Assignment 1 Report

This is an outline for your report to ease the amount of work required to create your report. Jupyter notebook supports markdown, and I recommend you to check out this cheat sheet. If you are not familiar with markdown.

Before delivery, **remember to convert this file to PDF**. You can do it in two ways:

1. Print the webpage (ctrl+P or cmd+P)
2. Export with latex. This is somewhat more difficult, but you'll get somehwat of a "prettier" PDF. Go to File -> Download as -> PDF via LaTeX. You might have to install nbconvert and pandoc through conda; `conda install nbconvert pandoc`.

Task 1

task 1a)

$$C^n(w) = -\left(y^n \ln(\hat{y}^n) + (1-y^n)\ln(1-\hat{y}^n)\right)$$

Using that $y \neq y(w_i)$:

$$\frac{\partial C^n(w)}{\partial w_i} = -y\frac{\partial \ln \hat{y}}{\partial w_i} - (1-y)\cdot\frac{\partial \ln(1-\hat{y})}{\partial w_i}$$

First term, using the chain-rule and the hint $\frac{\partial f(x^n)}{\partial w_i} = x_i^n f\left(x^n\right)\left(1-f\left(x^n\right)\right)$:

$$\frac{\partial \ln \hat{y}^n}{\partial w_i} = \frac{1}{\hat{y}^n}\cdot\frac{\partial \hat{y}^n}{\partial w_i} = \frac{1}{\hat{y}^n}\cdot x_i^n \cdot \hat{y}^n \cdot (1-\hat{y}^n) = \underline{\underline{x_i^n\left(1-\hat{y}^n\right)}}$$

Second term, using the chain-rule and the hint $\frac{\partial f(x^n)}{\partial w_i} = x_i^n f\left(x^n\right)\left(1-f\left(x^n\right)\right)$:

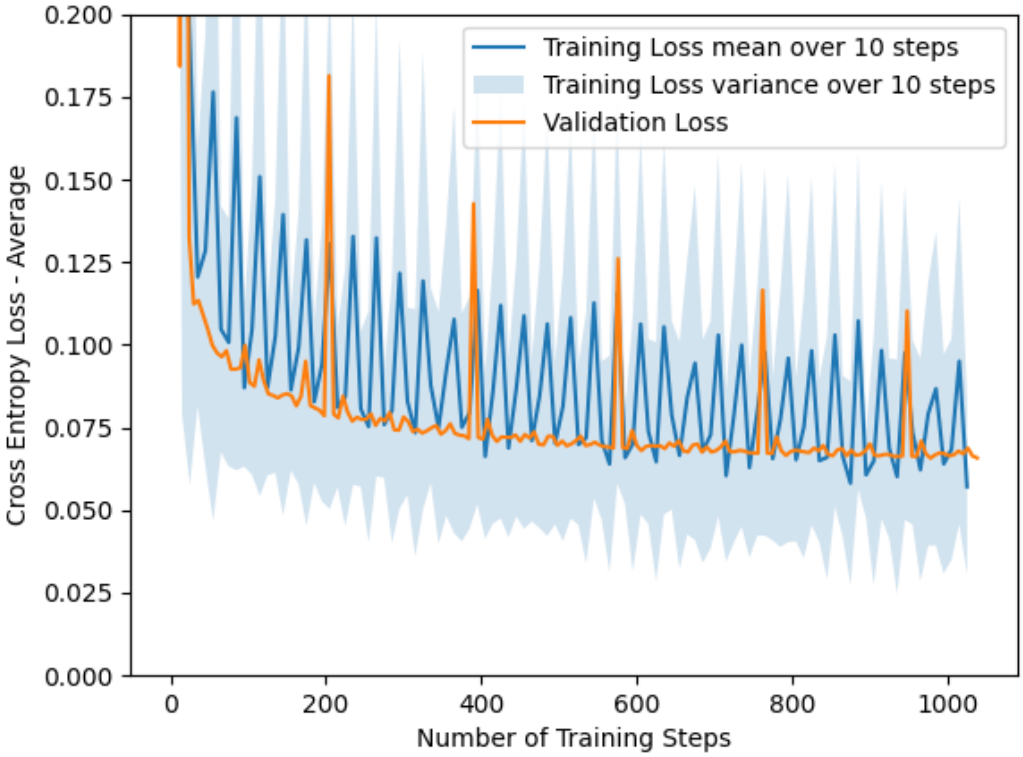
$$\frac{\partial \ln(1-\hat{y}^n)}{\partial w_i} = -\frac{1}{1-\hat{y}^n}\cdot\frac{\partial(1-\hat{y}^n)}{\partial w_i} = -\frac{1}{1-\hat{y}^n}\cdot x_i^n\cdot(1-\hat{y}^n)\cdot(1-1+\hat{y}^n) = \underline{\underline{-}}$$

