{①}{\underset{\in\,\mathbb{R}}{\frac{\partial J}{\partial y}}} \cdot \overset{②}{\underset{\in\,\mathbb{R}}{\frac{\partial y}{\partial h^2}}} \cdot \overset{③}{\underset{\in\,\mathbb{R}^{1\times 2}}{\frac{\partial h^2}{\partial a^1}}} \cdot \overset{④}{\underset{\in\,\mathbb{R}^{2\times 2}}{\frac{\partial a^1}{\partial h^1}}} \cdot \overset{⑤}{\underset{\in\,\mathbb{R}^{2\times 2}}{\frac{\partial h^1}{\partial b^1}}}$$

$$\|$$

$$\nabla_{\vec{b_1}} J = \left[\frac{\partial J}{\partial b_1^1} \quad \frac{\partial J}{\partial b_2^1}\right]$$



$$\frac{\partial J}{\partial b^1} = \underbrace{-(t-y)}_{①} \cdot \underbrace{y\cdot(1-y)}_{②} \cdot \underbrace{(W^2)^T}_{③} \cdot \underbrace{ReLU'(h^1)}_{④} \cdot \underbrace{I_{2\times 2}}_{⑤}$$

$$= -(t-y)\,y(1-y)\,(W^2)^T \odot \left[ReLU'(h_1^1) \quad ReLU'(h_2^1)\right] \cdot I_{2\times 2}$$

Note:

$$\underset{\in\,\mathbb{R}^{2\times 2}}{\frac{\partial a^1}{\partial h^1}} = \begin{bmatrix} \frac{\partial a_1^1}{\partial h_1^1} & \frac{\partial a_1^1}{\partial h_2^1} \\ \frac{\partial a_2^1}{\partial h_1^1} & \frac{\partial a_2^1}{\partial h_2^1} \end{bmatrix} = \begin{bmatrix} ReLU'(h_1^1) & 0 \\ 0 & ReLU'(h_2^1) \end{bmatrix}$$

$$\mathbb{R}^{2\times 2} \ni \frac{\partial h^1}{\partial b^1} = \begin{bmatrix} \frac{\partial h_1^1}{\partial b_1^1} & \frac{\partial h_1^1}{\partial b_2^1} \\ \frac{\partial h_2^1}{\partial b_1^1} & \frac{\partial h_2^1}{\partial b_2^1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_{2\times 2}$$

$$(*) = -(t-y)\,y\,(1-y)\left[W_1^2\, ReLU'(h_1^1) \quad W_2^2\, ReLU'(h_2^1)\right]$$

because

$$\underset{\in\,\mathbb{R}^{1\times 2}}{\left[W_1^2 \quad W_2^2\right]} \cdot \underset{\in\,\mathbb{R}^{2\times 2}}{\begin{bmatrix} ReLU'(h_1^1) & 0 \\ 0 & ReLU'(h_2^1) \end{bmatrix}} = \underset{\in\,\mathbb{R}^{1\times 2}}{\left[W_1^2\, ReLU'(h_1^1) \quad W_2^2\, ReLU'(h_2^1)\right]}$$

$$= (W^2)^T \odot \underbrace{\left[ReLU'(h_1^1) \quad ReLU'(h_2^1)\right]}_{= \operatorname{diag}\left(\frac{\partial a^1}{\partial h^1}\right)}$$

Now, let's try to compute $\frac{\partial J}{\partial W^1_{11}}$.

$$\frac{\partial J}{\partial W^1_{11}} = \frac{\partial\left[\frac{1}{2}(t-y)^2\right]}{\partial W^1_{11}} \underset{\substack{\uparrow \\ \text{OMIT} \\ W^{(1)}_{11}}}{=} -(t-y)\,\frac{\partial y}{\partial W_{11}} = \overbrace{-(t-y)\,y(1-y)}^{C} \cdot \frac{\partial h^2}{\partial W_{11}}$$

$$= C \cdot \frac{\partial\left((W^2)^T a^1 + b^2\right)}{\partial W_{11}} = C \cdot W_1^2 \cdot \frac{\partial a_1^1}{\partial W_{11}}$$

$$(W^2)^T a^1 = \begin{bmatrix} W_1^2 & W_2^2 \end{bmatrix} \cdot \begin{bmatrix} a_1^1 \\ a_2^1 \end{bmatrix} = W_1^2 a_1^1 + \underbrace{W_2^2 a_2^1}_{\substack{\text{DOES NOT} \\ \text{DEPEND ON} \\ W^1_{11}}}$$

$a_1^1 = \mathrm{ReLU}(h_1^1)$

$$= C \cdot W_1^2 \cdot \mathrm{ReLU}'(h_1^1) \cdot \frac{\partial h_1^1}{\partial W_{11}} = C \cdot W_1^2 \cdot \mathrm{ReLU}'(h_1^1) \cdot x_1$$

$$h_1^1 = \underbrace{W^1_{11} x_1} + W^1_{12} x_2 + b^1_1$$

Substituting the numerical values:

$$\frac{\partial J}{\partial W^1_{11}} = -\left(1-\frac{1}{2}\right) \cdot \frac{1}{2}\left(1-\frac{1}{2}\right) \cdot 1 \cdot 1 \cdot \frac{1}{2} = -\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = -\frac{1}{16}.$$

