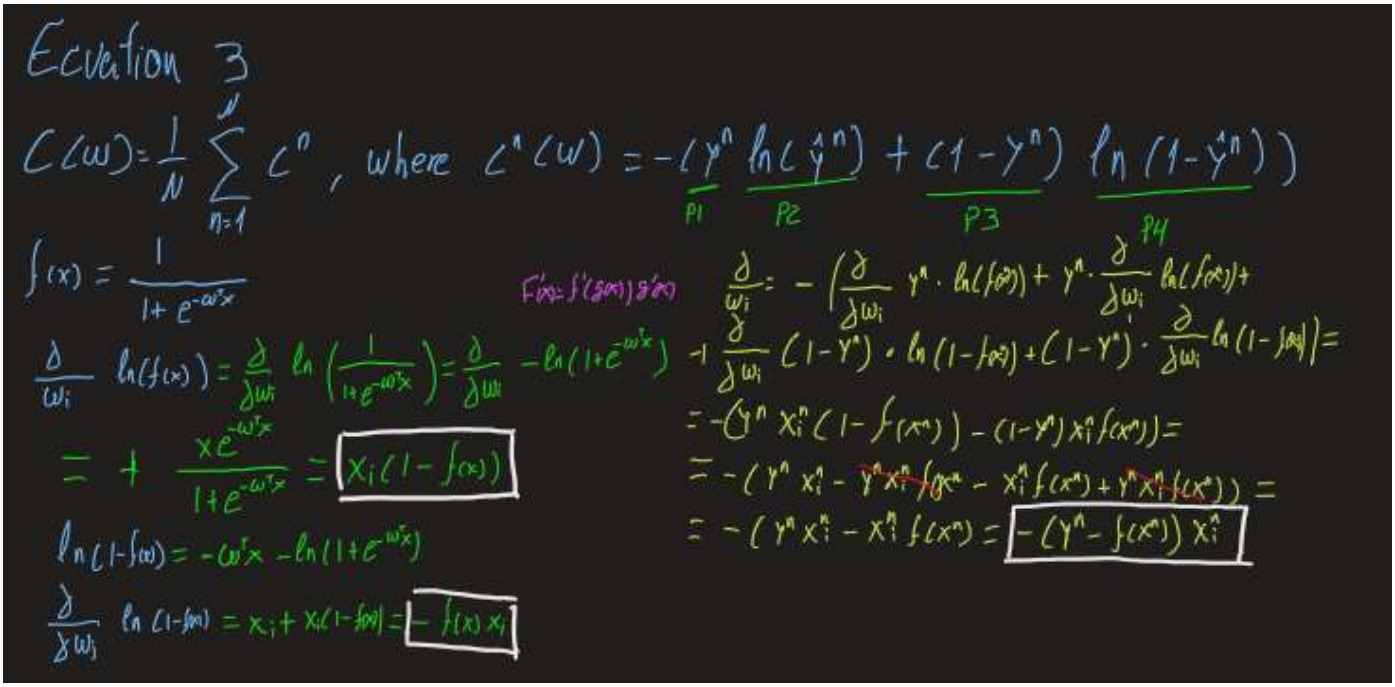
Assignment 1 Report

Authors:

- Alberto Robazza
- Ivan Garcia Alcaraz

Task 1

task 1a)



task 1b)

Calculous to makes the derivative more easily.

Ecuation 5

$$\hat{y}_k = \frac{e^{z_k}}{\sum_{k'}^{K} e^{z_{k'}}}, \text{ where } z_k = w_k^T \cdot x = \sum_{i}^{I} w_{k,i} \cdot x_i$$

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

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

$$= -\sum_{k=1}^{K} y_k^n z_k^n + \sum_{k=1}^{K} y_k^n \left(\ln\left(\sum_{k'}^{K} e^{z_{k'}^n}\right)\right) = \underbrace{-\sum_{k=1}^{K} y_k^n z_k^n}_{-g^n(w)} + \underbrace{\ln\left(\sum_{k'}^{K} e^{z_{k'}^n}\right)}_{+h^n(w)}$$

