Algorithms for Big Data (Fall 25)

Lecture 7

Heavy Hitters: Misra-Gries, CountMin and CountSketch

Ali Vakilian (vakilian@vt.edu)

Frequent Items Problem (F_∞-Moment)

Recall: What is F_∞?

- F_∞ is very brittle and hard to estimate with low memory.
- Even strong lower bounds even for very weak relative approximations.

Hence, we need to settle for weaker (additive) guarantees.

Heavy Hitters Problem: Find all items i such that $f_i > \frac{m}{k}$ for some fixed k.

Heavy hitters are **very** frequent items.

Finding Majority Element (interview question)

Offline: given an array/list A of m integers, is there an element that occurs more than $m/2$ times in A?

Streaming: is there an i such that $f_i > m/2$?

Boyer-Moore Voting

Lemma. If there exists a majority item i, the algorithm outputs $s = i$ and $c \geq f_i - \frac{m}{2}$.

Why it works?

Majority (in streams):

```
let c ← 0, s ← null

foreach item e_j in the stream do:
    if e_j = s then
        c ← c + 1
    else if c = 0
        c ← 1 and s ← e_j
    else
        c ← c − 1
return c and s
```

Boyer-Moore Voting

Lemma. If there exists a majority item i, the algorithm outputs $s = i$ and $c \geq f_i - \frac{m}{2}$.

Why it works?

What if no majority item exists?

How to verify?

Majority (in streams):

let $c \leftarrow 0, s \leftarrow null$

foreach item e_j in the stream do:

  if $\boldsymbol{e_j} = s$ **then**

    $c \leftarrow c + 1$

  else if $c = 0$

    $c \leftarrow 1$ and $s \leftarrow \boldsymbol{e_j}$

  else

    $c \leftarrow c - 1$

return c and s

Extension to k Heavy Hitters

Offline: given an array/list $\boldsymbol{A}$ of m integers, is there an element that occurs more than m/k times in $\boldsymbol{A}$?

Streaming: is there an i such that $f_i > m/k$?

Idea. Extending Boyer-More Voting algorithm to this more general setting.

Misra-Gries Algorithm ($f_i \geq m/k$)

Space: $O(k)$

Theorem. For each $i \in [n]$,

$$f_i - \frac{m}{k+1} \leq \hat{f}_i \leq f_i$$

$\implies$ any item with $f_i > \frac{m}{k}$ is in D.

Misra-Gries (k):

let D be an empty array of size k
foreach item e_j in the stream **do**
  if $\boldsymbol{e_j} \in D$ **then**
    $D[\boldsymbol{e_j}] \leftarrow D[\boldsymbol{e_j}] + 1$
  else if D has less than k items
    add $\boldsymbol{e_j}$ to D and set $D[\boldsymbol{e_j}] \leftarrow 1$
  else
    foreach $\ell \in D$ **do**
      $D[\ell] \leftarrow D[\ell] - 1$ (if 0, remove)
foreach $\ell \in D$, set $\hat{f}_\ell \leftarrow D[\ell]$ (zero for rest)

Misra-Gries Algorithm ($f_i \geq m/k$): Proof

Theorem. For each $i \in [n]$, $f_i - \frac{m}{k+1} \leq \hat{f}_i \leq f_i$.

Proof.

- $\hat{f}_i \leq f_i$ is easy.

Misra-Gries Algorithm ($f_i \geq m/k$): Proof

Theorem. For each $i \in [n]$, $f_i - \frac{m}{k+1} \leq \hat{f}_i \leq f_i$.

Proof.

Alternative view of algorithm.

- maintain count $C[i]$ for each i (initialized to 0). $\leq k$ are nonzero at anytime.
- when next element e_j arrives:
  - if $C[e_j] > 0$ then **increment** $C[e_j]$
  - else if $< k$ positive counters, then **set** $C[e_j] = 1$
  - else, **decrement all** positive counters (exactly k of them)
- output $\hat{f}_i = C[i]$ for each i

Misra-Gries Algorithm ($f_i \geq m/k$): Proof

Theorem. For each $i \in [n]$, $f_i - \frac{m}{k+1} \leq \hat{f}_i \leq f_i$.

Goal. $f_i - \hat{f}_i \leq \frac{m}{k+1}$

• Suppose **decrement all** occur ℓ times, then $\ell k + \ell \leq m \Longrightarrow \ell \leq \frac{m}{k+1}$

"each **decrement all** remove k previously added items and involves an insertion causing this operation. (each deals with $k+1$ distinct elements)"

Define $\alpha = f_i - \hat{f}_i$. It is initially zero (as both are equal to zero).

