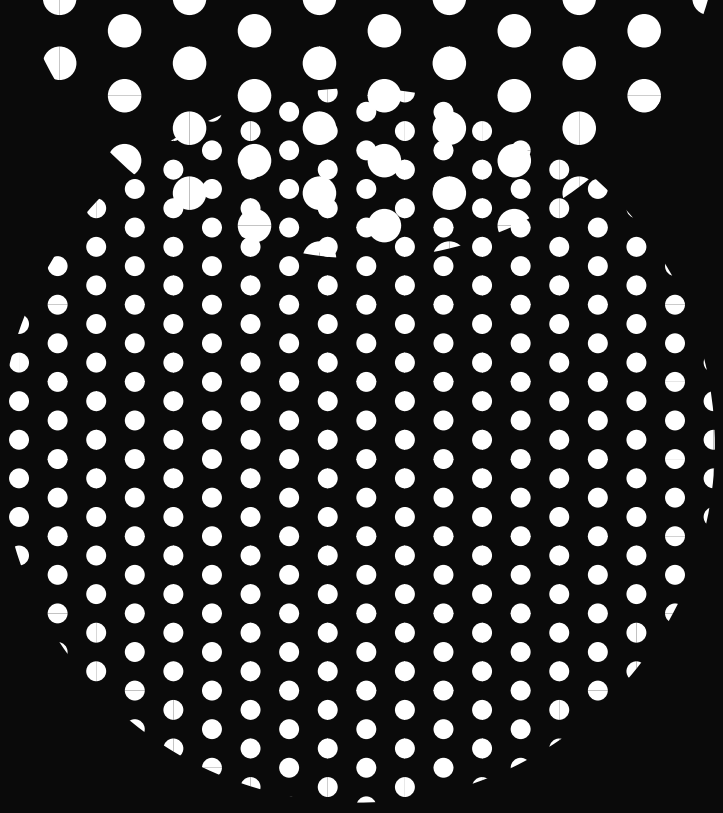


Problem 2: You are trying to solve a regression task and you want to choose between two approaches:

1. A simple linear regression model.
2. A feed forward neural network $f_{\boldsymbol{W}}(\boldsymbol{x})$ with L hidden layers, where each hidden layer $l \in \{1, ..., L\}$ has a weight matrix $\boldsymbol{W}_l \in \mathbb{R}^{D \times D}$ and a ReLU activation function. The output layer has a weight matrix $\boldsymbol{W}_{L+1} \in \mathbb{R}^{D \times 1}$ and no activation function.

In both models, there are no bias terms.

Your dataset $\mathcal{D}$ contains data points with nonnegative features $\boldsymbol{x}_n$ and the target y_n is continuous:

$$\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N, \qquad \boldsymbol{x}_n \in \mathbb{R}^D_{\geq 0}, \qquad y_n \in \mathbb{R}$$

Let $\boldsymbol{w}^*_{LS} \in \mathbb{R}^D$ be the optimal weights for the linear regression model corresponding to a *global* minimum of the following least squares optimization problem:

$$\boldsymbol{w}^*_{LS} = \arg\min_{\boldsymbol{w} \in \mathbb{R}^D} \mathcal{L}_{LS}(\boldsymbol{w}) = \arg\min_{\boldsymbol{w} \in \mathbb{R}^D} \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{w}^T \boldsymbol{x}_n - y_n)^2$$

Let $\boldsymbol{W}^*_{NN} = \{\boldsymbol{W}^*_1, \ldots, \boldsymbol{W}^*_{L+1}\}$ be the optimal weights for the neural network corresponding to a *global* minimum of the following optimization problem:

$$\boldsymbol{W}^*_{NN} = \arg\min_{\boldsymbol{W}} \mathcal{L}_{NN}(\boldsymbol{W}) = \arg\min_{\boldsymbol{W}} \frac{1}{2} \sum_{n=1}^{N} (f_{\boldsymbol{W}}(\boldsymbol{x}_n) - y_n)^2$$

a) Assume that the optimal $\boldsymbol{W}^*_{NN}$ you obtain are non-negative.
What will the relation $(<, \leq, =, \geq, >)$ between the neural network loss $\mathcal{L}_{NN}(\boldsymbol{W}^*_{NN})$ and the linear regression loss $\mathcal{L}_{LS}(\boldsymbol{w}^*_{LS})$ be? Provide a mathematical argument to justify your answer.

