Name:

NetID:

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

$\frac{\partial J}{\partial \theta_1} = 3\theta_1^2$	$\frac{\partial J}{\partial \theta_2} = -2\theta_2$
1) $3(2)^2 = 12$ $\theta_1 = 2 - (0.1)(12)$ $\theta_1 = 0.8$	$2(1) = 2$ $\theta_2 = 1 + (0.1)(2)$ $\theta_2 = 1.2$
$3(0.8)^2 = 1.92$ $\theta_1 = 0.8 - (0.1)(1.92)$ $\theta_1 = 0.608$	$2(1.2$ $\theta_2 = 1.2 + (0.1)(2.4)$ $\theta_2 = 1.44$

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

It's not differentiable

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.)

Loss function WRT weights

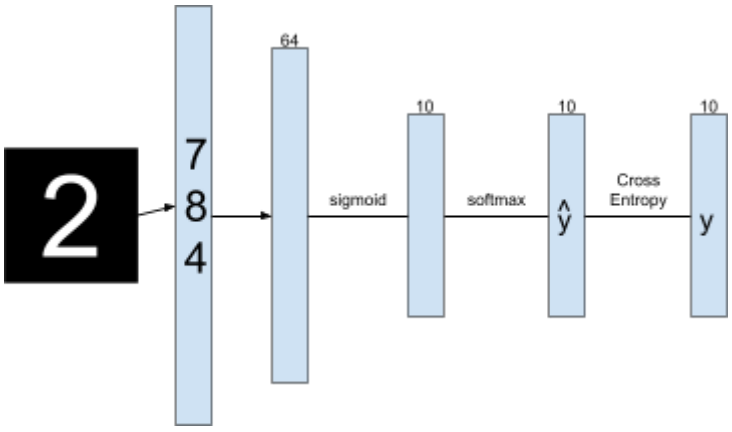
4.) What is grid search and what do we use it for? (2 pt.)

hyperparameter testing

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

backprop is used for efficient gradient calculation so we can use gradient descent

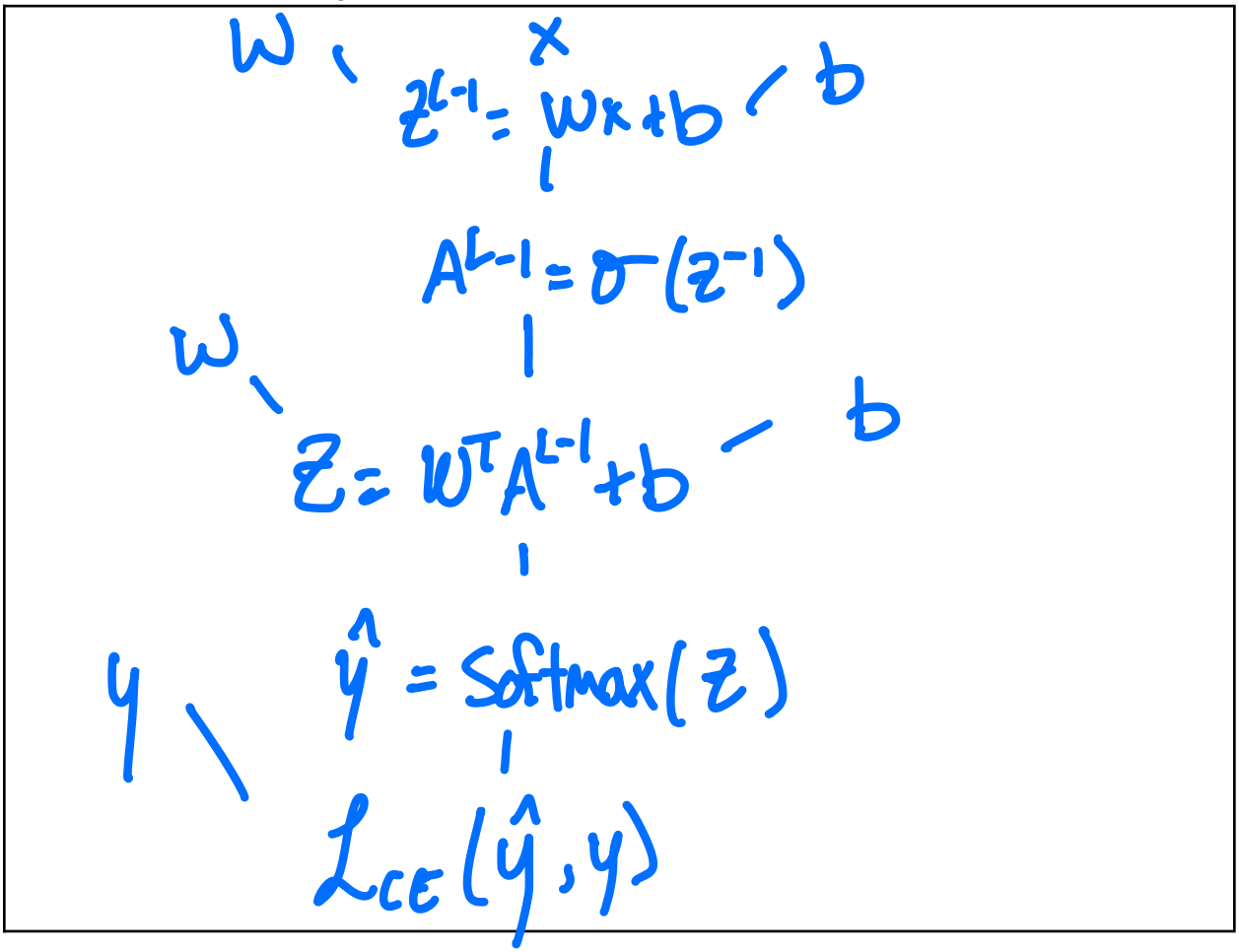
6.) Given the network diagram, answer the following questions.



For reference:

The derivative of the sigmoid is: σ(x)(1 - σ(x))
The derivative of Cross-Entropy Loss and Softmax is: ŷ - 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.)

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial A^{L-1}} \frac{\partial A^{L-1}}{\partial z^{-1}} \frac{\partial z^{-1}}{\partial b^{-1}}$$

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

$$\frac{\partial L}{\partial b} = (\hat{y} - y)\, w\, \sigma(z^{-1})(1 - \sigma(z^{-1}))$$

(1)

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

Sequential data

e.g. text

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

$W^T X + b$

$W^T X + W^T h + b$

Context Matrix for sequences

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

Vanishing goes to 0,
exploding to ∞

too small or big gradients

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.)

Cool!