# Homework 2: Neural Networks

CS 545: Machine Learning, Winter 2017

*Ben Wilson*

## Introduction

Similar to the first homework, in this assignment we were tasked with classifying handwritten digits from the MNIST data set. We still had 785 inputs and ten outputs, but this time we implemented a two-layer neural network to gain even better accuracy.

The neural network differs from the perceptron model by using a hidden layer of nodes that enable it to solve non-linear problems. However, training a neural network takes more time to train because it has more weights than a perceptron model and the algorithm to update the weights is more computationally expensive. All inputs are fully connected to each node in the hidden layer and each hidden node is fully connected to the outputs and the weights are updated by means of back propagation.

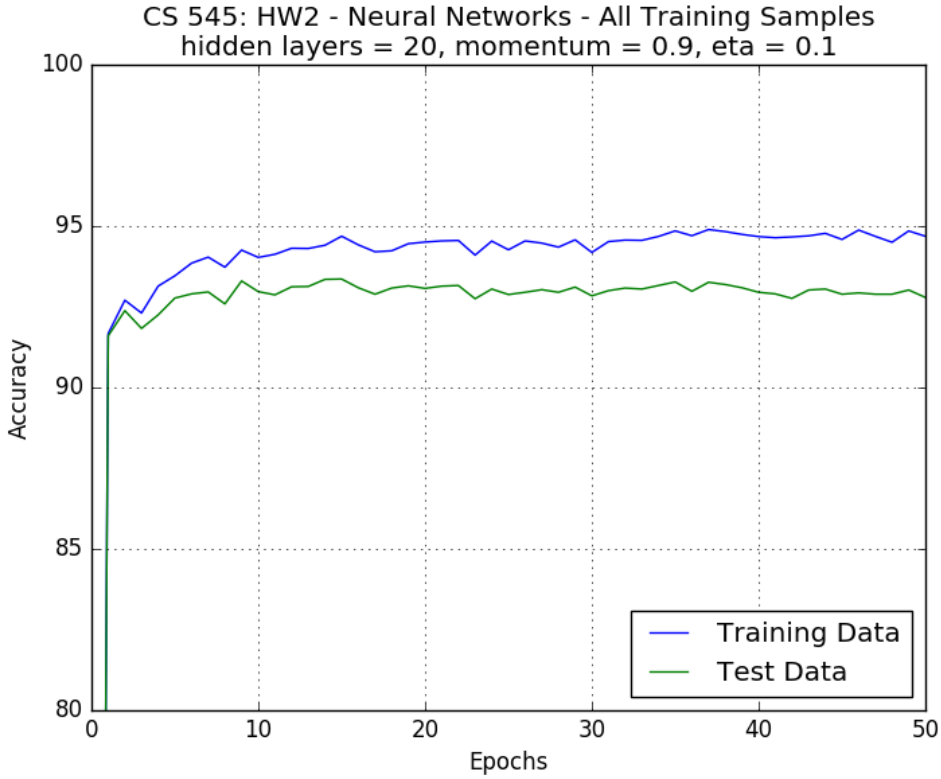
## Experiment 1: Vary the number of hidden units

For the first experiment in this assignment, the number of hidden units is varied from 20, 50, to 100. As the number of hidden units is increased, the test accuracy is also increased: 93% accuracy with 20 units, 96% with 50 units, and 97% with 100 units. Increasing the number of hidden units changes how the results converge: with 20 units there is a lot more oscillation and seems to bounce around 93%; with 50 units the results quickly converge after about 10 epochs and remains steady around 96% accuracy; with 100 units the accuracy reaches it's highest point in the first dozen epochs and then slowly decreases in accuracy. All three networks showed evidence of overfitting, but it's most clearly seen in the networks with 50 and 100 hidden units; in the 50 unit case it can be seen that the test accuracy stops increasing while the training accuracy keeps trending towards 100%; in the 100 unit case the test accuracy gets worse while the training accuracy gets better. This is a strong indication of overfitting. Compared to homework 1, the neural network performs much better. The best results achieved in homework 1 were around 88%, while the neural net can achieve 96% accuracy. Figures 1-3, on the following pages summarize these results.

| | | Predicted Class | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Actual Class | **0** | 962 | 0 | 0 | 1 | 0 | 4 | 5 | 5 | 2 | 1 |
| | **1** | 0 | 1118 | 1 | 5 | 0 | 1 | 3 | 0 | 5 | 2 |
| | **2** | 9 | 5 | 932 | 19 | 5 | 4 | 12 | 7 | 35 | 4 |
| | **3** | 8 | 1 | 14 | 915 | 0 | 19 | 3 | 4 | 40 | 6 |
| | **4** | 1 | 1 | 7 | 0 | 883 | 0 | 14 | 3 | 10 | 63 |
| | **5** | 13 | 1 | 3 | 31 | 2 | 782 | 15 | 7 | 31 | 7 |
| | **6** | 17 | 2 | 4 | 1 | 5 | 18 | 894 | 1 | 16 | 0 |
| | **7** | 1 | 8 | 8 | 5 | 1 | 1 | 0 | 946 | 14 | 44 |
| | **8** | 11 | 5 | 2 | 14 | 7 | 10 | 8 | 1 | 903 | 13 |
| | **9** | 8 | 7 | 0 | 14 | 12 | 2 | 0 | 7 | 15 | 944 |

(a) Confusion Matrix

.