How big can it get?

Misra-Gries Algorithm ($f_i \geq m/k$): Proof

Theorem. For each $i \in [n]$, $f_i - \frac{m}{k+1} \leq \hat{f}_i \leq f_i$.

Define $\alpha = f_i - \hat{f}_i$. It is initially zero (as both are equal to zero).

How big can it get?

- If $e_j = i$ and $C[i]$ is incremented, then α stays the same.
- If $e_j = i$ and $C[i]$ is not incremented, then α increases by one and k counters decremented (charge to one of the ℓ events).
- If $e_j \neq i$ and $C[i]$ is decremented, then α increases by one. This only happens in **decrement all** scenario (again charge to one of the ℓ events).

So, $\boldsymbol{\alpha \leq \ell \leq m/(k+1)}$

Wrap-up: Deterministic vs Randomized

- Cannot improve $O(k)$ space if one wants additive error of at most m/k.
- Somewhat rare to have a deterministic algorithm that is near-optimal.

Why may we still look for randomized solutions?

- Supporting deletions
- Extra properties of sketch-based solutions

Basic Hashing Idea

Heavy Hitters Problem: Find all items i such that $f_i \geq m/k$.

- Let $b_1, \ldots, b_k$ be the k heavy hitters (at most k)
- Suppose we pick a hash function $h: [n] \rightarrow [ck]$ for some $c > 1$
- h maps the heavy hitters into different buckets (k balls into ck bins)
- Then, ideally, we would like to use the count of items in each bucket as an estimate for the frequency of one heavy hitters.

Repeating this idea with independent hashes improves the estimate

CountMin Sketch [Cormode-Muthukrishnan]

- k pairwise independent hash functions $h_1, \dots, h_k$; each $[n] \to [w]$
- Equivalently, a table C with k rows and w columns.
- Store one counter per entry in the table, which keep the aggregate frequency of items mapped to the entry by the corresponding hash function. $C[\ell, s]$ is the counter for bucket s in hash function h_ℓ.
- Let $f \in \mathbb{R}^n$ be the final frequency vector. For $\ell \in [k], s \in [w]$,

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

- For every $\ell \in [k]$, $C[\ell, h_\ell(i)]$ is an over-estimate of f_i.
- We have k such estimate, how good is the quality of best of them?

CountMin Sketch in Streaming

- Each of k estimates for f_i is overcounting its frequency
- Picking the minimum such estimate is reasonable.

CountMin Sketch (stream):

let $h_1, \dots, h_k$ be pairwise independent hash functions from $[n] \to [w]$

foreach item $e_t = (i_t, \Delta_t)$ in the stream **do**:
  for $\ell = 1$ to k **do**:
    $C[\ell, h_\ell(i_t)] \leftarrow C[\ell, h_\ell(i_t)] + \Delta_t$

//frequency estimates
foreach $i \in [n]$, set $\tilde{f}_i = \min_{\ell \in [k]} C[\ell, h_\ell(i)]$

CountMin Sketch: Main Property

Theorem. Consider strict turnstile streaming (i.e., always $f \geq 0$). Let $k = \Omega\left(\log \frac{1}{\delta}\right)$ and $w > \frac{2}{\varepsilon}$. Then, for any fixed $i \in [n]$, $\boldsymbol{f_i \leq \tilde{f}_i}$, and

$$\mathbf{Pr}\left[\tilde{\boldsymbol{f}}_{\boldsymbol{i}} \geq \boldsymbol{f}_{\boldsymbol{i}} + \boldsymbol{\varepsilon}\|\boldsymbol{f}\|_{\mathbf{1}}\right] \leq \boldsymbol{\delta}$$

- Unlike Misra-Gries, CountMin overestimates.
- Items are not stored (can be recovered via queries).
- Handles deletion (works in strict turnstile model)
- Space complexity: $O(\frac{\log \frac{1}{\delta}}{\varepsilon} \cdot \log m)$ bits

CountMin: Analysis

- Consider an item i and fix a row ℓ.
- Define $Z_\ell = C[\ell, h_\ell(i)]$ the value of counter in row ℓ that i is hashed to.

