Assignment 1 Report TDT4265

Group 16

Hanna Waage Hjelmeland, Siri Holde Hegsvold, MTTK

Task 1

task 1a)

Fill in task 1a image of hand-written notes which are easy to read, or latex equations here

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

Hint: $\frac{\partial f(x^n)}{\partial w_i} = x_i^n f(x^n)(1 - f(x^n)) = x_i^n \hat{y}^n (1 - \hat{y}^n)$

$$\frac{\partial C^n}{\partial w_i}(w) = \frac{\partial}{\partial w_i}\Big(\underbrace{-(y^n \ln(\hat{y}^n)}_{\text{I}} + \underbrace{(1-y^n)\ln(1-\hat{y}^n)}_{\text{II}}\Big)$$

Recall: $\hat{y}^n = f(x^n)$.

Start with I:

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

$$= -\frac{y^n}{\hat{y}^n} \frac{\partial f(x^n)}{\partial w_i} = -\frac{y^n}{\hat{y}^n} \cdot x_i^n \hat{y}^n (1 - \hat{y}^n)$$

$$= -y^n x_i^n (1 - \hat{y}^n)$$

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

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

$$= -(1-y^n)(-x_i^n \hat{y}^n)$$

$$\Rightarrow \frac{\partial C^n}{\partial w_i}(w) = \text{I} + \text{II} = -y^n x_i^n(1-\hat{y}^n) - (1-y^n)(-x_i^n \hat{y}^n)$$

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

$$= -x_i^n\left(y^n - \hat{y}^n y^n - \hat{y}^n + \hat{y}^n y^n\right)$$

$$= -(y^n - \hat{y}^n)x_i^n \quad \square$$

Task 1b)

1.b)

$$C^n(w) = -\sum_{k=1}^{K} y_k^n \ln(\hat{y}_k^n)$$

$$\hat{y}_k = \frac{e^{z_k}}{\sum_{k'} z_{k'}} \qquad z_k = w_k^T x^n$$

$$\sum_{k'} e$$

$$\frac{\partial C^n}{\partial \omega_{kj}}(\omega) = \frac{\partial}{\partial \omega_{kj}}\left(-\sum_{k=1}^{K} y_k^n \ln(\hat{y}_k^n)\right)$$

$$= -\sum_{k=1}^{K} \frac{y_k^n}{\hat{y}_k^n} \cdot \frac{\partial}{\partial \omega_{kj}} \hat{y}_k^n$$

$$\frac{\partial}{\partial \omega_{kj}} \hat{y}_k^n = \frac{\partial}{\partial \omega_{kj}} \frac{e^{z_k}}{\sum_c e^{z_c}}$$

Using the quotient rule

$$\frac{d}{d\omega_j} \frac{f(\omega)}{g(\omega)} = \frac{\frac{\partial}{\partial \omega_{kj}}[f(\omega)]g(\omega) - f(\omega)\frac{\partial}{\partial \omega_{kj}}[g(\omega)]}{g^2(\omega)}$$

with $f(\omega) = e^{z_k} = e^{\omega_k^T x^n}$, $g(\omega) = \sum_c^K e^{\omega c^T x^n}$

$c = k$:

$$\frac{\partial}{\partial \omega_{kj}} e^{z_k} = \frac{\partial}{\partial \omega_{kj}}\left(e^{\omega_{k0}^T x_0^n} \cdot e^{\omega_{k1}^T x_1^n} \cdots e^{\omega_{kj}^T x_j^n} \cdots\right)$$

$$= x_j^n\left(e^{\omega_{k0}^T x_0^n} \cdot e^{\omega_{k1}^T x_1^n} \cdots e^{\omega_{kj}^T x_j^n} \cdots\right) = x_j^n e^{\omega_k^T x^n}$$

$c \neq k$: $\frac{\partial}{\partial w_{kj}} e^{z_{k'}} = 0$

Let $f(w) = e^{z_k} = e^{w_k^T x^n}$, $g(w) = \sum_{k'}^{K} e^{w_{k'}^T x^n}$

$\Rightarrow \frac{\partial}{\partial w_{kj}} f(w) = x_j^n e^{w_k^T x^n}$, $\frac{\partial}{\partial w_{kj}} g(w) = x_j^n e^{w_k^T x^n}$

$$\Rightarrow \frac{\partial}{\partial w_{kj}} \hat{y}_k^n = \frac{\partial}{\partial w_{kj}} \frac{f(w)}{g(w)} = \frac{\frac{\partial}{\partial w_{kj}}[f(w)]g(w) - f(w)\frac{\partial}{\partial w_{kj}}[g(w)]}{g^2(w)}$$

$$= \frac{x_j^n e^{w_k^T x^n} \sum_{c}^{K} e^{w_{k'}^T x^n} - e^{w_k^T x^n} x_j^n e^{w_k^T x^n}}{\left(\sum_{c}^{K} e^{w_{k'}^T x^n}\right)^2}$$

$$= x_j \left(\frac{e^{w_k^T x^n}}{\sum_{c}^{K} e^{w_{k'}^T x^n}} \cdot \frac{\left(\sum_{k'}^{K} e^{w_{k'}^T x^n} - e^{w_k^T x^n}\right)}{\sum_{c}^{K} e^{w_{k'}^T x^n}} \right)$$

