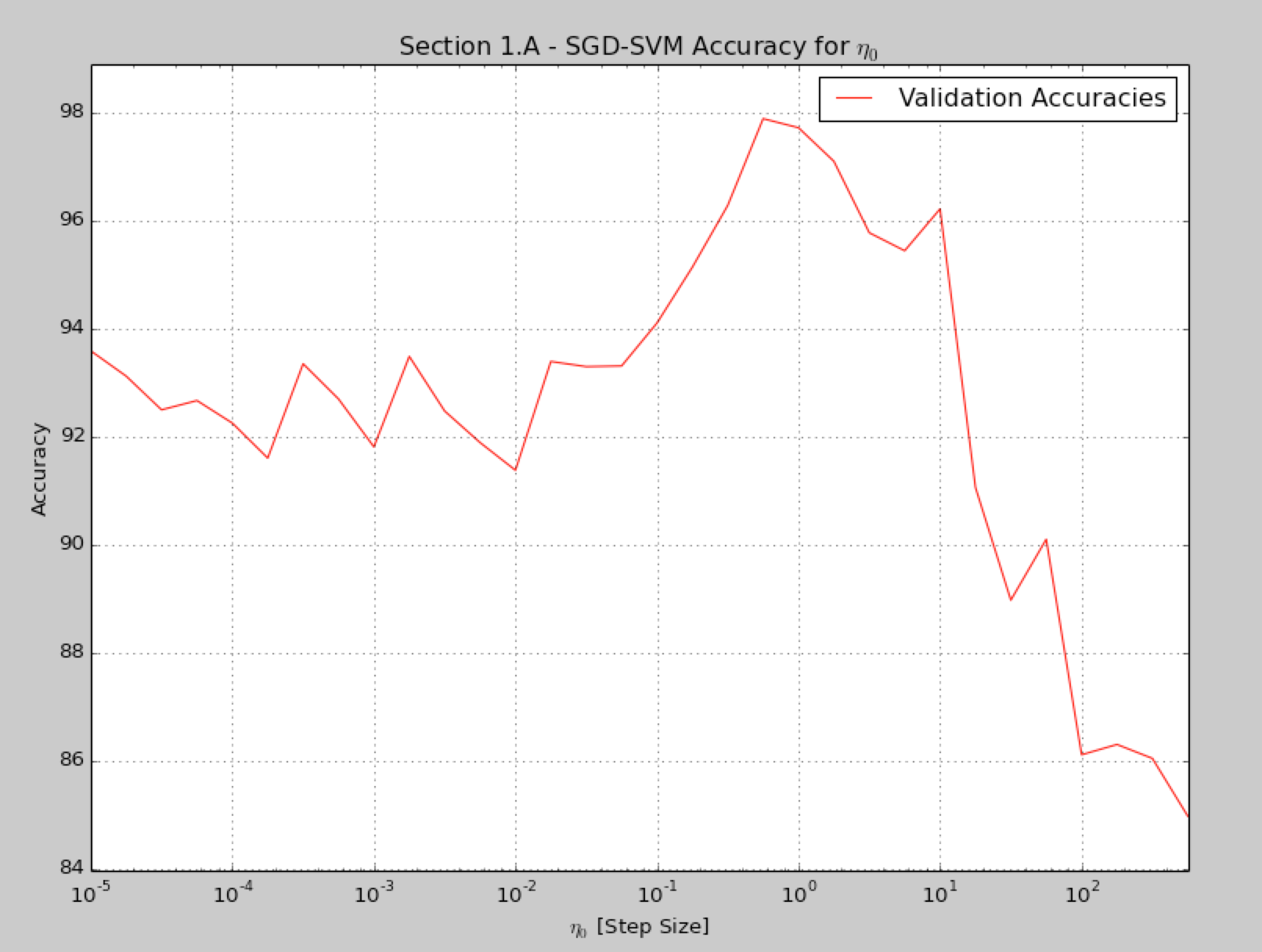
Assignment #4 – Programming Part

**Code Location:**

**Question 1**

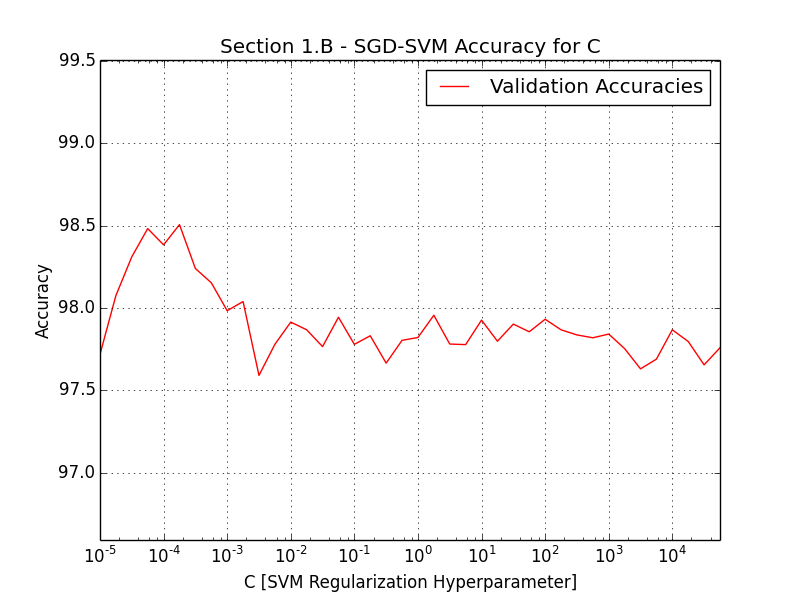
**a.** Searching for the step size that gives the highest validation accuracy over the range [10^-5, 10^3] with resolution of 0.25 yields the following graph:

****

The graph suggests that when the step size is too small for the amount of gradient steps SGD takes (T), the SVM’s weights don’t manage to update by the end of the training to fit the features MNIST model accurately. This phenomenon is mitigated as we approach the ideal step size (eta0=0.25, accuracy is almost 98%).

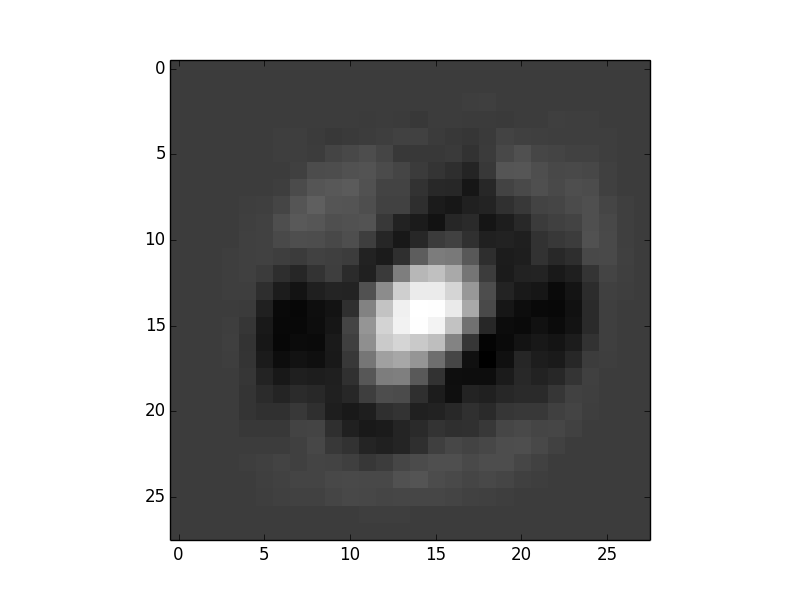
When the step size is too big, the SVM’s weights get big updates that may cause the gradient updates to skip the global minima of the loss function, yielding bad accuracy again (the algorithm fails to converge to the local minimum with such big step resolution). We’ve neglected results above 10^3 as they give extremely bad accuracies, and eventually cause numpy to overflow (even for float64 ndarrays).

**b.** Searching for the C regularization parameter that gives the highest validation accuracy over the range [10^-5, 10^5] with resolution of 0.25 yields the following graph:

****

The best C found is 5.7e-5, which yields an accuracy of almost 98.5% on the validation dataset (small value of C suggests the SVM is looking for a large-margin separating hyperplane).

**c.** With the optimal C, eta0 parameters found in Sections a, b, and T=20000, the classifier achieves an accuracy of 98.54% over the validation dataset.

