Practical Lecture 4 - Multi-layer Perceptron and the Backpropagation algorithm

Luis Sa-Couto[1] and Andreas Wichert[2]

INESC-ID, Instituto Superior Tecnico, Universidade de Lisboa
{luis.sa.couto,andreas.wichert}@tecnico.ulisboa.pt

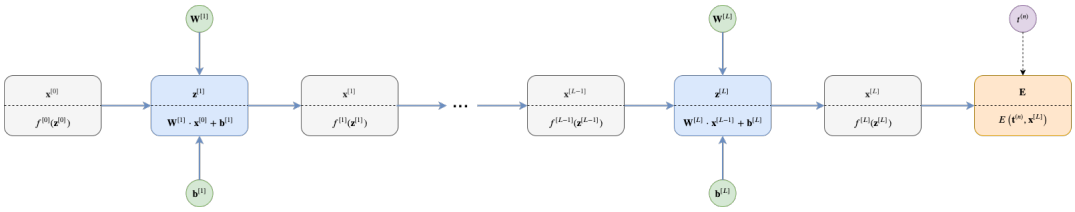
Throughout these notes we will use a superscript $[l]$ to refer to a quantity from layer l and a superscript (n) to refer to the n-th training example.

In general, the backpropagation algorithm can be described in three different phases.

Phase 1: Forward propagation

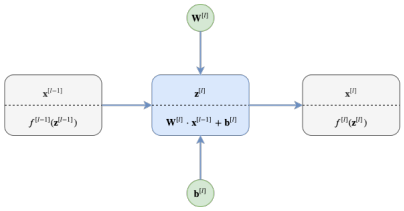
We can look at a multi-layer a perceptron as a set of functions that are applied in composition to an input. Following the figure below, for a general layer $l \in \{1, \ldots, L\}$, where L is the total number of layers, we denote the following quantities:

- $\mathbf{x}^{[l]}$ is the vector that contains the activations of the layer's units
- $\mathbf{z}^{[l]}$ is the vector that contains the net inputs to layer l units
- $\mathbf{W}^{[l]}$ is the weight matrix that makes the connections between layer $l-1$ and l. Specifically, w_{ij} is the weight that connects unit j of layer $l-1$ to unit i of layer l
- $\mathbf{b}^{[l]}$ is the bias vector that contains the bias of each unit in layer l
- $f^{[l]}$ is the activation function of the units in layer l



The figure also shows how to move forward from input $\mathbf{x}^{[0]}$ to the output $\mathbf{x}^{[L]}$ and the error computation between that output and the intended target $t^{(n)}$. The moving forward is what defines the forward propagation step. Each step follows the same logic behind the linear models we've seen before: a dot product between weights and input plus a bias. However, the multiple units in a layer serve as inputs to multiple perceptrons (units) in the next layer. So, instead of a weight vector, we have a weight matrix and instead of a bias scalar we have a bias vector.

Looking carefully at the figure we see the same pattern repeated when going from one layer to the next:



Concretely, we can write the forward propagation equations:

$$\mathbf{x}^{[0]} = Input$$
$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{x}^{[0]} + \mathbf{b}^{[1]}$$
$$\mathbf{x}^{[1]} = f^{[1]}\left(\mathbf{z}^{[1]}\right)$$
$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{x}^{[1]} + \mathbf{b}^{[2]}$$
$$\mathbf{x}^{[2]} = f^{[2]}\left(\mathbf{z}^{[2]}\right)$$
$$\vdots$$
$$\mathbf{z}^{[L-1]} = \mathbf{W}^{[L-1]}\mathbf{x}^{[L-2]} + \mathbf{b}^{[L-1]}$$
$$\mathbf{x}^{[L-1]} = f^{[L-1]}\left(\mathbf{z}^{[L-1]}\right)$$
$$\mathbf{z}^{[L]} = \mathbf{W}^{[L]}\mathbf{x}^{[L-1]} + \mathbf{b}^{[L]}$$
$$\mathbf{x}^{[L]} = f^{[L]}\left(\mathbf{z}^{[L]}\right)$$

And verify the same pattern:

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

So, to start the algorithm, for each example, we start at layer 1 and go all the way to layer L applying the forward recursion to get the final output.

Phase 2: Backward propagation

Just like in the linear models, the interesting question is how to choose the model's parameters. Like before, we apply gradient descent:

$$\mathbf{W}^{[l]} = \mathbf{W}^{[l]} - \eta\frac{\partial E}{\partial \mathbf{W}^{[l]}}$$

$$\mathbf{b}^{[l]} = \mathbf{b}^{[l]} - \eta\frac{\partial E}{\partial \mathbf{b}^{[l]}}$$

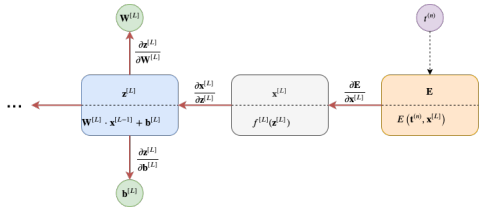
Much like before, we will need to use the chain rule. However, since we will be working with derivatives of vectors and matrices we will need a few simple rules to avoid mismatches in matrix dimensions.

1. Deriving matrix multiplication (symbol $\cdot$):
   (a) Derivative of a quantity on the right: transpose and multiply on the left

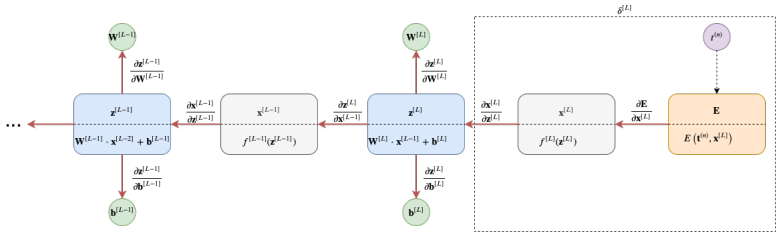
(b) Derivative of a quantity on the left: transpose and multiply on the right

2. Chain rule product corresponds to Hadamard (i.e. element-wise) product (symbol $\circ$)

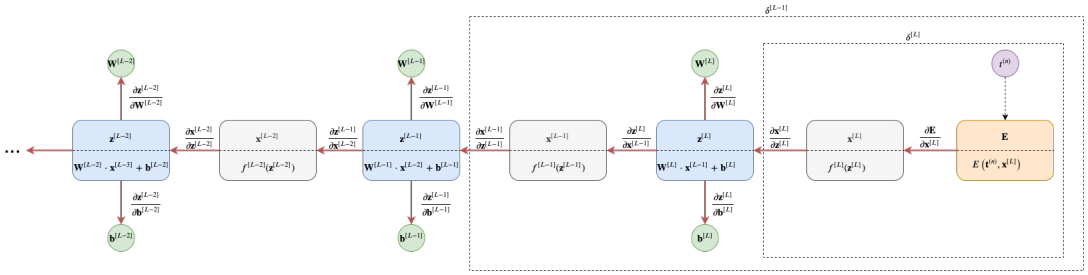
Applying the chain rule corresponds to travelling back in our network and multiplying the derivatives as we go. So, let us try to find the gradient for the last layer's parameters:



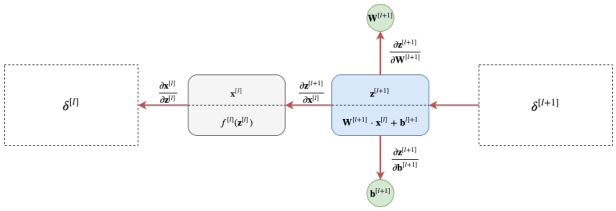
We see that it works just like the linear models. So, why stop there? We can keep going towards the layer before the last:



By doing so, we immediately notice that we can reuse part of the math we did for the last layer. We call this quantity the delta of the last layer (i.e. $\delta^{[L]}$). We can interpret it as the "error" that layer $L-1$ sees. Doing it for the next layer, we see the same pattern:



So, again, we have a delta which is the error that the current layer sees and we use it to get our gradient. The pattern keeps appearing until the first layer. So, we also have a recursion to go back... the backward recursion:



Now that we've seen it in pictures, let us write it all down. For the first layer, we have:

$$\frac{\partial E}{\partial \mathbf{W}^{[L]}} = \frac{\partial E}{\partial \mathbf{x}^{[L]}} \circ \frac{\partial \mathbf{x}^{[L]}}{\partial \mathbf{z}^{[L]}} \cdot \frac{\partial \mathbf{z}^{[L]}}{\partial \mathbf{W}^{[L]}}^T$$

$$\frac{\partial E}{\partial \mathbf{b}^{[L]}} = \frac{\partial E}{\partial \mathbf{x}^{[L]}} \circ \frac{\partial \mathbf{x}^{[L]}}{\partial \mathbf{z}^{[L]}} \cdot \frac{\partial \mathbf{z}^{[L]}}{\partial \mathbf{b}^{[L]}}^T$$

$$\delta^{[L]} = \frac{\partial E}{\partial \mathbf{x}^{[L]}} \circ \frac{\partial \mathbf{x}^{[L]}}{\partial \mathbf{z}^{[L]}}$$

For the second to last layer we have:

$$\frac{\partial E}{\partial \mathbf{W}^{[L-1]}} = \frac{\partial \mathbf{z}^{[L]}}{\partial \mathbf{x}^{[L-1]}}^T \cdot \left(\frac{\partial E}{\partial \mathbf{x}^{[L]}} \circ \frac{\partial \mathbf{x}^{[L]}}{\partial \mathbf{z}^{[L]}} \right) \circ \frac{\partial \mathbf{x}^{[L-1]}}{\partial \mathbf{z}^{[L-1]}} \cdot \frac{\partial \mathbf{z}^{[L-1]}}{\partial \mathbf{W}^{[L-1]}}^T$$

$$\frac{\partial E}{\partial \mathbf{b}^{[L-1]}} = \frac{\partial \mathbf{z}^{[L]}}{\partial \mathbf{x}^{[L-1]}}^T \cdot \left(\frac{\partial E}{\partial \mathbf{x}^{[L]}} \circ \frac{\partial \mathbf{x}^{[L]}}{\partial \mathbf{z}^{[L]}} \right) \circ \frac{\partial \mathbf{x}^{[L-1]}}{\partial \mathbf{z}^{[L-1]}} \cdot \frac{\partial \mathbf{z}^{[L-1]}}{\partial \mathbf{b}^{[L-1]}}^T$$

$$\delta^{[L-1]} = \frac{\partial \mathbf{z}^{[L]}}{\partial \mathbf{x}^{[L-1]}}^T \cdot \delta^{[L]} \circ \frac{\partial \mathbf{x}^{[L-1]}}{\partial \mathbf{z}^{[L-1]}}$$

For layer $L-2$ we have:

$$\frac{\partial E}{\partial \mathbf{W}^{[L-2]}} = \frac{\partial \mathbf{z}^{[L-1]}}{\partial \mathbf{x}^{[L-2]}}^T \cdot \left(\frac{\partial \mathbf{z}^{[L]}}{\partial \mathbf{x}^{[L-1]}}^T \cdot \left(\frac{\partial E}{\partial \mathbf{x}^{[L]}} \circ \frac{\partial \mathbf{x}^{[L]}}{\partial \mathbf{z}^{[L]}} \right) \circ \frac{\partial \mathbf{x}^{[L-1]}}{\partial \mathbf{z}^{[L-1]}} \right) \circ \frac{\partial \mathbf{x}^{[L-2]}}{\partial \mathbf{z}^{[L-2]}} \cdot \frac{\partial \mathbf{z}^{[L-2]}}{\partial \mathbf{W}^{[L-2]}}^T$$

$$\frac{\partial E}{\partial \mathbf{b}^{[L-2]}} = \frac{\partial \mathbf{z}^{[L-1]}}{\partial \mathbf{x}^{[L-2]}}^T \cdot \left(\frac{\partial \mathbf{z}^{[L]}}{\partial \mathbf{x}^{[L-1]}}^T \cdot \left(\frac{\partial E}{\partial \mathbf{x}^{[L]}} \circ \frac{\partial \mathbf{x}^{[L]}}{\partial \mathbf{z}^{[L]}} \right) \circ \frac{\partial \mathbf{x}^{[L-1]}}{\partial \mathbf{z}^{[L-1]}} \right) \circ \frac{\partial \mathbf{x}^{[L-2]}}{\partial \mathbf{z}^{[L-2]}} \cdot \frac{\partial \mathbf{z}^{[L-2]}}{\partial \mathbf{b}^{[L-2]}}^T$$

$$\delta^{[L-2]} = \frac{\partial \mathbf{z}^{[L-1]}}{\partial \mathbf{x}^{[L-2]}}^T \cdot \delta^{[L-1]} \circ \frac{\partial \mathbf{x}^{[L-2]}}{\partial \mathbf{z}^{[L-2]}}$$

In general, we can write down the backward recursion as follows:

$$\delta^{[L]} = \frac{\partial E}{\partial \mathbf{x}^{[L]}} \circ \frac{\partial \mathbf{x}^{[L]}}{\partial \mathbf{z}^{[L]}}$$

$$\delta^{[l]} = \frac{\partial \mathbf{z}^{[l+1]}}{\partial \mathbf{x}^{[l]}}^T \cdot \delta^{[l+1]} \circ \frac{\partial \mathbf{x}^{[l]}}{\partial \mathbf{z}^{[l]}}$$

To complete the backward step all we need to do is to go from layer L to layer 1 and compute all deltas. Having the deltas, we can go to the last step.

