Name:

NetID:

1.) Given the cost function $\mathbf{J(\theta_1, \theta_2, \theta_3) = \theta_1^3 - \theta_2^2 - 3\theta_3^4 + 1}$. Perform 2 steps of gradient descent starting at $\mathbf{\theta_1 = 2, \theta_2 = 1, and \theta_3 = 3}$. Use a **learning rate of 0.1** (5 pts.)

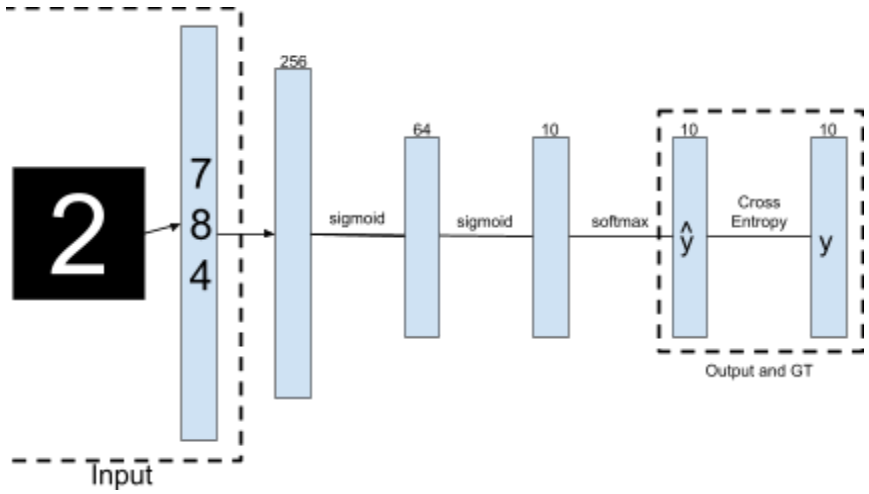
2.) Why can't we use gradient descent for the step function? (1 pt.)

3.) What is the gradient we're descending when we use gradient descent? What are we trying to optimize and what do we take the partial derivatives with respect to to do so? (2 pts.)

4.) What is grid search and what do we use it for? (2 pt.)

5.) How are gradient descent and backprop related? (2 pts.)

6.) Given the network diagram, answer the following questions.



For reference:

The derivative of the sigmoid is: $\sigma(x)(1 - \sigma(x))$
The derivative of Cross-Entropy Loss and Softmax is: $\hat{y} - y$

a.) Draw the computation graph for the network. (5 pts.)

b.) Give the derivative chain for calculating the gradient of the bias in the first layer. (2 pts.)

$$\frac{\partial \mathcal{L}}{\partial b} =$$

c.) Give the equation you'd use to actually calculate the gradients. (5 pts.)

7.) What type of data are RNNs/LSTMs/GRUs used for? (1 pt.)

8.) Write the “equations” for a fully connected layer and an RNN. Explain the reasoning behind the difference between the two. (3 pts.)

9.) What are vanishing and exploding gradients? Why do they occur? (2 pts.)

Bonus.) Next semester my current plan is to get rid of the quizzes and go from 10 homeworks to 5, combining the quizzes and homeworks. What are your thoughts? (1 bonus pt.)