Heavy Hitters

- Count-Min
- Count-Sketch

[Misra-Greis '82] $\quad f_i - \frac{m}{K} \leq \tilde{f}_i \leq f_i \qquad m = F_1$

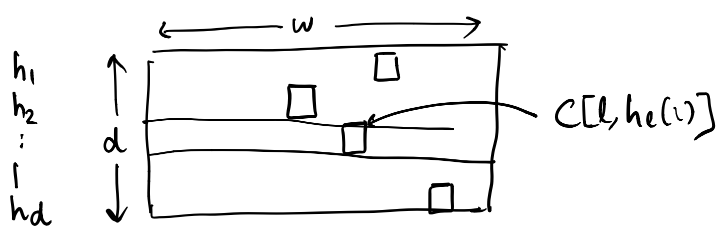
additive εF_1 approximation for $k = \frac{1}{\varepsilon}$

Storage $\frac{1}{\varepsilon}$

Count-Min [Cormode-Muthukrishnan] CM-sketch

Array of counters width w & depth d

For each row $i \in [d]$: $h_i : [n] \to [w]$



$h_1 \ldots h_d$: pairwise independent hash fn $[n] \to [w]$

For each element of stream
 $i \leftarrow$ current element
 for $\ell = 1$ to d
 $C[\ell, h_\ell(i)] \leftarrow C[\ell, h_\ell(i)] + 1$

Query (i)

$$\tilde{f}_i = \min_{\ell \in [d]} C[\ell, h_\ell(i)]$$

$$C[\ell, j] = \sum_{i : h_\ell(i) = j} f_i$$

Analysis : $\quad \mathbb{E}[C[\ell, h_\ell(i)]] \leq f_i + \frac{m}{w} \qquad m = F_1$

Say $h_\ell(i) = b$

$$C[\ell, h_\ell(i)] = \sum_{i' : h_\ell(i') = b} f_{i'}$$

$$\mathbb{E}[C[\ell, h_\ell(i)]] = f_i + \sum_{i' \neq i} \underbrace{\Pr[h_\ell(i') = b]} \cdot f_{i'}$$

$$= f_i + \frac{1}{w} \sum_{i' \neq i} f_{i'} \leq f_i + \frac{m}{w}$$

Also $C[\ell, h_\ell(i)] \geq f_i$

$$P\left[C[\ell, h_\ell(i)] \geq f_i + \frac{2m}{w}\right] = P\left[C[\ell, h_\ell(i)] - f_i \geq \frac{2m}{w}\right]$$

$$\leq \frac{\mathbb{E}[C[\ell, h_\ell(i)] - f_i]}{\frac{2m}{w}} \leq \frac{1}{2}$$

$h_1 \ldots h_d$ are independent

$$P\left[\tilde{f}_i \geq f_i + \frac{2m}{w}\right] = \prod_{\ell \in [d]} P\left[C[\ell, h_\ell(i)] \geq f_i + \frac{2m}{w}\right]$$

$$\leq \left(\frac{1}{2}\right)^d$$

$w = \frac{2}{\varepsilon} \Rightarrow \frac{2m}{w} = \varepsilon m$

$d = \log_2 \frac{1}{\delta}$, $\left(\frac{1}{2}\right)^d = \delta$

$$\Pr[\tilde{f}_i \geq f_i + \varepsilon m] \leq \delta$$

$f_i \leq \tilde{f}_i \leq f_i + \varepsilon m$ w. prob. $1 - \delta$

Space usage $O\left(\frac{1}{\varepsilon} \log_2 \frac{1}{\delta}\right)$

Count-Sketch [C, Chen, Farach-Colton '02]

$h_1 \dots h_d$ pairwise independent hash f^n $[n] \to [w]$

$s_1 \dots s_d$ —— " —— $[n] \to \{\pm 1\}$

For each element of stream

$i \leftarrow$ current elt.

