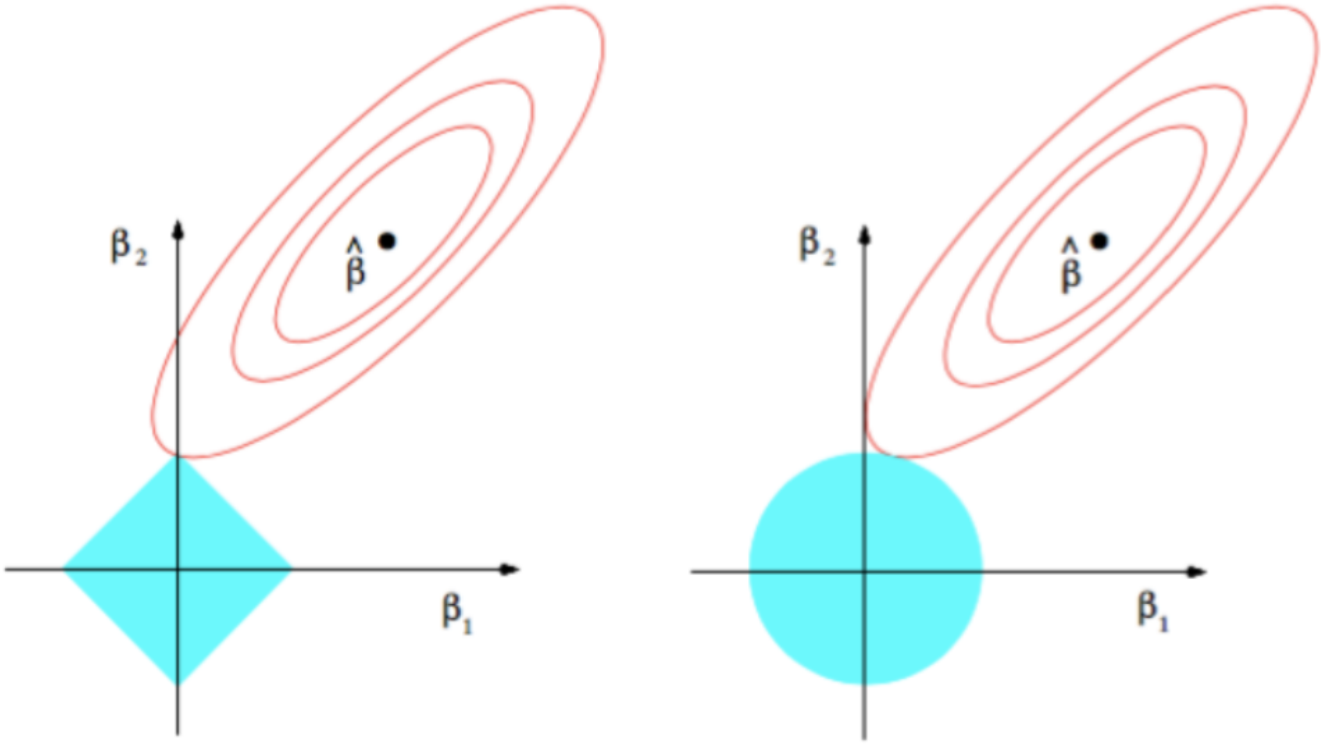
CSEP 546 HW 2

1a. False. The weight on a feature is not directly related with model error if the model is not standardized. Additionally, model error can decrease when a feature is removed. The feature may also be coorrelated to other features, causing weights to increase in other features when it is removed.

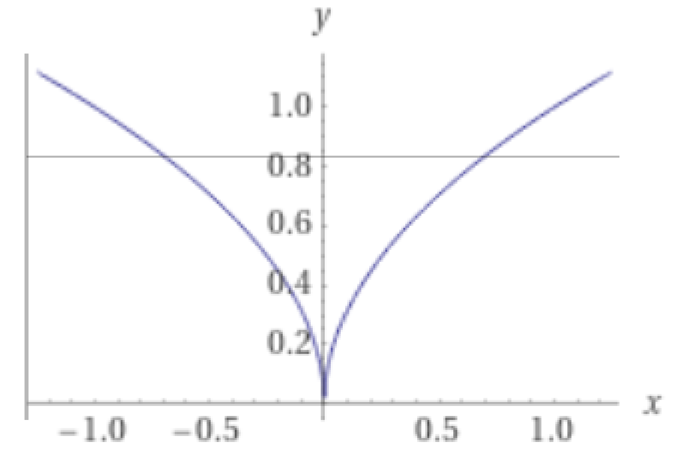
1b. Given the shape of the L1 norm, it is more likely to occur at sparse points since the solution surface is likely to find a touch point on the tip of the l1 diamond.

