School of Computer Science and Technology
University of Science and Technology of China

Exercise Sheet 3 for Algorithms for Big Data 2023 Spring

Solution

Exercise 1

Given a stream $\mathcal{S}$ of m integers from the set $[n] = \{1, \cdots, n\}$, let $\mathrm{HH}_k = \{i \in [n] : f_i > \frac{m}{k}\}$ be the set of k-heavy hitters, where f_i is the frequency (i.e. the number of occurrences) of i in $\mathcal{S}$.
Modify Misra-Gries algorithm to find a set H such that

$$\mathrm{HH}_k(\mathcal{S}) \subseteq H \subseteq \mathrm{HH}_{2k}(\mathcal{S}).$$

Your algorithm should only use one pass and use at most $O(k(\log n + \log m))$ bits of space.

Solution.
Algorithm description:

(a) Run Misra-Gries algorithm to maintain $2k-1$ items i and counters $\hat{f}_i$.

(b) Output $H = \{i \in [n] : \hat{f}_i > \frac{m}{2k}\}$.

Correctness:
By Theorem in Lecture 11, we have

$$f_i - \frac{m}{2k} \leqslant \hat{f}_i \leqslant f_i.$$

If $f_i > \frac{m}{k}$, then $\hat{f}_i \geqslant f_i - \frac{m}{2k} > \frac{m}{2k}$ and the algorithm will output i. If $f_i \leqslant \frac{m}{2k}$, then $\hat{f}_i \leqslant \frac{m}{2k}$ and the algorithm will not output i. Therefore, the algorithm outputs all items i that $f_i > \frac{m}{k}$ and every item i in H satisfies $f_i > \frac{m}{2k}$, i.e., $\mathrm{HH}_k(\mathcal{S}) \subseteq H \subseteq \mathrm{HH}_{2k}(\mathcal{S})$.

Exercise 2

Let X and Y be finite sets and let Y^X denote the set of all functions from X to Y. We will think of these functions as “hash” functions. A family $\mathcal{H} \subseteq Y^X$ is said to be strongly 2-universal if the following property holds, with $h \in \mathcal{H}$ picked uniformly at random:

$$\forall x, x' \in X \; \forall y, y' \in Y \left(x \neq x' \Rightarrow \Pr_h[h(x) = y \wedge h(x') = y'] = \frac{1}{|Y|^2} \right).$$

We are given a a stream $\mathcal{S}$ of elements of X, and suppose that $\mathcal{S}$ contains at most s distinct elements. Let $\mathcal{H} \subseteq Y^X$ be a strongly 2-universal hash family with $|Y| = cs^2$ for some constant $c > 0$. Suppose we use a random function $h \in \mathcal{H}$ to hash.
Prove that the probability of a collision (i.e., the event that two distinct elements of $\mathcal{S}$ hash to the same location) is at most $1/(2c)$.

Proof. We use h chosen uniformly at random from $\mathcal{H}$ to hash the stream $\mathcal{S}$. Let the random variable C be the total number of collisions between any pair of distinct numbers in the stream. We are asked to show that $\Pr[C \geq 1] \leq 1/(2c)$. For any pair of distinct numbers x and x', let $\chi_{\{h(x)=h(x')\}} = 1$ if

$h(x) = h(x')$, and $\chi_{\{h(x)=h(x')\}} = 0$ otherwise. Then the total number of pairwise collisions can be written as $C = \sum_{x \neq x'} \chi_{\{h(x)=h(x')\}}$, and by linearity of expectation,

$$\mathrm{E}[C] = \mathrm{E}\left[\sum_{x \neq x'} \chi_{\{h(x)=h(x')\}}\right] = \sum_{x \neq x'} \mathrm{E}\left[\chi_{\{h(x)=h(x')\}}\right].$$

From the definition of strong 2-universality follows with $y = y'$, that for $x \neq x'$, $\mathrm{E}\left[\chi_{\{h(x)=h(x')\}}\right] = \Pr[h(x) = h(x')] \leq \sum_y \Pr\left[(h(x) = y) \wedge (h(x') = y)\right] = \sum_y 1/|Y|^2 = 1/|Y| = 1/cs^2$. There are at most $\binom{s}{2}$ distinct pairs in the stream $\mathcal{S}$, therefore

$$\mathrm{E}[C] \leq \frac{\binom{s}{2}}{cs^2} \leq \frac{\frac{1}{2}s^2}{cs^2} = \frac{1}{2c}$$

By Markov's inequality, we obtain

$$\Pr[C \geq 1] \leq \frac{\mathrm{E}[C]}{1} = \frac{1}{2c}.$$

□

Exercise 3

Consider the following two problems for analyzing a stream $\mathcal{S}$ of integers from $[n]$ in the strict turnstile model.