for $\ell = 1$ to d

$$C[\ell, h_\ell(i)] \leftarrow C[\ell, h_\ell(i)] + s_\ell(i)$$

Query (i)

$$\tilde{f}_i = \underset{\ell = 1 \text{ to } d}{\text{median}} \{ C[\ell, h_\ell(i)] \cdot s_\ell(i) \}$$

<u>Analysis:</u>

Fix $i \in [n]$

$$Z_\ell = C[\ell, h_\ell(i)] \cdot s_\ell(i)$$

For $i' \in [n]$ $Y_{i'}$ indicator $\begin{cases} 1 & \text{if } h_\ell(i') = h_\ell(i) \\ 0 & \text{otherwise} \end{cases}$

$$\mathbb{E}[Y_{i'}^2] = \mathbb{E}[Y_{i'}] = \frac{1}{w}$$

$$Z_\ell = C[\ell, h_\ell(i)] \cdot s_\ell(i) = s_\ell(i) \sum_{i'} Y_{i'} \cdot f_{i'} \cdot s_\ell(i')$$

$$\mathbb{E}[Z_\ell] = f_i + \sum_{i' \neq i} \mathbb{E}[\underbrace{s_\ell(i)\, s_\ell(i') \cdot Y_{i'}}]\, f_{i'}$$

$\mathbb{E}[s_\ell(i')\, s_\ell(i)] = 0$ for $i' \neq i$

$Y_{i'}$ independent of $s_\ell(i)$, $s_\ell(i')$

$$\mathbb{E}[Z_\ell] = f_i$$

$$\text{Var}[Z_\ell] = \mathbb{E}\Big[\Big(\sum_{i' \neq i} s_\ell(i)\, s_\ell(i') \cdot Y_{i'}\, f_{i'}\Big)^2\Big]$$

$$= \mathbb{E}\Big[\sum_{i' \neq i} f_{i'}^2\, Y_{i'}^2 + \sum_{\substack{i' \neq i'' \\ i', i'' \neq i}} f_{i'} f_{i''}\, \underbrace{s_\ell(i')\, s_\ell(i'')\, Y_{i'}\, Y_{i''}}\Big]$$

$$= \sum_{i' \neq i} f_{i'}^2 \, \mathbb{E}[Y_{i'}^2]$$

$$\leq \frac{\|f\|_2^2}{w}$$

$$w = \frac{3}{\varepsilon^2}$$

$$\Pr[|Z_e - f_i| \geqslant \varepsilon \|f\|_2] \leq \frac{Var[Z_e]}{\varepsilon^2 \|f\|_2^2} \leq \frac{1}{\varepsilon^2 w} \leq \frac{1}{3}$$

Now via Chernoff bound

$$\Pr[|\text{median}\{Z_1 .. Z_d\} - f_i| \geqslant \varepsilon \|f\|_2] \leq e^{-cd} \leq \delta$$

by choosing $d = O(\log(\frac{1}{\delta}))$

Space: $O(\frac{1}{\varepsilon^2} \log(\frac{1}{\delta}))$

Comparison	Guarantee	Space
Misra-Greis	$f_i - \varepsilon\|f\|_1 \leq \tilde{f}_i \leq f_i$	$\frac{1}{\varepsilon}$
Count-Min	$\underbrace{f_i \leq \tilde{f}_i \leq f_i + \varepsilon \|f\|_1}_{\text{w. prob. } 1-\delta}$	$O(\frac{1}{\varepsilon} \log(\frac{1}{\delta}))$
Count-Sketch	$\|f_i - \tilde{f}_i\| \leq \varepsilon \|f\|_2$ w. prob. $1-\delta$	$O(\frac{1}{\varepsilon^2} \log(\frac{1}{\delta}))$

$$\|f\|_1 \geqslant \|f\|_2$$

all 1's $\quad n \quad \sqrt{n}$

Heavy tailed $\quad f_i \sim \frac{1}{\sqrt{i}}$

$$\|f\|_1 = \theta(\sqrt{n}) \qquad \|f\|_2 = \theta(1)$$

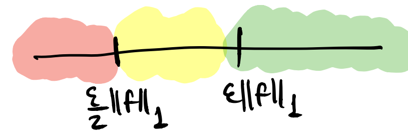
We solved: point queries
estimate f_i to within $\pm \varepsilon \|f\|_1$
$\pm \varepsilon \|f\|_2$

