CSCE 689: Special Topics in Modern Algorithms for Data Science Fall 2023

Lecture 13 — September 20, 2023

Prof. Samson Zhou *Scribe: Dawei Xiang*

1 $\mathcal{L}_2$ Heavy-Hitters problem.

We will first consider a $\mathcal{L}_2$ Heavy-Hitters problem, defined as follows:

Given a set {S} of m elements from a stream $[n]$, we have a frequency vector $f \in \mathbb{R}^n$ where f_i is the frequency of i-th item, and a threshold parameter $\varepsilon \in (0, 1)$. We want a list that includes:

- The items from $[n]$ that have frequency at least $\varepsilon \cdot \|f\|_2$
- No item with frequency less than $\frac{\varepsilon}{2} \cdot \|f\|_2$

Compared to the previously mentioned $\mathcal{L}_1$ Heavy-Hitters problem, we now are interested in the items that are still frequent but less frequent than the $\mathcal{L}_1$ Heavy-Hitters, because $\|f\|_2$ is always less than or equal to $\|f\|_1$. To solve this problem, we introduce the *CountSketch* algorithm.

2 CountSketch

In general, CountSketch uses the randomized signs of different items to cancel out their effect on the estimated frequency.

Initialization: First we create b buckets of counters and use a random hash function $h : [n] \to [b]$ to map the streaming item to its corresponding bucket. And we also assign a uniformly random sign function $s : [n] \to \{-1, +1\}$, i.e., $\mathbf{Pr}[s(i) = +1] = \mathbf{Pr}[s(i) = -1] = 1/2$ to assign a sign for each element.

Algorithm: For each insertion(or deletion) to the element x_i, we change the counter $h(x_i)$ by $s(x_i)$ (or $-s(x_i)$). At the end of the stream, output the quantity $s(x_i) \cdot h(x_i)$ as the estimated frequency for x_i.

Here we give an example: suppose the stream is $[1, 1, 2, 3, 5, 1, 2, 4]$, $h(x) = x \mod 3$ and $s(x_i) = 1$ for $x_i \leq 3$. We denote the b buckets as $\{c_1, c_2, c_3\}$. Then we have:

f_1	f_2	f_3	f_4	f_5
3	2	1	1	1

$s(x_1)$	$s(x_2)$	$s(x_3)$	$s(x_4)$	$s(x_5)$
1	1	1	−1	−1

c_1	c_2	c_3
2	1	1

Given a set S of m elements from $[n]$, let $\hat{f}_i$ be the estimated frequency for f_i. Suppose $h(i) = a$, then we have $\hat{f}_i = s(i) \cdot c_a$.

3 CountSketch error analysis

We are now considering the error of the estimated frequency $\hat{f}_i$ compared to the ground truth frequency f_i.

We have $c_a = \sum_{(j:h(j)=a)}(s(j) \cdot f_a)$ and the estimated frequency f_i of i is

$$\hat{f}_i = s(i) \cdot c_a \tag{1}$$
$$= s(i) \cdot \sum_{(j:h(j)=a)} (s(j) \cdot f_a) \tag{2}$$
$$= s(i) \cdot s(i) \cdot f_i + \sum_{(j \neq i:h(j)=a)} (s(i) \cdot s(j) \cdot f_j) \tag{3}$$

Since $s(i) \in \{-1, +1\}$, we have $s(i) \cdot s(i) = 1$ and

$$= f_i + \sum_{(j \neq i:h(j)=a)} (s(i) \cdot s(j) \cdot f_j) \tag{4}$$

3.1 Mean Analysis

Now we can write the error as $error_i = \hat{f}_i - f_i = \sum_{(j \neq i:h(j)=a)}(s(i) \cdot s(j) \cdot f_j)$. And

$$\mathbb{E}[error_i] = \mathbb{E}[\sum_{j \neq i:h(j)=a} (s(i) \cdot s(j) \cdot f_j)] \tag{5}$$
$$= \sum_{j \neq i} \mathbb{E}[(s(i) \cdot s(j) \cdot f_j \cdot \mathbf{Pr}[h_j = h_i])] \tag{6}$$

Because $\mathbb{E}[s_i] = 0$ and s_i and s_j are independent to other variables, we have

$$= \sum_{j \neq i} \mathbb{E}[(s(i)] \cdot \mathbb{E}[s(j)] \cdot \mathbb{E}[f_j \cdot \mathbf{Pr}[h_j = h_i])] \tag{7}$$
$$= 0 \tag{8}$$

This means $\hat{f}_i$ is an unbiased estimator for f_i.

3.2 Variance analysis

Now we consider the variance of $|error_i|$. We have

$$\mathbb{E}[error_i^2] = \mathbb{E}[(\sum_{j\neq i:h(j)=a}(s(i)\cdot s(j)\cdot f_j))^2] \tag{9}$$

$$= \mathbb{E}[s(i)^2 \cdot (\sum_{j\neq i:h(j)=a}(s(j)\cdot f_j))^2] \tag{10}$$

$$= \mathbb{E}[(\sum_{j\neq i:h(j)=a}(s(j)\cdot f_j))^2] \tag{11}$$

Because $\mathbb{E}[s_i \cdot s_j] = 0 \ when \ j \neq i$, we have

$$= \sum_{j\neq i}\mathbb{E}[f_j^2 \cdot \mathbf{Pr}[h_j = h_i]] \tag{12}$$

$$= \sum_{j\neq i} f_j^2 \cdot \mathbf{Pr}[h_j = h_i] \tag{13}$$

$$= \sum_{j\neq i} f_j^2 \cdot \frac{1}{b} \tag{14}$$

$$\leq \frac{\|f\|_2^2}{b} \tag{15}$$

Because $Var(x) = \mathbb{E}[x^2] - \mathbb{E}[x]^2 \leq \mathbb{E}[x^2]$, $Var(error_i)$ is bounded by $\frac{\|f\|_2^2}{b}$.

If we set $b = \frac{9k^2}{\varepsilon^2}$, then the variance can be bounded by $\frac{\varepsilon^2\|f\|_2^2}{9k^2}$. Recall Chebyshev's inequality,

$$P(|X-\mu| \geq m\sigma) \leq \frac{1}{m^2} \tag{16}$$

Let $\sigma = \frac{\varepsilon\|f\|_2}{3k}$ and $m = 3$, then the probability that error for f_i is more than $\frac{\varepsilon\|f\|_2}{k}$ is less than 1/9.

Thus we can answer the $\mathcal{L}_2$ Heavy-Hitters problem. If we pick the items based on their estimated frequency $\hat{f}_i$, all the items we picked will satisfy the $\mathcal{L}_2$ Heavy-Hitters requirements.

4 Success boosting

If we have fixed b and we want to increase the success probability of $\hat{f}_i$ so that we can guarantee correctness for all $i \in [n]$ by a union bound, we can repeat multiple times to get estimates $e_1, ...e_l$, and use the median as the final estimator, applying Chernoff bounds.