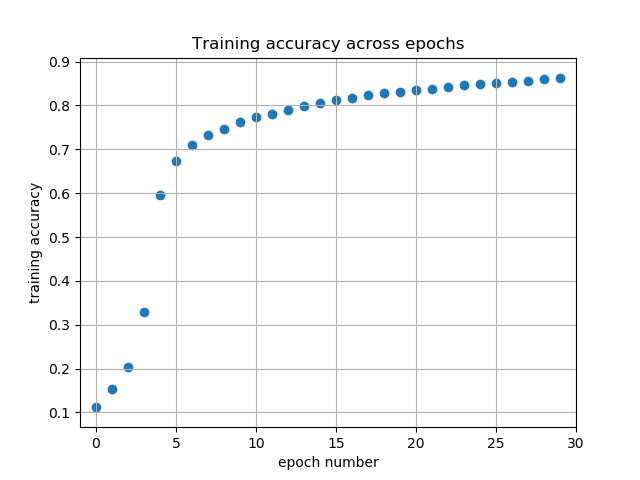
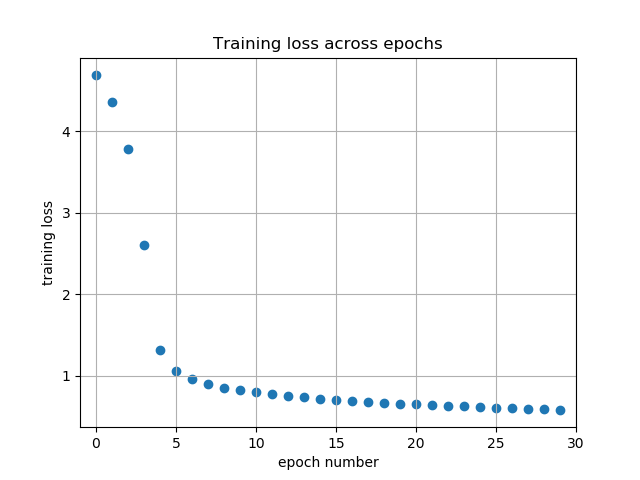
Assignment #4 – Programming Part

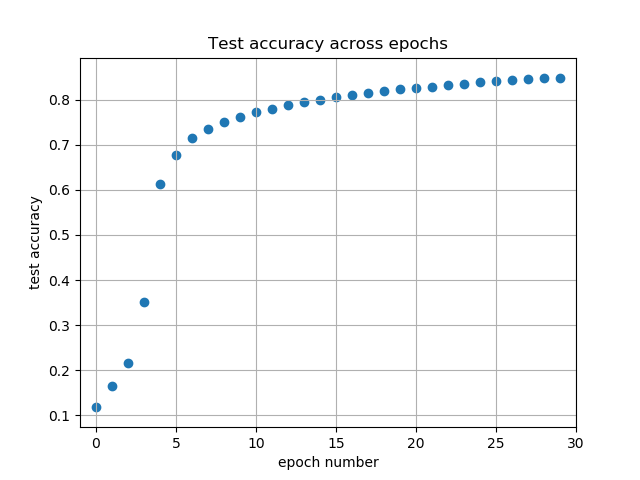
**Code Location:**

**Question 1**

**Question 2**

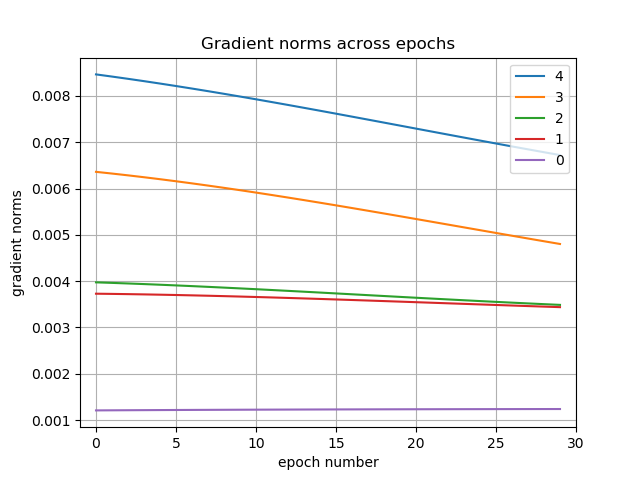
**b.** we got the following results:





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%

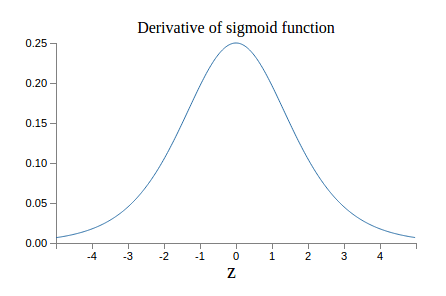
**d.** The plot for this section:

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.