Question

How much does $J(X,Y)$ change along a dimension $\rho^{[l]}$ in the space $(\mathbb{W},\mathbb{B})$, the space of all weights and bias values, at point $\mathbb{K}$, a point in the larger space ← larger by $m \cdot (\lambda^{[0]}+1)$ dimensions.
$(\mathbb{W},\mathbb{B},\mathbb{X},\mathbb{Y})$ of weights, biases, training example inputs, training example label.

$J(X,Y)$ has been computed at point $\mathbb{K} \in (\mathbb{W},\mathbb{B},\mathbb{X},\mathbb{Y})$

Find Gradient

We need to find the derivatives

$$\left.\frac{\partial J}{\partial \rho^{[l]}}\right|_{\mathbb{K}}$$

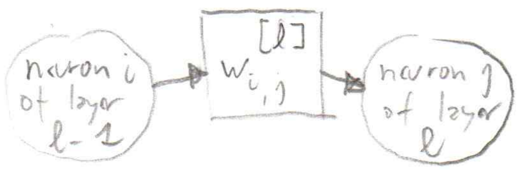
Notation for: the partial derivative of J along the dimension of "placeholder variable" $\rho^{[l]}$, which is related to layer $[l]$ and will be further defined, at point $\mathbb{K}$.
$\rho^{[l]}$ is one of the weights or bias values found in layer $[l]$.

$\boxed{\rho^{[l]}}$ is one of: $\boxed{w_{i,j}^{[l]}}$

$l \in \{1,2,3\}$ for a 3-layer net

$i \in \{1,\dots,\lambda^{[l-1]}\}$

$j \in \{1,\dots,\lambda^{[l]}\}$



$\boxed{b_j^{[l]}}$ $\quad j \in \{1,\dots,\lambda^{[l]}\}$

$$\left.\frac{\partial J}{\partial \rho^{[\ell]}}\right|_K = \frac{1}{m}\sum_{c\in\{1,\dots,m\}} \boxed{\left.\left(\frac{\partial}{\partial \rho^{[\ell]}} \operatorname{loss}\left(\hat{y}^{(c)}, y^{(c)}\right)\right)\right|_K}$$

this computes a mean

In the diagram, the above values are arranged in a matrix according to the $\rho^{[\ell]}$.

Developing only the expression of the term:

1) Use our definition of $\operatorname{loss}(-,-)$ based on negative log-likelihood for 2 classes
2) $y^{(c)}$ does not depend on $\rho^{[\ell]}$

$$- \left[y^{(c)} \frac{\partial}{\partial \rho^{[\ell]}} \log\left(\hat{y}^{(c)}\right) + \left(1-y^{(c)}\right) \frac{\partial}{\partial \rho^{[\ell]}} \log\left(1-\hat{y}^{(c)}\right) \right]$$

Chain rule — Chain rule

$$= - \left[y^{(c)} \cdot \frac{1}{\hat{y}^{(c)}} \cdot \frac{\partial}{\partial \rho^{[\ell]}} \hat{y}^{(c)} + \left(1-y^{(c)}\right) \frac{1}{1-\hat{y}^{(c)}} \cdot \frac{\partial}{\partial \rho^{[\ell]}} \left(1-\hat{y}^{(c)}\right) \right]$$

Using $\boxed{\hat{y}^{(c)} = a^{[3](c)}}$ ($a^{[3](c)}$ is of the shape 1×1 i.e. a scalar!)

Pulling in the minus

Differentiating the rightmost sum

$$= - \frac{y^{(c)}}{\hat{y}^{(c)}} \cdot \frac{\partial}{\partial \rho^{[\ell]}} a^{[3](c)} - \frac{1-y^{(c)}}{1-\hat{y}^{(c)}} \cdot \frac{\partial}{\partial \rho^{[\ell]}} \left(1 - a^{[3](c)}\right)$$

$$= \underbrace{\boxed{\left(\frac{1-y^{(c)}}{1-\hat{y}^{(c)}} - \frac{y^{(c)}}{\hat{y}^{(c)}} \right)}}_{\left.\frac{\partial}{\partial \eta} \operatorname{loss}\left(\eta, y^{(c)}\right)\right|_{\eta=\hat{y}^{(c)}}} \cdot \boxed{\frac{\partial}{\partial \rho^{[\ell]}} a^{[3](c)}}$$

scalar in this architecture!