Phase 3: Update the parameters

As we have seen in the last section, having the delta for a given layer, all we have to do to compute the gradients for its parameters is:

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

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

So, the last phase is to go from layer 1to layer L computing the gradients and performing the updates:

$$\mathbf{W}^{[l]} = \mathbf{W}^{[l]} - \eta \frac{\partial E}{\partial \mathbf{W}^{[l]}}$$

$$\mathbf{b}^{[l]} = \mathbf{b}^{[l]} - \eta \frac{\partial E}{\partial \mathbf{b}^{[l]}}$$

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:

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

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

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

$$\mathbf{b}^{[2]} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

We are now ready to do 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. Let us compute them beforehand:

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

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

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

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

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

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

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

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

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

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

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

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

All that is left is to update te weights for the output layer.

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{W}^{[2]}} &= \delta^{[2]} \cdot \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{W}^{[2]}}^T \\
&= \delta^{[2]} \cdot \left(\mathbf{x}^{[1]} \right)^T \\
&= \begin{pmatrix} -0.11394 \\ 0.13437 \end{pmatrix} \cdot \begin{pmatrix} \sigma\left(0.2\right) & \sigma\left(0.2\right) & \sigma\left(0.2\right) \end{pmatrix} \\
&= \begin{pmatrix} -0.062647 & -0.062647 & -0.062647 \\ 0.073881 & 0.073881 & 0.073881 \end{pmatrix} \\
\mathbf{W}^{[2]} &= \mathbf{W}^{[2]} - \eta \frac{\partial E}{\partial \mathbf{W}^{[2]}} \\
&= \begin{pmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix} - 1 \begin{pmatrix} -0.062647 & -0.062647 & -0.062647 \\ 0.073881 & 0.073881 & 0.073881 \end{pmatrix} \\
&= \begin{pmatrix} 0.162647 & 0.162647 & 0.162647 \\ 0.026119 & 0.026119 & 0.026119 \end{pmatrix} \\
\frac{\partial E}{\partial \mathbf{b}^{[2]}} &= \delta^{[2]} \cdot \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{b}^{[2]}}^T \\
&= \delta^{[2]} \\
&= \begin{pmatrix} -0.11394 \\ 0.13437 \end{pmatrix}
\end{aligned}$$

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

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

---

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:

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

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

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

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

To get the class label we need to get the MLP output for the point. To get that, we just need to do forward propagation:

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

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

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

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

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

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

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

---

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:

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

Solution:

Before moving forward, we recall and derive the tanh activation function.

$$\tanh(x) = \frac{2}{1 + \exp(-2x)} - 1$$

$$
\begin{aligned}
\frac{\partial}{\partial x}\tanh(x) &= \frac{\partial}{\partial x}\left(\frac{2}{1+\exp(-2x)} - 1\right) \\
&= \frac{\partial}{\partial x}\left(\frac{2}{1+\exp(-2x)}\right) \\
&= 2\frac{\partial}{\partial x}\left(\frac{1}{1+\exp(-2x)}\right) \\
&= 2\left(\frac{\partial\left(\frac{1}{1+\exp(-2x)}\right)}{\partial(1+\exp(-2x))}\frac{\partial(1+\exp(-2x))}{\partial(\exp(-2x))}\frac{\partial(\exp(-2x))}{\partial(-2x)}\frac{\partial(-2x)}{\partial x}\right) \\
&= 2\left(-\frac{1}{(1+\exp(-2x))^2}(1)\exp(-2x)(-2)\right) \\
&= 4\left(\frac{\exp(-2x)}{(1+\exp(-2x))^2}\right) \\
&= 4\left(\frac{1+\exp(-2x)-1}{(1+\exp(-2x))^2}\right) \\
&= 4\left(\frac{1+\exp(-2x)}{(1+\exp(-2x))^2} - \frac{1}{(1+\exp(-2x))^2}\right) \\
&= 4\left(\frac{1}{1+\exp(-2x)} - \frac{1}{(1+\exp(-2x))^2}\right) \\
&= \frac{4}{1+\exp(-2x)}\left(1 - \frac{1}{1+\exp(-2x)}\right) \\
&= 2\frac{2}{1+\exp(-2x)}\left(1 - \frac{1}{1+\exp(-2x)}\right) \\
&= \frac{2}{1+\exp(-2x)}\left(2 - \frac{2}{1+\exp(-2x)}\right) \\
&= (\tanh(x)+1)(2-(\tanh(x)+1)) \\
&= (\tanh(x)+1)(2-\tanh(x)-1) \\
&= (\tanh(x)+1)(1-\tanh(x)) \\
&= \tanh(x) - \tanh(x)^2 + 1 - \tanh(x) \\
&1 - \tanh(x)^2
\end{aligned}
$$

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:

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

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

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

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

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

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

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

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:

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

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

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

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

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

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

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

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:

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

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

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

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

Now the second:

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

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

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

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

All that is left is to update the parameters for the output layer.

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

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

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

---

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.

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:

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

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

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

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

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

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

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

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

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

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

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

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

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

Let us start with the first layer:

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

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

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

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

Now the second:

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

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

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

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

All that is left is to update the parameters for the output layer.

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{W}^{[3]}} &= \delta^{[3](1)} \cdot \frac{\partial \mathbf{z}^{[3](1)}{}^T}{\partial \mathbf{W}^{[3]}} + \delta^{[3](2)} \cdot \frac{\partial \mathbf{z}^{[3](2)}{}^T}{\partial \mathbf{W}^{[3]}} \\
&= \delta^{[3](1)} \cdot \left(\mathbf{x}^{[2](1)}\right)^T + \delta^{[3](2)} \cdot \left(\mathbf{x}^{[2](2)}\right)^T \\
&= \begin{pmatrix} 0.15822 \\ -0.81538 \\ 0.15822 \end{pmatrix} \cdot \begin{pmatrix} 0.2132 & 0.2132 & 0.2132 \end{pmatrix} + \begin{pmatrix} 0.2057 \\ 0.2057 \\ -0.7478 \end{pmatrix} \cdot \begin{pmatrix} 0.3971 & 0.3971 & 0.3971 \end{pmatrix} \\
&= \begin{pmatrix} 0.11539 & 0.11534 & 0.11534 \\ -0.09218 & -0.09218 & -0.09218 \\ -0.26323 & -0.26323 & -0.26323 \end{pmatrix} \\
\mathbf{W}^{[3]} &= \mathbf{W}^{[3]} - \eta \frac{\partial E}{\partial \mathbf{W}^{[3]}} \\
&= \begin{pmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0.11539 & 0.11534 & 0.11534 \\ -0.09218 & -0.09218 & -0.09218 \\ -0.26323 & -0.26323 & -0.26323 \end{pmatrix} \\
&= \begin{pmatrix} 0.08846 & 0.08846 & 0.08846 \\ 0.10922 & 0.10922 & 0.10922 \\ 0.12632 & 0.12632 & 0.12632 \end{pmatrix} \\
\frac{\partial E}{\partial \mathbf{b}^{[3]}} &= \delta^{[3](1)} \cdot \frac{\partial \mathbf{z}^{[3](1)}{}^T}{\partial \mathbf{b}^{[3]}} + \delta^{[3](2)} \cdot \frac{\partial \mathbf{z}^{[3](2)}{}^T}{\partial \mathbf{b}^{[3]}} \\
&= \delta^{[3](1)} + \delta^{[3](2)} \\
&= \begin{pmatrix} 0.3639 \\ -0.6097 \\ -0.5896 \end{pmatrix} \\
\mathbf{b}^{[3]} &= \mathbf{b}^{[3]} - \eta \frac{\partial E}{\partial \mathbf{b}^{[3]}} \\
&= \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0.3639 \\ -0.6097 \\ -0.5896 \end{pmatrix} \\
&= \begin{pmatrix} 0.0636 \\ 0.1609 \\ 0.1589 \end{pmatrix}
\end{aligned}
$$