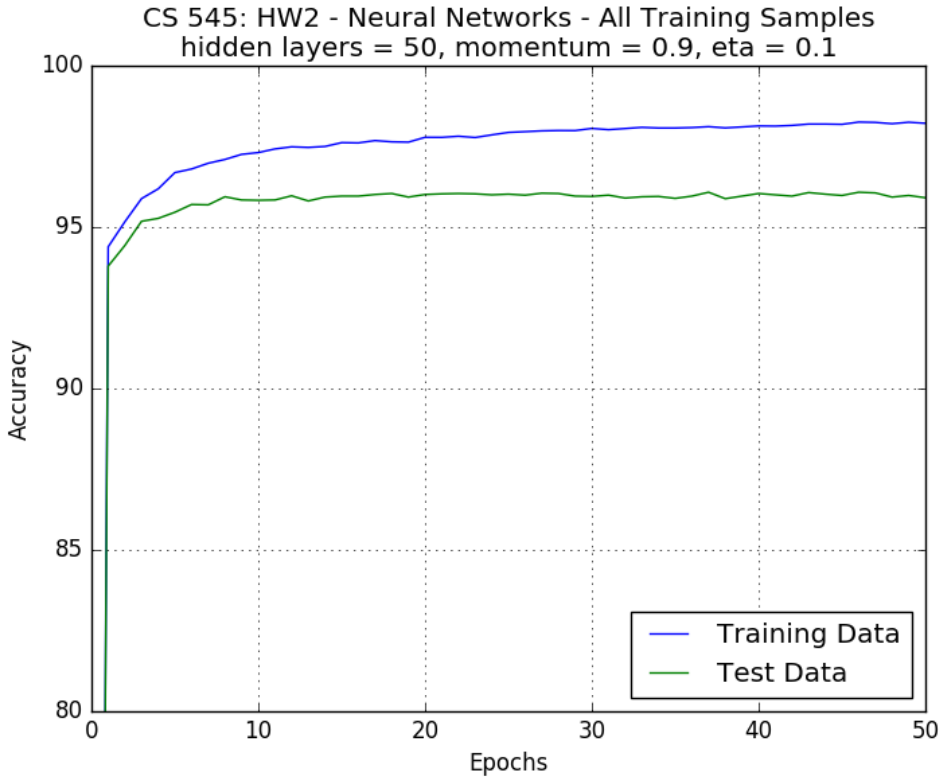

(b) Accuracy over epochs

Fig. 1: *hidden nodes* $= 20, \alpha = 0.9, \eta = 0.1$ – (a) Confusion matrix after 50 epochs. (b) The training data and test data are measure for accuracy after each epoch. In this case oscillations are seen in the training and test data. As the number of epochs increases, there is an increase in the amount of overfitting.

| | | Predicted Class | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Actual Class | **0** | 960 | 0 | 1 | 1 | 0 | 5 | 5 | 2 | 6 | 0 |
| | **1** | 0 | 1119 | 1 | 5 | 1 | 1 | 2 | 1 | 5 | 0 |
| | **2** | 3 | 4 | 982 | 7 | 2 | 0 | 3 | 10 | 19 | 2 |
| | **3** | 2 | 0 | 6 | 973 | 0 | 9 | 1 | 4 | 12 | 3 |
| | **4** | 1 | 0 | 6 | 1 | 943 | 0 | 2 | 0 | 3 | 26 |
| | **5** | 4 | 1 | 2 | 19 | 0 | 831 | 6 | 3 | 21 | 5 |
| | **6** | 11 | 3 | 3 | 0 | 1 | 5 | 919 | 0 | 14 | 2 |
| | **7** | 0 | 4 | 16 | 3 | 4 | 0 | 0 | 976 | 10 | 15 |
| | **8** | 9 | 2 | 4 | 2 | 4 | 7 | 3 | 4 | 933 | 6 |
| | **9** | 2 | 6 | 2 | 9 | 16 | 3 | 1 | 6 | 9 | 955 |

(a) Confusion Matrix

.



