AssignmentReport-Group1

February 3, 2022

1 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`.

2 Task 1

2.1 Task 1a)

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

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

$$\frac{\partial}{\partial z}\frac{1}{1+e^z} = -(1+e^z)^{-2}e^z$$

$$\frac{\partial \hat{y}^n}{\partial w_i} = \frac{\partial}{\partial w_i}\frac{1}{1+e^{-w^T x}} = -(1+e^{-w^T x})^{-2}e^{-w^T x}\frac{\partial}{\partial w_i}(-w^T x) = -(\hat{y}^n)^2(\frac{1}{\hat{y}^n} - 1)(-x_i) = \hat{y}^n(1-\hat{y}^n)x_i$$

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

2.2 Task 1b)

$$\frac{\partial C^n}{\partial w_{kj}} = -\frac{\partial}{\partial w_{kj}}\sum_{i=1}^{K} y_i^n \ln(\hat{y}_i^n) = -\sum_{i=1}^{K}\frac{y_i^n}{\hat{y}_i^n}\frac{\partial \hat{y}_i^n}{\partial w_{kj}} = -\sum_{i=1}^{K}\frac{y_i^n}{\hat{y}_i^n}\frac{\partial \hat{y}_i^n}{\partial z_k}\frac{\partial z_k}{\partial w_{kj}}$$

$$\frac{\partial \hat{y}_i^n}{\partial z_k}(\text{ for } i \neq k) = \frac{\partial}{\partial z_k}\frac{e^{z_i}}{e^{z_k} + \sum_{k' \neq k} e^{z_{k'}}} = e^{z_i}(-1)\left(\sum_{k'=1} e^{z_{k'}}\right)^{-2} e^{z_k} = -\hat{y}_i^n \hat{y}_k^n$$

$$\frac{\partial \hat{y}_k^n}{\partial z_k} = \frac{\partial}{\partial z_k} \frac{e^{z_k}}{\sum_{k'=1}^{K} e^{z_{k'}}} = \frac{1}{\sum_{k'=1}^{K} e^{z_{k'}}} \frac{\partial}{\partial z_k}(e^{z_k}) + e^{z_k} \frac{\partial}{\partial z_k} \left(\sum_{k'=1}^{K} e^{z_{k'}} \right)^{-1}$$

$$= \frac{e^{z_k}}{\sum_{k'=1}^{K} e^{z_{k'}}} + e^{z_k}(-1) \left(\sum_{k'=1}^{K} e^{z_{k'}} \right)^{-2} e^{z_k} = \frac{e^{z_k}}{\sum_{k'=1}^{K} e^{z_{k'}}} \left(1 - \frac{e^{z_k}}{\sum_{k'=1}^{K} e^{z_{k'}}} \right) = \hat{y}_k^n (1 - \hat{y}_k^n) = \hat{y}_k^n - \hat{y}_k^n \hat{y}_k^n$$

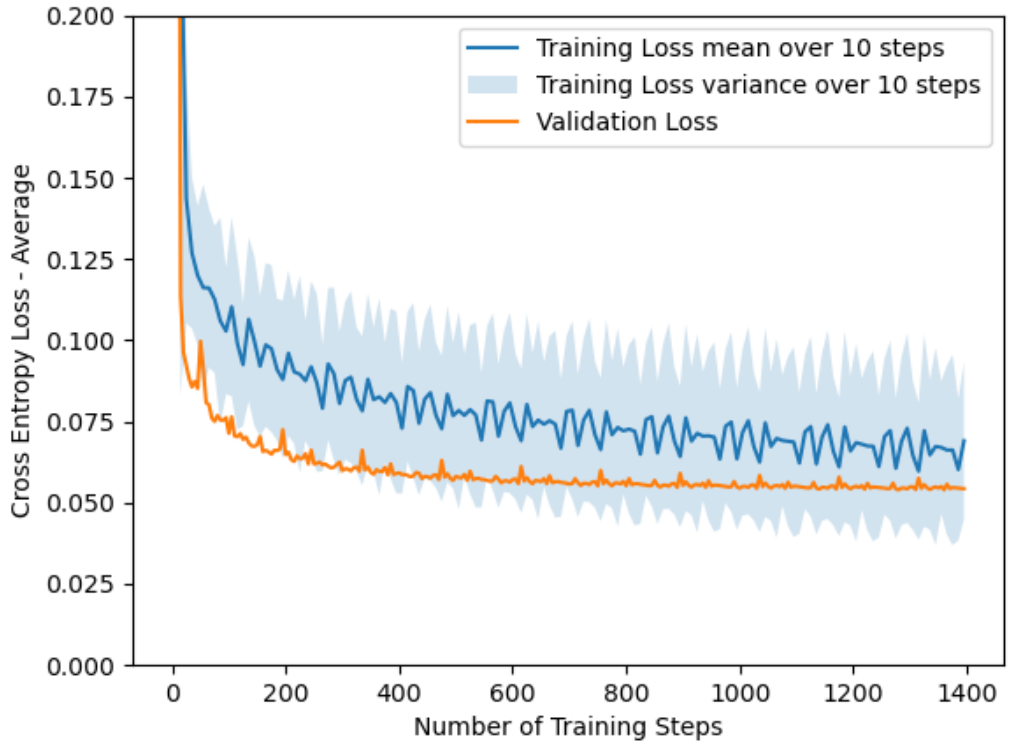
$$\frac{\partial z_k}{\partial w_{kj}} = x_j$$

$$\frac{\partial C^n}{\partial w_{kj}} = -\sum_{i=1}^{K} \frac{y_i^n}{\hat{y}_i^n} \frac{\partial \hat{y}_i^n}{\partial z_k} \frac{\partial z_k}{\partial w_{kj}} = -\frac{y_k^n}{\hat{y}_k^n} \hat{y}_k^n x_j - \sum_{i=1}^{K} \frac{y_i^n}{\hat{y}_i^n} (-\hat{y}_i^n \hat{y}_k^n) x_j = x_j \left(-y_k^n + \hat{y}_k^n \sum_{i=1}^{K} y_i^n \right) = -x_j (y_k^n - \hat{y}_k^n)$$