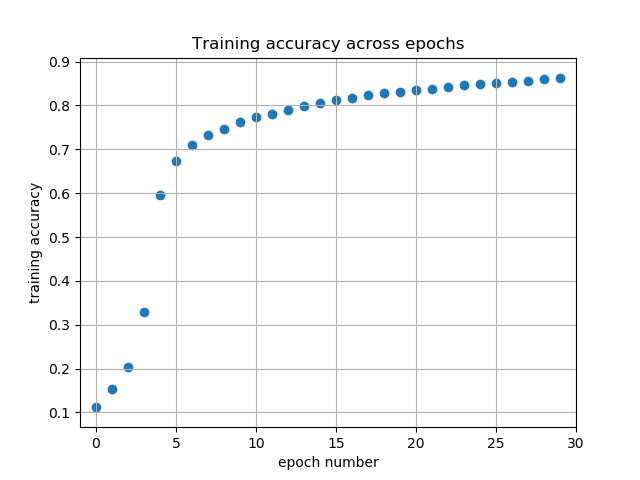
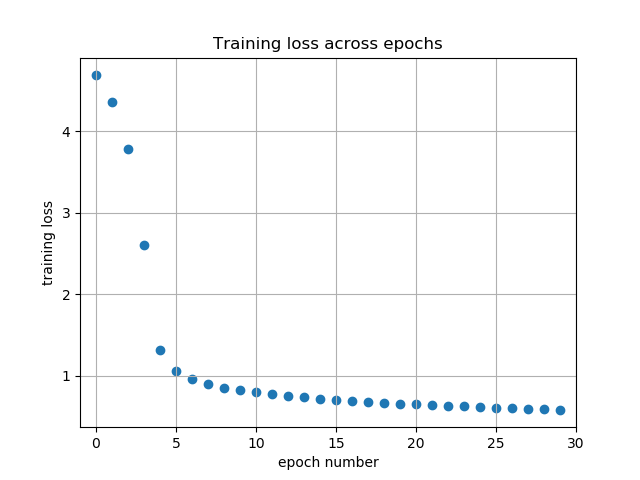
The SVM’s weights appear as follows: 

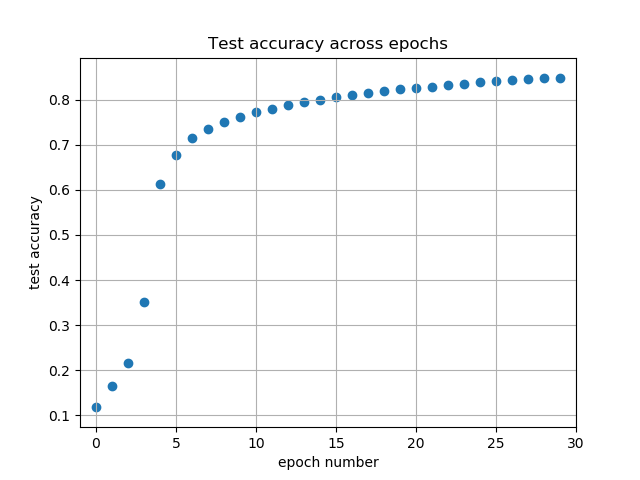
Compared to the Weight Matrix extracted in Ex3, this visualization looks smoother (lower frequencies), and the black / white areas that identify 0/8 digits appear to be more uniform.

**d.** The average accuracy for 10 iterations on the test set is: 99.20% (Best classifier, T=20,000).

**Question 2**

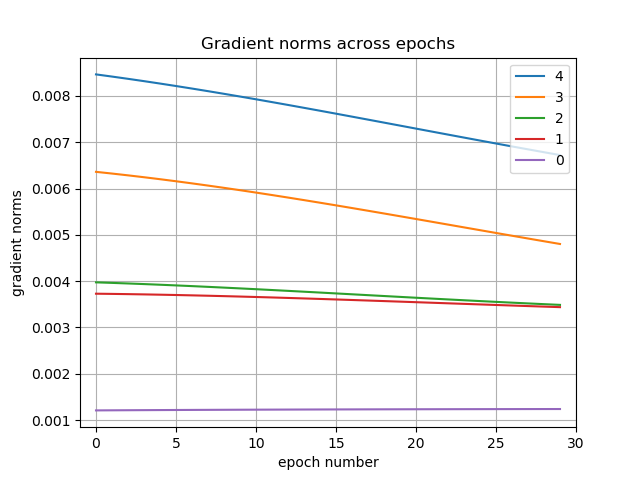
**b.** We got the following results:





As we can see from the graphs, as training progresses, the training and test accuracies increase and the training loss decreases.

**c.** In the last epoch we got a test accuracy of 91.59%

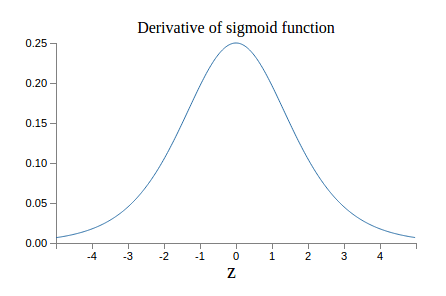
**d.** The plot for this section:

As we can see, gradients closer to the input (lower numbers) have lower norm.

The expression of the gradients is:

This implies that when we look at lower layer gradient, it will have more terms of (because unwrapping will yield more terms of ) .

The looks like this:



And the key feature here is that is always smaller than 0.25.

So if we take the product of many terms, because it is smaller than 1, the result will tend to exponentially decrease.

This can be a possible explanation for the vanishing gradient phenomenon.