Note that in the expression

$$\frac{1-y^{(e)}}{1-\hat{y}^{(e)}} - \frac{y^{(e)}}{\hat{y}^{(e)}}$$

$y^{(e)}$ and $1-y^{(e)}$ are "selectors": exactly one is 1.
This can be used during implementations to drop one of the terms and avoid floating-point division that may yield bad results. One should probably make sure to clamp the division results to reasonable values, too.

Now we need to work on the "red block" from earlier:

$$\boxed{\frac{d}{d\rho^{[e]}} a^{[3](e)}}$$

For layer [3], we use the standard sigmoid $\sigma(-)$ as activation function!

$$= \frac{d}{d\rho^{[e]}} \sigma\left(z^{[3](e)}\right)$$

Note that for layer [3] $z^{[3](e)}$ has shape 1×1 (a scalar)

$$= \underbrace{\frac{d}{d\eta}\sigma(\eta)\Big|_{z^{[3](e)}}} \cdot \underbrace{\frac{d}{d\rho^{[e]}} z^{[3](e)}}$$

Chain rule, introduce a new variable η for the expression of the derivative at $z^{[3](e)}$

$$= \sigma'\left(z^{[3](e)}\right) \cdot \frac{\partial}{\partial \rho^{[\ell]}} z^{[3](e)}$$

Switch to Lagrange's "prime notation"

Now use a property peculiar to σ : $\sigma' = \sigma \circ (1-\sigma)$

$$= \underbrace{\sigma\left(z^{[3](e)}\right) \cdot \left(1-\sigma\left(z^{[3](e)}\right)\right)}_{\text{this part can be computed and cached during forward computation}} \cdot \boxed{\frac{\partial}{\partial \rho^{[\ell]}} z^{[3](e)}}$$

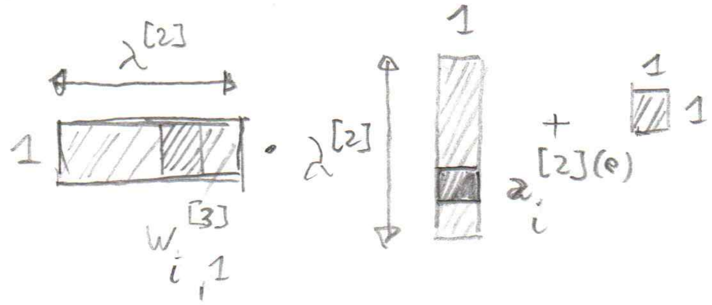
Now we just need to work on the "red block" above.

How does it look for layer 3?

$$\boxed{\frac{\partial}{\partial \rho^{\boxed{[3]}}} z^{[3](e)}} = \frac{\partial}{\partial \rho^{\boxed{[3]}}} \left(W^{[3]^T} \cdot a^{[2](e)} + b^{[3]} \right)$$

NB now at layer 3

$z^{[3](e)}$ is 1×1



▶ For $\boxed{\rho^{[3]} = w_{i,j}^{[3]}}$: $\frac{\partial}{\partial w_{i,1}^{[3]}} \left(\underbrace{W^{[3]^T} \cdot a^{[2](e)}}_{\text{Selects } a_i^{[2](e)}} + \underbrace{b^{[3]}}_{\to 0} \right)$

neuron of layer 2

neuron of layer 3, can only be 1

$$= \boxed{a_i^{[2](e)}}$$

- For $\boxed{\rho^{[3]} = b_1^{[3]}}$: $\frac{\partial}{\partial b_1^{[3]}} \Big(\underbrace{W^{[3]^T} \cdot a^{[2](e)}}_{\to 0} + \underbrace{b_1^{[3]}}_{\to 1} \Big)$

again, the index of the neuron for layer [3] can only be 1 ...