Now, we compute $\frac{\partial J}{\partial W^1}$ in MATRIX FORM.

$$\underset{\mathbb{R}^{1\times(2\times2)}}{\frac{\partial J}{\partial W^1}} = \underset{\mathbb{R}}{\overset{①}{\frac{\partial J}{\partial y}}} \cdot \underset{\mathbb{R}}{\overset{②}{\frac{\partial y}{\partial h^2}}} \cdot \underset{\mathbb{R}^{1\times2}}{\overset{③}{\frac{\partial h^2}{\partial a^1}}} \cdot \underset{\mathbb{R}^{2\times2}}{\overset{④}{\frac{\partial a^1}{\partial h^1}}} \cdot \underset{\mathbb{R}^{2\times(2\times2)}}{\overset{⑤}{\frac{\partial h^1}{\partial W^1}}}$$

① $\frac{\partial J}{\partial y} = -(t-y) \in \mathbb{R}$

② $\frac{\partial y}{\partial h^2} = y(1-y) \in \mathbb{R}$

③ $\frac{\partial h^2}{\partial a^1} = (W^2)^T \in \mathbb{R}^{1\times2}$

④ $\frac{\partial a^1}{\partial h^1} = \begin{bmatrix} \mathrm{ReLU}'(h_1^1) & 0 \\ 0 & \mathrm{ReLU}'(h_2^1) \end{bmatrix} \in \mathbb{R}^{2\times2}$

⑤ $\frac{\partial h^1}{\partial W^1} = \begin{bmatrix} \underset{\mathbb{R}^{2\times2}}{\frac{\partial h_1^1}{\partial W^1}} & \underset{\mathbb{R}^{2\times2}}{\frac{\partial h_2^1}{\partial W^1}} \end{bmatrix}$; $\cdot\ \frac{\partial h_1^1}{\partial W^1} = \begin{bmatrix} \frac{\partial h_1^1}{\partial W_{11}} & \frac{\partial h_1^1}{\partial W_{12}} \\ \frac{\partial h_1^1}{\partial W_{21}} & \frac{\partial h_1^1}{\partial W_{22}} \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ 0 & 0 \end{bmatrix}$

In the same way we obtain: $\dfrac{\partial h^1_2}{\partial W^1} = \begin{bmatrix} 0 & 0 \\ x_1 & x_2 \end{bmatrix}$.

TRICK: reduce by one dimension the tensor $\dfrac{\partial h^1}{\partial W^1}$, obtaining

$$\begin{bmatrix} x_1 & x_2 & 0 & 0 \\ 0 & 0 & x_1 & x_2 \end{bmatrix} \quad (*)$$

it's like doing the derivative wrt $[W^1_{11} \ W^1_{12} \ W^1_{21} \ W^1_{22}] \in \mathbb{R}^{1\times 4}$

instead of the derivative wrt $\begin{bmatrix} W^1_{11} & W^1_{12} \\ W^1_{21} & W^1_{22} \end{bmatrix} \in \mathbb{R}^{2\times 2}$.

So:

$$\frac{\partial J}{\partial W^1} = \overbrace{-(t-y)y(1-y)}^{=C}\,(W^2)^T \odot \left[ReLU'(h^1_1) \quad ReLU'(h^1_2)\right] \cdot \begin{bmatrix} x_1 & x_2 & 0 & 0 \\ 0 & 0 & x_1 & x_2 \end{bmatrix}$$

$$= C \cdot \left[W^2_1 ReLU'(h^1_1) \quad W^2_2 ReLU'(h^1_2)\right] \begin{bmatrix} x_1 & x_2 & 0 & 0 \\ 0 & 0 & x_1 & x_2 \end{bmatrix}$$

$$= C \cdot \left[x_1 \underbrace{W^2_1 ReLU'(h^1_1)}_{=c_1} \quad x_2 \underbrace{W^2_1 ReLU'(h^1_1)}_{=c_1} \quad x_1 \underbrace{W^2_2 ReLU'(h^1_2)}_{=c_2} \quad x_2 \underbrace{W^2_2 ReLU'(h^1_2)}_{=c_2} \right]$$

$$= C \cdot \begin{bmatrix} c_1 x_1 & c_1 x_2 \\ c_2 x_1 & c_2 x_2 \end{bmatrix} = C \cdot \underset{\in\,\mathbb{R}^{2\times 1}}{[c_1 \ c_2]^T} \cdot \underset{\in\,\mathbb{R}^{1\times 2}}{x^T}$$

$$= -(t-y)\cdot y(1-y)\cdot\left[(W^2)^T \odot ReLU'(h^1)\right]^T \cdot x^T$$

(*) Visualizing the tensor: