COL758: Advanced Algorithms **Spring 2019**

Lecture 18: March 14

Lecturer: Naveen Garg *Scribe: Pawan Kumar*

18.1 Streaming Algorithm (Recap)

Heavy-hitters

In the previous lecture result by Misra-Gries got us relation between all the token having frequency $\tilde{f}_j$ greater than $\frac{m}{k}$, known as heavy-hitters, and their actual frequency in the stream, f_j.

$$f_j - \frac{m}{k} \leq \tilde{f}_j \leq f_j \tag{18.1}$$

with space requirement: 2k (token, counter)

Number of distinct tokens

The problem of counting the number of distinct tokens, however requires use of hash functions $h(j)$, which maps a token from domain $\{1 \dots n\}$ to another token onto the domain $\{1 \dots n\}$ randomly, i.e. $h : i \to j$, thus randomizing the input stream. The algorithm 1 for this problem, covered in lecture 17, is

Algorithm 1: Distinct Token Algorithm

Result: Output: $2^{\zeta+1/2}$

$\zeta \leftarrow 0$;
for *each j in stream* **do**
  $\zeta \leftarrow \max(zeroes(h(j)), \zeta)$;
end

gives $2^{\zeta+1/2}$ as an approximate to the number of distinct token in the stream. The intuition here is that, we expect that the the ζ which is the maximum trailing zeros in binary representation of $h(j)$ will be close to $\log(d)$, where d is number of distinct token that appear in the stream.

18.2 Quality of Algorithm's Estimate

We begin with defining two random variable and then determining their corresponding expectation and variance. Then we will go on to use Markov and Chebyshev inequalities to determine the quality of the estimate.

Random Variables $X_{r,j}, Y_r$ and their expectations and variances:

We define indicator random variable $X_{r,j}$ as in equation 18.2, where $zeros(h(j))$ returns trailing zeros in the mapped token j

$$X_{r,j} = \begin{cases} 1, & \text{if } zeros(h(j)) \geq r \\ 0, & \text{if } zeros(h(j)) < r \end{cases} \tag{18.2}$$

Define another random variable Y_r as in equation 18.3, which sums over all distinct tokens in the streams that are mapped to have trailing zeros $\geq r$.

$$Y_r = \sum_j X_{r,j} \tag{18.3}$$

The expectation of Y_r is given by 18.4

$$E[Y_r] = \sum_j E[X_{r,j}] \tag{18.4}$$

the expectation of $X_{r,j}$ is given by

$$\begin{aligned} E[X_{r,j}] &= Pr[X_{r,j} = 1] \\ &= Pr[2^r \text{ divides } h(j)] \\ &= \frac{1}{2^r} \end{aligned} \tag{18.5}$$

Probability that a token j in $\{1 \ldots \text{n}\}$ got mapped by $h(j)$ to have r or more trailing zeros is $1/2^r$.

The maximum number of trailing zeros we can potentially have for domain $\{1 \ldots n\}$ is $\log(n)$. Say, there are d distinct tokens and their binary representation has r or more trailing zeros. The expectation $E[Y]$

$$\begin{aligned} E[Y_r] &= \sum_j X_{r,j} \\ &= \frac{d}{2^r} \end{aligned} \tag{18.6}$$

The probability of the random variable Y_r, $Pr[Y_r]$ forms a monotone, i.e. with increasing r it will have lesser and lesser tokens, $X_{r,j}$ to sum over.

$$Var[Y_r] = \sum_j Var[X_{r,j}] \tag{18.7}$$

from independence relation, $h()$ are chosen s.t. $X_{r,j}$ are independent

(18.8)

to get equation 18.7 we need to find $Var[X_{r,j}]$

$$\begin{aligned} Var[X_{r,j}] &= E[(X_{r,j} - E[X_{r,j}])^2] \\ &= E[X_{r,j}^2] - E[X_{r,j}]^2 \\ &= 1^2.Pr[X_{r,j} = 1] - \left(\frac{1}{2^r}\right)^2 \\ &= \frac{1}{2^r}\left(1 - \frac{1}{2^r}\right) \\ &\leq \frac{1}{2^r} \\ &= E[X] \end{aligned} \tag{18.9}$$

When equation 18.9 is used to get variance $Var[Y_r]$

$$Var[Y_r] \leq \frac{d}{2^r} \tag{18.10}$$

Quality of the estimate $\hat{d}$

Say, the reported value of $\hat{d} = 2^{\zeta+1/2}$ and actual distinct tokens are d. We want to determine how close is $\hat{d}$ to d, by determining $Pr[\hat{d} < d/3$ and $Pr[\hat{d} > 3d]$. Let's also represent $3d$ and $d/3$ as follow:

- a is the smallest integer s.t. $3d < 2^{a+1/2}$
- b is the largest integer s.t. $d/3 > 2^{b+1/2}$

Probability of $\hat{d} > 3d$:

$$\begin{aligned} Pr[\hat{d} > 3d] &= Pr[2^{\zeta+1/2} \geq 2^{a+1/2}] \\ &= [\zeta \geq a] \end{aligned} \tag{18.11}$$

($\geq$ shows up because of the choice of a being the smallest integer)

(18.12)

Equation 18.11 implies that there are some tokens $h(j)$ which are divisible by 2^{ζ} thus divisible by a lesser divisor 2^a as well. Which means random variable $X_{a,j}$ will be 1. Therefore, the random Variable Y_a as well will assume some non-zero value.
Using Markov's inequality, we have

$$Pr[Ya \geq 1] \leq E[Y_a] \tag{18.13}$$

$$\frac{d}{2^a} < \sqrt{2}/3 \tag{18.14}$$

Probability of $Pr[\hat{d} < d/3]$:

$$\begin{aligned} Pr[\hat{d} < d/3] &= Pr[2^{\zeta+1/2} \leq 2^{b+1/2}] \\ &= Pr[\zeta \leq b] \end{aligned} \tag{18.15}$$

Equation 18.15 implies that tokens $h(j)$ may be divisible by 2^b, and must not be divisible by 2^{b+1}, i.e. $Pr[Y_{b+1} = 0]$. It follows that

probability of $Y_{b+1} = 0$

$$Pr[Y_{b+1} = 0] = Pr[|Y_{b+1} - E[Y_{b+1}]| \geq \frac{d}{2^{b+1}}] \tag{18.16}$$

(18.17)

Using Chebyshev inquality,

$$Pr[|Y_{b+1} - E[Y_{b+1}]| \geq \frac{d}{2^r}] \leq \frac{Var(Y_{b+1})}{(\frac{d}{2^{b+1}})^2} \tag{18.18}$$

from 18.7, we can use result on $Var(Y_r)$

$$\begin{aligned} &\leq \frac{2^{b+1}}{d} \\ &< \frac{\sqrt{2}}{3} \end{aligned} \tag{18.19}$$

Both the failure probabilities 18.19 and 18.14 are bounded by a rather large value $\frac{\sqrt{2}}{3} \approx 47\%$. It can be improved by using median trick, where we put all the token through n hash functions, chosen randomly from a family of hash functions and take the median as our estimate for number of distinct tokens

18.3 Randomization and the Median Trick

To improve the estimate, we will evaluate k hash function over the stream and report the median as the final output for $\hat{d}$

Algorithm 2: Distinct Token Randomized Algorithm

Result: Output: $2^{median(\zeta^i)+1/2}$

```
{ζ^i} ← 0;
for each j in stream do
    for each {h_i()} do
        ζ_i ← max(zeroes(h_i(j)), ζ);
    end
end
```

Analysis:

Probability that median is bad

$$Pr[median > 3d] = Pr[\text{at least } \frac{k}{2} \ \zeta_i's \text{ are s.t. } 2^{\zeta_i+1/2} > 3d] \tag{18.20}$$

Chernoff Bounds:

$$X = X_i + \cdots + X_k \tag{18.21}$$

$$X_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases} \tag{18.22}$$

$$Pr[X > (1+\delta)kp] \leq \left[\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right]^{kp} \tag{18.23}$$

Note: $kp = E[X]$, it can be seen that with increasing k, number of hash functions, probability decreases rapidly. Here we have defined X_i for variable hash functions, $X_i = 1$ if $2^{\zeta_i+1/2} > 3d$

$$Pr[X_i = 1] \leq \frac{\sqrt{2}}{3} \tag{18.24}$$

$$Pr[median > 3d] \leq Pr[X > k/2] \qquad (\text{where, } X = \sum_{i=1}^{k} X_i) \tag{18.25}$$

$$= Pr[X \geq (1+\delta)\frac{k\sqrt{2}}{3}]$$

$$\leq \left[\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right]^{k\sqrt{2}/3} \tag{18.26}$$

By increasing k, we can make $Pr[\hat{d} > 3d]$ as small as desired.

if we want:

$$Pr[\hat{d} > 3d] \leq \epsilon$$

then,

$$k \sim \log \epsilon \tag{18.27}$$

Now, we can see that

$$Pr[\hat{d} > \text{range}] < \delta$$
$$Pr[\hat{d} < \text{-range}] < \delta$$

therefore:

$$Pr[\hat{d} \quad \text{is within range}] < 1 - 2\delta \tag{18.28}$$