DS-563 / CS-543

Lecture 3 / 2023-01-26

Today

- Heavy hitters
- Second moment estimation via AMS sketch

Setting:

X - universe from which elements of the stream come

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

$x \in X$

S = total number of items $= \sum_{x \in X} f(x)$

Heavy hitters: "find most frequent elements of the stream"

Our task: For some $\varepsilon \in (0,1)$, (parameter)

return $H \subseteq X$ s.t.

$$\forall x \in X: \quad \begin{array}{l} f(x) \geq 2\varepsilon \cdot S \Rightarrow x \in H \\ f(x) \leq \varepsilon \cdot S \Rightarrow x \notin H \end{array}$$

(In general, thresholds $0 < \alpha < \beta < 1$.)

In other words;

- We want to include all <u>heavy</u> elements x in our output
  $\left(\frac{f(x)}{s} \geq 2\varepsilon\right)$
- We don't want to output any <u>light</u> elements
  $\left(f(x)/s \leq \varepsilon\right)$
- "Gray zone!" elements in between, $\varepsilon < \frac{f(x)}{s} < 2\varepsilon$, can be ~~outpt~~ output, but don't have to be

---

<u>Approach</u>:

- find candidates $H' \subseteq X$
- use Count Min Sketch to verify:
  output all $x \in H'$ for which
  Count Min Sketch says "(fraction of x) $\geq 2\varepsilon$"

How to find <u>small</u> H'?

Warm-up:

- find element that occurs > 50% of time (called "leader")
- if no leader, output nothing or any element in X

Algorithm:

- remember at most one element ~~x~~ in X plus a count (= number of copies) — We call this "storage"
- initially, storage empty
- when new item x arrives:
    - if storage empty, store $(x, 1)$
    - otherwise, storage contains $(y, c) \in (X, \mathbb{Z}_+)$
        - if $x = y$:
          replace (y, c) with $(y, c+1)$
        - otherwise $(x \neq y)$:
          replace (y, c) with $(y, c-1)$
          and if $\boxed{c-1} = 0$, empty storage
- at the end of the stream:
  if storage non-empty, output the element from X in it

Why this works:

- The algorithm keeps forgetting pairs of <u>different</u> elements

- If x is a leader in multiset ~~$S = S' \cup \{x\}$~~

  $S = S' \cup \{y, z\}$, it's a leader in S'

  $y \neq z$

- Proof by induction:

  $S_0 =$ entire stream (multiset)

  $y_0 \neq z_0 \leftarrow$ first two elements removed

  $S_1 = S_0 \setminus \{y_0, z_0\}$

  single copies of y_0, z_0 removed

  If x is a leader in $S_0 \Rightarrow x$ is a leader in S_1

  Now: $y_1 \neq z_1 \leftarrow$ second two elements removed

  ⋮

If x is a leader in S_0, it will be in storage at the end of the stream.

Extension to finding elements more frequent than $1/k$:

- store at most k elements with counts (or $k-1$)
- if k different elements in storage:
  - decrease the count for each
  - remove those with count 0
- at the and of the stream, output elements in the storage

Back to original heavy hitters problem:

Say X = typical integers

~~W. p. 99%:~~ To solve it with probability 99%:

$$\text{space} \quad \underbrace{O(1/\varepsilon)}_{\text{finding candidates}} + \underbrace{O\left(\tfrac{1}{\varepsilon}\log(1/\varepsilon)\right)}_{\text{CountMin Sketch}}$$

Frequency moments

Goal: approximate $F_p = \sum (f(x))^p$

p-th moment

corner case $p=0$:

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

distinct elements

- 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

AMS = 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

Quick check:

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

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

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

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

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

this situation: <u>unbiased</u> estimator

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

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

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

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

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

Surviving / non-zero cases:

- $x = y = z = t$
- two pairs of different elements: e.g. $x = y \neq z = t$ (in general all permutations)

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

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

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

$$= 2 \sum_{x} \sum_{\substack{y \\ (y \neq x)}} f^2(x) f^2(y)$$

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

$$= 2 \sum_{x} f^2(x) \cdot F_2$$

$$= 2 F_2 \sum_{x} f^2(x) = 2 F_2^2$$

TO BE CONTINUED