---

c) Compute the MLP class for the query point

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

using both models from a) and b). Which has smallest squared error? Which model has better classification accuracy?

Solution:
Let us start with model a):

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

$$\mathbf{z}^{[1]} = \begin{pmatrix} 0.10131\ 0.1\ 0.10131\ 0.1 \\ 0.10131\ 0.1\ 0.10131\ 0.1 \\ 0.10131\ 0.1\ 0.10131\ 0.1 \\ 0.10131\ 0.1\ 0.10131\ 0.1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0.10131 \\ 0.10131 \\ 0.10131 \\ 0.10131 \end{pmatrix} = \begin{pmatrix} 0.4039 \\ 0.4039 \\ 0.4039 \\ 0.4039 \end{pmatrix}$$

$$\mathbf{x}^{[1]} = tanh\left(\begin{pmatrix} 0.4039 \\ 0.4039 \\ 0.4039 \\ 0.4039 \end{pmatrix} \right) = \begin{pmatrix} 0.3833 \\ 0.3833 \\ 0.3833 \\ 0.3833 \end{pmatrix}$$

$$\mathbf{z}^{[2]} = \begin{pmatrix} 0.101387\ 0.101387\ 0.101387\ 0.101387 \\ 0.101387\ 0.101387\ 0.101387\ 0.101387 \\ 0.101387\ 0.101387\ 0.101387\ 0.101387 \end{pmatrix} \begin{pmatrix} 0.3833 \\ 0.3833 \\ 0.3833 \\ 0.3833 \end{pmatrix} + \begin{pmatrix} 0.10476 \\ 0.10476 \\ 0.10476 \end{pmatrix} = \begin{pmatrix} 0.2602 \\ 0.2602 \\ 0.2602 \end{pmatrix}$$

$$\mathbf{x}^{[2]} = tanh\left(\begin{pmatrix} 0.2602 \\ 0.2602 \\ 0.2602 \end{pmatrix} \right) = \begin{pmatrix} 0.2545 \\ 0.2545 \\ 0.2545 \end{pmatrix}$$

$$\mathbf{z}^{[3]} = \begin{pmatrix} 0.096627\ 0.096627\ 0.096627 \\ 0.117384\ 0.117384\ 0.117384 \\ 0.096627\ 0.096627\ 0.096627 \end{pmatrix} \begin{pmatrix} 0.2545 \\ 0.2545 \\ 0.2545 \end{pmatrix} + \begin{pmatrix} 0.084178 \\ 0.181538 \\ 0.084178 \end{pmatrix} = \begin{pmatrix} 0.1579 \\ 0.2712 \\ 0.1579 \end{pmatrix}$$

$$\mathbf{x}^{[3]} = tanh\left(\begin{pmatrix} 0.1579 \\ 0.2712 \\ 0.1579 \end{pmatrix} \right) = \begin{pmatrix} 0.1567 \\ 0.2647 \\ 0.1567 \end{pmatrix}$$

$$E\left(\mathbf{t}, \mathbf{x}^{[3]}\right) = \frac{1}{2} \sum_{i=1}^{2} \left(\mathbf{x}^{[3]} - \mathbf{t}\right)^2 = 0.8058$$

Now model b):

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

$$\mathbf{z}^{[1]} = \begin{pmatrix} 0.10131\ 0.1\ 0.10436\ 0.1 \\ 0.10131\ 0.1\ 0.10436\ 0.1 \\ 0.10131\ 0.1\ 0.10436\ 0.1 \\ 0.10131\ 0.1\ 0.10436\ 0.1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0.10161 \\ 0.10161 \\ 0.10161 \\ 0.10161 \end{pmatrix} = \begin{pmatrix} 0.4073 \\ 0.4073 \\ 0.4073 \\ 0.4073 \end{pmatrix}$$

$$\mathbf{x}^{[1]} = tanh\left(\begin{pmatrix} 0.4073 \\ 0.4073 \\ 0.4073 \\ 0.4073 \end{pmatrix}\right) = \begin{pmatrix} 0.3862 \\ 0.3862 \\ 0.3862 \\ 0.3862 \end{pmatrix}$$

$$\mathbf{z}^{[2]} = \begin{pmatrix} 0.103656\ 0.103656\ 0.103656\ 0.103656 \\ 0.103656\ 0.103656\ 0.103656\ 0.103656 \\ 0.103656\ 0.103656\ 0.103656\ 0.103656 \end{pmatrix} \begin{pmatrix} 0.3862 \\ 0.3862 \\ 0.3862 \\ 0.3862 \end{pmatrix} + \begin{pmatrix} 0.107597 \\ 0.107597 \\ 0.107597 \end{pmatrix} = \begin{pmatrix} 0.2677 \\ 0.2677 \\ 0.2677 \end{pmatrix}$$

$$\mathbf{x}^{[2]} = tanh\left(\begin{pmatrix} 0.2677 \\ 0.2677 \\ 0.2677 \end{pmatrix}\right) = \begin{pmatrix} 0.2615 \\ 0.2615 \\ 0.2615 \end{pmatrix}$$

$$\mathbf{z}^{[3]} = \begin{pmatrix} 0.08846\ 0.08846\ 0.08846 \\ 0.10922\ 0.10922\ 0.10922 \\ 0.12632\ 0.12632\ 0.12632 \end{pmatrix} \begin{pmatrix} 0.2615 \\ 0.2615 \\ 0.2615 \end{pmatrix} + \begin{pmatrix} 0.0636 \\ 0.1609 \\ 0.1589 \end{pmatrix} = \begin{pmatrix} 0.1330 \\ 0.2467 \\ 0.2581 \end{pmatrix}$$

$$\mathbf{x}^{[3]} = tanh\left(\begin{pmatrix} 0.1330 \\ 0.2467 \\ 0.2581 \end{pmatrix}\right) = \begin{pmatrix} 0.1322 \\ 0.2418 \\ 0.2525 \end{pmatrix}$$

$$E\left(\mathbf{t}, \mathbf{x}^{[3]}\right) = \frac{1}{2}\sum_{i=1}^{2}\left(\mathbf{x}^{[3]} - \mathbf{t}\right)^2 = 0.6347$$

Model b) has a lower error and is the only one that classifies the point correctly. However, the difference between a correct classification and an incorrect classification seems to be small in terms of squared error. Perhaps we need a better error function...