$$= \boxed{1}$$

This ends the calculation for layer [3] as we now have all the info:

$$\frac{\partial J}{\partial \rho^{[3]}}\Bigg|_{\mathbb{K}} = \frac{1}{m} \sum_{e \in \{1 \dots m\}} \Bigg[\left(\frac{1-y^{(e)}}{1-\hat{y}^{(e)}} - \frac{y^{(e)}}{\hat{y}^{(e)}} \right)$$

factor from loss

$$\cdot\, \sigma\left(z_1^{[3](e)}\right) \cdot \left(1 - \sigma\left(z_1^{[3](e)}\right)\right)$$

factor from $act^{[3]}$

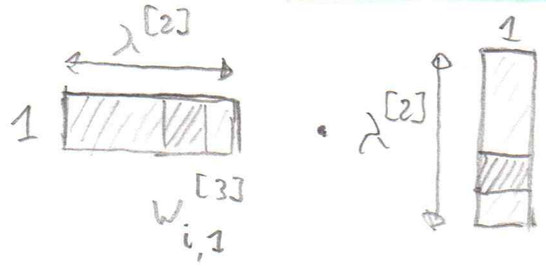
$$\cdot \begin{cases} \rho^{[3]} = b_1^{[3]} : & 1 \\ \rho^{[3]} = w_{i,1}^{[3]} : & a_i^{[2](e)} \end{cases} \Bigg]$$

How does it look for layer 2?

$$\boxed{\frac{\delta}{\delta \boxed{\rho^{[2]}}} z^{[3](e)}} = \frac{\delta}{\delta \boxed{\rho^{[2]}}} \left(W^{[3]^T} \cdot a^{[2](e)} + \underbrace{b^{[3]}}_{\text{resolves to } \emptyset \text{ when differentiating}} \right)$$

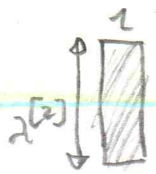
NB now at layer 2

$$= W^{[3]^T} \cdot \boxed{\frac{\delta}{\delta \rho^{[2]}} a^{[2](e)}}$$



Now we need to work on the "green block" above:

$$\boxed{\frac{\delta}{\delta \rho^{[2]}} a^{[2](e)}}$$

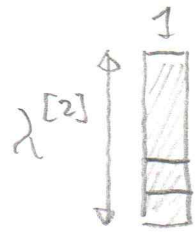


For layer [2], we use an <u>unspecified</u> activation function $act^{[2]}(\_)$:

$$= \frac{\partial}{\partial \rho^{[2]}} act^{[2]}\left(z^{[2](e)} \right)$$

element wise multiplication

$$= \underbrace{\frac{\partial}{\partial \eta} act^{[2]}(\eta) \Big|_{z^{[2](e)}}}_{\text{New variable } \eta \text{ for clarity}} \cdot \underbrace{\frac{\partial}{\partial \rho^{[2]}} z^{[2](e)}}$$



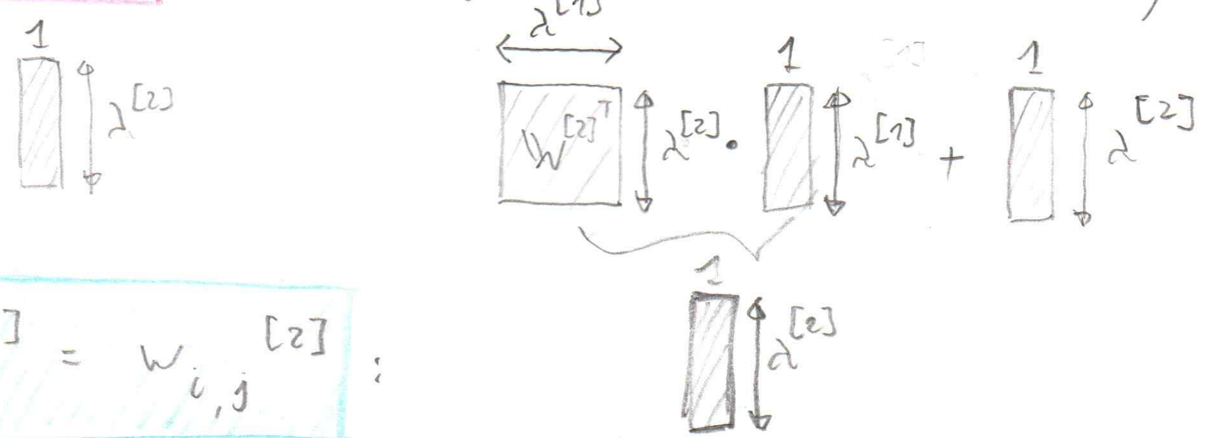
Shape of $z^{[2](e)}$. The derivative has the same shape