$$= x_j \left(\hat{y}_k^n \cdot (1 - \hat{y}_k^n) \right)$$

$$\frac{\partial}{\partial w_{kj}} C^n(w) = -\sum_{k'=1}^{K} \frac{y_{k'}^n}{\hat{y}_{k'}^n} \cdot \frac{\partial}{\partial w_{kj}} \hat{y}_{k'}^n$$

need to find $\frac{\partial}{\partial w_{kj}} \hat{y}_{k'}^n = \frac{\partial}{\partial w_{kj}} \frac{e^{z_{k'}}}{\sum_{c}^{K} e^{z_c}}$,

i.e. when $k' \neq k$. Using f, g as

i.e. when $k = k'$

previously, we obtain

$$\frac{\partial}{\partial w_{kj}} f(w) = \frac{\partial}{\partial w_{kj}} e^{z_{k'}} = 0 \quad \ldots e^{z_k} \text{ is contained here}$$

$$\frac{\partial}{\partial w_{kj}} g(w) = \frac{\partial}{\partial w_{kj}} \left(\sum_{c}^{k} e^{z_{c'}} \right) = x_j e^{w_k^T x^n}$$

$$\Rightarrow \frac{\partial}{\partial w_{kj}} \hat{y}_{k'}^n = \frac{0 - e^{w_{k'}^T x^n} x_j e^{w_k^T x^n}}{\left(\sum_{c}^{k} e^{w_{k'}^T x^n} \right)^2}$$

$$= - x_j \hat{y}_{k'}^n \hat{y}_k^n$$

$$\Rightarrow \frac{\partial C^n}{\partial w_{kj}} = - \sum_{k'=1}^{K} \frac{y_{k'}^n}{\hat{y}_{k'}^n} \frac{\partial \hat{y}_{k'}^n}{\partial w_{kj}}$$

$$= - \left(\sum_{k'=k} \frac{y_k^n}{\hat{y}_k^n} \frac{\partial \hat{y}_k^n}{\partial w_{kj}} + \sum_{k' \neq k} \frac{y_{k'}^n}{\hat{y}_{k'}^n} \frac{\partial \hat{y}_{k'}^n}{\partial w_{kj}} \right)$$

$$= -\left(\frac{y_k}{\hat{y}_k^n} \left(x_j^n \left(\hat{y}_k^n \cdot (1 - \hat{y}_k^n) \right) - \sum_{k' \neq k} \frac{y_{k'}}{\hat{y}_{k'}^n} x_j^n \hat{y}_{k'}^n \hat{y}_k^n \right) \right)$$

$$= \sum_{k' \neq k} y_{k'}^n x_j^n \hat{y}_k^n - x_j^n y_k^n (1 - \hat{y}_k^n)$$

$$= -x_j^n \left(y_k^n - \left(y_k^n \hat{y}_k^n + \sum_{k' \neq k} y_{k'}^n \hat{y}_k^n \right) \right)$$

