3.3 ReLU, softmax, early stopping

Table (1) and (2) displays the classification errors and number of training epochs for each of the three networks.

Table 1: Classification errors for each network

Data set	Network 1	Network 2	Network 3
Training set	0.4335	0.4405	0.3981
Validation set	0.5173	0.5244	0.4976
Test set	0.5132	0.5217	0.4962

Table 2: Number of training epochs and iterations for each network

Networks	Number of epochs
Network 1	195
Network 2	165
Network 3	353

It is clear that Network 3 is the best in terms of accuracy from table (1), since it has the lowest classification error. However, it is the network with highest number of epochs, meaning that it took the longest time to run. The network that performs worst is network 2 since it has the highest classification error, even tough it has the lowest number of epochs.

The reason for why network 3 outperforms network 1 and 2 is due to its architecture. Since the networks are deep networks, they will encounter problems with overfitting due to having more neurons. Overfitting can be reduced by using regularisation schemes, such as the L_2-regularisation that is introduced to network 3 with its L_2-regularisation parameter. By reducing overfitting network 3 will be able to learn faster, leading to a lower classification error.

The reason for why network 1 outperforms network 2 is due to having fewer neurons in the hidden layers, since network 1 have one less hidden layer with 50 neurons than network 2. This results in network 2 having more problems with overfitting and a lower accuracy, i.e higher classification error. Another explanation for why network 2 performs less accurate is due to unstable gradients. The ReLU function makes the vanishing gradient problem less severe since the gradient of the activation function does not decrease exponentially, as in the case for the sigmoid function. However the ReLU function does not saturate, thereby resulting in an increase of the weights which slows down the learning of the network. In order for the weights not to grow it is necessary to implement a regularisation scheme such as in network 3.

It is also noted that from table (2) that early stopping occurred for all networks, since the maximum number of epochs were set to 400