The derivative of a function g(w) is the following:

$$\frac{\partial}{\partial w_{kj}} g(w) = \frac{\partial}{\partial w_{kj}} \sum_{k=1}^{K} y_k^n z_k^n = \frac{\partial}{\partial w_{kj}} \sum_{k=1}^{K} y_k^n \sum_{i=1}^{I} w_{ki} \cdot x_i$$

The derivative of the previous formula takes the fact that only when $j=i$ the derivate of w_{ik} is non-zero so:

$$\frac{\partial}{\partial w_{kj}} g(w) = \boxed{y_k \cdot x_j}$$

The derivative of a function h(w) is the following:

$$\frac{\partial}{\partial w_{kj}} h(w) = \frac{\partial}{\partial w_{kj}} \ln\left(\sum_{k'}^{K} e^{z_{k'}}\right) = \frac{\partial}{\partial w_{kj}} \ln\left(\sum_{k'}^{K} e^{\sum_i w_{k'i} \cdot x_i}\right) = \frac{1}{\sum_{k'}^{K} e^{\sum_i w_{k'i} \cdot x_i}} \frac{\partial}{\partial w_{kj}} \sum_{k'}^{K} e^{\sum_i w_{k'i} \cdot x_i}$$

The derivative of the previous formula takes the fact that only when $j=i$ and $k=k'$ the derivate of w_{ik} is non-zero so:

$$\frac{\partial}{\partial w_{kj}} \sum_{k'}^{K} e^{\sum_i w_{k'i} x_i} = \boxed{x_j e^{\sum_i w_{ki} x_i}}$$

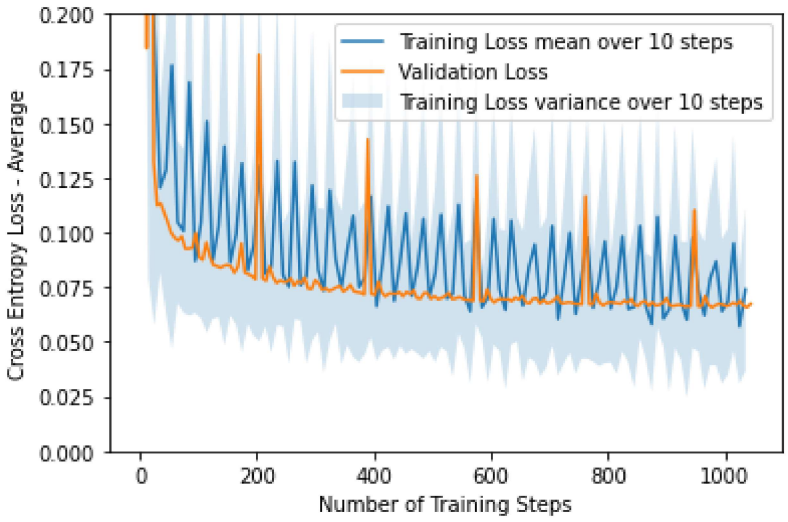
Now we merge the derivative of g(w) with the derivative of h(w) and we get the following:

$$\frac{1}{\sum_{k'}^{K} e^{\sum_i w_{k'i} \cdot x_i}} \frac{\partial}{\partial w_{kj}} \sum_{k'}^{K} e^{\sum_i w_{k'i} \cdot x_i} = \frac{e^{z_k}}{\sum_{k'}^{K} e^{\sum_i w_{k'i} \cdot x_i}} x_j = \hat{y}_k x_j$$

