

Streaming:

Models:
 - Turnstile vs Cash Register
 - Streaming vs Semi-streaming
 - One pass vs. Multipass.

- Counter
- $F_0, F_2 \dots$
- heavy hitters
- median.

AMS: $F_p: \boxed{F_p := \sum_i (x_i(t))^p}$

$\boxed{\bar{x}_p(t) = \text{frequency vector @ time } t.}$

$C = \sum_{i=1}^{|U|} x_i \cdot h_i$ output C^2 .

$h_i: \text{universal } U \rightarrow \{-1, 1\}$.

Note: $EC^2 = \sum_i x_i^2$ by unbiasedness & indep.

$VC^2 = E(C^2)^2 - (EC^2)^2$

$\uparrow EC^4 = \sum_i x_i^4 + 6 \sum_{i \neq j} x_i^2 x_j^2$

$\Rightarrow VC^2 = 4 \sum_{i \neq j} x_i^2 x_j^2 \leq 2(EC^2)^2$

$\uparrow$ variance same order as expectation.

good model.

Chebyshev:

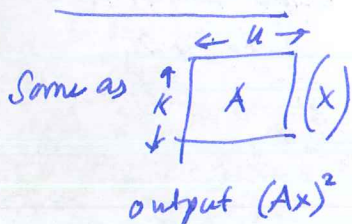
$P[|C^2 - EC^2| \geq \epsilon EC^2] \leq \frac{\text{Var}(C^2)}{\epsilon^2 EC^2} \leq \frac{2}{\epsilon^2}$

$\uparrow$ but repeat. $\text{Var}(Est) \leq k \cdot \text{orig variance}$
 $k \text{ times indep.}$ $Expect \leq k \cdot \text{orig expect.}$

or average (k time)

$\text{Var} \leq \frac{1}{k} \text{ orig var}$
 $Exp = \text{same}$

$\Downarrow$ tail prob $\leq \frac{2k}{\epsilon^2 k^2} \Rightarrow \text{set } k = \frac{2}{\epsilon^2} \cdot \frac{1}{\delta} \checkmark$



where rows of A drawn from 4-wise indep family. $\in \{\pm \frac{1}{\sqrt{k}}\}$

And $\frac{2}{\epsilon^2} \cdot \frac{1}{\delta}$ rows suffice. to show.

that $\|Ax\|^2 \leq \|x\|^2$. (TL!).

Tag of War sketch: get $E\langle Sx, Sy \rangle = \langle x, y \rangle$
 $\forall x, y$ $\text{Var}(\langle Sx, Sy \rangle) \leq \frac{2}{k} \|x\|^2 \|y\|^2$

$AB = \underbrace{A^T}_{\text{row}} \underbrace{B}_{\text{col}}$

$E \uparrow$ if entry $= \langle S_{ai}, S_{bj} \rangle = \langle a_i, b_j \rangle = (AB)_{ij}$ so its like ToW sketch!

Fo: exact requires $\Omega(\min(N, U))$ space.

apx: several ways.

Use the Flajolet Martin Idea:- smallest elt $\leq \frac{1}{d+1} \Rightarrow$ output $\frac{1}{\text{smallest}} - 1$.

↑ right expectation
high variance.

So instead:- keep track of s smallest values seen

(say hash fn goes from LL to $[0, 1]$). $\leftarrow$ actually to $[0, M]$ ~~but that's another detail for later.~~

Intuition: s th smallest element $\Theta \approx \frac{sM}{d} = L \in [0, M]$.
(if d distinct elts)

$\Rightarrow$ output ~~smallest~~ $\frac{MS}{L}$. (slightly off - but almost ok).

What about error prob?

Claim: (est off by factor of $1 \pm \epsilon$) $\leq O(1/s)$.

Pf: for estimator too small.

So say $\frac{MS}{L} < d/2 \Rightarrow$ position of s th smallest hash $> \frac{2MS}{d}$

estimator, where

$L \in [0, M]$ is position of
 s th smallest hash value.

$\Rightarrow$ # elts smaller than this $\leq s$
but expected # = $2s$.

$\Rightarrow$ we fell below expectation by
factor 4 2.

Calculation in
old lecture notes seem
incorrect.
Fix!

Heavy hitters: Count Min sketch.

• Misra Gries better analysis. Sps keep more counters. (Want k heavy hitters. Keep l regs).

Output $\hat{f}_1, \hat{f}_2, \dots, \hat{f}_k$. (Say n arrivals overall.)

Note: if d decrements, sum of counters = $n - d \cdot l$

$$= \sum_{i=1}^l \hat{f}_i \geq \sum_{i=1}^l (f_i - d) \geq \sum_{i=1}^k (f_i - d)$$

$$\Rightarrow d = \frac{F_{[k+1..n]}}{l-k} \Rightarrow \text{our loss} \leq F_{[k+1..n]} + d \cdot k = \left(1 + \frac{k}{l-k}\right) F_{[1..k]} - kd.$$

non neg of counters.