3x5 2x3

1) Consider a network with three layers: 5 inputs, 3 hidden units and 2 outputs where all units use a sigmoid activation function.

a) Initialize all connection weights to 0.1 and all biases to 0. Using the squared error loss do a **stochastic gradient descent** update (with learning rate $\eta = 1$) for the training example

$$\left\{\mathbf{x} = \begin{pmatrix}1 & 1 & 0 & 0 & 0\end{pmatrix}^T, \mathbf{t} = \begin{pmatrix}1\\0\end{pmatrix}\right\}$$

Solution:

We start by writting the connection weights and the biases:

1 Initialize

$$\mathbf{W}^{[1]} = \begin{pmatrix}0.1 & 0.1 & 0.1 & 0.1 & 0.1\\0.1 & 0.1 & 0.1 & 0.1 & 0.1\\0.1 & 0.1 & 0.1 & 0.1 & 0.1\end{pmatrix} \quad (3\times5)$$

$$\mathbf{b}^{[1]} = \begin{pmatrix}0\\0\\0\end{pmatrix} \quad (3\times1)$$

$$\mathbf{W}^{[2]} = \begin{pmatrix}0.1 & 0.1 & 0.1\\0.1 & 0.1 & 0.1\end{pmatrix} \quad (2\times3)$$

$$\mathbf{b}^{[2]} = \begin{pmatrix}0\\0\end{pmatrix} \quad (2\times1)$$

We are now ready to do forward propagation:

2 Forward Propagation

$$\mathbf{x}^{[0]} = \begin{pmatrix}1\\1\\0\\0\\0\end{pmatrix}$$

$$\mathbf{z}^{[1]} = \begin{pmatrix}0.1 & 0.1 & 0.1 & 0.1 & 0.1\\0.1 & 0.1 & 0.1 & 0.1 & 0.1\\0.1 & 0.1 & 0.1 & 0.1 & 0.1\end{pmatrix}\begin{pmatrix}1\\1\\0\\0\\0\end{pmatrix} + \begin{pmatrix}0\\0\\0\end{pmatrix} = \begin{pmatrix}0.2\\0.2\\0.2\end{pmatrix}$$

$$\mathbf{x}^{[1]} = \sigma\left(\begin{pmatrix}0.2\\0.2\\0.2\end{pmatrix}\right) = \begin{pmatrix}\sigma(0.2)\\\sigma(0.2)\\\sigma(0.2)\end{pmatrix}$$

$$\mathbf{z}^{[2]} = \begin{pmatrix}0.1 & 0.1 & 0.1\\0.1 & 0.1 & 0.1\end{pmatrix}\begin{pmatrix}\sigma(0.2)\\\sigma(0.2)\\\sigma(0.2)\end{pmatrix} + \begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}0.16495\\0.16495\end{pmatrix}$$

$$\mathbf{x}^{[2]} = \sigma\left(\begin{pmatrix}0.16495\\0.16495\end{pmatrix}\right) = \begin{pmatrix}\sigma(0.16495)\\\sigma(0.16495)\end{pmatrix} = \begin{pmatrix}0.5411\\0.5411\end{pmatrix}$$

Now we want to do the backward phase. Recall the squared error measure:

$$E\left(\mathbf{t}, \mathbf{x}^{[2]}\right) = \frac{1}{2}\sum_{i=1}^{1}\left(\mathbf{x}^{[2]} - \mathbf{t}\right)^2 = \frac{1}{2}\left(\mathbf{x}^{[2]} - \mathbf{t}\right)^2$$

In general, we will need to know how to derive all functions in our network.

$$\frac{\partial E}{\partial \mathbf{x}^{[2]}}\left(\mathbf{t}, \mathbf{x}^{[2]}\right) = \frac{1}{2}\left[2\left(\mathbf{x}^{[2]} - \mathbf{t}\right)\right] = \mathbf{x}^{[2]} - \mathbf{t}$$

$$x^{(l)}\left(z^{(l)}\right) = \sigma\left(z^{(l)}\right)$$

$$\frac{\partial \mathbf{x}^{[l]}}{\partial \mathbf{z}^{[l]}}\left(\mathbf{z}^{[l]}\right) = \sigma\left(\mathbf{z}^{[l]}\right)\left(1 - \sigma\left(\mathbf{z}^{[l]}\right)\right)$$

3 Necessary Gradients

$$z^{(l)}\left(w^{(l)}, b^{(l)}, x^{(l-1)}\right) = w^{(l)}x^{(l-1)} + b^{(l)}$$

$$\frac{\partial \mathbf{z}^{[l]}}{\partial \mathbf{W}^{[l]}}\left(\mathbf{W}^{[l]}, \mathbf{b}^{[l]}, \mathbf{x}^{[l-1]}\right) = \mathbf{x}^{[l-1]}$$

$$\frac{\partial \mathbf{z}^{[l]}}{\partial \mathbf{b}^{[l]}}\left(\mathbf{W}^{[l]}, \mathbf{b}^{[l]}, \mathbf{x}^{[l-1]}\right) = 1$$

$$\frac{\partial \mathbf{z}^{[l]}}{\partial \mathbf{x}^{[l-1]}}\left(\mathbf{W}^{[l]}, \mathbf{b}^{[l]}, \mathbf{x}^{[l-1]}\right) = \mathbf{W}^{[l]}$$

