AssignmentReport

February 9, 2024

1 Assignment 1 Report - Group 175

1.1 Group members:

Alfred Indrehus

Andreas Lothe Måkestad

2 Task 1

2.1 task 1a)

Task 1 a)

show that: $\frac{\partial C^n(w)}{\partial w_i} = -(y^n - \hat{y}^n) x_i^n$

where: $C(w) = \frac{1}{N} \sum_{n=1}^{N} -\left(y^n \ln(\hat{y}^n) + (1-y^n) \cdot \ln(1-\hat{y})\right)$

$\hat{y}^n = \sigma(w_i x_i^n + b_i) = \sigma(z) = \frac{1}{1+e^{-z}}$ $\quad z_i = w_i \cdot x_i^n + b$

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

$$\frac{\partial C^n}{\partial w_i} = \frac{\partial C^n}{\partial \hat{y}^n} \cdot \frac{\partial \hat{y}^n}{\partial w_i} = \frac{\partial C^n}{\partial \hat{y}^n} \cdot \frac{\partial \hat{y}^n}{\partial z} \cdot \frac{\partial z}{\partial w_i}$$

① ② ③

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

$$\frac{\partial \hat{y}^n}{\partial z_i} = \frac{\partial}{\partial z_i}\left(1+e^{-z_i}\right)^{-1} \Rightarrow \quad \frac{\partial y_0}{\partial z_i} = \frac{(e^{-z_i})}{(1+e^{-z_i})^2}$$

$$\frac{d\hat{y}_n}{dz_i} = \frac{(e^{-z_i})}{(1+e^{-z_i})^2}$$

$$\hat{y}_n^2 = \frac{1}{(1+e^{-z})^2} \Rightarrow \boxed{\frac{d\hat{y}_n}{dz_i} = \hat{y}_n(1-\hat{y}_n)} \quad (2)$$

$$e^{-z} = (1-\hat{y}_n)/\hat{y}_n$$

$$\frac{dz_i}{dw_i} = \frac{d}{dw_i}(w_i x_i + b_i) \Rightarrow \boxed{\frac{dz_i}{dw_i} = x_i} \quad (3)$$

$$\frac{dC^n}{dw_i} = ① \cdot ② \cdot ③$$

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

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

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

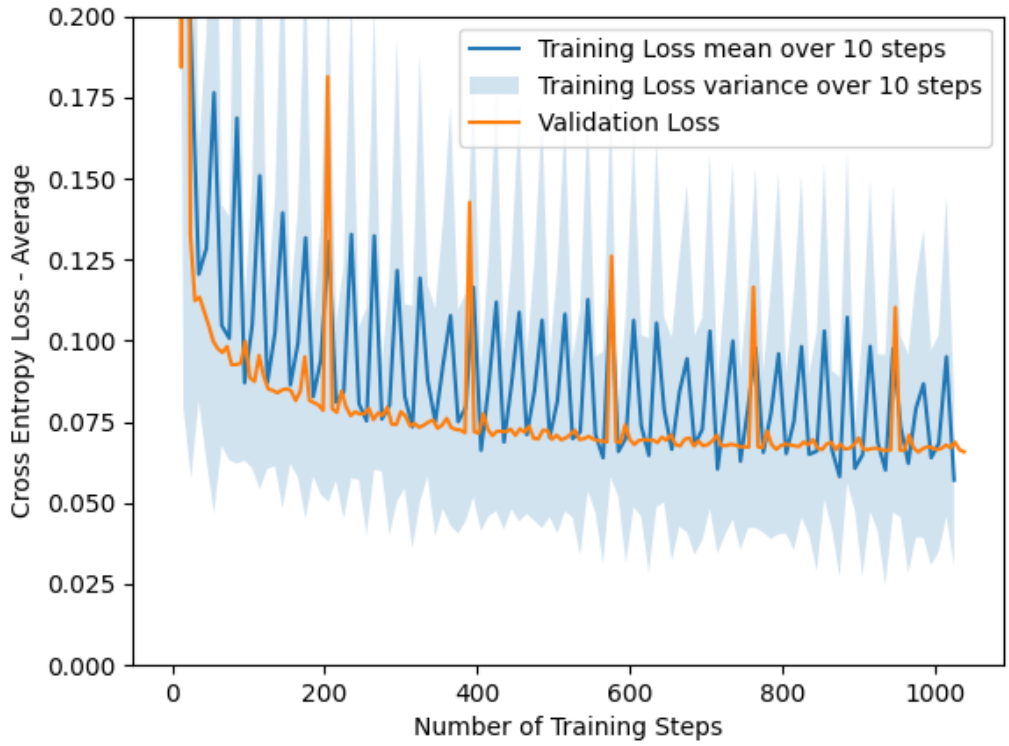
$$\frac{dC^n}{dw_i} = -\left(y^n - \hat{y}^n \right) x_i^n$$

