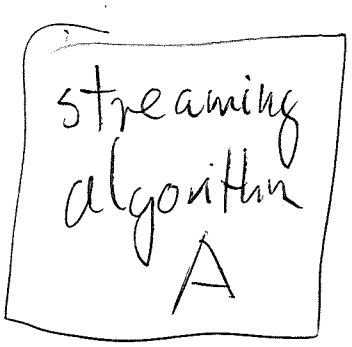
DS 563, Fall 2021, Lecture 2: Heavy Hitters

Streaming algorithms



- A reads and processes items one by one
- A should use much less space than the input stream size

Today: Heavy hitters (i.e. list frequent elements in the stream)

Setting: same as for Count Min Sketch

X – universe of items that we get to see

$f(x)$ – number of occurrences of $x \in X$

S – total number of items

Task: For some $\varepsilon \in (0,1)$ ← parameter

return $H \subseteq X$ s.t.

$\forall x \in X$: $f(x) \geq 2\varepsilon \cdot S \Rightarrow x \in H$

$f(x) \leq \varepsilon S \Rightarrow x \notin X$

Approach: find candidates, use Count Min sketch to verify

$H' \subseteq X$

output all $x \in H'$ for which Count Min sketch says fraction of occurrences $\geq 2\varepsilon$

How to find H'?

Warm-up: find element that occurs more than 1/2 of the time (called leader) or output nothing or any $z \in X$ if there is no leader

Algorithm:

- remember at most one element plus a count (= number of copies)
- when new item x arrives
  - if storage empty store x with a count of 1
  - otherwise, if same item in storage, increase the count
    if different items, throw away the new item + one copy of the item in storage (= decrease the count)
    If the count becomes 0, empty storage

Why works:

if x is a leader in multiset $Y \subseteq X$,
it will still be ~~[illegible]~~
a leader after removing two
different elements

If there is a leader,
it will be in the storage
at the end of the stream

Easy extension to finding elements
more frequent than $1/k$ fraction
- store at most k elements with counts
- if k different elements in storage,
  remove a copy of each

Homework: additional properties + estimate space usage