Aprendizagem 2023/24

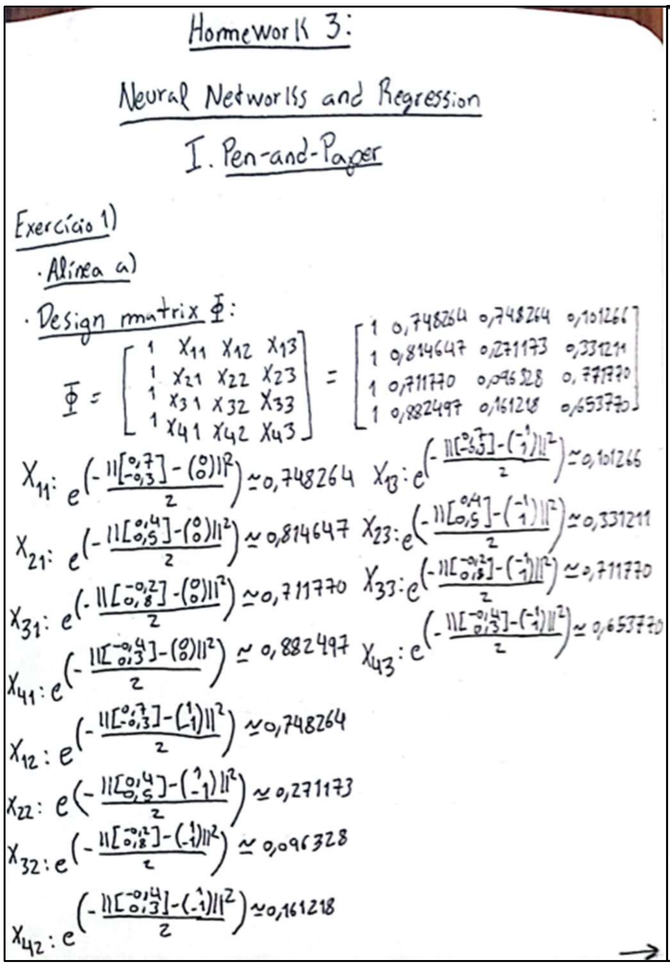
Homework III

I. Pen-and-paper [12v]

For questions in this group, show your numerical results with 5 decimals or scientific notation. *Hint*: we highly recommend the use of numpy (e.g., linalg.pinv for inverse) or other programmatic facilities to support the calculus involved in both questions (1) and (2).

1. Consider the problem of learning a regression model from 4 bivariate observations $\left\{\begin{pmatrix}0.7\\-0.3\end{pmatrix},\begin{pmatrix}0.4\\0.5\end{pmatrix},\begin{pmatrix}-0.2\\0.8\end{pmatrix},\begin{pmatrix}-0.4\\0.3\end{pmatrix}\right\}$ with targets $(0.8, 0.6, 0.3, 0.3)$.

   a. [4v] Given the radial basis function, $\phi_j(x) = exp\left(-\frac{\|x-c_j\|^2}{2}\right)$, that transforms the original space onto a new space characterized by the similarity of the original observations to the following data points, $\left\{c_1 = \begin{pmatrix}0\\0\end{pmatrix}, c_2 = \begin{pmatrix}1\\-1\end{pmatrix}, c_3 = \begin{pmatrix}-1\\1\end{pmatrix}\right\}$.



Homework 3:

Neural Networks and Regression

I. Pen-and-Paper

Exercício 1)

· Alínea a)

· Design matrix Φ:

$$\Phi = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{13} \\ 1 & X_{21} & X_{22} & X_{23} \\ 1 & X_{31} & X_{32} & X_{33} \\ 1 & X_{41} & X_{42} & X_{43} \end{bmatrix} = \begin{bmatrix} 1 & 0{,}748264 & 0{,}748264 & 0{,}101266 \\ 1 & 0{,}814647 & 0{,}271173 & 0{,}331211 \\ 1 & 0{,}711770 & 0{,}096328 & 0{,}711770 \\ 1 & 0{,}882497 & 0{,}161218 & 0{,}653770 \end{bmatrix}$$

$X_{11}: e^{\left(-\frac{\|[{0,7 \atop -0,3}]-({0 \atop 0})\|^2}{2}\right)} \simeq 0{,}748264$ $X_{13}: e^{\left(-\frac{\|[{0,7 \atop -0,3}]-({-1 \atop 1})\|^2}{2}\right)} \simeq 0{,}101266$

$X_{21}: e^{\left(-\frac{\|[{0,4 \atop 0,5}]-({0 \atop 0})\|^2}{2}\right)} \simeq 0{,}814647$ $X_{23}: e^{\left(-\frac{\|[{0,4 \atop 0,5}]-({-1 \atop 1})\|^2}{2}\right)} \simeq 0{,}331211$

$X_{31}: e^{\left(-\frac{\|[{-0,2 \atop 0,8}]-({0 \atop 0})\|^2}{2}\right)} \simeq 0{,}711770$ $X_{33}: e^{\left(-\frac{\|[{-0,2 \atop 0,8}]-({-1 \atop 1})\|^2}{2}\right)} \simeq 0{,}711770$

$X_{41}: e^{\left(-\frac{\|[{-0,4 \atop 0,3}]-({0 \atop 0})\|^2}{2}\right)} \simeq 0{,}882497$ $X_{43}: e^{\left(-\frac{\|[{-0,4 \atop 0,3}]-({-1 \atop 1})\|^2}{2}\right)} \simeq 0{,}653770$

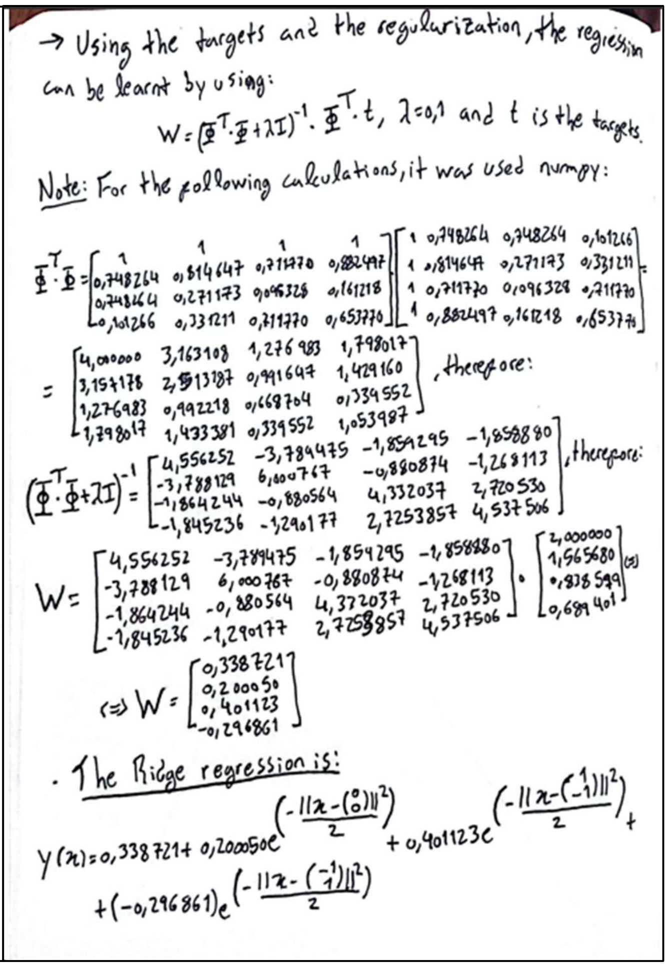
$X_{12}: e^{\left(-\frac{\|[{0,7 \atop -0,3}]-({1 \atop -1})\|^2}{2}\right)} \simeq 0{,}748264$

$X_{22}: e^{\left(-\frac{\|[{0,4 \atop 0,5}]-({1 \atop -1})\|^2}{2}\right)} \simeq 0{,}271173$

$X_{32}: e^{\left(-\frac{\|[{-0,2 \atop 0,8}]-({1 \atop -1})\|^2}{2}\right)} \simeq 0{,}096328$

$X_{42}: e^{\left(-\frac{\|[{-0,4 \atop 0,3}]-({1 \atop -1})\|^2}{2}\right)} \simeq 0{,}161218$

$\rightarrow$



$\rightarrow$ Using the targets and the regularization, the regression can be learnt by using:

$$W = (\Phi^T \cdot \Phi + \lambda I)^{-1} \cdot \Phi^T \cdot t,\ \lambda = 0{,}1 \text{ and } t \text{ is the targets.}$$

