Task 1:

a)

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

$\hat{y}^n = f(z_k)$

$f(x) = \frac{1}{1+e^{-w^T x}}$ z_k $\qquad \hat{y} = \frac{1}{1+e^{-z_k}}$

$z_k = w^T x = \sum_i w_i x_i$

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

$\frac{\partial \hat{y}^n}{\partial z_k} = (1+e^{-z_k})^{-1\,\prime} = -(1+e^{-z_k})^{-2} \cdot \frac{\partial}{\partial z_k}(1+e^{-z_k}) = -(1+e^{-z_k})^{-2}(-e^{-x}) = (1+e^{-z_k})^{-2} e^{-x} = \frac{e^{-x}}{(1+e^{-z_k})^2} = \frac{e^{-x}}{(1+e^{-z_k})(1+e^{-z_k})}$

$= \frac{1}{(1+e^{-z_k})} \cdot \frac{e^{-x}}{(1+e^{-z_k})} = \frac{1}{(1+e^{-z_k})} \cdot \frac{e^{-x}+1-1}{(1+e^{-z_k})} = \frac{1}{(1+e^{-z_k})} \cdot \left(1 - \frac{1}{(1+e^{-z_k})}\right)$

$\frac{\partial z_k}{\partial w_i} = x_i$

Gradient: $\frac{\partial C^n(w)}{\partial (w)} = \frac{\partial C^n(w)}{\partial \hat{y}^n} \cdot \frac{\partial \hat{y}^n}{\partial z_k} \cdot \frac{\partial z_k}{\partial w_i} = \left(-\frac{y^n}{\hat{y}^n} + \frac{(1-y^n)}{(1-\hat{y}^n)}\right) \cdot \left(\frac{1}{1+e^{-z_k}} \cdot \left(1 - \frac{1}{(1+e^{-z_k})}\right)\right) \cdot x_i$

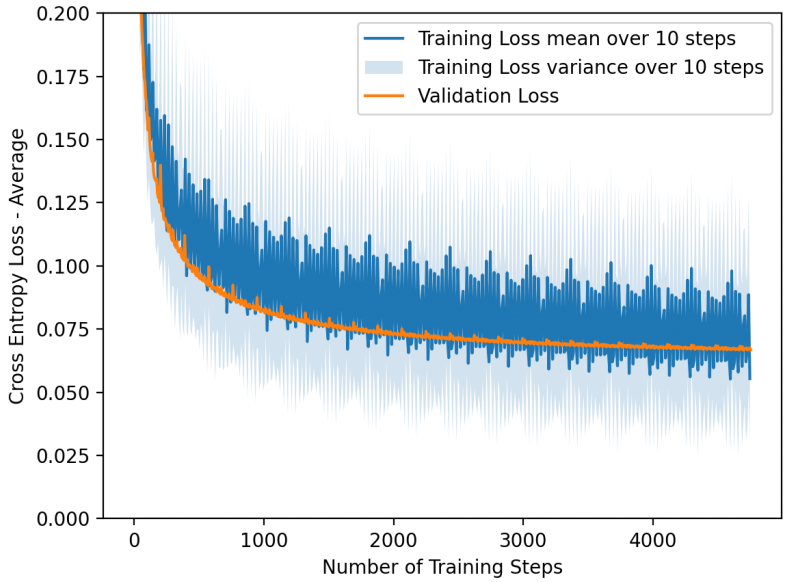
$= \left(-\frac{y^n}{\hat{y}^n} + \frac{(1-y^n)}{(1-\hat{y}^n)}\right) \cdot (\hat{y}^n \cdot (1-\hat{y}^n)) \cdot x_i =$

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

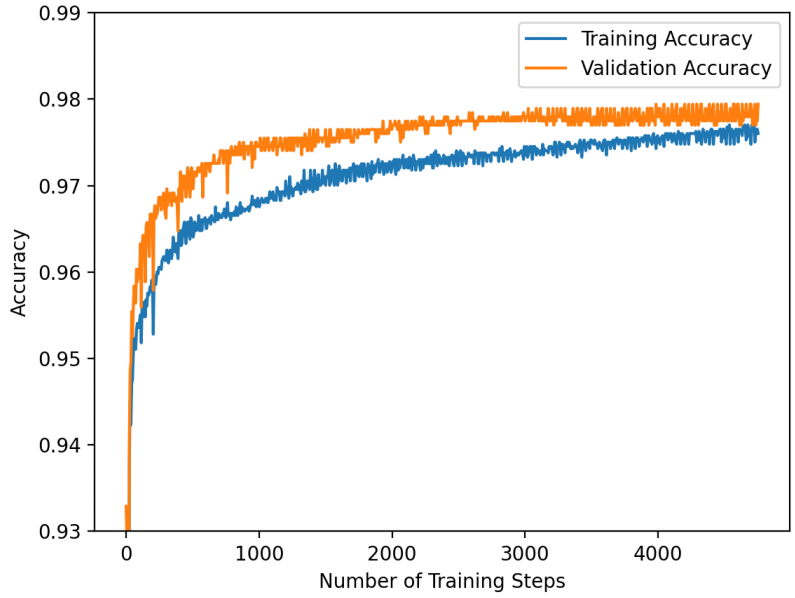
$= \left(-(y^n - y^n\hat{y}^n) + \hat{y}^n - y^n\hat{y}^n\right) x_i = \underline{\underline{(\hat{y}^n - y^n)\, x_i}}$