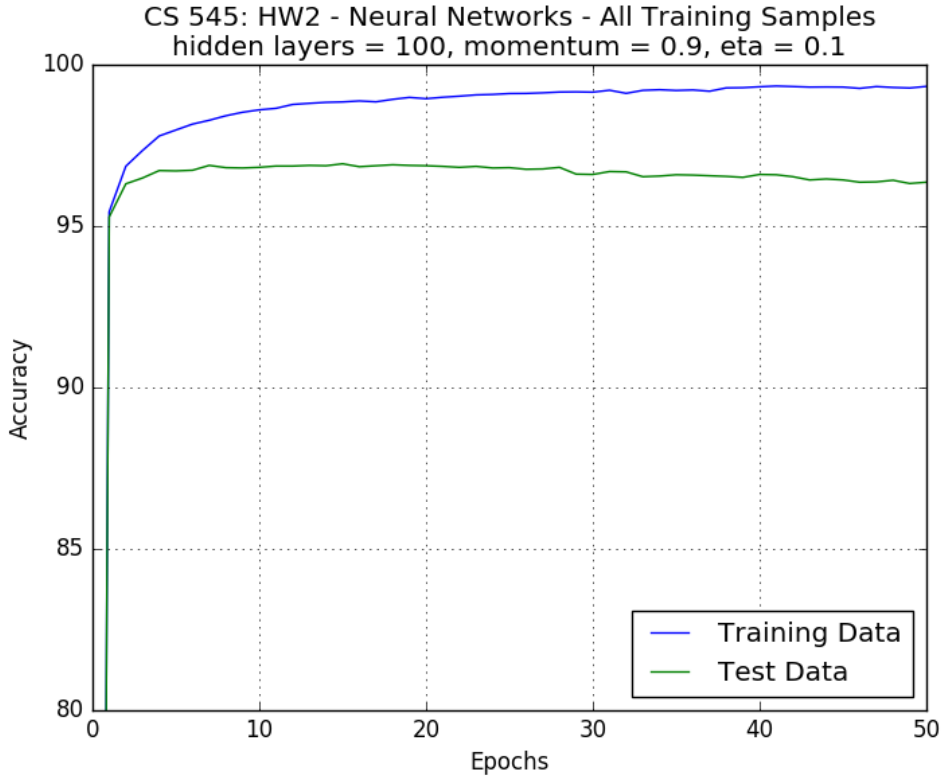

(b) Accuracy over epochs

Fig. 2: *hidden nodes* $= 50, \alpha = 0.9, \eta = 0.1$ – (a) Confusion matrix after 50 epochs. (b) In this case oscillations are small. The test accuracy seems to level out around 11 epochs, but the training accuracy steadily increases; this indicates overfitting.

| | | Predicted Class | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Actual Class | **0** | 970 | 1 | 1 | 1 | 0 | 1 | 3 | 1 | 1 | 1 |
| | **1** | 0 | 1123 | 5 | 2 | 0 | 2 | 0 | 0 | 3 | 0 |
| | **2** | 5 | 1 | 988 | 6 | 1 | 2 | 3 | 7 | 16 | 3 |
| | **3** | 0 | 1 | 5 | 967 | 0 | 14 | 0 | 3 | 8 | 12 |
| | **4** | 1 | 0 | 4 | 0 | 921 | 1 | 4 | 1 | 4 | 46 |
| | **5** | 4 | 1 | 0 | 13 | 0 | 849 | 8 | 3 | 9 | 5 |
| | **6** | 6 | 3 | 3 | 1 | 1 | 9 | 926 | 0 | 8 | 1 |
| | **7** | 2 | 5 | 11 | 3 | 3 | 0 | 0 | 982 | 6 | 16 |
| | **8** | 4 | 2 | 1 | 4 | 4 | 7 | 3 | 3 | 941 | 5 |
| | **9** | 5 | 6 | 1 | 5 | 8 | 1 | 0 | 4 | 10 | 969 |

(a) Confusion Matrix

.




(b) Accuracy over epochs

Fig. 3: *hidden nodes* $= 100, \alpha = 0.9, \eta = 0.1$ – (a) Confusion matrix after 50 epochs. (b) The training data and test data are measure for accuracy after each epoch. There does not appear to be any oscillations, but there is very noticeable overfitting. In this instance the test accuracy slowly get worse while the training accuracy gets better.

## Experiment 2: Vary the momentum value

In experiment 2, the momentum value, $\alpha$, is swept from 0, 0.25, and 0.5 while keeping the learning rate at 0.9 and the number of hidden units at 100 (the last test from experiment 1 is also included in this discussion). It's difficult to immediately tell what effect the momentum has on the accuracy of the neural network. It seems like the neural nets with smaller momentum seem to converge slightly slower but do so more smoothly. It's not obvious how this affects the number of epochs needed to converge. In the case momentum=0 and 0.25 it seems like it took more epochs to get higher accuracy. However, in the case of momentum=0.5 and 0.9, the best accuracy was after only a dozen or so epochs before it started to get worse. In all cases, overfitting can be seen form the fact that the test accuracy stops increasing (or decreases) while the training accuracy continues to increase. Figures 4-7 on the following pages summarize these results.

