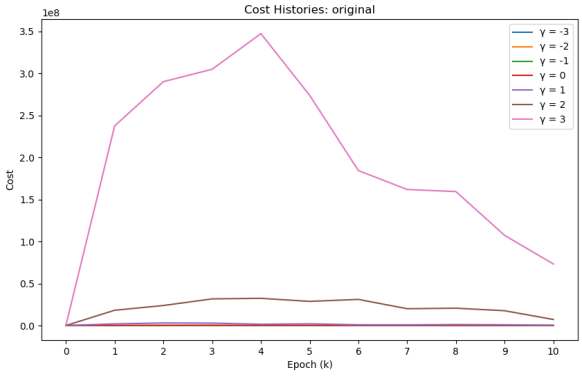
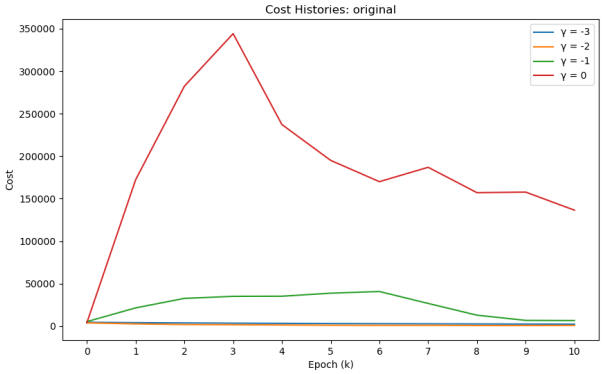
Andrew Hall
CS 589 HW7 Report

Cost History Analysis

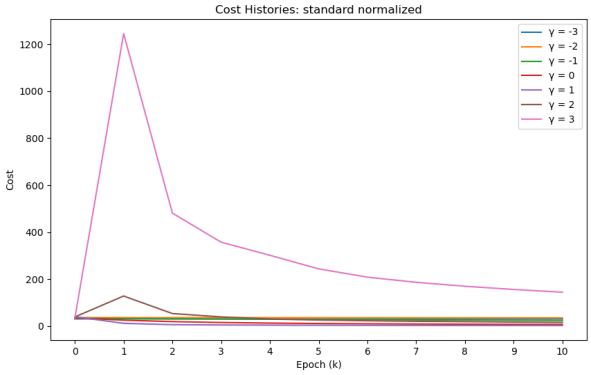
Finding the Best GD Learning Rate Magnitude γ for each preprocessed MNIST dataset
(Constant Learning Rate = 10^{γ})

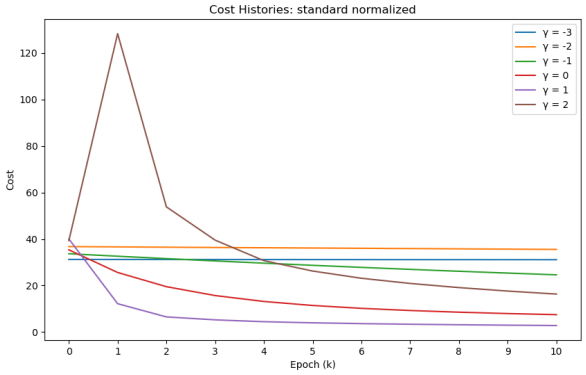
Original Data (No Preprocessing): $\gamma > 0$ (left) goes out of control,
(right) you can see $\gamma = -2$ is optimal but even $\gamma \in \{-1, 0\}$ diverge pretty badly.



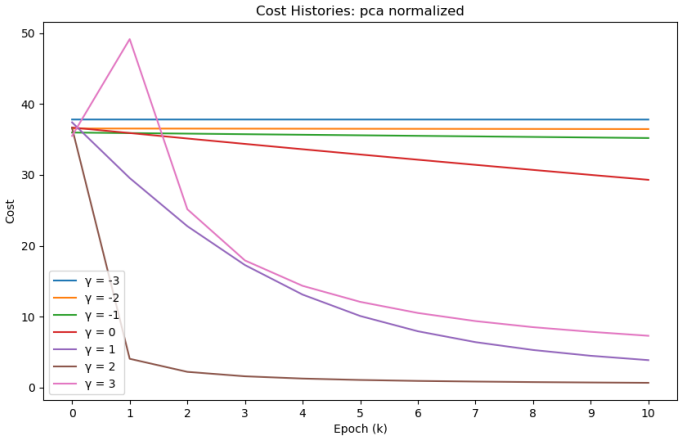


Standard Normalized Data: $\gamma \in \{2, 3\}$ diverges (left) but it's not nearly as bad as above. $\gamma = 1$ is optimal. The sensitivity is reduced, but it seems easier to have too small of a learning rate ($\gamma < 0$).