---

3) Let us repeat the exact same exercise as in 2) but this time we will change:

- The output units have a softmax activation function
- The error function is cross-entropy

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:

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

Solution:

Before moving forward, we recall and derive the $softmax$ activation function. This activation function will work differently from what we have seen so far. This happens because each unit's output depends not only on its input but on the inputs of the other units. Let us make this more concrete. This function receives a vector $\mathbf{z} \in \mathbb{R}^d$ and outputs a vector $\mathbf{x} \in \mathbb{R}^d$.

$$softmax\left(\begin{bmatrix} z_1 & z_2 & \cdots & z_d \end{bmatrix}^T\right) = \begin{bmatrix} x_1 & x_2 & \cdots & x_d \end{bmatrix}^T$$

In general, we have that $x_i = \frac{\exp(z_i)}{\sum_{k=1}^{d} \exp(z_k)}$. Notice how each x_i depends not only on z_i like before. So, let us try to find the derivative of a given x_i with respect to a general z_i.

$$\frac{\partial x_i}{\partial z_j} = \frac{\partial}{\partial z_j} \frac{\exp(z_i)}{\sum_{k=1}^{d} \exp(z_k)}$$

If we have $i = j$ we have to use the rule to derive a quotient:

$$
\begin{aligned}
\frac{\partial x_i}{\partial z_j} &= \frac{\partial}{\partial z_j} \frac{\exp(z_i)}{\sum_{k=1}^{d} \exp(z_k)} \\
&= \frac{\left(\frac{\partial}{\partial z_j} \exp(z_i)\right)\left(\sum_{k=1}^{d} \exp(z_k)\right) - (\exp(z_i))\left(\frac{\partial}{\partial z_j} \sum_{k=1}^{d} \exp(z_k)\right)}{\left(\sum_{k=1}^{d} \exp(z_k)\right)^2} \\
&= \frac{\exp(z_i) \sum_{k=1}^{d} \exp(z_k) - \exp(z_i) \exp(z_j)}{\left(\sum_{k=1}^{d} \exp(z_k)\right)^2} \\
&= \frac{\exp(z_i)\left(\sum_{k=1}^{d} \exp(z_k) - \exp(z_j)\right)}{\left(\sum_{k=1}^{d} \exp(z_k)\right)^2} \\
&= \frac{\exp(z_i)}{\sum_{k=1}^{d} \exp(z_k)} \frac{\sum_{k=1}^{d} \exp(z_k) - \exp(z_j)}{\sum_{k=1}^{d} \exp(z_k)} \\
&= \frac{\exp(z_i)}{\sum_{k=1}^{d} \exp(z_k)} \left(\frac{\sum_{k=1}^{d} \exp(z_k)}{\sum_{k=1}^{d} \exp(z_k)} - \frac{\exp(z_j)}{\sum_{k=1}^{d} \exp(z_k)}\right) \\
&= \frac{\exp(z_i)}{\sum_{k=1}^{d} \exp(z_k)} \left(1 - \frac{\exp(z_j)}{\sum_{k=1}^{d} \exp(z_k)}\right) \\
&= x_i (1 - x_j) \\
&= x_i (1 - x_i)
\end{aligned}
$$

If we have $i \neq j$ we can factor out the numerator:

$$
\begin{aligned}
\frac{\partial x_i}{\partial z_j} &= \frac{\partial}{\partial z_j} \frac{\exp(z_i)}{\sum_{k=1}^{d} \exp(z_k)} \\
&= \exp(z_i) \frac{\partial}{\partial z_j} \frac{1}{\sum_{k=1}^{d} \exp(z_k)} \\
&= \exp(z_i) \frac{\partial\left(\frac{1}{\sum_{k=1}^{d} \exp(z_k)}\right)}{\partial\left(\sum_{k=1}^{d} \exp(z_k)\right)} \frac{\partial\left(\sum_{k=1}^{d} \exp(z_k)\right)}{\partial z_j} \\
&= \exp(z_i) \left(-\frac{1}{\left(\sum_{k=1}^{d} \exp(z_k)\right)^2}\right) \exp(z_j) \\
&= -\frac{\exp(z_i)}{\sum_{k=1}^{d} \exp(z_k)} \frac{\exp(z_j)}{\sum_{k=1}^{d} \exp(z_k)} \\
&= -x_i x_j
\end{aligned}
$$

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:

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

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

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

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

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

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

$$\mathbf{x}^{[3]} = softmax\left(\begin{pmatrix} 0.16396 \\ 0.16396 \\ 0.16396 \end{pmatrix}\right) = \begin{pmatrix} 0.3333 \\ 0.3333 \\ 0.3333 \end{pmatrix}$$

Recall the cross-entropy loss:

$$E\left(\mathbf{t}, \mathbf{x}^{[3]}\right) = -\sum_{i=1}^{d} t_i \log x_i^{[3]}$$

What will differ is the delta of the first layer. For the remaining layers we can use the same approach as before. So, for a given $z_i^{[3]}$ the derivative of the error function will be as follows.

