Counting Frequent Elements

stream $e_1, e_2, \dots, e_N \in [1, n]$

Denote by f_i the frequency of item $i \in [n]$

- Would like to know: which items occur frequently?

  (last class: majority)

- Maybe also estimate the frequencies?

Intuition for feasibility

- Suppose we're interested in items i with frequency $\geq t$
- $\leq \frac{N}{t}$ many such items

  ↳ in principle $\frac{N}{t}$ space is therefore a reasonable target

Algorithm Misra Gries (k)

Let D be an empty array ← attention: space-efficient implementation of "array"

For each item e_j in the stream do

    if e_j is a key in D, then

        $D[e_j] \leftarrow D[e_j] + 1$

    else

        if the number of keys in D is < k then

            add e_j to D with value 1

        else

            For each key ℓ in D do

                $D[\ell] \leftarrow D[\ell] - 1$

            } "decrement-all"

            remove all keys from D with value 0

return key-value pairs in D

Space complexity: $O(k \log N)$

<u>Theorem:</u> For each $i \in [n]$: $f_i - \frac{N}{k+1} \leq \hat{f}_i \leq f_i$

estimate returned by algorithm

all items not in D receive $\hat{f}_i = 0$

In particular, any item i with $f_i > \frac{N}{k}$ will be in D

<u>Proof:</u> "$\leq$" trivial

Since we only increase the counter of an item i whenever its true frequency f_i increases, $\hat{f}_i \leq f_i$ holds.

"$\geq$" #dec: number of times that decrement-all occurs over the course of the algorithm

$\#dec \cdot \underbrace{(k+1)} \leq N$

We can assign to each "decrement-all" uniquely $(k+1)$ items that we have read from the stream

→ in particular, k items that previously increased counters by 1 (and are then decremented) + the one item first read from stream def. Under-count of item i : $\Delta_i = f_i - \hat{f}_i$

Analyze change in Δ_i when item e_j is read from stream: (relevant cases)

(i) ($e_j = i$) : f_i and $\hat{f}_i$ increase both by 1 => Δ_i is unchanged

$i \in D$ (counter for i exists)

(ii) ($e_j = i$, $i \notin D$, counters already full) : f_i increases by 1, but $\hat{f}_i$ does not

→ Δ_i increases by 1

↳ triggers a decrement event

(iii) ($e_j \neq i$, counters full):

f_i is unchanged, but $\hat{f}_i$ (if it exists) is decremented by 1

→ Δ_i could increase by 1

↳ decrement event

Overall: $\Delta_i = f_i - \hat{f}_i \leq \#dec \leq \frac{N}{k+1}$

$$\Leftrightarrow f_i - \frac{N}{k+1} \leq \hat{f}_i$$