Note: For the following calculations, it was used numpy:

$$\Phi^T \cdot \Phi = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0{,}748264 & 0{,}814647 & 0{,}711770 & 0{,}882497 \\ 0{,}748264 & 0{,}271173 & 0{,}096328 & 0{,}161218 \\ 0{,}101266 & 0{,}331211 & 0{,}711770 & 0{,}653770 \end{bmatrix} \begin{bmatrix} 1 & 0{,}748264 & 0{,}748264 & 0{,}101266 \\ 1 & 0{,}814647 & 0{,}271173 & 0{,}331211 \\ 1 & 0{,}711770 & 0{,}096328 & 0{,}711770 \\ 1 & 0{,}882497 & 0{,}161218 & 0{,}653770 \end{bmatrix} =$$

$$= \begin{bmatrix} 4{,}000000 & 3{,}163108 & 1{,}276983 & 1{,}798017 \\ 3{,}157178 & 2{,}513787 & 0{,}991647 & 1{,}429160 \\ 1{,}276983 & 0{,}992218 & 0{,}668704 & 0{,}334552 \\ 1{,}798017 & 1{,}433381 & 0{,}339552 & 1{,}053987 \end{bmatrix}, \text{therefore:}$$

$$(\Phi^T \cdot \Phi + \lambda I)^{-1} = \begin{bmatrix} 4{,}556252 & -3{,}784475 & -1{,}854295 & -1{,}858880 \\ -3{,}788129 & 6{,}000767 & -0{,}880874 & -1{,}268113 \\ -1{,}864244 & -0{,}880564 & 4{,}332037 & 2{,}720530 \\ -1{,}845236 & -1{,}290177 & 2{,}7253857 & 4{,}537506 \end{bmatrix}, \text{therefore:}$$

$$W = \begin{bmatrix} 4{,}556252 & -3{,}789475 & -1{,}854295 & -1{,}858880 \\ -3{,}788129 & 6{,}000767 & -0{,}880874 & -1{,}268113 \\ -1{,}864244 & -0{,}880564 & 4{,}372037 & 2{,}720530 \\ -1{,}845236 & -1{,}290177 & 2{,}7258857 & 4{,}537506 \end{bmatrix} \cdot \begin{bmatrix} 2{,}000000 \\ 1{,}565680 \\ 0{,}838599 \\ 0{,}689401 \end{bmatrix} \Leftrightarrow$$

$$\Leftrightarrow W = \begin{bmatrix} 0{,}338721 \\ 0{,}200050 \\ 0{,}401123 \\ -0{,}296861 \end{bmatrix}$$

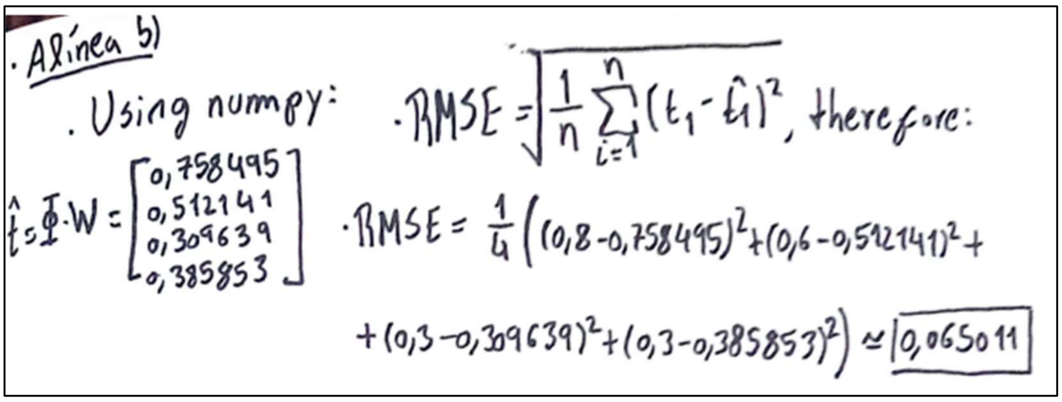
· The Ridge regression is:

$$y(x) = 0{,}338721 + 0{,}200050 e^{\left(-\frac{\|x-({0 \atop 0})\|^2}{2}\right)} + 0{,}401123 e^{\left(-\frac{\|x-({1 \atop -1})\|^2}{2}\right)} + (-0{,}296861) e^{\left(-\frac{\|x-({-1 \atop 1})\|^2}{2}\right)}$$

Aprendizagem 2023/24

Homework III

b. [2v] Compute the training RMSE for the learnt regression.



· Alínea b)

. Using numpy:

$$\cdot \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_1 - \hat{t}_1)^2}, \text{ therefore:}$$

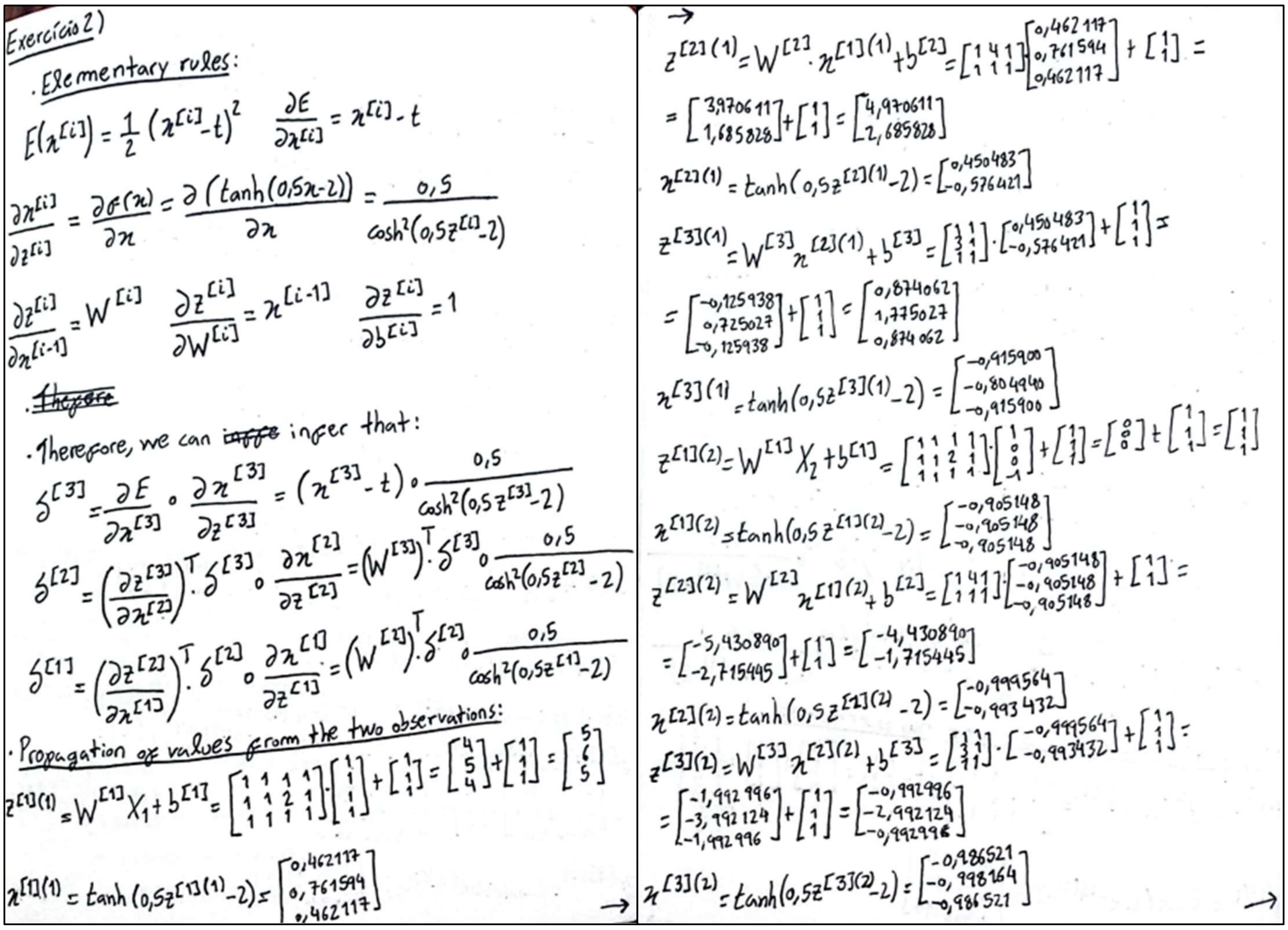
$$\hat{t} = \Phi \cdot W = \begin{bmatrix} 0{,}758495 \\ 0{,}512141 \\ 0{,}309639 \\ 0{,}385853 \end{bmatrix}$$

$$\cdot \text{RMSE} = \frac{1}{4}\Big((0{,}8 - 0{,}758495)^2 + (0{,}6 - 0{,}512141)^2 + $$
$$+ (0{,}3 - 0{,}309639)^2 + (0{,}3 - 0{,}385853)^2\Big) \simeq \boxed{0{,}065011}$$

