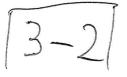
lecture 3/2023-01-26 DS-563/ CS-543 Today - Heavy hitters -Second moment estimation via AMS sketch Setting: X-universe from which elements of the stream come f(x) = # of occurrences of X in the stream = x EX = frequency of x $5 = total number of items = \sum_{X \in X} f(X)$ Heavy hitters: "find most frequent elements of the stream" Our task: For some EE (0,1) petarn HEXs.t. $\begin{array}{ccc} & f(x) & 7 & 2 \\ \hline \end{array} \\ & f(x) & f(x) & 2 \\ \hline \end{array} \\ & f(x) & (x) & (x) \\ & (x$ (In general, thresholds old < B< 1.) 3-1

In other words; - We want to include all heavy elements X $\left(\frac{f(x)}{2}, \frac{7}{2}\epsilon\right)$ - We don't want to output any light elements $(f(x)/s \leqslant \varepsilon)$ - "Grag zone!" elements in between 2 (fb) (25, can be output, but doult have to be Approach: - find candidates H'EX - Use Count Min Sketch to verify: output all XGH' for which Count Min Sketch says "(fraction of x) 7,2E" How to find small 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 & in X] we call this plus a count (= number of copies) J" 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 (x7y): replace (y, c) with (y, c-1) and if C-1=0, empty storage - at the end of the stream: if storage non-empty, output the element from X in it 3-3

Why this works:
- The algorithm keeps forgetting pairs
of clifferent elements
- If x is a leader in multiset set

$$S = 5'$$
 is $\{y_1z\}_1$ it's a leader in s'
 $y \neq z$
- Proof by induction:
 $S_0 = entine$ stream (multiset)
 $y_0 \neq z_0 \leq first$ two elements removed
 $S_1 = S_0 \setminus \{y_{0,1}z_0\}$
Single copies of $y_{0,1}z_0$ removed
If x is a leader in $S_0 \Rightarrow x$ is a leader
in S_1
Now: $y_1 \neq z_1 \in second$ two elements
removed
 \vdots
 $3-4$

If x is a leader in So
$$t = 14$$
 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 O
- at the and of the stream, output elements
in the storage
Back to original heavy hitting problem:
Say X = typical integers
two graphics To solve it with probability 93%:
Space $O(1/\epsilon) + O(\frac{1}{\epsilon} \log (1/\epsilon))$
finding CountMin Sketch
condidutes

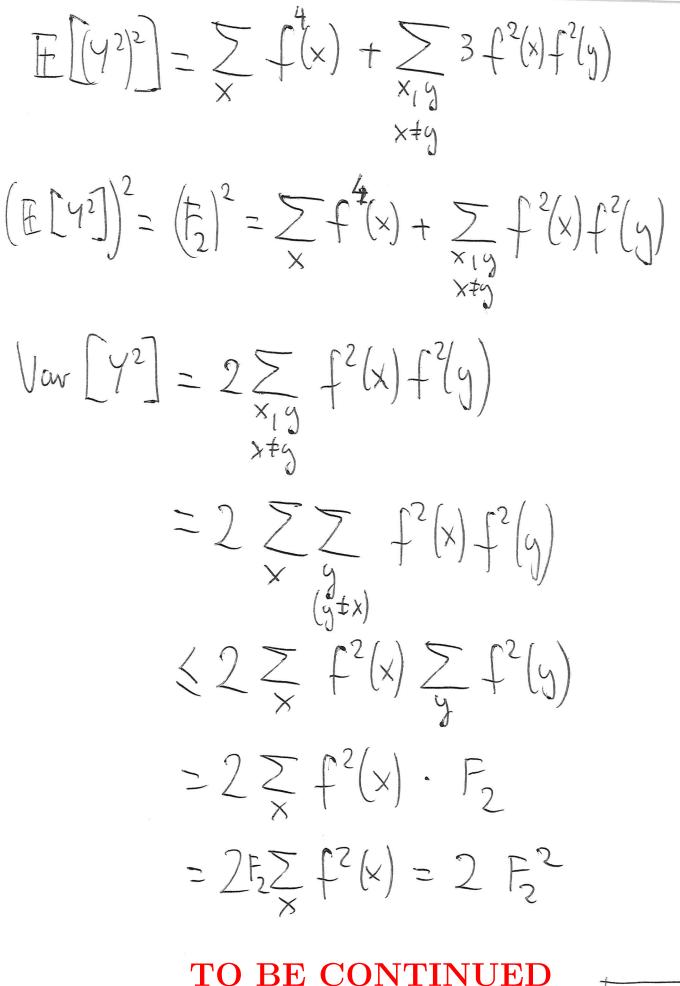
3-51

Frequency moments
Goal: upproximate
$$F_p = \sum (f(x))^p$$

 p -th moment
corner case $p=0$:
 $F_p = \left[\left\{ x \in X : f(x) \neq 0 \right\} \right]$
 $\#$ 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
 I
Alon-Motics-Szegedy, 1996
(clussic streaming paper, won important awards!)
 $h: X \Rightarrow \{-1, +1\} \in Handown hash function
 $selected from uniform$
 $distribution on all
functions $[3-6]$$$

Quich check: E[Y]? $E[Y] = \sum_{x \in X} f(x) \cdot E[h(x)] = 0$ $\mathbb{E}[Y^2] = \mathbb{E}\left[(\mathbb{Z}h(x)f(x))^2\right]$ $= \mathbb{E}\left[\sum_{x} h^{2}(x) f^{2}(x) + \sum_{x,y} h(x) h(y) f(x) f(y)\right]$ $= \sum_{x} f^{2}(x) + \sum_{x,y} f(x) f(y) E[h(x) \cdot h(y)]$ $= \sum_{x} f^{2}(x) = F_{2} \leq exactly what we want in expectation$ this situation: unbiased estimator Question: Is having an unbissed estimator good though?

	-		1
-		/	1
			I
		_	-7



3 -9