$$= \underbrace{act^{[2]'}\left(z^{[2](e)}\right)}_{\text{can be computed and cached during fwd computation}} \cdot \boxed{\frac{\partial}{\partial \rho^{[2]}} z^{[2](e)}}$$

Switch to Lagrange's "prime notation"

Now we just need to work on the "red block" above:

$$\boxed{\frac{\partial}{\partial \rho^{[2]}} z^{[2](e)}} = \frac{\partial}{\partial \rho^{[2]}}\left(W^{[2]^T} \cdot a^{[1](e)} + b^{[2]}\right)$$

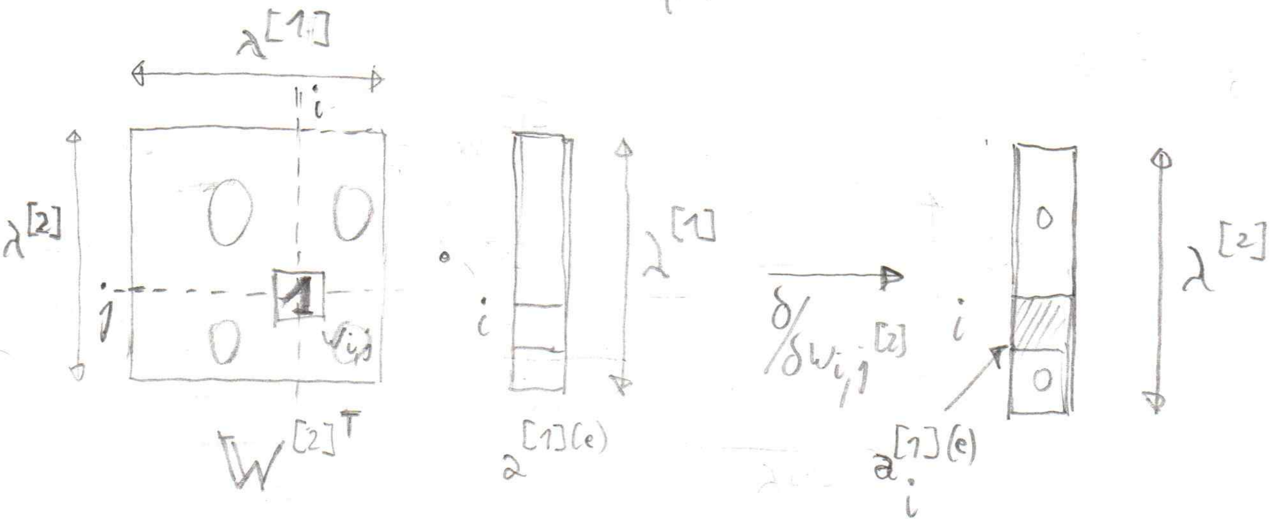


▶ For $\boxed{\rho^{[2]} = w_{i,j}^{[2]}}$:

$$\frac{\partial}{\partial w_{i,j}^{[2]}}\left(\underbrace{W^{[2]^T} \cdot a^{[1](e)}}_{\text{Selects } a_i^{[1](e)} \text{ on differentiation}} + \underbrace{b^{[2]}}_{\text{yields } \emptyset \text{ on differentiation}}\right) = \boxed{a_i^{[1](e)}}$$

inside a column vector of ∅s

From neuron i in [1] to neuron j in [2]