$$\mathbb{E}[Z_\ell] = f_i + \sum_{j \neq i} \Pr[h_\ell(j) = h_\ell(i)] \cdot f_j$$

$$= f_i + \sum_{j \neq i} \frac{1}{w} \cdot f_j \qquad \text{// pairwise independence of } h_\ell$$

$$\leq f_i + \varepsilon \|f\|_1/2 \qquad \text{// } w > 2/\varepsilon$$

Applying Markov, $\Pr[Z_\ell - f_i \geq \varepsilon \|f\|_1] \leq 1/2$

Since k hash functions are independent,

$$\Pr[\min_{\ell \in [k]} Z_\ell \geq f_i + \varepsilon \|f\|_1] \leq \frac{1}{2^k} \leq \delta \qquad \text{// } k = \Omega\left(\log \frac{1}{\delta}\right)$$

CountMin Sketch

Space Complexity: $O(\frac{1}{\varepsilon}\log n \log m)$ bits

Theorem. Consider strict turnstile streaming (i.e., always $f \geq 0$). Let $k = \Omega\left(\log\frac{1}{\delta}\right)$ and $w > \frac{2}{\varepsilon}$. Then, for any fixed $i \in [n]$, $\boldsymbol{f_i \leq \tilde{f}_i}$, and

$$\mathbf{Pr}\left[\tilde{\boldsymbol{f}}_i \geq \boldsymbol{f}_i + \boldsymbol{\varepsilon}\|\boldsymbol{f}\|_1\right] \leq \boldsymbol{\delta}$$

- Setting $\delta = 1/n^2$, a CountMin with $O(\log n)$ rows and $O(1/\varepsilon)$ columns, for every $i \in [n]$,

$$\Pr\left[\tilde{f}_i > f_i + \varepsilon\|f\|_1\right] \leq 1/n^2$$

- By union bound over all n items, with probability $\geq 1 - 1/n$, for all $i \in [n]$

$$\tilde{f}_i \leq f_i + \varepsilon\|f\|_1$$

CountMin is a Linear Sketch

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{bmatrix} = \begin{bmatrix} f_4 \\ f_2 + f_3 \\ f_1 + f_5 \end{bmatrix}$$

Item **1** is hashed to bucket 3

Item **2** is hashed to bucket 2

Item **3** is hashed to bucket 2

Item **4** is hashed to bucket 1

Item **5** is hashed to bucket 3

hash function h_i as a Matrix-Vector Multiplication

$\Pi_{w \times m} \, \boldsymbol{f}_{m \times 1}$

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{bmatrix} = \begin{bmatrix} f_4 \\ f_2 + f_3 \\ f_1 + f_5 \\ f_1 + f_5 \\ f_2 + f_4 \\ f_3 \\ f_1 + f_2 \\ f_3 + f_4 \\ f_5 \end{bmatrix}$$

CountMin as a Matrix-Vector Multiplication

$\Pi_{(k \cdot w) \times m} \, \boldsymbol{f}_{m \times 1}$

CountSketch

- Simialr to CountMin, keeps track of a table of $k \times w$ counters
- Inspired by AMS sketch, assign u.a.r signs $\{-1, +1\}$ to items
- Counters can get even negative

CountSketch (stream):

let $h_1, \ldots, h_k$ be pairwise independent hash functions from $[n] \to [w]$

let $g_1, \ldots, g_k$ be pairwise independent hash functions from $[n] \to \{-1, +1\}$

foreach item $e_t = (i_t, \Delta_t)$ in the stream **do**:

for $\ell = 1$ to k **do**:

$$C[\ell, h_\ell(i_t)] \leftarrow C[\ell, h_\ell(i_t)] + g_\ell(i_t) \cdot \Delta_t$$

//frequency estimates

foreach $i \in [n]$, set $\tilde{f}_i = \operatorname{median}_{\ell \in [k]}\{g_\ell(i) \cdot C[\ell, h_\ell(i)]\}$

Why CountSkecth is a Linear Sketch?

$$\begin{bmatrix} 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{bmatrix} = \begin{bmatrix} -f_4 \\ -f_2 + f_3 \\ f_1 + f_5 \end{bmatrix}$$

