CountMin and Count Sketches

Lecture 10
February 14, 2019
Heavy Hitters Problem

**Heavy Hitters Problem:** Find all items $i$ such that $f_i > m/k$ for some fixed $k$.

Heavy hitters are **very** frequent items.

We saw Misra-Gries deterministic algorithm that in $O(k)$ space finds the heavy hitters assuming they exist. Two pass algorithm correctly identifies heavy hitters.
(Strict) Turnstile Model

- Turnstile model: each update is \((i_j, \Delta_j)\) where \(\Delta_j\) can be positive or negative.
- Strict turnstile: need \(x_i \geq 0\) at all time for all \(i\).

In terms of frequent items we want additive error to \(x_i\).
Basic Hashing/Sampling Idea

**Heavy Hitters Problem:** Find all items \( i \) such that \( f_i > m/k \).

- Let \( b_1, b_2, \ldots, b_k \) be the \( k \) heavy hitters
- Suppose we pick \( h : [n] \to [ck] \) for some \( c > 1 \)
- \( h \) spreads \( b_1, \ldots, b_k \) among the buckets (\( k \) balls into \( ck \) bins)
- In ideal situation each bucket can be used to count a separate heavy hitter
Part I

CountMin Sketch
CountMin Sketch

[Cormode-Muthukrishnan]

**CountMin-Sketch**\((w,d)\):

\(h_1, h_2, \ldots, h_d\) are pair-wise independent hash functions from \([n] \rightarrow [w]\).

While (stream is not empty) do

\(e_t = (i_t, \Delta_t)\) is current item

for \(\ell = 1\) to \(d\) do

\[ C[\ell, h_\ell(i_j)] \leftarrow C[\ell, h_\ell(i_j)] + \Delta_t \]

endWhile

For \(i \in [n]\) set \(\tilde{x}_i = \min_{\ell=1}^{d} C[\ell, h_\ell(i)]\).

Counter \(C[\ell, j]\) simply counts the sum of all \(x_i\) such that \(h_\ell(i) = j\).

That is,

\[ C[\ell, j] = \sum_{i : h_\ell(i) = j} x_i. \]
Intuition

- Suppose there are \( k \) heavy hitters \( b_1, b_2, \ldots, b_k \)
- Consider \( b_i \): Hash function \( h_\ell \) sends \( b_i \) to \( h_\ell(b_i) \). \( C[\ell, h(b_i)] \) counts \( x_{b_i} \) and also other items that hash to same bucket \( h(b_i) \) so we always overcount (since strict turnstile model)
- Repeating with many hash functions and taking \textit{minimum} is right thing to do: for \( b_i \) the goal is to avoid other heavy hitters colliding with it
Lemma

Let \( d = \Omega(\log \frac{1}{\delta}) \) and \( w > \frac{2}{\epsilon} \). Then for any fixed \( i \in [n] \), \( x_i \leq \tilde{x}_i \) and

\[
\Pr[\tilde{x}_i \geq x_i + \epsilon \|x\|_1] \leq \delta.
\]
Property of CountMin Sketch

**Lemma**

Let $d = \Omega(\log \frac{1}{\delta})$ and $w > \frac{2}{\epsilon}$. Then for any fixed $i \in [n]$, $x_i \leq \tilde{x}_i$ and

$$\Pr[\tilde{x}_i \geq x_i + \epsilon \|x\|_1] \leq \delta.$$ 

- Unlike Misra-Greis we have over estimates
- Actual items are not stored (requires work to recover heavy hitters)
- Works in strict turnstile model and hence can handle deletions
- Space usage is $O\left(\frac{\log(1/\delta)}{\epsilon}\right)$ counters and hence $O\left(\frac{\log(1/\delta)}{\epsilon} \log m\right)$ bits
Fix $\ell$: $h_\ell(i)$ is the bucket that $h_\ell$ hashes $i$ to.
Fix $\ell$: $h_\ell(i)$ is the bucket that $h_\ell$ hashes $i$ to.

$Z_\ell = C[\ell, h_\ell(i)]$ is the counter value that $i$ is hashed to.
Analysis

Fix $\ell$: $h_\ell(i)$ is the bucket that $h_\ell$ hashes $i$ to.

$Z_\ell = C[\ell, h_\ell(i)]$ is the counter value that $i$ is hashed to.

$E[Z_\ell] = x_i + \sum_{i', i' \neq i} \Pr[h_\ell(i') = h_\ell(i)]x_{i'}$
Fix $\ell$: $h_\ell(i)$ is the bucket that $h_\ell$ hashes $i$ to.

