AA Lecture 6. 1/28/2021

HW1 due
HW2 on Course Works.

Heavy hitters:

f_x = frequency of x, $x \in [n]$.

x is ϕ-HH if $f_x \geq \phi \cdot \sum_y f_y = \phi m$.

Eg: $n = 2^{32}$.

CountMin Algo: linear sketch of f_x: Af.

of size: $O\left(\frac{\lg n}{\varepsilon\phi}\right)$ words

- returns set $H \subseteq [n]$ s.t., with Prob. $\geq 1 - 1/n$
  - if x is ϕ-HH then $x \in H$
  - if x is not $(1-\varepsilon)\phi$-HH, $x \notin H$.

$O(\lg m)$ bits.

Morris: can do approx $O\left(\frac{\lg \lg m}{\varepsilon}\right)$ bits

- estimator: estimate $\tilde{f}_x$ for any given x.

to output H:

for each $x \in [n]$
  $\tilde{f}_x = \min_i S_i[h_i(x)]$.
  if $\tilde{f}_x \geq \ldots$
    add x to H.
Output H.

Time: at least n (= universe of items)

very too large. (Note: $n >> m$)

Obs: $|H| \leq \frac{1}{(1-\epsilon)\phi}$.

Goal: to compute ϕ-HH in time $< cn$.

Thm: can get the same guarantees as CountMin (on HH), using:

- $O\left(\frac{\lg^2 n}{\phi}\right)$ time to find HHs.
- $O\left(\frac{\lg^2 n}{\epsilon\phi}\right)$ space (in words).

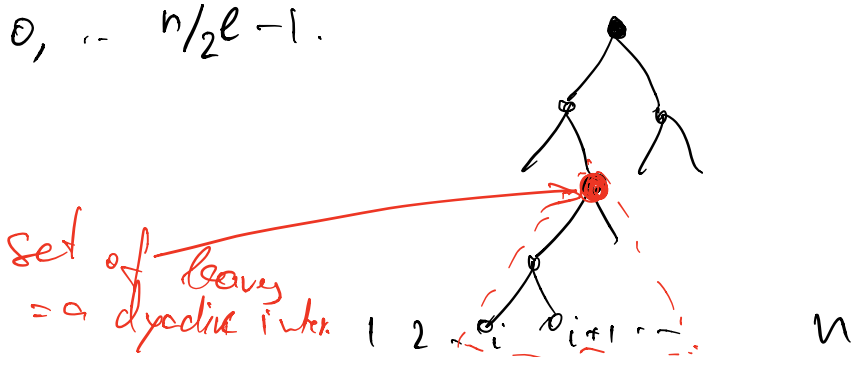
Proof: n is a power of 2.

Def: dyadic interval on $[n]$:

an interval $\left[i \cdot 2^{\ell}, (i+1)2^{\ell}\right]$

$\ell = 0, 1, 2, \ldots \lg_2 n$

$i = 0, \ldots n/2^{\ell} - 1$.

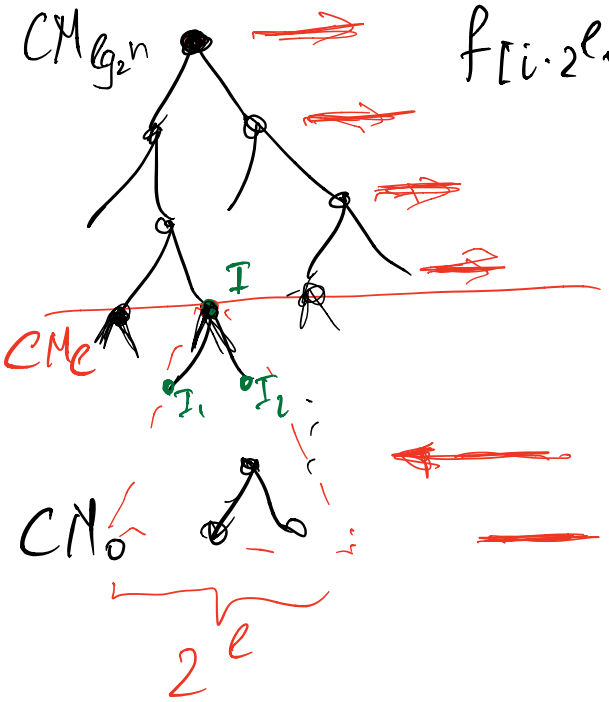


Fast Count Min:

— store $\lg_2 n + 1$ sketches CM

CM_ℓ : universe dyadic intervals @ level ℓ

$[1, 2^\ell], [2^{\ell+1}, 2 \cdot 2^\ell], \ldots$



$f_{[i \cdot 2^{\ell+1}, (i+1)2^\ell]}$ = sum frequencies of leafs.

each node has interval of len $= 2^\ell$

items $[1,2], [3,4], \ldots$

$f_{[1,2]} = f_1 + f_2$.

