DS-563/CS-543 2025-01-30

Today: AMS sketch for F_2

Review: Frequency moments F_p

- Input/data: multiset of items from X
- Frequencies $f: X \to \mathbb{N}$

  $f(x) = \#\text{occurrences of } x \in X$
- p-th frequency moment

  $p \in (0, \infty)$

  $$F_p(\text{the data}) = \sum_{x \in X} |f(x)|^p$$

  absolute value because in general $f(x)$ can be negative

Streaming algorithms

big stream of data

$\boxed{\text{Streaming algorithm}} \longleftarrow 3, 11, 4, 3, 3, 1, 4, \ldots$

Today: we want to estimate $F_2(\text{stream})$ in small space

Solution: AMS sketch

Alon-Mattias-Szegedy (1996)

This paper was very impactful.
It inspired a lot of interest
in the streaming model.

Basic estimator:

$h : X \to \{-1, +1\}$ is "fully random"

each $h(x)$ independent
and uniformly distributed

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

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

Hmm...

Not very interesting

But...

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

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

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

independence

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

this scenario: Y^2 is an unbiased estimator for F_2

Question: Is having an unbiased estimator good enough?

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

$$\square = \mathbb{E}[(Y^2)^2] = \mathbb{E}[Y^4]$$

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

$$= \sum_{x,y,z,t} \underbrace{\mathbb{E}\left[h(x)h(y)h(z)h(t)\right]} \cdot f(x)f(y)f(z)f(t)$$

Up to four different elements x, y, z, t.
If one of them occurs odd number of times, this expectation is 0.

Non-zero cases:

1. $x = y = z = t$
2. two pairs of different elements (example: $x = t \neq y = z$)

$$\square = \underbrace{\sum_{x} f^4(x)}_{1} + \underbrace{3 \sum_{\substack{x,y \\ x \neq y}} f^2(x) f^2(y)}_{2}$$

$$\triangle = \left(\mathbb{E}[Y^2]\right)^2 = F_2^2 = \left(\sum_x f^2(x)\right)^2$$

$$= \sum_x f^4(x) + \sum_{\substack{x,y \\ x \neq y}} f^2(x) f^2(y)$$

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

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

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

[How can we use this?]

$$\boxed{\begin{array}{l} \text{Chebyshev's inequality} \\ X - \text{random variable with finite expectation \& variance} \\ \Pr\left[|X - \mathbb{E}[X]| \geq a\sqrt{\mathrm{Var}[X]} \right] \leq \frac{1}{a^2} \quad \text{for any } a > 0 \end{array}}$$

Use k independent copies: $Y_1, Y_2, \dots, Y_k$

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

$$\mathbb{E}[Z] = \frac{k\,\mathbb{E}[Y^2]}{k} = \mathbb{E}[Y^2] = F_2$$

$$\mathrm{Var}[Z] = \frac{1}{k^2}\,\mathrm{Var}\left[\sum_{i=1}^{k} Y_i^2\right] \underset{\substack{\uparrow \\ \text{independent} \\ \text{variables}}}{=} \frac{1}{k^2}\cdot\sum_{i=1}^{k}\mathrm{Var}[Y_i^2]$$

$$= \frac{k}{k^2}\,\mathrm{Var}[Y^2] \leq \frac{2F_2^2}{k}$$

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

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

Via Chebyshev's inequality:

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

$$\leq \Pr\left[\left|Z - \mathbb{E}[Z]\right| \geq 3\sqrt{\mathrm{Var}[Z]}\right] \leq \frac{1}{3^2} = \frac{1}{9}$$

$$\boxed{\text{We will finish the discussion of AMS sketch next time}}$$