Written by: Professor Wu
DS 4420: Machine Learning 2

# Assignment 4

**Question 1.** Load the data `complex_distribution.csv` provided on Canvas. We are going to call the original data $X$.

1) (20pt) Use KDE with distributions covered in class to find $p(x)$.

2) (20pt) Once you have $p(x)$ from KDE, you are going to
   - use $p(x)$ and rejection sampling to generate your own samples. Let's call your generated samples $\hat{X}$.
   - use $p(x)$ and KDE sampling to generate your own samples. Let's call your generated samples $\bar{X}$.
   - Plot 3 histograms of $X$, $\bar{X}$ and $\hat{X}$ side by side. All 3 distributions should look very similar.

3) (20pt) Use the original samples $X$ and approximate
   (a) Expectation, $\mathbb{E}[X]$
   (b) Population Variance, $Var[X]$ or equivalently $\mathbb{E}[(x-\mu)^2]$.
   (c) $\mathbb{E}[3x+2], \mathbb{E}[x^2-5x], \mathbb{E}[2e^x+1]$

4) (20pt) Use the samples you generated $\hat{X}$ and approximate
   (a) Expectation, $\mathbb{E}[X]$
   (b) Population Variance, $Var[X]$ or equivalently $\mathbb{E}[(x-\mu)^2]$.
   (c) $\mathbb{E}[3x+2], \mathbb{E}[x^2-5x], \mathbb{E}[2e^x+1]$

5) (20pt) Use the samples you generated $\bar{X}$ and approximate
   (a) Expectation, $\mathbb{E}[X]$
   (b) Population Variance, $Var[X]$ or equivalently $\mathbb{E}[(x-\mu)^2]$.
   (c) $\mathbb{E}[3x+2], \mathbb{E}[x^2-5x], \mathbb{E}[2e^x+1]$

**Professor's Note:** The purpose of this exercise is for you to convince yourself of some conclusions we arrived at during class.

1) If you have the samples, you know how to get $p(x)$ using KDE.
2) Conversely, if you have $p(x)$ and don't have the samples, you can generate your own samples using KDE sampling or rejection sampling.
3) Regardless if your samples came from Rejection, KDE, or the original, you should be able to approximate $\mathbb{E}[f(x)]$ for any function $f(x)$.
4) Remember that the expectation is approximately the average value where

$$\mathbb{E}[X] \approx \frac{1}{n}\sum_i x_i,$$

therefore, this tells us that

$$\mathbb{E}[f(X)] \approx \frac{1}{n}\sum_i f(x_i),$$

so if you want to calculate $\mathbb{E} = [2x-1]$, it would be

$$\mathbb{E}[2x-1] \approx \frac{1}{n}\sum_i (2x_i - 1)$$

**Question 2.** The following integral is not the easiest integral to integrate. The solution is approximately

$$\int_0^2 \ln(x+1)\, dx \approx 1.29.$$

Use a sampling technique to obtain the same result.

**Professor's Note:**

1) Remember that there are 2 ways to find the expectation/average of samples

$$\mathbb{E}[X] = \int_{\mathcal{X}} x\, p(x)\, dx \quad \text{and} \quad \mathbb{E}[X] \approx \frac{1}{n}\sum_i x_i$$

2) This implies that an integral can be approximated with

$$\mathbb{E}[X] = \underbrace{\int_{\mathcal{X}} x\, p(x)\, dx}_{\text{annoy to do}} \approx \underbrace{\frac{1}{n}\sum_i x_i}_{\text{easy to do}}.$$

and that any function can also be approximated via

$$\underbrace{\mathbb{E}[f(x)] = \int_{\mathcal{X}} f(x)\, p(x)\, dx \approx \frac{1}{n}\sum_i f(x_i).}_{\text{This observation enable us to approximate integrals by just averaging.}}$$

3) If you have a uniform distribution between the range of 0 and 2, then $p(x) = \frac{1}{2}$. Draw this out if you don't see how I got $p(x) = \frac{1}{2}$.
4) Using this uniform distribution we have

$$\mathbb{E}[f(x)] = \int_0^2 f(x)p(x)\, dx = \int_0^2 f(x)\frac{1}{2}\, dx \approx \frac{1}{n}\sum_i f(x).$$

5) This tells us that

$$\mathbb{E}[f(x)] = \int_0^2 f(x)\frac{1}{2}\, dx \approx \frac{1}{n}\sum_i f(x_i),$$

6) Implying the conclusion that

$$2\mathbb{E}[f(x)] = \underbrace{\int_0^2 f(x)\, dx}_{\text{integral of any function}} \approx \frac{2}{n}\sum_i f(x_i),$$

7) This conclusion tells us that if we randomly generate numbers between 0 and 2 with a uniform distribution. Two times the average value of $f(x_i)$ is the solution of the integral.
8) This is the beauty of mathematics, from the chaos of random numbers, we achieve order.

**Question 3.** Us the same logic and use sampling to approximate the integral

$$\int_0^1 x^2 - x + 1\, dx$$