Q.E.D

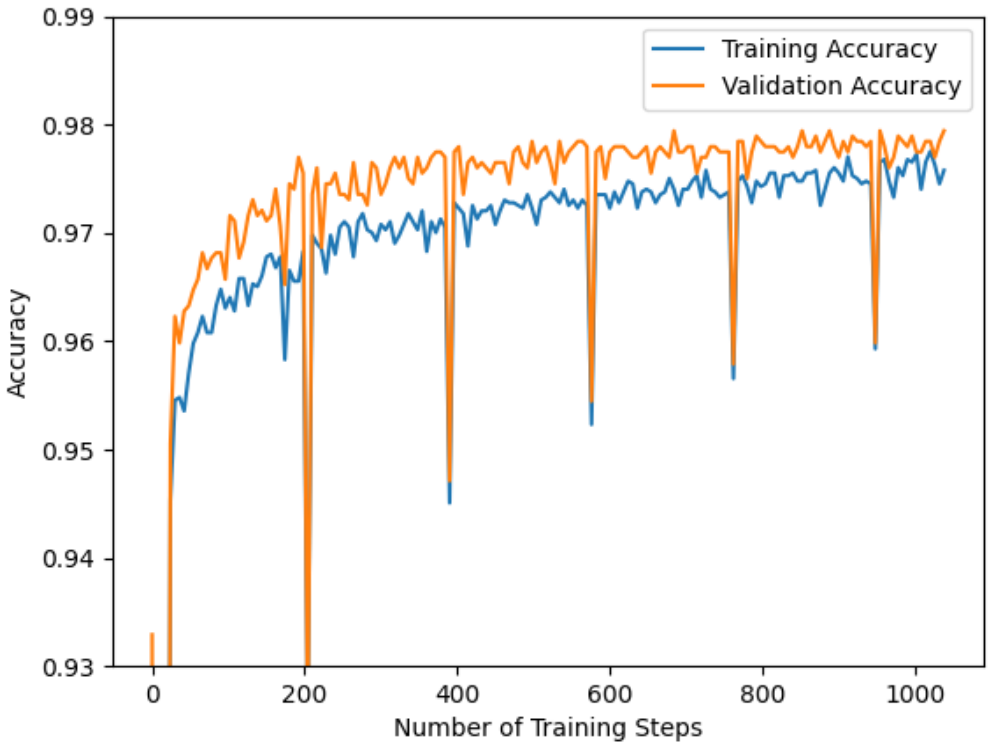
2.2 task 1a)