2. [6v] Consider a MLP classifier of three outcomes – A, B and C – characterized by the weights,

$$W^{[1]} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, b^{[1]} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, W^{[2]} = \begin{pmatrix} 1 & 4 & 1 \\ 1 & 1 & 1 \end{pmatrix}, b^{[2]} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, W^{[3]} = \begin{pmatrix} 1 & 1 \\ 3 & 1 \\ 1 & 1 \end{pmatrix}, b^{[3]} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

the activation $f(x) = \frac{e^{0.5x-2} - e^{-0.5x+2}}{e^{0.5x-2} + e^{-0.5x+2}} = tanh(0.5x - 2)$ for every unit, and squared error loss $\frac{1}{2}\|\mathbf{z} - \hat{\mathbf{z}}\|_2^2$. Perform one batch gradient descent update (with learning rate $\eta = 0.1$) for training observations $\mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$ and $\mathbf{x}_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}$ with targets B and A, respectively.



Exercício 2)

. Elementary rules:

$$E(x^{[i]}) = \frac{1}{2}(x^{[i]}-t)^2 \quad \frac{\partial E}{\partial x^{[i]}} = x^{[i]} - t$$

$$\frac{\partial x^{[i]}}{\partial z^{[i]}} = \frac{\partial f(x)}{\partial x} = \frac{\partial(\tanh(0{,}5x-2))}{\partial x} = \frac{0{,}5}{\cosh^2(0{,}5z^{[i]}-2)}$$

$$\frac{\partial z^{[i]}}{\partial x^{[i-1]}} = W^{[i]} \quad \frac{\partial z^{[i]}}{\partial W^{[i]}} = x^{[i-1]} \quad \frac{\partial z^{[i]}}{\partial b^{[i]}} = 1$$

. ~~Therefore~~

. Therefore, we can ~~inffe~~ infer that:

$$\delta^{[3]} = \frac{\partial E}{\partial x^{[3]}} \circ \frac{\partial x^{[3]}}{\partial z^{[3]}} = (x^{[3]} - t) \circ \frac{0{,}5}{\cosh^2(0{,}5z^{[3]}-2)}$$

$$\delta^{[2]} = \left(\frac{\partial z^{[3]}}{\partial x^{[2]}}\right)^T \cdot \delta^{[3]} \circ \frac{\partial x^{[2]}}{\partial z^{[2]}} = (W^{[3]})^T \cdot \delta^{[3]} \circ \frac{0{,}5}{\cosh^2(0{,}5z^{[2]}-2)}$$

$$\delta^{[1]} = \left(\frac{\partial z^{[2]}}{\partial x^{[1]}}\right)^T \cdot \delta^{[2]} \circ \frac{\partial x^{[1]}}{\partial z^{[1]}} = (W^{[2]})^T \cdot \delta^{[2]} \circ \frac{0{,}5}{\cosh^2(0{,}5z^{[1]}-2)}$$

. Propagation of values from the two observations:

$$z^{[1](1)} = W^{[1]}X_1 + b^{[1]} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 4 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 5 \\ 6 \\ 5 \end{bmatrix}$$

$$x^{[1](1)} = \tanh(0{,}5z^{[1](1)}-2) = \begin{bmatrix} 0{,}462117 \\ 0{,}761594 \\ 0{,}462117 \end{bmatrix} \rightarrow$$

$\rightarrow$

$$z^{[2](1)} = W^{[2]} \cdot x^{[1](1)} + b^{[2]} = \begin{bmatrix} 1 & 4 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0{,}462117 \\ 0{,}761594 \\ 0{,}462117 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} =$$

$$= \begin{bmatrix} 3{,}970611 \\ 1{,}685828 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 4{,}970611 \\ 2{,}685828 \end{bmatrix}$$

$$x^{[2](1)} = \tanh(0{,}5z^{[2](1)}-2) = \begin{bmatrix} 0{,}450483 \\ -0{,}576421 \end{bmatrix}$$

$$z^{[3](1)} = W^{[3]} x^{[2](1)} + b^{[3]} = \begin{bmatrix} 1 & 1 \\ 3 & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0{,}450483 \\ -0{,}576421 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} =$$

$$= \begin{bmatrix} -0{,}125938 \\ 0{,}725027 \\ -0{,}125938 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0{,}874062 \\ 1{,}725027 \\ 0{,}874062 \end{bmatrix}$$

$$x^{[3](1)} = \tanh(0{,}5z^{[3](1)}-2) = \begin{bmatrix} -0{,}915900 \\ -0{,}804940 \\ -0{,}915900 \end{bmatrix}$$

$$z^{[1](2)} = W^{[1]}X_2 + b^{[1]} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$x^{[1](2)} = \tanh(0{,}5z^{[1](2)}-2) = \begin{bmatrix} -0{,}905148 \\ -0{,}905148 \\ -0{,}905148 \end{bmatrix}$$

$$z^{[2](2)} = W^{[2]} x^{[1](2)} + b^{[2]} = \begin{bmatrix} 1 & 4 & 1 \\ 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} -0{,}905148 \\ -0{,}905148 \\ -0{,}905148 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} =$$

$$= \begin{bmatrix} -5{,}430890 \\ -2{,}715445 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -4{,}430890 \\ -1{,}715445 \end{bmatrix}$$

$$x^{[2](2)} = \tanh(0{,}5z^{[2](2)}-2) = \begin{bmatrix} -0{,}999564 \\ -0{,}993432 \end{bmatrix}$$

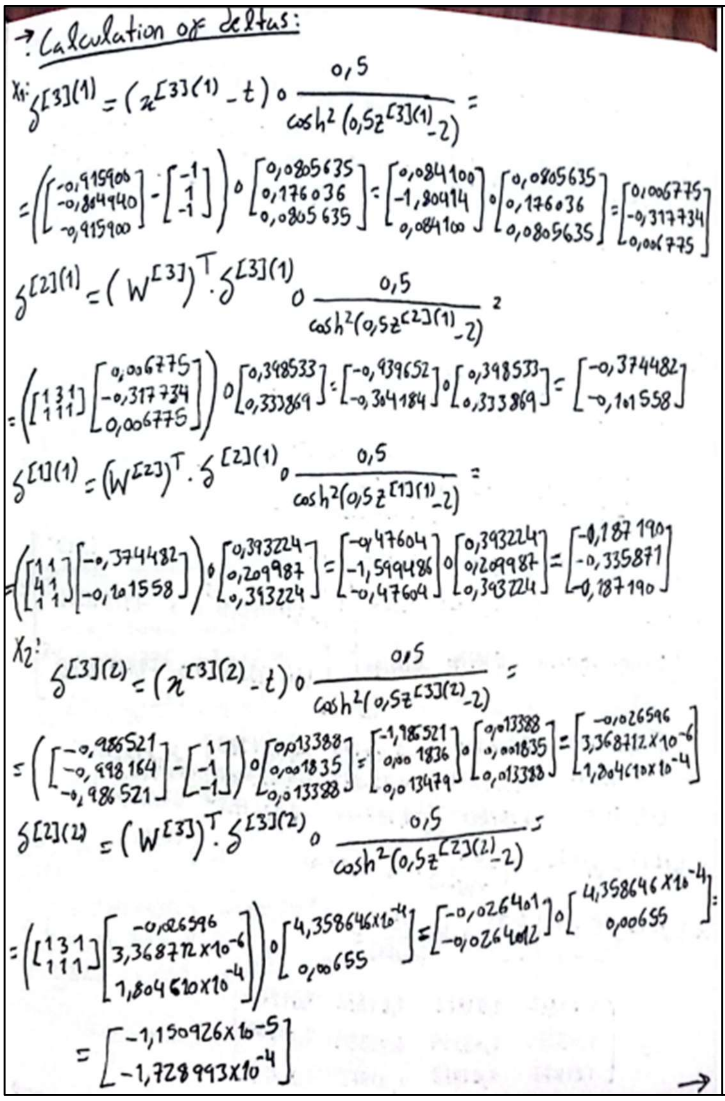
$$z^{[3](2)} = W^{[3]} x^{[2](2)} + b^{[3]} = \begin{bmatrix} 1 & 1 \\ 3 & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} -0{,}999564 \\ -0{,}993432 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} =$$

$$= \begin{bmatrix} -1{,}992996 \\ -3{,}992124 \\ -1{,}992996 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -0{,}992996 \\ -2{,}992124 \\ -0{,}992996 \end{bmatrix}$$

$$x^{[3](2)} = \tanh(0{,}5z^{[3](2)}-2) = \begin{bmatrix} -0{,}986521 \\ -0{,}998164 \\ -0{,}986521 \end{bmatrix} \rightarrow$$