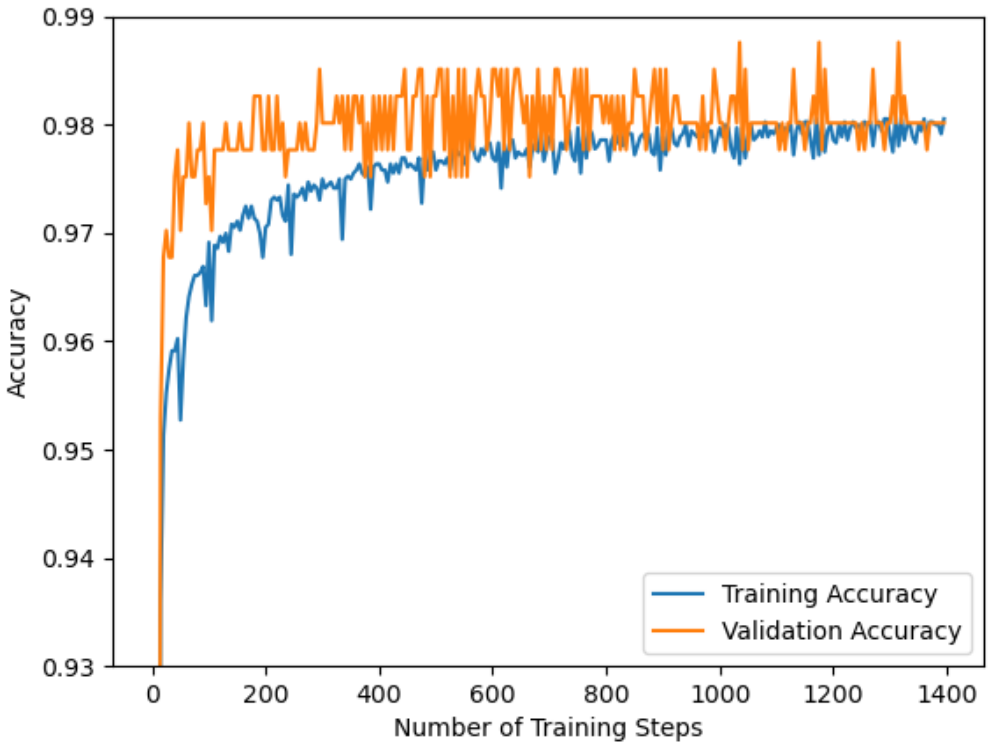
3 Task 2

The first tasks were done by using data that was sampled stochastically using the given code on github. This caused less pikes in the validation accuracy. However the stochastic sampling was turned off in task 2e to get some spikes to compare shuffling vs not shuffling.