3 Task 2

3.1 Task 2b)



3.2 Task 2c)

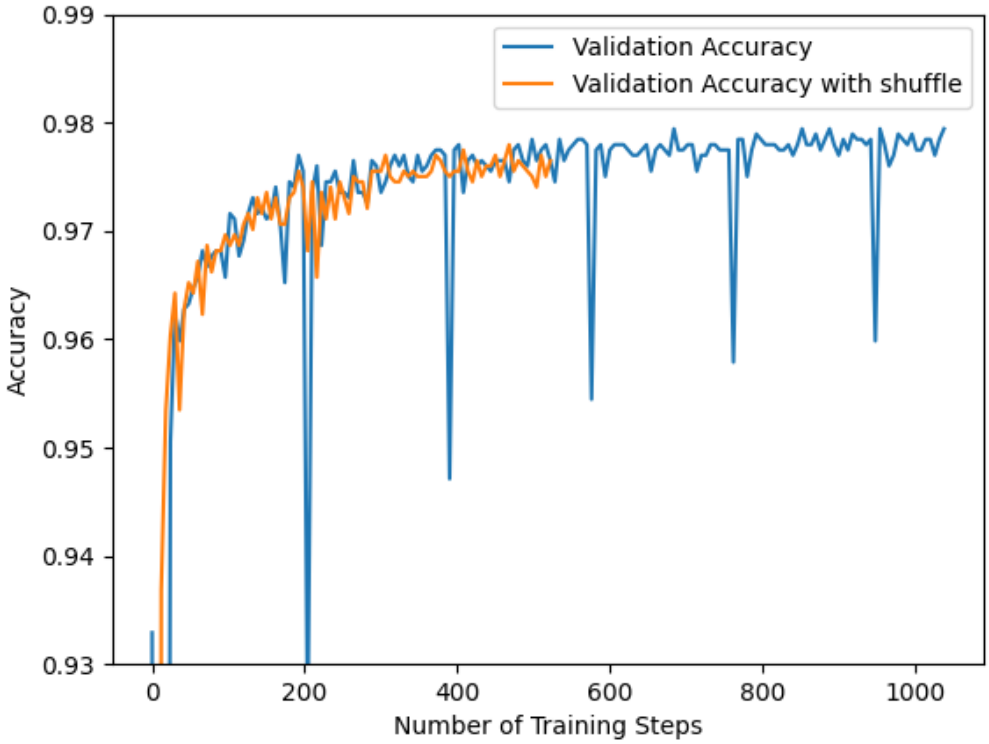


3.3 Task 2d)

Early stopping kicks in after 33 epochs.

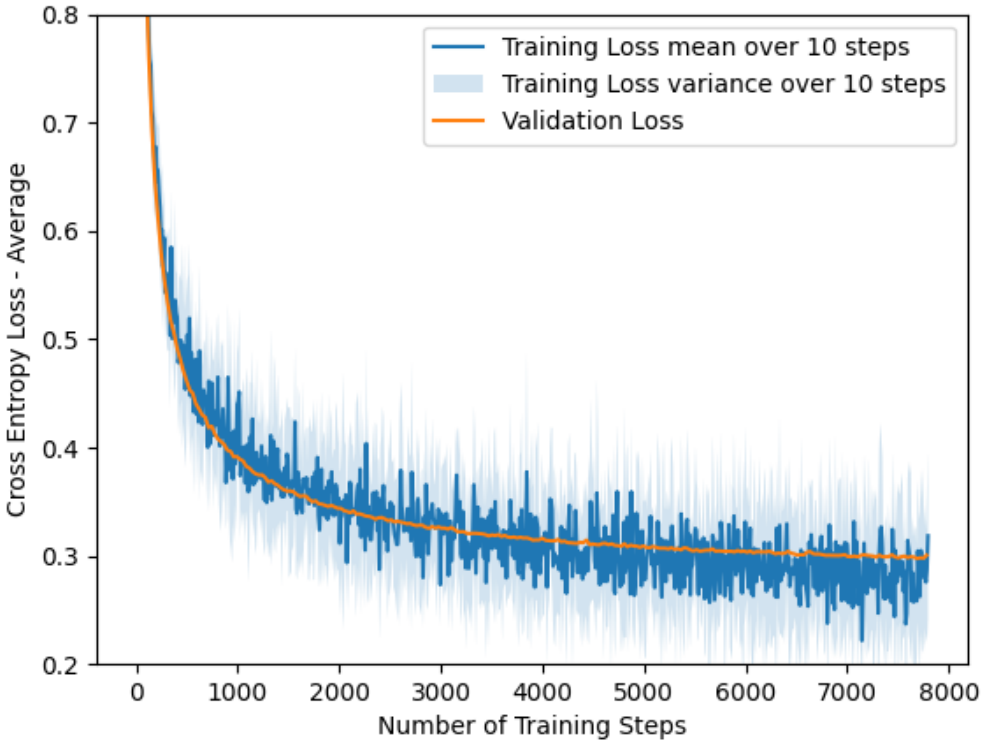
3.4 Task 2e)

The graph has notably fewer spikes and is in general more stable. This happens because shuffling the dataset makes it less prone to overfitting, causing the model to not be as sensitive to the data it is trained on. If for instance the data contains certain orders or patterns, shuffling the dataset will make the model avoid learning these patterns.