Aprendizagem 2023/24

Homework III



→ Calculation of deltas:

x_1: $$\delta^{[3](1)} = (x^{[3](1)} - t) \circ \frac{0,5}{\cosh^2(0,5z^{[3](1)}-2)} =$$

$$= \left(\begin{bmatrix}-0,915900\\-0,864940\\-0,915900\end{bmatrix} - \begin{bmatrix}-1\\1\\-1\end{bmatrix}\right) \circ \begin{bmatrix}0,0805635\\0,176036\\0,0805635\end{bmatrix} = \begin{bmatrix}0,084100\\-1,80414\\0,084100\end{bmatrix} \circ \begin{bmatrix}0,0805635\\0,176036\\0,0805635\end{bmatrix} = \begin{bmatrix}0,006775\\-0,317734\\0,006775\end{bmatrix}$$

$$\delta^{[2](1)} = (W^{[3]})^T \cdot \delta^{[3](1)} \circ \frac{0,5}{\cosh^2(0,5z^{[2](1)}-2)} =$$

$$= \left(\begin{bmatrix}1&3&1\\1&1&1\end{bmatrix}\begin{bmatrix}0,006775\\-0,317734\\0,006775\end{bmatrix}\right) \circ \begin{bmatrix}0,348533\\0,333869\end{bmatrix} = \begin{bmatrix}-0,939652\\-0,304184\end{bmatrix} \circ \begin{bmatrix}0,398533\\0,333869\end{bmatrix} = \begin{bmatrix}-0,374482\\-0,101558\end{bmatrix}$$

$$\delta^{[1](1)} = (W^{[2]})^T \cdot \delta^{[2](1)} \circ \frac{0,5}{\cosh^2(0,5z^{[1](1)}-2)} =$$

$$= \left(\begin{bmatrix}1&1\\4&1\\1&1\end{bmatrix}\begin{bmatrix}-0,374482\\-0,101558\end{bmatrix}\right) \circ \begin{bmatrix}0,393224\\0,209987\\0,393224\end{bmatrix} = \begin{bmatrix}-0,47604\\-1,599486\\-0,47604\end{bmatrix} \circ \begin{bmatrix}0,393224\\0,209987\\0,393224\end{bmatrix} = \begin{bmatrix}-0,187190\\-0,335871\\-0,187190\end{bmatrix}$$

x_2: $$\delta^{[3](2)} = (x^{[3](2)} - t) \circ \frac{0,5}{\cosh^2(0,5z^{[3](2)}-2)} =$$

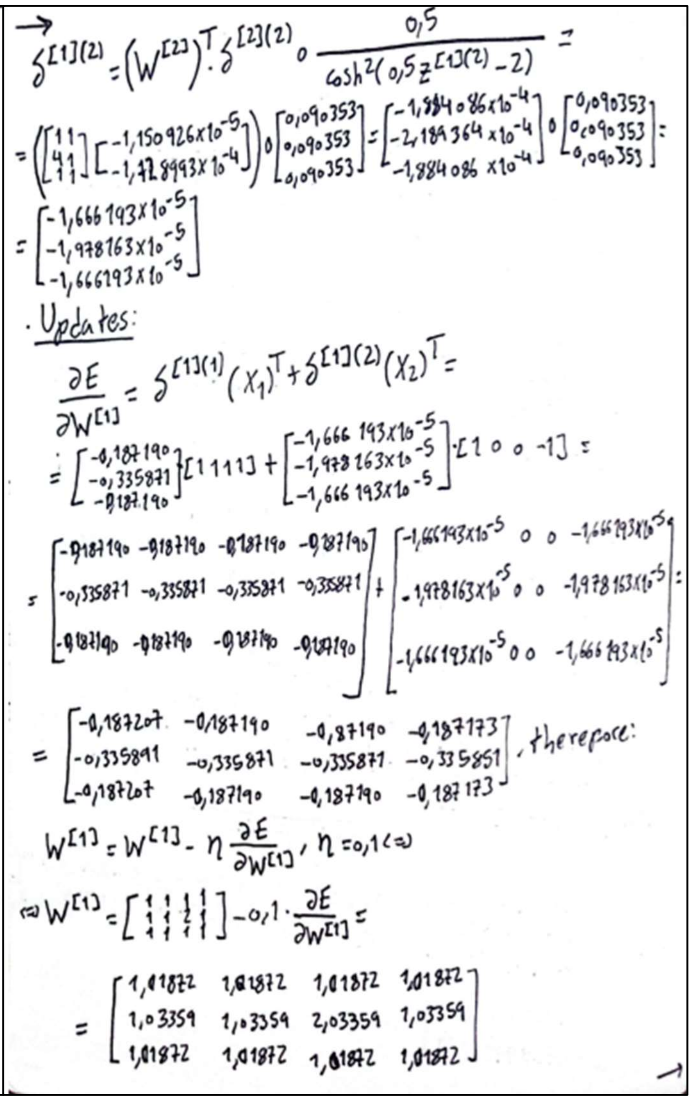
$$= \left(\begin{bmatrix}-0,986521\\-0,998164\\-0,986521\end{bmatrix} - \begin{bmatrix}1\\-1\\-1\end{bmatrix}\right) \circ \begin{bmatrix}0,013388\\0,001835\\0,013388\end{bmatrix} = \begin{bmatrix}-1,986521\\0,001836\\0,013479\end{bmatrix} \circ \begin{bmatrix}0,013388\\0,001835\\0,013388\end{bmatrix} = \begin{bmatrix}-0,026596\\3,368712\times10^{-6}\\1,804610\times10^{-4}\end{bmatrix}$$

$$\delta^{[2](2)} = (W^{[3]})^T \cdot \delta^{[3](2)} \circ \frac{0,5}{\cosh^2(0,5z^{[2](2)}-2)} =$$

$$= \left(\begin{bmatrix}1&3&1\\1&1&1\end{bmatrix}\begin{bmatrix}-0,026596\\3,368712\times10^{-6}\\1,804610\times10^{-4}\end{bmatrix}\right) \circ \begin{bmatrix}4,358646\times10^{-4}\\0,00655\end{bmatrix} = \begin{bmatrix}-0,026401\\-0,0264012\end{bmatrix} \circ \begin{bmatrix}4,358646\times10^{-4}\\0,00655\end{bmatrix} =$$

$$= \begin{bmatrix}-1,150926\times10^{-5}\\-1,728993\times10^{-4}\end{bmatrix}$$

→



→

$$\delta^{[1](2)} = (W^{[2]})^T \cdot \delta^{[2](2)} \circ \frac{0,5}{\cosh^2(0,5z^{[1](2)}-2)} =$$

$$= \left(\begin{bmatrix}1&1\\4&1\\1&1\end{bmatrix}\begin{bmatrix}-1,150926\times10^{-5}\\-1,728993\times10^{-4}\end{bmatrix}\right) \circ \begin{bmatrix}0,090353\\0,090353\\0,090353\end{bmatrix} = \begin{bmatrix}-1,884086\times10^{-4}\\-2,189364\times10^{-4}\\-1,884086\times10^{-4}\end{bmatrix} \circ \begin{bmatrix}0,090353\\0,090353\\0,090353\end{bmatrix} =$$

$$= \begin{bmatrix}-1,666193\times10^{-5}\\-1,978163\times10^{-5}\\-1,666193\times10^{-5}\end{bmatrix}$$