3.1 Task 2b)

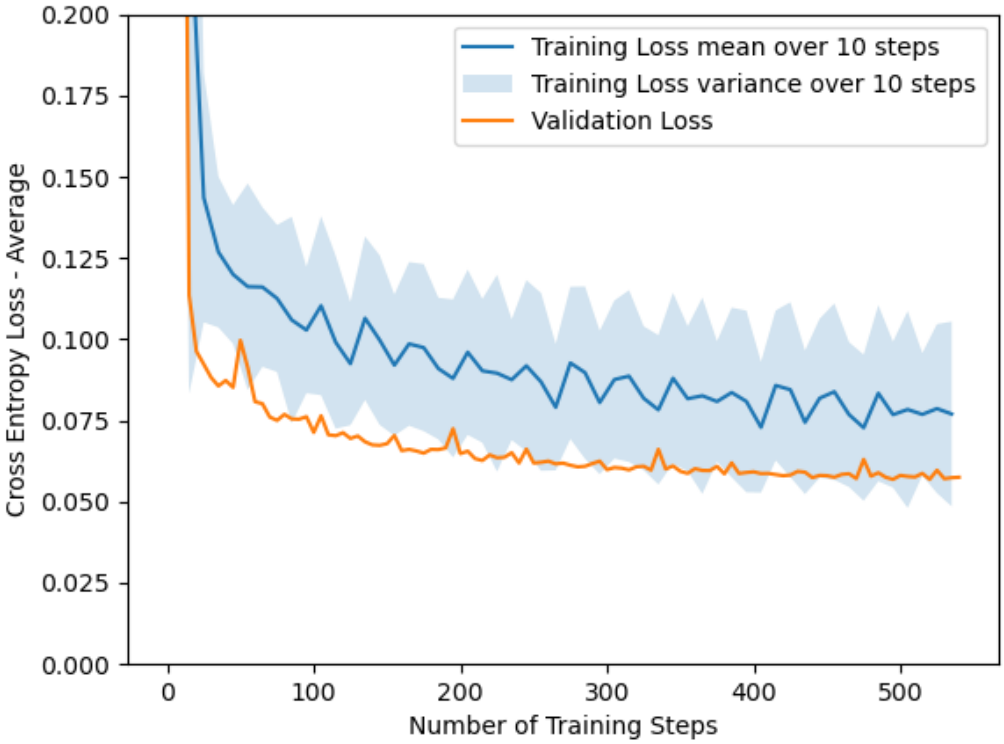


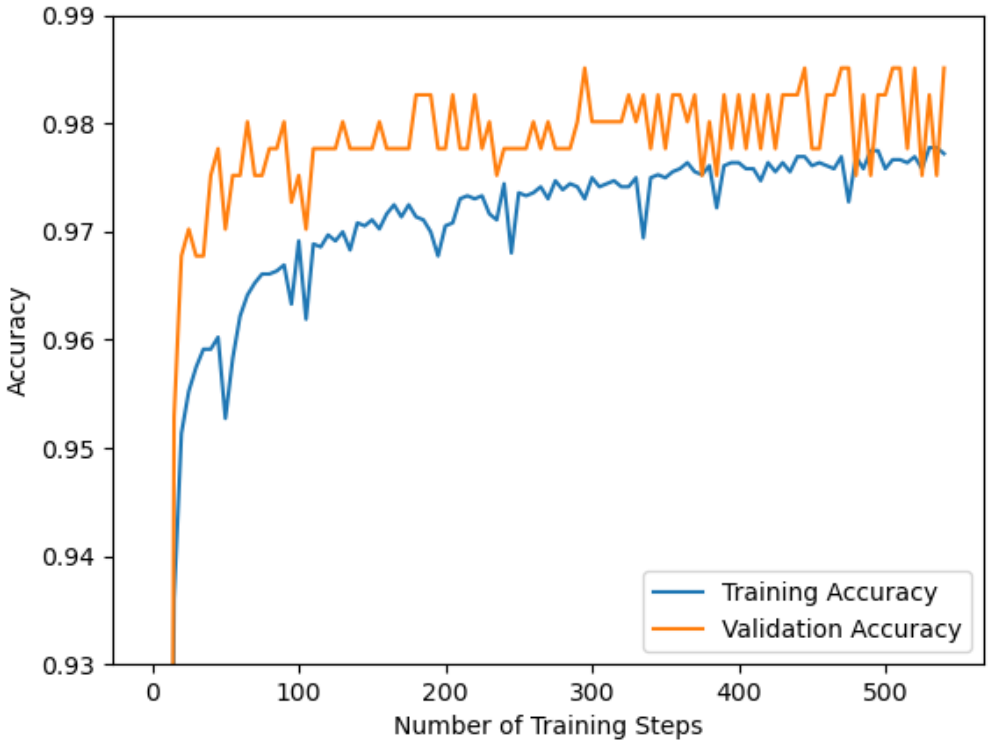
3.2 Task 2c)



3.3 Task 2d)

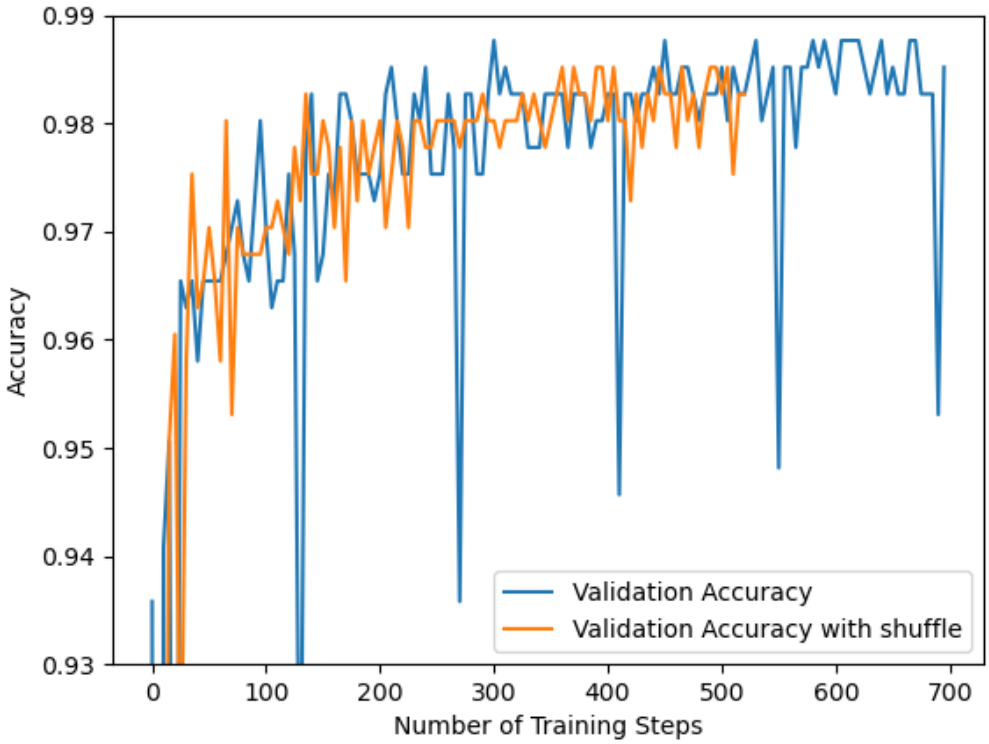
It stops after around 550 steps which is in the 19th epoch.


