DS-563 / CS-543 2024-01-30

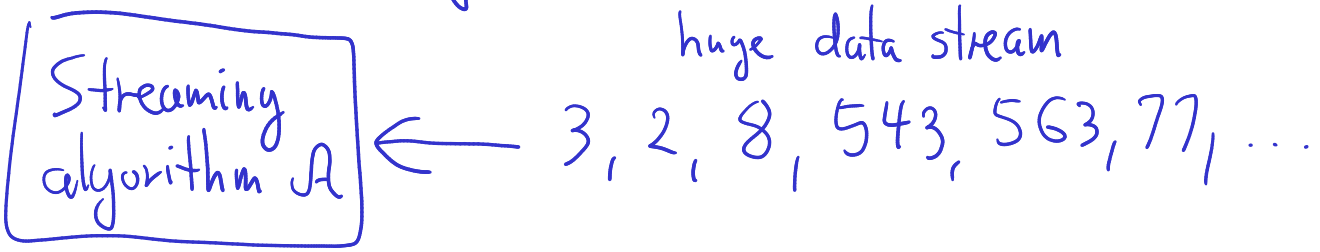
[News: HW2 is out]

## Today

- Deterministic fraction estimates (Misra-Gries)
- ~~Moments & AMS sketch~~ Next time

---

## Model: Streaming algorithms




- Process elements one by one
- Use small space (don't retain all data)

For now: insertions only

---

Goal: after seeing a collection of items provide fraction estimates for any element $x$ $\longrightarrow \frac{f(x)}{\sum_{y \in X} f(y)}$

($f(z)$ = #occurrences of $z$)

---

Misra-Gries algorithm:

- parameter $k \in \mathbb{N}$, $k \geq 2$
- maintains a small collection of items with counts

pairs $(x, c) \in X \times \mathbb{N}$

($X$: input universe)

Input processing:

$C \leftarrow \phi$, $S \leftarrow 0$

for each item $x$:

  $S \leftarrow S+1$

  if pair $(x, t)$ in $C$:

    replace it with $(x, t+1)$

  else:

    insert $(x, 1)$ into $C$

  if $|C| = k$:

    for each $(y, t)$ in $C$:

      replace it with $(y, t-1)$

    remove all pairs $(\ldots, 0)$ from $C$

Estimates:

For $x$ s.t. no $(x, \ldots)$ in $C$: $0$

For $x$ s.t. $(x, t) \in C$: $\frac{t}{S}$

Intuition:

- Maintain a subset of the input
- At any point if there are $k$ different items in the collection remove one copy of each

## Quick check:

Can underestimate? Yes. Drops some elements.

Can overestimate? No. Does not insert extra elements.

## How much can it underestimate?

- Consider $x \in X$
- estimate: $\frac{f(x) - \Delta_x}{s}$ ← $\Delta_x$ = # removed occurrences of $x$
- every time $x$ is removed, $k-1$ other items are removed
- So $k \cdot \Delta_x \leq s \Rightarrow \frac{\Delta_x}{s} \leq \frac{1}{k} \leq \varepsilon$

  ↑ set $k = \lceil 1/\varepsilon \rceil$

Conclusion:

$$\frac{f(x)}{s} - \varepsilon \leq \text{estimate} \leq \frac{f(x)}{s}$$

Space usage: $O(1/\varepsilon)$ elements of $X$ and counters

## Follow-up questions:

- Can handle deletions?
- Is linear sketch?
- Can combine summaries computed for separate subsets?

[See "Resources" for follow-up research question on Piazza]

# Heavy hitters = finding frequent elements

For instance:

For $0 < \alpha < \beta \leq 1$,

return subset $H \subseteq X$ s.t. for each $x \in X$,

if $\frac{f(x)}{s} \geq \beta \Rightarrow x \in H$

and if $\frac{f(x)}{s} < \alpha \Rightarrow x \notin H$

- Misra-Gries can be used to solve this problem

---

Discussion section tomorrow:

- Range query: "what fraction of items is in $[a, b]$?"
- You'll see how to use CountMin Sketch or Misra-Gries for this purpose.