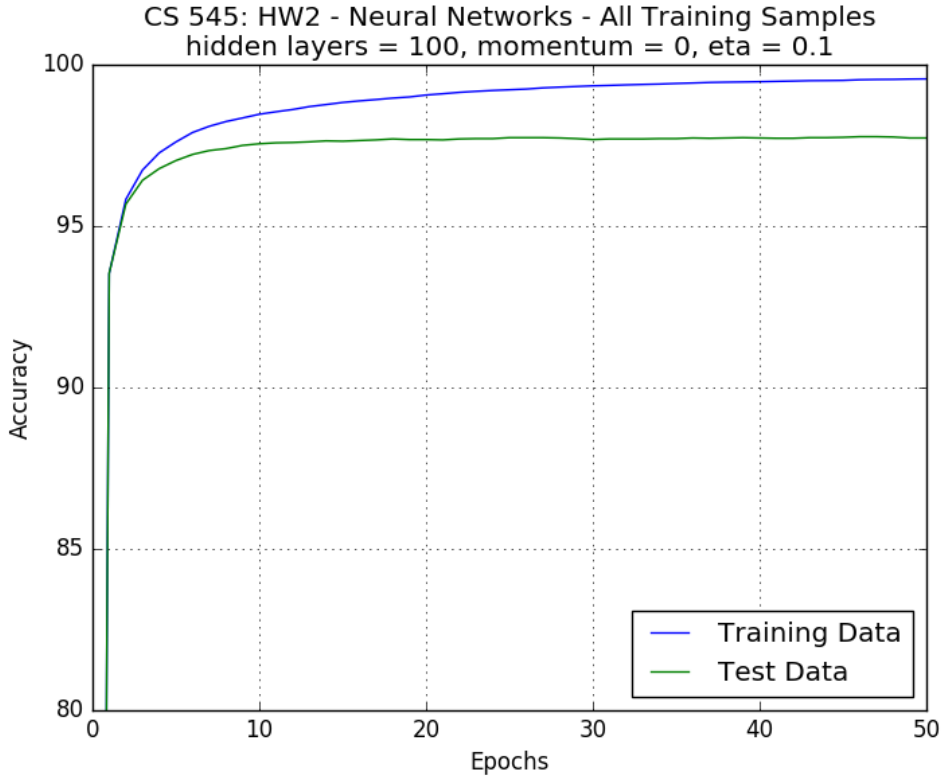
## Experiment 3: Vary the number of training examples

For experiment 3, the number of training samples was limited to 25% and 50%. By reducing the size of the training set, the accuracy of the test data was reduced. However, in both cases, the training accuracy was much higher (above 99%). This indicates that neural net was highly overfit. In both cases, the accuracy of the test data reached it's peak at about 12 epochs. After that point, an increase in epochs resulted in worse test accuracy. Figures 8 and 9 on the following pages display these results.

| | | Predicted Class | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Actual Class | **0** | 973 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 4 | 0 |
| | **1** | 0 | 1123 | 3 | 4 | 0 | 1 | 2 | 0 | 2 | 0 |
| | **2** | 3 | 2 | 1005 | 4 | 1 | 1 | 4 | 5 | 5 | 2 |
| | **3** | 0 | 0 | 3 | 987 | 0 | 9 | 0 | 2 | 4 | 5 |
| | **4** | 1 | 0 | 0 | 1 | 960 | 0 | 5 | 0 | 1 | 14 |
| | **5** | 2 | 0 | 0 | 12 | 0 | 860 | 6 | 1 | 6 | 5 |
| | **6** | 5 | 3 | 1 | 0 | 2 | 2 | 942 | 0 | 3 | 0 |
| | **7** | 2 | 5 | 10 | 2 | 1 | 1 | 0 | 998 | 2 | 7 |
| | **8** | 5 | 1 | 3 | 4 | 5 | 4 | 2 | 3 | 944 | 3 |
| | **9** | 2 | 6 | 0 | 5 | 6 | 0 | 0 | 5 | 4 | 981 |

(a) Confusion Matrix

.