$Z_\ell = C[\ell, h_\ell(i)]$ is the counter value that $i$ is hashed to.

$$E[Z_\ell] = x_i + \sum_{i' \neq i} \Pr[h_\ell(i') = h_\ell(i)]x_{i'}$$

By pairwise-independence

$$E[Z_\ell] = x_i + \sum_{i' \neq i} x_{i'}/w \leq x_i + \epsilon \|x\|_1/2$$
Analysis

Fix $\ell$: $h_\ell(i)$ is the bucket that $h_\ell$ hashes $i$ to.

$Z_\ell = C[\ell, h_\ell(i)]$ is the counter value that $i$ is hashed to.

$$\mathbb{E}[Z_\ell] = x_i + \sum_{i' \neq i} \Pr[h_\ell(i') = h_\ell(i)]x_{i'}$$

By pairwise-independence

$$\mathbb{E}[Z_\ell] = x_i + \sum_{i' \neq i} x_{i'}/w \leq x_i + \epsilon \|x\|_1/2$$

Via Markov applied to $Z_\ell - x_i$ (we use strict turnstile here)

$$\Pr[Z_\ell] \geq x_i + \epsilon \|x\|_1 \leq 1/2$$
Analysis

Fix $\ell$: $h_\ell(i)$ is the bucket that $h_\ell$ hashes $i$ to.

$Z_\ell = C[\ell, h_\ell(i)]$ is the counter value that $i$ is hashed to.

$E[Z_\ell] = x_i + \sum_{i' \neq i} \Pr[h_\ell(i') = h_\ell(i)] x_{i'}$

By pairwise-independence

$E[Z_\ell] = x_i + \sum_{i' \neq i} x_{i'}/w \leq x_i + \epsilon \|x\|_1/2$

Via Markov applied to $Z_\ell - x_i$ (we use strict turnstile here)

$\Pr[Z_\ell] \geq x_i + \epsilon \|x\|_1 \leq 1/2$

Since the $d$ hash functions are independent

$\Pr[\min_\ell Z_\ell \geq x_i + \epsilon \|x\|_1] \leq 1/2^d \leq \delta$
Lemma

Let $d = \Omega(\log \frac{1}{\delta})$ and $w > \frac{2}{\epsilon}$. Then for any fixed $i \in [n]$, $x_i \leq \tilde{x}_i$ and

$$\Pr[\tilde{x}_i \geq x_i + \epsilon \|x\|_1] \leq \delta.$$ 

Choose $d = 2 \ln n$ and $w = 2/\epsilon$: we have

$$\Pr[\tilde{x}_i \geq x_i + \epsilon \|x\|_1] \leq \frac{1}{n^2}.$$ 

By union bound, with probability $(1 - 1/n)$, for all $i \in [n]$,

$$\tilde{x}_i \leq x_i + \epsilon \|x\|_1$$
Lemma

Let \( d = \Omega(\log \frac{1}{\delta}) \) and \( w > \frac{2}{\epsilon} \). Then for any fixed \( i \in [n] \), \( x_i \leq \tilde{x}_i \) and

\[
\Pr[\tilde{x}_i \geq x_i + \epsilon \|x\|_1] \leq \delta.
\]

Choose \( d = 2 \ln n \) and \( w = 2/\epsilon \): we have

\[
\Pr[\tilde{x}_i \geq x_i + \epsilon \|x\|_1] \leq \frac{1}{n^2}.
\]

By union bound, with probability \((1 - 1/n)\), for all \( i \in [n] \),

\[
\tilde{x}_i \leq x_i + \epsilon \|x\|_1
\]

Total space \( O\left(\frac{1}{\epsilon} \log n\right) \) counters and hence \( O\left(\frac{1}{\epsilon} \log n \log m\right) \) bits.
CountMin as a Linear Sketch

**Question**: Why is CountMin a linear sketch?

Recall that for $1 \leq \ell \leq d$ and $1 \leq s \leq w$:

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

Thus, once hash function $h_\ell$ is fixed:

$$C[\ell, s] = \langle u, x \rangle$$

where $u$ is a row vector in $\{0, 1\}^n$ such that $u_i = 1$ if $h_\ell(i) = s$ and $u_i = 0$ otherwise.

Thus, once hash functions are fixed, the counter values can be written as $Mx$ where $M \in \{0, 1\}^{wd \times n}$ is the sketch matrix.
**CountMin as a Linear Sketch**

**Question:** Why is CountMin a linear sketch?

Recall that for $1 \leq \ell \leq d$ and $1 \leq s \leq w$:

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

Thus, once hash function $h_\ell$ is fixed:

$$C[\ell, s] = \langle u, x \rangle$$

where $u$ is a row vector in $\{0, 1\}^n$ such that $u_i = 1$ if $h_\ell(i) = s$ and $u_i = 0$ otherwise.

