ST3901: Statistical Learning and Decision Making
Exercise 1

The goal of this homework is to get started in using Python for data processing. We will practice generating random data, representing functions, computing statistics from data, empirical distributions, and selecting subsets of data with conditions. Mathematically, the most important part is to work with conditional distributions. We will also practice making simple plots.

1 Generate Randome Variables and Compute Empirical Distributions

We will start with generating some data samples with a given distribution. For the following two PMFs,

```
PMF1 = np.asarray([4.0, 1.0, 6.0, 1.0, 3.0, 1.0, 5.0, 1.0])
PMF1 = PMF1 / np.sum(PMF1)
PMF2 = np.asarray([2.0, 3.0, 2.0, 5.0, 5.0, 2.0, 3.0, 2.0])
PMF2 = PMF2 / np.sum(PMF2)
```

Task 1: Generate 1000 samples from PMF1, and 500 samples from PMF2. You can either write your own random sample generator by using the inverse CDF approach we talked about in class, or use the np.random.choice() function.
Task 2: Write a function compareHIST(D, p), where D is an 1-D array of data samples, and p is a valid PMF. Compute and plot the empirical distribution of D using matplotlib.pyplot.hist(), and plot p against it in the same plot for comparison.
Task 3: Mix the two datasets generated in Task 1 into an array of 1500 samples. Compute the ensemble distribution of this mixture from PMF1 and PMF2, and compare that with the empirical distribution of the mixed dataset, by using the compareHIST() function in Task 2.

2 Data Manipulation and Conditional Distributions

Use the following code to generate labels for the two datasets generated from the previous tasks, stack with the data samples, and mix into a dataset of length 1500.

```
Labels = np.ones([1500])
Labels[1000:] = 2

Dataset = np.stack((np.concatenate((D1, D2)), Labels))
```

```
Dataset = Dataset[:, np.random.permutation(1500)]
```

Task 4: Compute the empirical distribution of the mixture dataset, and compute the correct ensemble, use compareHIST() to verify.
Task 5: Select the sub-sequences with Labels == 1 and Labels ==2. Again verify this using compareHIST().
Task 6: For each value in $x \in \{0, 1, \ldots, 7\}$, compute both the conditional probability $P_{\text{label}|X}(1|x)$ using the Bayes rule, and using the empirical distributions (by selecting the right sub-sequences). Compare your results in a single plot.

3 Convergence of Empirical Average

Task 7: Randomly select a function f by choosing $f(x)$ for $x \in \{0, \ldots, 7\}$, with whatever distribution you like.

For number of samples varying in the set steps = [10, 20, 40, 80, 160, 320, 630, 1280, 2560, 5120, 10240], generate this number of random samples from a given PMF (use PMF1 in the previous tasks). Compute the empirical average

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

Make a plot of of the empirical average as the number of samples increases, and observe the empirical average approach to the ensemble average.

Repeat this experiment 20 times, and observe that in all cases the empirical average converge to the same limit.
Task 8: Only if you know this from somewhere else, give a theoretical range of the empirical averages computed from Task 7. Plot them in the same picture. You should see something like the following.