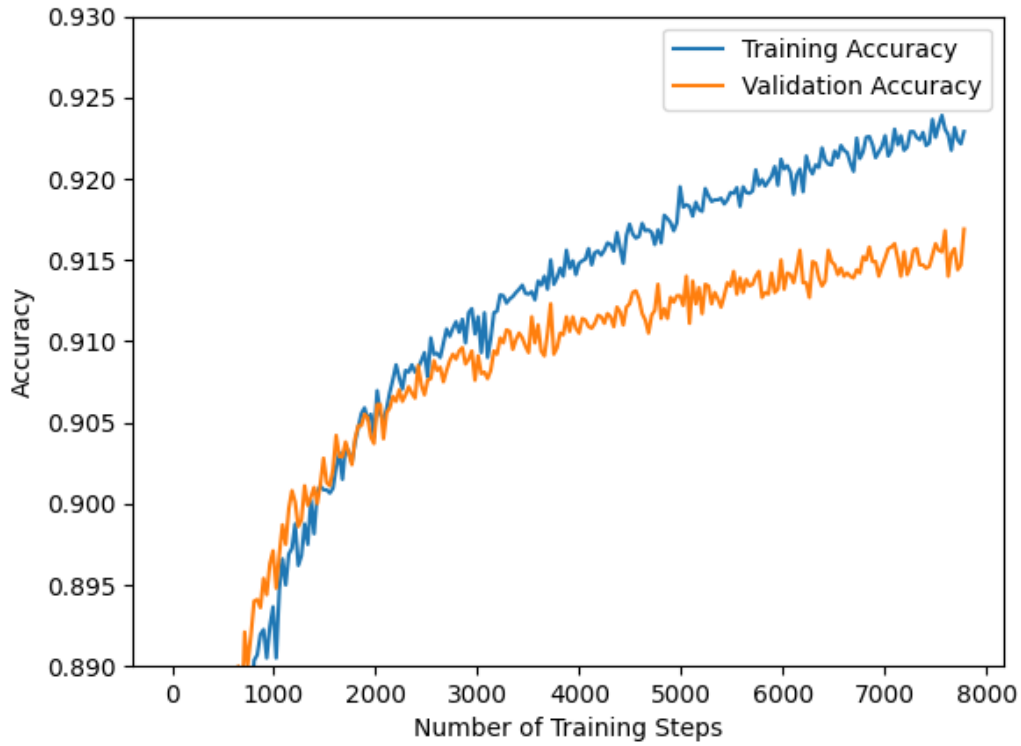

Validation Accuracy
Validation Accuracy with shuffle
Accuracy
Number of Training Steps
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0
200
400
600
800
1000

4 Task 3

4.1 Task 3b)



4.2 Task 3c)



4.3 Task 3d)

The difference in accuracy on the training data and the validation data seems to start diverging at around 3000 steps. Here, the training accuracy continues to increase with the same slope, while the accuracy on the validation accuracy starts to slow down, but not by much. The difference between the two could indicate slight overfitting, however, with the difference being so small, one could argue it is a good fit.

