University of Cyprus
Master in Data Science

DSC 531

2^{nd} **Assignment**

Deadline: Thursday, 02/12/2021

Your solutions for the programming part of this assignment should be in an R script (with comments for your code wherever needed). For any other comments on your results and also for any mathematical calculations needed, I require them in a simple LaTeX document. This should not exceed the length of 2-3 pages.

Exercise 1: (5 points). One method for pseudorandom number generation works as follows: starting with $X_0 \in \{0, 1, \ldots, 99\}$, define X_n for $n \in \mathbb{N}$ to be the middle two digits of the four-digit number X_{n-1}^2. If X_{n-1}^2 does not have four digits, it is padded with leading zeros. For example, if $X_0 = 64$, we have $X_0^2 = 4096$ and thus $X_1 = 09 = 9$. In the next step, we find $X_1^2 = 81 = 0081$ and thus $X_2 = 08 = 8$.

(a) Write a function, that for a given n and X_0, computes the n terms $X_1, X_2, \ldots, X_n$ using this pseudorandom number generator. You should do all the necessary checks for the input arguments and return the relevant errors if necessary; the user thinks that the input for both arguments should just be a real number.

(b) Try your method for various values of n and X_0. Can you see an obvious issue of the above method? Why is this the case?

Exercise 2: (3 points). Write a program which uses the inversion method to generate random numbers with probability density function

$$f(x) = \begin{cases} 1/x^2, & \text{if } x \geq 1 \\ 0, & \text{otherwise} \end{cases}.$$

To test your program, plot a histogram and a density plot of 10,000 random numbers from your function and compare this with the actual density f.

Exercise 3: (5 points). Let the random variable X follow a standard normal distribution, with its density as defined in the lecture notes.

(a) Write a function in R that employs the rejection algorithm in order to simulate n samples from the standard normal, with the proposal being the standard Cauchy distribution. The density of the standard Cauchy distribution is

$$g(x) = \frac{1}{\pi(1 + x^2)}, \quad x \in \mathbb{R}.$$

The function should also return the total number of draws.

(b) Test your program (for $n = 10,000$) by generating a histogram of the output and by comparing the histogram with the theoretical density of the standard normal distribution.

(c) Notice the total number of draws in part (b) of the exercise. Derive now mathematically the expected number of draws C. Does this agree with the results from part (b) of the exercise?

Exercise 4: (5 points). Let $\phi(x)$ be the probability density function of the standard normal distribution and consider the integral

$$J = \int_{-\infty}^{\infty} (x + \alpha)^2 \phi(x) \mathrm{d}x.$$

(a) Evaluate the integral. (It requires two lines of simple calculations and to use the definition of the expected value)

(b) Compute J by Monte Carlo integration. Use 100 and 1000 simulations and let $\alpha = 0, 1, 2, 3, 4$.

(c) Compute now J through importance sampling based on the candidate (proposal) $g(x) = \phi(x - \alpha)$. Use the same values for n and α as in part (b).

(d) How does the value of α affect the results in (b) and (c)? In addition, do you notice any differences between your results from (b) and those from (c)? If yes, which method gives better results?

Exercise 5: (2 points). Implement the random walk Metropolis-Hastings algorithm for sampling from the target distribution $\mathcal{N}(100, 1)$ on R, using proposals $Y = X + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 1)$. Experiment with different starting values $X_0 \in \mathbb{R}$ and create a plot of a path $X_0, X_1, \ldots, X_n$ which actually illustrates the need for a burn-in period.