CS 6550: Randomized Algorithms **Spring 2019**

Lecture 4: Streaming: Frequency moments

January 17, 2019

Lecturer: Eric Vigoda *Scribes: Mengfei Yang, Yatharth Dubey*

Theorem 4.1 *Chernoff bounds: Let $X_1, X_2, ..., X_n$ be independent variables, where $0 \le X_i \le 1$. Let*

$$X = \sum_{i=1}^{n} X_i, \quad \mu = E[X]$$

.
Then for $0 \le \delta \le 1$,

$$Pr[X \ge (1+\delta)\mu] \le e^{-\frac{\delta^2\mu}{3}}$$

$$Pr[X \le (1-\delta)\mu] \le e^{-\frac{\delta^2\mu}{2}}$$

4.1 Warm-up example: Median estimate

4.1.1 Problem definition

Definition 4.2 *$\epsilon-$approximate median: Given unordered list $S = [X_1, X_2, ..., X_m]$, for simplicity, assume $X_i's$ are distinct.*
The rank of y is given by

$$rank(y) = |\{x \in S : x \le y\}|$$

The goal is to find an $\epsilon-$approximate median of S. That is, given $\epsilon > 0$, find $y \in S$ where

$$\frac{m}{2} - \epsilon m < rank(y) < \frac{m}{2} + \epsilon m$$

4.1.2 Solution

The intuition is choose some random elements from the list, and output the median of these elements. Then prove this median is $\epsilon-approximate$ median.
algorithm 1 select $t \ge \frac{2}{\epsilon^2} log\frac{1}{\delta}$ random elements from S, then sort these random elements and output the median.

Algorithm 1: Find median

input : An unordered list $S = [X_1, X_2, ..., X_m]$.
output: One integer represents the median

1. $R = [r_1, r_2, ..., r_t] \longleftarrow$choose t random elements from S;
2. sort(R);
3. **return** median(R)

4.1.3 Analysis

Claim 4.3 *Assume p is the median returned by Algorithm 1.*

$$Pr[p \ is \ \epsilon - approximate \ median] \geq 1 - \delta$$

This means p is an $(\epsilon, \delta) - approximation$ of the median.

Proof: Divide S into 3 parts:

$$\begin{aligned} S_L &= \{y \in S : rank(y) \leq \frac{m}{2} - \epsilon m\} \\ S_M &= \{y \in S : \frac{m}{2} - \epsilon m < rank(y) < \frac{m}{2} + \epsilon m\} \\ S_U &= \{y \in S : rank(y) \geq \frac{m}{2} + \epsilon m\} \end{aligned}$$

If both $|R \cap S_L| < \frac{t}{2}$ and $|R \cap S_U| < \frac{t}{2}$ hold, then $p = r_{\frac{t}{2}} \in S_M$, which means p is $\epsilon - approximate$ median.
We will only show $|R \cap S_L| < \frac{t}{2}$. The other inequality will follow by an analogous argument.
Set random variables X_i to indicate whether element r_i is belong to S_L, and X to the summation of X_i.

$$X_i = \begin{cases} 1, & \text{if } r_i \in S_L; \\ 0, & \text{otherwise.} \end{cases}$$

$$X = \sum_{i=1}^{t} X_i$$

$$E[X_i] = \frac{\frac{m}{2} - \epsilon m}{m} = \frac{1}{2} - \epsilon$$

$$\mu = E[X] = t(\frac{1}{2} - \epsilon)$$

Now we can use Chernoff bounds

$$\begin{aligned} Pr[X \geq \frac{t}{2}] &= Pr[X \geq \mu + \epsilon t] \\ &\leq Pr[X \geq \mu(1 + 2\epsilon)] \\ &\leq e^{-(2\epsilon)^2 \frac{(\frac{1}{2} - \epsilon)t}{3}} \\ &\leq e^{-\frac{4\epsilon^2}{7} t} \\ &\leq \frac{\delta}{2} \end{aligned}$$

Hence,

$$Pr[|R \cap S_L| \geq \frac{t}{2}] \leq \frac{\delta}{2}$$

Similarly,

$$Pr[|R \cap S_U| \geq \frac{t}{2}] \leq \frac{\delta}{2}$$

$$Pr[|R \cap S_L| \leq \frac{t}{2} \text{ and } |R \cap S_U| \leq \frac{t}{2}] \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta$$

$$Pr[p \text{ is } \epsilon - approximate \text{ median}] \geq 1 - \delta$$

■

4.2 Streaming

4.2.1 Problem definition

Definition 4.4 *Streaming: get one-by-one m elements $X_1, X_2, ..., X_m$, where $X_i \in \{1, 2, ..., n\}$(X_i is repeatable). m is huge so we can't store the entire stream.*

Let f_i be the frequency of number i in the stream. Set $f = (f_1, f_2, ..., f_n)$

Definition 4.5 *Reservoir Sampling: choose an element S uniformly at random from $\{X_1, X_2, ..., X_m\}$ without knowing m beforehand.*

The problem is, give a function $g(f_i)$, where $g(0) = 0$, compute $\sum_{i=1}^{n} g(f_i)$

4.2.2 Solution

Algorithm 2 resolves Reservoir Sampling problem. Detailed analysis is given in later section.