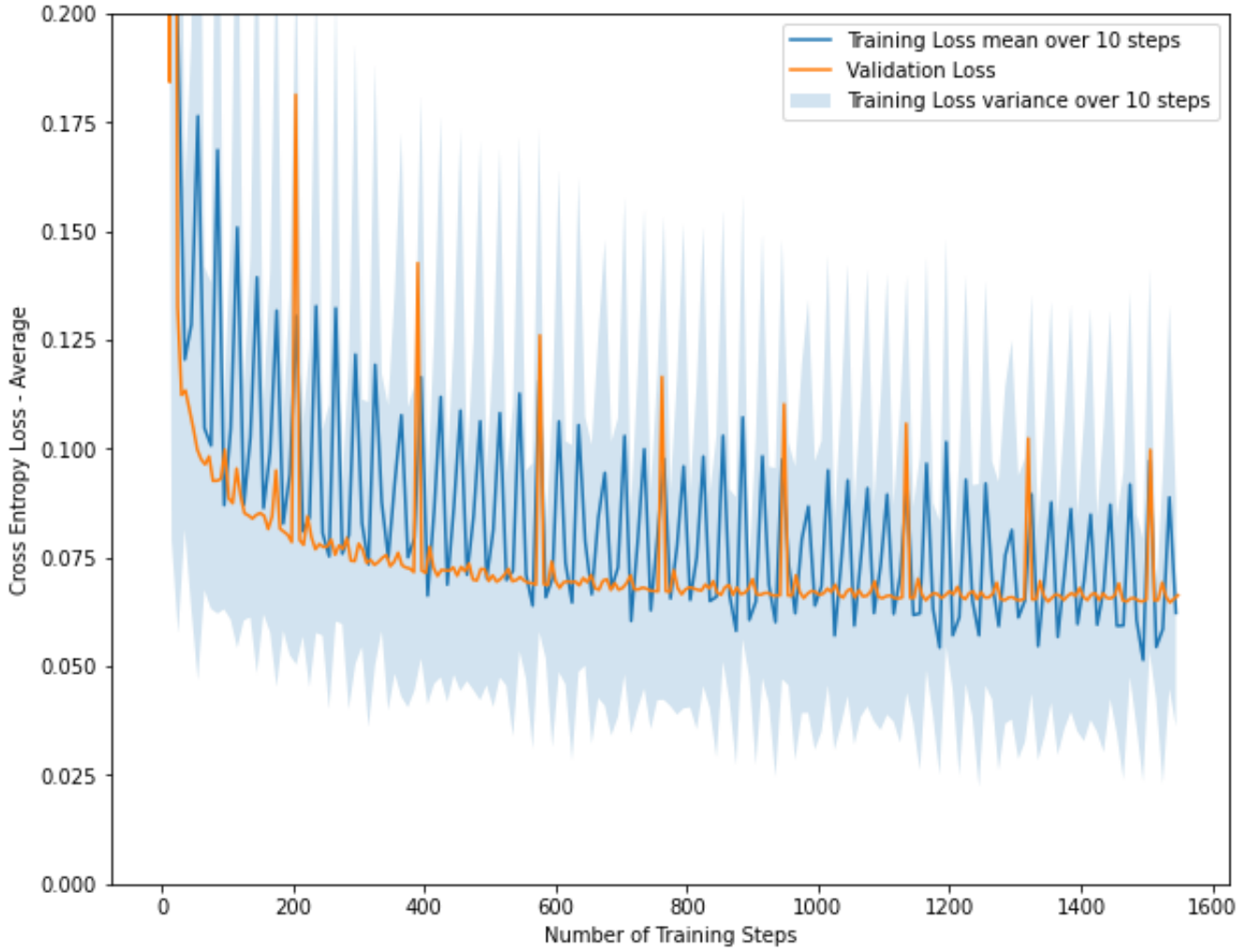
$$= -x_j^n \left(y_k^n - \sum_{k'=1}^{K} y_{k'}^n \hat{y}_k^n \right) = -x_j^n \left(y_k^n - \hat{y}_k^n \underbrace{\sum_{k'=1}^{K} y_{k'}^n}_{=1} \right)$$

$$= \underline{\underline{-x_j^n \left(y_k^n - \hat{y}_k^n \right)}} \quad \square$$

Task 2

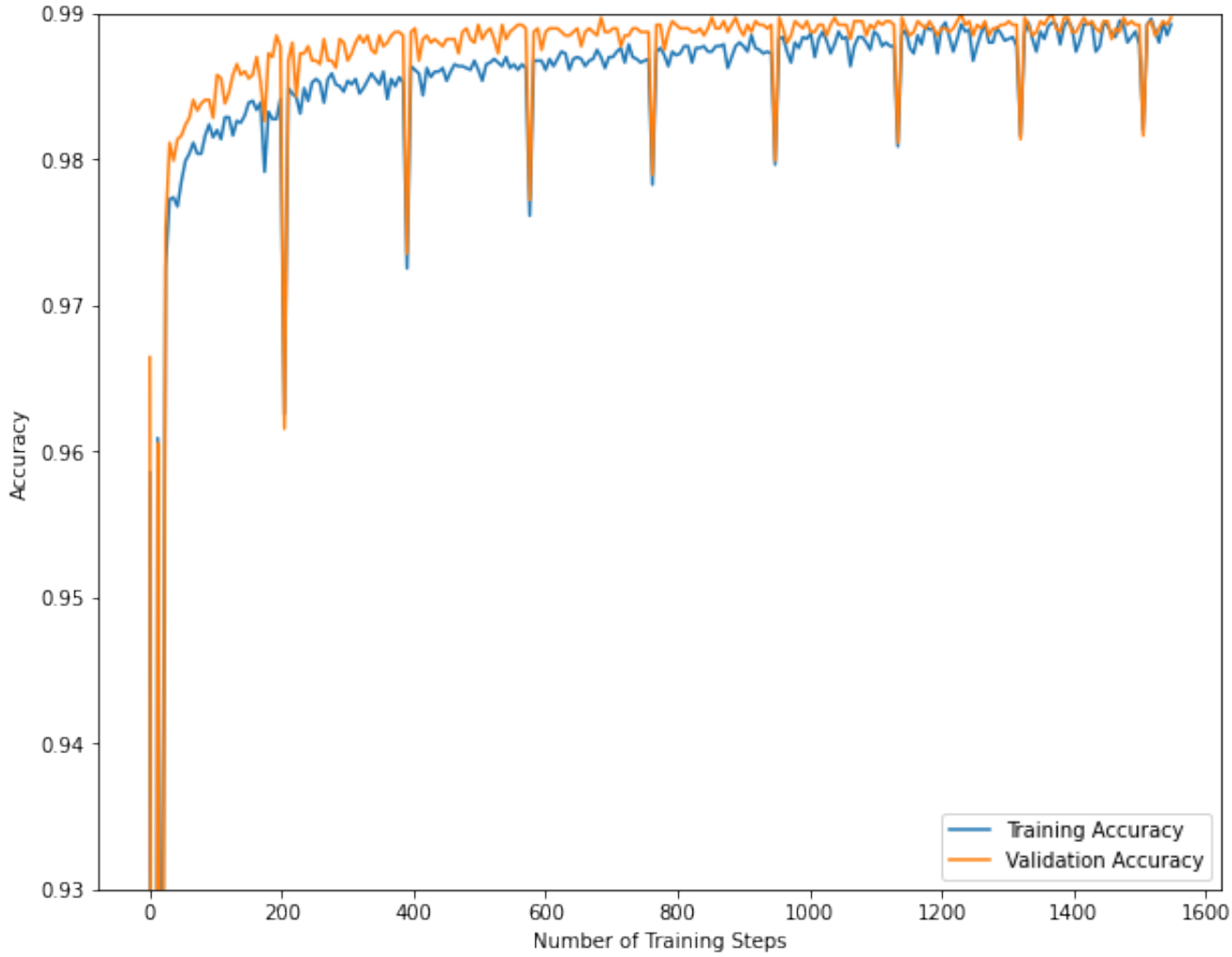
Task 2b)

```
Train shape: X: (4005, 784), Y: (4005, 1)
Validation shape: X: (2042, 784), Y: (2042, 1)
Final Train Cross Entropy Loss: 0.06509106268432195
Final Validation Cross Entropy Loss: 0.06522830574814967
Train accuracy: 0.9792759051186017
Validation accuracy: 0.9789422135161606
```



