CS170–Spring 2019 — Homework 13 Solutions

Ran Liao, SID 3034504227

April 28, 2019

Collaborators:Jingyi Xu, Renee Pu

1 Study Group

Name	SID
Ran Liao	3034504227
Jingyi Xu	3032003885
Renee Pu	3032083302

2 One-to-One Functions

Denote P to be the probability that a randomly chosen hash function h from $\mathcal{H}$ is one-to-one.

$$
\begin{aligned}
P &= 1 - P(\bigcup_{1\le i<j\le n} h(i) = h(j)) \\
&\ge 1 - \sum_{1\le i<j\le n} P(h(i) = h(j)) \\
&= 1 - \frac{n(n-1)}{2} \cdot \frac{1}{n^3} \\
&= 1 - \frac{n-1}{2} \cdot \frac{1}{n^2} \\
&\ge 1 - \frac{n}{2} \cdot \frac{1}{n^2} \\
&= 1 - \frac{1}{2n} \\
&\ge 1 - \frac{1}{n}
\end{aligned}
$$

3 Approximate Median

(a) **Main Idea**

Set $t = \frac{1}{2\epsilon^2}\ln(\frac{2}{\delta})$. Run the reservoir sampling of t elements without replacement algorithm in notes. Output the median elements in these t samples as result.

(b) **Proof of Correctness**

Create the random variables as follows,

$$X_i = \begin{cases} 1 & i\text{th sample is in } \frac{1}{2} \text{ percentile} \\ 0 & \text{otherwise} \end{cases}$$

Therefore, $p = E[X+i] = \frac{1}{2}$.

According to Hoeffding Bound,

$$Pr\Big[\Big|\frac{1}{t}\cdot\sum_{t=1}^{t} X_i - p\Big| \geq \epsilon\Big] \leq 2e^{-2\epsilon^2 t}$$

$$Pr\Big[\Big|\frac{1}{t}\cdot\sum_{t=1}^{t} X_i - \frac{1}{2}\Big| \geq \epsilon\Big] \leq 2e^{-2\epsilon^2 t}$$

$$Pr\Big[\Big|\frac{1}{t}\cdot\sum_{t=1}^{t} X_i - \frac{1}{2}\Big| \geq \epsilon\Big] \leq \delta$$

$$Pr\Big[\Big|\frac{1}{t}\cdot\sum_{t=1}^{t} X_i - \frac{1}{2}\Big| \leq \epsilon\Big] \geq 1-\delta$$

$$Pr\Big[-\epsilon \leq \frac{1}{t}\cdot\sum_{t=1}^{t} X_i - \frac{1}{2} \leq \epsilon\Big] \geq 1-\delta$$

$$Pr\Big[\frac{1}{2}-\epsilon \leq \frac{1}{t}\cdot\sum_{t=1}^{t} X_i \leq \frac{1}{2}+\epsilon\Big] \geq 1-\delta$$

Therefore, less than $\frac{1}{2}$ of the samples will be from the $\frac{1}{2}-\epsilon$ and $\frac{1}{2}+\epsilon$ percentile.

(c) **Space Complexity**

The reservoir will need $O(t) = O(\frac{1}{2\epsilon^2}\ln(\frac{2}{\delta}))$ space.

4 Count-Median-Sketch

(a) For a fixed choice of the functions h_i,

$$M[i, h_i(a)] = f_a + \sum_{b \neq a : h_i(b) = h_i(a)} f_b$$

The expectation of $M[i, h_i(a)]$ over the random choice of the function h_i is,

$$\begin{aligned}
\mathbb{E}[M[i, h_i(a)]] &= \mathbb{E}[f_a + \sum_{b \neq a : h_i(b) = h_i(a)} f_b] \\
&= f_a + \sum_{b \neq a} Pr[h_i(a) = h_i(b)] \cdot f_b \\
&= f_a + \frac{1}{B} \sum_{b \neq a} f_b \\
&\leq f_a + \frac{n}{B}
\end{aligned}$$

Applying Markov's inequality to the random variable $M[i, h_i(a)] - f_a$,

$$\begin{aligned}
P(M[i, h_i(a)] - f_a \geq \frac{2n}{B}) &\leq \frac{\mathbb{E}[M[i, h_i(a)] - f_a]}{\frac{2n}{B}} \\
P(M[i, h_i(a)] \geq f_a + \frac{2n}{B}) &\leq \frac{\frac{n}{B}}{\frac{2n}{B}} \\
P(M[i, h_i(a)] \geq f_a + \frac{2n}{B}) &\leq \frac{1}{2}
\end{aligned}$$

Therefore,

$$\begin{aligned}
P(\text{median}_{i=1\ldots,l} M[i, h_i(a)] \geq f_a + \frac{2n}{B}) &= P(\text{half of elements fit this equality}) \\
&= \left(P(M[i, h_i(a)] \geq f_a + \frac{2n}{B})\right)^{\frac{l}{2}} \\
&\leq \frac{1}{2^{\frac{l}{2}}}
\end{aligned}$$

(b) Generally, if there're more additions than deletions. The majority of $M[i, h_i(a)] - f_a$ are non-negative. Therefore, it's still reasonable to use the Markov's inequality to solve this problem. And according to (a), we only need to half of elements to meet the inequality. Therefore, the deduction should work most of times.

(c) No, because the *Count-Min-Sketch* only works when all $M[i, h_i(a)] - f_a$ are non-negative. This requirement may not be satisfied if deletions are allowed. It can be violated easily.