Put together:

$$\frac{\partial C^n}{\partial w_i} = -y^n x_i^n(1-\hat{y}^n) + (1-y^n)x_i^n\hat{y}^n = x_i^n(-y^n + y^n\hat{y}^n + \hat{y}^n - y^n\hat{y}^n) = x_i^n(\hat{y}^n -$$

Task 2

Task 2b)



Task 2c)

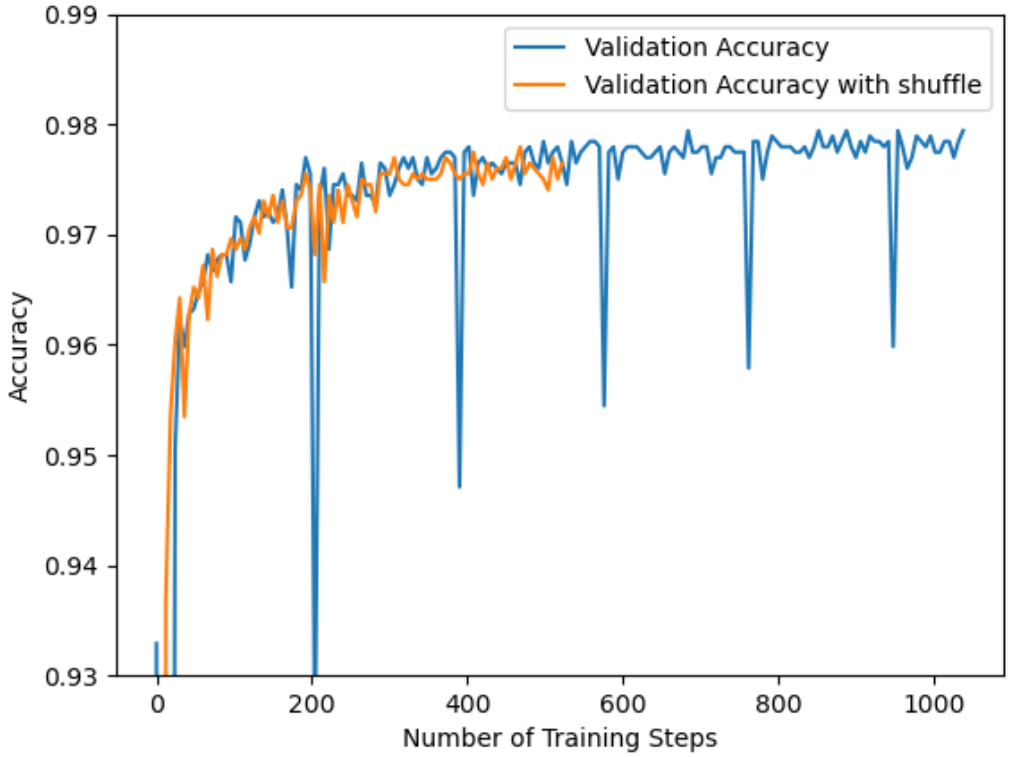
Task 2d)

Early stopping in Epoch 33.

Task 2e)

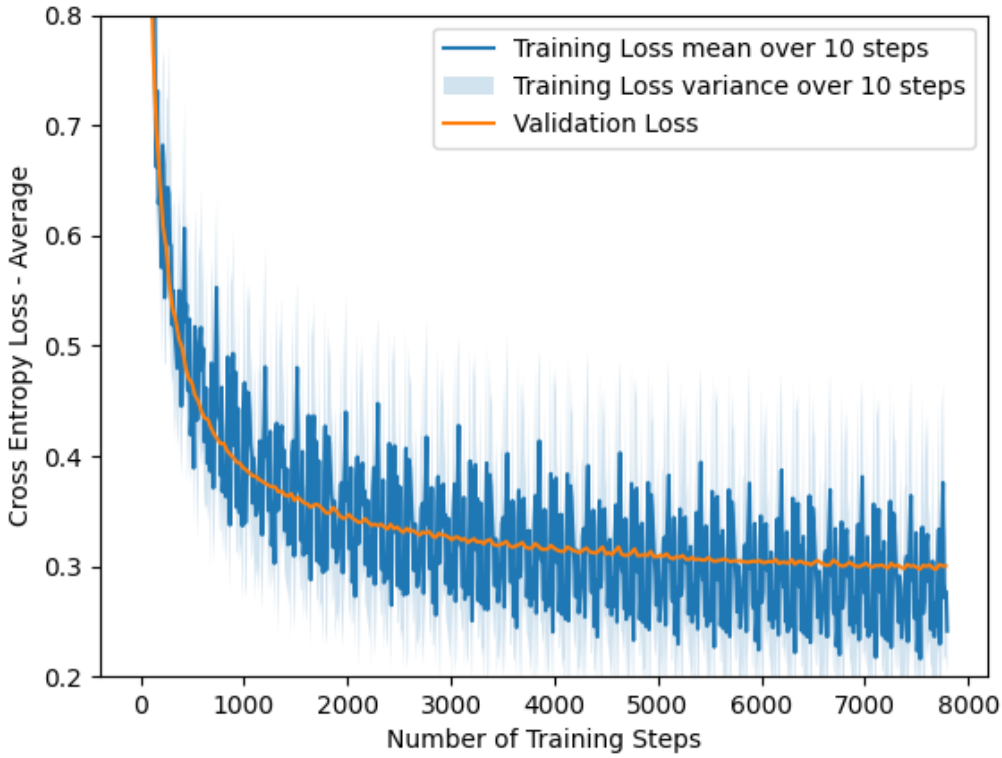
You should notice that the validation accuracy has fewer ”spikes”. Why does this happen?