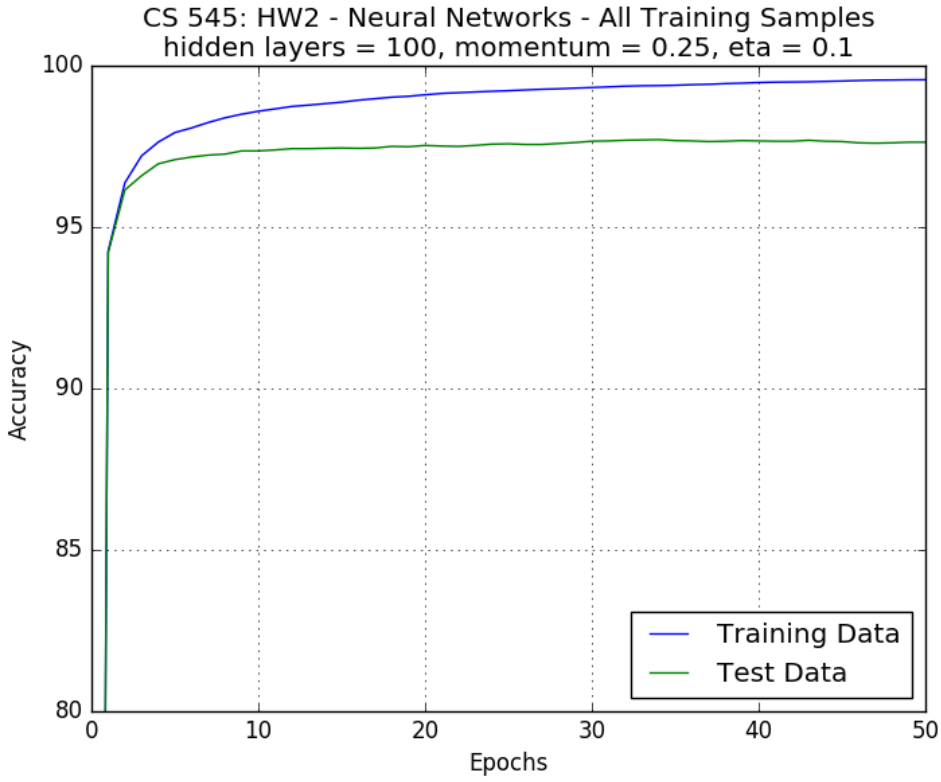

(b) Accuracy over epochs

Fig. 4: *hidden nodes* $= 100, \alpha = 0, \eta = 0.1$ – (a) Confusion matrix after 50 epochs. (b) The training data and test data are measure for accuracy after each epoch. With the momentum set to 0, the accuracy of the training and test sets smoothly increase. Overfitting becomes more noticeable as the epochs increase and the test and training accuracy diverge.

| | | Predicted Class | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Actual Class | **0** | 973 | 0 | 0 | 1 | 0 | 1 | 2 | 1 | 2 | 0 |
| | **1** | 0 | 1126 | 2 | 2 | 0 | 1 | 2 | 0 | 2 | 0 |
| | **2** | 7 | 2 | 998 | 6 | 4 | 0 | 3 | 6 | 6 | 0 |
| | **3** | 0 | 0 | 2 | 992 | 0 | 4 | 1 | 4 | 3 | 4 |
| | **4** | 2 | 0 | 4 | 0 | 960 | 0 | 4 | 0 | 1 | 11 |
| | **5** | 2 | 0 | 0 | 8 | 2 | 861 | 6 | 2 | 8 | 3 |
| | **6** | 6 | 3 | 2 | 2 | 2 | 3 | 937 | 0 | 3 | 0 |
| | **7** | 0 | 3 | 12 | 2 | 3 | 0 | 0 | 990 | 6 | 12 |
| | **8** | 5 | 1 | 1 | 2 | 3 | 4 | 4 | 2 | 947 | 5 |
| | **9** | 3 | 6 | 0 | 5 | 9 | 0 | 0 | 5 | 2 | 979 |

(a) Confusion Matrix

.