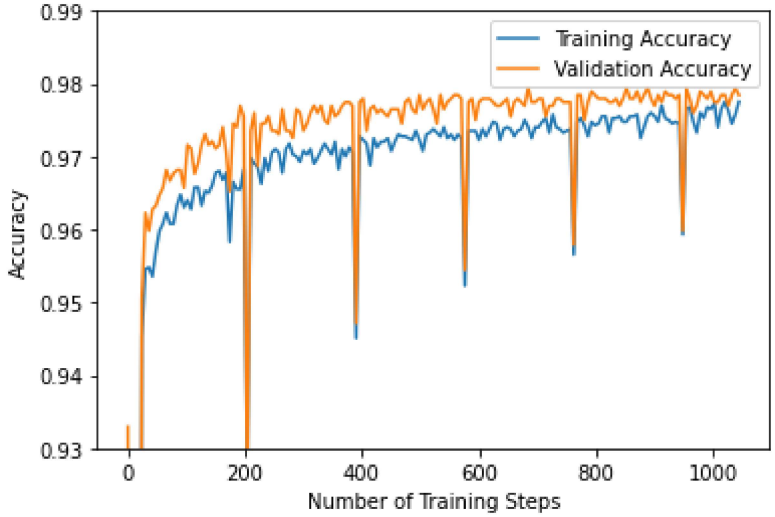
$$\frac{\partial}{\partial w_{kj}} C^n(w) = \frac{\partial}{\partial w_{kj}} g^n(w) + \frac{\partial}{\partial w_{kj}} h^n(w) = -y_k^n x_j^n + \hat{y}_k^n x_j^n = \boxed{-x_j^n (y_k^n - \hat{y}_k^n)}$$

Task 2

Task 2b)



Task 2c)

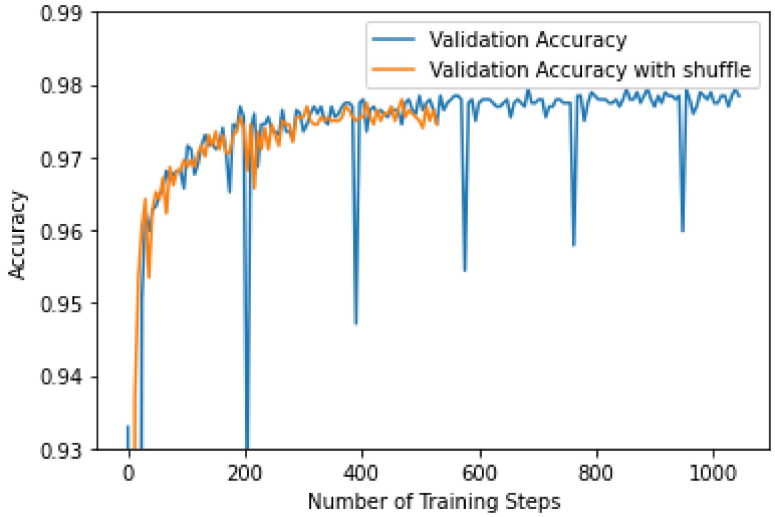


Task 2d)

In our test, we were bumbed out after 17 epochs

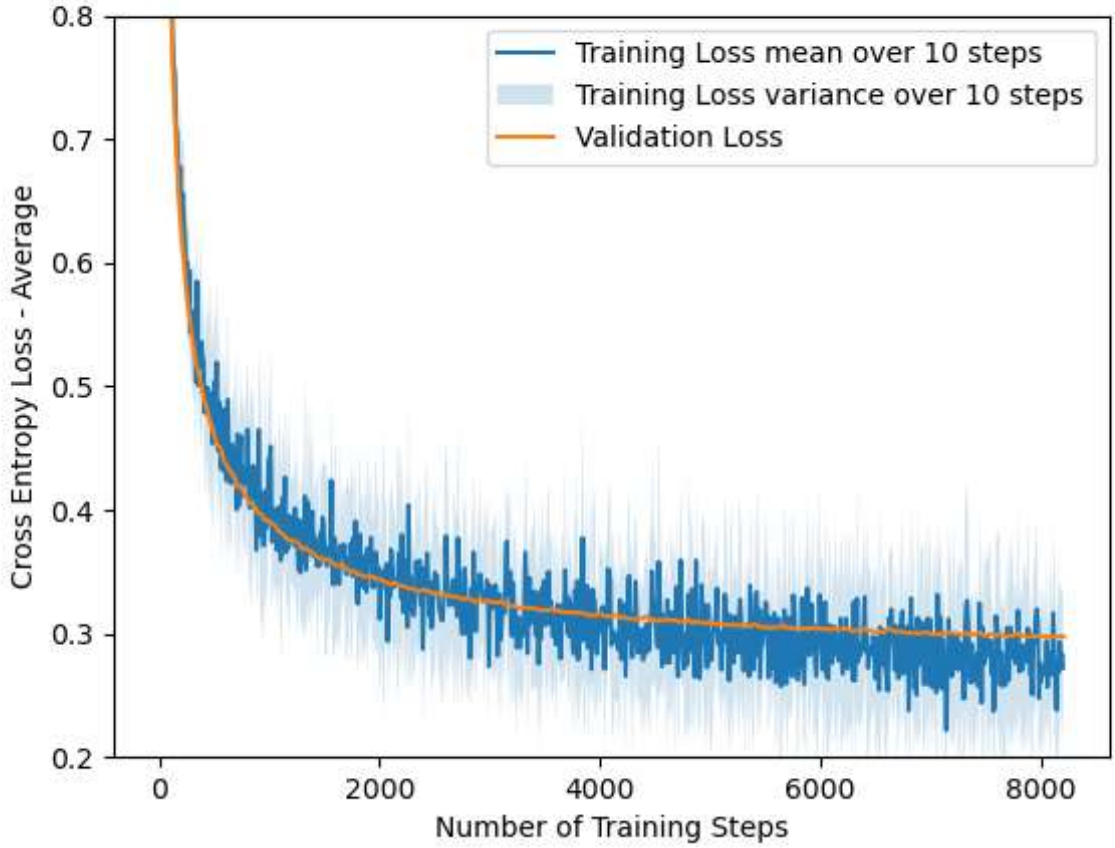
Task 2e)

