CMPSCI590D: Algorithms for Data Science

Homework 1

Instructor: Barna Saha — Posted: Sept 18th, Due: Sept 29th at 4 pm

Do not look up materials on the Web. You can consult the reference books mentioned on the course website, and also the class slides for solving the homework problems. You may work in a group of size at most 3. Every person is allowed to talk to at most 2 others. All communications are bidirected: A talks with B, automatically implies B talks with A. Therefore, those who are working in a group of 3 are not allowed to talk with anyone else outside the group. Mention any collaboration clearly in the submission. Submit one homework solution per group. No late homework will be accepted.

For programming assignments, submit your code with a detailed readme file that contains instruction for running it. Also include any test dataset that you have used and results obtained to show correctness of your implementation.

*Total Point:*100*, Bonus Point:*20

Exercise 1. *A monkey types on a 26-letter keyboard that has lowercase letters only. Each letter is chosen independently and uniformly at random from the alphabet. If the monkey types* $1,000,000$ *letters, what is the expected number of times the sequence "proof" appears? [10]*

Exercise 2. *A permutation on the numbers* $[1, n]$ *can be represented as a function* $\pi : [1, n] \rightarrow [1, n]$*, where* $\pi(i)$ *is the position of* i *in the ordering given by the permutation. A fixed point of a permutation* $\pi : [1, n] \rightarrow [1, n]$ *is a value for which* $\pi(x) = x$*. Find the expected number of fixed points of a permutation chosen uniformly at random from all permutations. [20]*

Exercise 3. *Suppose you play a simple game with your friend where you flip a coin. If the coin is heads, your friend pays you a dollar. If it's tails, you pay your friend a dollar.*

(a) *Suppose you play the game* 100 *times, what is your expected pay off? [5]*

(b) *Suppose your friend decides to trick you, and swaps the fair coin for a biased coin that comes up tails with probability* 0.7*. What is your expected pay off if you play* 100 *times? [5]*

(c) *Use Markov Inequality to give an upper bound on the probability that your friend gets more than* 50 *after* 100 *rounds. [10]*

Exercise 4. *Suppose that we roll a standard fair die* 100 *times, Let* X *be the random variable denoting the sum of numbers that appear over the* 100 *rolls. Use Chebyshev's inequality to bound* $\Pr[|X - 350| \geq 50]$*. [10]*

Exercise 5. *Suppose you throw* m *balls into* n *bins, each ball equally likely to go into any of the* n *bins; imagine* $m \geq n$*. Let random variable* B_i *denote the number of balls in bin* i*. What is* $E[B_i]$*?*

- *Suppose* $m = 100n \ln n$*. Use the Chernoff bound to show that the number of balls in bin* i *does not differ from the expectation by more than (say)* $25 \ln n$ *with probability at least* $1 - \frac{1}{n^2}$*. Hence, show that the load of the heaviest and lightest bins differ by at most a constant factor with probability at least* $1 - \frac{1}{n}$*. [20]*

- *For general* $m = \Omega(n \ln n)$*, show that the number of balls in all the bins lie in the range* $\frac{m}{n} \pm O(\sqrt{\frac{m}{n}} \ln n)$ *with probability at least* $1 - \frac{1}{n}$*. [20]*

Exercise 6. *In the reservoir sampling mini-exercise 1, you saw as you run your algorithm multiple times, probability of selecting every item becomes uniform. Suppose, again you have 100 items coming in with reservoir of size 1. What is the minimum value* t *of number of runs that you would need to iterate your algorithm such that according to Chernoff bound each item will be selected between* $\frac{t}{100}(1 \pm \frac{1}{3})$ *times with probability* 0.99*? [20]*