Like having $\lg_2 n + 1$ parrallel streams. each with its own CM sketch (indep. of each other)

Obs: fix d interval $I = I_1 \cup I_2$.

level ℓ ↖ $\ell - 1$

if node is HH

$f(I) = f(I_1) + f(I_2)$.

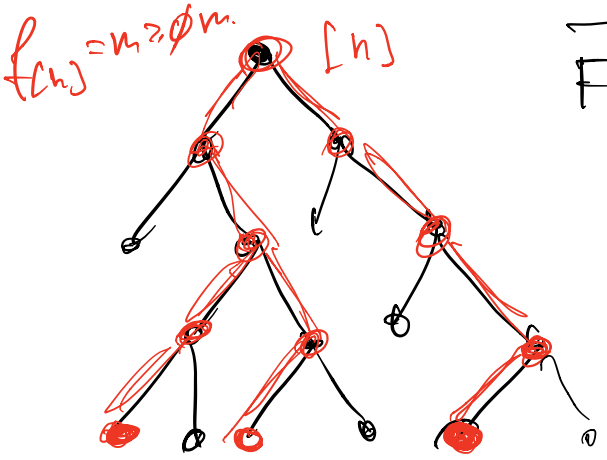
if I_1 or I_2 are ϕ-HHs $\Rightarrow I$ is a ϕ-HH.

$\Rightarrow$ parent is HH.

Pf: for every level ℓ $\sum_{I \text{ dyad. interv. @ level } \ell} f_I = m$.

if I_1 is ϕ-HH $\Rightarrow f_{I_1} \geq \phi \cdot m$

$\Rightarrow f_I = f_{I_1} + f_{I_2} \geq \phi m$. ☒



Facts: 1) for every node, can check if HH

2) in every level $\leq \frac{1}{(1-\varepsilon)\phi}$ HHs.

3) if node HH $\Rightarrow$ parent is too.

Algo to find all HH:

HH(node v):

if v is leaf, add to H, return.

Let c_1 & c_2 be children.

$\rightarrow$ use CM of c_1.

if c_1 is HH, HH(c_1)
if c_2 is HH, HH(c_2)

Space: $O(\lg n)$ CM sketches $\Rightarrow O\left(\frac{\lg^2 n}{\varepsilon\phi}\right)$.

Est time: $O(\lg n) \cdot 2 \cdot \frac{1}{(1-\varepsilon)\phi}$ calls to ed. f_x.

$\Rightarrow O\left(\frac{\lg^2 n}{(1-\varepsilon)\phi}\right) = O\left(\frac{\lg^2 n}{\phi}\right)$.

Max frequency: $\|f\|_\infty = \max_x f_x$.

Other norms of f?

$\|f\|_1 = \sum_x f_x = m$.

$\|f\|_2 = \left(\sum_x f_x^2\right)^{1/2}$ ← TODAY.

Frequency moments:

$F_p = \sum_x f_x^p$ $\quad F_2 = \|f\|_2^2$.

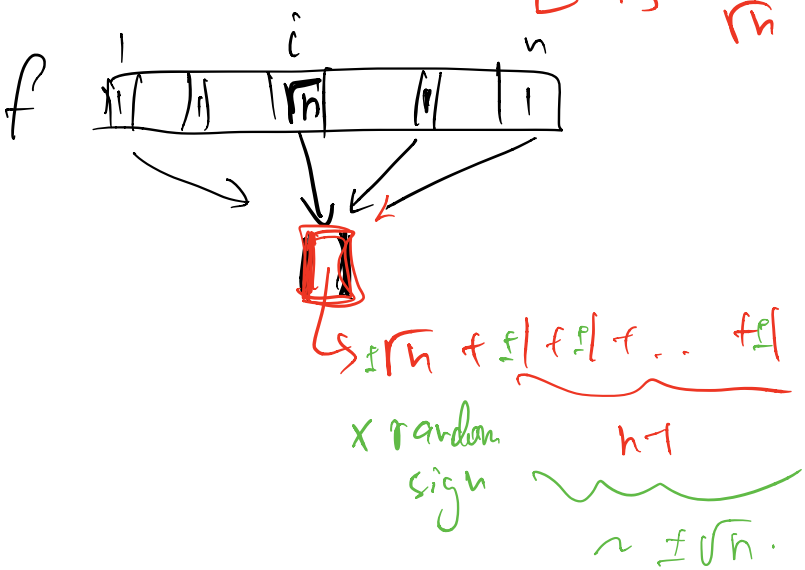
Today: $F_2 / \|f\|_2$.

