Name:

NetID:

Note: This quiz will be a set of practice problems for the upcoming exam. Complete this at your leisure outside of class but please bring in an **INDIVIDUAL** copy of the quiz **ON PAPER** to turn in at the start of the exam. While doing the work feel free to work with your classmates, but I'd like you all to turn in a copy individually. You don't necessarily need to print and fill out this exact document, just writing the answers on your own piece of paper will be fine.

1.) Given the cost function **$J(\theta_1, \theta_2, \theta_3) = \theta_1^5 - \theta_2^3 - \theta_3^2 - 5$**. Perform 2 steps of gradient descent starting at **$\theta_1 = 2$** and **$\theta_2 = 1$** and **$\theta_3 = 3$**.

$\alpha = 0.1$

$\frac{\partial J}{\partial \theta_1} = 5\theta_1^4$	$\frac{\partial J}{\partial \theta_2} = -3\theta_2^2$	$\frac{\partial J}{\partial \theta_3} = -2\theta_3$
$\theta_1 = 2$ $5(2)^4 = 80$ $\theta_1 = 2 - (0.1)(80)$ $\theta_1 = -6$	$\theta_2 = 1$ $-3(1)^2 = -3$ $\theta_2 = 1 - (0.1)(-3)$ $\theta_2 = 1.3$	$\theta_3 = 3$ $-2(3) = -6$ $\theta_3 = 3 - (0.1)(-6)$ $\theta_3 = 3.6$
$5(-6)^4 = 6480$ $\theta_1 = -6 - (0.1)(6480)$ $\boxed{\theta_1 = -654}$	$-3(1.3)^2 = -5.07$ $\theta_2 = 1.3 - (0.1)(-5.07)$ $\boxed{\theta_2 = 1.807}$	$-2(3.6) = -7.2$ $\theta_3 = 3.6 - (0.1)(-7.2)$ $\boxed{\theta_3 = 4.32}$

2.) Explain why are sigmoids particularly prone to vanishing gradients, especially in the case of RNNs

Sigmoid gradient is between 0 and 0.25 so very tiny. RNNs require backprop over timesteps and therefore increase tinyness

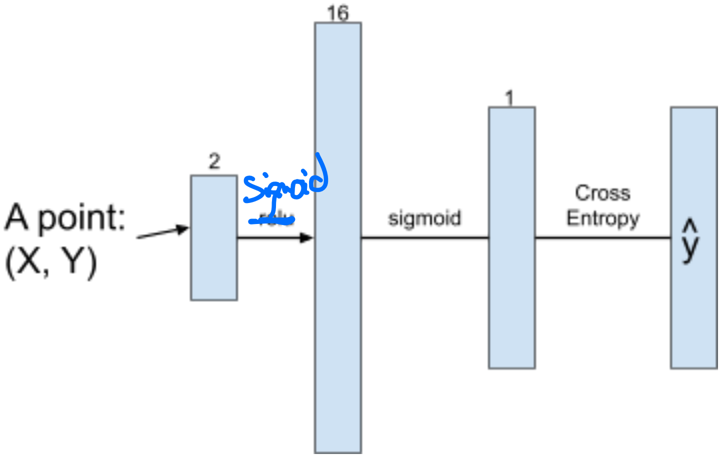
3.) What is the gradient we're descending when we use gradient descent? What are we trying to optimize and what do we take the partial derivatives with respect to to do so?

See Exam 02 solutions

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

See Exam 02 solutions

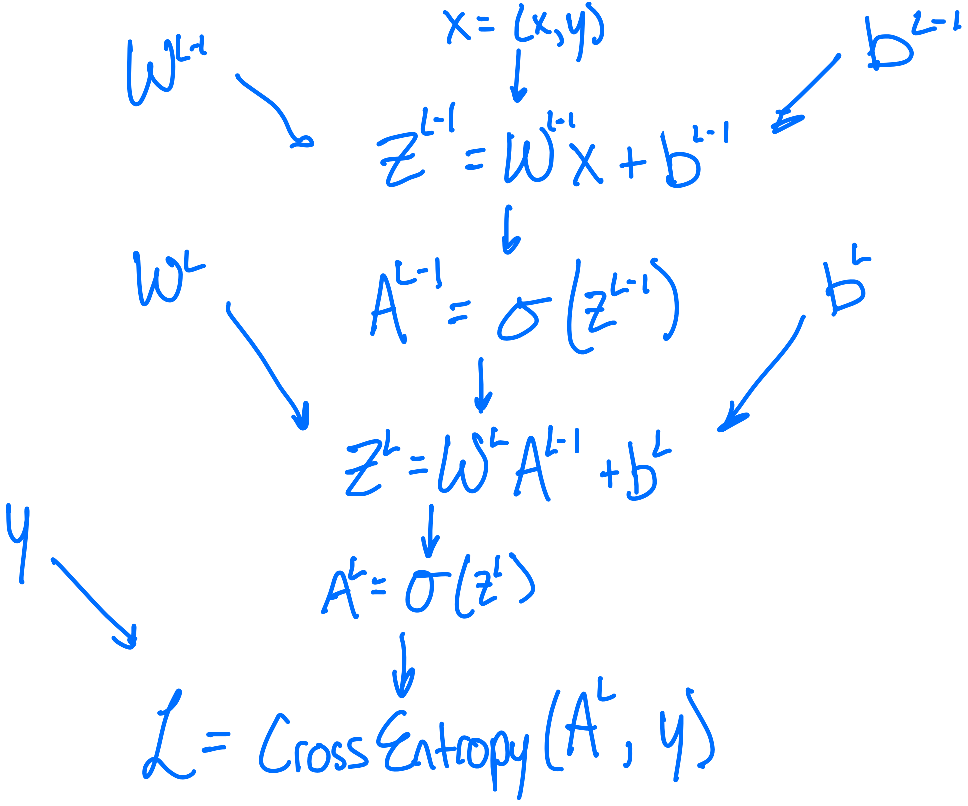
5.) Given the network diagram, answer the following questions.



For reference:

The derivative of the sigmoid is: σ(x)(1 - σ(x))
The derivative of Cross-Entropy Loss is: -((Y / ŷ) - ((1 - Y) / (1 - ŷ)))

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



W^{L-1}
$X = (x, y)$
b^{L-1}
$Z^{L-1} = W^{L-1}X + b^{L-1}$
W^{L}
$A^{L-1} = \sigma(Z^{L-1})$
b^{L}
$Z^{L} = W^{L}A^{L-1} + b^{L}$
y
$A^{L} = \sigma(Z^{L})$
$\mathcal{L} = \text{CrossEntropy}(A^{L}, y)$

b.) Give the derivative chain for calculating the gradient of the bias in the ~~hidden~~ 1st hidden layer. (2 pts.)

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

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

$$\frac{\partial L}{\partial b^{L-1}} = -\underbrace{\left(\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right)} \cdot \underbrace{\sigma(z^{L})(1-\sigma(z^{L}))}$$

$$\cdot \underbrace{W^{L}} \cdot \underbrace{\sigma(z^{L-1})(1-\sigma(z^{L-1}))}$$

$$\cdot \underbrace{1}$$

6.) What's an example of sequential data and how do RNNs capture the context?

a string "I like cats"

they use a hidden "context" matrix that they pass over timesteps

7.) Why do we call them **recurrent** neural networks? What makes them “recurrent”?

Because they have a "loop"
and read information over time

Bonus.) Next semester my current plan is to drop exam 03, I think it’s too tight to fit everything in. What are your thoughts on that?