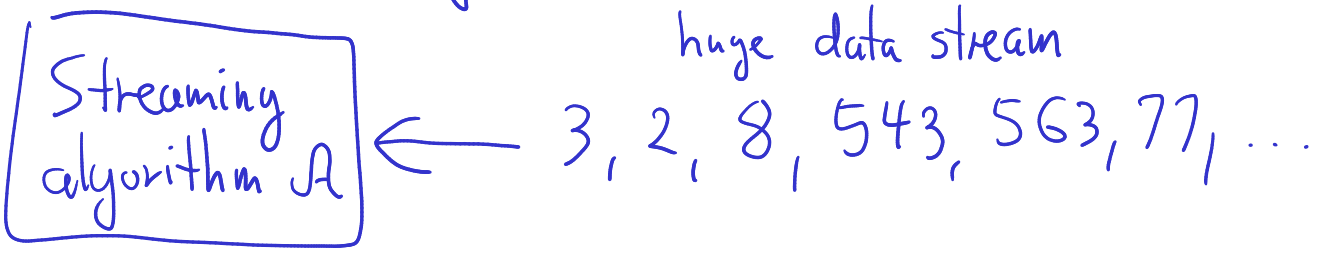
DS-563 / CS-543 2024-02-01

Today

- Frequency moments
- AMS sketch for F_2

Model: Streaming algorithms



X - universe of elements of the stream

for $x \in X$, $f(x) = \#$ occurrences of x

Moments: α-th frequency moment for $\alpha \in (0, \infty)$

$$F_\alpha(\text{input}) = \sum_{x \in X} (f(x))^\alpha$$

example: $F_2 = \sum_{x \in X} f^2(x)$

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

↖ also "number of distinct elements"

Why moments?

- important statistical tool
- naturally appears in some contexts
  - tracking network traffic
  - database planning
- can be used to approximate other functions (e.g., entropy)

Now: AMS sketch for F_2

||

Alon-Matias-Szegedy (1996)

(classic streaming paper, won important awards!)

$h : X \to \{-1, +1\}$ ← random hash function selected from uniform distribution on all functions

Basic estimator:

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

Quick check:

$\mathbb{E}[Y] = ?$ $\quad \mathbb{E}[Y] = \sum_{x \in X} f(x) \cdot \underbrace{\mathbb{E}[h(x)]}_{=0} = 0$

$$\mathbb{E}[Y^2] = \mathbb{E}\left[\left(\sum_x h(x) f(x)\right)^2\right]$$

$$= \mathbb{E}\left[\sum_x \underbrace{h^2(x)}_{=1} f^2(x) + \sum_{\substack{x,y \\ x \neq y}} h(x) h(y) f(x) f(y)\right]$$

$$= \sum_x f^2(x) + \sum_{\substack{x,y \\ x \neq y}} f(x) f(y) \underbrace{\mathbb{E}[h(x) h(y)]}_{=0}$$

$$= \sum_x f^2(x) = F_2 \leftarrow \text{exactly what we want in expectation}$$

this situation: unbiased estimator

[Question: Is having an unbiased estimator good enough?]

$$\mathrm{Var}[Y^2] = \mathbb{E}[(Y^2)^2] - (\mathbb{E}[Y^2])^2$$

$$\mathbb{E}[(Y^2)^2] = \mathbb{E}[Y^4] = \mathbb{E}\Big[\sum_{x,y,z,t} h(x)h(y)h(z)h(t)\cdot f(x)f(y)f(z)f(t)\Big]$$

$$= \sum_{x,y,z,t} \underbrace{\mathbb{E}[h(x)h(y)h(z)h(t)]}_{\text{up to four different elements } x,y,z,t.\ \text{If one of them occurs odd number of times, this expectation is 0.}} \cdot f(x)f(y)f(z)f(t)$$

Non-zero cases:

① $x=y=z=t$

② two pairs of different elements

example: $x=t\neq y=z$

$$\mathbb{E}[(Y^2)^2] = \underbrace{\sum_{x} f^4(x)}_{①} + \underbrace{3\sum_{\substack{x,y\\ x\neq y}} f^2(x)f^2(y)}_{②}$$

$$(\mathbb{E}[Y^2])^2 = (F_2)^2 = \sum_{x} f^4(x) + \sum_{\substack{x,y\\ x\neq y}} f^2(x)f^2(y)$$

$$\mathrm{Var}[Y^2] = 2 \sum_{\substack{x,y \\ x \neq y}} f^2(x) f^2(y)$$

$$\leq 2 \sum_{x,y} f^2(x) f^2(y)$$

$$= 2 \underbrace{\left(\sum_x f^2(x) \right)}_{= F_2} \underbrace{\left(\sum_y f^2(y) \right)}_{= F_2}$$

$$= 2 F_2^2$$

[How can we use this?]

Chebyshev's inequality

X - random variable with finite expectation & variance

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

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

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

$$\mathbb{E}[Z] = \frac{k \cdot \mathbb{E}[Y]}{k} = \mathbb{E}[Y] = F_2$$

$$\mathrm{Var}[Z] = \frac{1}{k^2}\,\mathrm{Var}\left[\sum_{i=1}^{k} Y_i^2\right]$$

$$= \frac{1}{k^2}\sum_{i=1}^{k} \mathrm{Var}\left[Y_i^2\right]$$

independent variables

$$= \frac{k}{k^2}\,\mathrm{Var}\left[Y^2\right]$$

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

Set $k = \lceil 18/\varepsilon^2 \rceil \Rightarrow \mathrm{Var}[Z] \leq \frac{\varepsilon^2 F_2^2}{9}$

Chebyshev's inequality gives:

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

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

In other words:

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

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

This guarantee: $(1+\varepsilon)$-multiplicative approximation

---

Implementation of each Y_i:

- start from empty counter
- for each x in the stream,
  add $h_i(x)$ ← insertions

(if deletions allowed, for a deletion of x, subtract $h_i(x)$)

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

Hash functions: 4-wise independence suffices to get all expectations right

(e.g., $\mathbb{E}[h(x)h(y)h(z)h(t)]$)

For $X = [n]$ ← integers using $O(1)$ words of space, each hash function needs $O(1)$ words of space

How: via low degree polynomials (discussion section?)

Better probability of success?

See Question 3 on Homework 1