▶ For $\boxed{\rho^{[2]} = b_j^{[2]}}$:

$$\frac{\delta}{\delta b_j^{[2]}}\left(\underbrace{W^{[2]^T}\cdot a^{[1](e)}}_{\text{yields } 0} + \underbrace{b^{[2]}}_{\text{yields "1 at place } j\text{"}}\right) = \boxed{1 \text{ inside a column vector of } 0\text{s}}$$

the above has the shape

$$\begin{array}{cc} & 1 \\ & \left.\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}\right\} a^{[2]} \end{array}$$

with the 1 in row j.

This ends the calculation for layer [2] as we now have all the info.

(9)

$$\frac{\partial J}{\partial \rho^{[2]}}\Bigg|_K = \frac{1}{m} \sum_{e \in \{1,..,m\}} \boxed{\left(\frac{1-y^{(e)}}{1-\hat{y}^{(e)}} - \frac{y^{(e)}}{\hat{y}^{(e)}}\right)}$$

scalar factor from loss computation

For a given $\rho^{[2]}$, this is a scalar.

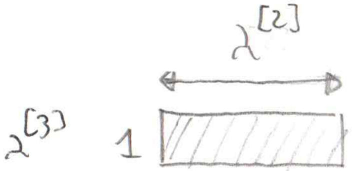
$$\cdot \; \boxed{\sigma\left(z^{[3](e)}\right) \cdot \left(1 - \sigma\left(z^{[3](e)}\right)\right)}$$

scalar factor from differentiating through $act^{[3]} = \sigma(\cdot)$

$$\cdot \; \boxed{W^{[3]^T}}$$

from differentiating $z^{[3](e)}$

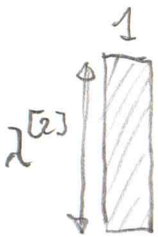
has shape



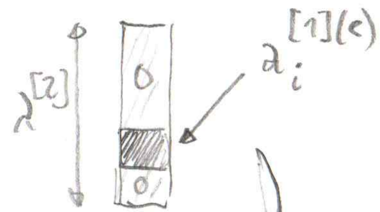
$$\cdot \left(\boxed{act^{[2]'}\left(z^{[2](e)}\right)} \right.$$

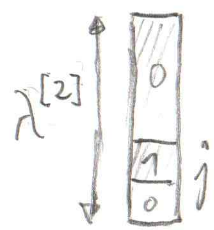
vector factor from differentiating through $act^{[2]}$

has shape



$$\cdot \left. \begin{cases} \text{if } \rho^{[2]} = w_{i,1}^{[2]} : & \text{the vector} \\ \text{if } \rho^{[2]} = b_j^{[2]} : & \text{the vector} \end{cases} \right)$$

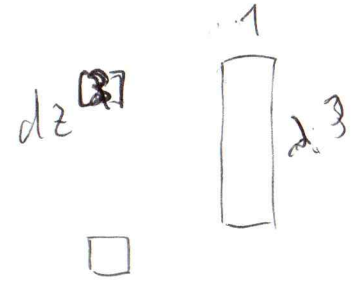




And J' us :

