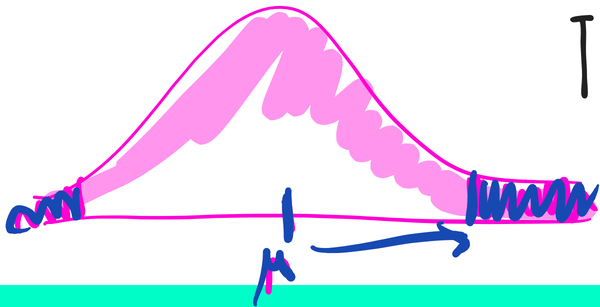
Heavy Hitters (continued)
Tail Bounds



Anna Karlin

Problem

- Input: sequence of n elements $x_1, x_2, \ldots, x_n$ from a known universe U (e.g., 8-byte integers).

- Goal: perform a computation on the input, in a single left to right pass where
  - Elements processed in real time
  - Can’t store the full data. => minimal storage requirement to maintain working “summary”

Heavy Hitters: Keys that occur many times

x_1 32, x_2 12, x_3 14, 32, 7, x_6 12, 32, 7, 32, 12, x_{11} 4,

$f_{32}^{6} = 2$ $\quad f_{32}^{7} = 3$

$x_1, x_2, \ldots, x_n$

Applications:

- Determining popular products
- Computing frequent search queries
- Identifying heavy TCP

$\boxed{k = 100}$

$\forall\ t \le n$ $\quad \boxed{f_x^t}$: # times element x has appeared in $x_1, x_2, \ldots, x_t$

$f_x^n \ge 0.01\,n$

$\le$ 100 such elements

Goal: Find/output all elements x with $f_x^n > \frac{n}{k}$

(These elts are "heavy hitters")

$\Downarrow$

much smaller than n space impossible

$\boxed{\text{Output has size } O(k)}$

$\boxed{\text{Provably impossible to solve this problem exactly with sublinear space}}$

Modified goal: (solve ε-HH problem)

① If $f_x^n \geq \frac{n}{k}$, x added to HH list

② If some elt, say y, added to list, then w.p. $\geq 1-\delta$,

$$f_y^n \geq \frac{n}{k} - \varepsilon n$$

outputting a list of elts we claim are HHs.

if an elt y in list

w. prob $\geq 1-\delta$

$$f_y^n \geq \frac{n}{k} - \varepsilon n$$

Example:

$k = 20$ $\quad \frac{n}{k} = \frac{n}{20} = 0.05n$

$\varepsilon \doteq 0.01$

$$f_y^n \geq 0.05n - 0.01n = 0.04n$$

Count-min sketch

- Maintain a short summary of the information that still enables answering queries.
- Cousin of the Bloom filter
  - Bloom Filter solves the “membership problem”.
  - We want to extend it to solve a counting problem.

COUNT-MIN SKETCH

Modified goal: (solve ε-HH problem)

1. If $f_x^n \geq \frac{n}{k}$, x added to HH list
2. If some elt, say y, added to list, then w.p. $\geq 1-\delta$
$$f_y^n \geq \frac{n}{k} - \varepsilon n$$

designer specifies k, ε, δ $\Rightarrow$ $\boxed{b, \ell}$

Keep 2D array
ℓ hash tables, each of size b.

initialize tables with all 0s

when elt x shows up.

Update (x): $\forall \; 1 \le j \le \ell$ increment $t_j[h_j(x)]$

Count (x): return $\boxed{\min\limits_{1 \le j \le \ell} t_j[h_j(x)]}$

if $\text{Count}(x) \ge \frac{n}{k}$, add x to HH list

$$\boxed{t_j[h_j(x)] \ge f_x^t}$$

$$\boxed{\text{Count}(x) \ge f_x^t}$$

		1	2	3	4	5				b
h_1	1	1		4						
h_2	2		1		1	3				
h_ℓ	ℓ									

t_1 first row
t_2 second row

Example $\ell = 2$

x_1	x_2	x_3	x_4	x_5
x	y	x	z	x

Count(x)
return 3

$$\boxed{Z_2^5 = t_2[h_2(x)] = 3}$$

Suppose

$h_1(x) = 3 \quad h_2(x) = 5$
$h_1(y) = 3 \quad h_2(y) = 2$
$h_1(z) = 1 \quad h_2(z) = 4$

initialize tables with all 0s

when elt x shows up.