4 Delta from last layer

To start the recursion, we need the delta from the last layer:

$$\begin{aligned}\delta^{[2]} &= \frac{\partial E}{\partial \mathbf{x}^{[2]}} \circ \frac{\partial \mathbf{x}^{[2]}}{\partial \mathbf{z}^{[2]}}\\ &= \left(\mathbf{x}^{[2]} - \mathbf{t}\right) \circ \sigma\left(\mathbf{z}^{[2]}\right)\left(1 - \sigma\left(\mathbf{z}^{[2]}\right)\right)\\ &= \left(\begin{pmatrix}0.5411\\0.5411\end{pmatrix} - \begin{pmatrix}1\\0\end{pmatrix}\right) \circ \begin{pmatrix}0.16495\\0.16495\end{pmatrix} \circ \left(1 - \begin{pmatrix}0.16495\\0.16495\end{pmatrix}\right)\\ &= \begin{pmatrix}-0.11394\\0.13437\end{pmatrix}\end{aligned}$$

5 Delta for hidden layer

Now, we can use the recursion to compute the delta from the hidden layer:

$$\begin{aligned}\delta^{[1]} &= \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{x}^{[1]}}^T \cdot \delta^{[2]} \circ \frac{\partial \mathbf{x}^{[l]}}{\partial \mathbf{z}^{[l]}}\\ &= \left(\mathbf{W}^{[2]}\right)^T \cdot \delta^{[2]} \circ \sigma\left(\mathbf{z}^{[1]}\right) \circ \left(1 - \sigma\left(\mathbf{z}^{[1]}\right)\right)\\ &= \left(\mathbf{W}^{[2]}\right)^T \cdot \delta^{[2]} \circ \sigma\left(\mathbf{z}^{[1]}\right) \circ \left(1 - \sigma\left(\mathbf{z}^{[1]}\right)\right)\\ &= \begin{pmatrix}0.1 & 0.1\\0.1 & 0.1\\0.1 & 0.1\end{pmatrix} \cdot \begin{pmatrix}-0.11394\\0.13437\end{pmatrix} \circ \sigma\left(\begin{pmatrix}0.2\\0.2\\0.2\end{pmatrix}\right) \circ \left(1 - \sigma\left(\begin{pmatrix}0.2\\0.2\\0.2\end{pmatrix}\right)\right)\\ &= \begin{pmatrix}0.00050575\\0.00050575\\0.00050575\end{pmatrix}\end{aligned}$$

6 Update values

Finally, we can go to the last phase and perform the updates. We start with the first layer:

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{W}^{[1]}} &= \delta^{[1]} \cdot \frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{W}^{[1]}}^T \\ &= \delta^{[1]} \cdot \left(\mathbf{x}^{[0]}\right)^T \\ &= \begin{pmatrix} 0.00050575 \\ 0.00050575 \\ 0.00050575 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0.00050575 & 0.00050575 & 0 & 0 & 0 \\ 0.00050575 & 0.00050575 & 0 & 0 & 0 \\ 0.00050575 & 0.00050575 & 0 & 0 & 0 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{W}^{[1]} &= \mathbf{W}^{[1]} - \eta \frac{\partial E}{\partial \mathbf{W}^{[1]}} \\ &= \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix} - 1 \begin{pmatrix} 0.00050575 & 0.00050575 & 0 & 0 & 0 \\ 0.00050575 & 0.00050575 & 0 & 0 & 0 \\ 0.00050575 & 0.00050575 & 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0.09949 & 0.09949 & 0.1 & 0.1 & 0.1 \\ 0.09949 & 0.09949 & 0.1 & 0.1 & 0.1 \\ 0.09949 & 0.09949 & 0.1 & 0.1 & 0.1 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{b}^{[1]}} &= \delta^{[1]} \cdot \frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{b}^{[1]}}^T \\ &= \delta^{[1]} \\ &= \begin{pmatrix} 0.00050575 \\ 0.00050575 \\ 0.00050575 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{b}^{[1]} &= \mathbf{b}^{[1]} - \eta \frac{\partial E}{\partial \mathbf{b}^{[1]}} \\ &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} - 1 \begin{pmatrix} 0.00050575 \\ 0.00050575 \\ 0.00050575 \end{pmatrix} \\ &= \begin{pmatrix} -0.00050575 \\ -0.00050575 \\ -0.00050575 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{W}^{[2]}} &= \delta^{[2]} \cdot \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{W}^{[2]}}^T \\ &= \delta^{[2]} \cdot \left(\mathbf{x}^{[1]}\right)^T \\ &= \begin{pmatrix} -0.11394 \\ 0.13437 \end{pmatrix} \cdot \begin{pmatrix} \sigma(0.2) & \sigma(0.2) & \sigma(0.2) \end{pmatrix} \\ &= \begin{pmatrix} -0.062647 & -0.062647 & -0.062647 \\ 0.073881 & 0.073881 & 0.073881 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{W}^{[2]} &= \mathbf{W}^{[2]} - \eta \frac{\partial E}{\partial \mathbf{W}^{[2]}} \\ &= \begin{pmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix} - 1 \begin{pmatrix} -0.062647 & -0.062647 & -0.062647 \\ 0.073881 & 0.073881 & 0.073881 \end{pmatrix} \\ &= \begin{pmatrix} 0.162647 & 0.162647 & 0.162647 \\ 0.026119 & 0.026119 & 0.026119 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{b}^{[2]}} &= \delta^{[2]} \cdot \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{b}^{[2]}}^T \\ &= \delta^{[2]} \\ &= \begin{pmatrix} -0.11394 \\ 0.13437 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{b}^{[2]} &= \mathbf{b}^{[2]} - \eta \frac{\partial E}{\partial \mathbf{b}^{[2]}} \\ &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} - 1 \begin{pmatrix} -0.11394 \\ 0.13437 \end{pmatrix} \\ &= \begin{pmatrix} 0.11394 \\ -0.13437 \end{pmatrix}\end{aligned}$$

$$\frac{\partial E}{\partial \mathbf{b}^{[1]}} = \delta^{[1]} \cdot \frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{b}^{[1]}}^T$$

b) Compute the MLP class for the query point $\mathbf{x} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \end{pmatrix}^T$.