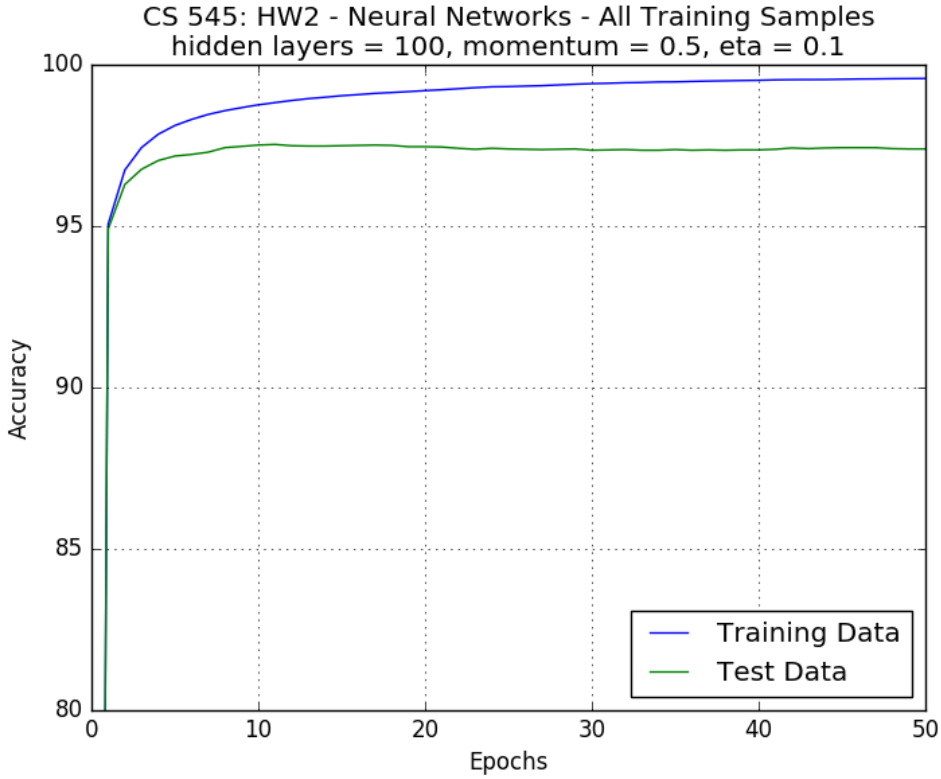

(b) Accuracy over epochs

Fig. 5: *hidden nodes* $= 100, \alpha = 0.25, \eta = 0.1$ – (a) Confusion matrix after 50 epochs. (b) The training data and test data are measure for accuracy after each epoch. The test data continues to increase in accuracy with more epochs, but the training accuracy is diverging at a faster rate.

| | | Predicted Class | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Actual Class | **0** | 970 | 1 | 1 | 0 | 0 | 1 | 3 | 1 | 3 | 0 |
| | **1** | 0 | 1123 | 1 | 4 | 0 | 1 | 3 | 1 | 2 | 0 |
| | **2** | 4 | 3 | 1003 | 3 | 3 | 0 | 3 | 5 | 7 | 1 |
| | **3** | 2 | 1 | 4 | 988 | 0 | 5 | 0 | 3 | 3 | 4 |
| | **4** | 0 | 0 | 2 | 1 | 961 | 0 | 3 | 0 | 1 | 14 |
| | **5** | 3 | 0 | 0 | 15 | 0 | 853 | 9 | 1 | 7 | 4 |
| | **6** | 7 | 3 | 1 | 1 | 2 | 8 | 932 | 0 | 4 | 0 |
| | **7** | 1 | 4 | 8 | 3 | 5 | 0 | 0 | 986 | 4 | 17 |
| | **8** | 5 | 2 | 1 | 2 | 4 | 2 | 4 | 4 | 945 | 5 |
| | **9** | 4 | 4 | 0 | 6 | 8 | 3 | 1 | 4 | 1 | 978 |

(a) Confusion Matrix

.



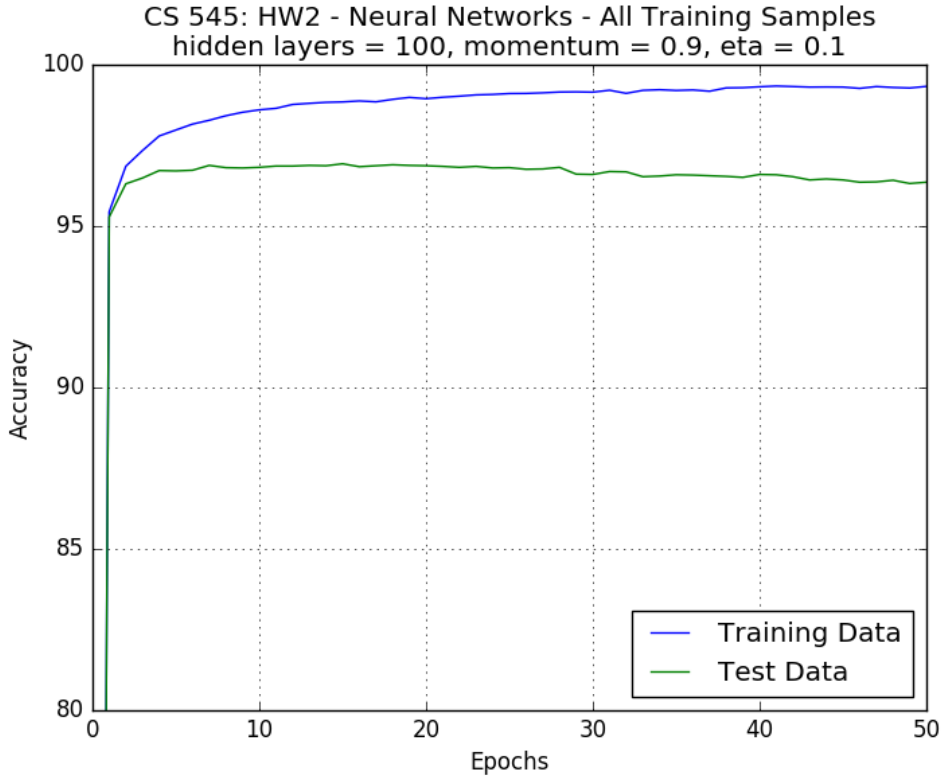

(b) Accuracy over epochs

Fig. 6: *hidden nodes* $= 100, \alpha = 0.5, \eta = 0.1$ – (a) Confusion matrix after 50 epochs. (b) The training data and test data are measure for accuracy after each epoch. The test accuracy quickly converges, but then begins to decrease slightly. The test accuracy continues to increase and indicates overfitting as the epochs are increased.

