

1a)

$$\begin{aligned}
 \omega_{ji} &= \omega_{ji} - \eta \frac{\partial C}{\partial \omega_{ji}} \\
 &= \omega_{ji} - \eta \frac{\partial C}{\partial z_j} \frac{\partial z_j}{\partial \omega_{ji}} \\
 &= \omega_{ji} - \eta \frac{\partial C}{\partial z_k} \cdot \frac{\partial z_k}{\partial a_j} \cdot \frac{\partial a_j}{\partial z_j} \cdot \frac{\partial z_j}{\partial \omega_{ji}} \\
 &= \omega_{ji} - \eta \left(\sum_k \omega_{kj} \delta_k \right) \cdot f'(z_j) \cdot x_i \\
 &= \omega_{ji} - \eta f'(z_j) \sum_k \omega_{kj} \delta_k \cdot x_i \\
 &= \omega_{ji} - \eta \delta_j x_i \\
 \frac{\partial C}{\partial z_k} &= \delta_k \quad \frac{\partial z_k}{\partial a_j} = \frac{\partial \sum_l \omega_{lj} z_l}{\partial a_j} = \sum_l \omega_{lj} = \sum_k \omega_{kj} \\
 \frac{\partial a_j}{\partial z_j} &= \frac{\partial f(z_j)}{\partial z_j} = f'(z_j) \\
 \frac{\partial z_i}{\partial \omega_{ji}} &= \frac{\partial z_i - \omega_{ji} x_i}{\partial \omega_{ji}} = -x_i \\
 \omega_{ji} &
 \end{aligned}$$

1b)

$$1b, \quad X = B\text{-size} \times 785$$

$$W_{ji} = 785 \times 64$$

$$Z_j = \sum_{i=0}^d x_i w_{ji} \Rightarrow Z_j = B\text{-size} \times 64$$

$$\Downarrow$$

$$Z_j = X \cdot W_{ji} = B\text{-size} \times 64$$

$$A_j = f(Z_j) = B\text{-size} \times 64$$

$$\Delta_j = f'(Z_j) \sum_k w_{kj} \delta_k$$

$$W_{kj} = 64 \times 10$$

$$\Delta_k = -(\bar{y}_k - \hat{y}_k) = -(y_k - f(Z_k))$$

$$Z_k = A_j W_{kj} = B\text{-size} \times 10$$

$$y_k = B\text{-size} \times 10$$

$$\Delta_k = B\text{-size} \times 10$$

$$\delta_k =$$

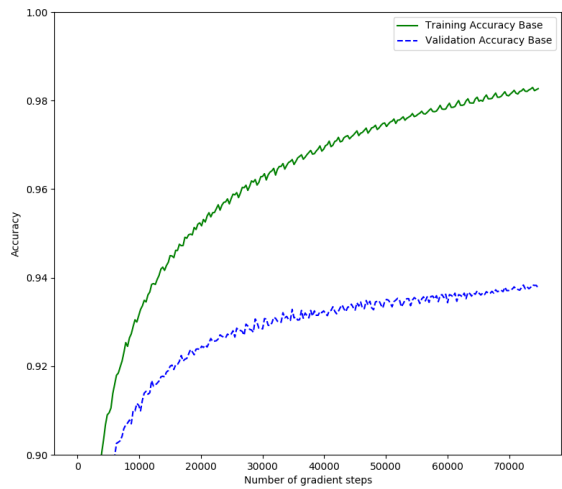
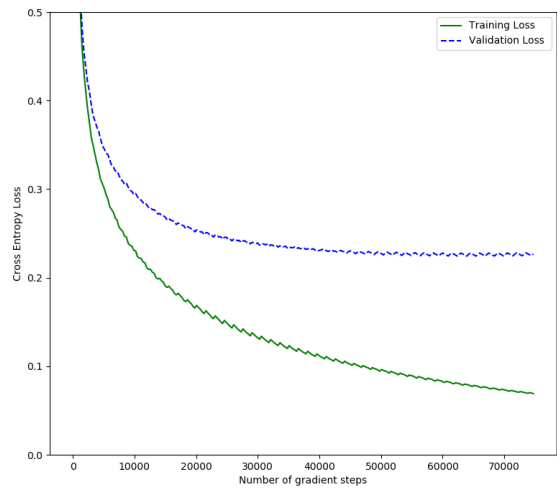
$$\Delta_j = f'(Z_j) \odot \Delta_k W_{kj}^T = B\text{-size} \times 64$$