Solution:
We use the weights and biases from the previous exercise:

$$\mathbf{W}^{[1]} = \begin{pmatrix} 0.09949 & 0.09949 & 0.1 & 0.1 & 0.1 \\ 0.09949 & 0.09949 & 0.1 & 0.1 & 0.1 \\ 0.09949 & 0.09949 & 0.1 & 0.1 & 0.1 \end{pmatrix}$$

$$\mathbf{b}^{[1]} = \begin{pmatrix} 0.00050575 \\ 0.00050575 \\ 0.00050575 \end{pmatrix}$$

$$\mathbf{W}^{[2]} = \begin{pmatrix} -0.062647 & -0.062647 & -0.062647 \\ 0.073881 & 0.073881 & 0.073881 \end{pmatrix}$$

$$\mathbf{b}^{[2]} = \begin{pmatrix} 0.11394 \\ -0.13437 \end{pmatrix}$$

Just perform forward propagation like before, starting with the query point.

$$\mathbf{x}^{[0]} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

$$\mathbf{z}^{[1]} = \begin{pmatrix} 0.09949 & 0.09949 & 0.1 & 0.1 & 0.1 \\ 0.09949 & 0.09949 & 0.1 & 0.1 & 0.1 \\ 0.09949 & 0.09949 & 0.1 & 0.1 & 0.1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 0.00050575 \\ 0.00050575 \\ 0.00050575 \end{pmatrix} = \begin{pmatrix} 0.19949 \\ 0.19949 \\ 0.19949 \end{pmatrix}$$

$$\mathbf{x}^{[1]} = \sigma\left(\begin{pmatrix} 0.19949 \\ 0.19949 \\ 0.19949 \end{pmatrix}\right) = \begin{pmatrix} 0.54971 \\ 0.54971 \\ 0.54971 \end{pmatrix}$$

$$\mathbf{z}^{[2]} = \begin{pmatrix} -0.062647 & -0.062647 & -0.062647 \\ 0.073881 & 0.073881 & 0.073881 \end{pmatrix} \begin{pmatrix} 0.54971 \\ 0.54971 \\ 0.54971 \end{pmatrix} + \begin{pmatrix} 0.11394 \\ -0.13437 \end{pmatrix} = \begin{pmatrix} 0.382162 \\ -0.091297 \end{pmatrix}$$

$$\mathbf{x}^{[2]} = \sigma\left(\begin{pmatrix} 0.382162 \\ -0.091297 \end{pmatrix}\right) = \begin{pmatrix} 0.59439 \\ 0.47719 \end{pmatrix}$$

Now, we just need to choose the label with highest output:

$$label = \arg\max_i \mathbf{x}^{[2]} = \arg\max_i \begin{pmatrix} \mathbf{x}_1^{[2]} \\ \mathbf{x}_2^{[2]} \end{pmatrix} = \arg\max_i \begin{pmatrix} \boxed{0.59439} \\ 0.47719 \end{pmatrix} = 1$$

2) Consider a network with four layers with the following numbers of units: $4, 4, 3, 3$. Assume all units use the hyperbolic tangent activation function.

a) Initialize all connection weights and biases to 0.1. Using the squared error loss do a **stochastic gradient descent** update (with learning rate $\eta = 0.1$) for the training example:

$$\left\{ \mathbf{x} = (1\ 0\ 1\ 0)^T, \mathbf{t} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}$$

$$\frac{d}{dx} \tanh(x) = 1 - \tanh^2(x)$$

We start by writting the connection weights and the biases:

$$\mathbf{W}^{[1]} = \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix}$$

$$\mathbf{b}^{[1]} = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{pmatrix}$$

$$\mathbf{W}^{[2]} = \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix}$$

$$\mathbf{b}^{[2]} = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix}$$

$$\mathbf{W}^{[3]} = \begin{pmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix}$$

$$\mathbf{b}^{[3]} = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix}$$

We are now ready to do forward propagation:

(2)

$$\mathbf{x}^{[0]} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

$$\mathbf{z}^{[1]} = \begin{pmatrix} 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} = \begin{pmatrix} 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \end{pmatrix}$$

$$\mathbf{x}^{[1]} = tanh\left(\begin{pmatrix} 0.3 \\ 0.3 \\ 0.3 \\ 0.3 \end{pmatrix}\right) = \begin{pmatrix} 0.2913 \\ 0.2913 \\ 0.2913 \\ 0.2913 \end{pmatrix}$$

$$\mathbf{z}^{[2]} = \begin{pmatrix} 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \end{pmatrix} \begin{pmatrix} 0.2913 \\ 0.2913 \\ 0.2913 \\ 0.2913 \end{pmatrix} + \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} = \begin{pmatrix} 0.2165 \\ 0.2165 \\ 0.2165 \end{pmatrix}$$

$$\mathbf{x}^{[2]} = tanh\left(\begin{pmatrix} 0.2165 \\ 0.2165 \\ 0.2165 \end{pmatrix}\right) = \begin{pmatrix} 0.2132 \\ 0.2132 \\ 0.2132 \end{pmatrix}$$

$$\mathbf{z}^{[3]} = \begin{pmatrix} 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \end{pmatrix} \begin{pmatrix} 0.2132 \\ 0.2132 \\ 0.2132 \end{pmatrix} + \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} = \begin{pmatrix} 0.16396 \\ 0.16396 \\ 0.16396 \end{pmatrix}$$

$$\mathbf{x}^{[3]} = tanh\left(\begin{pmatrix} 0.16396 \\ 0.16396 \\ 0.16396 \end{pmatrix}\right) = \begin{pmatrix} 0.16251 \\ 0.16251 \\ 0.16251 \end{pmatrix}$$