Update (x): $\forall\ 1 \le j \le \ell$ increment $t_j[h_j(x)]$

Count (x): return $\min\limits_{1 \le j \le \ell} t_j[h_j(x)]$

if $\text{Count}(x) \geq \frac{n}{k}$, add x to HH list

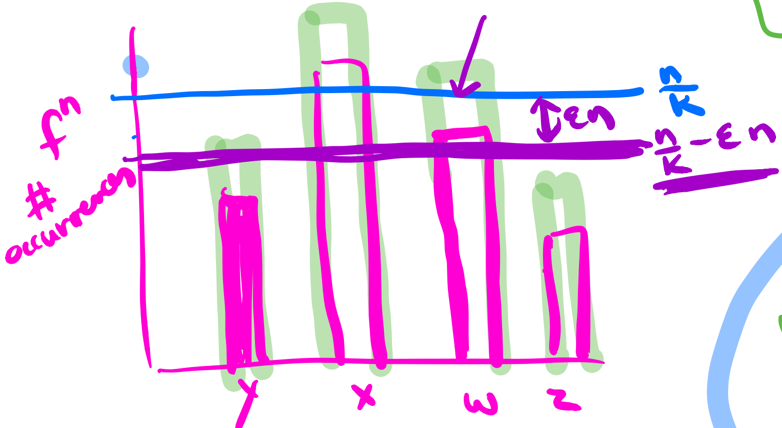
Observations

$\forall j,\ \forall x,\ \forall t$

$t_j[h_j(x)] \geq f_x^t$

$\Rightarrow$

$$\text{Count}(x) = \min_{1 \le j \le \ell} t_j[h_j(x)] \geq f_x^t$$



$x_1 \ldots\ldots x_n$

green show current Count

$\forall y$ that is output

w.p. $\geq 1-\delta$ $\quad f_y^n \geq \frac{n}{k} - \varepsilon n$

$\frac{n}{k} = 0.05n$

$\boxed{0.04n}$

$\varepsilon n = 0.01n$

$b, \ell.$

Assumptions

① hash functions behave like random maps

$h_1, \ldots, h_\ell : U \longrightarrow \{0, 1, \ldots, b-1\}$

$\forall x \neq y \quad \Pr(h_j(x) = h_j(y)) = \frac{1}{b}$

② hash fns $h_1, \ldots, h_\ell$ are indep of each other.

initialize tables with all 0s

when elt x shows up.

Update (x): $\forall \; 1 \leq j \leq \ell$ increment $t_j[h_j(x)]$

Count (x): return $\min\limits_{1 \leq j \leq \ell} t_j[h_j(x)]$

if Count$(x) \geq \frac{n}{k}$, add x to HH list

Fix time t; $x_1, x_2, \dots, x_t$ have just arrived

$$Z_j^t \triangleq t_j[h_j(x)]$$

$$Z_j^t \geq f_x^t$$

$$Z_j^t = f_x^t + \sum_{y \neq x} f_y^t \, w_{xy}$$

$$\boxed{E(Z_j^t) = f_x^t + \sum_{\text{distinct } y \neq x} f_y^t \, E(w_{xy})}$$

$y \in \{x_1, \dots, x_t\}$

$$w_{xy} = \begin{cases} 1 & h_j(x) = h_j(y) \\ 0 & \text{otherwise} \end{cases}$$

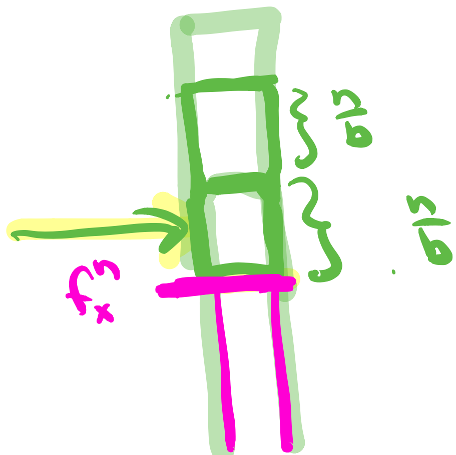
$$= f_x^t + \sum_{y \neq x} f_y^t \cdot \frac{1}{b}$$