. Updates:

$$\frac{\partial E}{\partial W^{[1]}} = \delta^{[1](1)}(x_1)^T + \delta^{[1](2)}(x_2)^T =$$

$$= \begin{bmatrix}-0,187190\\-0,335871\\-0,187190\end{bmatrix}\begin{bmatrix}1&1&1&1\end{bmatrix} + \begin{bmatrix}-1,666193\times10^{-5}\\-1,978163\times10^{-5}\\-1,666193\times10^{-5}\end{bmatrix} \cdot \begin{bmatrix}1&0&0&-1\end{bmatrix} =$$

$$= \begin{bmatrix}-0,187190&-0,187190&-0,187190&-0,187190\\-0,335871&-0,335871&-0,335871&-0,335871\\-0,187190&-0,187190&-0,187190&-0,187190\end{bmatrix} + \begin{bmatrix}-1,666193\times10^{-5}&0&0&-1,666193\times10^{-5}\\-1,978163\times10^{-5}&0&0&-1,978163\times10^{-5}\\-1,666193\times10^{-5}&0&0&-1,666193\times10^{-5}\end{bmatrix} =$$

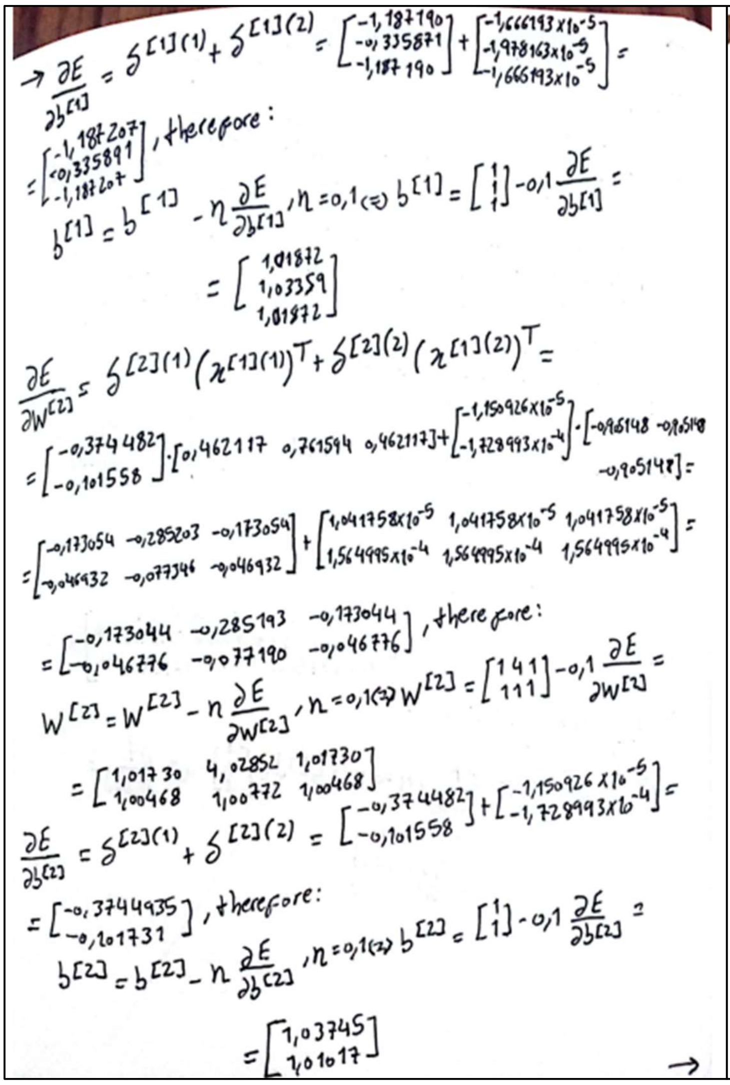
$$= \begin{bmatrix}-0,187207&-0,187190&-0,87190&-0,187173\\-0,335891&-0,335871&-0,335871&-0,335851\\-0,187207&-0,187190&-0,187190&-0,187173\end{bmatrix}$$, therefore:

$$W^{[1]} = W^{[1]} - \eta\frac{\partial E}{\partial W^{[1]}}, \ \eta = 0,1 \Leftrightarrow$$

$$\Leftrightarrow W^{[1]} = \begin{bmatrix}1&1&1&1\\1&1&2&1\\1&1&1&1\end{bmatrix} - 0,1 \cdot \frac{\partial E}{\partial W^{[1]}} =$$

$$= \begin{bmatrix}1,01872&1,01872&1,01872&1,01872\\1,03359&1,03359&2,03359&1,03359\\1,01872&1,01872&1,01872&1,01872\end{bmatrix}$$

→



$$\rightarrow \frac{\partial E}{\partial b^{[1]}} = \delta^{[1](1)} + \delta^{[1](2)} = \begin{bmatrix} -1{,}187190 \\ -0{,}335871 \\ -1{,}187190 \end{bmatrix} + \begin{bmatrix} -1{,}666193\times10^{-5} \\ -1{,}978163\times10^{-5} \\ -1{,}666193\times10^{-5} \end{bmatrix} =$$

$$= \begin{bmatrix} -1{,}187207 \\ -0{,}335891 \\ -1{,}187207 \end{bmatrix}, \text{therefore:}$$

$$b^{[1]} = b^{[1]} - \eta \frac{\partial E}{\partial b^{[1]}}, \eta = 0{,}1 \Leftrightarrow b^{[1]} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - 0{,}1 \frac{\partial E}{\partial b^{[1]}} =$$

$$= \begin{bmatrix} 1{,}01872 \\ 1{,}03359 \\ 1{,}01872 \end{bmatrix}$$

$$\frac{\partial E}{\partial W^{[2]}} = \delta^{[2](1)} \left(x^{[1](1)}\right)^T + \delta^{[2](2)} \left(x^{[1](2)}\right)^T =$$

$$= \begin{bmatrix} -0{,}374482 \\ -0{,}101558 \end{bmatrix} \cdot \begin{bmatrix} 0{,}462117 & 0{,}761594 & 0{,}462117 \end{bmatrix} + \begin{bmatrix} -1{,}150926\times10^{-5} \\ -1{,}728993\times10^{-4} \end{bmatrix} \cdot \begin{bmatrix} -0{,}905148 & -0{,}905148 & -0{,}905148 \end{bmatrix} =$$

$$= \begin{bmatrix} -0{,}173054 & -0{,}285203 & -0{,}173054 \\ -0{,}046932 & -0{,}077346 & -0{,}046932 \end{bmatrix} + \begin{bmatrix} 1{,}041758\times10^{-5} & 1{,}041758\times10^{-5} & 1{,}041758\times10^{-5} \\ 1{,}564995\times10^{-4} & 1{,}564995\times10^{-4} & 1{,}564995\times10^{-4} \end{bmatrix} =$$

$$= \begin{bmatrix} -0{,}173044 & -0{,}285193 & -0{,}173044 \\ -0{,}046776 & -0{,}077190 & -0{,}046776 \end{bmatrix}, \text{therefore:}$$

$$W^{[2]} = W^{[2]} - \eta \frac{\partial E}{\partial W^{[2]}}, \eta = 0{,}1 \Leftrightarrow W^{[2]} = \begin{bmatrix} 1 & 4 & 1 \\ 1 & 1 & 1 \end{bmatrix} - 0{,}1 \frac{\partial E}{\partial W^{[2]}} =$$

$$= \begin{bmatrix} 1{,}01730 & 4{,}02852 & 1{,}01730 \\ 1{,}00468 & 1{,}00772 & 1{,}00468 \end{bmatrix}$$