Now we want to do the backward phase. Recall the squared error measure:

$$E\left(\mathbf{t}, \mathbf{x}^{[L]}\right) = \frac{1}{2}\sum_{i=1}^{1}\left(\mathbf{x}^{[L]} - \mathbf{t}\right)^2 = \frac{1}{2}\left(\mathbf{x}^{[L]} - \mathbf{t}\right)^2$$

In general, we will need to know how to derive all functions in our network. Let us compute them beforehand:

$$\frac{\partial E}{\partial \mathbf{x}^{[L]}}\left(\mathbf{t}, \mathbf{x}^{[L]}\right) = \frac{\partial E}{\partial \left(\mathbf{x}^{[L]} - \mathbf{t}\right)^2}\frac{\partial \left(\mathbf{x}^{[L]} - \mathbf{t}\right)^2}{\partial \left(\mathbf{x}^{[L]} - \mathbf{t}\right)}\frac{\partial \left(\mathbf{x}^{[L]} - \mathbf{t}\right)}{\partial \mathbf{x}^{[L]}} = \frac{1}{2}\left[2\left(\mathbf{x}^{[L]} - \mathbf{t}\right)\right] = \mathbf{x}^{[L]} - \mathbf{t}$$

(3)

$$\frac{\partial \mathbf{x}^{[l]}}{\partial \mathbf{z}^{[l]}}\left(\mathbf{z}^{[l]}\right) = 1 - tanh\left(\mathbf{z}^{[l]}\right)^2$$

$$\frac{\partial \mathbf{z}^{[l]}}{\partial \mathbf{W}^{[l]}}\left(\mathbf{W}^{[l]}, \mathbf{b}^{[l]}, \mathbf{x}^{[l-1]}\right) = \mathbf{x}^{[l-1]}$$

$$\frac{\partial \mathbf{z}^{[l]}}{\partial \mathbf{b}^{[l]}}\left(\mathbf{W}^{[l]}, \mathbf{b}^{[l]}, \mathbf{x}^{[l-1]}\right) = 1$$

$$\frac{\partial \mathbf{z}^{[l]}}{\partial \mathbf{x}^{[l-1]}}\left(\mathbf{W}^{[l]}, \mathbf{b}^{[l]}, \mathbf{x}^{[l-1]}\right) = \mathbf{W}^{[l]}$$

(4)

To start the recursion, we need the delta from the last layer:

$$\begin{aligned}
\delta^{[3]} &= \frac{\partial E}{\partial \mathbf{x}^{[3]}} \circ \frac{\partial \mathbf{x}^{[3]}}{\partial \mathbf{z}^{[3]}} \\
&= \left(\mathbf{x}^{[3]} - \mathbf{t}\right) \circ \left(1 - tanh\left(\mathbf{z}^{[3]}\right)^2\right) \\
&= \left(\begin{pmatrix} 0.16251 \\ 0.16251 \\ 0.16251 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}\right) \circ \left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} \tanh(0.16396)^2 \\ \tanh(0.16396)^2 \\ \tanh(0.16396)^2 \end{pmatrix}\right) \\
&= \begin{pmatrix} 0.15822 \\ -0.81538 \\ 0.15822 \end{pmatrix}
\end{aligned}$$

Now, we can use the recursion to compute the delta from the hidden layers:

$$
\begin{aligned}
\delta^{[2]} &= \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{x}^{[2]}}^T \cdot \delta^{[3]} \circ \frac{\partial \mathbf{x}^{[2]}}{\partial \mathbf{z}^{[2]}} \\
&= \left(\mathbf{W}^{[3]}\right)^T \cdot \delta^{[3]} \circ \left(1 - tanh\left(\mathbf{z}^{[2]}\right)^2\right) \\
&= \begin{pmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix} \cdot \begin{pmatrix} 0.15822 \\ -0.81538 \\ 0.15822 \end{pmatrix} \circ \left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} \tanh(0.2165)^2 \\ \tanh(0.2165)^2 \\ \tanh(0.2165)^2 \end{pmatrix}\right) \\
&= \begin{pmatrix} -0.0476 \\ -0.0476 \\ -0.0476 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\delta^{[1]} &= \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{x}^{[1]}}^T \cdot \delta^{[2]} \circ \frac{\partial \mathbf{x}^{[1]}}{\partial \mathbf{z}^{[1]}} \\
&= \left(\mathbf{W}^{[2]}\right)^T \cdot \delta^{[2]} \circ \left(1 - tanh\left(\mathbf{z}^{[1]}\right)^2\right) \\
&= \begin{pmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix} \cdot \begin{pmatrix} -0.0476 \\ -0.0476 \\ -0.0476 \end{pmatrix} \circ \left(\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} \tanh(0.3)^2 \\ \tanh(0.3)^2 \\ \tanh(0.3)^2 \\ \tanh(0.3)^2 \end{pmatrix}\right) \\
&= \begin{pmatrix} -0.0131 \\ -0.0131 \\ -0.0131 \\ -0.0131 \end{pmatrix}
\end{aligned}
$$