$$= f_x^t + \frac{1}{b} \boxed{\sum_{y \neq x} f_y^t}$$

$$= t - f_x^t \leq t \underset{\text{at end}}{\leq} n$$

$$\leq f_x^t + \frac{n}{b}$$

$$\boxed{E(Z_j^t) - f_x^t \leq \frac{n}{b}}$$



$$\Pr\left(z_j^t - f_x^t \geq 2\tfrac{n}{b}\right) = ?$$

By Markov's Inequality, if $X \geq 0$ then

$$\Pr\left(X > 2E(X)\right) \leq \tfrac{1}{2}$$

$\Rightarrow$ Let $X \triangleq z_j^t - f_x^t$

$\Rightarrow X \geq 0$ z_j^t is overestimate of f_x^t

$$\Pr\left(z_j^t - f_x^t \geq 2E(X)\right) \leq \tfrac{1}{2}$$

$\forall j, \forall t \leq n$

$$(*) \quad \Pr\left(z_j^t - f_x^t \geq \tfrac{2n}{b}\right) \leq \Pr\left(z_j^t - f_x^t \geq 2E(X)\right) \leq \tfrac{1}{2}$$

$E(X) \leq \frac{n}{b}$

Putting it all together

$$\Pr\left(\text{Count}(x) - f_x^n \geq \tfrac{2n}{b}\right)$$

$$= \Pr\left(\min(z_1^n, \ldots, z_\ell^n) - f_x^n \geq \tfrac{2n}{b}\right)$$

$$= \Pr\left(z_1^n - f_x^n \geq \tfrac{2n}{b},\ z_2^n - f_x^n \geq \tfrac{2n}{b},\ \ldots,\ z_\ell^n - f_x^n \geq \tfrac{2n}{b}\right)$$

$$= \prod_{j=1}^{\ell} \Pr\left(z_j^n - f_x^n \geq \tfrac{2n}{b}\right) \leq \underbrace{\tfrac{1}{2} \cdot \tfrac{1}{2} \cdots \tfrac{1}{2}}_{\ell} = \tfrac{1}{2^\ell}$$

$(*)$

independence of hash fns for different tables

$$\boxed{\Pr\left(\text{Count}(x) - f_x^n \geq \tfrac{2n}{b}\right) \leq \tfrac{1}{2^\ell}}$$

$k, \varepsilon, \delta.$

Modified goal: (solve ε-HH problem)

① If $f_x^n \ge \frac{n}{k}$, x added to HH list

② If some elt, say y, added to list, then w.p. $\ge 1-\delta$

HH

$f_y^n \ge \frac{n}{k} - \varepsilon n$

$\varepsilon \cancel{n} = \frac{2\cancel{n}}{b}$

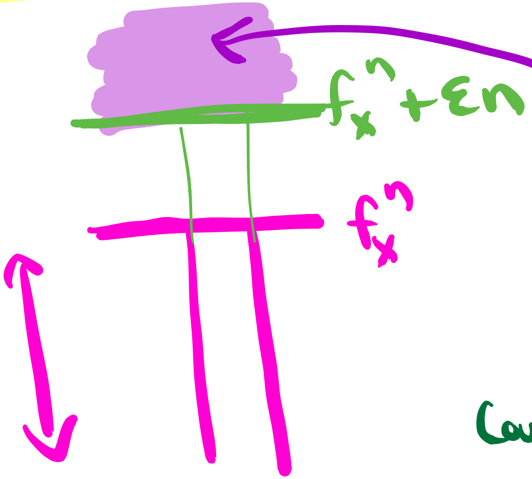
$\Rightarrow \boxed{b = \frac{2}{\varepsilon}}$ (*)

$\delta = \frac{1}{2^\ell}$

$\Rightarrow 2^\ell = \frac{1}{\delta}$

$\Rightarrow \boxed{\ell = \log_2\left(\frac{1}{\delta}\right)}$ (*)

$$\Pr\Big(\text{Count}(x) - f_x^n \ge \underbrace{\boxed{\frac{2n}{b}}}_{\varepsilon n}\Big) \le \underbrace{\boxed{\frac{1}{2^\ell}}}_{\delta}$$



Prob Count(x) ends up in purple $\le \delta$

If y added to HH list, $\Rightarrow$

Count(y) $\ge \frac{n}{k} \Rightarrow$ w.p. $\ge 1-\delta$ $f_y^n \ge \frac{n}{k} - \varepsilon n$

Count-Min Sketch

- Elegant small space data structure. $O(b \cdot \ell + k)$
- Space used is independent of n.
- Is implemented in several real systems.
  - AT&T used in network switches to analyze network traffic.
  - Google uses a version on top of Map Reduce parallel processing infrastructure and in log analysis.
- Huge literature on sketching and streaming algorithms (algorithms like Distinct Elements, Heavy Hitters and many many other very cool algorithms).

