Heavy Hitters
Tail Bounds
(continued)

Anna Karlin
Problem

- **Input**: sequence of $n$ elements $x_1, x_2, ..., x_n$ 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**

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

\[ f_{32} = 4 \]

Goal: Find/output all elements with \( f_x \geq \frac{n}{k} \)
(These are "heavy hitters")
Output has size $O(k)$

Provably impossible to solve this problem exactly with sublinear space

Modified goal: (solve \(E+H+H\) problem)

1. If \(f_x^n \geq \frac{n}{k}\), \(x\) added to \(H+H\) list
2. If some elt, say \(y\), added to \(H+H\), then \(wp. \geq 1 - \varepsilon\)
   \(\frac{f_y^n}{n} \geq \frac{n}{k} - 3\varepsilon n\)

\[\frac{n}{k} = 0.05n\]
\[\varepsilon n = 0.01n\]
\[0.04n\]
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 3-HTH problem)
1. If \( f_x \geq \frac{n}{k} \), \( x \) added to HTH list
2. If some elt, say \( y \), added to list, then wp. \( \geq 1 - \delta \)
   \[ f_y \geq \frac{n}{k} - 3\epsilon n \]

designer specifies \( k, \epsilon, \delta \) => \( b, \ell \)

keep 2D array
1 hash tables, each of size \( b \)
Initialize tables with all Os when elt $x$ 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 HT list.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Example** $l=2$

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$y$</td>
<td>$x$</td>
<td>$z$</td>
<td>$y$</td>
</tr>
</tbody>
</table>

$\text{Count}(x) = 3$
Assumptions

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

2. Hash functions are independent of each other.

Initialize tables with all 0s

When elt \( x \) shows up,

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

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

if \( \text{Count}(x) \geq \frac{n}{k} \), add \( x \) to HT list.
Fix time $t$. \((x_1, \ldots, x_n)\) have just arrived. 

\[
\zeta_j^t \equiv t_j \left[ h_j(x) \right]
\]

\[
E(\zeta_j^t) = f_x^t + \sum_{y \neq x} f_y^t E(\omega_{xy})
\]

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

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

\[
\leq f_x^t + \frac{n}{b}
\]

\[
E(\zeta_j^t) - f_x^t \leq \frac{n}{b}
\]
\[ \Pr( \frac{Z_j^+ - f_x^+}{\sqrt{n}} > -z) = \ ? \]

\[ E(\frac{Z_j^+ - f_x^+}{\sqrt{n}}) \leq \frac{\alpha}{\beta} \]

\[ \Pr( X > 2E(X)) \leq \frac{1}{2} \]

\[ \Pr(\frac{Z_j^+ - f_x^+}{\sqrt{n}} > 2\frac{\alpha}{\beta}) \leq \frac{1}{2} \forall \alpha \]

\[ \Pr(\text{Count}(x) - f_x^+ > 2\frac{\alpha}{\beta}) \]

\[ = \Pr(\max(Z_1, \ldots, Z_k) - f_x^+ > 2\frac{\alpha}{\beta}) \]

\[ = \Pr(\frac{Z_1 - f_x^+ > 2\frac{\alpha}{\beta}}{\sqrt{n}}, \frac{Z_2 - f_x^+ > 2\frac{\alpha}{\beta}}{\sqrt{n}}, \ldots, \frac{Z_k - f_x^+ > 2\frac{\alpha}{\beta}}{\sqrt{n}}) \]

\[ = \prod_{j=1}^{k} \Pr(\frac{Z_j - f_x^+ > 2\frac{\alpha}{\beta}}{\sqrt{n}}) \leq \frac{1}{2} \text{ (each term)} \]
Modified goal: (solve $E$-H+H 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 wp. $\geq 1 - \xi$

$\xi = \frac{b}{k} - 3n$

$\Pr(\text{Count}(x) - f_x^n \geq \xi n) \leq \delta$

$\xi n = \frac{b n}{d} \Rightarrow b = \frac{\xi n}{d}$

$\delta = \frac{1}{2^k} \Rightarrow k = \log_{\frac{1}{2}}(\frac{1}{\delta})$
Count-Min Sketch

• Elegant small space data structure.

• 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).
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)?

\[ \Rightarrow E(X) > \frac{100 \cdot \frac{1}{2}}{50} \rightarrow \]

\[ \Rightarrow 50\% \]

\[ \leq \]

\[ \leq \]

\[ \leq \]

\[ a) \ 100\% \]

\[ b) \ 50\% \]

\[ c) \ 25\% \]

\[ d) \ no \ bound \]
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).

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)?

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

- a) 100%
- b) 50%
- c) 25%
- d) no bound
Markov's Inequality (Intuition)

The score distribution of an exam is modeled 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)?

\[
\frac{1}{4}
\]

What if you could get a negative score?

\[
\begin{array}{c}
X \\
E(X) = 50 \\
\geq 100 \\
0.01 - 4900 \\
0.99 \quad 100 \\
0.999999
\end{array}
\]

\[
\leq
\]

\[
a) \ 100\% \\
b) \ 50\% \\
c) \ 25\% \\
d) \ no \ bound
\]
Markov’s Inequality: Let $X \geq 0$ be a nonnegative random variable (discrete or continuous), and let $k > 0$. Then,

$$P(X \geq k) \leq \frac{E[X]}{k}$$
**Markov’s Inequality**

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

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

Alternatively,

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

$c \geq 1$
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):

\[
E[X] = \int_{0}^{\infty} xf_X(x)\,dx = \int_{0}^{k} xf_X(x)\,dx + \int_{k}^{\infty} xf_X(x)\,dx \\
\geq \int_{k}^{\infty} xf_X(x)\,dx \geq \int_{k}^{\infty} kf_X(x)\,dx = k \int_{k}^{\infty} f_X(x)\,dx = kP(X \geq k)
\]

Rearranging gives

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