Task 2:

b)



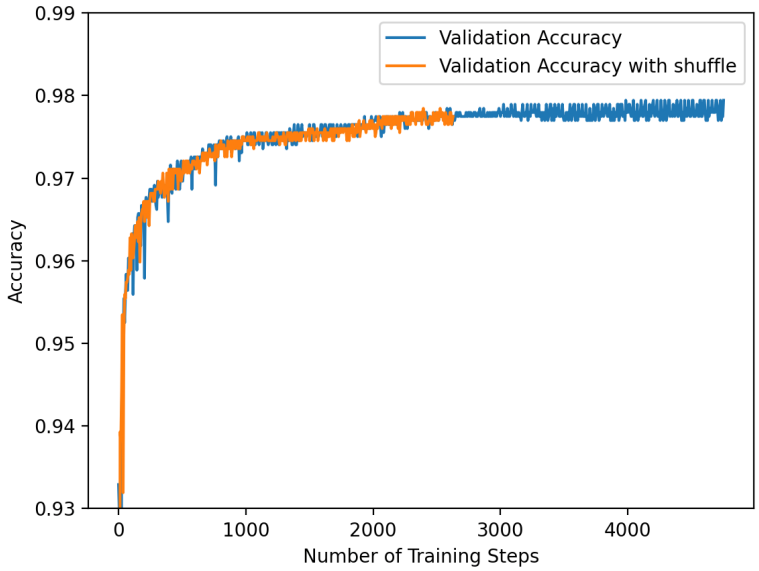
c)



d) The early stopping kicks in after 153 epochs

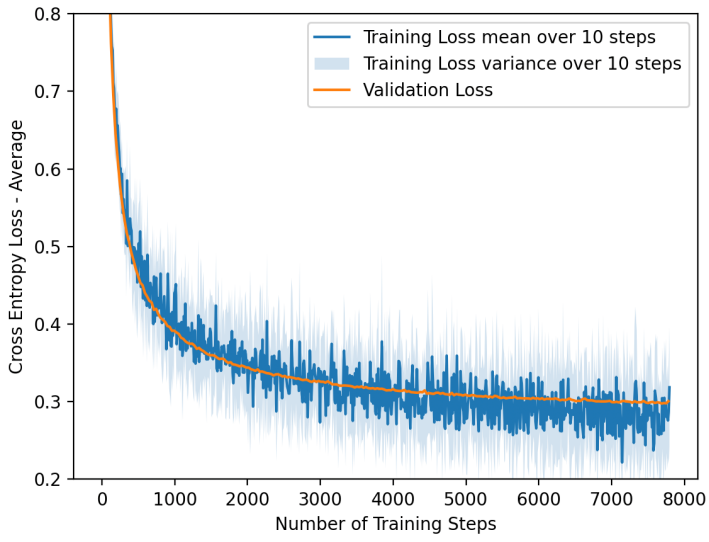
```
Train shape: X: (4005, 784), Y: (4005, 1)
Validation shape: X: (2042, 784), Y: (2042, 1)
Early stopping at epoch 153
Final Train Cross Entropy Loss: [0.07122371]
Final Validation Cross Entropy Loss: [0.06669121]
Train accuracy: 0.9760299625468165
Validation accuracy: 0.9794319294809011
```

e)


Validation Accuracy
Validation Accuracy with shuffle
Accuracy
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0
1000
2000
3000
4000
Number of Training Steps

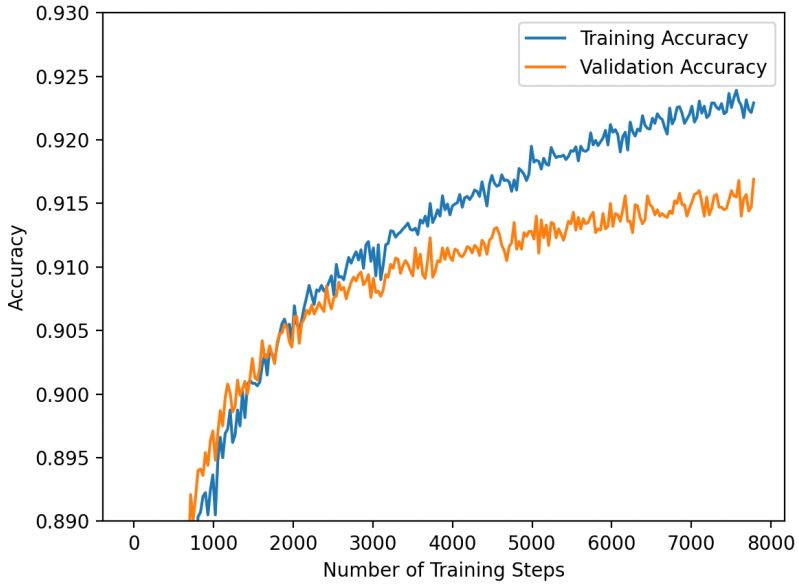
Task 3:

b)