- By shuffling the dataset before each epoch, the data to train the model is more randomly distributed preventing possible correlations given by the order of the samples in the original dataset.

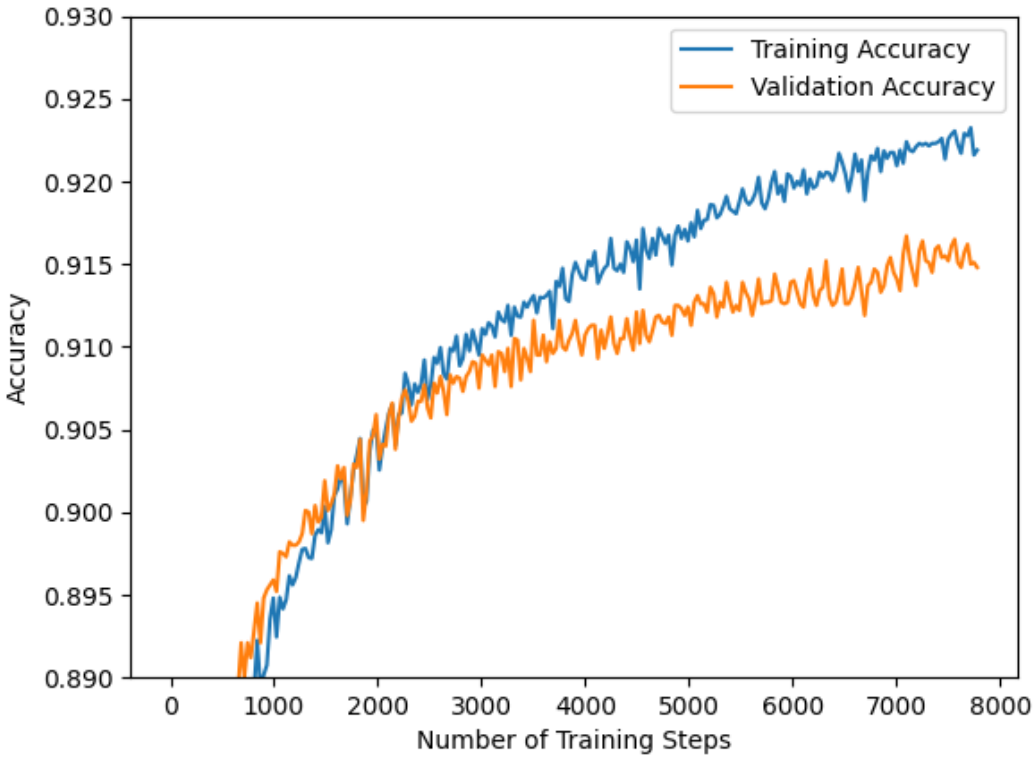


Task 3

Task 3b)



Task 3c)



Task 3d)

At the start of the training the validation accuracy is higher than the training accuracy. Then the training accuracy increases more and more while the validation accuracy plateaus. This is an indicator for overfitting.

