Algorithms for Big Data Fall 2020

Homework 1 Solutions

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

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

Solution: (a) The random variable X is given as

$$X = \frac{1}{t}\sum_{i=1}^{t} X_i.$$

By linearity of expectation:

$$\mathbb{E}[X] = \frac{1}{t}\sum_{i=1}^{t} \mathbb{E}[X_i]$$

Since, X_is are iid samples and $X_i \sim \mathsf{Bern}(p)$

$$\begin{aligned}\mathbb{E}[X] &= \frac{1}{t}\sum_{i=1}^{t} \mathbb{E}[X_i] \\ &= \frac{1}{t}\sum_{i=1}^{t} p \\ &= p.\end{aligned}$$

The variance, $\mathrm{Var}(X)$ can be written as:

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{Var}(\frac{1}{t}\sum_{i=1}^{t} X_i) \\ &= \frac{1}{t^2}\mathrm{Var}(\sum_{i=1}^{t} X_i) \\ &= \frac{1}{t^2}(\sum_{i=1}^{t} \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j))\end{aligned}$$

Since, X_is are iids, $\text{Cov}(X_i, X_j) = 0 \ \forall \ i \neq j$

$$\begin{aligned}
\text{Var}(X) &= \frac{1}{t^2} \sum_{i=1}^{t} \text{Var}(X_i) \\
&= \frac{1}{t^2} t \, \text{Var}(X_i) \\
&= \frac{p(1-p)}{t}.
\end{aligned}$$

(b) Observe that the inequality in part(b) of the problem can be rewritten as:

$$\Pr(|X - p| \leq p\varepsilon) \geq \frac{9}{10}.$$

Reversing the inequality, we get,

$$\Pr(|X - p| \geq p\varepsilon) \leq \frac{1}{10}. \tag{1}$$

According to Chebyshev's Inequality,

$$P(|X - \mu| \geq \alpha) \leq \frac{\text{Var}(X)}{\alpha^2}.$$

Therefore, for Inequality 1 to hold, we need

$$\begin{aligned}
\frac{\text{Var}(X)}{\alpha^2} &\leq \frac{1}{10} \\
\frac{p(1-p)}{tp^2\varepsilon^2} &\leq \frac{1}{10}
\end{aligned}$$

Therefore,

$$t \geq \frac{10(1-p)}{p\varepsilon^2}.$$

Observe that using Hoeffding's inequality you can get a much better bound,

$$t \geq \frac{\log(20)}{2p^2\varepsilon^2} \quad \text{(verify this)}$$

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.

Solution: Let X be the number of balls needed to fill n bins such that each bin has at least one ball. The key idea is to split the random process into 'rounds' based on the number of bins that are non-empty. We define random variables $X_1, X_2, \ldots, X_n$ as follows:

Let X_i denote the number of balls needed so that one of the remaining $n - (i-1)$ bins gets its first ball, given that exactly $i - 1$ bins already have at least one ball.

The probability that one of the remaining $n - (i-1)$ bins gets a ball is exactly $\dfrac{n-(i-1)}{n}$. Therefore,

$$\Pr(X_i = 1) = \frac{n-(i-1)}{n} = P_i$$

In other words, by independence of the balls, X_i has a geometric distribution with

$$\mathbb{E}[X_i] = \frac{1}{P_i}.$$

By linearity of expectation:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{n} \mathbb{E}[X_i] \\
&= \frac{1}{P_1} + \frac{1}{P_2} + \cdots + \frac{1}{P_n} \\
&= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} \\
&= n \cdot \sum_{i=1}^{n} \frac{1}{i}.
\end{aligned}$$

Use the approximation $\sum_{i=1}^{n} 1/i = c \log n$ for a constant c. Also, we have that

$$\Pr[X \geq 10 \cdot \mathbb{E}[X]] \;\leq\; 1/10$$

using Markov's inequality. Therefore, we conclude that $X \leq 10cn \log n = O(n \log n)$ with probability at least 0.9.

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.

Solution: The problem asks us to find a function $f(n)$ as small as possible such that,

$$\Pr(\max_{i \in [n]} X_i \leq f(n)) \geq 1 - \frac{1}{n}$$

or

$$\Pr(\max_{i \in [n]} X_i \geq f(n)) \leq \frac{1}{n}.$$

Observe that stating $\max_{i \in [n]} X_i \geq f(n)$ is equivalent to stating that there is at least one i such that $X_i \geq f(n)$. This is easy to prove. Hint: Try proving that if any one of the two statements is true, the other will be true as well.

Let E_i be the event: $X_i \geq f(n)$. Therefore,

$$\Pr(\max_{i \in [n]} X_i \geq f(n)) = \Pr(\bigcup_{i=1}^{n} E_i)$$

Using Union bound, we get,

$$\Pr(\bigcup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} \Pr(E_i).$$

According to Lemma 5 from the Lecture 2 notes, if $X \sim \mathcal{N}(0,1)$, then for any $\lambda > 0$, we have,

$$\Pr(|X| > \lambda) \leq e^{\frac{-\lambda^2}{2}}.$$

Since, $\mathcal{N}(0,1)$ is symmetrical with respect to 0, $\Pr(X \geq \lambda) = \frac{1}{2}\Pr(|X| \geq \lambda)$. Therefore,

$$\Pr(X \geq \lambda) \leq \frac{1}{2} e^{\frac{-\lambda^2}{2}}$$

Next, we use this inequality to get a tight bound on the sum as shown below:

$$\sum_{i=1}^{n} \Pr(E_i) = \sum_{i=1}^{n} \Pr(X_i \geq f(n)) \leq \frac{n}{2} e^{-\frac{f(n)^2}{2}}$$

Hence,

$$\Pr(\max_{i \in [n]} X_i \geq f(n)) \leq \frac{n}{2} e^{-\frac{f(n)^2}{2}}$$

Equating $\frac{n}{2} e^{-\frac{f(n)^2}{2}}$ with $\frac{1}{n}$, we get

$$f(n) = \sqrt{4\log(n) - 2\log(2)}.$$