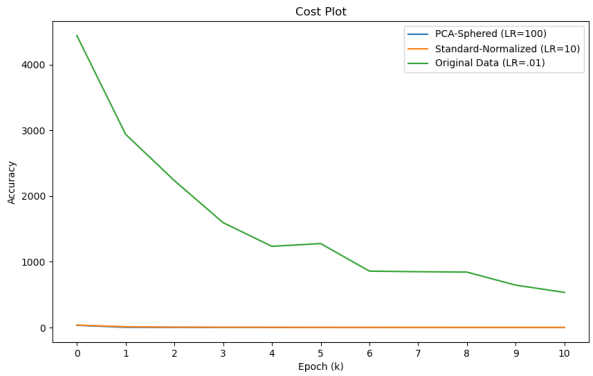
PCA-Sphered Data: This normalization method offers the most resilience to large magnitude LR, $\gamma = 2$ is optimal.

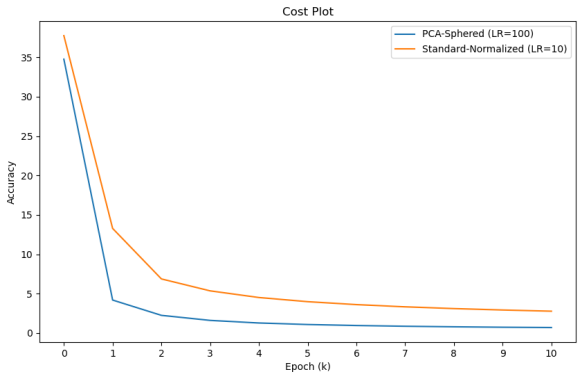


Visualizing the Difference Between Normalization Methods

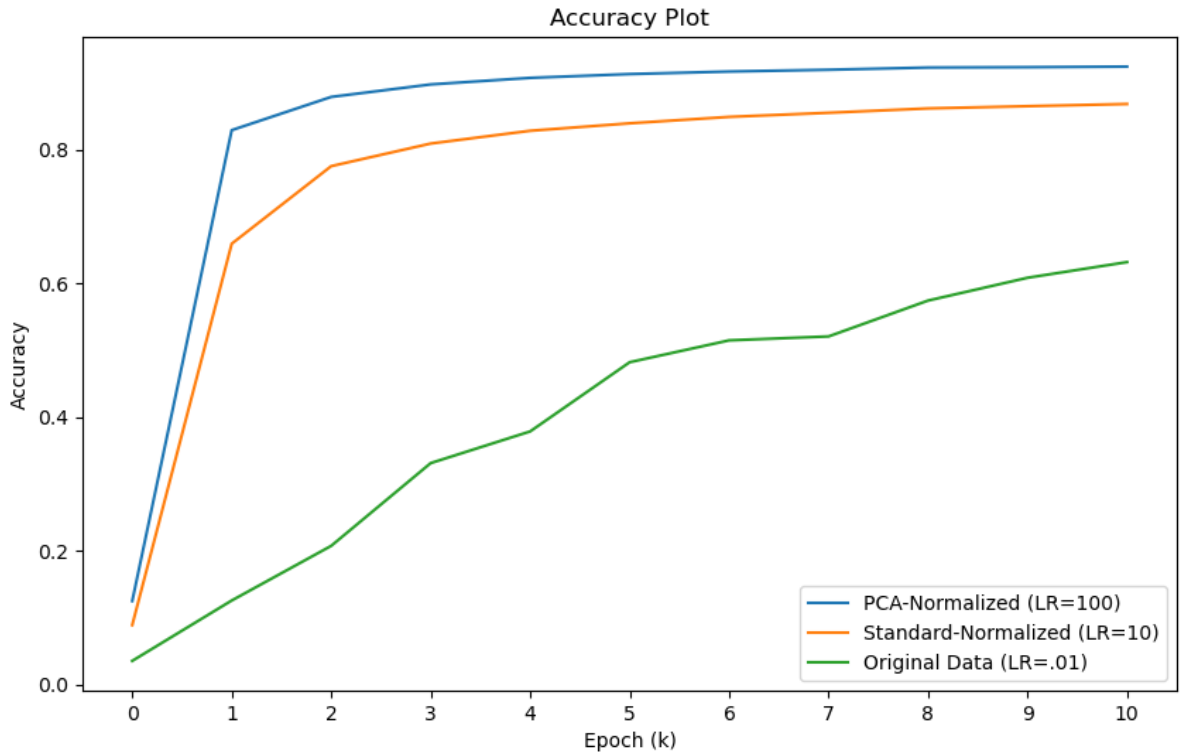
Each Plot Shows the Optimal Magnitude of Learning Rate for Each Dataset

As can be seen on the left, the unpreprocessed data has a huge cost to begin and exhibits instability, on the right you can see the difference between the preprocessed data more clearly. Note that weight initialization is random which may account for some of the difference in starting cost.