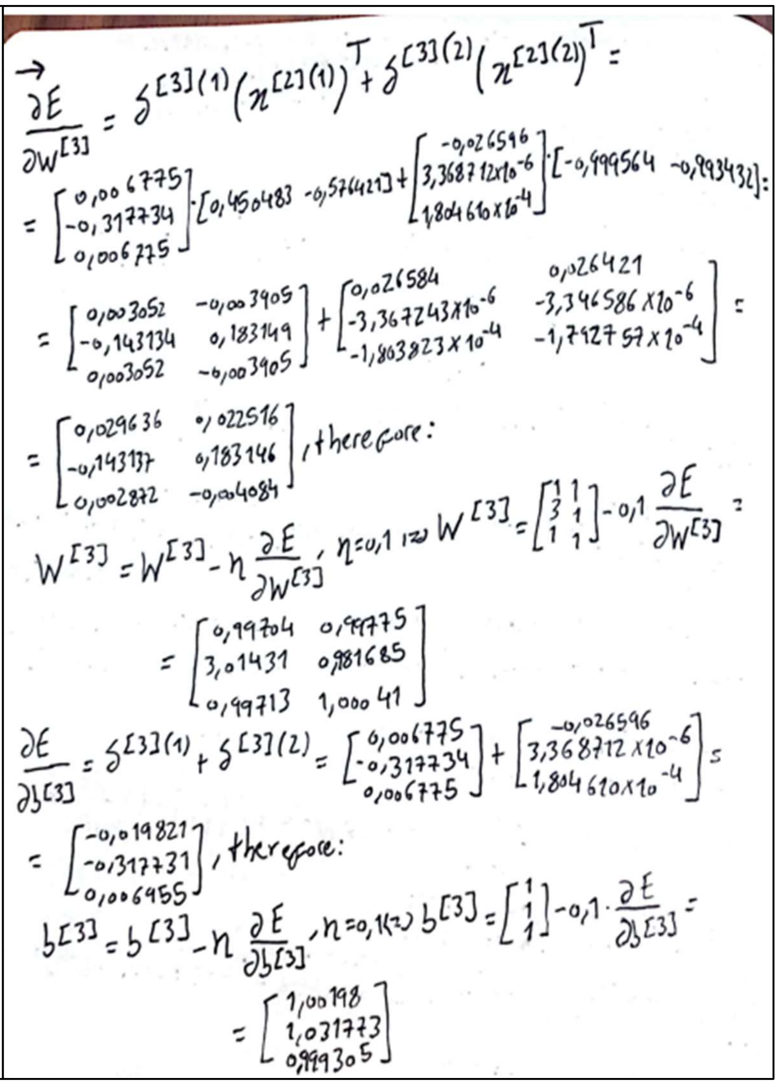
$$\frac{\partial E}{\partial b^{[2]}} = \delta^{[2](1)} + \delta^{[2](2)} = \begin{bmatrix} -0{,}374482 \\ -0{,}101558 \end{bmatrix} + \begin{bmatrix} -1{,}150926\times10^{-5} \\ -1{,}728993\times10^{-4} \end{bmatrix} =$$

$$= \begin{bmatrix} -0{,}3744935 \\ -0{,}101731 \end{bmatrix}, \text{therefore:}$$

$$b^{[2]} = b^{[2]} - \eta \frac{\partial E}{\partial b^{[2]}}, \eta = 0{,}1 \Leftrightarrow b^{[2]} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 0{,}1 \frac{\partial E}{\partial b^{[2]}} =$$

$$= \begin{bmatrix} 1{,}03745 \\ 1{,}01017 \end{bmatrix}$$

$\rightarrow$



$$\rightarrow$$

$$\frac{\partial E}{\partial W^{[3]}} = \delta^{[3](1)} \left(x^{[2](1)}\right)^T + \delta^{[3](2)} \left(x^{[2](2)}\right)^T =$$

$$= \begin{bmatrix} 0{,}006775 \\ -0{,}317734 \\ 0{,}006775 \end{bmatrix} \cdot \begin{bmatrix} 0{,}450483 & -0{,}576421 \end{bmatrix} + \begin{bmatrix} -0{,}026596 \\ 3{,}368712\times10^{-6} \\ 1{,}804610\times10^{-4} \end{bmatrix} \cdot \begin{bmatrix} -0{,}999564 & -0{,}993432 \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}003052 & -0{,}003905 \\ -0{,}143134 & 0{,}183149 \\ 0{,}003052 & -0{,}003905 \end{bmatrix} + \begin{bmatrix} 0{,}026584 & 0{,}026421 \\ -3{,}367243\times10^{-6} & -3{,}346586\times10^{-6} \\ -1{,}803823\times10^{-4} & -1{,}792757\times10^{-4} \end{bmatrix} =$$

$$= \begin{bmatrix} 0{,}029636 & 0{,}022516 \\ -0{,}143137 & 0{,}183146 \\ 0{,}002872 & -0{,}004084 \end{bmatrix}, \text{therefore:}$$

$$W^{[3]} = W^{[3]} - \eta \frac{\partial E}{\partial W^{[3]}}, \eta = 0{,}1 \Leftrightarrow W^{[3]} = \begin{bmatrix} 1 & 1 \\ 3 & 1 \\ 1 & 1 \end{bmatrix} - 0{,}1 \frac{\partial E}{\partial W^{[3]}} =$$

$$= \begin{bmatrix} 0{,}99704 & 0{,}99775 \\ 3{,}01431 & 0{,}981685 \\ 0{,}99713 & 1{,}00041 \end{bmatrix}$$

$$\frac{\partial E}{\partial b^{[3]}} = \delta^{[3](1)} + \delta^{[3](2)} = \begin{bmatrix} 0{,}006775 \\ -0{,}317734 \\ 0{,}006775 \end{bmatrix} + \begin{bmatrix} -0{,}026596 \\ 3{,}368712\times10^{-6} \\ 1{,}804610\times10^{-4} \end{bmatrix} =$$

$$= \begin{bmatrix} -0{,}019821 \\ -0{,}317731 \\ 0{,}006955 \end{bmatrix}, \text{therefore:}$$

$$b^{[3]} = b^{[3]} - \eta \frac{\partial E}{\partial b^{[3]}}, \eta = 0{,}1 \Leftrightarrow b^{[3]} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - 0{,}1 \cdot \frac{\partial E}{\partial b^{[3]}} =$$

$$= \begin{bmatrix} 1{,}00198 \\ 1{,}031773 \\ 0{,}993305 \end{bmatrix}$$

TÉCNICO LISBOA

Aprendizagem 2023/24

Homework III

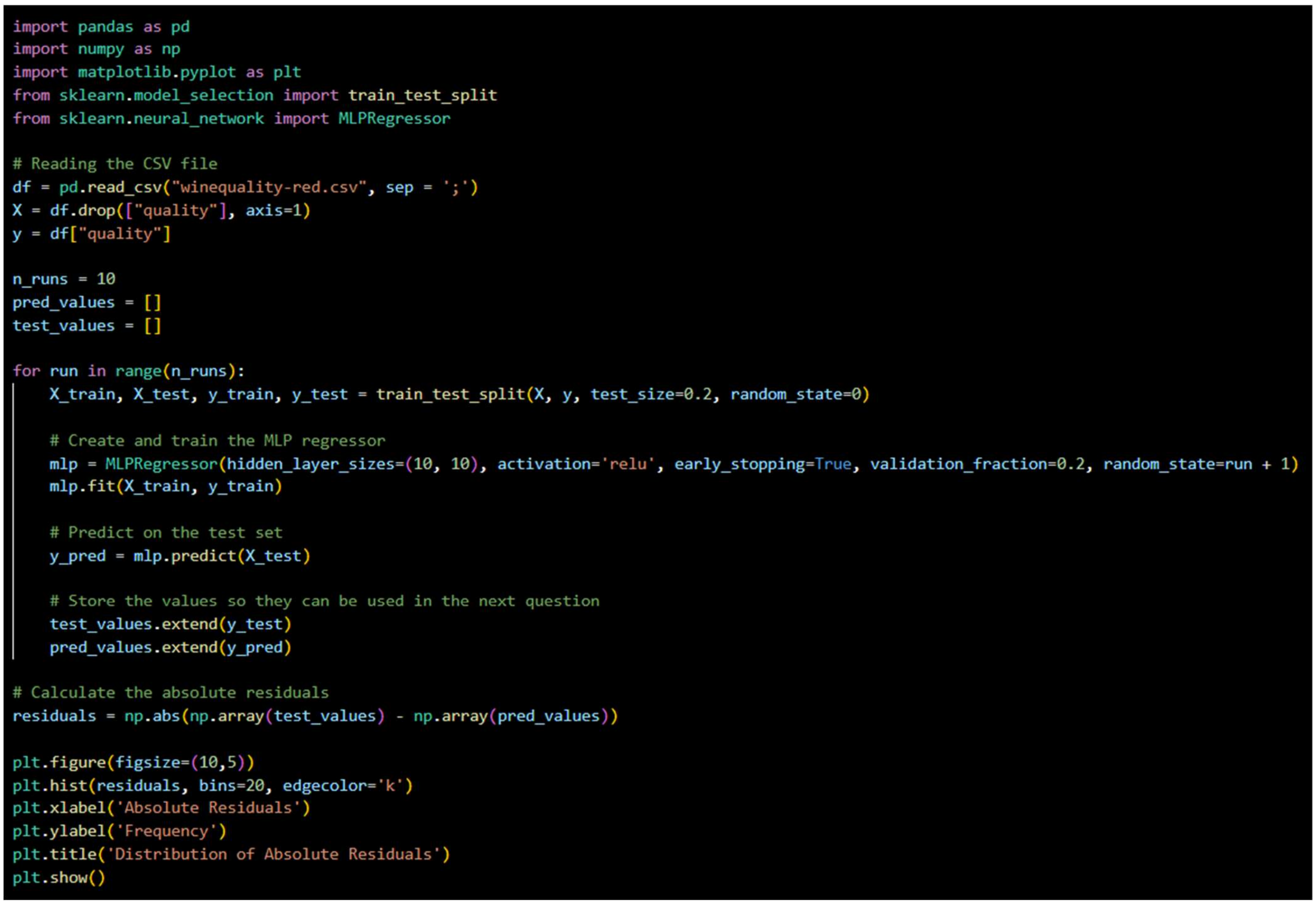
II. Programming and critical analysis [8v]

Consider the winequality-red.csv dataset (available at the webpage) where the goal is to estimate the quality (sensory appreciation) of a wine based on physicochemical inputs.