Task 2c)

Implemented calculate_accuracy. Resulting plot:



Task 2d)

With early stopping, we get:

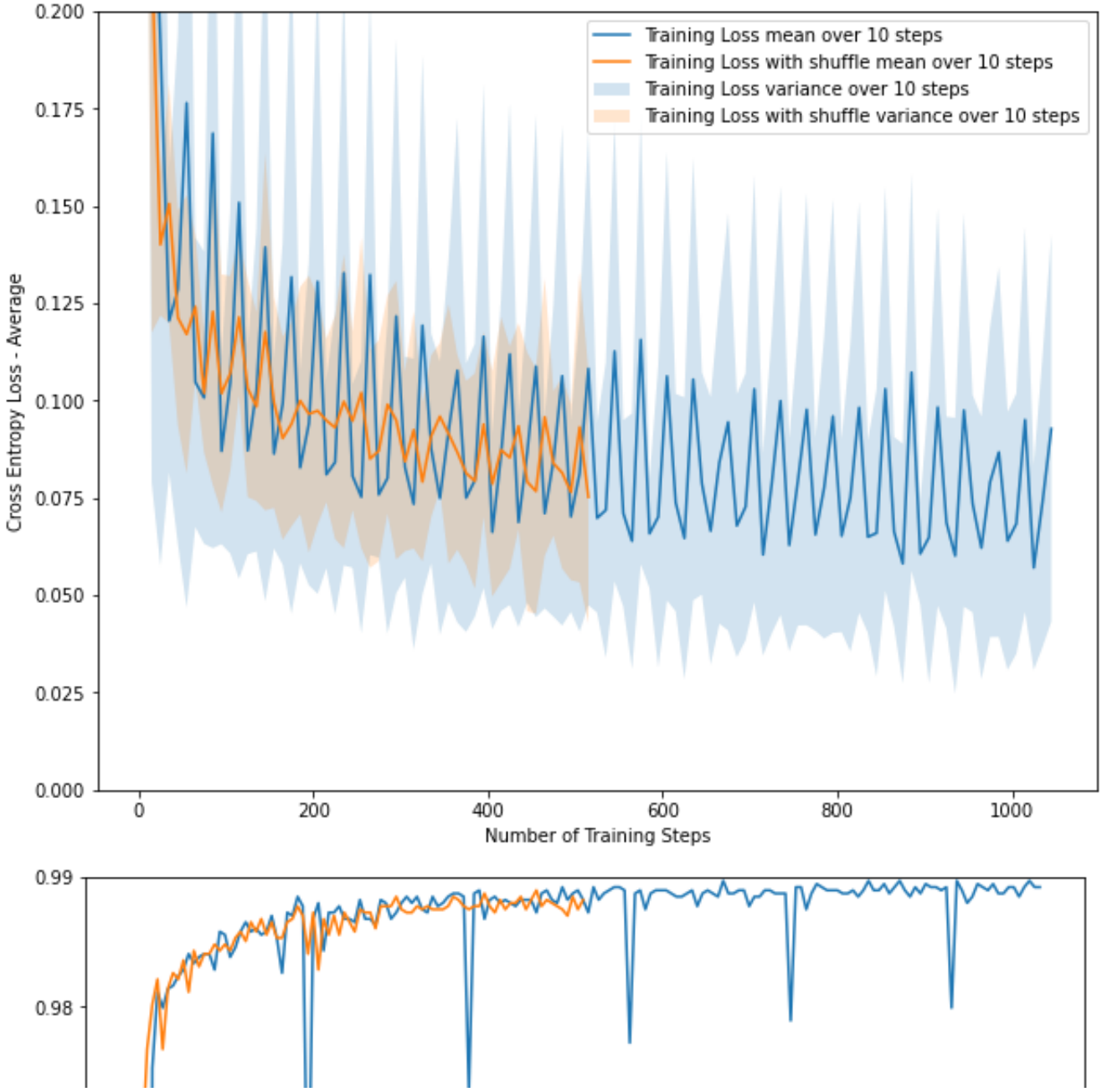
```
Train shape: X: (4005, 784), Y: (4005, 1)
Validation shape: X: (2042, 784), Y: (2042, 1)
Early stopping at epoch:  33
```

