Learning results

I achieved best results on the validation set with a two layer network with 700 hidden units per layer, relu activation and a softmax outputlayer. I used a learning rate of 0.001 and a batchsize of 100 with an adam descent with parameters 0.9 and 0.995. For the regularizationterm $\frac{\alpha}{2}||W||^2$ I used $\alpha = 0.00025$.

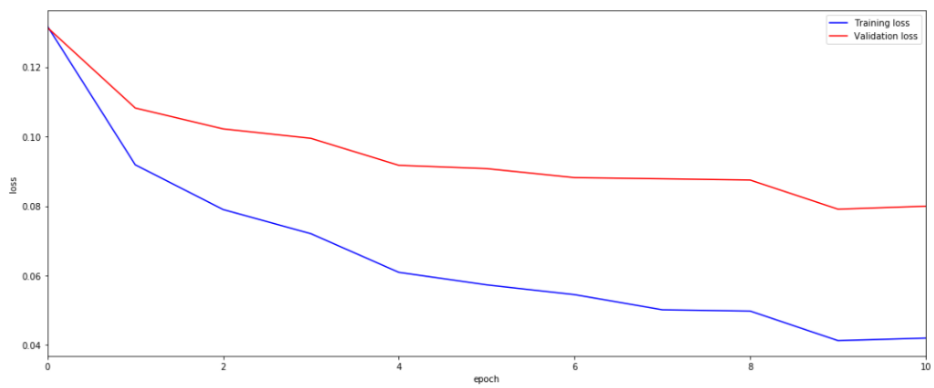

Figure 1: training and validation loss for above network.

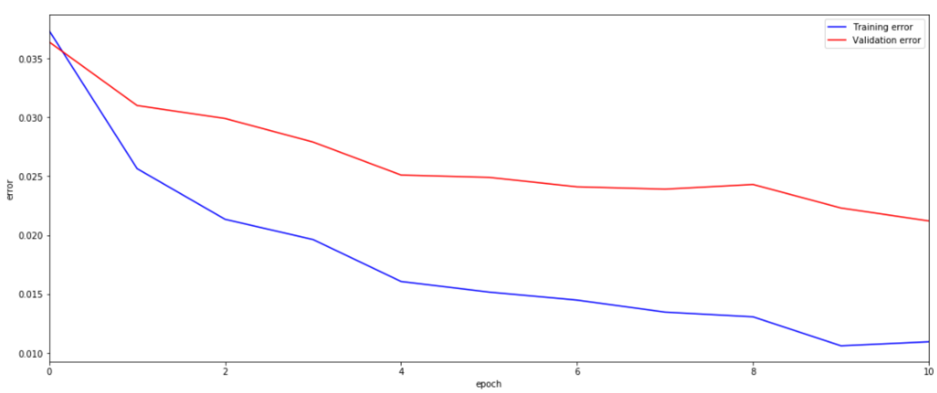

Figure 2: training and validation error for above network.
Final training error: 1.09%. Final validation error: 2.12%

When training the network on training and validation set and valuating on the test set, the network achived an error rate of 3.11%. On the combined validation and training set the network achieved a final error rate of 1.9%. While the training on the training set had a monoton decreasing error rate, the training on training and validation set combined had jumps in both error and loss that I can't explain.

Implementation Details

Derivation of the sigmoid function

$$
\begin{aligned}
s(x) &:= \frac{1}{1+\exp(-x)} \\
s'(x) &= \frac{\exp(-x)}{(1+\exp(-x))^2} \\
&= \frac{1}{1+\exp(-x)} \frac{\exp(-x)}{1+\exp(-x)} \\
&= s(x)\frac{\exp(-x)+1-1}{1+\exp(-x)} \\
&= s(x)\left(\frac{1+\exp(-x)}{1+\exp(-x)} \frac{-1}{1+\exp(-x)}\right) \\
&= s(x)(1-s(x))
\end{aligned}
$$

Derivation of the softmax function

$$
\begin{aligned}
\hat{y}_j(z) &:= \frac{\exp(z_j)}{\sum_k \exp(z_k)} \\
L(\hat{y}) &:= -\sum_j y_j \log(\hat{y}_j) \\
\frac{\partial \hat{y}_j}{\partial z_i} &= \begin{cases} \frac{\exp(z_j)}{\sum_k \exp(z_k)} - \frac{\exp(z_j)^2}{\left(\sum_k \exp(z_k)\right)^2} = \hat{y}_j(1-\hat{y}_j) & \text{for i =j} \\ \frac{-\exp(z_j)\exp(z_i)}{\left(\sum_k \exp(z_k)\right)^2} = -\hat{y}_j\hat{y}_i & \text{for i} \neq \text{j} \end{cases} \\
\frac{\partial L}{\partial z_i} &= -\sum_k y_k \frac{\log \hat{y}_k}{\partial z_i} = -\sum_k y_k \frac{1}{\hat{y}_k}\frac{\hat{y}_k}{\partial z_i} = -y_i(1-\hat{y}_i) - \sum_{k\neq i} y_k \frac{1}{\hat{y}_k}(-\hat{y}_k\hat{y}_i) \\
&= -y_i + y_i\hat{y}_i + \sum_{k\neq i} y_k\hat{y}_i = \left(\sum_k y_k\right)\hat{y}_i - y_i \\
&= \hat{y}_i - y_i \qquad \text{since y is one hot vector}
\end{aligned}
$$

Derivation of the L2 regularization term

$$
\begin{aligned}
\frac{\partial}{\partial w_{i,j}}(L + \frac{\alpha}{2}||W||_2^2) &= \frac{\partial L}{\partial w_{i,j}} + \frac{\partial}{\partial w_{i,j}}\frac{\alpha}{2}\sum_{k,l} w_{k,l}^2 \\
&= \frac{\partial L}{\partial w_{i,j}} + \alpha w_{i,j}
\end{aligned}
$$