5 Task 4

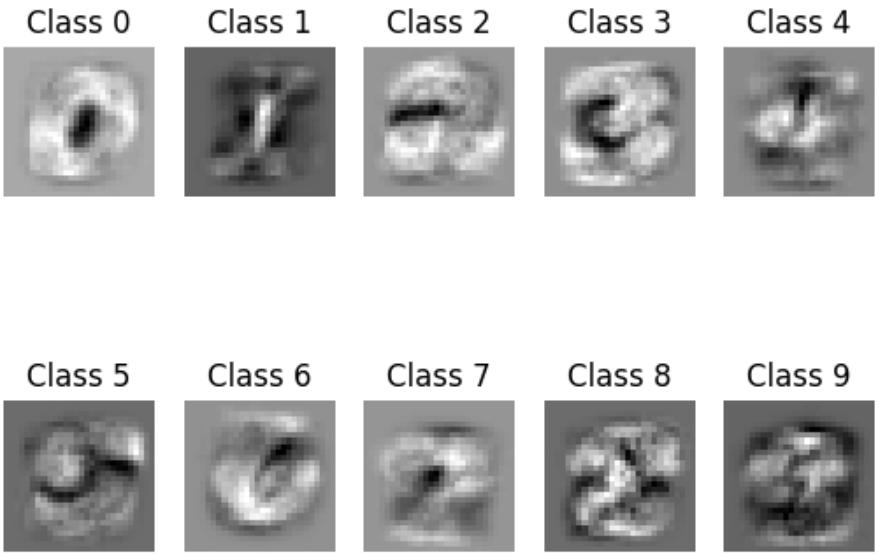
5.1 Task 4a)

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

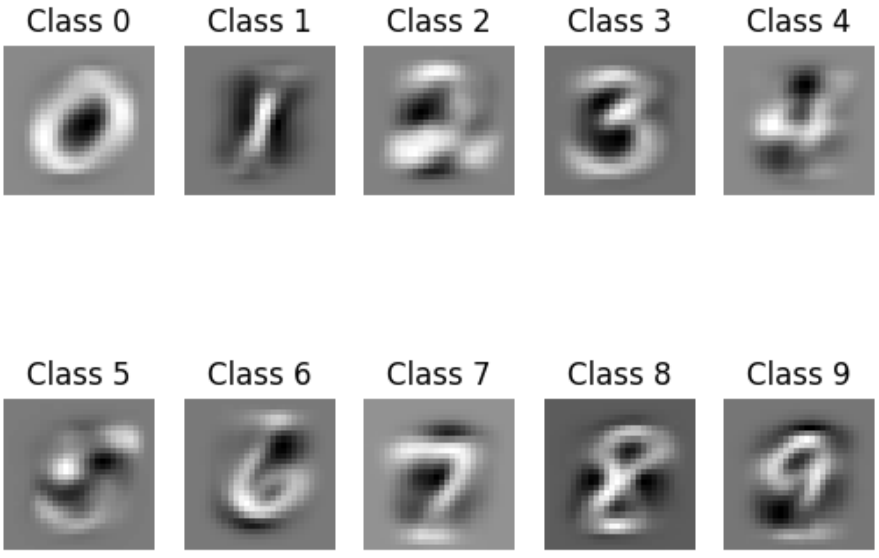
5.2 Task 4b)

The weights with the higher lambda is less noisy because of the increased generalization. This means that the model won't be prone to outliers and it wont become too complex, causing the weights to better reflect the generalized data, since larger weights are penalized.