Thus, once hash functions are fixed, the counter values can be written as $Mx$ where $M \in \{0, 1\}^{wd \times n}$ is the sketch matrix.
Part II

Count Sketch
Count Sketch

[Charikar-Chen-Farach-Colton]

\textbf{Count-Sketch}($w, d$):
\begin{itemize}
  \item $h_1, h_2, \ldots, h_d$ are pair-wise independent hash functions from $[n] \rightarrow [w]$.
  \item $g_1, g_2, \ldots, g_d$ are pair-wise independent hash functions from $[n] \rightarrow \{-1, 1\}$.
\end{itemize}

While (stream is not empty) do
  \begin{itemize}
    \item $e_t = (i_t, \Delta_t)$ is current item
    \item for $\ell = 1$ to $d$ do
      \begin{itemize}
        \item $C[\ell, h_\ell(i_j)] \leftarrow C[\ell, h_\ell(i_j)] + g(i_t)\Delta_t$
      \end{itemize}
  \end{itemize}
endWhile

For $i \in [n]$
  \begin{itemize}
    \item set $\tilde{x}_i = \text{median}\{g_1(i)C[1, h_1(i)], \ldots, g_\ell(i)C[\ell, h_\ell(i)]\}$.
  \end{itemize}

Like CountMin, Count sketch has $wd$ counters. Now counter values can become negative even if $x$ is positive.
Intuition

- Each hash function $h_\ell$ spreads the elements across $w$ buckets.
- The hash function $g_\ell$ induces cancellations (inspired by $F_2$ estimation algorithm).
- Since answer may be negative even if $x \geq 0$, we take the median.

**Exercise:** Show that Count sketch is also a linear sketch.
Lemma

Let \( d \geq 4 \log \frac{1}{\delta} \) and \( w > \frac{3}{\epsilon^2} \). Then for any fixed \( i \in [n] \),
\[
E[\tilde{x}_i] = x_i \quad \text{and} \quad \Pr[|\tilde{x}_i - x_i| \geq \epsilon \|x\|_2] \leq \delta.
\]
Lemma

Let $d \geq 4 \log \frac{1}{\delta}$ and $w > \frac{3}{\epsilon^2}$. Then for any fixed $i \in [n]$, $E[\tilde{x}_i] = x_i$ and

$$\Pr[|\tilde{x}_i - x_i| \geq \epsilon \|x\|_2] \leq \delta.$$ 

Comparison to CountMin

- Error guarantee is with respect to $\|x\|_2$ instead of $\|x\|_1$. For $x \geq 0$, $\|x\|_2 \leq \|x\|_1$ and in some cases $\|x\|_2 \ll \|x\|_1$.
- Space increases to $O\left(\frac{1}{\epsilon^2} \log n\right)$ counters from $O\left(\frac{1}{\epsilon} \log n\right)$ counters.
Fix an \( i \in [n] \). Let \( Z_\ell = g_\ell(i) C[\ell, h_\ell(i)] \).
Fix an $i \in [n]$. Let $Z_\ell = g_\ell(i)C[\ell, h_\ell(i)]$.

For $i' \in [n]$ let $Y_{i'}$ be the indicator random variable that is 1 if $h_\ell(i) = h_\ell(i')$; that is $i$ and $i'$ collide in $h_\ell$.

$E[Y_{i'}] = E[Y_{i'}^2] = 1/\nu$ from pairwise independence of $h_\ell$. 


Fix an $i \in [n]$. Let $Z_\ell = g_\ell(i)C[\ell, h_\ell(i)]$.

For $i' \in [n]$ let $Y_{i'}$ be the indicator random variable that is 1 if $h_\ell(i) = h_\ell(i')$; that is $i$ and $i'$ collide in $h_\ell$. $E[Y_{i'}] = E[Y_{i'}^2] = 1/w$ from pairwise independence of $h_\ell$.

$$Z_\ell = g_\ell(i)C[\ell, h_\ell(i)] = g_\ell(i) \sum_{i'} g_\ell(i')x_{i'} Y_{i'}$$
Fix an $i \in [n]$. Let $Z_\ell = g_\ell(i) C[\ell, h_\ell(i)]$.

