CS 480 - Backpropagation Workshop **Name:** ______________________

1. Make a computation graph for the function $f(x,y) = (3x + xy)^3$. Then use the backpropagation algorithm to find the gradient ∇f at $(x,y) = (2,-2)$. Hint: *When you get the end of the backpropagation algorithm, the multivariable chain rule says that if both a and b are functions of x, then*

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial a}\frac{\partial a}{\partial x} + \frac{\partial f}{\partial b}\frac{\partial b}{\partial x}.$$

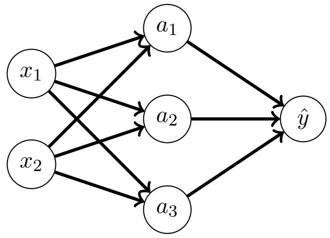
2. Make a computation graph for the following loss function. Label the edges with the partial derivatives for each step in the computation.

$$L(\mathbf{w}) = \max(0, 1 - \text{margin}), \ \textit{where}$$

$$\text{margin} = y \cdot \text{score}, \ \textit{and}$$

$$\text{score} = \mathbf{w} \cdot \mathbf{x}.$$

3. Consider the neural network with one hidden layer shown below:



$$\mathbf{a} = \text{ReLU}(W^{(1)}\mathbf{x} + \mathbf{b}^{(1)})$$

$$\hat{y} = \sigma(W^{(2)}\mathbf{a} + \mathbf{b}^{(2)})$$

where $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ is an input data vector, $\hat{y}$ is a prediction (between 0 and 1), and σ is the sigmoid activation function $\sigma(x) = \dfrac{1}{1+e^{-x}}$.

(a) What are the dimensions of the matrices $W^{(1)}$ $W^{(2)}$, and the vectors $\mathbf{b}^{(1)}$, $\mathbf{b}^{(2)}$?

(b) How many (real valued) parameters does this model have?

(c) One reason that the sigmoid function is a popular choice for the activation function of the final layer in a neural network is because the final output can be interpreted as a probability. If that is the case, then how should we interpret the output of $W^{(2)}\mathbf{a} + \mathbf{b}^{(2)}$ before we apply the sigmoid function? Hint: *Another formula for* σ *is* $\sigma(x) = \dfrac{e^x}{e^x+1}$. *Where have we seen that formula, and what did* x *represent then?*

(d) **Extra credit.** Suppose we want to train our neural network using several different data vectors $\mathbf{x}^{(i)}$ and corresponding classifications $y^{(i)}$ where each $y^{(i)}$ is either 1 or 0. We can use the cross entropy loss function (another name for the logistic loss function) which is the sum of the loss from each observation:

$$L^{(i)} = -y^{(i)}\log(\hat{y}^{(i)}) - (1-y^{(i)})\log(1-\hat{y}^{(i)}).$$

Make a computation graph for this loss function (for one i). Hint: *You'll need a big blank sheet of paper!*