**Programming Assignment 2**

**Theory:**

Y

Z3

H2

H1

Z1

Z2

X

We have the equations for each layer as follow:

To compute the gradient for each weight, we can do the following:

Note that

Thus, from matrix multiplication and simplification

Where T is the true label

Next

From these, we get

The derivative of ReLU function is

Where

Also,

So, we have

Similar gradients can be derived for the first weights and bias corresponding to the input data and first hidden layer

**Model:**

X

Z1

H1

Z2

H2

Z3

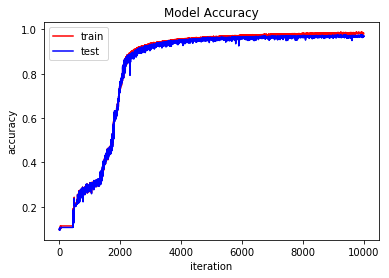
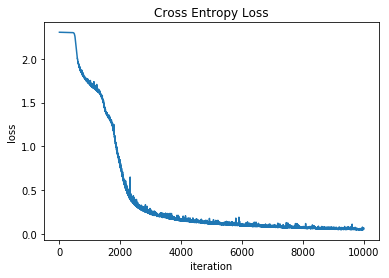
Y

This model takes in a mini batch of 50 figures with each figure contains a 784-vector of values between 0 and 1. It has 2 hidden layers (100 nodes each) with ReLU activation function and outputs a 10-vector of the classification probability for each digit by a softmax activation function.

For the experimental setting, I initialize the weights to be small positive values and the bias to be zeros. The batch size is 50.

For the hyperparameters, I use learning rate of 0.1 and 10 epochs where each epoch contains 1000 iterations of 50 different figures that are passed in and trained in the model.

The training and testing accuracy and the loss plots are shown below.



The classification error is shown below.