$$W_{ji} := W_{ji} - \alpha X^T \Delta_j = 785 \times 64$$



$$\left[f(Z_j) = \frac{1}{1 + e^{-Z}} \right]$$

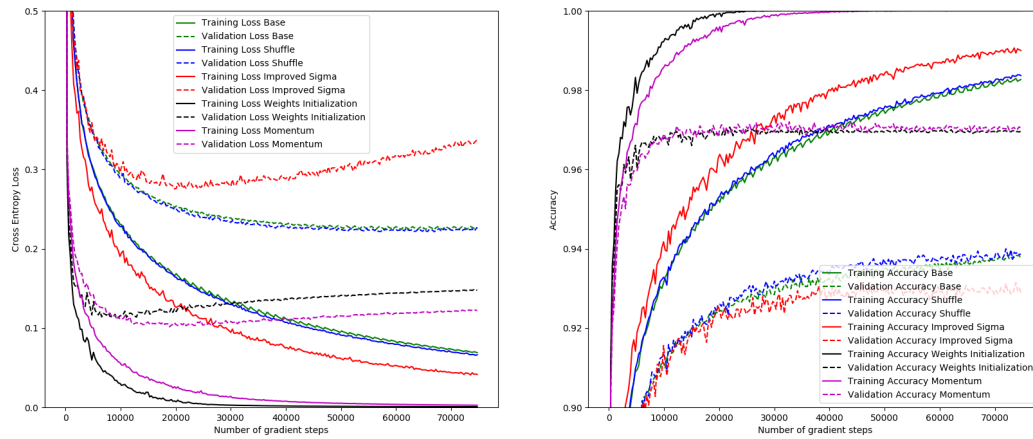
2c)



2d)

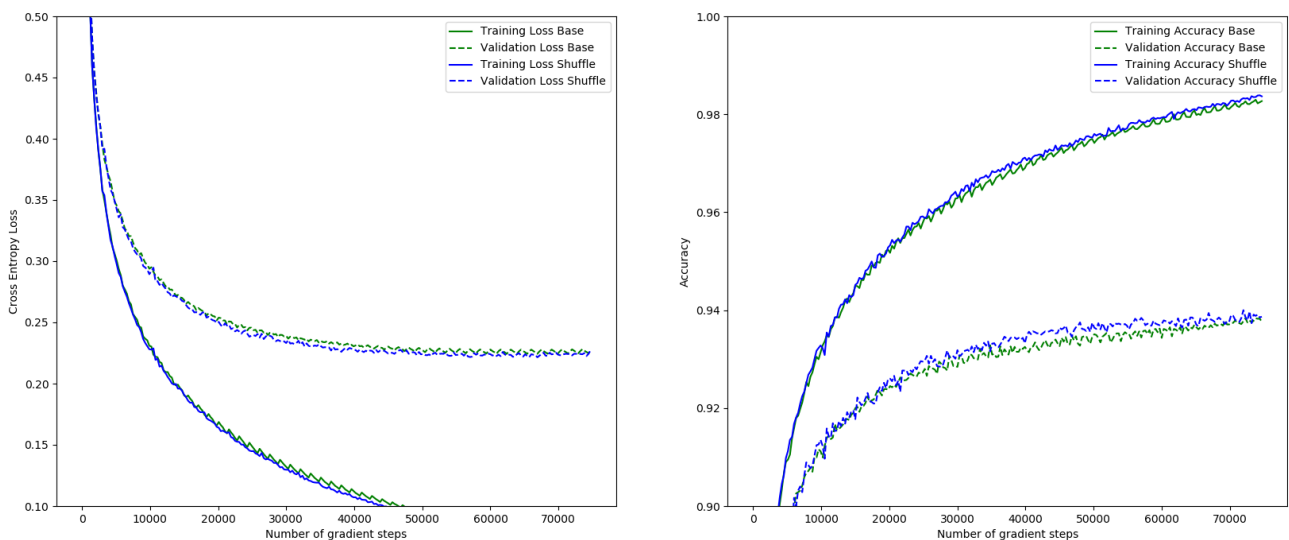
As we are using the bias trick all the weights and biases are included in the weight matrices. There are two weight matrices with the dimensions $[785, 64]$ and $[64, 10]$. In total there are $785 * 64 + 64 * 10 = 50880$ parameters.

This plot contains all plots created in task 3) for a general overview. The changes on the accuracy and the loss is discussed in detail by taking a look at each specific plot in the following subsections.



