DS-563 / CS-543 Lecture 4 / 2023-01-31

Today

- Wrap up AMS sketch for F_2
- Distinct elements

Setting

Stream of data from X

$f(x)$ = frequency of x = # times x appears in the stream

m = length of the stream | $F_2 = \sum f^2(x)$

Last time

$h : X \to \{-1, 1\}$ ← fully random

$$Y = \sum_{x \in X} h(x) \cdot f(x)$$

$$\mathbb{E}[Y^2] = F_2$$

$$\mathrm{Var}[Y^2] \leq 2 F_2^2$$

[How can we use this?]

Chebyshev's Inequality

X – random variable, finite expectation & variance

$$\Pr\left(|X - \mathbb{E}[X]| \geq a\sqrt{\mathrm{Var}[X]}\right) \leq \frac{1}{a^2} \quad \text{for any } a > 0$$

k independent copies: $Y_1, Y_2, Y_3, Y_4, \ldots, Y_k$

$$\text{Output } Z = \frac{\sum_{i=1}^{k} Y_i^2}{k}$$

$$\mathbb{E}[Z] = F_2$$

independent variables

$$\mathrm{Var}[Z] = \sum_i \mathrm{Var}\left[\frac{Y_i^2}{k}\right] = \frac{1}{k^2}\sum_i \mathrm{Var}[Y_i^2] \leq \frac{k}{k^2} \cdot 2F_2^2$$

$$= \frac{2F_2^2}{k}$$

Set $k = \lceil 18/\varepsilon^2 \rceil$

$$\mathrm{Var}[Z] \leq \frac{\varepsilon^2 F_2^2}{9}$$

Chebyshev's inequality gives:

$$\Pr\left(|Z - F_2| \geq \varepsilon F_2\right) \leq \Pr\left(|Z - \mathbb{E}[Z]| \geq 3\sqrt{\frac{\varepsilon^2 F_2^2}{9}}\right)$$

$$\leq \Pr\left(|Z - \mathbb{E}[Z]| \geq 3\sqrt{\mathrm{Var}[Z]}\right) \leq \frac{1}{3^2}$$

In other words:

w.p. at least $8/9$:

$$(1-\varepsilon)F_2 \leq Z \leq (1+\varepsilon)F_2$$

This is called <u>$(1+\varepsilon)$-multiplicative approximation</u>

Implementation of each Y_i:

- start from empty counter
- for x in the stream add $h(x)$

Space usage: $O(1/\varepsilon^2)$ counters

Hash functions: 4-wise independence suffices to get all expectations right such as $\mathbb{E}[h(x)h(y)h(z)h(t)]$

For $X=[n]$, $O(1)$ words of space each

How: Discussion tomorrow (bonus: secret sharing)

Improve probability of success

$$8/9 \Rightarrow 1-\delta$$

- Run $O(\log(1/\delta))$ independent copies
- Return <u>median</u> of results

Why this works: Homework 1 (most likely)

Distinct elements

Goal: Compute $(1+\varepsilon)$-multiplicative approximation to

$$F_0 = |\{x \in X : f(x) \neq 0\}|$$

Algorithm

h = random hash function from X to $\mathbb{N}$ s.t.

$$\Pr(h(x) = i) = 2^{-(i+1)}$$

$h(x) =$	0	1	2	3	...
$\Pr(\checkmark)$	1/2	1/4	1/8	1/16	...

Initially:

$z \leftarrow 0$

$A \leftarrow \phi$

For each element x in the stream:

if $h(x) \geq z$

$A \leftarrow A \cup \{(x, h(x))\}$

while $|A| \geq c/\varepsilon^2$:

$z \leftarrow z + 1$

remove from A all pairs (y, g) s.t. $g < z$

some large constant $(c > 576)$

Output: $|A| \cdot 2^z$

Analysis:

Random variables:

Z – final value ~~of~~ of z

$X_{i,x} = \begin{cases} 1 & \text{if } h(x) \geq i \\ 0 & \text{o.w.} \end{cases}$ (for $x \in X$, $i \in \mathbb{N}$)

$$Y_i = \sum_{x : f(x) > 0} X_{i,x} \quad (\text{for } i \in \mathbb{N})$$

Output of the algorithm: $Y_Z \cdot 2^Z$

Incorrect output: $|Y_Z \cdot 2^Z - F_0| > \varepsilon F_0$

$$\Updownarrow$$

$$\left| Y_Z - \frac{F_0}{2^Z} \right| > \frac{\varepsilon F_0}{2^Z}$$

If ~~then~~ $Z = 0$: $|A| = Y_0 = F_0$

We output exact value

$$(*) = \Pr(\text{incorrect output}) = \sum_{i=1}^{\infty} \Pr\left(\left| Y_i - \frac{F_0}{2^i} \right| > \frac{\varepsilon F_0}{2^i} \wedge Z = i \right)$$

Select s (integer) s.t.

$$\frac{12}{\varepsilon^2} \leq \frac{F_0}{2^s} < \frac{24}{\varepsilon^2}$$

If $s \leq 1$, $F_0 \ll c/\varepsilon^2$ and $Z = 0 \Rightarrow$ algorithm outputs exact value

$$\circledast \leq \underbrace{\sum_{i=1}^{s-1} \Pr\left(\left| Y_i - \frac{F_0}{2^i} \right| > \frac{\varepsilon F_0}{2^i} \right)}_{\Delta} + \underbrace{\sum_{i=s}^{\infty} \Pr(Z = i)}_{\square}$$

$$\square = \Pr(Z \geq s) = \Pr\left(Y_{s-1} \geq c/\varepsilon^2 \right)$$

Markov's

$$\leq \frac{\mathbb{E}[Y_{s-1}]}{c/\varepsilon^2} = \frac{F_0/2^{s-1}}{c/\varepsilon^2} = \frac{2\varepsilon^2}{c} \cdot \frac{F_0}{2^s}$$

$$\leq \frac{2\varepsilon^2}{c} \cdot \frac{24}{\varepsilon^2} = \frac{48}{c} < \frac{1}{12}$$

selection of s

large c

$\Delta \leq$?

$$\mathbb{E}[Y_i] = \frac{F_0}{2^i}$$

$$\mathrm{Var}[Y_i] = \sum_{x: f(x) > 0} \mathrm{Var}[X_{i,x}] \leq \sum_{x: f(x) > 0} \mathbb{E}[X_{i,x}^2]$$

$$= \sum_{x: f(x) > 0} \mathbb{E}[X_{i,x}] = \mathbb{E}[Y_i] = \frac{F_0}{2^i}$$

$$\Delta = \sum_{i=1}^{s-1} \Pr\left(|Y_i - \mathbb{E}[Y_i]| > \frac{\varepsilon F_0}{2^i}\right) = \sum_{i=1}^{s-1} \Pr\left(|Y_i - \mathbb{E}[Y_i]| > \varepsilon \sqrt{\frac{F_0}{2^i}} \sqrt{\mathrm{Var}[Y_i]}\right)$$

$$\underset{\substack{\uparrow \\ \text{Chebyshev's} \\ \text{inequality}}}{\leq} \sum_{i=1}^{s-1} \frac{2^i}{\varepsilon^2 F_0} < \frac{2^s}{\varepsilon^2 F_0} \underset{\substack{\uparrow \\ \text{selection} \\ \text{of } s}}{\leq} \frac{1}{\varepsilon^2} \cdot \frac{\varepsilon^2}{12} = \frac{1}{12}$$

$$(*) \leq \Delta + \square \leq \frac{1}{12} + \frac{1}{12} = \frac{1}{6}$$

Conclusion: The algorithm outputs a $(1+\varepsilon)$-multiplicative approximation w.p. at least $5/6$

Space usage: $\underbrace{O(1/\varepsilon^2) \text{ elements of } X}$ $+$ $\underbrace{O(1/\varepsilon^2) \text{ integers}}$

can be reduced by hashing into a small range (2-wise independence is sufficient)

at most $O(\log \log m)$ bits with high probability for further space savings (m = stream length)

Optimal: $O(1/\varepsilon^2 + \log n)$ bits for $X = [n]$

Hash functions h?

- 2-wise independence suffices, needed for $\mathrm{Var}[Y_i]$

- non-uniform distribution on some range? see Homework 1