Practical part:

Question 1:

Network architecture:

tanh

Softmax

ReLU

100

784

100

10

Hyperparameters:

In the ADAM implementation the hyperparameters were

In the training process, I set the batch\_size parameter to 100

Optimization:

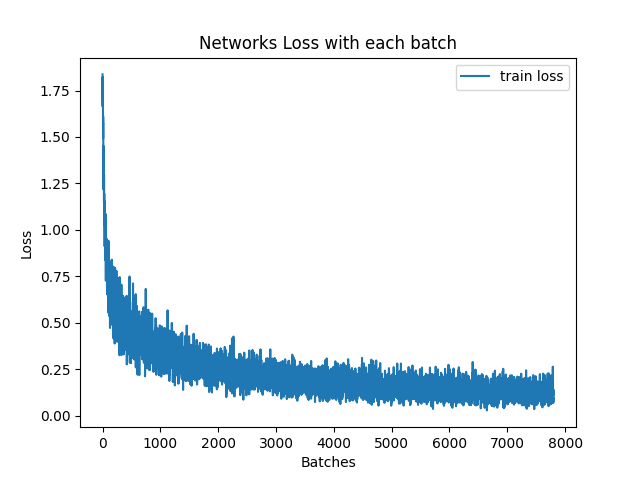
Implemented the ADAM optimizer in the backward function

Calculated the gradients w.r.t. to each weight matrix, bias and updated the weights and biases accordingly

Short summary & conclusions:

Although I calculated the gradient of softmax in Question 1 (Theoretical part) the training process seemed to be negatively affected from the gradient of the softmax and the network loss seemed to improve better when the softmax gradient weren’t multiplied

Chart, histogram

Description automatically generated

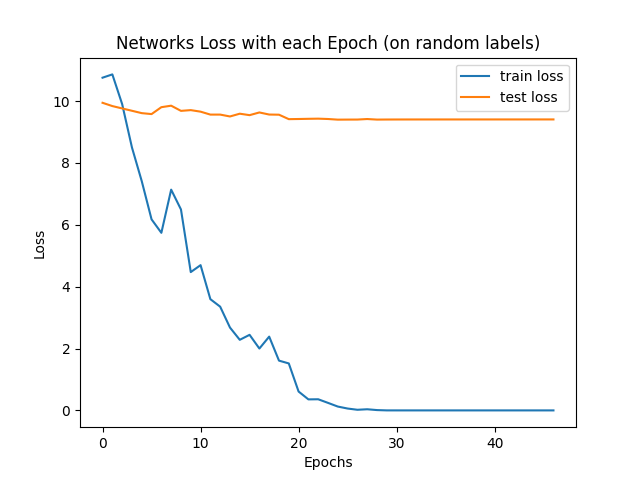
Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Question 2:



Chart, line chart, histogram

Description automatically generatedalthough the loss on the training set is improving and almost 0, the loss on the test set is almost the same as the initial loss.

And that’s due to the face that the network attempted to learn the labels that’s being generated at random.

And of course we can’t learn the “toss of a coin”.