Hash functions

Say hashing n 32 bit integers into table of size b.

Pick prime number $p > \min(n, 2^{32})$

$$\boxed{\mathcal{H}} = \left\{ h_{e,g}(x) = \left[(ex+g) \bmod p\right] \bmod b \;\middle|\; \begin{matrix} 1 \le e \le p-1 \\ 0 \le g \le p-1 \end{matrix} \right\}$$

$\in 0, \ldots, b-1$

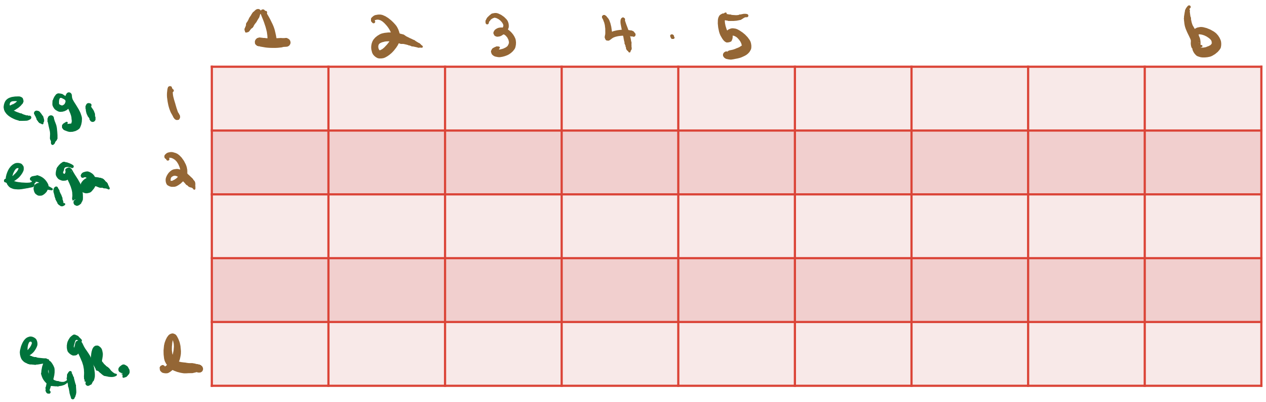
family of hash fns: $(p-1) \cdot p$ # of fns in family.

If h is chosen uniformly at random from $\mathcal{H}$

$$\forall x \neq y \quad \Pr(h(x) = h(y)) \le \frac{2}{b}.$$

ℓ diffent hash fns

each selected uniformly at random
and independently from $\mathcal{H}$



6.1 Tail bounds

Most Slides by Joshua Fan and Alex Tsun

Agenda

- Markov's Inequality
- ~~Chebyshev's Inequality~~
- ~~The law of large numbers~~

Markov's Inequality (Intuition)

The score distribution of an exam is modelled by a rv X with range $\Omega_X \subseteq [0,110]$ (for extra credit).

If the average was $E[X] = 50$, at most what percentage of the class could have gotten 100 (or higher)?

at most 50% of class could have gotten score of 100 or higher.

$\leq$

a) 100%

$\boxed{\text{b) } 50\%}$

c) 25%

d) no bound

Pf by $\rightarrow\leftarrow$

suppose > 50% got score ≥ 100

$$E(X) = \sum_{\text{scores}} \underbrace{\text{score}}_{\geq 100} \underbrace{\Pr(X=\text{score})}_{>50\%} > 50$$

$$> 100 \cdot \tfrac{1}{2} > 50$$

Markov's Inequality (Intuition)

The score distribution of an exam is modelled by a rv X with range $\Omega_X \subseteq [0,110]$ (for extra credit).

If the average was $E[X] = 50$, at most what percentage of the class could have gotten 100 (or higher)?