$$
\begin{aligned}
\delta_i^{[3]} &= \frac{\partial E\left(\mathbf{t}, \mathbf{x}^{[3]}\right)}{\partial z_i} \\
&= \frac{\partial}{\partial z_i}\left(-\sum_{k=1}^{d} t_k \log x_k^{[3]}\right) \\
&= -\sum_{k=1}^{d} t_k \frac{\partial}{\partial z_i} \log x_k^{[3]} \\
&= -\sum_{k=1}^{d} t_k \frac{1}{x_k^{[3]}} \frac{\partial x_k^{[3]}}{\partial z_i} \\
&= -\sum_{k=i} t_k \frac{1}{x_k^{[3]}} \frac{\partial x_k^{[3]}}{\partial z_i} - \sum_{k\neq i} t_k \frac{1}{x_k^{[3]}} \frac{\partial x_k^{[3]}}{\partial z_i} \\
&= -\sum_{k=i} t_k \frac{1}{x_k^{[3]}} \left(x_i^{[3]}\left(1 - x_i^{[3]}\right)\right) - \sum_{k\neq i} t_k \frac{1}{x_k^{[3]}} \left(-x_k^{[3]} x_i^{[3]}\right) \\
&= -t_i \frac{1}{x_i^{[3]}} \left(x_i^{[3]}\left(1 - x_i^{[3]}\right)\right) - \sum_{k\neq i} t_k \frac{1}{x_k^{[3]}} \left(-x_k^{[3]} x_i^{[3]}\right) \\
&= -t_i \left(1 - x_i^{[3]}\right) + \sum_{k\neq i} t_k x_i^{[3]} \\
&= -t_i + t_i x_i^{[3]} + \sum_{k\neq i} t_k x_i^{[3]} \\
&= -t_i + x_i^{[3]} \left(t_i + \sum_{k\neq i} t_k\right) \\
&= -t_i + x_i^{[3]} \left(\sum_{k=1}^{d} t_k\right) \\
&= -t_i + x_i^{[3]} \\
&= x_i^{[3]} - t_i
\end{aligned}
$$

The remaining derivatives can be computed as before:

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

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

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

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

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

$$\begin{aligned}
\delta^{[3]} &= \begin{pmatrix} \delta_1^{[3]} \\ \delta_2^{[3]} \\ \delta_3^{[3]} \end{pmatrix} = \begin{pmatrix} \frac{\partial E\left(\mathbf{t},\mathbf{x}^{[3]}\right)}{\partial z_1} \\ \frac{\partial E\left(\mathbf{t},\mathbf{x}^{[3]}\right)}{\partial z_2} \\ \frac{\partial E\left(\mathbf{t},\mathbf{x}^{[3]}\right)}{\partial z_3} \end{pmatrix} \\
&= \begin{pmatrix} x_1^{[3]} - t_1 \\ x_2^{[3]} - t_2 \\ x_3^{[3]} - t_3 \end{pmatrix} \\
&= \left(\mathbf{x}^{[3]} - \mathbf{t}\right) \\
&= \begin{pmatrix} 0.3333 \\ 0.3333 \\ 0.3333 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} 0.3333 \\ -0.6666 \\ 0.3333 \end{pmatrix}
\end{aligned}$$

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

$$\begin{aligned}
\delta^{[2]} &= \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{x}^{[2]}}^T \cdot \delta^{[3]} \circ \frac{\partial \mathbf{x}^{[2]}}{\partial \mathbf{z}^{[2]}} \\
&= \left(\mathbf{W}^{[3]}\right)^T \cdot \delta^{[3]} \circ \left(1 - tanh\left(\mathbf{z}^{[2]}\right)^2\right) \\
&= \begin{pmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix} \cdot \begin{pmatrix} 0.3333 \\ -0.6667 \\ 0.3333 \end{pmatrix} \circ \left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} \tanh\left(0.2165\right)^2 \\ \tanh\left(0.2165\right)^2 \\ \tanh\left(0.2165\right)^2 \end{pmatrix} \right) \\
&= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
\end{aligned}$$

$$\begin{aligned}
\delta^{[1]} &= \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{x}^{[1]}}^T \cdot \delta^{[2]} \circ \frac{\partial \mathbf{x}^{[1]}}{\partial \mathbf{z}^{[1]}} \\
&= \left(\mathbf{W}^{[2]}\right)^T \cdot \delta^{[2]} \circ \left(1 - tanh\left(\mathbf{z}^{[1]}\right)^2\right) \\
&= \begin{pmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \circ \left(\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} \tanh\left(0.3\right)^2 \\ \tanh\left(0.3\right)^2 \\ \tanh\left(0.3\right)^2 \\ \tanh\left(0.3\right)^2 \end{pmatrix} \right) \\
&= \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}
\end{aligned}$$

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

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

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

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{b}^{[1]}} &= \delta^{[1]} \cdot \frac{\partial \mathbf{z}^{[1]}}{\partial \mathbf{b}^{[1]}}^T \\ &= \delta^{[1]} \\ &= \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}\end{aligned}$$

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

Now the second:

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{W}^{[2]}} &= \delta^{[2]} \cdot \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{W}^{[2]}}^T \\ &= \delta^{[2]} \cdot \left(\mathbf{x}^{[1]}\right)^T \\ &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 0.2913 & 0.2913 & 0.2913 & 0.2913 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{W}^{[2]} &= \mathbf{W}^{[2]} - \eta \frac{\partial E}{\partial \mathbf{W}^{[2]}} \\ &= \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{b}^{[2]}} &= \delta^{[2]} \cdot \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{b}^{[2]}}^T \\ &= \delta^{[2]} \\ &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{b}^{[2]} &= \mathbf{b}^{[2]} - \eta \frac{\partial E}{\partial \mathbf{b}^{[2]}} \\ &= \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix}\end{aligned}$$

All that is left is to update the parameters for the output layer.

$$
\begin{aligned}
\frac{\partial E}{\partial \mathbf{W}^{[3]}} &= \delta^{[3]} \cdot \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{W}^{[3]}}^T \\
&= \delta^{[3]} \cdot \left(\mathbf{x}^{[2]}\right)^T \\
&= \begin{pmatrix} 0.3333 \\ -0.6666 \\ 0.3333 \end{pmatrix} \cdot \begin{pmatrix} 0.2132 & 0.2132 & 0.2132 \end{pmatrix} \\
&= \begin{pmatrix} 0.0711 & 0.0711 & 0.0711 \\ -0.1421 & -0.1421 & -0.1421 \\ 0.0711 & 0.0711 & 0.0711 \end{pmatrix} \\
\mathbf{W}^{[3]} &= \mathbf{W}^{[3]} - \eta \frac{\partial E}{\partial \mathbf{W}^{[3]}} \\
&= \begin{pmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0.0711 & 0.0711 & 0.0711 \\ -0.1421 & -0.1421 & -0.1421 \\ 0.0711 & 0.0711 & 0.0711 \end{pmatrix} \\
&= \begin{pmatrix} 0.0929 & 0.0929 & 0.0929 \\ 0.1142 & 0.1142 & 0.1142 \\ 0.0929 & 0.0929 & 0.0929 \end{pmatrix} \\
\frac{\partial E}{\partial \mathbf{b}^{[3]}} &= \delta^{[3]} \cdot \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{b}^{[3]}}^T \\
&= \delta^{[3]} \\
&= \begin{pmatrix} 0.3333 \\ -0.6666 \\ 0.3333 \end{pmatrix} \\
\mathbf{b}^{[3]} &= \mathbf{b}^{[3]} - \eta \frac{\partial E}{\partial \mathbf{b}^{[3]}} \\
&= \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0.3333 \\ -0.6666 \\ 0.3333 \end{pmatrix} \\
&= \begin{pmatrix} 0.0667 \\ 0.1667 \\ 0.0667 \end{pmatrix}
\end{aligned}
$$

---

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:

Since 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.