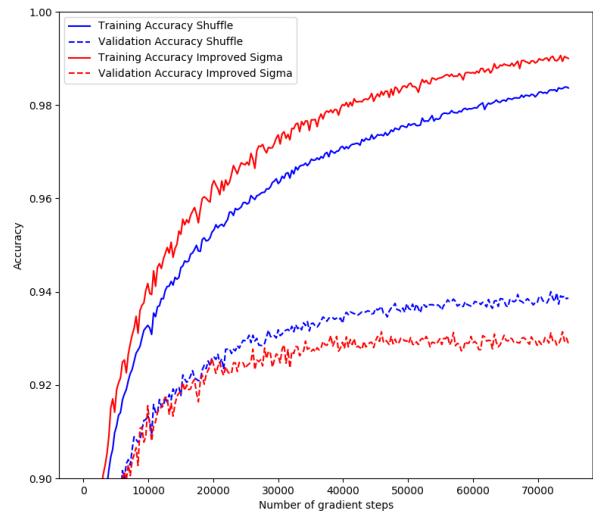
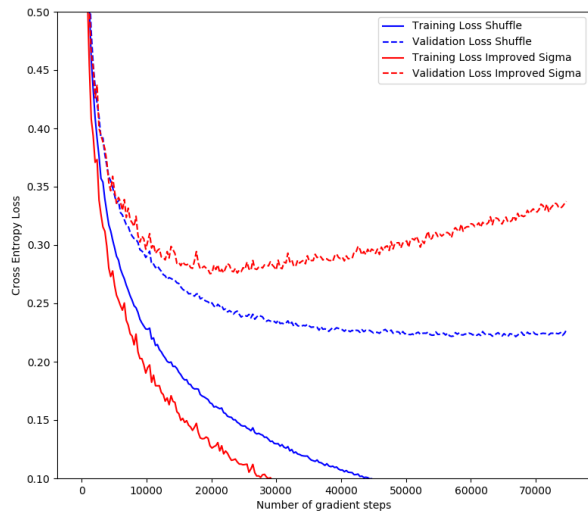
3a)

When shuffling the training data the convergence happens a little faster and the accuracy is marginally higher.



3b)

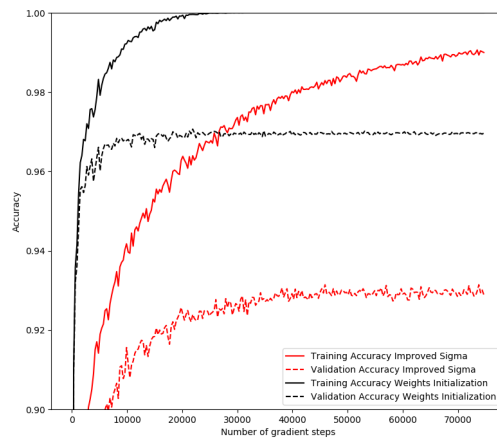
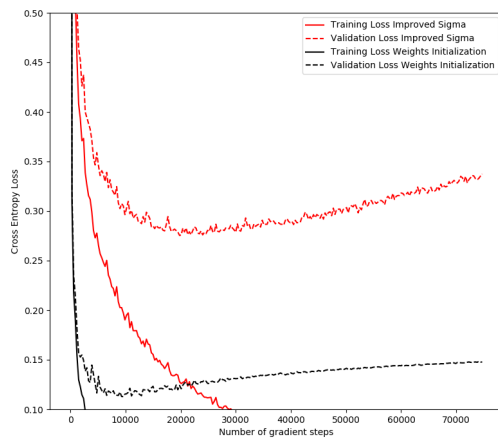
Using an improved sigmoid function increases the convergences speed significantly. This can be seen by the substantially lower training loss in comparison to the setup only using the shuffling. A major drawback is the worsened validation accuracy, this must be improved to achieve a better model.



3c)

Fan-in:

- First Layer = 785
- Second Layer = 64

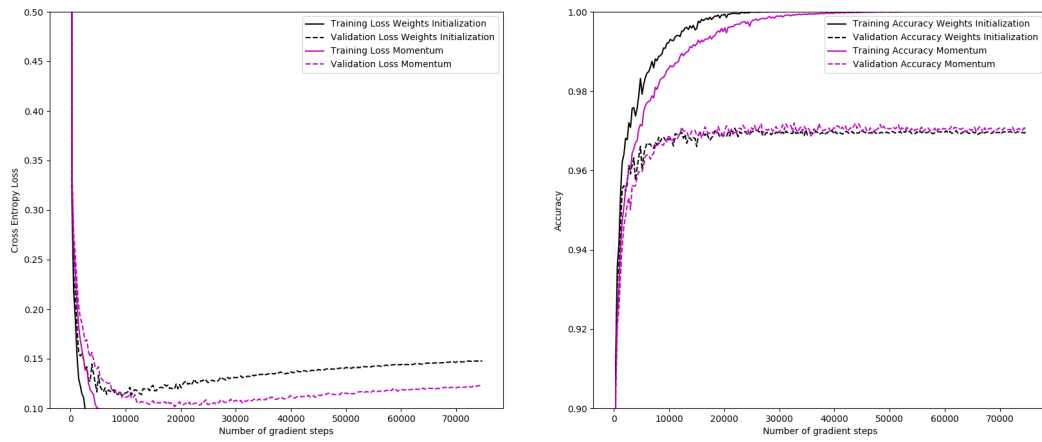


Using the improved weight initialization leads to a substantial increase in the convergence speed and the accuracy.