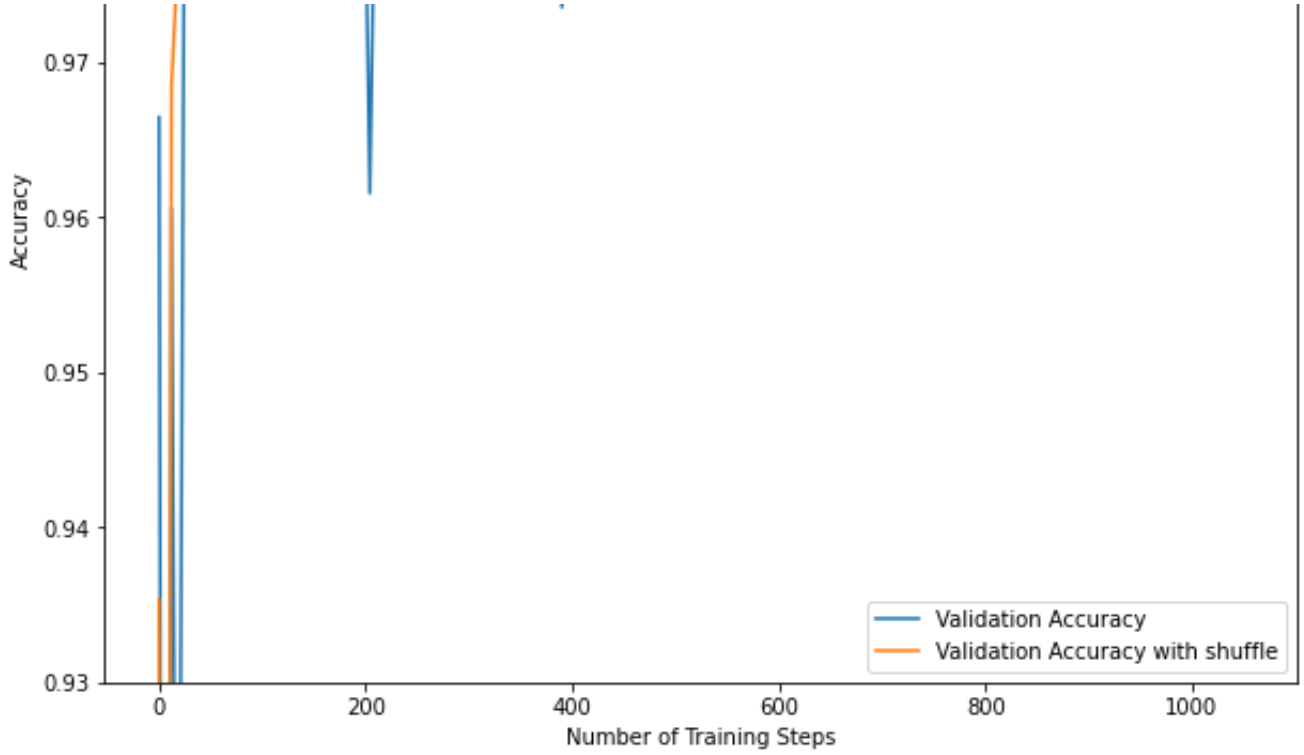
Result is, evidently: Eary stopping at epoch 33.

Task 2e)

Result:

```
Train shape: X: (4005, 784), Y: (4005, 1)
Validation shape: X: (2042, 784), Y: (2042, 1)
Early stopping at epoch:  33
Early stopping at epoch:  16
```