Training Loss mean over 10 steps
Training Loss variance over 10 steps
Validation Loss
Cross Entropy Loss - Average
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
100
200
300
400
500
Number of Training Steps


Accuracy
0.99
0.98
0.97
0.96
0.95
0.94
0.93
Training Accuracy
Validation Accuracy
0
100
200
300
400
500
Number of Training Steps

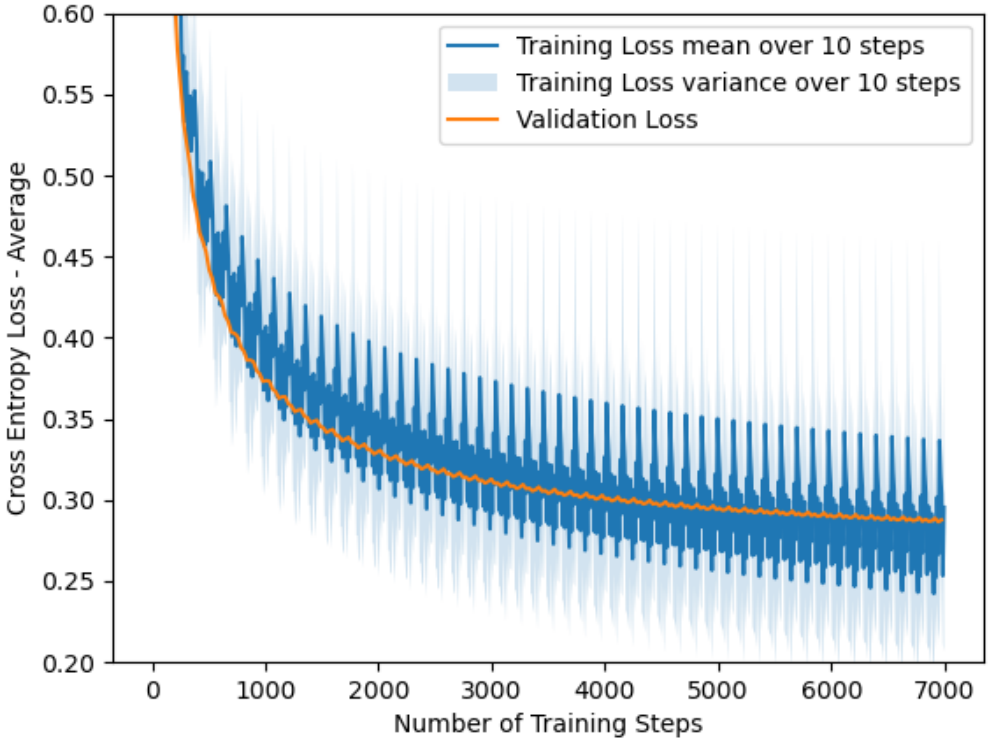
3.4 Task 2e)

When shuffling the training set one calculates the gradient on different parts of the data set each epoch making it a better representation of a real situation which again makes it better at the validation set.