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:

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

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

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

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

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

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

$$\mathbf{x}^{[3]} = tanh\left(\begin{pmatrix} 0.2191 \\ 0.2191 \\ 0.2191 \end{pmatrix} \right) = \begin{pmatrix} 0.3333 \\ 0.3333 \\ 0.3333 \end{pmatrix}$$

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

$$\begin{aligned} \delta^{[3]} &= \begin{pmatrix} \delta_1^{[3]} \\ \delta_2^{[3]} \\ \delta_3^{[3]} \end{pmatrix} = \begin{pmatrix} \frac{\partial E\left(\mathbf{t}, \mathbf{x}^{[3]}\right)}{\partial z_1} \\ \frac{\partial E\left(\mathbf{t}, \mathbf{x}^{[3]}\right)}{\partial z_2} \\ \frac{\partial E\left(\mathbf{t}, \mathbf{x}^{[3]}\right)}{\partial z_3} \end{pmatrix} \\ &= \begin{pmatrix} x_1^{[3]} - t_1 \\ x_2^{[3]} - t_2 \\ x_3^{[3]} - t_3 \end{pmatrix} \\ &= \left(\mathbf{x}^{[3]} - \mathbf{t} \right) \\ &= \begin{pmatrix} 0.3333 \\ 0.3333 \\ 0.3333 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 0.3333 \\ 0.3333 \\ -0.6666 \end{pmatrix} \end{aligned}$$

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

$$\begin{aligned} \delta^{[2]} &= \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{x}^{[2]}}^T \cdot \delta^{[3]} \circ \frac{\partial \mathbf{x}^{[2]}}{\partial \mathbf{z}^{[2]}} \\ &= \left(\mathbf{W}^{[3]} \right)^T \cdot \delta^{[3]} \circ \left(1 - tanh\left(\mathbf{z}^{[2]} \right)^2 \right) \\ &= \begin{pmatrix} 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \end{pmatrix} \cdot \begin{pmatrix} 0.3333 \\ 0.3333 \\ -0.6666 \end{pmatrix} \circ \left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} \tanh\left(0.4202\right)^2 \\ \tanh\left(0.4202\right)^2 \\ \tanh\left(0.4202\right)^2 \end{pmatrix} \right) \\ &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
\delta^{[1]} &= \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{x}^{[1]}}^T \cdot \delta^{[2]} \circ \frac{\partial \mathbf{x}^{[1]}}{\partial \mathbf{z}^{[1]}} \\
&= \left(\mathbf{W}^{[2]}\right)^T \cdot \delta^{[2]} \circ \left(1 - tanh\left(\mathbf{z}^{[1]}\right)^2\right) \\
&= \begin{pmatrix} 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \circ \left(\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} \tanh(1.1)^2 \\ \tanh(1.1)^2 \\ \tanh(1.1)^2 \\ \tanh(1.1)^2 \end{pmatrix} \right) \\
&= \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}
\end{aligned}
$$

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

Let us start with the first layer:

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

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

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{b}^{[1]}} &= \delta^{[1](1)} \cdot \frac{\partial \mathbf{z}^{[1](1)}{}^{T}}{\partial \mathbf{b}^{[1]}} + \delta^{[1](2)} \cdot \frac{\partial \mathbf{z}^{[1](2)}{}^{T}}{\partial \mathbf{b}^{[1]}} \\ &= \delta^{[1](1)} + \delta^{[1](2)} \\ &= \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}\end{aligned}$$

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

Now the second:

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{W}^{[2]}} &= \delta^{[2](1)} \cdot \frac{\partial \mathbf{z}^{[2](1)}{}^{T}}{\partial \mathbf{W}^{[2]}} + \delta^{[2](2)} \cdot \frac{\partial \mathbf{z}^{[2](2)}{}^{T}}{\partial \mathbf{W}^{[2]}} \\ &= \delta^{[2](1)} \cdot \left(\mathbf{x}^{[1](1)}\right)^{T} + \delta^{[2](2)} \cdot \left(\mathbf{x}^{[1](2)}\right)^{T} \\ &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 0.2913 & 0.2913 & 0.2913 & 0.2913 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 0.8005 & 0.8005 & 0.8005 & 0.8005 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{W}^{[2]} &= \mathbf{W}^{[2]} - \eta \frac{\partial E}{\partial \mathbf{W}^{[2]}} \\ &= \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{b}^{[2]}} &= \delta^{[2](1)} \cdot \frac{\partial \mathbf{z}^{[2](1)}{}^T}{\partial \mathbf{b}^{[2]}} + \delta^{[2](2)} \cdot \frac{\partial \mathbf{z}^{[2](2)}{}^T}{\partial \mathbf{b}^{[2]}} \\ &= \delta^{[2](1)} + \delta^{[2](2)} \\ &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{b}^{[2]} &= \mathbf{b}^{[2]} - \eta \frac{\partial E}{\partial \mathbf{b}^{[2]}} \\ &= \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix}\end{aligned}$$

All that is left is to update the parameters for the output layer.

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{W}^{[3]}} &= \delta^{[3](1)} \cdot \frac{\partial \mathbf{z}^{[3](1)}{}^T}{\partial \mathbf{W}^{[3]}} + \delta^{[3](2)} \cdot \frac{\partial \mathbf{z}^{[3](2)}{}^T}{\partial \mathbf{W}^{[3]}} \\ &= \delta^{[3](1)} \cdot \left(\mathbf{x}^{[2](1)}\right)^T + \delta^{[3](2)} \cdot \left(\mathbf{x}^{[2](2)}\right)^T \\ &= \begin{pmatrix} 0.3333 \\ -0.6666 \\ 0.3333 \end{pmatrix} \cdot \begin{pmatrix} 0.2132 & 0.2132 & 0.2132 \end{pmatrix} + \begin{pmatrix} 0.3333 \\ 0.3333 \\ -0.6666 \end{pmatrix} \cdot \begin{pmatrix} 0.3971 & 0.3971 & 0.3971 \end{pmatrix} \\ &= \begin{pmatrix} 0.2034 & 0.2034 & 0.2034 \\ -0.0098 & -0.0098 & -0.0098 \\ -0.1937 & -0.1937 & -0.1937 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{W}^{[3]} &= \mathbf{W}^{[3]} - \eta \frac{\partial E}{\partial \mathbf{W}^{[3]}} \\ &= \begin{pmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0.2034 & 0.2034 & 0.2034 \\ -0.0098 & -0.0098 & -0.0098 \\ -0.1937 & -0.1937 & -0.1937 \end{pmatrix} \\ &= \begin{pmatrix} 0.0797 & 0.0797 & 0.0797 \\ 0.1009 & 0.1009 & 0.1009 \\ 0.1194 & 0.1194 & 0.1194 \end{pmatrix}\end{aligned}$$