Algorithm 2: Reservoir Sampling

input : streaming elements $X_1, X_2, ...$
output: one randomly chosen integer.

set $S \leftarrow X_1$;
for $t > 1$ **do**
  upon seeing t^{th} elements X_t, with probability $\frac{1}{t}$ set $S = X_t$
return S

To compute $\sum_{i=1}^{n} g(f_i)$, we introduce the unbiased estimator: a random variable X, where

$$E[X] = \sum_{i=1}^{n} g(f_i)$$

Algorithm 3 is AMS algorithm [AMS], shows how to calculate X.

Algorithm 3: AMS algorithm

input : streaming elements $X_1, X_2, ...$
output: Integer X

use Reservoir Sampling to choose random index $J \in \{1, 2, ..., m\}$;
$r \leftarrow |\{j \geq J : X_j = X_J\}|$; `// of occurrences of` x_J `after J`
$X \leftarrow m \times (g(r) - g(r-1))$;
return X;

4.2.3 Analysis

For algorithm 2, $S = X_i$ means set S while seeing i^{th} element, and never set S after that. The probability that $S = X_i$ for some time $t \geq i$ is

$$\begin{aligned} Pr[S = X_i] &= \frac{1}{i} \times (1 - \frac{1}{i+1}) \times (1 - \frac{1}{i+2}) \times ... \times (1 - \frac{1}{t}) \\ &= \frac{1}{i} \times \frac{i}{i+1} \times \frac{i+1}{i+2} \times ... \times \frac{t-1}{t} \\ &= \frac{1}{t} \end{aligned}$$

We only need to keep track of one number S, so it takes $O(logn)$ bits of space to get S. It will take $O(klogn)$ bits of space to get k samples.
Now analyze algorithm 3, X is the output of this algorithm.

Claim 4.6

$$E[X] = \sum_{i=1}^{n} g(f_i)$$

Proof:

$$\begin{aligned} E[X] &= Pr[X_J = i]E[X|X_J = i] \\ &= \sum_i \frac{f_i}{m} \sum_{r=1}^{f_i} \frac{m(g(r) - g(r-1))}{f_i} \\ &= \sum_i g(f_i) \end{aligned}$$

■

4.3 Example: Frequency Moments

For integer $k \geq 1$, the k-th frequency moment is denoted

$$F_k = \sum_{i=1}^{n} f_i^k.$$

Computing an (ϵ, δ)-approximation of F_k will be the goal of this section. Note that $g(r) = r^k$, where g plays the same role as in the previous section. We can now apply the AMS algorithms from the previous section to this function g. Then

$$X = m(r^k - (r-1)^k).$$

We know that $\mathrm{E}[X] = F_k$. Then, we can conduct l independent trials to get $X_1, ..., X_l$, and output $\frac{1}{l}\sum_{i=1}^{l} X_i$, an unbiased estimator for $\mathrm{E}[X]$ and therefore for F_k. To show that these are close with high probability, we plan to show that $\mathrm{Var}[X]$ is small and apply Chebyshev's Inequality. For this we employ the following lemma.

Lemma 4.7 $Var[X] \leq kn^{1-1/k}F_k^2$.

Then for

$$l = \frac{3\mathrm{Var}[X]}{\epsilon^2 \mathrm{E}[X]^2} \leq \frac{3kn^{1-1/k}F_k^2}{\epsilon^2 F_k^2} = 3kn^{1-\frac{1}{k}}\epsilon^{-2},$$

let $Y = \frac{1}{l}\sum_{i=1}^{l} X_i$, the mean of l independent trials. Now we compute the expected value and variance of Y.

$$\mathrm{E}[Y] = \mathrm{E}[X_i] = F_k$$

$$\mathrm{Var}[Y] = \frac{1}{l^2}\sum_{i=1}^{l} \mathrm{Var}[X_i] = \frac{\mathrm{Var}[X]}{l} = \frac{\epsilon^2 \mathrm{E}[X]^2}{3} = \frac{\epsilon^2 F_k^2}{3}$$

Then by Chebyshev's Inequality, we have

$$\Pr\left[|Y - \mathrm{E}[Y]| \geq \epsilon \mathrm{E}[Y]\right] = \Pr[|Y - F_k| \geq \epsilon F_k]$$

$$\leq \frac{\mathrm{Var}[Y]}{(\epsilon F_k)^2} = \frac{1}{3}.$$

So with probability at least $2/3$, Y is an ϵ-approximation of F_k. How can we boost this probability to at least $1-\delta$? We repeat the above procedure T times and take the median of the T estimates. Suppose we do this $T = c\log(1/\delta)$ times and get estimates $Y_1, ..., Y_T$. Then, consider the indicator random variable for Y_j being an ϵ-approximation

$$Z_j = \begin{cases} 1 & |Y_j - F_k| \leq \epsilon F_k \\ 0 & \text{otherwise} \end{cases}.$$

Then, for $Z = \sum_j Z_j$, we have $\mathrm{E}[Z] \geq \frac{2}{3}t$. Note that if $Z \geq \frac{t}{2}$, the median of $Y_1, ..., Y_T$ must be an ϵ-approximation. We now analyze the probability of this event

$$\begin{aligned} \Pr[Z < \frac{t}{2}] &\leq \Pr[Z \leq \mathrm{E}[Z](1 - \frac{1}{6})] \\ &\leq e^{-\frac{1}{6^2}\frac{t}{3}\frac{1}{3}} \\ &= e^{-\frac{t}{6^2 3^2}} \\ &\leq \delta, \end{aligned}$$

where the last inequality holds for $c \geq 6^2 3^2$.

References

[1] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci., 58(1):137147, 1999.*