| | | Predicted Class | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Actual Class | **0** | 970 | 1 | 1 | 1 | 0 | 1 | 3 | 1 | 1 | 1 |
| | **1** | 0 | 1123 | 5 | 2 | 0 | 2 | 0 | 0 | 3 | 0 |
| | **2** | 5 | 1 | 988 | 6 | 1 | 2 | 3 | 7 | 16 | 3 |
| | **3** | 0 | 1 | 5 | 967 | 0 | 14 | 0 | 3 | 8 | 12 |
| | **4** | 1 | 0 | 4 | 0 | 921 | 1 | 4 | 1 | 4 | 46 |
| | **5** | 4 | 1 | 0 | 13 | 0 | 849 | 8 | 3 | 9 | 5 |
| | **6** | 6 | 3 | 3 | 1 | 1 | 9 | 926 | 0 | 8 | 1 |
| | **7** | 2 | 5 | 11 | 3 | 3 | 0 | 0 | 982 | 6 | 16 |
| | **8** | 4 | 2 | 1 | 4 | 4 | 7 | 3 | 3 | 941 | 5 |
| | **9** | 5 | 6 | 1 | 5 | 8 | 1 | 0 | 4 | 10 | 969 |

(a) Confusion Matrix

.



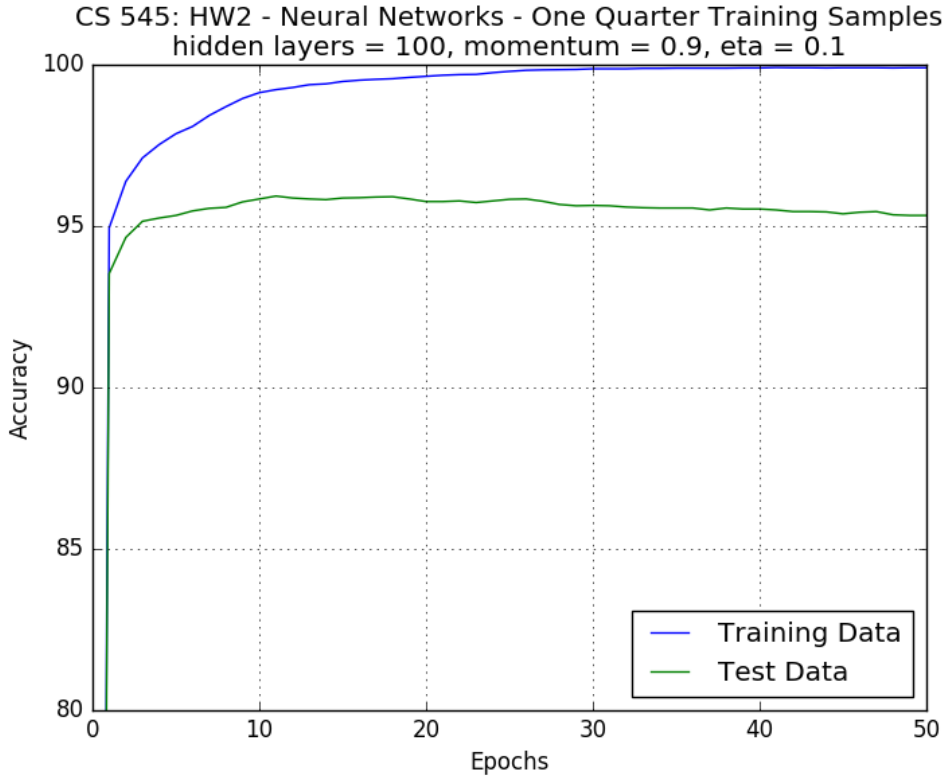

(b) Accuracy over epochs

Fig. 7: *hidden nodes* $= 100, \alpha = 0.9, \eta = 0.1$ – (a) Confusion matrix after 50 epochs. (b) The training data and test data are measure for accuracy after each epoch. There does not appear to be any oscillations, but there is very noticeable overfitting. In this instance the test accuracy slowly get worse while the training accuracy gets better.

| | | Predicted Class | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Actual Class | **0** | 966 | 0 | 1 | 2 | 0 | 1 | 3 | 3 | 3 | 1 |
| | **1** | 1 | 1115 | 6 | 3 | 0 | 1 | 3 | 0 | 6 | 0 |
| | **2** | 6 | 2 | 978 | 5 | 2 | 1 | 5 | 14 | 16 | 3 |
| | **3** | 2 | 0 | 13 | 953 | 0 | 8 | 1 | 7 | 19 | 7 |
| | **4** | 1 | 2 | 5 | 0 | 917 | 0 | 8 | 1 | 7 | 41 |
| | **5** | 7 | 1 | 2 | 23 | 0 | 816 | 12 | 5 | 22 | 4 |
| | **6** | 11 | 4 | 2 | 0 | 6 | 7 | 912 | 0 | 14 | 2 |
| | **7** | 1 | 3 | 10 | 4 | 5 | 0 | 2 | 985 | 10 | 8 |
| | **8** | 10 | 1 | 2 | 5 | 4 | 4 | 3 | 5 | 932 | 8 |
| | **9** | 5 | 4 | 2 | 4 | 9 | 3 | 0 | 9 | 14 | 959 |

(a) Confusion Matrix

.



