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. $\Rightarrow$ minimal storage requirement to maintain working “summary”
Heavy Hitters: Keys that occur many times

\[ x_1, x_2, x_3, \ldots, x_n \]

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

Goal: Find/output all elements \( x \) with \( f_x^+ \geq \frac{n}{k} \) (These are "heavy hitters")

\[ f_{x_1}^+ = 6, f_{x_2}^+ = 3 \]

\[ x_1, x_2, \ldots, x_n \]

\[ k = 100, \quad f_{x_i} \geq 0.01n \]

\[ \leq 100 \text{ such elements} \]
Output has size $O(k)$

Provably impossible to solve this problem exactly with sublinear space

Modified goal: (solve $\epsilon$-HHT problem)

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

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

   \[
   f_y^n \geq \frac{n}{k} - \epsilon \cdot n
   \]

Example:

$k = 20, \quad \frac{2}{k} = \frac{20}{20} = 0.05n$

$\epsilon = 0.01$

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

much smaller than $n$ space impossible
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 \( f_x \geq \frac{n}{k} \), \( x \) added to HH list
2. If some elt, say \( y \), added to list, then w.p. \( \geq 1 - \delta \)
   \[ f_y \geq \frac{n}{k} - \epsilon n \]

**designer specifies** \( k, \epsilon, \delta \) \( \Rightarrow \) \( b, \epsilon \)

Keep \( 2D \) away.

2 hash tables, each of size \( b \).
Initialize tables with all 0s when elt $x$ shows up.

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

Count($x$): return $\min_{1 \leq j \leq l} t_j[h_j(x)]$

If Count($x$) $\geq \frac{n}{k}$, add $x$ to hit list.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example $l=2$

Suppose

$\text{Count}(x)$

Return 3

$\sum_{x \in S} \text{Count}(h_0(x)) = 3$
Initialize tables with all 0s when element $x$ shows up.

**Update** $(x)$: For $1 \leq j \leq L$, increment $t_j[h_j(x)]$.

**Count** $(x)$: Return $\min_{1 \leq j \leq L} t_j[h_j(x)]$.

If $\text{Count}(x) \geq \frac{n}{k}$, add $x$ to HHT list.

**Observations**

\[ A_j, A_x, A_t \]

\[ t_j[h_j(x)] = f_x \]

\[ \Rightarrow \]

\[ \text{Count}(x) = \min_{1 \leq j \leq L} t_j[h_j(x)] = f_x \]

**Graph**

Green shows current Count

For any $x$, that is output

$\omega_p \geq 1 - \delta$ if $f_y \geq \frac{n}{k} - \varepsilon n$
Assumptions

1. Hash functions behave like random maps
   \[ h_1, \ldots, h_b : \{0, 1, \ldots, b-1\} \]
   \[ \forall x \neq y \quad \Pr(h_j(x) = h_j(y)) = \frac{1}{b} \]

2. Hash functions are independent of each other.

Initialize tables with all Os

When elt x shows up:

Update \((x)\): \(\forall 1 \leq j \leq b\) increment \(t_j[h_j(x)]\)

Count \((x)\): return \(\min_{1 \leq j \leq b} t_j[h_j(x)]\)

if \(\text{Count}(x) \geq \frac{n}{k}\), add \(x\) to HT list.
Fix time $t_j$ have just arrived

$Z^+_j \equiv t_j \left[ h_j(x) \right]$

$$E(Z^+_j) = f^+_x + \sum_{y \neq x} f^+_y E(w_{xy})$$

$$= f^+_x + \sum_{y \neq x} f^+_y \cdot \frac{1}{b}$$

$$= f^+_x + \frac{1}{b} \left[ \sum_{y \neq x} f^+_y \right]$$

$$= f^+_x + \frac{1}{b} \left[ f^+_x + \frac{n}{b} \right]$$

$$\leq f^+_x + \frac{n}{b}$$

$$E(Z^+_j) - f^+_x \leq \frac{n}{b}$$
Let $X \triangleq z_j^t - f_x^t$ be non-negative. By Markov's Inequality,

$$\Pr(X > 2\mathbb{E}(X)) = \frac{1}{2}$$

for any $X > 0$. Therefore,

$$\Pr(z_j^t - f_x^t > 2\mathbb{E}(X)) = \frac{1}{2}$$

Putting it all together,

$$\Pr(\text{Count}(x) - f_x^a > \frac{2n}{b})$$

is equivalent to

$$\Pr(\min_{j} (z_j^n - f_x^n) - f_x^n > \frac{2n}{b})$$

which can be expressed as

$$\Pr(Z_j^n - f_x^n > \frac{2n}{b}, Z_j^n - f_x^n > \frac{2n}{b}, \ldots, Z_j^n - f_x^n > \frac{2n}{b})$$

for each $j$. Using independence of hash keys for different tables, we have

$$\prod_{j=1}^n \Pr(Z_j^n - f_x^n > \frac{2n}{b}) = \frac{1}{2} \cdot \frac{1}{2} \cdots = \frac{1}{2^n}$$

