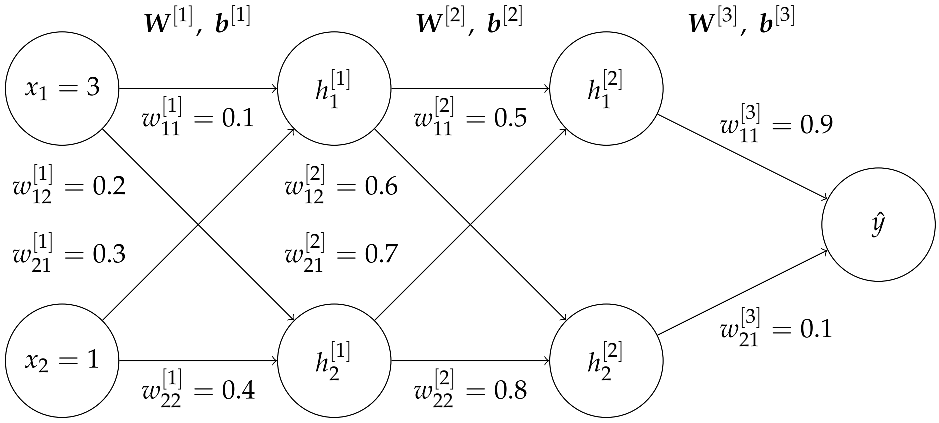
Task 1: Multi-Layer Perceptron (MLP) **(17 Points)**

For a regression task, the fully connected neural network (MLP) outlined below is used as a model approach. The activation function used in all of the hidden layers is the ReLU-activation function, while the identity function $f(x) = x$ is used at the output layer/node. The following loss function is given

$$L(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2$$

where $\hat{y}^{(i)}$ denotes the model output for the given input $\boldsymbol{x}^{(i)}$ and sample i of the batch with size m.



a) How does the input vector $\mathbf{x}$, weight matrix $\boldsymbol{W}^{[1]}, \boldsymbol{W}^{[2]}, \boldsymbol{W}^{[3]}$ and bias vector $\boldsymbol{b}^{[1]}, \boldsymbol{b}^{[2]}, \boldsymbol{b}^{[3]}$ look like and what are the individual shapes? All bias values of the bias vector $\boldsymbol{b}^{[1]}$ in the first layer are initialized with $\boldsymbol{b}^{[1]} = 0.1$, for the bias vector of the second layer $\boldsymbol{b}^{[2]} = 0.2$, and for the bias vector of the third layer $\boldsymbol{b}^{[3]} = 0.3$. How many parameters does the network have in total? *(4 Points)*

b) Perform the forward propagation of a single sample $\boldsymbol{x} \in \mathbb{R}^2$ (batch size $m = 1$), given in the above graph, and compute $\hat{y}$, considering the given weight matrices $\boldsymbol{W}^{[i]}$ and bias vectors $\boldsymbol{b}^{[i]}$.

Proceed step-by-step and use a vectorized form! *(5 Points)*

c) What is the name of the optimization method using a batch size of $m = 1$ and what are the advantages and disadvantages? *(2 Points)*

d) Compute the backward propagation (in vectorized form!) for the given MLP, in order to derive all the gradients required for a complete parameter update. Consider the above mentioned loss function $L(\theta)$ and a ground truth value of $y = 1$. *(5 Points)*

e) Perform the weight update for $w_{11}^{[3]}$ using a learning rate $\alpha = 0.01$. *(1 Point)*