COMS W6998-5: Algorithms through Geometric Lens September 13, 2017

Lecture 3– Heavy Hitters, CountSketch Algorithm

Instructor: *Alex Andoni* Scribe: *Elahe Vahdani*

We have a stream of items: $e_1, e_2, \cdots, e_m$ where $e_i \in [n]$. Let X be the frequency vector: $X = (x_1, x_2, \cdots, x_n)$ where x_i is number of times we've seen $e_j = i$.

Definition 1. *$i \in [n]$ is a ϕ-Heavy Hitter if $x_i \geq \phi \sum_{j=1}^{n} x_j = \phi ||X||_1$.*
ϕ is a parameter $\in (0, 1)$.

The problem is to find ϕ-heavy hitters for a fixed ϕ.

Idea:

Pick the following hash functions from 2-wise independent family :

1) $h : [n] \to [w]$

2) $r : [n] \to \{1, -1\}$ ($r_i \in \{1, -1\}$ for $i \in [n]$)

Space to store h and r: $\mathcal{O}(\log(n))$.

Let S be a hash table of size w so that: $S(j) = \sum_{i:h(i)=j} x_i r_i$ for $j \in [w]$.

Update on seeing element $e_t = i$:

$$S(h(i)) := S(h(i)) + r_i$$

To see if i is heavy hitter, look at $S(h(i))$.

Analysis:

Define $\delta_i = S(h(i)) - x_i r_i$.

Claim 2. $\mathbb{E}[\delta_i] = 0$.

Randomness is over r_{-i} where $r_{-i} = (r_1, r_2, \cdots, r_{i-1}, r_{i+1}, \cdots, r_n)$.

Proof.

$$\mathbb{E}[\delta_i] = \mathbb{E}[\sum_{\substack{j \neq i: \\ h(j)=h(i)}} x_j r_j] = \sum_{\substack{j \neq i: \\ h(j)=h(i)}} x_j \mathbb{E}[r_j] = 0 \tag{1}$$

□

Bounding Variance:

Define indicator variable χ:

$$\chi[h(j), h(i)] = \begin{cases} 1 & \text{if } h(i) = h(j) \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

$$\begin{aligned}
\text{Var}_{r_{-i},h}[\delta_i] &= \text{Var}_{r_{-i},h}[\sum_{\substack{j\neq i: \\ h(j)=h(i)}} x_j r_j] && \text{(3a)} \\
&\leq \mathbb{E}_{r_{-i},h}[(\sum_{\substack{j\neq i: \\ h(j)=h(i)}} x_j r_j)^2] = \mathbb{E}_h[\mathbb{E}_{r_{-i}}[(\sum_{\substack{j\neq i: \\ h(j)=h(i)}} x_j r_j)^2]] && \text{(3b)} \\
&= \mathbb{E}_h[\sum_{\substack{j_1,j_2\neq i: \\ h(j_1)=h(j_2)=h(i)}} x_{j_1} x_{j_2} \mathbb{E}_{r_{-i}}[r_{j_1} r_{j_2}]] = \mathbb{E}_h[\sum_{\substack{j\neq i: \\ h(j)=h(i)}} {x_j}^2] && \text{(3c)} \\
&= \mathbb{E}_h[\sum_{j\neq i} {x_j}^2 \chi[h(j), h(i)]] && \text{(3d)} \\
&= \sum_{j\neq i} {x_j}^2 \mathbb{E}_h[\chi[h(j), h(i)]] && \text{(3e)} \\
&= \sum_{j\neq i} \frac{{x_j}^2}{w} \leq \frac{||X||_2^2}{w} && \text{(3f)}
\end{aligned}$$

Chebyshev Bound:

$$|\delta_i|^2 \leq 10\frac{||X||_2^2}{w} \quad \text{with probability } \geq 0.9 \tag{4a}$$

$$|\delta_i| \leq \sqrt{\frac{10}{w}}||X||_2 \leq \sqrt{\frac{10}{w}}||X||_1 \quad \text{with probability } \geq 0.9 \tag{4b}$$

$$\tag{4c}$$

Ideally: want $|\delta_i| \leq \epsilon\phi||X||_1$ for small epsilon $\implies$ fix w such that $\sqrt{\frac{10}{w}} = \epsilon\phi \implies w = \mathcal{O}(\frac{1}{\epsilon^2\phi^2})$. So:

Claim 3.

$$|S(h(i))| = |x_i| \pm \epsilon\phi||X||_1 \quad \textit{with probability } \geq 0.9 \tag{5}$$

To find ϕ-heavy hitters, for each i check whether $|S(h(i))| > \phi(1-\epsilon)||X||_1$.

Issue: if i is ϕ-heavy hitter, then many other elements that collide with i in the same bucket, $S(h(i))$, will also be considered ϕ-heavy hitter! To avoid this issue, we use **Median Trick**:
Repeat L times:

$$L \text{ hash tables: } S_k[1 \cdots w]$$

$$L \text{ hash functions: } h_k : [n] \to [w]$$
$$L \text{ vector } r\text{: } r_{i,k} \in \{1, -1\}$$

Count Sketch(ϕ, ϵ)

Set $w = \mathcal{O}(\frac{1}{\epsilon^2\phi^2})$, $L = \mathcal{O}(\log(n))$.