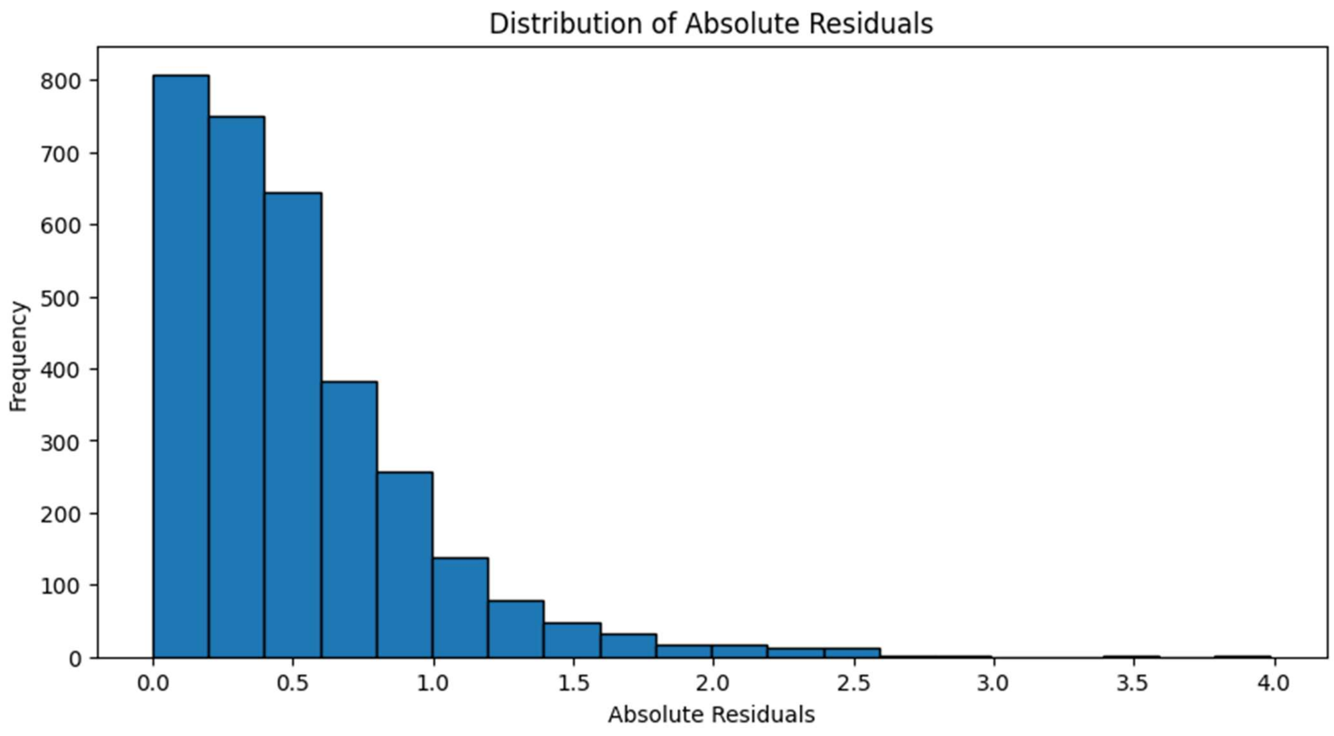
Using a 80-20 training-test split with a fixed seed (random_state=0), you are asked to learn MLP regressors to answer the following questions.

Given their stochastic behavior, average the performance of each MLP from 10 runs (for reproducibility consider seeding the MLPs with random_state $\in \{1..10\}$).

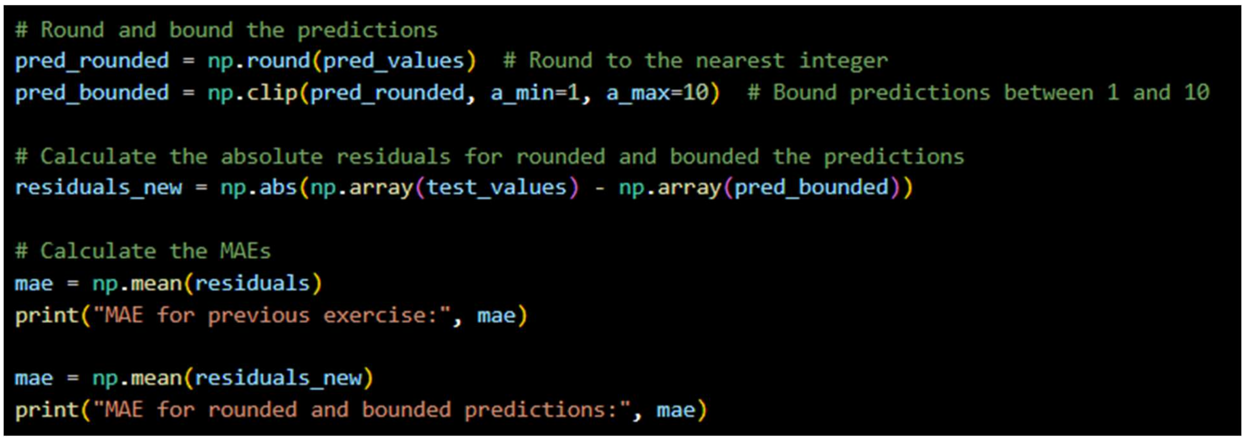
1) [3.5v] Learn a MLP regressor with 2 hidden layers of size 10, rectifier linear unit activation on all nodes, and early stopping with 20% of training data set aside for validation. All remaining parameters (e.g., loss, batch size, regularization term, solver) should be set as default. Plot the distribution of the residues (in absolute value) using a histogram.



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPRegressor

# Reading the CSV file
df = pd.read_csv("winequality-red.csv", sep = ';')
X = df.drop(["quality"], axis=1)
y = df["quality"]

n_runs = 10
pred_values = []
test_values = []

for run in range(n_runs):
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

    # Create and train the MLP regressor
    mlp = MLPRegressor(hidden_layer_sizes=(10, 10), activation='relu', early_stopping=True, validation_fraction=0.2, random_state=run + 1)
    mlp.fit(X_train, y_train)

    # Predict on the test set
    y_pred = mlp.predict(X_test)

    # Store the values so they can be used in the next question
    test_values.extend(y_test)
    pred_values.extend(y_pred)