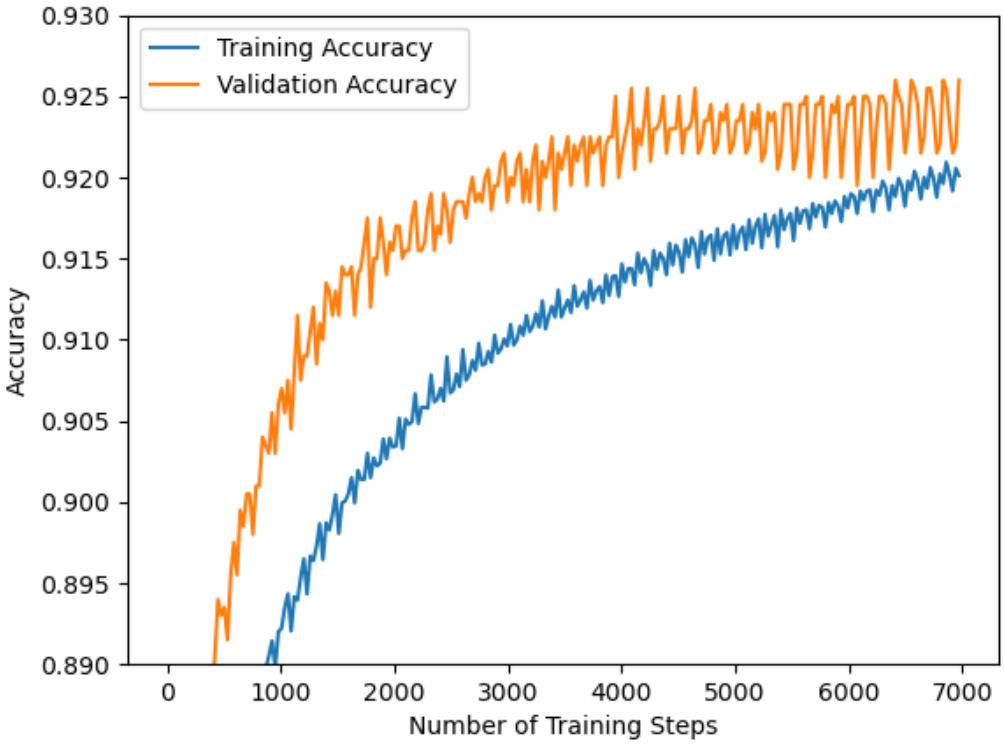


4 Task 3

4.1 Task 3b)



4.2 Task 3c)



4.3 Task 3d)

From the plot of training accuracy compared to calidation accuracy, we can see some signs of overfitting. After around 3800 training steps, the validation accuracy stops improving and rather begins oscillating around the value 0.924. However, the training accuracy is still increasing. This is undesired because we want our predictor to perform well on new data, rather than just the training data.

5 Task 4

5.1 Task 4a)

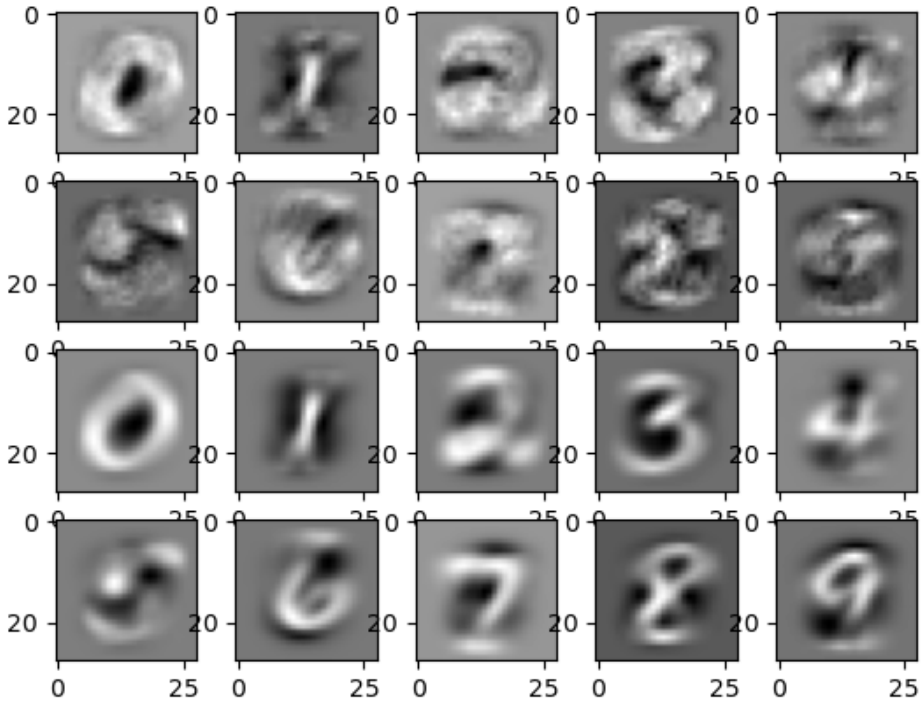
$$\frac{\partial J(w)}{\partial w_{kj}} = \frac{\partial}{\partial w_{kj}}(C(w) + \lambda R(w)) = \frac{\partial C}{\partial w_{kj}} + \lambda \frac{\partial R}{\partial w_{kj}}$$

