Algorithms for Big Data Fall 2020

Homework 1

Due: Friday 10/09/20, 5pm PT

- **Solving 2 of the following 3 problems** will lead to full credit. You may attempt all 3 problems, but the grading will be based on the 2 problems with the highest scores.
- Email the solutions to both the instructor and TA (emails listed on the course website).
- You may work in groups of size 1-3. If you do, please hand-in a single assignment with everyone's names on it. It is strongly encouraged to type up the solutions in Latex.
- If the question asks to prove something, you must write out a formal mathematical proof.
- If the question involves analyzing an algorithm, you must formally explain the time and/or space usage, along with the approximation guarantees (when applicable).
- When you are asked to prove a bound, it suffices to prove it up to multiplicative constants, i.e., using $O(\cdot)$, $\Theta(\cdot)$, or $\Omega(\cdot)$ notation. No need to optimize (multiplicative) constants!
- You may use other resources, but you must cite them. If you use any external sources, you still must provide a complete and self-contained proof/result for the homework solution.

1 Problem 1: Coin Flipping

Let $\mathsf{Bern}(p)$ for parameter $p \in (0,1)$ denote the distribution that equals one with probability p and equals 0 with probability $1-p$. In other words, if $X_i \sim \mathsf{Bern}(p)$, then X_i can be thought of as a coin that comes up heads with probability p and tails with probability $1-p$. Assume that p is an unknown parameter. You want to determine the value of p by using some number t of i.i.d. samples $X_1, X_2, \ldots, X_t \sim \mathsf{Bern}(p)$. Consider the estimate $X = \frac{1}{t}\sum_{i=1}^{t} X_i$ for the value of p.

(a) Compute the expectation $\mathbb{E}[X]$ and the variance $\mathrm{Var}(X)$ as a function of p and t.

(b) For an accuracy parameter $\varepsilon \in (0,1)$, determine a value of t as a function of p and ε such that

$$\Pr\left[(1-\varepsilon)p \le X \le (1+\varepsilon)p\right] \ge \frac{9}{10}.$$

2 Problem 2: Balls and Bins

Consider n bins, where several balls are thrown. We throw each ball independently in a uniformly random bin. Let X be a random variable equal to the number of balls we need to throw until every bin contains at least one ball. Show that

$$\mathbb{E}[X] = n \cdot \sum_{i=1}^{n} \frac{1}{i}.$$

Use that $\sum_{i=1}^{n} \frac{1}{i} = \Theta(\log n)$ to conclude that $X = O(n \log n)$ with probability at least 9/10.

3 Problem 3: Maximum Element

Let $\mathcal{N}(0,1)$ denote the standard normal distribution. Let $X_1, X_2, \ldots, X_n \sim \mathcal{N}(0,1)$ be n random variables sampled i.i.d. uniformly. Determine a function $f(n)$ that is as small as possible such that with probability at least $1 - 1/n$, it holds that

$$\max_{i \in [n]} X_i \leq f(n).$$

In other words, provide an upper bound on $\|\vec{X}\|_\infty$ that holds with probability at least $1-1/n$, where $\vec{X} = (X_1 \; X_2 \; \cdots \; X_n)$ is an n-dimensional vector.