Task 4

Task 4a)

$$J(w) = C(w) + \lambda R(w), \qquad R(w) = \|w\|^2 = \sum_{i,j} w_{i,j}^2$$

$$\frac{\partial J}{\partial w} = \frac{\partial C}{\partial w} + \frac{\partial}{\partial w}(\lambda\, R(w))$$

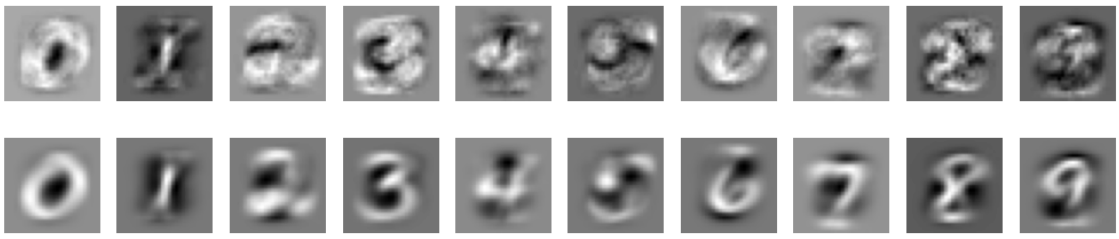
Using from before: $\frac{\partial C^n}{\partial w_{kj}} = -x_j^n \cdot (y_k^n - \hat{y}_k^n)$, and $\frac{\partial}{\partial w}(\lambda\, R(w)) = 2\lambda w$.

$$\frac{\partial J}{\partial w_{kj}} = -x_j^n \cdot (y_k^n - \hat{y}_k^n) + 2\lambda w_{kj}$$

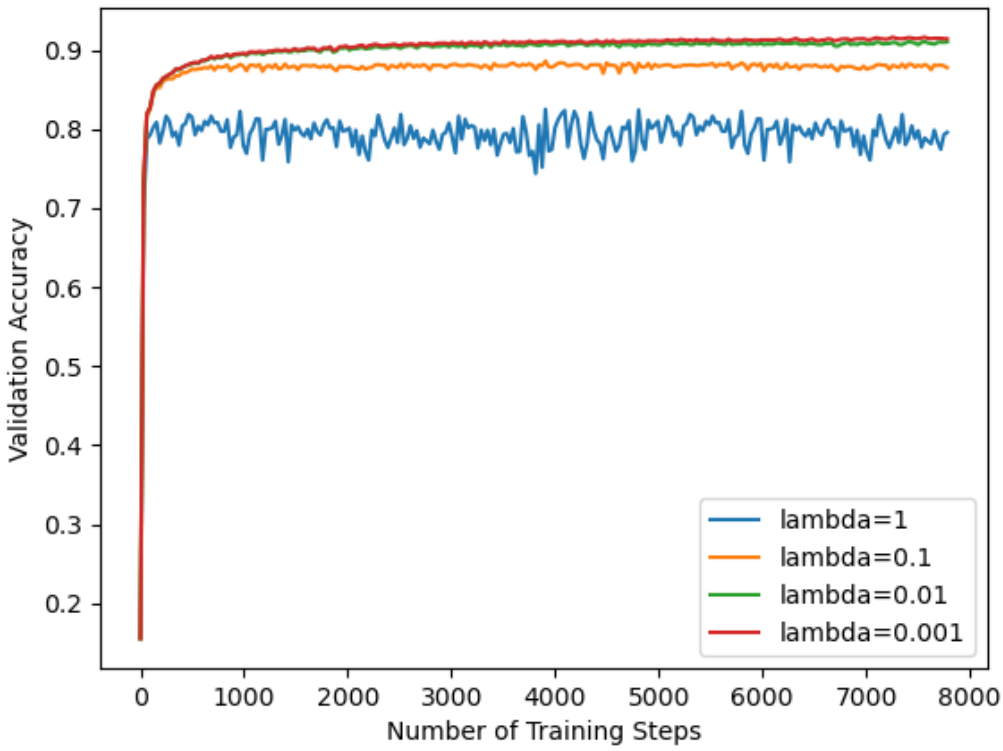
Task 4b)

First row: $\lambda = 0$, second row: $\lambda = 1$. Why are the weights for the model with λ = 1.0 less noisy?

Regularization is leading that the overall squared weight sum is leading to zero and therefore the absolute weights become smaller. However, to reduce the cost function, the right weights still need to be active in order to reduce the classification error. Combining both effects (right weights active, everything else going to zero), leading to significantly smaller noises for the weights.



Task 4c)



Task 4d)

You will notice that the validation accuracy degrades when applying any amount of regularization. What do you think is the reason for this?

With too strong regularization, the model will tend to underfitting so the model complexity and the validation accuracy decreases. Common regularization parameters often range between 0 and 0.1.

In addition, the accuracy only take into account the percentage of total correct predictions and might not be the correct metric.

Task 4e)

What do you observe?

For a higher regularization factor λ the L2-norm of the weights become smaller, which is exspected since λ controls how much the L2-norm is weighted in the cost function. For bigger λ the L2-norm term becomes more important in the cost function and therefore will become smaller in order to minimize the lost function.