Coding Test **Time Limit: 5 days**

Q1(a) Write a function to generate an m+1 dimensional data set, of size n, consisting of m continuous independent variables (X) and one dependent variable (Y) defined as

$$y_i = x_i\beta + e$$

where

- e is a Gaussian distribution with mean 0 and standard deviation (σ), representing the unexplained variation in Y
- β is a random vector of dimensionality m + 1, representing the coefficients of the linear relationship between X and Y, and $\forall i \in [1, n], x_{i0} = 1$

The function should take the following parameters:

- σ: The spread of noise in the output variable
- n: The size of the data set
- m: number of independent variables

Output from the function should be:

- X: An $n \times m+1$ numpy array of independent variable values (with a 1 in the first column)
- Y: The $n \times 1$ numpy array of output values
- β: The random coefficients used to generate Y from X.

Q1(b) Write a function that learns the parameters of a linear regression line given inputs

- X: An $n \times m$ numpy array of independent variable values
- Y : The $n \times 1$ numpy array of output values
- k: the number of interactions (epochs)
- τ : the threshold on change in Cost function value from the previous to current

iteration

- λ: the learning rate for Gradient Descent

The function should implement the Gradient Descent algorithm that initializes β with random values and then updates these values in each interaction by moving in the direction defined by the partial derivative of the cost function with respect to each of the coefficients. The function should use only one loop that ends after a number of iterations (k) or a threshold on the change in cost function value (τ).

The output should be an m + 1 dimensional vector of coefficients and the final cost function value.

Q2. Let's say you are at the bottom of a staircase with a die. With each throw of the die, you either move down one step (if you get a 1 or 2 on the dice) or move up one step (if you get a 3, 4, or 5 on the dice). If you throw a 6 on the die, you throw the die again and move up the staircase by the number you get on that second throw. Note if you are at the base of the staircase, you cannot move down! What is the probability that you will reach higher than the 200th step after 250 throws of the die?

Change the code so that you have a function that takes as a parameter, the number of throws. The function has another parameter that takes a probability distribution over all outcomes from a dice throw. For example (0.2,0.3,0.2,0.1,0.1,0.1) would suggest that the probability of getting a 1 is 0.2, 2 is 0.3 etc. Calculate the probability of reaching a step higher than the 200th one for the following cases:

(a) Number of throws: 100 throws, distribution=(0.2,0.2,0.2,0.2,0.1,0.1)
(b) Number of throws: 200 throws, distribution=(0.1,0.01,0.01,0.08, 0.4,0.4)
(c) Number of throws: 40 throws, distribution=(0.3,0.01,0.01,0.01,0.01,0.66)
(d) Number of throws: 1000 throws, distribution=(0.17,0.17,0.17,0.16,0.16,0.17)

Special Notes: Assume that the base of the staircase is numbered as step 0. You cannot use numpy. Only library that can be imported for this question is random.

Q3. Write a program to replace the following list of key phrases with underscore in between them in given text:

list_of_keyphrases = ['Prince Charles', 'Prince William', 'Meghan Markle', 'United Kingdom', 'North America', 'Duke and Duchess of Sussex', 'Queen Elizabeth II']

text = 'On January 8, Prince Harry and Meghan Markle, the Duke and Duchess of Sussex, unveiled their controversial plan to walk away from royal roles. We intend to step back as ‘senior’ members of the royal family and work to become financially independent while continuing to fully support her majesty the queen, they said in a joint statement. We now plan to balance our time between the United Kingdom and North America, continuing to honor our duty to the Queen, the commonwealth and our patronages. This geographic balance will enable us to raise our son with an appreciation for the royal tradition into which he was born, while also providing our family with the space to focus on the next chapter, including the launch of our new charitable entity, the statement added. Apparently, the announcement on the Sussex Royal Instagram page blindsided the Queen and other family members who had no idea it was coming, it sent tabloids into overdrive. Meanwhile, the Queen summoned Senior Royals to an emergency summit to discuss the future of the Duke and Duchess of Sussex. Billed as the Sandringham summit, the meeting took place at the Queen's estate in Norfolk and involved Queen Elizabeth II, Harry his father, Prince Charles and his brother Prince William, with Meghan Markle reportedly joining the discussions by phone from Canada. Soon after, the queen released a statement, that said, My family and I are entirely supportive of Harry and Meghan Markle desire to create a new life as a young family. Although we would have preferred them to remain full-time working members of the Royal family, we respect and understand their wish to live a more independent life as a family while remaining a valued part of my family.'

Q4. Given to you is the historical price of the Apple Stock (AAPL - AAPL - Google Sheets) on the NASDAQ stock exchange. Let us define daily returns as:

Daily Returns= (Adjusted Close of the current day/Adjusted close of the previous day)-1

Your task is to use pandas to read the CSV file, find the daily returns, and see if the data fits any possible probability distribution. For the same, you are encouraged to make use of plenty of visualisation libraries and any other possible libraries you think might be useful to test your hypothesis(claim). Post testing the same, give at least three calculations of probabilities which show that the empirical probability is **similar** to the one obtained by the probability distribution you hypothesised earlier.
In case you believe that no distribution reasonably fits the data, please explain why you think the same and give atleast three probability calculations which illustrate that there is **no similarity** between the empirical probabilities and the values obtained from the distribution.

Please illustrate your calculations and state your hypothesis within the ipynb notebook itself. Also, comment on the final conclusions within the notebook itself.

PS: Bonus points will be there for a statistically rigorous python notebook for this question.