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

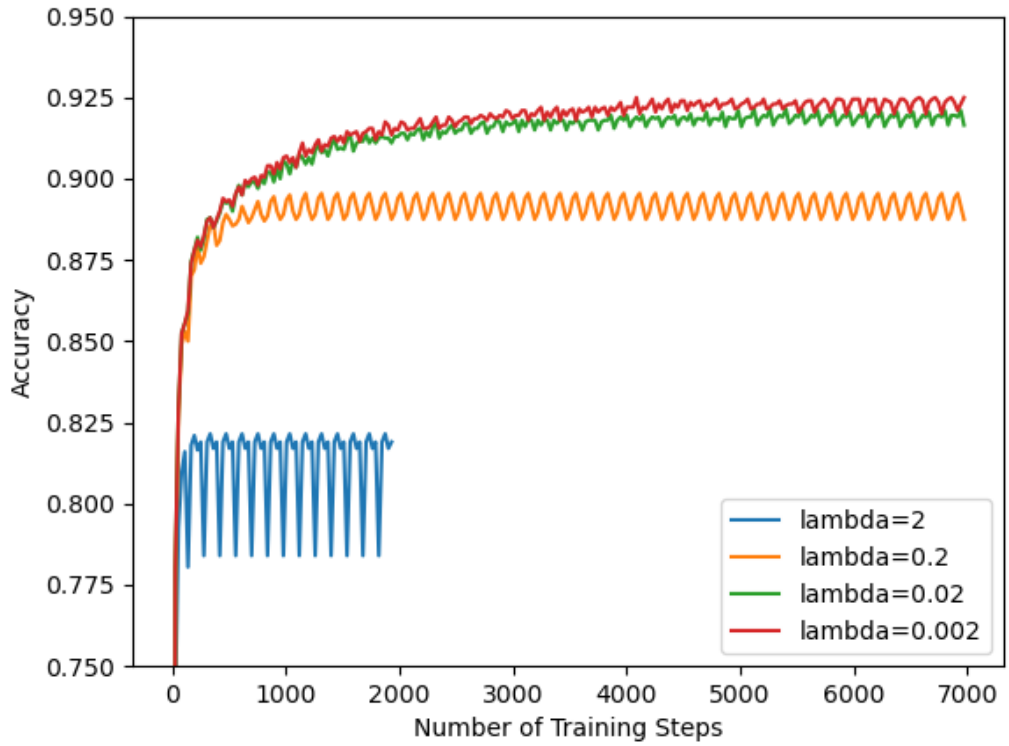
5.2 Task 4b)

When using regularization ($\lambda = 2$) the algorithm tries to keep the weights small while making good guesses. It will then prioritize to use the weights to extract the most important features of

the image. E.g the weights for recognizing a 1 with regularization just recognizes straight line i the middle, while the weights without regularization seems to try to recognize line with multiple gradients.



5.3 Task 4c)



5.4 Task 4d)

With regularization the error have to be sufficiently large so that the regularization part of the gradient does not dominate so that the weights can be changed in the direction of better training accuracy. At some point the weights will become sufficiently large so that the gradient becomes very small while there still is errors to correct. This effect stops it from achieving better training accuracy which indirectly affect the validation accuracy.

5.5 Task 4e)

The L2-norm of the weights drastically decrease with increasing λ.