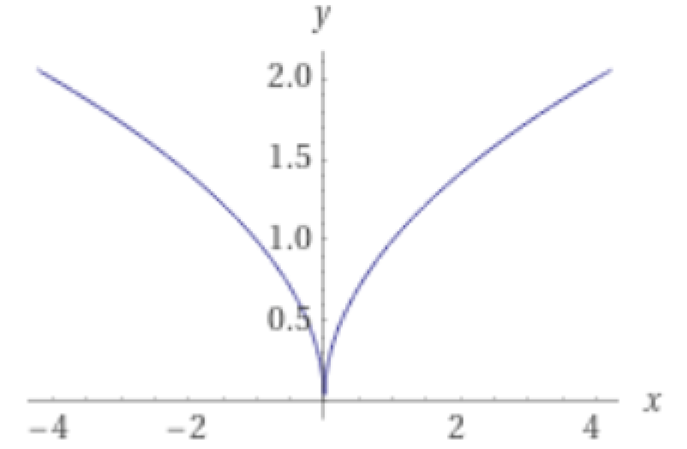


1c: Compared to the traditional lasso and ridge regularizers, this function penalizes weights much less. Additionally, it keeps the desirable propertities of lasso in term of sparsity while being a less harsh regularizor.

The downside is that this function would have difficulty converging. As the step size lowers, the function will keep overshooting. The function also does not give a closed form solution.

Notes: Concave function





1d. True, if the step-size is too large we may not converge and may even diverge. This is because while our step gives the maximum change, it is in the wrong direction and will overshoot the minimum on each step.

1e: While stochastic gradient descent only takes a single example (a batch size of 1) per iteration, it works given that the dataset size is large enough and there are enough iterations. As stated by theories of convex minimization, stochastic graident descent should converge when the objective function is convex.

1f: The benefit of SGD is that it is much faster since only a single training sample is used versus all the samples in the training set. This means that SGD converges much faster compared to GD. However, the error function is not as well minimized since SGD parameters values reach the optimal and keep oscillating. In contract, GD parameters are much more stable.

2a:

PctForeignBorn: This is susceptible to historical policy choices as political sentiments often determine immigration policies

PolicBudgPerPop: This is susceptible to policy as police budgets are determined by local politics

pctUrban: This is susceptible to planning policies as well as real estate properties within the cities

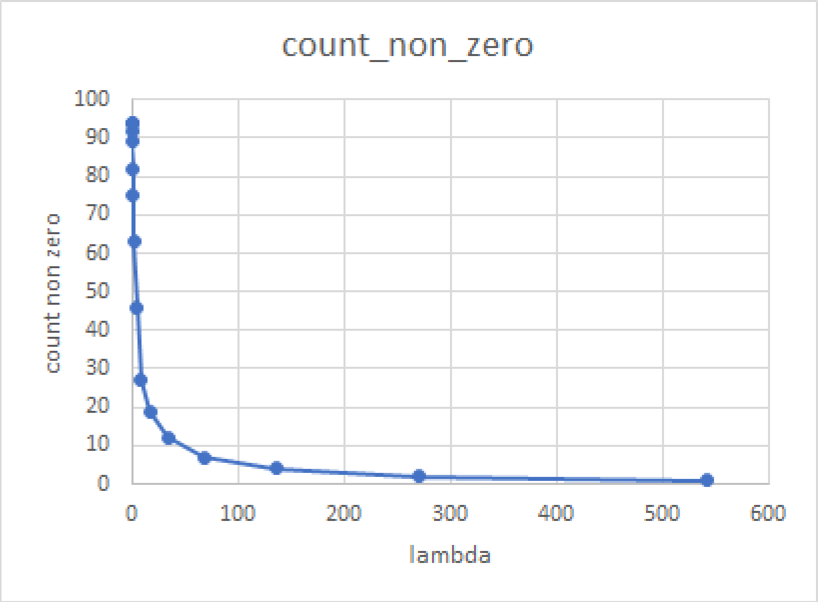
2b:

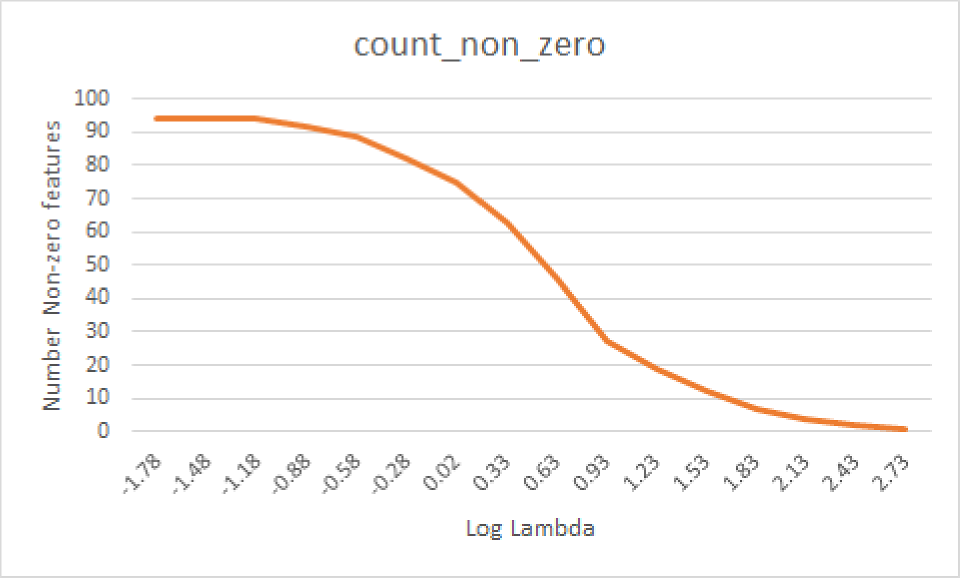
NumInShelters Violent crime in an area might depress the area economically, increasing the number of people in shelters.

PctVacantBoarded Violent crime in an area might increase the number of boarded vacant homes, rather than the other way around.

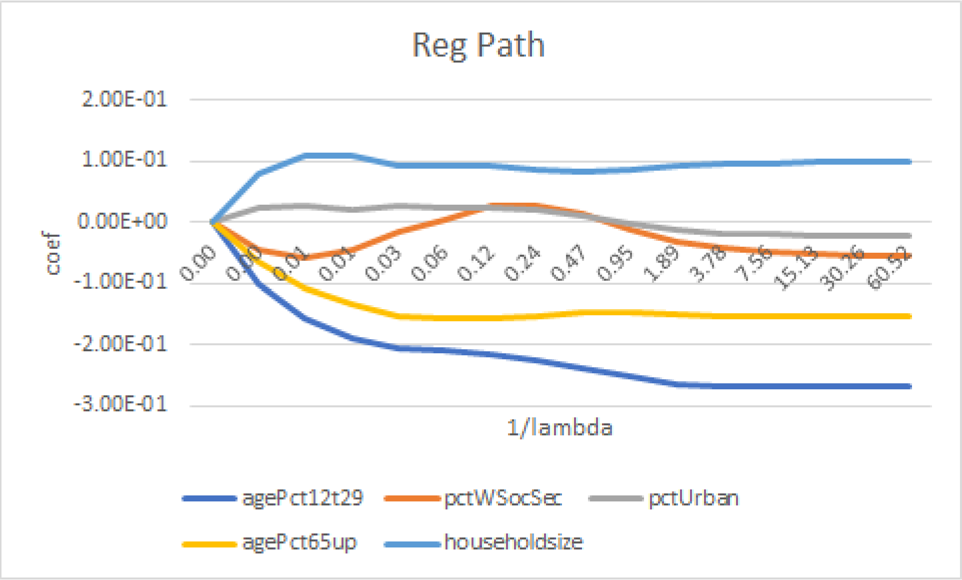
PctUnemployed Violent crime might cause unemployment due to depressed economic conditions rather than unemployment driving violent crime.

2c:

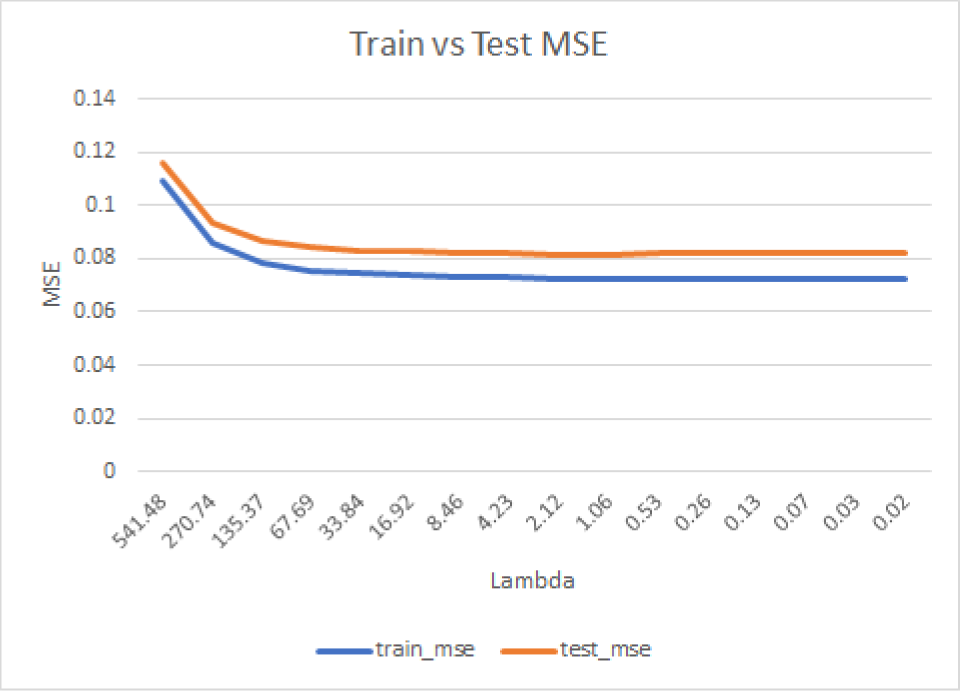



count_non_zero
Number Non-zero features
100
90
80
70
60
50
40
30
20
10
0
-1.78
-1.48
-1.18
-0.88
-0.58
-0.28
0.02
0.33
0.63
0.93
1.23
1.53
1.83
2.13
2.43
2.73
Log Lambda

2d:


Reg Path
coef
2.00E-01
1.00E-01
0.00E+00
-1.00E-01
-2.00E-01
-3.00E-01
0.00
0.00
0.01
0.01
0.03
0.06
0.12
0.24
0.47
0.95
1.89
3.78
7.56
15.13
30.26
60.52
1/lambda
agePct12t29
pctWSocSec
pctUrban
agePct65up
householdsize

2e:



2f:

Results for Lambda

Largest: 0.068725 (Index 45) as PctIlleg

Smallest: -0.069215 (Index 39) as PctKids2Par

MSE: 0.074 vs 0.072 final convergence

Feature Count: 12

Compared to a smaller lambda, we have much less features, yet only a small difference in train mse (0.074 vs 0.072). Given that there is a total of 12 features, the model is much more interpretable than if we had a smaller lambda. The largest positive feature is percent illegal immigrant and the largest negative feature is percent of households with two parents.

2g:

Causation is different than correlation. We cannot infer that agePct65up causes reduced crime given that there might be confounding variables such as house value. Generally, older folk may own more valuable homes.

3a:

Let $u_i = u_i(w,b)$

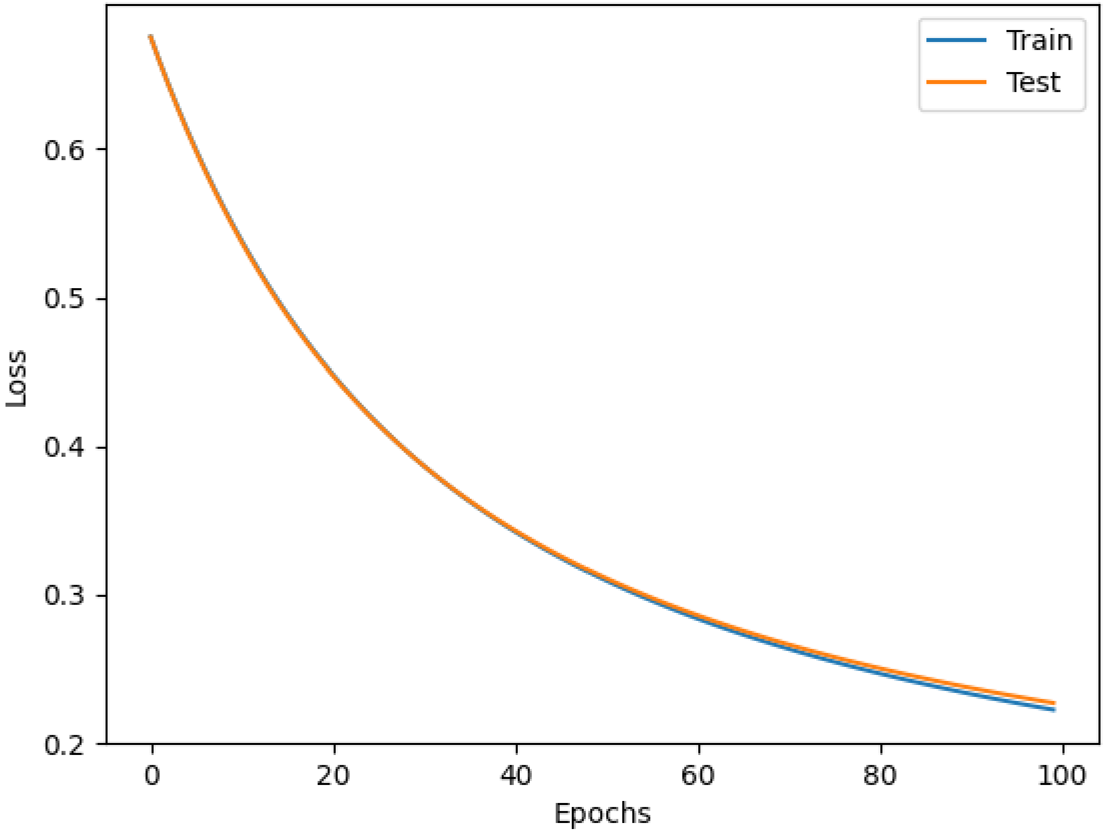
$J(w,b) = \frac{1}{n}\sum_{i=1}^{n}\log(1+e^{-y_i(b+x_i^T w)}) + \lambda||w||_2^2$

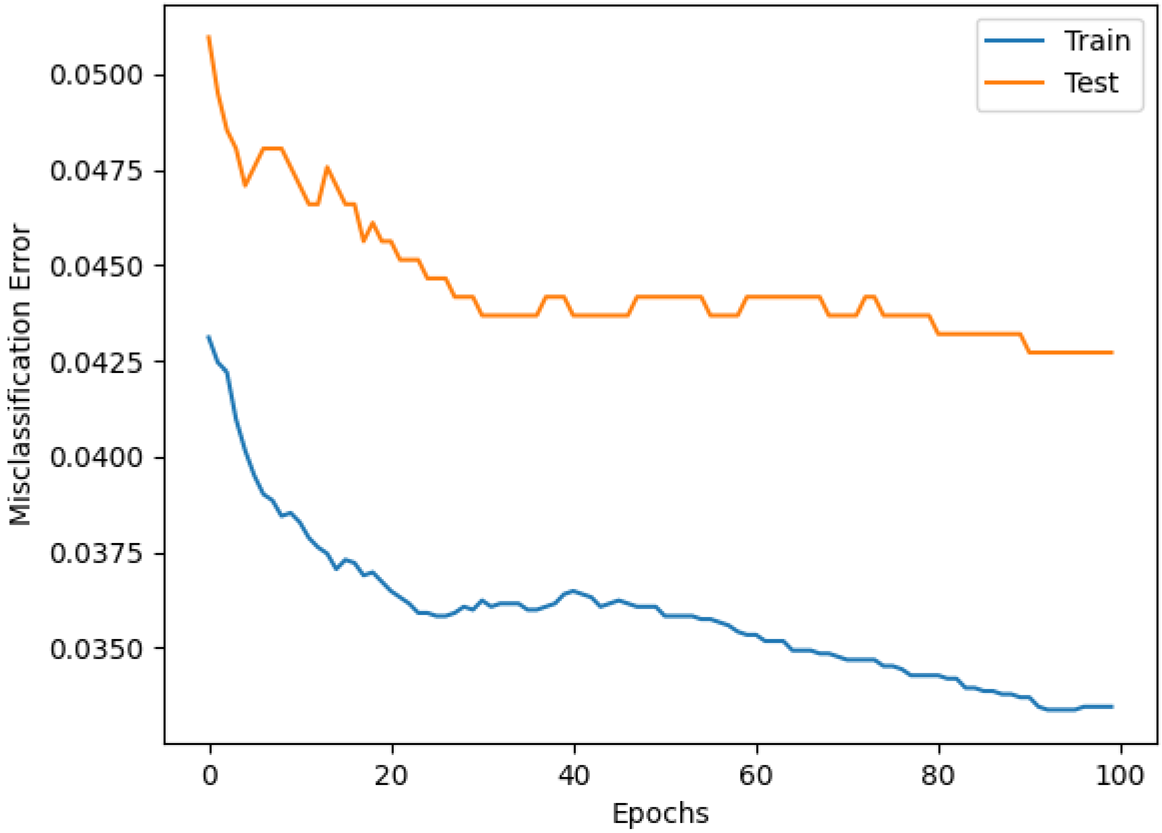
Using Chain Rule:

$\nabla_w J(w,b) = \frac{1}{n}\sum_{i=1}^{n} u_i(1-\frac{1}{u_i})y_i X^T + 2\lambda||w||$

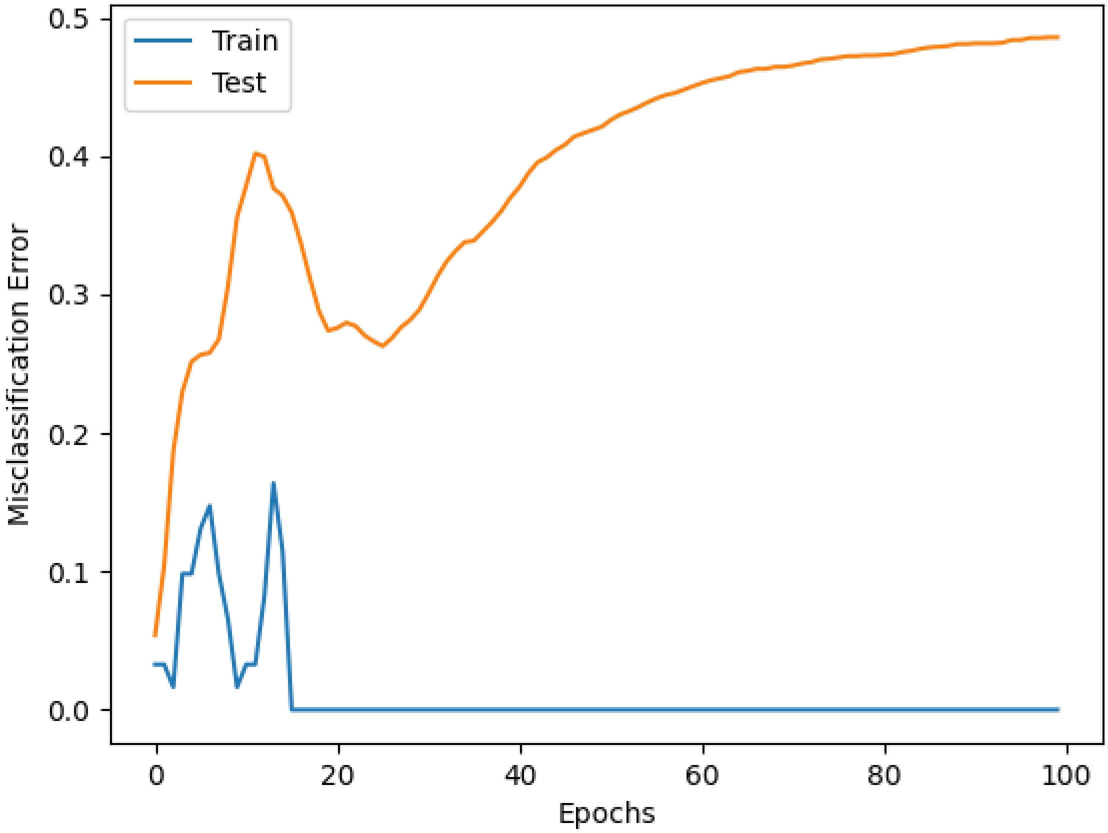
$\nabla_b J(w,b) = \frac{1}{n}\sum_{i=1}^{n} u_i(1-\frac{1}{u_i})y_i$