$\frac{1}{2}$

If the average was $E[X] = 25$, at most what percentage of the class could have gotten 100 (or higher)?

at most $\frac{1}{4}$ can get score ≥ 100

$\leq$

a) 100%

b) 50%

$\boxed{\text{c) 25\%}}$

d) no bound

Markov's Inequality (Intuition)

The score distribution of an exam is modelled by a rv X with range $\Omega_X \subseteq [0,110]$ (for extra credit).

Ω_X all #s ≤ 110

If the average was $E[X] = 50$, at most what percentage of the class could have gotten 100 (or higher)?

$$\frac{1}{2}$$

If the average was $E[X] = 25$, at most what percentage of the class could have gotten 100 (or higher)?

$$\frac{1}{4}$$

$\leq$

a) 100%

b) 50%

c) 25%

d) no bound

What if you could get a negative score?

Any r.v. with $\boxed{E(X) = 50}$

$0.01(-4900) + 0.99 \cdot 100 = 50$

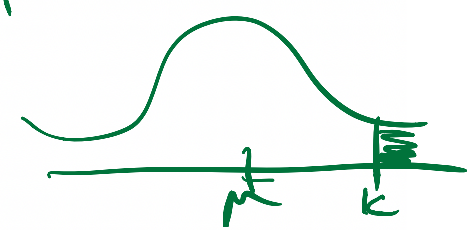
0.01	−4900
0.99	100
0.99999	

Markov's Inequality

Markov's Inequality: Let $X \geq 0$ be a **nonnegative** random variable (discrete or continuous), and let $k > 0$. Then,

bound on upper tail

$$P(X \geq k) \leq \frac{E[X]}{k}$$



$$E(X) = \sum_{\substack{x \mid x<k \\ x \in \Omega_X}} x \Pr(X=x) + \sum_{\substack{x \mid x \geq k \\ x \in \Omega_X}} x \Pr(X=x)$$

≥ 0

$$\geq \sum_{x \mid x \geq k} x \Pr(X=x) \geq k \sum_{x \mid x \geq k} \Pr(X=x) = k \Pr(X \geq k)$$

$x \geq k$

$E(X) \geq k \Pr(X \geq k)$

$\equiv \quad \Pr(X \geq k) \leq \frac{E(X)}{k}$

Markov's Inequality

Markov's Inequality: Let $X \geq 0$ be a **nonnegative** random variable (discrete or continuous), and let $k > 0$. Then,

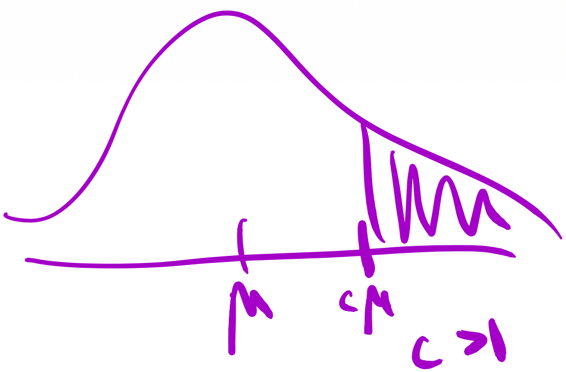
$$\boxed{P(X \geq k) \leq \frac{E[X]}{k}} \implies \Pr(X \geq cE(X)) \leq \frac{\cancel{E(X)}}{c\cancel{E(X)}} = \frac{1}{c}$$

Alternatively, Setting $k = cE(X)$

$$\boxed{P(X \geq cE[X]) \leq \frac{1}{c}}$$

$c < 1$

$\frac{1}{c} > 1$



$c = 2$

$\Pr(X \geq 2E(X)) \leq \frac{1}{2}$

for a $X \geq 0$

Markov's Inequality (Proof)

Markov's Inequality: Let $X \geq 0$ be a **nonnegative** rv and let $k > 0$. Then,

$$P(X \geq k) \leq \frac{E[X]}{k}$$

Proof (Markov):

$X \geq 0$

$$E[X] = \int_0^\infty x f_X(x)dx = \int_0^k x f_X(x)dx + \int_k^\infty x f_X(x)dx$$

$$\geq \int_k^\infty x f_X(x)dx \geq \int_k^\infty k f_X(x)dx = k\int_k^\infty f_X(x)dx = kP(X \geq k)$$

Rearranging gives

$$P(X \geq k) \leq \frac{E[X]}{k}$$