# Assignment Seven
# ECE 4200/5240

- Provide credit to **any sources** other than the course staff that helped you solve the problems. This includes **all students** you talked to regarding the problems.
- You can look up definitions/basics online (e.g., wikipedia, stack-exchange, etc)
- **The due date is 11/12/2021, 23.59.59 eastern time**.
- Submission rules are the same as previous assignments.

**Problem 1. (10 points).** The tanh function is $\tanh(y) = (e^y - e^{-y})/(e^y + e^{-y})$. Consider the function $\tanh(w_0 + w_1x_1 + w_2x_2)$, with five inputs, and a scalar output.

1. Draw the computational graph of the function (you can use tanh in your computation graph).
2. What is the derivative of $\tanh(y)$ with respect to $y$.
3. Suppose $(w_0, w_1, w_2, x_1, x_2) = (-2, -3, 1, 2, 3)$. Compute the forward function values, and back-propagation of gradients.

**Problem 2. (15 points).** Consider one layer of a ReLU network. The feature vector is $d$ dimensional $\overrightarrow{x}$. The linear transformation is a $m \times d$ dimensional matrix $W$. The output of the ReLU network is a $m$ dimensional vector $y$ given by $\max\{\mathbf{0}, W\overrightarrow{x}\}$. This is a component-wise max function.

- Suppose $\overrightarrow{x}$ is fixed, and all its entries are non-zero.
- Suppose the entries in $W$ are all independent, and distributed accoding to a Gaussian distribution with mean 0, and standard deviation 1 (a $N(0,1)$ distribution).

1. Show that the expected number of non-zero entries in the output is $m/2$.
2. Suppose $\|\overrightarrow{x}\|_2^2 = \sigma^2$, what is the distribution of each entry in $Wx$ (the output before applying ReLU function)?
3. What is the mean of each entry in $y$ (after ReLU function)?

**Problem 3. (10 points).** Consider the setting as in the previous problem, with $m = 2$, and $d = 2$. Let

$$W = \begin{bmatrix} 1 & 2 \\ -2 & 3 \end{bmatrix}, \overrightarrow{x} \begin{bmatrix} 2 \\ -3 \end{bmatrix}.$$

Consider the function $L = \max\left\{\sigma(W_{(1)}\overrightarrow{x}), \sigma(W_{(2)}\overrightarrow{x})\right\}$, where $\sigma$ is the Sigmoid function and $W_{(i)}$ denotes the $i$th row of $W$. Please draw the computational graph for this function, and compute the gradients (which will be Jacobians at some nodes!).

**Problem 4. (10 points).** Given inputs $z_1, z_2 \in \mathbb{R}$, the softmax function is the following:

$$\hat{y} = \frac{e^{z_1}}{e^{z_1} + e^{z_2}}.$$

Let $y \in \{0, 1\}$, then define the cross-entropy loss between $y$ and $\hat{y}$ be

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

Prove that:

$$\frac{\partial L(y, \hat{y})}{\partial z_1} = \hat{y} - y, \frac{\partial L(y, \hat{y})}{\partial z_2} = y - \hat{y}.$$

**Problem 5. (15 points).** Consider datapoints in Figure 1: $(-2, 0), (2, 0)$ are crosses, and $(0, 2), (0, -2)$ are circles. Let the crosses be labeled $+1$, and the circles be labeled $-1$. In this problem the goal

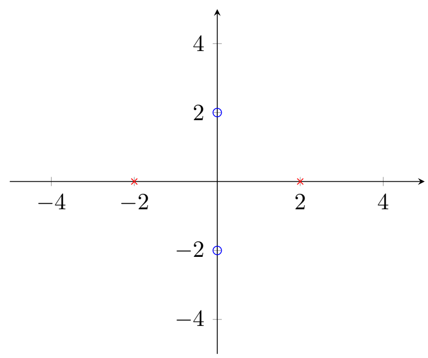
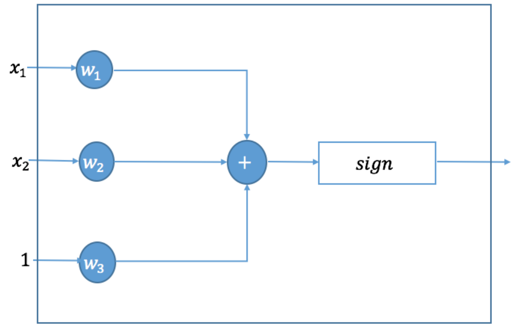



Figure 1: Neural Networks

is to design a neural network with no error on this dataset.

To make things simple, consider the following generalization. We first append a $+1$ to each input and form a new dataset as follows: $(-2, 0, 1), (2, 0, 1)$ are labeled $+1$, and $(0, 2, 1), (0, -2, 1)$ are labeled $-1$. Note that the last feature is redundant.

We consider the following basic units for our neural networks: Linear transformation followed by hard thresholding. Each unit has three parameters $w_1, w_2, w_3$. The output of the unit is the sign of the inner product of the parameters with the input.




1. Design a neural network with these units that make no error on the datapoints above. (Hint: You can take two units in the first layer, and one in the output layer, a total of three units).

2. Show that if you design a neural network with ONLY one such unit, then the points cannot be all classified correctly.

**Problem 6. (40 points).** See attached notebook for details.