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{b}^{[3]}} &= \delta^{[3](1)} \cdot \frac{\partial \mathbf{z}^{[3](1)}}{\partial \mathbf{b}^{[3]}}^T + \delta^{[3](2)} \cdot \frac{\partial \mathbf{z}^{[3](2)}}{\partial \mathbf{b}^{[3]}}^T \\
&= \delta^{[3](1)} + \delta^{[3](2)} \\
&= \begin{pmatrix} 0.3333 \\ -0.6666 \\ 0.3333 \end{pmatrix} + \begin{pmatrix} 0.3333 \\ 0.3333 \\ -0.6666 \end{pmatrix} \\
&= \begin{pmatrix} 0.6666 \\ -0.3333 \\ -0.3333 \end{pmatrix} \\
\mathbf{b}^{[3]} &= \mathbf{b}^{[3]} - \eta \frac{\partial E}{\partial \mathbf{b}^{[3]}} \\
&= \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} - 0.1 \begin{pmatrix} 0.6666 \\ -0.3333 \\ -0.3333 \end{pmatrix} \\
&= \begin{pmatrix} 0.3333 \\ 1.3333 \\ 1.3333 \end{pmatrix}
\end{aligned}$$

---

c) Compute the MLP class for the query point

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

using both models from a) and b). Which has smallest cross-entropy loss? Which model has better classification accuracy?

Solution:
Let us start with model a):

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

$$\mathbf{z}^{[1]} = \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} = \begin{pmatrix} 0.4 \\ 0.4 \\ 0.4 \\ 0.4 \end{pmatrix}$$

$$\mathbf{x}^{[1]} = tanh \left(\begin{pmatrix} 0.4 \\ 0.4 \\ 0.4 \\ 0.4 \end{pmatrix} \right) = \begin{pmatrix} 0.3799 \\ 0.3799 \\ 0.3799 \\ 0.3799 \end{pmatrix}$$

$$\mathbf{z}^{[2]} = \begin{pmatrix} 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \end{pmatrix} \begin{pmatrix} 0.3799 \\ 0.3799 \\ 0.3799 \\ 0.3799 \end{pmatrix} + \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} = \begin{pmatrix} 0.2519 \\ 0.2519 \\ 0.2519 \end{pmatrix}$$

$$\mathbf{x}^{[2]} = tanh\left(\begin{pmatrix} 0.2519 \\ 0.2519 \\ 0.2519 \end{pmatrix} \right) = \begin{pmatrix} 0.2468 \\ 0.2468 \\ 0.2468 \end{pmatrix}$$

$$\mathbf{z}^{[3]} = \begin{pmatrix} 0.0929\ 0.0929\ 0.0929 \\ 0.1142\ 0.1142\ 0.1142 \\ 0.0929\ 0.0929\ 0.0929 \end{pmatrix} \begin{pmatrix} 0.2468 \\ 0.2468 \\ 0.2468 \end{pmatrix} + \begin{pmatrix} 0.0667 \\ 0.1667 \\ 0.0667 \end{pmatrix} = \begin{pmatrix} 0.1354 \\ 0.2512 \\ 0.1354 \end{pmatrix}$$

$$\mathbf{x}^{[3]} = tanh\left(\begin{pmatrix} 0.1354 \\ 0.2512 \\ 0.1354 \end{pmatrix} \right) = \begin{pmatrix} 0.3202 \\ 0.3595 \\ 0.3202 \end{pmatrix}$$

$$E\left(\mathbf{t}, \mathbf{x}^{[3]}\right) = \frac{1}{2} \sum_{i=1}^{2} \left(\mathbf{x}^{[3]} - \mathbf{t}\right)^2 = 1.1387$$

Now model b):

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

$$\mathbf{z}^{[1]} = \begin{pmatrix} 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} = \begin{pmatrix} 0.4 \\ 0.4 \\ 0.4 \\ 0.4 \end{pmatrix}$$

$$\mathbf{x}^{[1]} = tanh\left(\begin{pmatrix} 0.4 \\ 0.4 \\ 0.4 \\ 0.4 \end{pmatrix} \right) = \begin{pmatrix} 0.3799 \\ 0.3799 \\ 0.3799 \\ 0.3799 \end{pmatrix}$$

$$\mathbf{z}^{[2]} = \begin{pmatrix} 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \\ 0.1\ 0.1\ 0.1\ 0.1 \end{pmatrix} \begin{pmatrix} 0.3799 \\ 0.3799 \\ 0.3799 \\ 0.3799 \end{pmatrix} + \begin{pmatrix} 0.1 \\ 0.1 \\ 0.1 \end{pmatrix} = \begin{pmatrix} 0.2519 \\ 0.2519 \\ 0.2519 \end{pmatrix}$$

$$\mathbf{x}^{[2]} = tanh\left(\begin{pmatrix} 0.2519 \\ 0.2519 \\ 0.2519 \end{pmatrix} \right) = \begin{pmatrix} 0.2468 \\ 0.2468 \\ 0.2468 \end{pmatrix}$$

$$\mathbf{z}^{[3]} = \begin{pmatrix} 0.0797\ 0.0797\ 0.0797 \\ 0.1009\ 0.1009\ 0.1009 \\ 0.1194\ 0.1194\ 0.1194 \end{pmatrix} \begin{pmatrix} 0.2468 \\ 0.2468 \\ 0.2468 \end{pmatrix} + \begin{pmatrix} 0.3333 \\ 1.3333 \\ 1.3333 \end{pmatrix} = \begin{pmatrix} 0.0923 \\ 0.2081 \\ 0.2217 \end{pmatrix}$$

$$\mathbf{x}^{[3]} = softmax \left(\begin{pmatrix} 0.0923 \\ 0.2081 \\ 0.2217 \end{pmatrix} \right) = \begin{pmatrix} 0.3067 \\ 0.3443 \\ 0.3490 \end{pmatrix}$$

$$E\left(\mathbf{t}, \mathbf{x}^{[3]}\right) = 1.0526$$

Model b) has a lower error and is the only one that classifies the point correctly.

---

3 Thinking Questions

a) What are the main differences between using squared error and cross-entropy?

b) Try to repeat the problems in the lecture with invented (differentiable) activation functions.

c) How do you think the MLP decision boundary will look like in two-dimensional problems?

d) Think about how the number of parameters to estimate grows with the number of layers. Do we need more or less data to estimate them properly?

e) Can we use MLP for a regression problem?

f) Try to apply an MLP to exercises from previous lectures. Notice that it can solve non-linearly separable problems.