Input is more stable, the model is trained on different pictures compared to the immediatly validatd ones

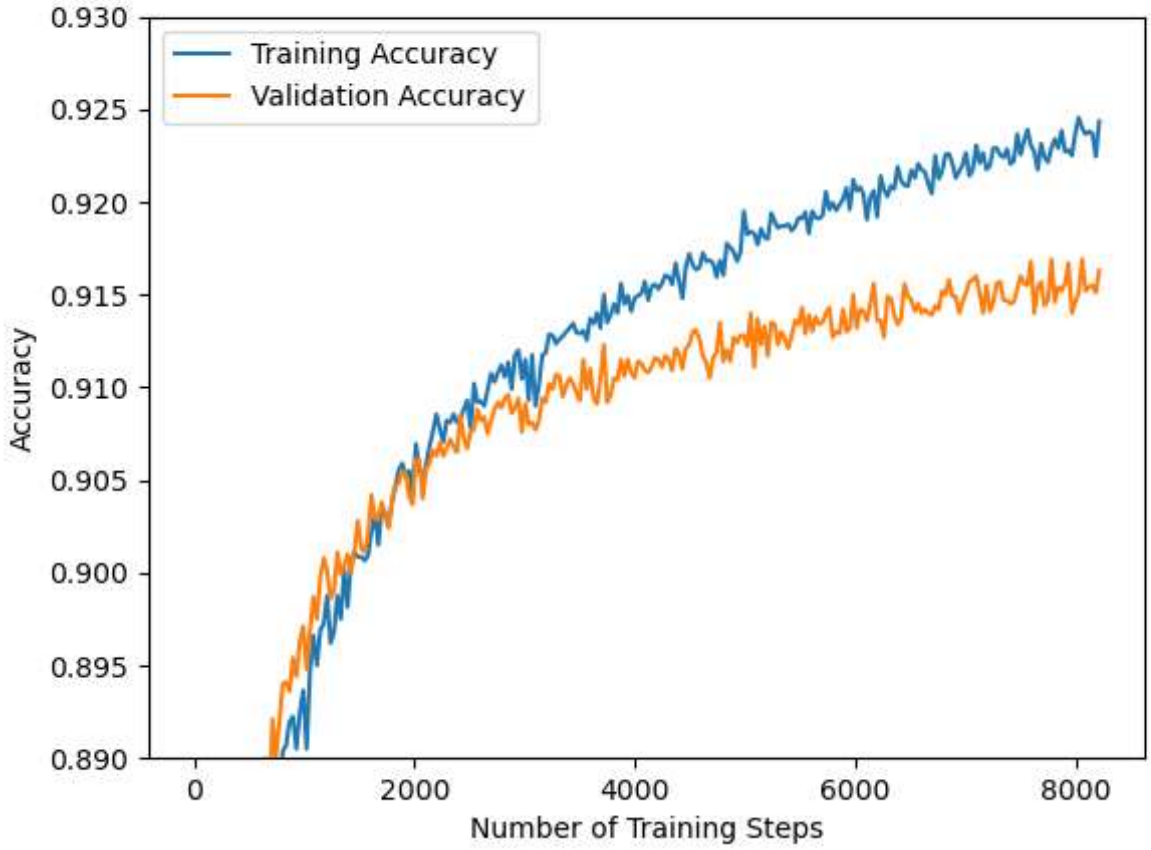


Task 3

Task 3b)