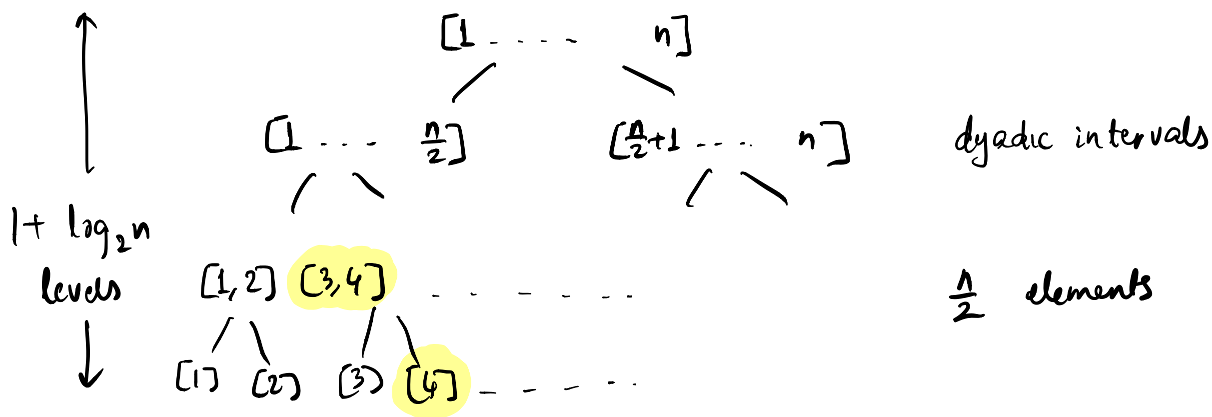
Heavy Hitters:

Output L such that

$f_i \geq \varepsilon \|f\|_1 \Rightarrow i \in L$

$f_i < \frac{\varepsilon}{2} \|f\|_1 \Rightarrow i \notin L$





One CM sketch for each level.

α-heavy hitters w. failure prob. $\leq \delta$

each CM sketch has error parameter $\varepsilon = \frac{\alpha}{4}$

failure prob. $\eta = \frac{\delta}{(\log n)} \cdot \frac{\alpha}{4}$

At each level j of tree, track L_j: heavy hitters at level j

L_j: contain all α-heavy hitters at its level
no one below $\frac{\alpha}{2}$ heavy

For each of 2 children, point query child using CM sketch at level $j+1$

If child has point query $\geq \left(\frac{3\alpha}{4}\right) \|x\|_2$

include in L_{j+1}

L: list at leaf level.

Conditioned on correctness (no failure)

L_j has size $\leq \frac{2}{\alpha}$

we query $\leq \frac{4}{\alpha}$ intervals at level $j+1$

At most $Q \leq \frac{4}{\alpha} \log n$

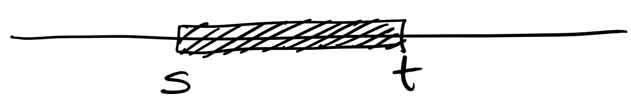
We chose parameters so that each CM sketch has

failure prob $\leq \frac{\delta}{Q} = \frac{\delta \cdot \alpha}{4 \log n}$

union bound, failure prob. $\leq \delta$

Interval queries.

$\sum_{i \in [s,t]} f_i$



interval queries via point queries:

query time linear in interval size

error scales linearly in interval size

Instead:

Use interval based data structure

Each interval $[s,t]$ is union of at most $2 \log n$ dyadic intervals

Now query time $= O(\log n)$ (Query time of CM sketch)

error $\leq O(\log n) \cdot$ (error of CM-sketch at each level)