Przemysław Uznański 07/10/2019

Algorithms for Big Data

Fall Semester 2019

Exercise Set 1

In the following we are concerned in designing a (memory/query) efficient algorithm for a following problem: we are given (in an offline form[1]) a binary array $A[1..n]$ where $\forall_i A[i] \in \{0,1\}$. Our goal is to estimate (up to some additive error ε) the value of $Y = \frac{1}{n}\sum_i A[i]$ using only little additional memory.

Exercise 1: (2 pts)

Show that simple *random sampling* performs well: select independently $i_1, i_2, \ldots, i_k \in [n]$. Show that $\frac{1}{k}(A[i_1] + \ldots + A[i_k])$ is an unbiased estimator of Y.[2] Use Hoeffding bound to bound k, the number of samples necessary, so that the estimation holds:

- with probability 9/10,
- with probability $1 - 1/n$? (So called *with high probability.*)

Exercise 2: (2 pts)

Use Chebyshev's inequality (instead of Hoeffding bound) to bound k from Exercise 1. How many samples do we need so that the estimation holds:

- with probability 9/10,
- with probability $1 - 1/n$?

Exercise 3: (2 pts)

Consider 9/10 probability estimation from previous exercise. Consider t fully independent repetitions of the same estimation procedure, with values $Y_1, Y_2, \ldots, Y_t$. Show that for $t = \Theta(\log n)$, the value of $\text{median}(Y_1, \ldots, Y_t)$ is an $\pm\varepsilon$ estimation of Y with high probability. What is the total number of samples needed?

Exercise 4: (2 pts)

- Prove Markov's inequality.
- Show that Chebyshev's inequality follows from Markov's inequality.

[1] offline: read-only

[2] X is an *unbiased estimator* of Y iff $\mathbb{E}[X] = Y$.

Theorem 1 (Markov's inequality) *Let $X \geq 0$ be a random variable. Then for any $k \geq 1$:*

$$\Pr(X \geq k \cdot \mathbb{E}[X]) \leq \frac{1}{k}.$$

Theorem 2 (Chebyshev's inequality) *Let X be a random variable. For any $k > 0$:*

$$\Pr(|X - \mathbb{E}[X]| \geq k \cdot \sqrt{\mathrm{Var}[X]}) \leq \frac{1}{k^2}.$$

Theorem 3 (Hoeffding bound) *Let $X_1, X_2, \ldots, X_n \in \{0, 1\}$ be* ***fully independent random variables****. Let $X = \sum_i X_i$. Then:*

$$\Pr(|X - \mathbb{E}[X]| \geq t) \leq 2\exp(-\frac{t^2}{n}).$$