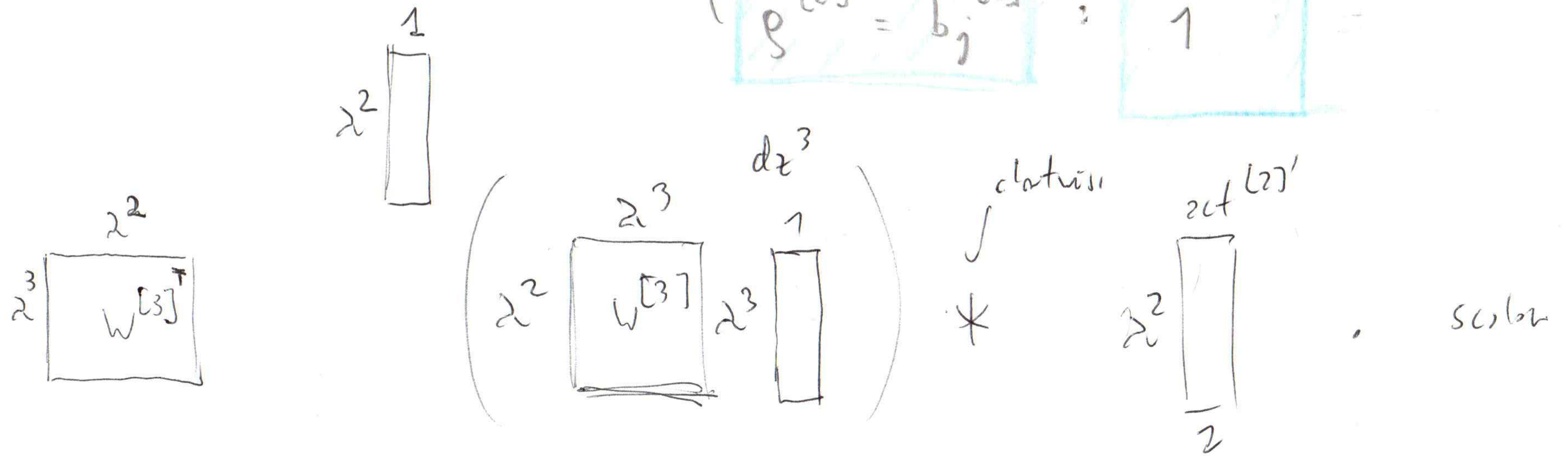
(10)

$$\frac{\partial J}{\partial \rho^{[2]}}\Bigg|_k = \frac{1}{m}\sum_{e\in\{1,\dots,m\}} \underbrace{\left(\frac{1-y^{(e)}}{1-\hat{y}^{(e)}} - \frac{y^{(e)}}{\hat{y}^{(e)}}\right) \cdot \sigma\left(z^{[3](e)}\right)\cdot\left(1-\sigma\left(z^{[3](e)}\right)\right)}_{dz^{[3]}}$$



this seems wrong! you need the transpose!

$$\cdot\, W^{[3]^T} \cdot act^{[2]'}\left(z_j^{[2](e)}\right) \cdot \begin{cases} \rho^{[2]} = w_{i,j}^{[2]} : & a_i^{[1](e)} \\ \rho^{[2]} = b_j^{[2]} : & 1 \end{cases}$$



How does it look for layer 3 ?

$$\frac{\partial}{\partial \rho^{\boxed{[1]}}} z^{[3](e)} = W^{[3]^T} \cdot \frac{\partial}{\partial \rho^{[1]}} a^{[2](e)}$$

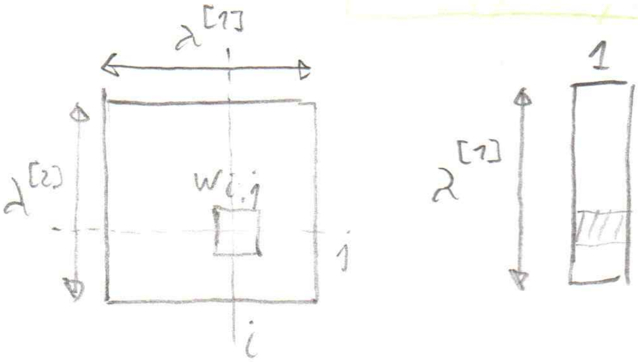
NB now at layer 1

$$\frac{\partial}{\partial \rho^{[1]}} a^{[2](e)} = act^{[2]'}\left(z^{[2](e)}\right) \cdot \frac{\partial}{\partial \rho^{[1]}} z^{[2](e)}$$

