Deep Learning
Assignment 1 (37 Points)

1. **The process of sampling from a discrete distribution with 2 outcomes was described on Page 29 of Lecture 2. Describe the sampling algorithm that is able to sample from a discrete distribution with K outcomes. Write a Python program to generate 100 samples from the following distribution:**
   P(X = 1) = 0.7
   P(X = 2) = 0.2
   P(X = 3) = 0.1
   Count the number of each outcome. Are they close in number to what the distribution predicts?

   (5 Points)

2. **How are Deep Learning systems different from older Machine Learning systems?**

   (2 Points)

3. **There are 4 balls in an urn. Each ball is either red or black. You start by believing that the probabilities that the urn contains 0, 1, 2, 3, 4 red balls are all equal. You then reach into the urn and pull out a ball at random. It is red. Compute the new probabilities that the urn contains 0, 1, 2, 3 or 4 red balls.**

   Hint: Use the Law of Conditional Probabilities

   (10 Points)

4. **Answer the following questions:**

(a) Derive the expression $\frac{\partial y}{\partial x}$ for the derivative of the Sigmoid Function

$$y = \frac{1}{1 + exp(-x)}$$

(2 Points)

(b) Derive the expression $\frac{\partial y_k}{\partial a_i}$ for the derivative of the Softmax Function (both for $i = k \; and \; i \neq k$)

$$y_k = \frac{exp(a_k)}{\sum_{i=1}^{K} exp(a_i)}$$

(4 Points)

(c) Using the results of Part (b), show that the derivatives $\frac{\partial \mathcal{L}}{\partial a_k}$ and $\frac{\partial \mathcal{L}}{\partial w_{kj}}$ for K-ary classification problem are given by (see Page 48 of Lecture 3 slides):

$$\frac{\partial \mathcal{L}}{\partial a_k} = (y_k - t_k) \; and \; \frac{\partial \mathcal{L}}{\partial w_{kj}} = x_j \, (y_k - t_k)$$

Hint: Use the Chain Rule for multiple variables (Page 6 of Lecture 2)

(6 Points)

(d) Show that for the case K = 2, the Softmax output Y of the Logistic Regression system, is equivalent to a system in which Y is computed using the Logistic Sigmoid.

(2 Points)

5. **Consider the Linear Model with K = 2, as described on Page 23 of Lecture 3. Suppose that the Cross Entropy Loss Function was replaced by the Mean Square Error (MSE) Loss Function (see Lecture 3, Page 11 for a definition of MSE Loss Function).**
   Compute the gradient $\frac{\partial \mathcal{L}}{\partial w_i}$ for the MSE Loss Function.
   Hint: Follow the steps shown on Lecture 3, Page 31
   (6 Points)