For $i' \in [n]$ let $Y_{i'}$ be the indicator random variable that is 1 if $h_\ell(i) = h_\ell(i')$; that is $i$ and $i'$ collide in $h_\ell$. $E[Y_{i'}] = E[Y_{i'}^2] = 1/w$ from pairwise independence of $h_\ell$.

$$Z_\ell = g_\ell(i) C[\ell, h_\ell(i)] = g_\ell(i) \sum_{i'} g_\ell(i') x_{i'} Y_{i'}$$

Therefore,

$$E[Z_\ell] = x_i + \sum_{i' \neq i} E[g_\ell(i) g_\ell(i') Y_{i'}] x_{i'} = x_i,$$

because $E[g_\ell(i) g_\ell(i')] = 0$ for $i \neq i'$ from pairwise independence of $g_\ell$ and $Y_{i'}$ is independent of $g_\ell(i)$ and $g_\ell(i')$. 
\[ Z_\ell = g_\ell(i) C[\ell, h_\ell(i)]. \text{ And } E[Z_\ell] = x_i. \]
Analysis

\[ Z_\ell = g_\ell(i) C[\ell, h_\ell(i)] \]. And \( E[Z_\ell] = x_i \).

\[
\text{Var}(Z_\ell) = E[(Z_\ell - x_i)^2]
\]
\[
= E \left[ \left( \sum_{i' \neq i} g_\ell(i) g_\ell(i') Y_{i'} x_{i'} \right)^2 \right]
\]
\[
= E \left[ \sum_{i' \neq i} x_{i'}^2 Y_{i'}^2 + \sum_{i' \neq i''} x_{i'} x_{i''} g_\ell(i') g_\ell(i'') Y_{i'} Y_{i''} \right]
\]
\[
= \sum_{i' \neq i} x_{i'}^2 E[Y_{i'}^2]
\]
\[
\leq \|x\|_2^2 / w.
\]
Analysis

\[ Z_\ell = g_\ell(i) C[\ell, h_\ell(i)]. \]

We have seen: \( E[Z_\ell] = x_i \) and \( Var(Z_\ell) \leq \|x\|_2^2/w. \)
Analysis

\[ Z_\ell = g_\ell(i) C[\ell, h_\ell(i)] . \]

We have seen: \( E[Z_\ell] = x_i \) and \( \text{Var}(Z_\ell) \leq \|x\|_2^2/w. \)

Using Chebyshev:

\[
\Pr[|Z_\ell - x_i| \geq \epsilon \|x\|_2] \leq \frac{\text{Var}(Z_\ell)}{\epsilon^2 \|x\|_2^2} \leq \frac{1}{\epsilon^2 w} \leq 1/3. 
\]
Analysis

\[ Z_\ell = g_\ell(i) C[\ell, h_\ell(i)]. \]

We have seen: \( E[Z_\ell] = x_i \) and \( \text{Var}(Z_\ell) \leq \|x\|^2_2 / w. \)

Using Chebyshev:

\[
\Pr[|Z_\ell - x_i| \geq \epsilon \|x\|_2] \leq \frac{\text{Var}(Z_\ell)}{\epsilon^2 \|x\|^2_2} \leq \frac{1}{\epsilon^2 w} \leq 1/3. 
\]

Via the Chernoff bound,

\[
\Pr[|\text{median}\{Z_1, \ldots, Z_d\} - x_i| \geq \epsilon \|x\|_2] \leq e^{-cd} \leq \delta. 
\]
Lemma

Let \( d \geq 4 \log \frac{1}{\delta} \) and \( w > \frac{3}{\epsilon^2} \). Then for any fixed \( i \in [n] \), \( E[\tilde{x}_i] = x_i \) and \( \Pr[|\tilde{x}_i - x_i| \geq \epsilon \|x\|_2] \leq \delta \).

Choose \( d = \theta(\ln n) \) and \( w = \frac{3}{\epsilon^2} \): we have

\[
\Pr[|\tilde{x}_i - x_i| \geq \epsilon \|x\|_2] \leq \frac{1}{n^2}.
\]

By union bound, with probability \((1 - 1/n)\), for all \( i \in [n] \),

\[
|\tilde{x}_i - x_i| \leq \epsilon \|x\|_2
\]
Summarizing

Lemma

Let $d \geq 4 \log \frac{1}{\delta}$ and $w > \frac{3}{\epsilon^2}$. Then for any fixed $i \in [n]$, $E[\tilde{x}_i] = x_i$ and $\Pr[|\tilde{x}_i - x_i| \geq \epsilon \|x\|_2] \leq \delta$.

Choose $d = \theta(\ln n)$ and $w = 3/\epsilon^2$: we have

$$\Pr[|\tilde{x}_i - x_i| \geq \epsilon \|x\|_2] \leq 1/n^2.$$ 

By union bound, with probability $(1 - 1/n)$, for all $i \in [n]$, $|\tilde{x}_i - x_i| \leq \epsilon \|x\|_2$.

Total space $O(\frac{1}{\epsilon^2} \log n)$ counters and hence $O(\frac{1}{\epsilon^2} \log n \log m)$ bits.