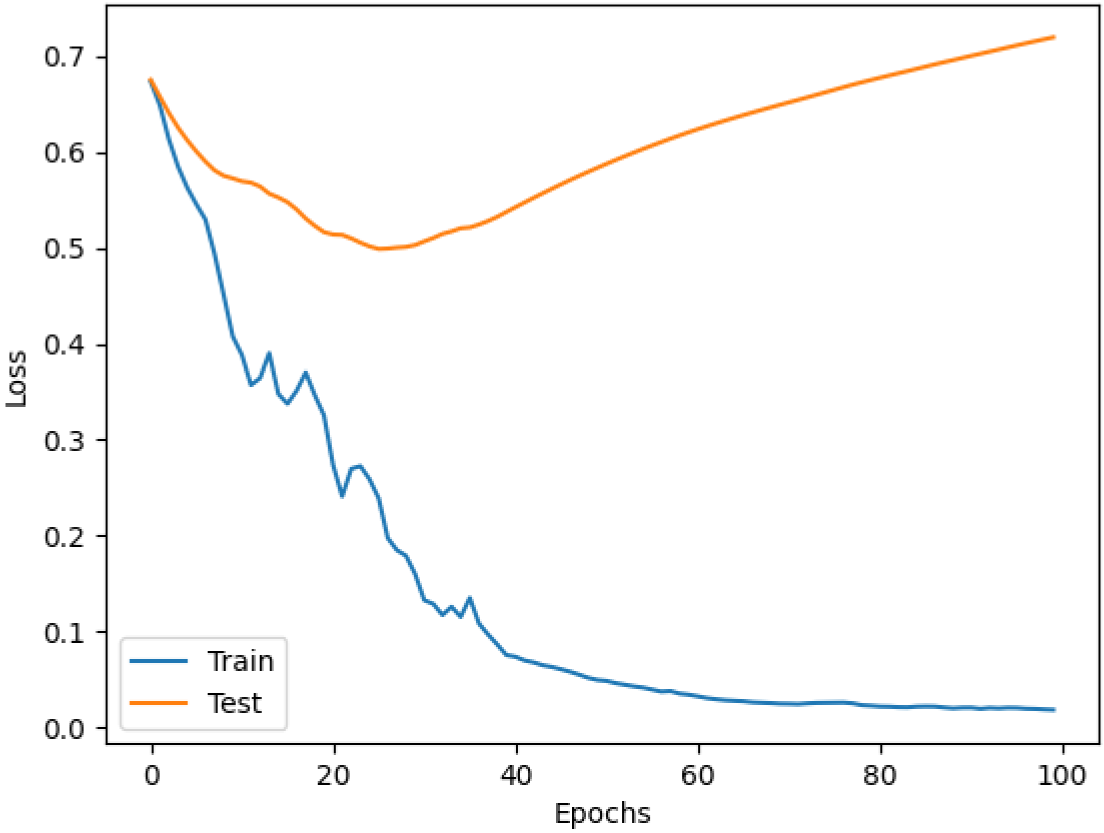
3b:



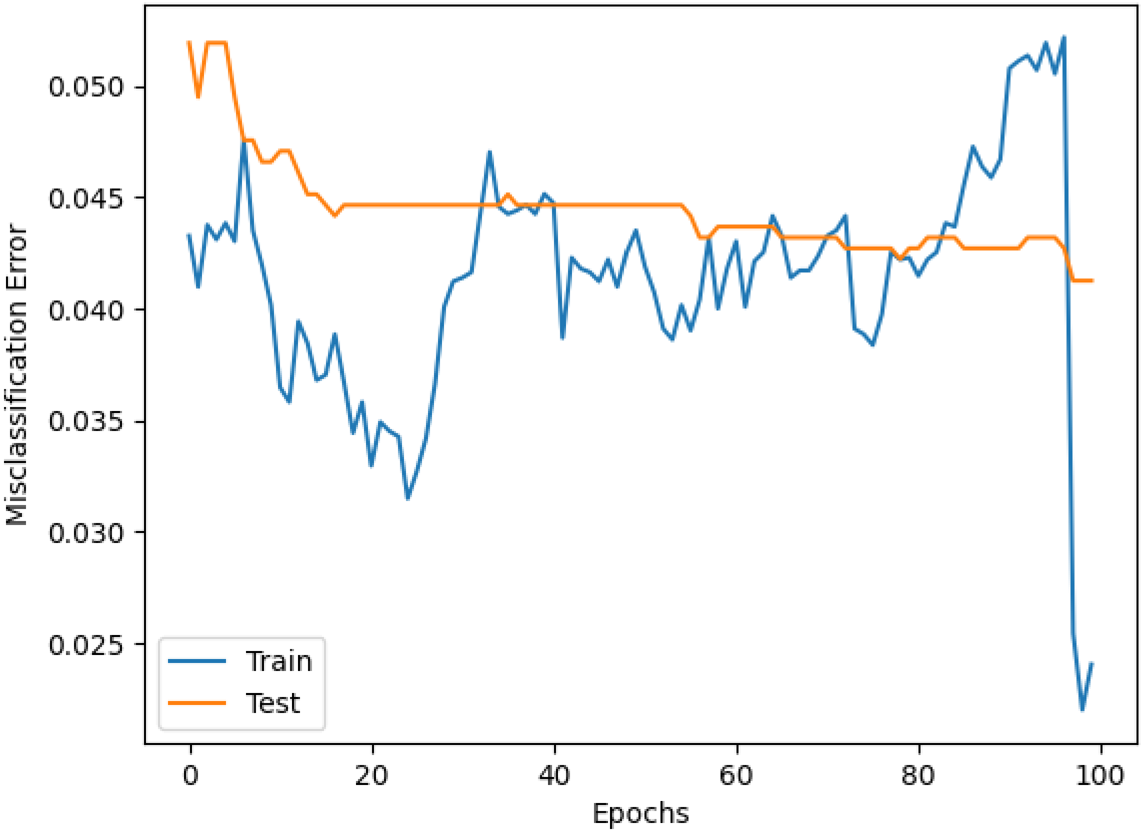

Train
Test
Misclassification Error
0.0500
0.0475
0.0450
0.0425
0.0400
0.0375
0.0350
0
20
40
60
80
100
Epochs

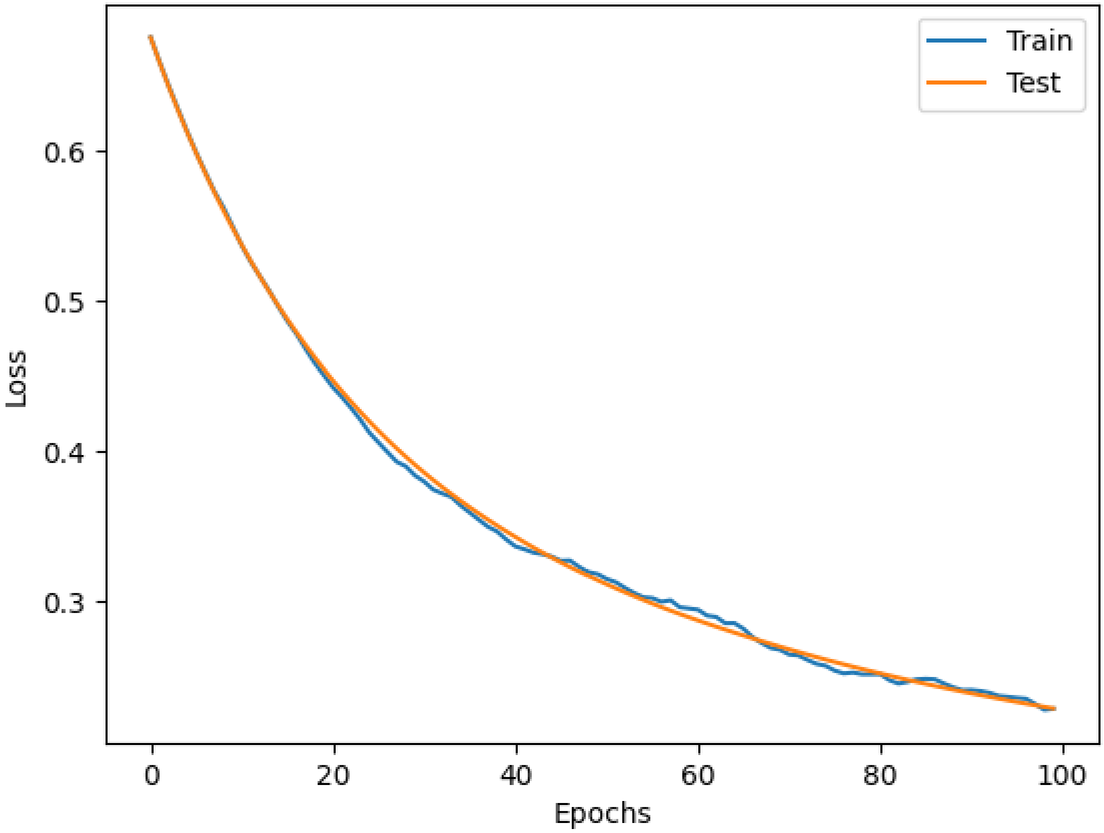
3c:


Train
Test
Misclassification Error
0.0
0.1
0.2
0.3
0.4
0.5
Epochs
0
20
40
60
80
100



3d:


Misclassification Error
0.050
0.045
0.040
0.035
0.030
0.025
Train
Test
0
20
40
60
80
100
Epochs



4a:

30 hours in total