Example: $f^{(1)} = (1, 1, \ldots, 1)$ $\quad F_2 = n$

vs $f^{(2)} = (1, 1, \ldots \sqrt{n}, \ldots, 1)$. $\quad F_2 = 2n - 1$. $\times 2$

Can't distin. f^1 vs f^2 using CM. unless space $\Theta(\sqrt{n})$.

D is $\frac{1}{\sqrt{n}}$-HH.



Tug-of-War:

- init: $z = 0$
  $\sigma : [n] \to \{\pm 1\}$ random.
- sketch: $z = \sum_x \sigma_x f_x$
  @ seeing item x: $z := z + \sigma_x$.
- estimator (for F_2): z^2.

Analysis:

Claim 1: $\mathbb{E}[z^2] = F_2$.

pf: $\mathbb{E}_\sigma[z^2] = \mathbb{E}\left[\left(\sum_x \sigma_x f_x\right)^2\right]$

$= \mathbb{E}\left[\sum_{x,y} \sigma_x \sigma_y f_x f_y\right]$

$= \sum_{x,y} \mathbb{E}[\sigma_x \sigma_y] \cdot f_x f_y$

if $x=y$, $\Rightarrow 1$

$x \neq y$, $\Rightarrow 0$

$= \sum_x f_x \cdot f_x = F_2$. ☒

Claim 2: $\mathrm{var}[z^2] \le O(F_2^2)$.

pf: $\mathrm{Var}[z^2] = \mathbb{E}[z^4] - (\mathbb{E}[z^2])^2$

$= \mathbb{E}\left[\left(\sum_x \sigma_x f_x\right)^4\right]$

$= \sum_{x,y,p,q} f_x f_y f_p f_q \cdot \mathbb{E}[\sigma_x \sigma_y \sigma_p \sigma_q]$

$= \sum_x f_x^4 + 3\sum_{x,y} f_x^2 f_y^2$.

$$\leq \left(\sum_x f_x^2\right)^2 + 3\left(\sum_x f_x^2\right)^2.$$

$$= 4F_2^2. \qquad \boxtimes$$

$\mathbb{E}[z^2] = F_2$

$\mathrm{Var}[z^2] \leq 4F_2^2.$

By Chebyshev: $\Pr\left[|z^2 - F_2| > 3\cdot\sqrt{4F_2^2}\right] \leq 1/9.$

$$z^2 = F_2 \pm 6F_2 \text{ with prob } \geq 8/9.$$

To get better concentration (for $1\pm\varepsilon$ factor approx) just "repeat" $k = \Theta(1/\varepsilon^2)$ times.

To W+:

- keep $k = \Theta(1/\varepsilon^2)$ counters $z_1, \ldots z_k$
- each $z_i = \sum_x \sigma_x^i f_x$, where $\sigma^i : [n] \to \{\pm 1\}$ iid random.
- Estimator: $z^2 = \frac{1}{k}\sum_{i=1}^{k} z_i^2.$

Claim 1: $\mathbb{E}[z^2] = F_2$.

Claim 2: $\mathrm{Var}[z^2] \le \frac{1}{k} \cdot 4F_2^2$.

By Chebyshev:

$$\Pr\left[\underbrace{|z^2 - F_2| > 3 \cdot \sqrt{\frac{4F_2^2}{k}}}_{\text{if this is false}}\right] \le 1/9.$$

$$\Rightarrow \boxed{z^2 = F_2 \pm \frac{6}{\sqrt{k}} F_2 = F_2 \cdot \left(1 \pm \frac{6}{\sqrt{k}}\right).}$$

set $k = 36/\varepsilon^2$.

$\Rightarrow z^2 = (1 \pm \varepsilon) F_2$ with probability $\ge 8/9$.