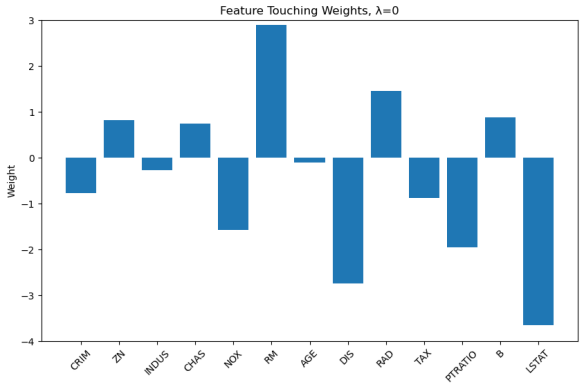


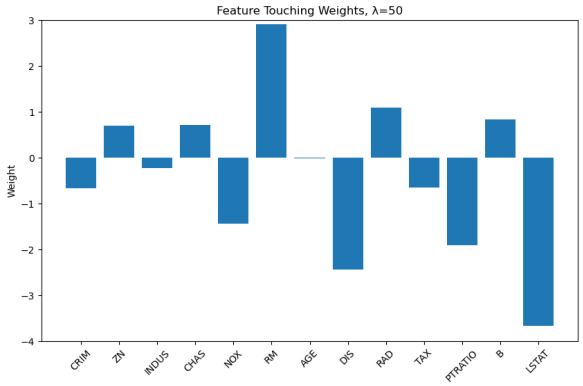


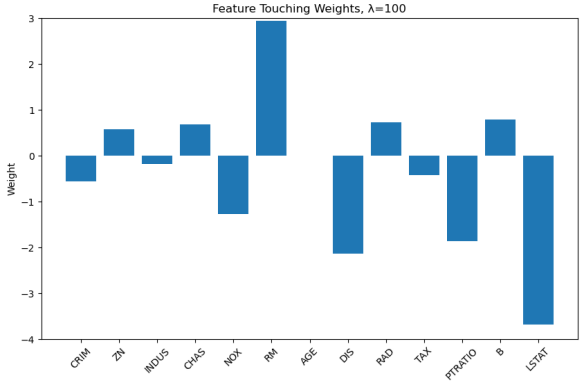
Given the accuracy plot below, it is interesting to see that the PCA normalization not only allows the model to converge faster but also to a higher accuracy. This shows the importance of data preprocessing in machine learning. Proper normalization has the potential to increase the speed of learning as well as the performance of the final model, not to mention providing stability to the optimization landscape.

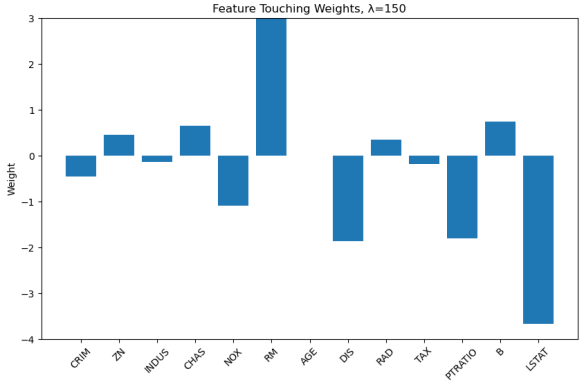


Regularization Strength Analysis on Boston Housing Dataset

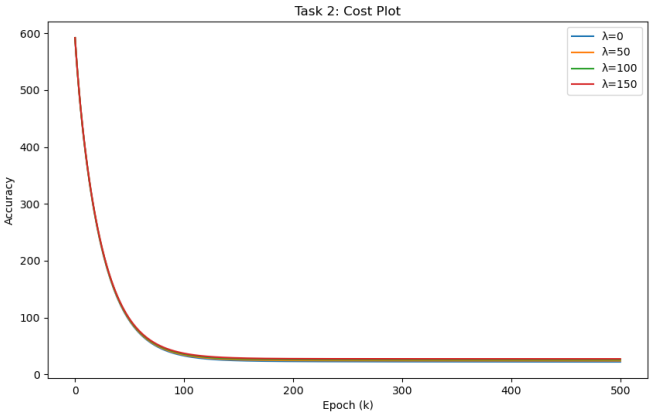








I used 0.1 as a learning rate, 500 max iterations and zero-initialization for all weights in each model trained. The constant weight initialization scheme helped illustrate the effects of regularization since the same feature touching weight appears to be diminishing with each increase in λ across the different models (see RAD & TAX). Although it was interesting seeing different random initializations ending up with approximately the same magnitude of weights but with seemingly random sign, sometimes being positive and sometimes negative. I also noticed that the magnitude of important weights slightly increases with regularization strength (most noticeable in the RM feature). I think I was successful in recreating the histogram plots.



The cost plot above shows that each model did successfully converge. The following table shows the final cost each one was able to achieve. It shows that regularization does affect the achievable cost.

λ	0	50	100	150
Final Cost	22.17	24.09	25.83	27.39