Finally, we can go to the last phase and perform the updates. We start with the first layer:

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{W}^{[1]}} &= \delta^{[1]} \cdot \frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{W}^{[1]}}^T \\
&= \delta^{[1]} \cdot \left(\mathbf{x}^{[0]}\right)^T \\
&= \begin{pmatrix} -0.0131 \\ -0.0131 \\ -0.0131 \\ -0.0131 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 1 & 0 \end{pmatrix} \\
&= \begin{pmatrix} -0.0131 & 0 & -0.0131 & 0 \\ -0.0131 & 0 & -0.0131 & 0 \\ -0.0131 & 0 & -0.0131 & 0 \\ -0.0131 & 0 & -0.0131 & 0 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{W}^{[1]} &= \mathbf{W}^{[1]} - \eta \frac{\partial E}{\partial \mathbf{W}^{[1]}} \\
&= \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} -0.0131 & 0 & -0.0131 & 0 \\ -0.0131 & 0 & -0.0131 & 0 \\ -0.0131 & 0 & -0.0131 & 0 \\ -0.0131 & 0 & -0.0131 & 0 \end{pmatrix} \\
&= \begin{pmatrix} 0.10131 & 0.1 & 0.10131 & 0.1 \\ 0.10131 & 0.1 & 0.10131 & 0.1 \\ 0.10131 & 0.1 & 0.10131 & 0.1 \\ 0.10131 & 0.1 & 0.10131 & 0.1 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{b}^{[1]}} &= \delta^{[1]} \cdot \frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{b}^{[1]}}^T \\
&= \delta^{[1]} \\
&= \begin{pmatrix} -0.0131 \\ -0.0131 \\ -0.0131 \\ -0.0131 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{b}^{[1]} &= \mathbf{b}^{[1]} - \eta \frac{\partial E}{\partial \mathbf{b}^{[1]}} \\
&= \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} -0.0131 \\ -0.0131 \\ -0.0131 \\ -0.0131 \end{pmatrix} \\
&= \begin{pmatrix} 0.10131 \\ 0.10131 \\ 0.10131 \\ 0.10131 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{W}^{[2]}} &= \delta^{[2]} \cdot \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{W}^{[2]}}^T \\
&= \delta^{[2]} \cdot \left(\mathbf{x}^{[1]}\right)^T \\
&= \begin{pmatrix} -0.0476 \\ -0.0476 \\ -0.0476 \end{pmatrix} \cdot \begin{pmatrix} 0.2913 & 0.2913 & 0.2913 & 0.2913 \end{pmatrix} \\
&= \begin{pmatrix} -0.01387 & -0.01387 & -0.01387 & -0.01387 \\ -0.01387 & -0.01387 & -0.01387 & -0.01387 \\ -0.01387 & -0.01387 & -0.01387 & -0.01387 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{W}^{[2]} &= \mathbf{W}^{[2]} - \eta \frac{\partial E}{\partial \mathbf{W}^{[2]}} \\
&= \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} -0.01387 & -0.01387 & -0.01387 & -0.01387 \\ -0.01387 & -0.01387 & -0.01387 & -0.01387 \\ -0.01387 & -0.01387 & -0.01387 & -0.01387 \end{pmatrix} \\
&= \begin{pmatrix} 0.101387 & 0.101387 & 0.101387 & 0.101387 \\ 0.101387 & 0.101387 & 0.101387 & 0.101387 \\ 0.101387 & 0.101387 & 0.101387 & 0.101387 \end{pmatrix}
\end{aligned}
$$

$$\frac{\partial E}{\partial \mathbf{b}^{[2]}} = \delta^{[2]} \cdot \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{b}^{[2]}}^T$$
$$= \delta^{[2]}$$
$$= \begin{pmatrix} -0.0476 \\ -0.0476 \\ -0.0476 \end{pmatrix}$$

$$\mathbf{b}^{[2]} = \mathbf{b}^{[2]} - \eta \frac{\partial E}{\partial \mathbf{b}^{[2]}}$$
$$= \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} -0.0476 \\ -0.0476 \\ -0.0476 \end{pmatrix}$$
$$= \begin{pmatrix} 0.10476 \\ 0.10476 \\ 0.10476 \end{pmatrix}$$

$$\frac{\partial E}{\partial \mathbf{W}^{[3]}} = \delta^{[3]} \cdot \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{W}^{[3]}}^T$$
$$= \delta^{[3]} \cdot \left(\mathbf{x}^{[2]}\right)^T$$
$$= \begin{pmatrix} 0.15822 \\ -0.81538 \\ 0.15822 \end{pmatrix} \cdot \begin{pmatrix} 0.2132 & 0.2132 & 0.2132 \end{pmatrix}$$
$$= \begin{pmatrix} 0.03373 & 0.03373 & 0.03373 \\ -0.17384 & -0.17384 & -0.17384 \\ 0.03373 & 0.03373 & 0.03373 \end{pmatrix}$$