- Initialize $S_k[j] = 0 \quad \forall k \in [L], j \in [w]$
- Update on seeing $e_j = i$:

$$\text{for } k = 1, \cdots, L\text{: } S_k[h_k(i)] := S_k[h_k(i)] + r_{i,k}$$

- Estimator: for $i = 1, \cdots, n$:

$$\hat{x_i} = \text{median}_{k=1,\cdots,L}\{S_k(h_k(i))\}$$

- Output i if $\hat{x_i} > \phi(1-\epsilon)||X||_1$

Theorem 4. *Count Sketch(ϕ, ϵ) outputs a list T such that, with probability $\geq (1 - \frac{1}{n})$:*

1) if i is ϕ-HH, then $i \in T$.

2) if $i \in T \implies i$ is $\phi(1-2\epsilon)$-HH.

Proof. Using Chernoff-Hoeffding bound:

Chernoff Bound:

$X_1, X_2, \cdots, X_m \in [0, 1]$ are i.i.d random variables, and $\mu = \mathbb{E}[\sum_{i=1}^m X_i]$, then:

$$\Pr[|\sum_{i=1}^{m} X_i - \mu| > \delta\mu] \leq 2e^{\frac{-\delta^2\mu^2}{m}}$$

To prove the Theorem, first we prove the following claim:

Claim 5. *Fix i, then $|\hat{x_i} - x_i| \leq \epsilon\phi||X||_1$ with probability $\geq 1 - \frac{1}{n^2}$.*

Proof. Let $m = L$. Define $X_1, X_2, \cdots, X_L$:

$$X_k = \begin{cases} 1 & \text{if experiment } \ k \ \text{ succeed :} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Experiment k:

$$|S_k(h_k(i)) - x_i r_i| \leq \epsilon\phi||X||_1$$

$\implies \mu = \mathbb{E}[\sum_{j=1}^{L} X_j] \geq 0.9L$ based on Claim 3.

Set $\delta = 0.01$. Using Chernoff Bound:

$$\Pr[|\sum_{i=1}^{m} X_i - 0.9L| > 0.10.9L] \leq 2e^{\frac{-0.010.81L^2}{L}} = 2e^{-0.0081L} \leq \frac{1}{n^2}$$

as long as $L = \frac{3}{0.0081}\ln(n) \implies$.

$$\sum_{j=1}^{L} X_j \geq 0.9L - 0.009L \geq 0.8L \quad \text{with probability} \quad \geq 1 - \frac{1}{n^2}$$

$\implies$ 80% of estimators $(S_k(h_k(i)))$ fall in range $[x_i - \epsilon\phi||X||_1, x_i + \epsilon\phi||X||_1]$ with probability $\geq 1 - \frac{1}{n^2}$, therefore median also will be in the aforementioned range with $1 - \frac{1}{n^2}$ probability which completes the proof for claim. □

Now to prove the Theorem:

$$\Pr[\text{ all estimators } \hat{x_i} \text{ are withing the allowed range }] \tag{7a}$$
$$= 1 - \Pr[\exists\ i \text{ such that } \hat{x_i} \text{ is misestimated }] \tag{7b}$$
$$\geq 1 - \Pr[\hat{x_1} \text{ is misestimated }] - \Pr[\hat{x_2} \text{ is misestimated }] - \cdots \geq 1 - n\frac{1}{n^2} = 1 - \frac{1}{n}. \tag{7c}$$

So, if i is ϕ-HH $\implies \hat{x_i} \geq \phi||X||_1 - \epsilon\phi||X||_1$, and algorithm will output it (similarly, for other condition). Space: $L \cdot w = \mathcal{O}(\log(n)\frac{1}{\epsilon^2\phi^2})$, and running time: $\mathcal{O}(nL)$ (n elements, and for each one we find the median of L variables). □

A variant of this sketch, is **Count Min** sketch which uses $\mathcal{O}(\log(n)\frac{1}{\epsilon\phi})space$.

Dynamic Graphs (in streaming)

Model: Graph G has n nodes. The stream is updates to edges of G: $e_j = ((u, v),$ add or delete). Initialization: zero edges.

Problem. Maintain a small sketch of G, so that can answer connectivity queries: are $u^\star$ and $v^\star$ connected in current G?

Total space: $\mathcal{O}(n \cdot \log(n)^{\mathcal{O}(1)}) << n^2$ (Semi streaming Model).

We will use l_0 sampling. Maintain a vector $X \in \{1, 0, -1\}^n$ under following updates: when see e_j, either increase or decrease x_i by 1.

l_0 sampling:

Problem. For vector $X = (x_1, x_2, \cdots, x_n)$, output $i \in [n]$ which is random from set $\{j : x_j \neq 0\}$ (Support(X)).