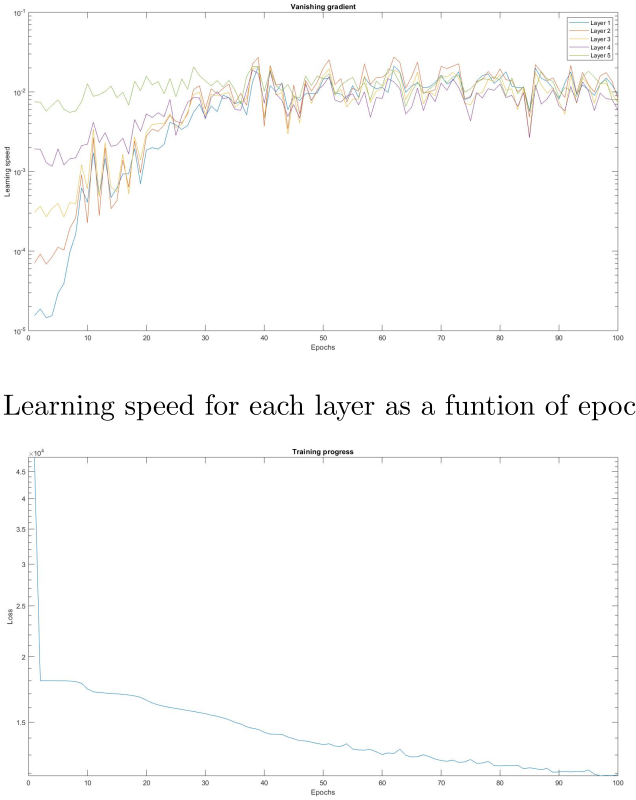

(a) Learning speed for each layer as a funtion of epochs

(b) The Loss functions as a function of epochs

Figure 1: a) The learning speed of each layer in the network which consisting of 4 hidden layers (20 neurons each) and 1 output layer .b) The loss function of the network as a function of epochs

Theoretical: Deeper networks have a more severe *vanishing gradient* issue. The gradient for the weights $\frac{\partial H}{\partial w_i^l}$ closer to the inputs are much smaller than the gradients closer to the output (when using sigmoid activation function). In consequence this leads to much *slower learning rates* for earlier layers.

Experimental results: The experiment consisted of 1 networks with 4 different hidden layers and 1 output layer. Figure 1a shows that the learning speed for thresholds in earlier layers are much lower. The learning phase really kicks in around epoch 5-10 and then starts to converge at a stable value for all layers.
Figure 1b shows that the loss drastically reduces in the beginning and than starts to slow down. This agrees with the theory that networks update the weights/thresholds according to the computed gradient, which should be smaller each epoch.