Computer Programming, Assignment 4

Monsoon 2021, IIIT-H
Suresh Purini

Some problems below require use of graph plotting tools such as GNU Plot, Octave, Google Sheets, Microsoft Excel etc. You may use any tool you are comfortable with.

1 Randomness in Computation

1. Write a program to throw a dice which can generate one of the 6 faces from 1 to 6 uniformly at random. Throw the dice a million times and compute the histogram. Plot the histogram and check how far is it from uniform distribution.

2. Write a program to throw two 6-faced dice. Sum the face values of the two dice, which will be some value between 2 to 12. Throw the two dice a million times and compute the histogram. Plot the histogram and check how far is it from the theoretical distribution (i.e., if the dice are unbiased).

3. Write a program to estimate the value of π empirically using the following methodology. Consider a square centered at origin $(0, 0)$ with the following corner points: $(1, 1)$, $(-1, 1)$, $(-1, -1)$ and $(1, -1)$. Now, consider a circle of unit radius centered at origin. If you sample a point within the square, the probability that it falls within the circle is given by $\frac{\pi}{4}$ (why?). Draw a large number (N) of sample points and plot how the estimate of π improves with N using a suitable plotting tool.

2 Streaming Computations

4. The mean μ and the variance σ^2 of a sequence of N numbers $x_1, \cdots, x_N$ is defined as follows.

$$\begin{aligned} \mu &= \frac{\sum_{i=1}^{N} x_i}{N} \\ \sigma^2 &= \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \end{aligned}$$

The necessary data files for the following problem will be posted separately.

(a) Write a program to compute the mean using constant amount of memory. Here constant amount means anything which is not a function of the input sequence length N.

(b) **Challenge Problem:** Can you compute the variance also in constant amount of memory? Write a program to compute the variance approximately and check how close is your approximation to the actual variance by plotting two curves with increasing i as i moves from 1 to N as the program sees more and more data.

(c) **Challenge Problem:** Similar to the above problem write a program which computes the percentage of numbers which fall in the range $[0.8\mu, 1.2\mu]$.

3 Cryptography and Bit Manipulation

5. **Exclusive OR Generator (XORG)** Pick a random 127-bit seed $x_1, x_2, \cdots, x_{127}$. The subsequent bits are constructed as follows.

$$x_i \;\; = \;\; x_{i-1} \oplus x_{i-127} \;\; \text{for } i \geq 128.$$

(a) Compute the the probability distribution of $0s$ and $1s$ in $x_{128}, \cdots, x_N$ for $N = 10^6$. Compare this probability distribution as against when 0s and 1s are generated using $rand()$ % 2 approach.

(b) Compute $P(x_i = 0/x_{i-1} = 0)$ and $P(x_i = 0/x_{i-1} = 1)$ for both the aforementioned approaches.

(c) We can use XORG generator to encrypt a sequence of data bits $b_1, \cdots, b_N$. The encryption and decryption functions are as follows.

$$e_i = b_i \oplus x_{i+127} b_i = e_i \oplus x_{i+127}$$

The secret key for encryption and decryption is the seed of the XORG generator.

4 Data on the Disk

6. Write a program to compute the combined character frequency from a given list of files.

7. Write a program to compute the combined word frequency from a given list of files.

8. **Performance Contest** Two files contain sorted list of names. Write a program which generates an output sorted file combining the names from both the files.

9. **Challenge Problem** Sort a big file containing list of names. The size of the file is around 50 GB.

5 Bioinformatics

10. Will be posted.

11. **Challenge Problem** Will be posted.