- Item **1** is hashed to bucket **3**
- Item **2** is hashed to bucket **2** × -1
- Item **3** is hashed to bucket **2**
- Item **4** is hashed to bucket **1** × -1
- Item **5** is hashed to bucket **3**

$$\begin{bmatrix} 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix} \times \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{bmatrix} = \begin{bmatrix} -f_4 \\ -f_2 + f_3 \\ f_1 + f_5 \\ -f_1 + f_5 \\ f_2 - f_4 \\ -f_3 \\ f_1 - f_2 \\ -f_3 + f_4 \\ -f_5 \end{bmatrix}$$

hash function h_i as a Matrix-Vector Multiplication

$$\Pi_{w \times m} \, \boldsymbol{f}_{m \times 1}$$

CountMin as a Matrix-Vector Multiplication

$$\Pi_{(k \cdot w) \times m} \, \boldsymbol{f}_{m \times 1}$$

CountSketch: Main Property

Theorem. Consider strict turnstile streaming (i.e., always $f \geq 0$). Let $k = \Omega\left(\log \frac{1}{\delta}\right)$ and $w > \frac{3}{\varepsilon^2}$. Then, for any fixed $i \in [n]$, $\mathbb{E}\left[\tilde{f}_i\right] = f_i$, and

$$\Pr\left[\left|\tilde{f}_i - f_i\right| \geq \varepsilon \|f\|_2\right] \leq \delta$$

Comparison to CountMin

- Error is w.r.t. $\|f\|_2$ instead of $\|f\|_1$. Note $\|f\|_2 \leq \|f\|_1$, and in some cases $\|f\|_2 \ll \|f\|_1$
- **Space complexity:** $O(\frac{1}{\varepsilon^2} \cdot \log n)$ bits

CountSketch Analysis

- Consider an item i and fix a row ℓ.
- Define $Z_\ell = g_\ell(i) C[\ell, h_\ell(i)]$ the value of counter in row ℓ that i is hashed to.

For $j \in [n]$ let Y_j be the indicator r.v. that is 1 if $h_\ell(i) = h_\ell(j)$; i.e., i and j collide in h_ℓ

$\mathbb{E}[Y_j] = \mathbb{E}[Y_j^2] = 1/w$ from pairwise independence of h_ℓ

$Z_\ell = g_\ell(i) C[\ell, h_\ell(i)] = g_\ell(i) f_i + \sum_{j \neq i} g_\ell(i) f_j Y_j$

$\mathbb{E}[Z_\ell] = f_i + \sum_{j \neq i} \mathbb{E}[g_\ell(i) g_\ell(j) Y_j] \cdot f_j$

$= f_i$ // pairwise independence of g_ℓ

Since $\mathbb{E}[g_\ell(i) g_\ell(j) Y_j] = \mathbb{E}[g_\ell(i) g_\ell(j)]\, \mathbb{E}[Y_j] = 0$

CountSketch Analysis: Variance

- Define $Z_\ell = g_\ell(i)C[\ell, h_\ell(i)]$ the value of counter in row ℓ that i is hashed to.

For $j \in [n]$ let Y_j be the indicator r.v. that is 1 if $h_\ell(i) = h_\ell(j)$; i.e., i and j collide in h_ℓ

$\mathbb{E}[Y_j] = \mathbb{E}[Y_j^2] = 1/w$ from pairwise independence of h_ℓ

$$= \mathbb{E}[(Z_\ell - f_i)^2]$$

$$= \mathbb{E}\left[\left(\sum_{j\neq i} g_\ell(i) g_\ell(j) Y_j f_j\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{j\neq i} g_\ell(i)^2 g_\ell(j)^2 Y_j^2 f_j^2 + \sum_{j,j'\neq i} g_\ell(i)^2 g_\ell(j) g_\ell(j') Y_j Y_{j'} f_j f_{j'}\right]$$

$$= \sum_{j\neq i} f_j^2\, \mathbb{E}[Y_j^2]$$

$$\leq \|f\|_2^2/w$$

Using Chebyshev, $\Pr[|Z_\ell - f_i| \geq \varepsilon \|f\|_2] \leq \frac{\mathrm{Var}(Z\_\ell)}{\varepsilon^2 \|f\|_2^2} \leq \frac{1}{\varepsilon^2 w} \leq 1/3$

Countsketch: Concentration

Using Chebyshev, $\Pr[|Z_\ell - f_i| \geq \varepsilon \|f\|_2] \leq \frac{\mathrm{Var}(Z\_\ell)}{\varepsilon^2 \|f\|_2^2} \leq \frac{1}{\varepsilon^2 w} \leq 1/3$

Then, by Chernoff bound,

$$\Pr[|\mathrm{median}\{Z_1, \ldots, Z_k\} - f_i| \geq \varepsilon \|f\|_2] \leq e^{-\Omega(k)} \leq \delta$$