Heavy Hitters
Tail Bounds

Anna Karlin

(continued)
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

Goal: Find/output all elements with \( f_x \geq \frac{n}{k} \)

(These sets are "heavy hitters")
Output has size $O(k)$

Provably impossible to solve this problem exactly with sublinear space

Modified goal: (solve $\mathcal{E}$-HTH problem)

1. If $f_x \geq \frac{n}{k}$, $x$ added to $\mathcal{H}$ list
2. If some elt, say $y$, added to list, then wp. $\geq 1-\varepsilon$
   
   $f^n_y \geq \frac{n}{k} - \varepsilon n$
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-ETH problem)

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

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

$$f_y \geq \frac{n}{K} - \varepsilon n$$

designer specifies $k, \varepsilon, \delta \Rightarrow b, \varepsilon$

keep 2D away
1 hash tables, each $y$ 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></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>h1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>h2</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>h3</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example \(l=2\)

\[\begin{array}{c}
\begin{array}{c}
x_1 \ x_2 \ x_3 \ x_4 \\
x \ y \ x \ z
\end{array}
\end{array}\]

\[\begin{array}{c}
\begin{array}{c}
h_1 \ h_2 \\
x \ y \ z
\end{array}
\end{array}\]

\[\begin{array}{c}
\begin{array}{c}
3 \ 5 \\
3 \ 2 \\
1 \ 4
\end{array}
\end{array}\]
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 \( h_1, \ldots, h_e \) are independent of each other.

Initialize tables with all 0s

When \( \text{elt } x \) shows up:

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

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

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

$$Z_j^t \triangleq t_j \left[ \bar{h}_{x_j}(x) \right]$$
**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).
Hash functions
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)?

$\leq$

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)? $\frac{1}{2}$
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)?

\[ \leq \]

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)?

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

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

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

Alternatively,

$$P(X \geq kE[X]) \leq \frac{1}{k}$$
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):

$$X \geq 0$$

$$E[X] = \int_{0}^{\infty} xf_{X}(x)dx = \int_{0}^{k} xf_{X}(x)dx + \int_{k}^{\infty} xf_{X}(x)dx$$

$$\geq \int_{k}^{\infty} xf_{X}(x)dx \geq \int_{k}^{\infty} k f_{X}(x)dx = k \int_{k}^{\infty} f_{X}(x)dx = kP(X \geq k)$$

Rearranging gives

$$P(X \geq k) \leq \frac{E[X]}{k}$$
Chebyshev’s Inequality: Let $X$ be any random variable, with mean $\mu = E[X]$ and (finite) variance. Let $\alpha > 0$.

$$P(|X - \mu| \geq \alpha) \leq \frac{Var(X)}{\alpha^2}$$
Chebyshev’s Inequality

**Chebyshev’s Inequality:** Let $X$ be any random variable, with mean $\mu = E[X]$ and (finite) variance. Let $\alpha > 0$.

$$P(|X - \mu| \geq \alpha) \leq \frac{Var(X)}{\alpha^2}$$

Alternatively, if $k > 0$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$
Chebyshev’s Inequality (picture for Gaussian)

Chebyshev’s Inequality: Let $X$ be any random variable, with mean $\mu = E[X]$ and (finite) variance. Let $k > 0$.

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$
Chebyshev’s Inequality (picture for Gaussian)

**Chebyshev’s Inequality:** Let $X$ be any random variable, with mean $\mu = E[X]$ and (finite) variance. Let $k > 0$.

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

\[ \leq \frac{1}{2^2} \]
Chebyshev’s Inequality (Proof)

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

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

Chebyshev’s Inequality: Let $X$ be any random variable, with mean $\mu = E[X]$ and (finite) variance. Let $\alpha > 0$.

$$P(|X - \mu| \geq \alpha) \leq \frac{Var(X)}{\alpha^2}$$
Markov's Inequality: Let \( X \geq 0 \) be a nonnegative rv and let \( k > 0 \). Then,

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

Chebyshev's Inequality: Let \( X \) be any random variable, with mean \( \mu = E[X] \) and (finite) variance. Let \( \alpha > 0 \).

\[
P(|X - \mu| \geq \alpha) \leq \frac{Var(X)}{\alpha^2}
\]

Proof (Chebyshev): \((X - \mu)^2\) is a nonnegative random variable.
Chebyshev’s Inequality (Proof)

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

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

Chebyshev’s Inequality: Let $X$ be any random variable, with mean $\mu = E[X]$ and (finite) variance. Let $\alpha > 0$.

$$P(|X - \mu| \geq \alpha) \leq \frac{Var(X)}{\alpha^2}$$

Proof (Chebyshev): $(X - \mu)^2$ is a nonnegative random variable.

$$P(|X - \mu| \geq \alpha) = P((X - \mu)^2 \geq \alpha^2)$$

$$\leq \frac{E[(X - \mu)^2]}{\alpha^2} \quad \text{[Markov]}$$

$$= \frac{Var(X)}{\alpha^2}$$
**The Law of Large Numbers**

**Weak Law of Large Numbers (WLLN):** Let $X_1, X_2, \ldots, X_n$ be a sequence of iid random variables with mean $\mu$. Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ be the sample mean. Then, $\bar{X}_n$ converges in probability to $\mu$. That is, for any $\epsilon > 0$,

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$
Proof of the WLLN

Weak Law of Large Numbers (WLLN): Let $X_1, X_2, \ldots, X_n$ be a sequence of iid random variables with mean $\mu$. Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ be the sample mean. Then, $\bar{X}_n$ converges in probability to $\mu$. That is, for any $\epsilon > 0$, \[
\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0
\]

Proof: Recall $E[\bar{X}_n] = \mu$ and $Var(\bar{X}_n) = \sigma^2/n$. 
Proof of the WLLN

Weak Law of Large Numbers (WLLN): Let $X_1, X_2, \ldots, X_n$ be a sequence of iid random variables with mean $\mu$. Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ be the sample mean. Then, $\bar{X}_n$ converges in probability to $\mu$. That is, for any $\epsilon > 0$,

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

Proof: Recall $E[\bar{X}_n] = \mu$ and $Var(\bar{X}_n) = \sigma^2/n$. By Chebyshev's inequality,

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{Var(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ (as } n \to \infty)$$