Now we just need to work on the "red block" above:

$$\frac{\partial}{\partial \rho^{\boxed{[1]}}} z^{[2](e)} = \frac{\partial}{\partial \rho^{\boxed{[1]}}} \left(W^{[2]^T} \cdot a^{[1](e)} + \underbrace{b^{[2]}}_{\text{resolves to } \emptyset} \right)$$

$$= W^{[2]^T} \cdot \frac{\partial}{\partial \rho^{[1]}} a^{[1](e)}$$



Now we need to work on the "green block" above:

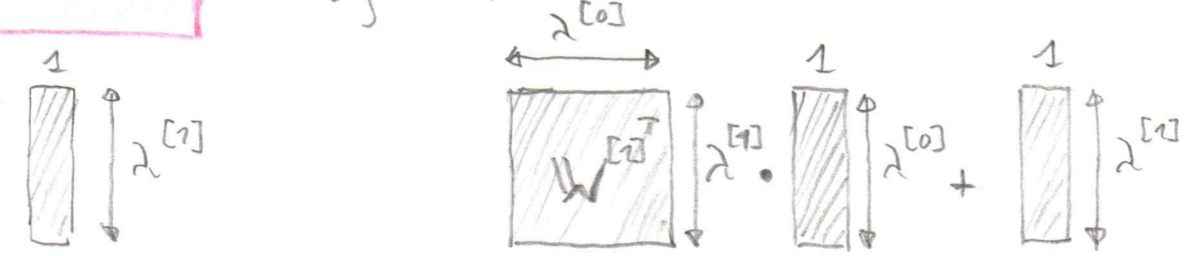
$$\frac{\partial}{\partial \rho^{[1]}} a^{[1](e)}$$

For layer [1], we use an <u>unspecified</u> activation function $act^{[1]}(\_)$

$$= \underbrace{act^{[1]'}\left(z^{[1](e)}\right)}_{\text{can be computed and cached during fwd computation}} \cdot \boxed{\frac{\partial}{\partial p^{[1]}} z^{[1](e)}}$$

Now we just need to work on the "red block" above:

$$\boxed{\frac{\partial}{\partial p^{[1]}} z^{[1](e)}} = \frac{\partial}{\partial p^{[1]}}\left(W^{[1]^T} \cdot a^{[0](e)} + b^{[1]}\right)$$



▷ For $\boxed{p^{[1]} = w_{i,j}^{[1]}}$:

$$\frac{\partial}{\partial w_{i,j}^{[1]}}\left(\underbrace{W^{[1]^T} \cdot a^{[0](e)}}_{\text{selects "}a_i^{[0](e)}\text{ at place } j\text{"}} + \underbrace{b^{[1]}}_{\text{yields } \emptyset}\right) = \boxed{a_i^{[0](e)} \text{ inside a column vector of } \emptyset\text{s}}$$

▷ For $\boxed{p^{[1]} = b_j^{[1]}}$:

$$\frac{\partial}{\partial b_j^{[1]}}\left(\underbrace{W^{[1]^T} \cdot a^{[0](e)}}_{\text{yields } \emptyset} + \underbrace{b^{[1]}}_{\text{yields "1 at place } j\text{"}}\right) = \boxed{1 \text{ inside a column vector of } \emptyset\text{s}}$$

And thus

$$\frac{\partial J}{\partial \rho^{[1]}}\bigg|_{\mathbb{K}} = \frac{1}{m}\sum_{e\in\{1,..,m\}} \left(\frac{1-y^{(e)}}{1-\hat{y}^{(e)}} - \frac{y^{(e)}}{\hat{y}^{(e)}}\right)$$

$$\cdot\, \sigma\left(z^{[3](e)}\right)\cdot\left(1-\sigma\left(z^{[3](e)}\right)\right) \quad \square$$

$$\cdot\, \mathbb{W}^{[3]^T} \cdot act^{[2]'}\left(z^{[2](e)}\right)$$

$$\cdot\, \mathbb{W}^{[2]^T} \cdot act^{[1]'}\left(z^{[1](e)}\right)$$

$$\cdot \begin{cases} \rho^{[1]} = w_{i,j}^{[1]} : & \boxed{a_i^{[0](e)}} \\ \rho^{[1]} = b_j^{[1]} : & 1 \end{cases}$$