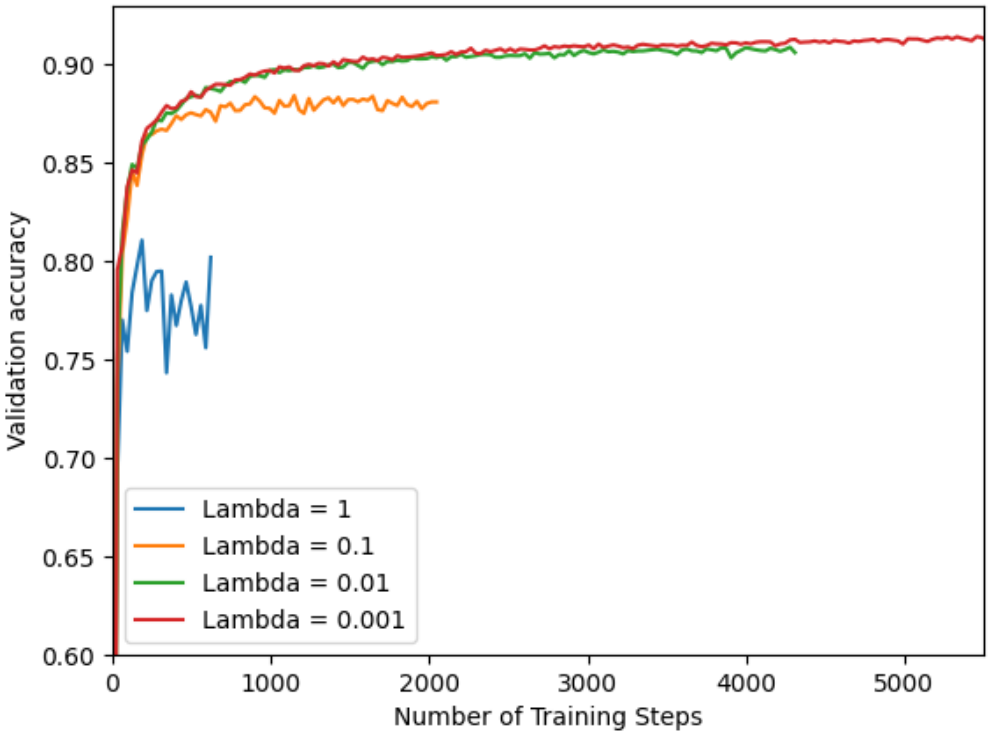
Lambda = 0



Lambda = 1



5.3 Task 4c)

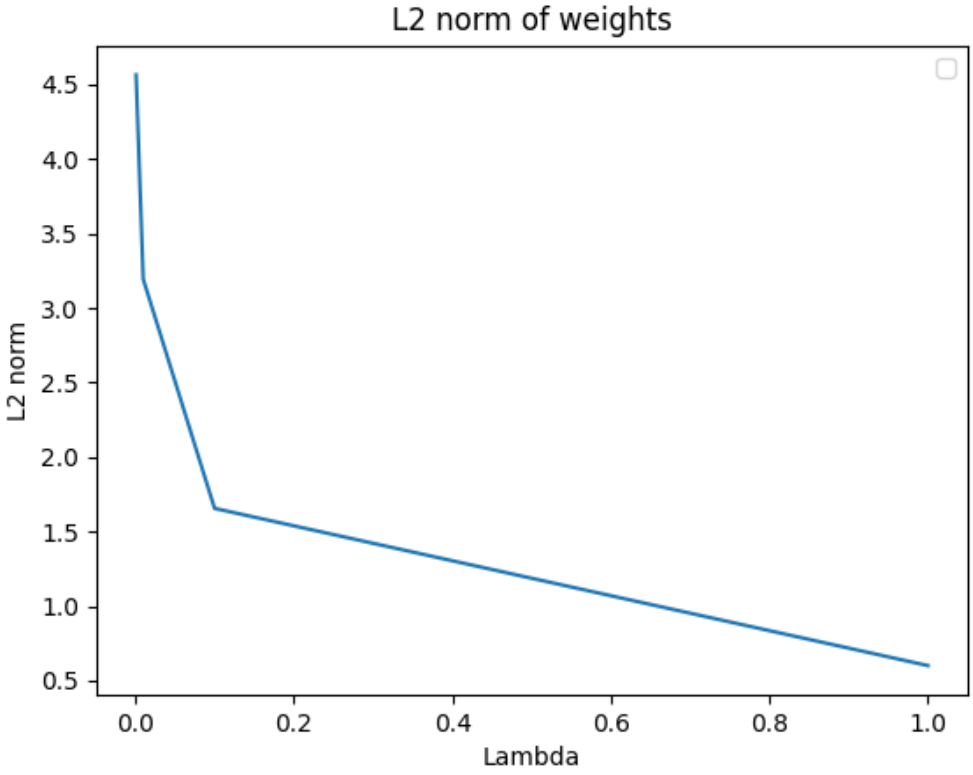


5.4 Task 4d)

The reason the validation accuracy degrades with regularization could be because the model simply becomes too generalized, thus resulting in underfitting.

5.5 Task 4e)

One can see from the graph over the l2 norm that the length of the the weight vector decreases as lambda increases. This makes sense from the definition of L2-regularization, since larger values of lambda causes the second term in equation (9) to decrease even more.


L2 norm of weights
L2 norm
Lambda
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.2
0.4
0.6
0.8
1.0