# Calculate the absolute residuals
residuals = np.abs(np.array(test_values) - np.array(pred_values))

plt.figure(figsize=(10,5))
plt.hist(residuals, bins=20, edgecolor='k')
plt.xlabel('Absolute Residuals')
plt.ylabel('Frequency')
plt.title('Distribution of Absolute Residuals')
plt.show()
```



2) [1.5v] Since we are in the presence of a *integer regression* task, a recommended trick is to round and bound estimates. Assess the impact of these operations on the MAE of the MLP learnt in previous question.



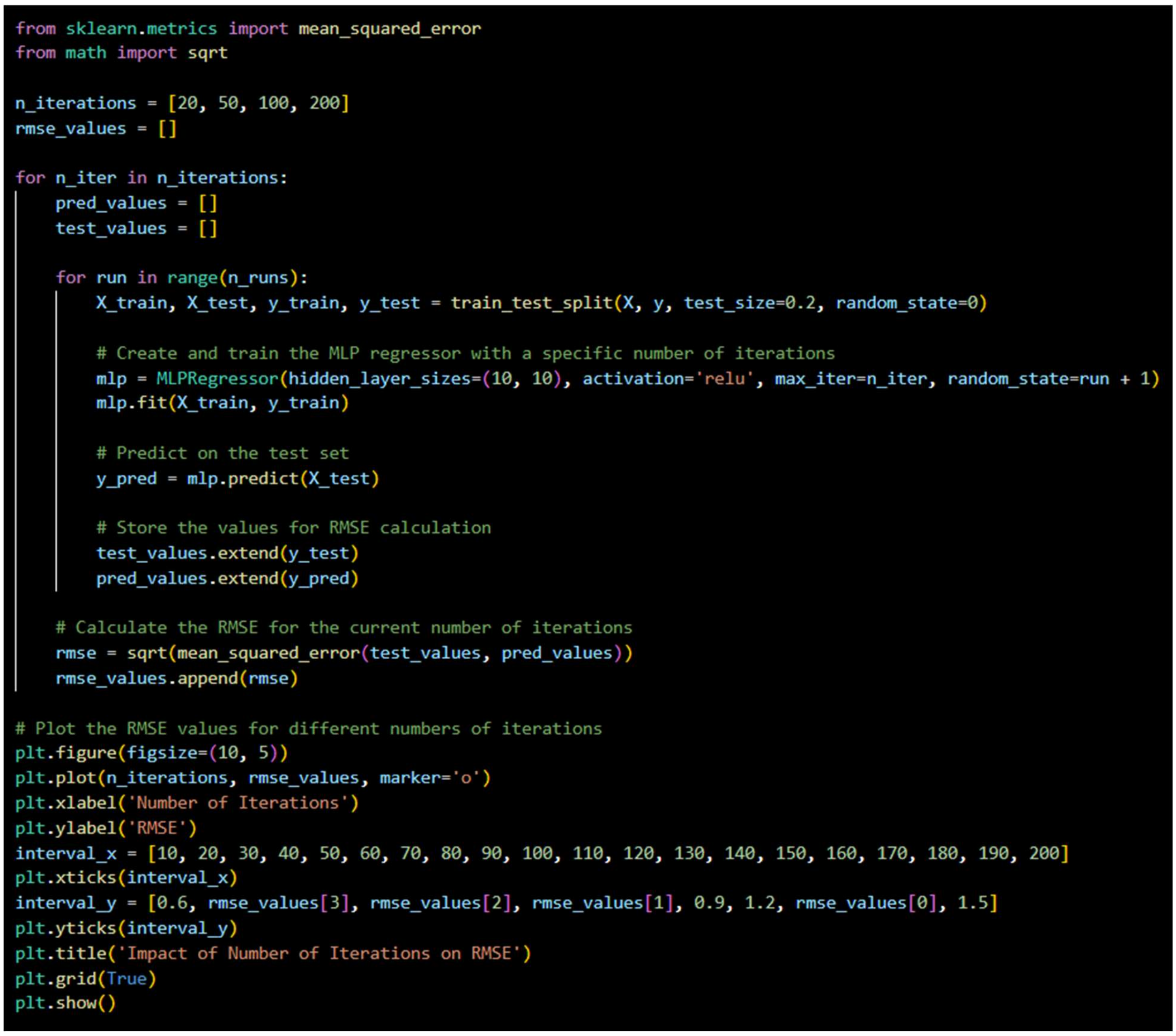
```
# Round and bound the predictions
pred_rounded = np.round(pred_values)  # Round to the nearest integer
pred_bounded = np.clip(pred_rounded, a_min=1, a_max=10)  # Bound predictions between 1 and 10

# Calculate the absolute residuals for rounded and bounded the predictions
residuals_new = np.abs(np.array(test_values) - np.array(pred_bounded))

# Calculate the MAEs
mae = np.mean(residuals)
print("MAE for previous exercise:", mae)

mae = np.mean(residuals_new)
print("MAE for rounded and bounded predictions:", mae)
```

```
MAE for previous exercise: 0.5097171955009514
MAE for rounded and bounded predictions: 0.43875
```

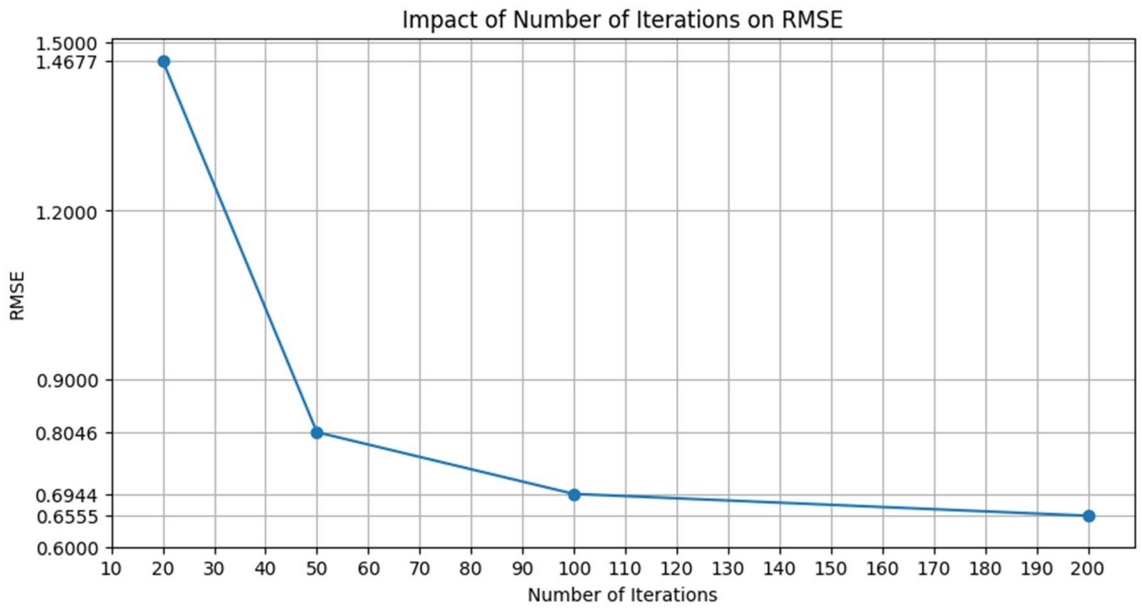
Comment about results:

By rounding and bounding the estimates, it is observed that the MAE (Mean Absolute Error) reduces in relation to the previous exercise, which indicates that there was a better performance of this new model.

3) [1.5v] Similarly assess the impact on RMSE from replacing early stopping by a well-defined number of iterations in {20,50,100,200} (where one iteration corresponds to a batch).



```python
from sklearn.metrics import mean_squared_error
from math import sqrt

n_iterations = [20, 50, 100, 200]
rmse_values = []

for n_iter in n_iterations:
    pred_values = []
    test_values = []

    for run in range(n_runs):
        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

        # Create and train the MLP regressor with a specific number of iterations
        mlp = MLPRegressor(hidden_layer_sizes=(10, 10), activation='relu', max_iter=n_iter, random_state=run + 1)
        mlp.fit(X_train, y_train)

        # Predict on the test set
        y_pred = mlp.predict(X_test)

        # Store the values for RMSE calculation
        test_values.extend(y_test)
        pred_values.extend(y_pred)

    # Calculate the RMSE for the current number of iterations
    rmse = sqrt(mean_squared_error(test_values, pred_values))
    rmse_values.append(rmse)

# Plot the RMSE values for different numbers of iterations
plt.figure(figsize=(10, 5))
plt.plot(n_iterations, rmse_values, marker='o')
plt.xlabel('Number of Iterations')
plt.ylabel('RMSE')
interval_x = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200]
plt.xticks(interval_x)
interval_y = [0.6, rmse_values[3], rmse_values[2], rmse_values[1], 0.9, 1.2, rmse_values[0], 1.5]
plt.yticks(interval_y)
plt.title('Impact of Number of Iterations on RMSE')
plt.grid(True)
plt.show()
```



4) [1.5v] Critically comment the results obtained in previous question, hypothesizing at least one reason why early stopping favors and/or worsens performance.

When the number of iterations is low (20 iterations), the model doesn't have a chance to adjust the weights to completely fit the training data, which leads to a relatively high RMSE (≈1.47). As the number of iterations increases, the model has more opportunities to adjust the weights and reduce the error in the training data, thus decreasing the RMSE. However, if training continues for too many iterations, the model may start to adjust too much to the training data. As it is possible to observe from the exercise above, for a number of iterations equal to 200 a relatively low RMSE (≈0.66) was obtained. When this happens, there may be an increase in error in the test data, thus being in the presence of overfitting.

With this, we conclude that early stopping, on the one hand, is an effective technique to avoid overfitting. However, when we apply this technique, the model may not be able to converge to its optimal performance, due to its training being stopped too early.

END