Therefore,

$$\Pr(\text{Count}(x) - f_x^a > \frac{2n}{b}) = \frac{1}{2^n}$$
Modified goal: (solve E-H H problem)

1. If \( f_x^n \geq \frac{n}{k} \), \( x \) added to \( H \) list
2. If some elt, say \( y \), added to \( H \) list, then up. \( \geq 1 - \delta \) \( \implies \) \( f_y^n \geq \frac{2^n}{k} - \frac{\epsilon}{n} \)

Pr(\( \text{Count}(x) - f_x^n \geq \frac{2^n}{b} \)) \( \leq \frac{1}{2^c} \)

\( \epsilon x = \frac{2x}{b} \)

\( \Rightarrow \) \( b = \frac{\epsilon x}{3} \)

\( \delta = \frac{1}{2^c} \)

\( \Rightarrow \) \( 2^{c} = \frac{1}{\delta} \)

\( \Rightarrow \) \( l = \log_2(\frac{1}{\delta}) \)

Prob \( \text{Count}(x) \) ends up in purple \( \leq \frac{\delta}{n} \)

If \( y \) added to \( H \) list, \( \Rightarrow \)

\( \text{Count}(y) \geq \frac{n}{k} \) \( \Rightarrow \) \( w_p \geq 1 - \frac{\epsilon}{k} \)

\( f_x^n \geq \frac{n}{k} - \frac{\epsilon}{n} \)
Count-Min Sketch

- Elegant small space data structure.
- Space used is independent of $n$. $O(b \cdot \log k)$
- 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 $p > \min(n, 2^{32})$.

$$H = \left\{ h_e(x) = \left[ (e \cdot x + g) \mod p \right] \mod b \right\} \text{ for } 0 \leq g \leq p-1, 1 \leq e \leq p-1$$

family of hash fn $f$ is $(p-1)$, $p$ # of fn in family.

If $h$ is chosen uniformly at random from $H$:

$$\forall x \neq y \quad \Pr(h(x) = h(y)) \leq \frac{2}{b}.$$
A different hash function should be selected uniformly at random 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 100 or higher.

Proof by:

Suppose $>50\%$ got score $\geq 100$

$$E(X) = \sum \frac{\text{score}}{\text{scored} \geq 100} \cdot \frac{\Pr(X=\text{score})}{\geq 50\%} \geq 100 \cdot \frac{1}{2} > 50$$

\[ a) \quad 100\% \]
\[ b) \quad 50\% \]
\[ c) \quad 25\% \]
\[ d) \quad \text{no bound} \]
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).

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

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

\[
\frac{1}{2}
\]

At most \(\frac{1}{4}\) can get score \(\geq 100\).
Markov’s Inequality (Intuition)

The score distribution of an exam is modeled by a random variable $X$ 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)?

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

What if you could get a negative score?

Any random variable $X$ with $E(X) = 50$

\[ 0.01 (-4900) + 0.99 \cdot 100 = 50 \]
**Markov’s Inequality**

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

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

**Bound on upper tail**

$$E(X) = \sum_{x \in \mathbb{X}} x \cdot \Pr(X = x) + \sum_{x \mid x \geq k} x \cdot \Pr(X = x)$$

$$\geq 0$$

$$\geq \sum_{x \mid x \geq k} x \cdot \Pr(X = x) \geq k \sum_{x \mid x \geq k} \Pr(X = x) = k \Pr(X \geq k)$$
**Markov’s Inequality**

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

\[ E(X) \geq k \Pr(X \geq k) = \Pr(X \geq k) \leq \frac{E(X)}{k} \]

Alternatively,

\[ P(X \geq k) \leq \frac{E[X]}{k} \Rightarrow \Pr(X \geq cE(X)) \leq \frac{E(X)}{cE(X)} = \frac{1}{c} \]

For $c > 1$,

\[ \Pr(X \geq 2E(X)) \leq \frac{1}{2} \quad \text{for } X \geq 0 \]

\[ c = 2 \]

\[ \Pr(X \geq 2E(X)) \leq \frac{1}{2} \]
**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):**

\[
X \geq 0 \quad E[X] = \int_0^\infty x f_X(x) \, dx = \int_0^k x f_X(x) \, dx + \int_k^\infty x f_X(x) \, dx
\]
\[
\geq \int_k^\infty x f_X(x) \, dx \geq \int_k^\infty k f_X(x) \, dx = k \int_k^\infty f_X(x) \, dx = k P(X \geq k)
\]

Rearranging gives

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