Training Loss mean over 10 steps
Training Loss variance over 10 steps
Validation Loss
Cross Entropy Loss - Average
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
1000
2000
3000
4000
5000
6000
7000
8000
Number of Training Steps

---

c)



d)

We can see that the model starts to overfit in 3c because the validation accuracy stops increasing at the same rate as the training accuracy. The model likely overfits on the traning data, so that when it must predict on unseen data (validation data) it expects the distribution of the data to look like the training data distribution (which it doesn't with some deviation).

Task 4

a)

a) $C(w) = \frac{1}{N}\sum^{N} C^n(w)$, $C^n(w) = -\sum y_k^n \ln(\hat{y}_k^n)$

$\hat{y}_k^n = \frac{e^{z_k}}{\sum^{k} e^{z_k}}$, $z_k = w_k^T \cdot x$

$J(w) = C(w) + \lambda R(w)$, $R(w) = \|w\|^2$

$\frac{dJ}{dw} = \underbrace{\frac{dC}{d\hat{y}_k} \cdot \frac{d\hat{y}_k}{dz_k} \cdot \frac{dz_k}{dw}} + \frac{d\lambda R(w)}{dw} = -x_j^n(y_k^n - \hat{y}_k^n) + \lambda 2w$

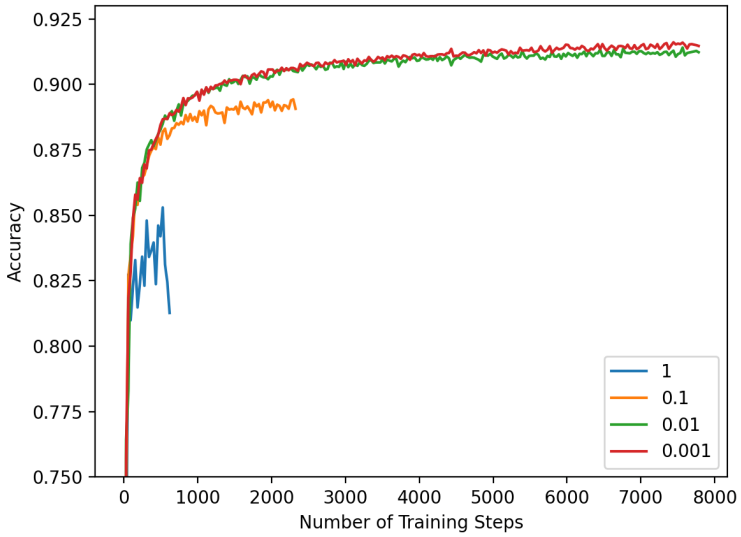
$\frac{dC}{dw} = -x_j^n(y_k^n - \hat{y}_k^n)$

b)
Weights with lambda=0.0:

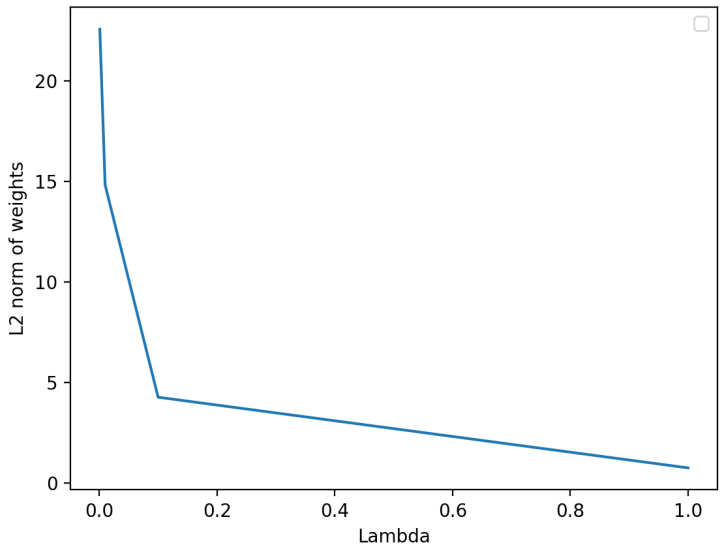
Weights with lambda=1.0:

The weights for the model with lambda=1.0 are less noisy because the regularization penalizes the model for learning the patterns for the training data to precisely.

c)


0.925
0.900
0.875
0.850
0.825
0.800
0.775
0.750
Accuracy
0
1000
2000
3000
4000
5000
6000
7000
8000
Number of Training Steps
1
0.1
0.01
0.001

d)


20
15
10
5
0
L2 norm of weights
0.0
0.2
0.4
0.6
0.8
1.0
Lambda