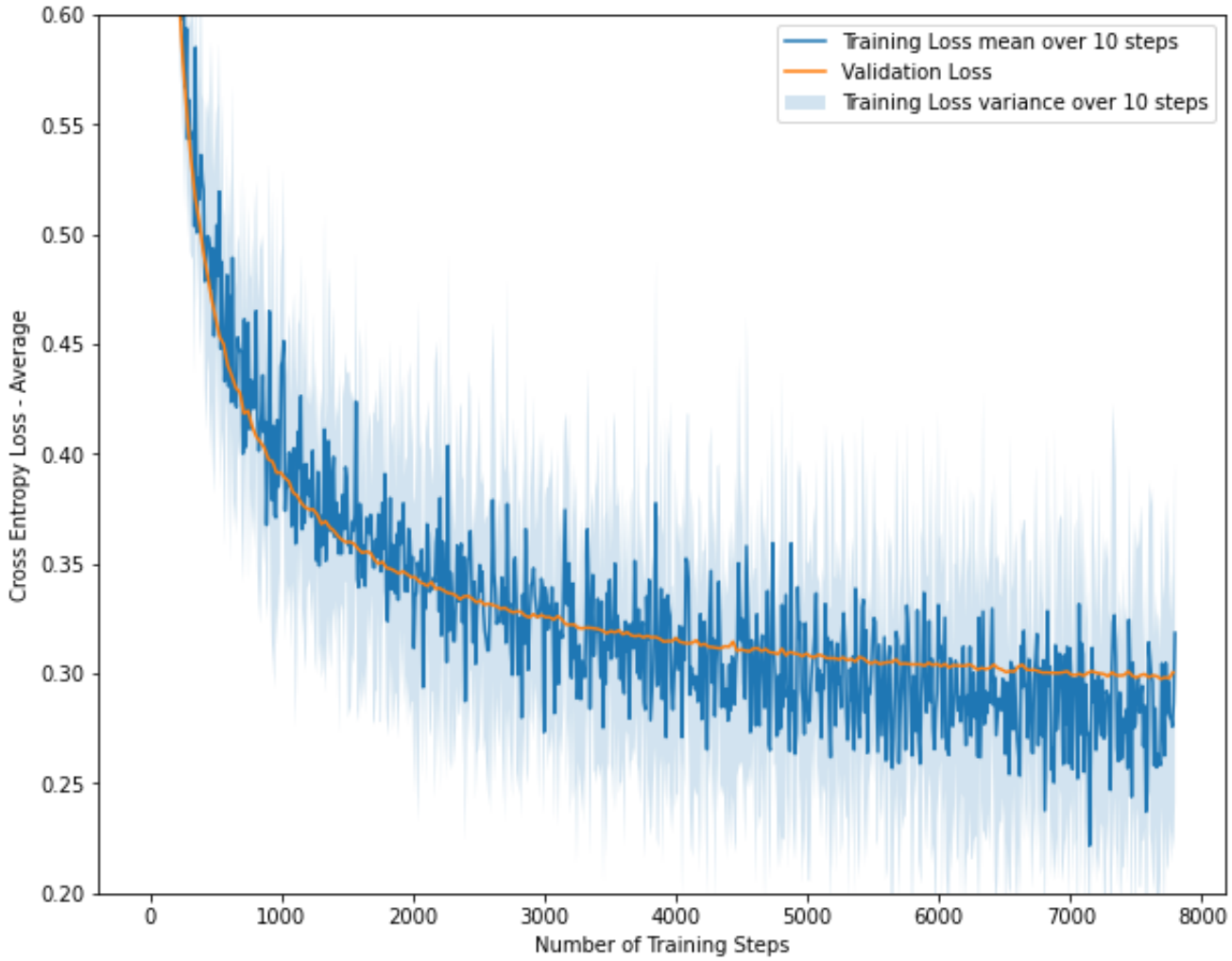
When using shuffle before every epoch, we get less "spikey behaviour". If we do not do this, at every epoch the network tries the updated weights on the same set as before, and these naturally fits badly. But, when using shuffling, it is a "new set" that the weights have no relation to, and this leads to the spikes disappearing as we get more generalization from the start. We also see that it leads to faster training, as the early stopping kicks in at 16 as opposed to 33.

Task 3

Task 3b)

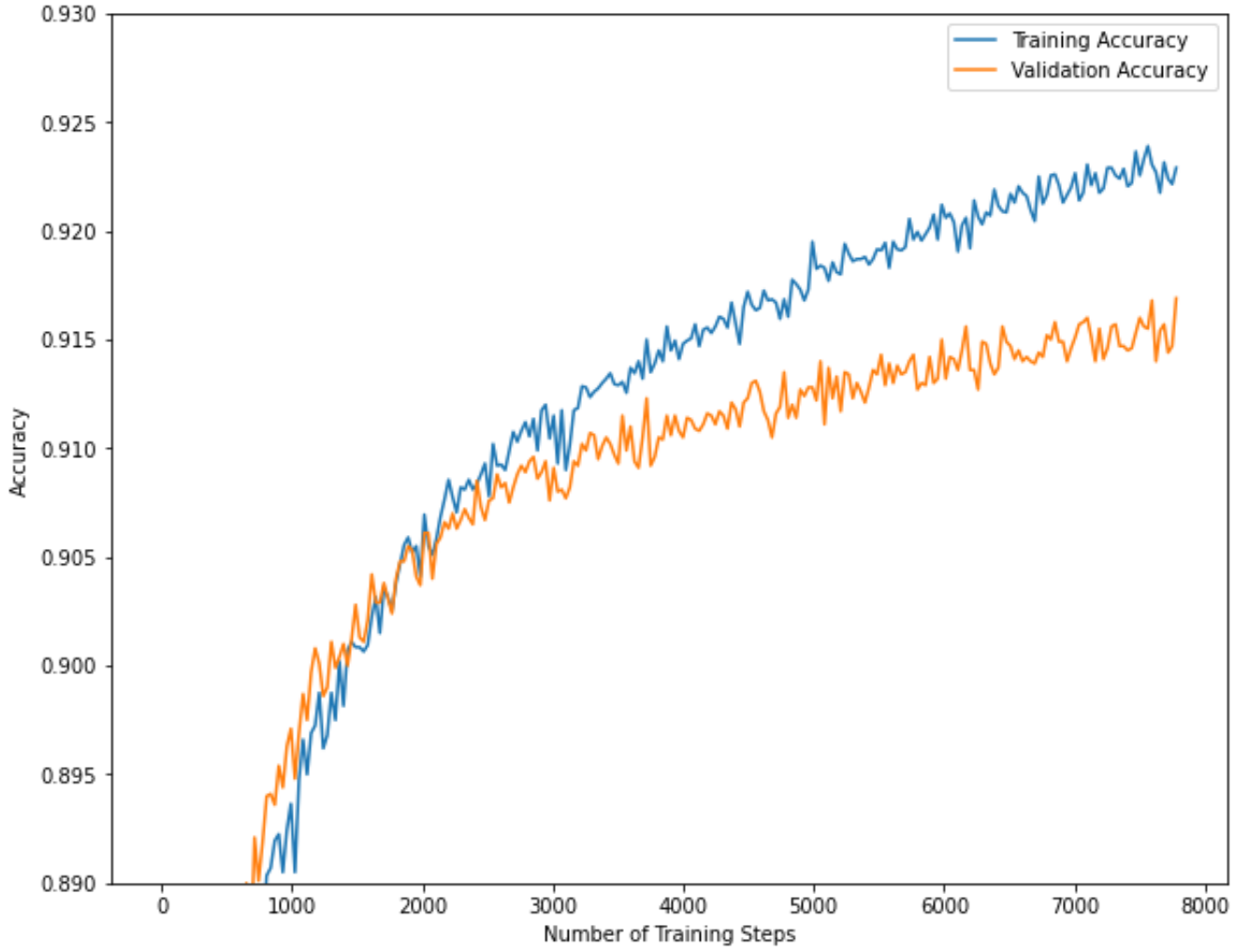
Result, Softmax

```
Final Train Cross Entropy Loss: 0.2800335875350492
Final Validation Cross Entropy Loss: 0.29902402393704086
Final Train accuracy: 0.9225
Final Validation accuracy: 0.9157
```