Task 3c)



Task 3d)

In Figure 6 the training accuracy is better than the validation accuracy after 3000 training steps. And the increase distance between the training accuracy and the validation accuracy could be a sign of overfitting, because the model perfomrs better in one than the other.

Task 4

Task 4a)

Task 4a:

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

The $\frac{\partial C}{\partial w}$ was done in Task 1

The $\frac{\partial \|w\|^2}{\partial w} = \frac{\partial}{\partial w} \sum_{i,j} w_{i,j}^2$

In the matrix generated by $w_{i,j}$ only the terms where $i' = i$ or $j' = j$ are the ones were the derivate is non-zero

So

$$\frac{\partial}{\partial w} = \boxed{2w}$$

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

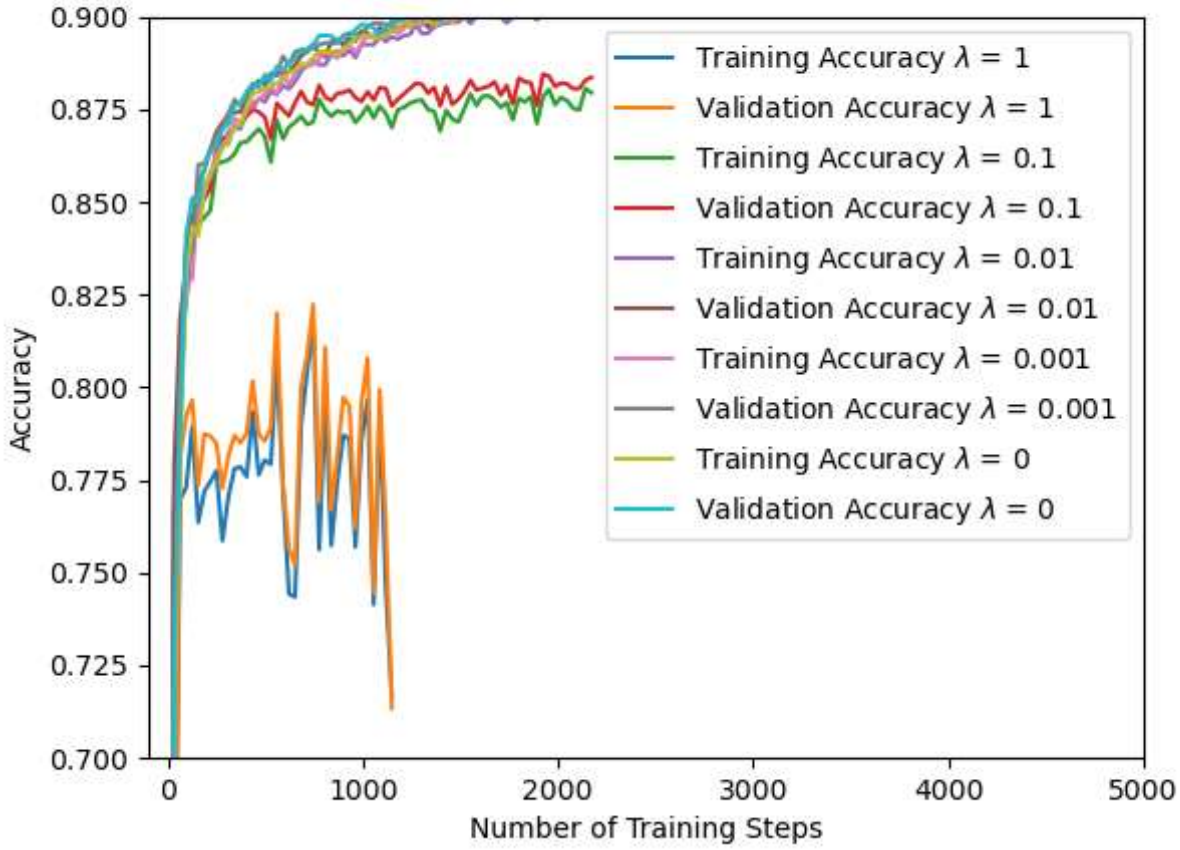
Task 4b)