(k, ℓ_2)**-point query problem**: given a query(i), $1 \leq i \leq n$, find a value $\tilde{x}_i \in [x_i - \frac{\|\boldsymbol{x}\|_2}{k}, x_i + \frac{\|\boldsymbol{x}\|_2}{k}]$.

(k, ℓ_2)**-heavy hitters problem**: given a query(), find a set $L \subset [n]$ such that $|L| = O(k^2)$ and if $x_i > \frac{\|\boldsymbol{x}\|_2}{k}$ then $i \in L$.

Suppose there exists an algorithm $\mathcal{A}$ for the (k, ℓ_2)-point query problem with failure probability $\frac{\delta}{n}$ and using s bits of space.

Give an algorithm $\mathcal{A}'$ for the $(\frac{k}{4}, \ell_2)$-heavy hitters problem with failure probability δ and using small space.

Solution.

Algorithm description:

Iteratively query all points with algorithm $\mathcal{A}$ and obtain $\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_n$. Remember $\frac{k^2}{4}$ indices with the biggest estimators and denote it as L. Output L.

Correctness:

Let Y_i be the event that the query on $\tilde{x}_i$ fails, i.e., the event that $\tilde{x}_i \notin \left[x_i - \frac{\|\boldsymbol{x}\|_2}{k}, x_i + \frac{\|\boldsymbol{x}\|_2}{k}\right]$. Denote $Y = \bigcap_{i=1}^n \overline{Y_i} = \overline{\bigcup_{i=1}^n Y_i}$, i.e., the event that all n point queries succeed. Since the failure probability of algorithm $\mathcal{A}$ is at most $\frac{\delta}{n}$, we can see that $\Pr[Y_i] \leq \frac{\delta}{n}$. By union bound, $\Pr[Y] = 1 - \Pr\left[\bigcup_{i=1}^n Y_i\right] \geq 1 - \sum_{i=1}^n \Pr\left[Y_i\right] \geq 1 - \delta$.

We show that our algorithm $\mathcal{A}'$ always succeeds when all point queries succeed. Assume $\tilde{x}_i \in \left[x_i - \frac{\|\boldsymbol{x}\|_2}{k}, x_i + \frac{\|\boldsymbol{x}\|_2}{k}\right]$ holds for all $i = 1, 2, \ldots, n$.

- If x_i is a $(\frac{k}{4}, \ell_2)$ heavy hitter, then $x_i \geq \frac{4}{k}\|\boldsymbol{x}\|_2$. Therefore $\tilde{x}_i \geq x_i - \frac{1}{k}\|\boldsymbol{x}\|_2 \geq \frac{3}{k}\|\boldsymbol{x}\|_2$.
- If $\tilde{x}_i \geq \frac{3}{k}$, then $x_i \geq \tilde{x}_i - \frac{1}{k}\|\boldsymbol{x}\|_2 \geq \frac{2}{k}$. Thus

$$\|\boldsymbol{x}\|_2^2 = \sum_{i=1}^n x_i^2 \geq \sum_{i: x_i \geq \frac{2}{k}\|\boldsymbol{x}\|_2} x_i^2 \geq \frac{4}{k^2}\|\boldsymbol{x}\|_2^2 \left|\left\{i : x_i \geq \frac{2}{k}\|\boldsymbol{x}\|_2\right\}\right|$$

$$\left|\left\{i : \tilde{x}_i \geq \frac{3}{k}\right\}\right| \leq \left|\left\{i : x_i \geq \frac{2}{k}\|\boldsymbol{x}\|_2\right\}\right| \leq \frac{k^2}{4}$$

Therefore, if i is a $(\frac{k}{4}, \ell_2)$-heavy hitter, then $\tilde{x}_i$ is among the $\frac{k^2}{4}$ biggest entries of $\tilde{\boldsymbol{x}}$, i.e., $i \in L$. Therefore, the set L output by our algorithm solves the (k, ℓ_2)-heavy hitters problem.