Note that for any non-negative x and non-negative W it holds $\text{ReLU}(xW) = xW$

Therefore, since our data points have non-negative features x_i and the optimal weights W^*_{NN} are non-negative, every ReLU layer is equivalent to a linear layer when plugging in the optimal weights. This means we can write:

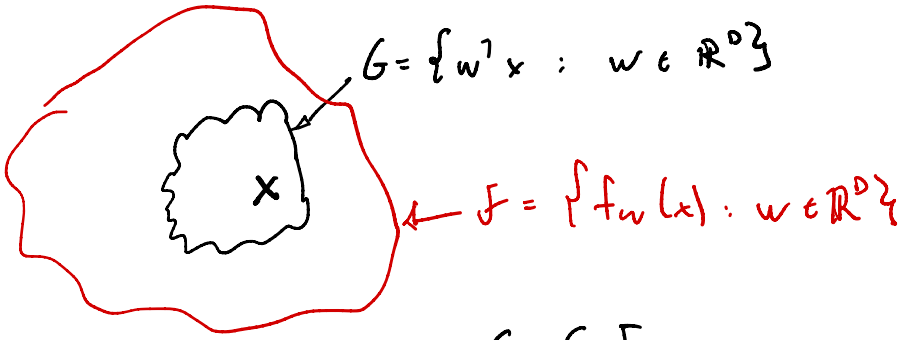
$$\begin{aligned} f_{W^*_{NN}}(x_i) &= \text{ReLU}(\text{ReLU}(\text{ReLU}(x_i^T W_1^*) W_2^*) \cdots W_L^*) W_{L+1}^* \\ &= x_i^T W_1^* W_2^* \cdots W_{L+1}^* \\ &= x_i^T W_{NN}^* \end{aligned}$$

where we defined $W^*_{NN} = W_1^* W_2^* \cdots W_{L+1}^*$. From this we can see that the neural network with optimal weights behaves like a linear regression with a different set of weights W^*_{NN}. Note also that linear regression is a special case of the above neural network, i.e for any weights W_{LS} you can find weights W_{NN} that produces the same output.

Given the above facts and since the optimal weights correspond to a global minima, we can conclude that $\mathcal{L}_{NN}(W^*_{NN}) = \mathcal{L}_{LS}(w^*_{LS})$ and the optimal weights found by solving the least squares optimization problem will be

$w^*_{LS} = w^*_{NN}$.

*
Not every NN can be represented with LR model (putting negative weights)



$G \subset F$

$$\min_{f \in F} \frac{1}{2} \sum_{n=1}^{N} (f(x_n) - y_n)^2 \geq \min_{f \in F} \frac{1}{2} \sum_{n=2}^{N} (f(x_n) - y_n)^2$$

b) In contrast to (a), now assume that the optimal weights $\boldsymbol{w}^*_{LS}$ you obtain are non-negative. What will the relation $(<, \leq, =, \geq, >)$ between the linear regression loss $\mathcal{L}_{LS}(\boldsymbol{w}^*_{LS})$ and the neural network loss $\mathcal{L}_{NN}(\boldsymbol{W}^*_{NN})$ be? Provide a mathematical argument to justify your answer.

As stated in (a) linear regression is a special case of the above neural network, i.e. for any weights w_{LS} you can find weights W_{NN} that produces the same output. That is, everything can be learned with a linear regression can be learned equally well with neural network. However, the reverse direction doesn't hold, since in principle neural networks can learn more complicated functions compared to linear regression. Moreover, the given fact that w^*_{LS} are non-negative does not tell us anything about the optimal weights of the neural network W^*_{NN}. Therefore it holds $\mathcal{L}_{NN}(W^*_{NN}) \leq \mathcal{L}_{LS}(w^*_{LS})$ since the neural network can potentially find a better fit for the data (e.g. by taking advantage of non-linearity).