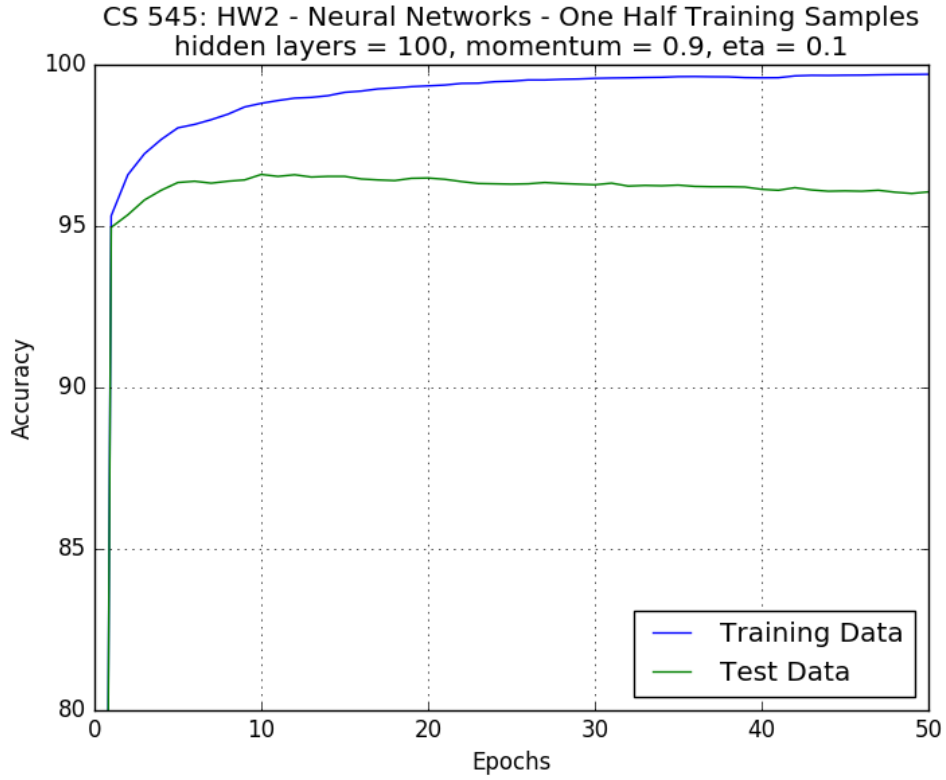

(b) Accuracy over epochs

Fig. 8: 25% *sample size*, *hidden nodes* $= 100, \alpha = 0.5, \eta = 0.1$ – (a) Confusion matrix after 50 epochs. (b) The training data and test data are measure for accuracy after each epoch. The test accuracy quickly converges, but then begins to decrease slightly. The test accuracy continues to increase and indicates overfitting as the epochs are increased.

| | | Predicted Class | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Actual Class | **0** | 960 | 1 | 4 | 1 | 0 | 1 | 3 | 1 | 3 | 6 |
| | **1** | 0 | 1115 | 1 | 2 | 1 | 1 | 5 | 3 | 7 | 0 |
| | **2** | 6 | 2 | 984 | 5 | 3 | 1 | 5 | 10 | 14 | 2 |
| | **3** | 0 | 0 | 11 | 972 | 1 | 7 | 0 | 3 | 10 | 6 |
| | **4** | 2 | 0 | 0 | 0 | 937 | 0 | 6 | 2 | 5 | 30 |
| | **5** | 4 | 0 | 0 | 21 | 1 | 835 | 7 | 3 | 17 | 4 |
| | **6** | 7 | 3 | 3 | 1 | 1 | 6 | 927 | 0 | 9 | 1 |
| | **7** | 1 | 4 | 12 | 3 | 3 | 1 | 0 | 983 | 8 | 13 |
| | **8** | 5 | 1 | 3 | 4 | 6 | 2 | 5 | 2 | 939 | 7 |
| | **9** | 5 | 5 | 1 | 7 | 11 | 2 | 0 | 7 | 17 | 954 |

(a) Confusion Matrix

.




(b) Accuracy over epochs

Fig. 9: 50% *sample size*, *hidden nodes* $= 100, \alpha = 0.5, \eta = 0.1$ – (a) Confusion matrix after 50 epochs. (b) The training data and test data are measure for accuracy after each epoch. The test accuracy quickly converges, but then begins to decrease slightly. The test accuracy continues to increase and indicates overfitting as the epochs are increased.