CS787: Advanced Algorithms	**Scribe:** Bhumesh Kumar
Lecture 21: Steaming Algorithms	**Date:** April 16, 2019

21.1 Streaming model

We recall the streaming model of interest, specified by the universe of possible streamed entries U such that $|U| = n$, an m length stream $i_1, i_2, \ldots, i_m$ where $i_t \in U \; \forall t$. We are interested in the frequency summary statistic $f_x = |\{t : i_t = x\}|$ and determining the heavy hitters ($f_X \geq \theta_m$ for some given θ_m) corresponding to them.

Count Sketch

1. Initialize $\text{count}_{j,p} = 0 \; \forall \; j \in [K], p \in [P]$
2. At time t, for all $p \in [P]$, update $count_{h_p(x),p} += g_p(i_t)$
3. At the end, for element x, return $\text{average}_{p \in [P]} \text{count}_{h_p(x),p}$

Introducing, hash functions and count function for the sketch. let $h_p : U \to [k]$ be pairwise independent hash functions and $g_p : U \to \{-1, +1\}$

$$\text{count}_{j,p} = \sum x : h_p(x) = j \text{Sign for } x \text{ in table } p f_x$$

$$\mathbb{E}[\text{count}_{h_p(x),p} g_p(x)] = f_x$$

Note that to analyze this kind of sketching we cannot apply Chernoff bound, which requires full independence between the random variables. We instead leverage Chebyshev's inequality.

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \geq \lambda) \leq \frac{\text{Var}(Z)}{\lambda^2}$$

For any x and p, denote $Y := \text{count}_{h_p(x),p} g_p(x) - f_x$. Then

$$\begin{aligned}
\text{Var}[Y] &= \mathbb{E}[Y^2] \qquad (\text{Since } \mathbb{E}[Y] = 0) \\
\text{Var}[Y] &\leq \sum_{y \in U, y \neq x} f_y^2 \times \frac{1}{k} \qquad (\text{Pairwise independence helps us drill the cross terms to zero}) \\
&\leq \sum_{y \in U} \frac{f_y^2}{k} \leq \frac{m^2}{k} \qquad (\text{Since } \sum_{y \in U} f_y \leq m)
\end{aligned}$$

If $\text{Var}[Z] \leq \alpha(\mathbb{E}[Z])^2$, then Chebyshev's inequality lends us

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq \epsilon\mathbb{E}[Z]] \leq \frac{\alpha(\mathbb{E}[Z])^2}{\epsilon^2(\mathbb{E}[Z])^2} = \frac{\alpha}{\epsilon^2} =: \delta$$

In the same spirit,

$$\text{Var}[\text{average}_{p\in[P]}\text{count}_{h_p(x),p}g_p(x)] \leq \frac{m^2}{kP} \approx \theta^2 m^2 \delta\epsilon^2$$
$$\text{Choosing } kP = \frac{1}{\theta^2\delta\epsilon^2}$$
$$\text{Space complexity for count min sketch } \approx \frac{1}{\theta}\frac{1}{\epsilon}\log\frac{1}{\delta}$$

21.2 Median of means

Let Z be an estimator such that $\mathbb{P}[Z \in (1 \pm \epsilon)\mu] \geq \frac{3}{4}$. Let $Z_1, Z_2, \ldots, Z_q$ be independent draws from distribution of Z

$$\begin{aligned}
\mathbb{P}[\text{median}(Z_1, Z_2, \ldots, Z_q) \in (1 \pm \epsilon)\mu] &\geq \mathbb{P}[\text{At least } \frac{q}{2}\ Z_i's\text{lie in}(1 \pm \epsilon)\mu \\
&= \mathbb{P}[\sum_i Q_i \geq \frac{q}{2}] \\
&= 1 - \mathbb{P}[\sum_i Q_i < \frac{q}{2}] \\
&\geq 1 - \mathbb{P}[|\sum_i Q_i - \mathbb{E}[\sum_i Q_i]| > \frac{q}{4}] \\
&\geq 1 - \exp(-(\frac{1}{3})^2\frac{3q}{4}) \\
&= 1 - \exp(-O(q))
\end{aligned}$$

Setting $q = \log\frac{1}{\delta}$, we get

$$\mathbb{P}[\text{median}(Z_1, Z_2, \ldots, Z_q) \in (1 \pm \epsilon)\mu] \geq 1 - \delta$$

Frequency moments
p^{th} moment $F_p = \sum_{x\in U} f_x^p$
In particular, 2^{nd} moment $F_p = \sum_{x\in U} f_x^2$
Tug of War Sketch

1. Initialize $\forall p \in [P]\text{count}_p = 0$
2. At step t, $\forall p \in [P]\text{update count}_p + = g_p(i_t)$

3. Return mean (or median of means) of $(\text{count}_p)^2$

Define $Z_p = \text{count}_p^2$

$$
\begin{aligned}
\mathbb{E}[Z_p] &= \mathbb{E}[(\sum_{x\in U} g_p(x) f_x)^2] \\
&= \sum_{x\in U} g_p^2(x) f_x^2 + 2 \sum_{x\neq x', x, x'\in U} f_x f_{x'} g_p(x) g_p(x') = \sum_{x\in U} f_x^2 = F_2
\end{aligned}
$$

For any fixed p

$$
\begin{aligned}
\mathbb{E}[Z_p^2] &= \mathbb{E}[(\sum_{x\in U} g_p(x) f_x)^4] \\
&= \sum_{x\in U} f_x^4 + 6 \sum_{x\neq x', x, x'\in U} f_x^2 f_{x'}^2 \qquad \text{(Using 4-independence between random variables)}
\end{aligned}
$$

We know,

$$
\begin{aligned}
F_2^2 &= (\sum_x f_x^4 + 2\sum x, x' f_x^2 f_{x'}^2)^2 \\
\therefore \mathbb{E}[Z_p] &\leq 3F_2^2 \\
\text{Var}(Z_p) &= \mathbb{E}[Z_p^2] - (\mathbb{E}[Z_p])^2 \\
&\leq 2F_2^2 - F_2^2 \\
&= 2F_2^2
\end{aligned}
$$

If we take $\frac{1}{\epsilon^2}$ copies we will get a variance reduction to $2\epsilon^2 F_2^2$ by taking their means. If we further take the median of these means ($\log \frac{1}{\delta}$ copies), we can use the median of means result to conclude that the resulting estimator with probability $1-\delta$ is around the true mean.

In the next, we'll see how to develop sketches for the special case of 0^{th} moment, which corresponds to the number of distinct elements in the stream.