Mandatory: Streaming

Philip Bille Inge Li Gørtz Eva Rotenberg

1 Frequency Estimation Consider the following algorithm:

Algorithm 1: Count-a-Lot.

Input : stream $a_1, a_2, a_3, \ldots, a_m$ of elements $a_i \in [n]$, and the number of bins k.
Initialize each bin $b \in [k]$ with an element $e_b \leftarrow$ `null` and a counter $c_b \leftarrow 0$.
for *each element a_i in the stream* **do**
  if *a_i is the element in a bin b* **then**
    increment b's counter $c_b \leftarrow c_b + 1$
  else
    find the bucket $b_{\min}$ with the smallest counter value (breaking ties arbitrarily),
    replace its element $e_{b_{\min}} \leftarrow a_i$,
    increment its counter $c_{b_{\min}} \leftarrow c_{b_{\min}} + 1$.
end
Output: for each $i \in [n]$ output $\hat{f}_i$ as follows: if $e_b = i$ for some b, then $\hat{f}_i = c_b$, otherwise, $\hat{f}_i = 0$.

Here $[x]$ denotes the set of integers from 1 to x.

1.1 Let $c_{\min}$ be the value of the smallest counter after the entire stream has been processed, and let $e_b(i)$ be the element of bin b when $c_b = i$. Show that for any $i \in [c_{\min}]$ and any pair of bins $b \neq b'$ we have $e_b(i) \neq e_{b'}(i)$.

1.2 Consider an element i with $\hat{f}_i = 0$. Show that the true frequency f_i is such that $0 \leq f_i < m/k$. Note that this implies $|\hat{f}_i - f_i| < m/k$.

1.3 Consider an element i with $\hat{f}_i > 0$. Show that $|\hat{f}_i - f_i| < m/k$.

Heavy hitters Recall that an element i in a stream is a *heavy hitter* if it has frequency $f_i > m/k$ for some given k. We call an element an *infrequent element* if it has frequency less than $\frac{m}{3k}$.

2 Approximate Heavy Hitters using Count-a-Lot In this exercise we want a deterministic algorithm that outputs *all* heavy hitters and *no* infrequent elements.

2.1 Modify the Count-a-Lot algorithm such that it returns all heavy hitters and no infrequent elements. Remember to argue that your algorithm is correct.

2.2 Analyse the space usage, update time, and reporting time of your algorithm.