CountMin and Count Sketches

Lecture 09

September 22, 2020

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.

- Identifies correct heavy hitters if they exist but can make a mistake if they don't and need second pass to verify
- Cannot handle deletions

(Strict) Turnstile Model

- Turnstile model: each update is (i_j, Δ_j) where Δ_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
- Use multiple independent hash functions to improve estimate

Part I

CountMin Sketch

CountMin Sketch: Offline view

- d independent hash functions $h_1, h_2, \ldots, h_d$. Each hash function is pair-wise independent
- Each $h_\ell : [n] \to [w]$ (hence maps to w buckets)
- Store one number per bucket and hence total of dw numbers which can be viewed as 2-day array (d rows, w columns). $C[\ell, s]$ is the counter for bucket s for hash function h_ℓ.
- Let $x \in \mathbb{R}^n$ be the given vector. For $1 \le \ell \le d$, $1 \le s \le w$

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

hence it keeps track of sum of all coordinates that h_ℓ maps to bucket s

CountMin Sketch

[Cormode-Muthukrishnan]

```
CountMin-Sketch(w, d):
    h_1, h_2, ..., h_d are pair-wise independent hash functions
        from [n] → [w].
    While (stream is not empty) do
        e_t = (i_t, Δ_t) is current item
        for ℓ = 1 to d do
            C[ℓ, h_ℓ(i_j)] ← C[ℓ, h_ℓ(i_j)] + Δ_t
    endWhile
    For i ∈ [n] set x̃_i = min_{ℓ=1}^{d} C[ℓ, h_ℓ(i)].
```

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

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

Intuition

- Suppose there are k heavy hitters $b_1, b_2, \ldots, b_k$
- Consider b_i: Hash function h_ℓ 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 *minimum* is right thing to do: for b_i the goal is to avoid other heavy hitters colliding with it

Property of CountMin Sketch

Lemma

Consider strict turnstile mode ($x \geq 0$). *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

Consider strict turnstile mode ($x \geq 0$)*. 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 $O(\frac{\log(1/\delta)}{\epsilon})$ counters and hence $O(\frac{\log(1/\delta)}{\epsilon} \log m)$ bits

Analysis

Fix ℓ and $i \in [n]$: $h_\ell(i)$ is the bucket that h_ℓ hashes i to.

Analysis

Fix ℓ and $i \in [n]$: $h_\ell(i)$ is the bucket that h_ℓ hashes i to.

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

Analysis

Fix ℓ and $i \in [n]$: $h_\ell(i)$ is the bucket that h_ℓ hashes i to.

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

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

Analysis

Fix ℓ and $i \in [n]$: $h_\ell(i)$ is the bucket that h_ℓ hashes i to.

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

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

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

Analysis

Fix ℓ and $i \in [n]$: $h_\ell(i)$ is the bucket that h_ℓ hashes i to.

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

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

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

Via Markov applied to $Z_\ell - x_i$ (we use strict turnstile here)
$\mathbf{Pr}[Z_\ell - x_i] \geq \epsilon \|x\|_1 \leq 1/2$

Analysis

Fix ℓ and $i \in [n]$: $h_\ell(i)$ is the bucket that h_ℓ hashes i to.

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

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

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

Via Markov applied to $Z_\ell - x_i$ (we use strict turnstile here)
$\Pr[Z_\ell - x_i] \geq \epsilon \|x\|_1 \leq 1/2$

Since the d hash functions are independent
$\Pr[\min_\ell Z_\ell \geq x_i + \epsilon \|x\|_1] \leq 1/2^d \leq \delta$

Summarizing

Lemma

Let $d > \log \frac{1}{\delta})$ 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 $w = 2/\epsilon$. Then

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

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

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

Summarizing

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*

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

Corollary

With $d = \Omega(\ln n)$ *and* $w = 2/\epsilon$, *with probability* $(1 - \frac{1}{n})$ *for all* $i \in [n]$:

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

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

CountMin as a Linear Sketch

Question: Why is CountMin a linear sketch?

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_ℓ 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

- Similar to CountMin use d hash functions each with w buckets etch and hence array of dw counters
- Inspired by F_2 estimation use additional $\{-1, 1\}$ hash functions which creates negative values
- Use median estimate

Count Sketch

[Charikar-Chen-FarachColton]

Count-Sketch(w, d):

```
    h_1, h_2, ..., h_d are pair-wise independent hash functions
        from [n] → [w].
    g_1, g_2, ..., g_d are pair-wise independent hash functions
        from [n] → {−1, 1}.
    While (stream is not empty) do
        e_t = (i_t, Δ_t) is current item
        for ℓ = 1 to d do
            C[ℓ, h_ℓ(i_j)] ← C[ℓ, h_ℓ(i_j)] + g(i_t)Δ_t
    endWhile
    For i ∈ [n]
        set x̃_i = median{g_1(i)C[1, h_1(i)], ..., g_ℓ(i)C[ℓ, h_ℓ(i)]}.
```

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

Intuition

- Each hash function h_ℓ spreads the elements across w buckets
- The has function g_ℓ 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.

Property of Count Sketch

Lemma

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

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

Property of Count Sketch

Lemma

Let $d \geq 4\log\frac{1}{\delta}$ *and* $w > \frac{3}{\epsilon^2}$. *Then for any fixed* $i \in [n]$, $\mathbf{E}[\tilde{x}_i] = x_i$ *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 $x \geq \mathbf{0}$, $\|x\|_2 \leq \|x\|_1$ and in some cases $\|x\|_2 \ll \|x\|_1$.
- Space increases to $O(\frac{1}{\epsilon^2}\log n)$ counters from $O(\frac{1}{\epsilon}\log n)$ counters

Analysis

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

Analysis

Fix an $i \in [n]$ and $\ell \in [d]$. 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_ℓ.
$E[Y_{i'}] = E[Y_{i'}^2] = 1/w$ from pairwise independence of h_ℓ.

Analysis

Fix an $i \in [n]$ and $\ell \in [d]$. 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_ℓ.
$E[Y_{i'}] = E[Y_{i'}^2] = 1/w$ from pairwise independence of h_ℓ.

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

Analysis

Fix an $i \in [n]$ and $\ell \in [d]$. 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_ℓ.
$E[Y_{i'}] = E[Y_{i'}^2] = 1/w$ from pairwise independence of h_ℓ.

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

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_ℓ and $Y_{i'}$ is independent of $g_\ell(i)$ and $g_\ell(i')$.

Analysis

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

Analysis

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

$$
\begin{aligned}
Var(Z_\ell) &= \mathbf{E}\left[(Z_\ell - x_i)^2\right] \\
&= \mathbf{E}\left[\Big(\sum_{i' \neq i} g_\ell(i) g_\ell(i') Y_{i'} x_{i'}\Big)^2\right] \\
&= \mathbf{E}\left[\sum_{i' \neq i} x_{i'}^2 Y_{i'}^2 + \sum_{i' \neq i''} x_{i'} x_{i''} g_\ell(i') g_\ell(i'') Y_{i'} Y_{i''}\right] \\
&= \sum_{i' \neq i} x_{i'}^2 \, \mathbf{E}\left[Y_{i'}^2\right] \\
&\leq \|x\|_2^2 / w.
\end{aligned}
$$

Analysis

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

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

Analysis

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

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

Using Chebyshev:

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

Analysis

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

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

Using Chebyshev:

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

Via the Chernoff bound,

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

Summarizing

Lemma

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

Corollary

With $d = \Theta(\ln n)$ *and* $w = 3/\epsilon^2$, *with probability* $(1 - \frac{1}{n})$ *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.