Exercise 4
In the class, we have seen a KMV algorithm, denoted by $\mathcal{A}$, for estimating the number of distinct elements. The algorithm $\mathcal{A}$ has failure probability $\frac{1}{3}$. Design a new algorithm for the Distinct Elements problem with failure probability at most δ, for any $\delta < \frac{1}{3}$. Your algorithm can use $\mathcal{A}$ as a subroutine.

Solution.
Algorithm description:

(a) Independently run $s = \lceil 18 \ln \frac{1}{\delta} \rceil$ copies of $\mathcal{A}$ and obtain the estimators $t_1, t_2, \ldots, t_s$.

(b) Output $t^* \triangleq \mathrm{median}(t_1, t_2, \ldots, t_s)$.

Correctness:
Denote these s subroutines as $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_s$ where $\mathcal{A}_i$ outputs t_i. Denote the exact number of distinct elements as t. So $\mathcal{A}_i$ fails if and only if $|t_i - t| > \varepsilon$. Since each subroutine $\mathcal{A}_i$ has a failure probability $\frac{1}{3}$, we have $\Pr[|t_i - t| > \varepsilon] \le \frac{1}{3}$. We then show that our algorithm fails with probability at most δ, i.e., $\Pr[|t^* - t| > \varepsilon] \le \delta$. Denote $Y_i \triangleq [|t_i - t| > \varepsilon] = [\mathcal{A}_i \text{ fails}] = \begin{cases} 1 & |t_i - t| > \varepsilon \\ 0 & \text{otherwise} \end{cases}$ and $Y \triangleq \sum_{i=1}^{s} Y_i$.
We first show that $t^* < t - \varepsilon \implies Y \ge \frac{s}{2}$. Without loss of generality, assume $t_1 \le t_2 \le \cdots \le t_s$. Since t^* is the median, by the definition of the median, we have $t_1 \le t_2 \le \cdots \le t_{\lceil \frac{s}{2} \rceil} \le t^*$. So when $t^* < t - \varepsilon$, $t_1 \le t_2 \le \cdots \le t_{\lceil \frac{s}{2} \rceil} < t - \varepsilon$ must hold. This indicates that $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_{\lceil \frac{s}{2} \rceil}$ fails, which indicates that at least $\lceil \frac{s}{2} \rceil$ of the s subroutines fail. This indicates that $t^* < t - \varepsilon \implies Y \ge \lceil \frac{s}{2} \rceil \ge \frac{s}{2}$.
We can show that $t^* > t + \varepsilon \implies Y \ge \frac{s}{2}$ in a similar way, by assuming $t_1 \ge t_2 \ge \cdots \ge t_s$ without loss of generality. Therefore, $|t^* - t| > \varepsilon \iff t^* - t < -\varepsilon \vee t^* - t > \varepsilon \implies Y \ge \frac{s}{2}$ must hold.
Since $Y_i = [|t_i - t| > \varepsilon]$, $\mathrm{E}[Y_i] = \Pr[|t_i - t| > \varepsilon] \le \frac{1}{3}$. By linearity of expectation, $\mathrm{E}[Y] = \mathrm{E}[\sum_{i=1}^{s} Y_i] = \sum_{i=1}^{s} \mathrm{E}[Y_i] \le \frac{s}{3}$. Hence $|t^* - t| > \varepsilon \implies Y \ge \frac{s}{2} \implies Y - \mathrm{E}[Y] \ge \frac{s}{6}$ and thus $\Pr[|t^* - t| > \varepsilon] \le \Pr\left[Y - \mathrm{E}[Y] \ge \frac{s}{6}\right]$.
By Chernoff bound, we have $\Pr\left[Y - \mathrm{E}[Y] \ge \frac{s}{6}\right] \le \exp\left(-2\left(\frac{s}{6}\right)^2 \frac{1}{s}\right)$. Since $s \ge 18 \ln \frac{1}{\delta}$, we can see that $\Pr[|t^* - t| > \varepsilon] \le \delta$.