$$\mathbf{W}^{[3]} = \mathbf{W}^{[3]} - \eta \frac{\partial E}{\partial \mathbf{W}^{[3]}}$$
$$= \begin{pmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0.03373 & 0.03373 & 0.03373 \\ -0.17384 & -0.17384 & -0.17384 \\ 0.03373 & 0.03373 & 0.03373 \end{pmatrix}$$
$$= \begin{pmatrix} 0.096627 & 0.096627 & 0.096627 \\ 0.117384 & 0.117384 & 0.117384 \\ 0.096627 & 0.096627 & 0.096627 \end{pmatrix}$$

$$\frac{\partial E}{\partial \mathbf{b}^{[3]}} = \delta^{[3]} \cdot \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{b}^{[3]}}^T$$
$$= \delta^{[3]}$$
$$= \begin{pmatrix} 0.15822 \\ -0.81538 \\ 0.15822 \end{pmatrix}$$

$$\mathbf{b}^{[3]} = \mathbf{b}^{[3]} - \eta \frac{\partial E}{\partial \mathbf{b}^{[3]}}$$
$$= \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0.15822 \\ -0.81538 \\ 0.15822 \end{pmatrix}$$
$$= \begin{pmatrix} 0.084178 \\ 0.181538 \\ 0.084178 \end{pmatrix}$$

b) Reusing the computations from the previous exercise do a **gradient descent** update (with learning rate $\eta = 0.1$) for the batch with the training example from the a) and the following:

$$\left\{ \mathbf{x} = \begin{pmatrix} 0 & 0 & 10 & 0 \end{pmatrix}^T, \mathbf{t} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

Solution:
Recall the squared error measure:

$$E\left(\mathbf{t}, \mathbf{x}^{[L]}\right) = \frac{1}{2} \sum_{i=1}^{2} \left(\mathbf{x}^{[L]} - \mathbf{t}\right)^2$$

Notice that the derivative of the sum is equal to the sum of the derivatives. For this reason, all we have to do is to compute the derivative for the new example and the final gradient will be the sum of both individual derivatives: the one for the new example and the one from the previous exercise.

Here, we do another iteration of gradient descent update, just like before. The difference comes in the parameter update stage.

We start by writting the connection weights and the biases:

$$\mathbf{W}^{[1]} = \begin{pmatrix} 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \end{pmatrix} \quad \mathbf{b}^{[1]} = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{pmatrix}$$

1

$$\mathbf{W}^{[2]} = \begin{pmatrix} 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \end{pmatrix} \quad \mathbf{b}^{[2]} = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix}$$

$$\mathbf{W}^{[3]} = \begin{pmatrix} 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \end{pmatrix} \quad \mathbf{b}^{[3]} = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix}$$

2

We are now ready to do forward propagation:

$$\mathbf{x}^{[0]} = \begin{pmatrix} 0 \\ 0 \\ 10 \\ 0 \end{pmatrix}$$

$$\mathbf{z}^{[1]} = \begin{pmatrix} 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 10 \\ 0 \end{pmatrix} + \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} = \begin{pmatrix} 1.1 \\ 1.1 \\ 1.1 \\ 1.1 \end{pmatrix}$$

$$\mathbf{x}^{[1]} = tanh\left(\begin{pmatrix} 1.1 \\ 1.1 \\ 1.1 \\ 1.1 \end{pmatrix}\right) = \begin{pmatrix} 0.8005 \\ 0.8005 \\ 0.8005 \\ 0.8005 \end{pmatrix}$$

$$\mathbf{z}^{[2]} = \begin{pmatrix} 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \end{pmatrix} \begin{pmatrix} 0.8005 \\ 0.8005 \\ 0.8005 \\ 0.8005 \end{pmatrix} + \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} = \begin{pmatrix} 0.4202 \\ 0.4202 \\ 0.4202 \end{pmatrix}$$

$$\mathbf{x}^{[2]} = tanh\left(\begin{pmatrix} 0.4202 \\ 0.4202 \\ 0.4202 \end{pmatrix}\right) = \begin{pmatrix} 0.3971 \\ 0.3971 \\ 0.3971 \end{pmatrix}$$

$$\mathbf{z}^{[3]} = \begin{pmatrix} 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \end{pmatrix} \begin{pmatrix} 0.3971 \\ 0.3971 \\ 0.3971 \end{pmatrix} + \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} = \begin{pmatrix} 0.2191 \\ 0.2191 \\ 0.2191 \end{pmatrix}$$

$$\mathbf{x}^{[3]} = tanh\left(\begin{pmatrix} 0.2191 \\ 0.2191 \\ 0.2191 \end{pmatrix}\right) = \begin{pmatrix} 0.2157 \\ 0.2157 \\ 0.2157 \end{pmatrix}$$

3

Derivatives already calculated in previous exercise.

4

To start the recursion, we need the delta from the last layer:

$$\begin{aligned} \delta^{[3]} &= \frac{\partial E}{\partial \mathbf{x}^{[3]}} \circ \frac{\partial \mathbf{x}^{[3]}}{\partial \mathbf{z}^{[3]}} \\ &= \left(\mathbf{x}^{[3]} - \mathbf{t}\right) \circ \left(1 - tanh\left(\mathbf{z}^{[3]}\right)^2\right) \\ &= \left(\begin{pmatrix} 0.2157 \\ 0.2157 \\ 0.2157 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\right) \circ \left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} \tanh(0.2191)^2 \\ \tanh(0.2191)^2 \\ \tanh(0.2191)^2 \end{pmatrix}\right) \\ &= \begin{pmatrix} 0.2057 \\ 0.2057 \\ -0.7478 \end{pmatrix} \end{aligned}$$