FILL IN ANSWER

With stronger L2 normalization (high values for lambda), the model tries to be simpler, the difference within different values would be lower, than the noise is reduced

Task 4c)

FILL IN ANSWER

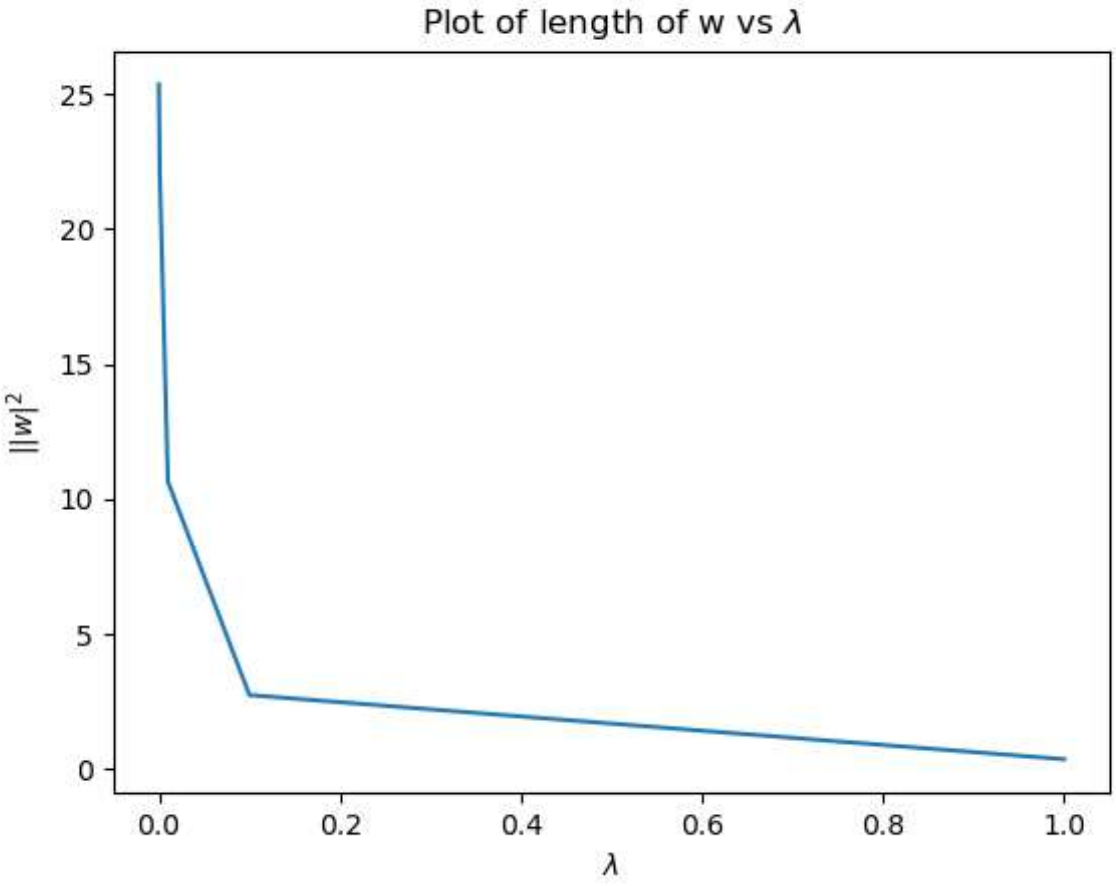


Task 4d)

Since L2 normalization penalize high values, the model tries to reduce them. It brings to a worse fitting model since it wouldn't align to the points

Task 4e)

On increasing lambda, l2 value lowers down. This is probabilly caused by the penalization of L2 on higher lambda values


Plot of length of w vs λ
25
20
15
10
5
0
||w||²
0.0
0.2
0.4
0.6
0.8
1.0
λ