Task 3c)

Plot accuracy



Task 3d)

Yes, we see signs of overfitting from around training step 2000. From there, the validation accuracy *does increase*, hence it is not stopped in the early stopping-mechanism, but it increases a lot slower than the training accuracy which indicates that we are on our way to overfitting.

Task 4

Task 4a)

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

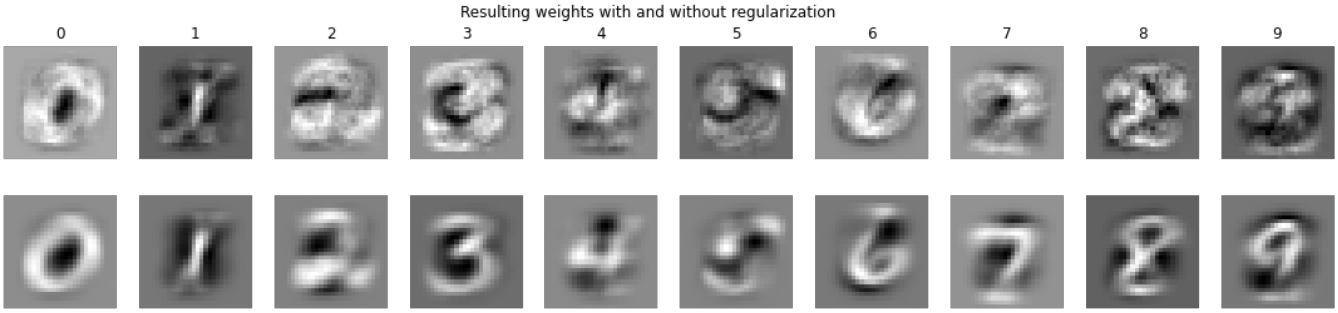
$$\frac{\partial R}{\partial w}(w) = \frac{\partial}{\partial w}\|w\|^2 = 2\lambda w$$

$$\Rightarrow \frac{\partial J}{\partial w}(w) = -x_j^i(y_k^i - \hat{y}_k^i) + 2\lambda w$$

Task 4b)

Plotting of softmax weights (Task 4b)

