Probability and Statistics, S2024	Posted: April 30, 2024
Project 1	
Lecturer: Hyang-Won Lee	**Due: May 8, 2024**

Notes

- Collaboration strongly encouraged, but not mandatory. If you collaborate with your fellow students, please print the name(s) of your collaborators in your solution.

1. In this project, we want to estimate the value of π through random sampling. First, consider the unit square $S = \{(x, y)|0 \leq x \leq 1, 0 \leq y \leq 1\}$ and the circle $C = \{(x, y)|(x-1/2)^2+(y-1/2)^2 \leq 1/4\}$ which resides inside the unit square. Suppose that we sample n points from the unit square S, uniformly and independently. Let X_i be 1 if the ith sample falls into the circle C and 0 otherwise. Define $S_n = X_1 + \cdots + X_n$. (이 문제에서는 무작위 샘플링을 통해 π값을 추정하길 원한다. 먼저, 단위정사각형 $S = \{(x, y)|0 \leq x \leq 1, 0 \leq y \leq 1\}$와 이 안에 원 $C = \{(x, y)|(x-1/2)^2+(y-1/2)^2 \leq 1/4\}$가 있다. 단위정사각형 S에서 균일하고 독립적으로 n개의 점을 샘플링했을 때, i 번째 샘플이 C에 들어가면 1, 그렇지 않으면 0인 확률변수 X_i를 정의하자. 또한, $S_n = X_1 + \cdots + X_n$이라 하자)

   (a) What is the PMF of X_i?

   (b) Calculate the expectation of X_i.

   (c) Calculate the mean and variance of $\frac{S_n}{n}$. What happens as n grows to infinity? Discuss what it means.

   (d) Based on your findings in (c), devise an estimator of the value of π. ((c)에서 발견한 사실을 이용하여 π값을 추정하는 방법을 기술해 보시오)

   (e) Implement the estimator in (d) by writing a python program. ((d)에서 설계한 추정기를 프로그램으로 작성해보시오)

   (f) Does your estimator approach the true value of π as the number of samples n grows? Plot the behavior of your estimator versus n and discuss the result. (위에서 만든 추정기가 실제로 π값으로 수렴합니까? n이 증가함에 따라 값이 어떻게 변하는지 그려보고, 결과를 논하시오)

2. **(optional)** In this project, we want to estimate the value of integral via sampling. Let $g(x)$ be a function of interest and suppose that we want to calculate the integral

$$\theta = \int_0^1 g(x)dx$$

Let X be a continuous random variable uniformly distributed on the interval $[0, 1]$. Then the expectation of random variable $g(X)$ is written as

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)dx = \int_0^1 g(x)dx,$$

If $X_1, ..., X_n$ be i.i.d. random variables uniformly distributed on the interval $[0, 1]$, it is clear that $g(X_1), ..., g(X_n)$ are also i.i.d. random variables. Then, we may expect the sample mean to converge to the true mean, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \rightarrow \mathbb{E}[g(X)] = \theta \text{ (as } n \rightarrow \infty)$$

One can use this fact in order to estimate the integral of a function which does not have a closed-form expression.

(a) The following integral does not have a closed-form expression, and thus has to be calculated numerically.

$$\theta = \int_0^1 g(x)dx = \int_0^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Write a computer program in python that

- Samples uniform random numbers $X_1, ..., X_n$ from $[0, 1]$
- Calculates the sample mean $\frac{1}{n}\sum_{i=1}^{n} g(X_i)$.

The above integrand is indeed the standard normal PDF, and so, its value can be found from the standard normal distribution table. Compare the sample mean and the value from standard normal distribution table. What happens as n increases? Explain why you have such results.

(b) How would you calculate the following integral via sampling?

$$\int_a^b g(x)dx$$

3. **(optional)** A randomized algorithm is said to give an (ϵ, δ)-approximation for the value of V if the output X of the algorithm satisfies (어떤 무작위 알고리듬이 있다. 이 알고리듬의 출력은 무작위성이 있고, 만약 출력 X가 다음의 부등식을 만족하면 (ϵ, δ)-근사 알고리듬이라고 부른다. 여기서 V는 알고리듬이 원래 찾고자 의도하는 값이고, 아래 부등식의 의미는 원래 찾고자하는 값에서 ϵ 이상 벗어날 확률이 δ보다 작다는 것을 의미한다. 따라서, ϵ과 δ가 작을수록 이 알고리듬은 높은 확률로 원래 찾고자 하던 값에 가까운 값을 찾게 되는 것이다)

$$\mathbb{P}(|X - V| \geq \epsilon) \leq \delta.$$

The following inequality is often a useful tool in the analysis of randomized algorithm.

- (Chebyshev Inequality) If X is a random variable with mean μ and variance σ^2, then

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2} \text{ for all } a > 0. \tag{1}$$

(a) By using the Chebyshev Inequality[1] in (1), calculate the number of samples n that guarantees (ϵ, δ)-approximation of π, where $\epsilon = 0.01$ and $\delta = 0.01$. You may want to use the fact $p(1-p) \leq 0.25$ for $p \in [0, 1]$.

[1] There is even tighter inequalities, such as Chernoff bound and Hoeffding's inequality, but Chebyshev should be enough for us to get a glimpse of the behavior of the sample mean of i.i.d. random variables