5

Now, we can use the recursion to compute the delta from the hidden layers:

$$\begin{aligned} \delta^{[2]} &= \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{x}^{[2]}}^T \cdot \delta^{[3]} \circ \frac{\partial \mathbf{x}^{[2]}}{\partial \mathbf{z}^{[2]}} \\ &= \left(\mathbf{W}^{[3]}\right)^T \cdot \delta^{[3]} \circ \left(1 - tanh\left(\mathbf{z}^{[2]}\right)^2\right) \\ &= \begin{pmatrix} 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \end{pmatrix} \cdot \begin{pmatrix} 0.2057 \\ 0.2057 \\ -0.7478 \end{pmatrix} \circ \left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} \tanh(0.4202)^2 \\ \tanh(0.4202)^2 \\ \tanh(0.4202)^2 \end{pmatrix}\right) \\ &= \begin{pmatrix} -0.0283 \\ -0.0283 \\ -0.0283 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} \delta^{[1]} &= \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{x}^{[1]}}^T \cdot \delta^{[2]} \circ \frac{\partial \mathbf{x}^{[1]}}{\partial \mathbf{z}^{[1]}} \\ &= \left(\mathbf{W}^{[2]}\right)^T \cdot \delta^{[2]} \circ \left(1 - tanh\left(\mathbf{z}^{[1]}\right)^2\right) \\ &= \begin{pmatrix} 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \end{pmatrix} \cdot \begin{pmatrix} -0.0283 \\ -0.0283 \\ -0.0283 \end{pmatrix} \circ \left(\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} \tanh(1.1)^2 \\ \tanh(1.1)^2 \\ \tanh(1.1)^2 \\ \tanh(1.1)^2 \end{pmatrix}\right) \\ &= \begin{pmatrix} -0.00305 \\ -0.00305 \\ -0.00305 \\ -0.00305 \end{pmatrix} \end{aligned}$$

Finally, we can go to the last phase and perform the updates. Recall that the gradient will be the sum of the individual gradiens!

Here comes the difference:

6

Let us start with the first layer:

$$\begin{aligned} \frac{\partial E}{\partial \mathbf{W}^{[1]}} &= \delta^{[1](1)} \cdot \frac{\partial \mathbf{z}^{[1](1)}}{\partial \mathbf{W}^{[1]}}^T + \delta^{[1](2)} \cdot \frac{\partial \mathbf{z}^{[1](2)}}{\partial \mathbf{W}^{[1]}}^T \\ &= \delta^{[1](1)} \cdot \left(\mathbf{x}^{[0](1)}\right)^T + \delta^{[1](2)} \cdot \left(\mathbf{x}^{[0](2)}\right)^T \\ &= \begin{pmatrix} -0.0131 \\ -0.0131 \\ -0.0131 \\ -0.0131 \end{pmatrix} \cdot (1\ 0\ 1\ 0) + \begin{pmatrix} -0.00305 \\ -0.00305 \\ -0.00305 \\ -0.00305 \end{pmatrix} \cdot (0\ 0\ 10\ 0) \\ &= \begin{pmatrix} -0.0131\ 0\ -0.0436\ 0\ 0 \\ -0.0131\ 0\ -0.0436\ 0\ 0 \\ -0.0131\ 0\ -0.0436\ 0\ 0 \\ -0.0131\ 0\ -0.0436\ 0\ 0 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} \mathbf{W}^{[1]} &= \mathbf{W}^{[1]} - \eta \frac{\partial E}{\partial \mathbf{W}^{[1]}} \\ &= \begin{pmatrix} 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} -0.0131\ 0\ -0.0436\ 0\ 0 \\ -0.0131\ 0\ -0.0436\ 0\ 0 \\ -0.0131\ 0\ -0.0436\ 0\ 0 \\ -0.0131\ 0\ -0.0436\ 0\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 0.10131\ 0.1\ 0.10436\ 0.1 \\ 0.10131\ 0.1\ 0.10436\ 0.1 \\ 0.10131\ 0.1\ 0.10436\ 0.1 \\ 0.10131\ 0.1\ 0.10436\ 0.1 \end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{b}^{[1]}} &= \boxed{\delta^{[1](1)} \cdot \frac{\partial \mathbf{z}^{[1](1)}}{\partial \mathbf{b}^{[1]}}^T} + \delta^{[1](2)} \cdot \frac{\partial \mathbf{z}^{[1](2)}}{\partial \mathbf{b}^{[1]}}^T \\
&= \delta^{[1](1)} + \delta^{[1](2)} \\
&= \begin{pmatrix} -0.0131 \\ -0.0131 \\ -0.0131 \\ -0.0131 \end{pmatrix} + \begin{pmatrix} -0.00305 \\ -0.00305 \\ -0.00305 \\ -0.00305 \end{pmatrix} \\
&= \begin{pmatrix} -0.0161 \\ -0.0161 \\ -0.0161 \\ -0.0161 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{b}^{[1]} &= \mathbf{b}^{[1]} - \eta \frac{\partial E}{\partial \mathbf{b}^{[1]}} \\
&= \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} -0.0161 \\ -0.0161 \\ -0.0161 \\ -0.0161 \end{pmatrix} \\
&= \begin{pmatrix} 0.10161 \\ 0.10161 \\ 0.10161 \\ 0.10161 \end{pmatrix}
\end{aligned}
$$

Now the second:

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{W}^{[2]}} &= \boxed{\delta^{[2](1)} \cdot \frac{\partial \mathbf{z}^{[2](1)}}{\partial \mathbf{W}^{[2]}}^T} + \delta^{[2](2)} \cdot \frac{\partial \mathbf{z}^{[2](2)}}{\partial \mathbf{W}^{[2]}}^T \\
&= \delta^{[2](1)} \cdot \left(\mathbf{x}^{[1](1)}\right)^T + \delta^{[2](2)} \cdot \left(\mathbf{x}^{[1](2)}\right)^T \\
&= \begin{pmatrix} -0.0476 \\ -0.0476 \\ -0.0476 \end{pmatrix} \cdot \begin{pmatrix} 0.2913 & 0.2913 & 0.2913 & 0.2913 \end{pmatrix} + \begin{pmatrix} -0.0283 \\ -0.0283 \\ -0.0283 \end{pmatrix} \cdot \begin{pmatrix} 0.8005 & 0.8005 & 0.8005 & 0.8005 \end{pmatrix} \\
&= \begin{pmatrix} -0.03656 & -0.03656 & -0.03656 & -0.03656 \\ -0.03656 & -0.03656 & -0.03656 & -0.03656 \\ -0.03656 & -0.03656 & -0.03656 & -0.03656 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{W}^{[2]} &= \mathbf{W}^{[2]} - \eta \frac{\partial E}{\partial \mathbf{W}^{[2]}} \\
&= \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} -0.03656 & -0.03656 & -0.03656 & -0.03656 \\ -0.03656 & -0.03656 & -0.03656 & -0.03656 \\ -0.03656 & -0.03656 & -0.03656 & -0.03656 \end{pmatrix} \\
&= \begin{pmatrix} 0.103656 & 0.103656 & 0.103656 & 0.103656 \\ 0.103656 & 0.103656 & 0.103656 & 0.103656 \\ 0.103656 & 0.103656 & 0.103656 & 0.103656 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{b}^{[2]}} &= \boxed{\delta^{[2](1)} \cdot \frac{\partial \mathbf{z}^{[2](1)}}{\partial \mathbf{b}^{[2]}}^T} + \delta^{[2](2)} \cdot \frac{\partial \mathbf{z}^{[2](2)}}{\partial \mathbf{b}^{[2]}}^T \\
&= \delta^{[2](1)} + \delta^{[2](2)} \\
&= \begin{pmatrix} -0.0476 \\ -0.0476 \\ -0.0476 \end{pmatrix} + \begin{pmatrix} -0.02835 \\ -0.02835 \\ -0.02835 \end{pmatrix} \\
&= \begin{pmatrix} -0.07597 \\ -0.07597 \\ -0.07597 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{b}^{[2]} &= \mathbf{b}^{[2]} - \eta \frac{\partial E}{\partial \mathbf{b}^{[2]}} \\
&= \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} -0.07597 \\ -0.07597 \\ -0.07597 \end{pmatrix} \\
&= \begin{pmatrix} 0.107597 \\ 0.107597 \\ 0.107597 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{W}^{[3]}} &= \delta^{[3](1)} \cdot \frac{\partial \mathbf{z}^{[3](1)}}{\partial \mathbf{W}^{[3]}}^T + \delta^{[3](2)} \cdot \frac{\partial \mathbf{z}^{[3](2)}}{\partial \mathbf{W}^{[3]}}^T \\
&= \delta^{[3](1)} \cdot \left(\mathbf{x}^{[2](1)}\right)^T + \delta^{[3](2)} \cdot \left(\mathbf{x}^{[2](2)}\right)^T \\
&= \begin{pmatrix} 0.15822 \\ -0.81538 \\ 0.15822 \end{pmatrix} \cdot \begin{pmatrix} 0.2132 & 0.2132 & 0.2132 \end{pmatrix} + \begin{pmatrix} 0.2057 \\ 0.2057 \\ -0.7478 \end{pmatrix} \cdot \begin{pmatrix} 0.3971 & 0.3971 & 0.3971 \end{pmatrix} \\
&= \begin{pmatrix} 0.11539 & 0.11534 & 0.11534 \\ -0.09218 & -0.09218 & -0.09218 \\ -0.26323 & -0.26323 & -0.26323 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{W}^{[3]} &= \mathbf{W}^{[3]} - \eta \frac{\partial E}{\partial \mathbf{W}^{[3]}} \\
&= \begin{pmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0.11539 & 0.11534 & 0.11534 \\ -0.09218 & -0.09218 & -0.09218 \\ -0.26323 & -0.26323 & -0.26323 \end{pmatrix} \\
&= \begin{pmatrix} 0.08846 & 0.08846 & 0.08846 \\ 0.10922 & 0.10922 & 0.10922 \\ 0.12632 & 0.12632 & 0.12632 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{b}^{[3]}} &= \delta^{[3](1)} \cdot \frac{\partial \mathbf{z}^{[3](1)}}{\partial \mathbf{b}^{[3]}}^T + \delta^{[3](2)} \cdot \frac{\partial \mathbf{z}^{[3](2)}}{\partial \mathbf{b}^{[3]}}^T \\
&= \delta^{[3](1)} + \delta^{[3](2)} \\
&= \begin{pmatrix} 0.3639 \\ -0.6097 \\ -0.5896 \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{b}^{[3]} &= \mathbf{b}^{[3]} - \eta \frac{\partial E}{\partial \mathbf{b}^{[3]}} \\
&= \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0.3639 \\ -0.6097 \\ -0.5896 \end{pmatrix} \\
&= \begin{pmatrix} 0.0636 \\ 0.1609 \\ 0.1589 \end{pmatrix}
\end{aligned}
$$