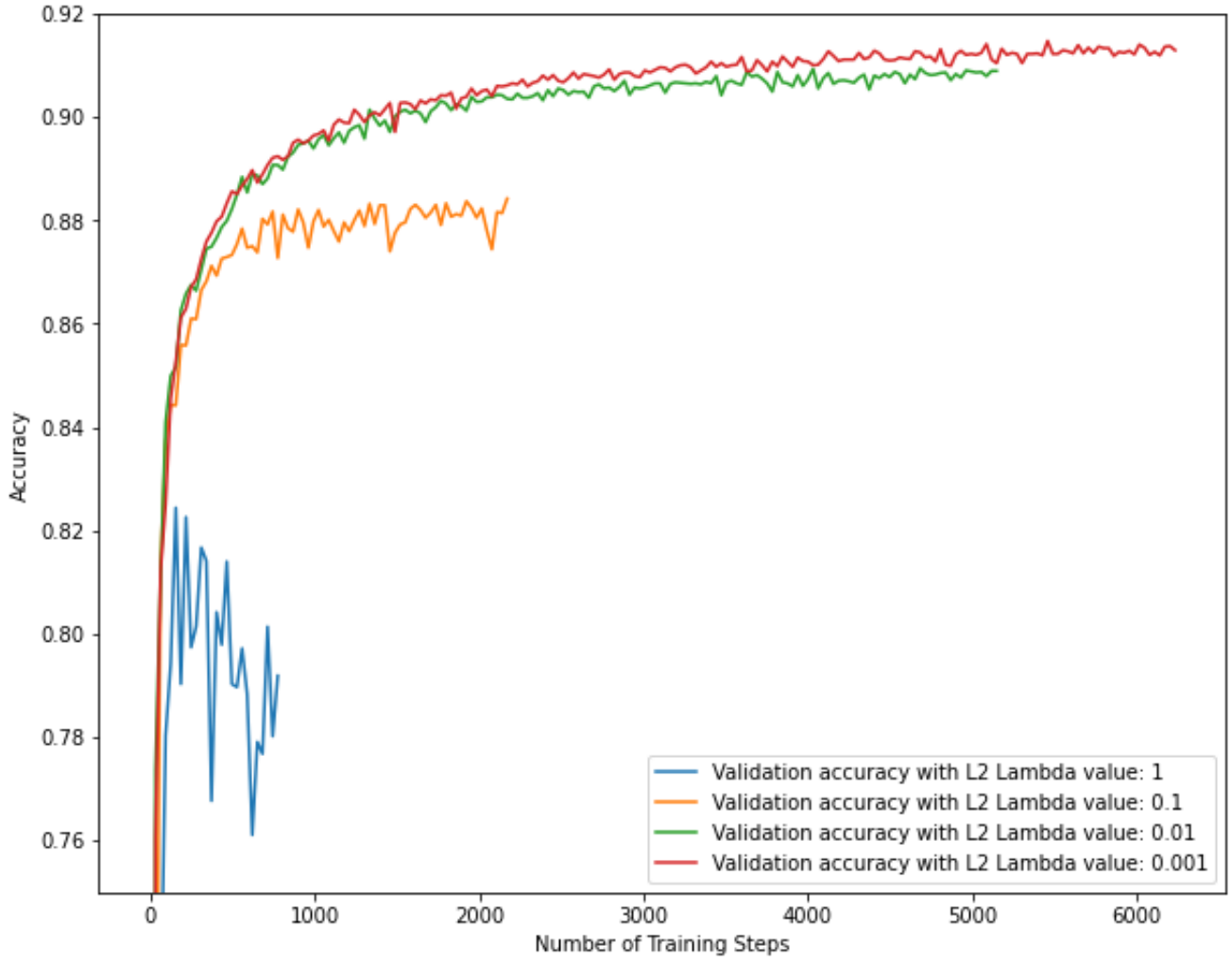
NB: the first row is the original without regularization. The second row is with regularization.



The weights for the model with lambda = 1 is more noisy. This is due to the fact that large values for the weights are penalized, and thus we do not see hard edges and very definite shapes.

Task 4c)

Validation accuracies for different values of lambda



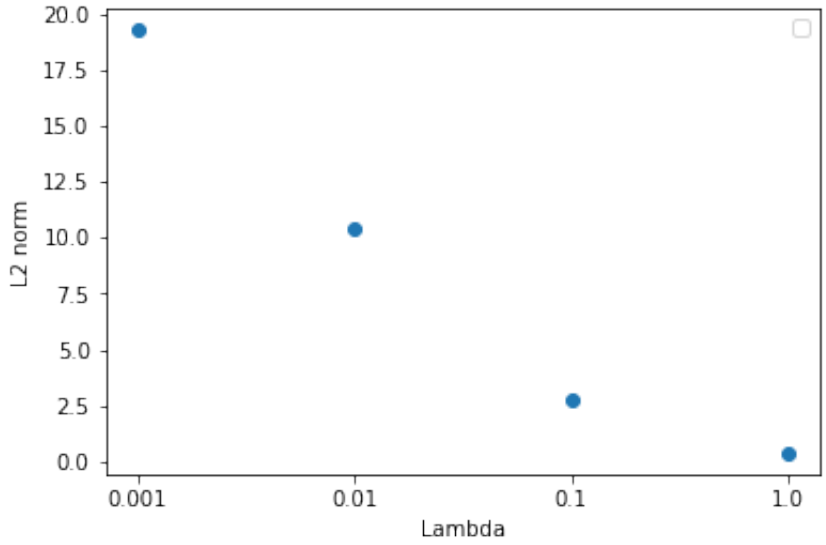
Task 4d)

The accuracy degrades when we apply regularization. This can be connected to task 4b), where we see that the weights are more noisy with lambda = 1. Thus, as we apply more regularization, the different predictions will get more blurry and, with a higher likelihood, blend in with eachother. In othe words, it becomes harder for the model to separate the digits.

Task 4e)

Plotting of the l2 norm for each weight

```
No handles with labels found to put in legend.
```



We observe, as expected, that the L2 norm of the weights decrease as we introduce more and more penalization of the size of the weights.