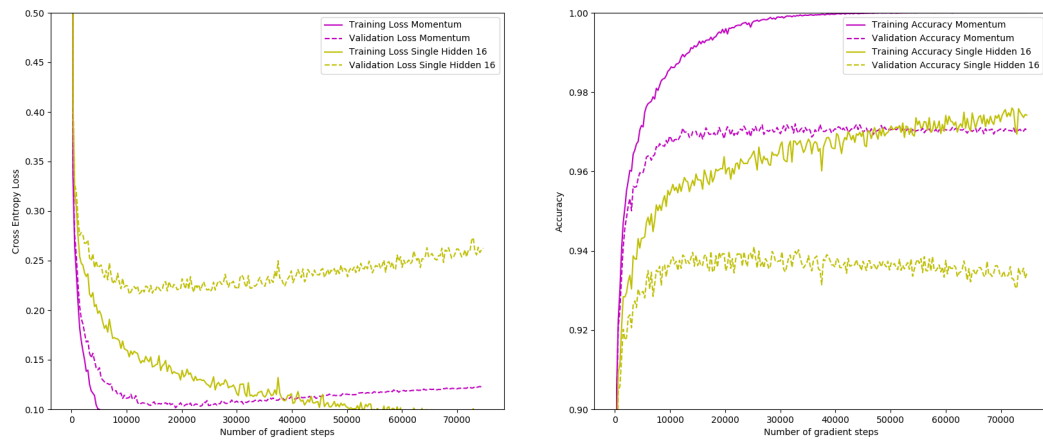
This can be seen in the significant and quick decrease of the cross entropy loss in the training set and the significant increase of the accuracy in the validation set.

3d)

Implementing the momentum yields in a slightly higher validation accuracy with the cost of slightly decreased convergence speed.



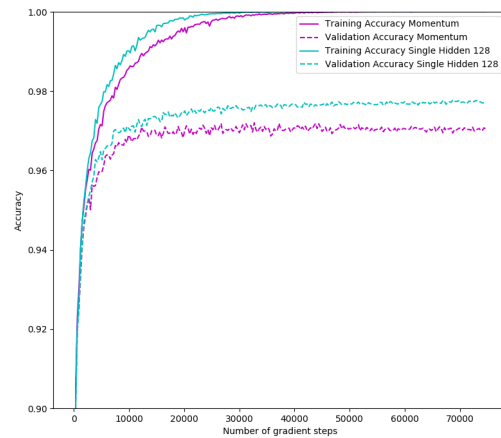
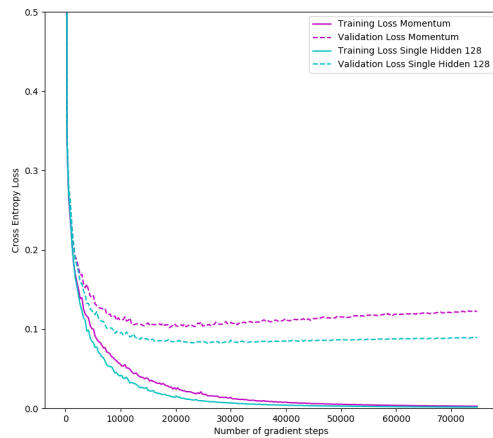
4a)



When a single hidden layer with 16 weights is used the accuracy of the neural network decreases substantially. The coverage speed is also slowed down.
On the opposite it is quicker to train a neural network when the hidden layer consists of only 16 weights.

4b)

Using a bigger hidden layer yields a significantly higher accuracy. The convergence (the point when the validation accuracy is approximately stable) takes more gradient steps. Additionally it takes more computation to train the bigger network as it contains a lot more parameters.



4d)

The network from 3) had weight matrices with the shape $[785, 64]$ and $[64, 10]$. This results in a total number of $785 * 64 + 64 * 10 = 50880$ parameters.

Using a neural network with two hidden layers with 60 nodes result in a similar amount of parameters. The weight matrices for such a network have the shape $[785, 60]$, $[60, 60]$, $[60, 10]$. A total of $785 * 60 + 60 * 60 + 60 * 10 = 51360$ parameters.

The validation accuracy of a network consisting of two hidden layers with 60 nodes each results in a similar validation accuracy. The convergence speed is also comparable to the network in 3)

In the provided code the gradient approximation test fails. After talking to the lab instructors this result is still considered correct, as the accuracy is still right.

