CSCE 689: Special Topics in Modern Algorithms for Data Science

Lecture 11

Samson Zhou

Presentation Schedule

- September 25: Team DAP, Team Bokun, Team Jason
- September 27: Galaxy AI, Team STMI
- September 29: Jung, Anmol, Chunkai

Last Time: Misra Gries

- Goal: Given a set S of m elements from $[n]$ and a parameter k, output the items from $[n]$ that have frequency at least $\frac{m}{k}$

- Initialize k items $V_1, \ldots, V_k$ with count $c_1, \ldots, c_k = 0$
- For updates $1, \ldots, m$:
  - If $V_t = x_i$ for some t, increment counter c_t, i.e., $c_t = c_t + 1$
  - Else if $c_t = 0$ for some t, set $V_t = x_i$
  - Else decrement all counters c_j, i.e., $c_j = c_j - 1$ for all $j \in [k]$

Last Time: Misra Gries

- Drawbacks: Misra-Gries may return false positives, i.e., items that are not frequent

- In fact, no algorithm using $o(n)$ space can output ONLY the items with frequency at least $\frac{n}{k}$

- Intuition: Hard to decide whether coordinate has frequency $\frac{n}{k}$ or $\frac{n}{k} - 1$

Last Time: (ε, k)-Frequent Items Problem

- Goal: Given a set S of m elements from $[n]$, an accuracy parameter $\varepsilon \in (0,1)$, and a parameter k, output a list that includes:
  - The items from $[n]$ that have frequency at least $\frac{m}{k}$
  - No items with frequency less than $(1-\varepsilon)\frac{m}{k}$

Last Time: Misra Gries for (ε, k)-Frequent Items Problem

- Set $r = \left\lceil \frac{k}{\varepsilon} \right\rceil$

- Initialize r items $V_1, \ldots, V_r$ with count $c_1, \ldots, c_r = 0$
- For updates $1, \ldots, m$:
  - If $V_t = x_i$ for some t, increment counter c_t, i.e., $c_t = c_t + 1$
  - Else if $c_t = 0$ for some t, set $V_t = x_i$
  - Else decrement all counters c_j, i.e., $c_j = c_j - 1$ for all $j \in [r]$
- Output coordinates V_t with $c_t \geq (1 - \varepsilon) \cdot \frac{m}{k}$

Last Time: Misra Gries for (ε, k)-Frequent Items Problem

- Claim: For all estimated frequencies $\widehat{f_i}$ by Misra-Gries, we have

$$f_i - \frac{\varepsilon m}{k} \leq \widehat{f_i} \leq f_i$$

- If $f_i \geq \frac{m}{k}$, then $\widehat{f_i} \geq f_i - \frac{\varepsilon m}{k}$ and if $f_i < (1-\varepsilon) \cdot \frac{m}{k}$, then $\widehat{f_i} < f_i - \frac{\varepsilon m}{k}$
- Returning coordinates V_t with $c_t \geq (1-\varepsilon) \cdot \frac{m}{k}$ means:
  - i with $f_i \geq \frac{m}{k}$ will be returned
  - NO i with $f_i < (1-\varepsilon) \cdot \frac{m}{k}$ will be returned

Last Time: Misra Gries for (ε, k)-Frequent Items Problem

- Summary: Misra-Gries can be used to solve the (ε, k)-frequent items problem
- Misra-Gries uses $O\left(\frac{k}{\varepsilon}\log n\right)$ bits of space
- Misra-Gries is a deterministic algorithm
- Misra-Gries *never* overestimates the true frequency

Insertion-Deletion Streams

- Stream of length $m = \Theta(n)$
- Universe of size $[n]$, underlying vector $f \in R^n$
- Each update increases or decreases a coordinate in f

f_1	f_2	f_3	f_4	f_5	f_6	f_7
0	0	0	0	0	0	0

- "Decrease f_6"

f_1	f_2	f_3	f_4	f_5	f_6	f_7
0	0	0	0	0	-1	0

Insertion-Deletion Streams

- Database Management: In database management, insertion-deletion streams are used to track changes made to the database over time

- Transaction logs often utilize this concept to record insertions and deletions to ensure data integrity and support features like rollbacks and recovery

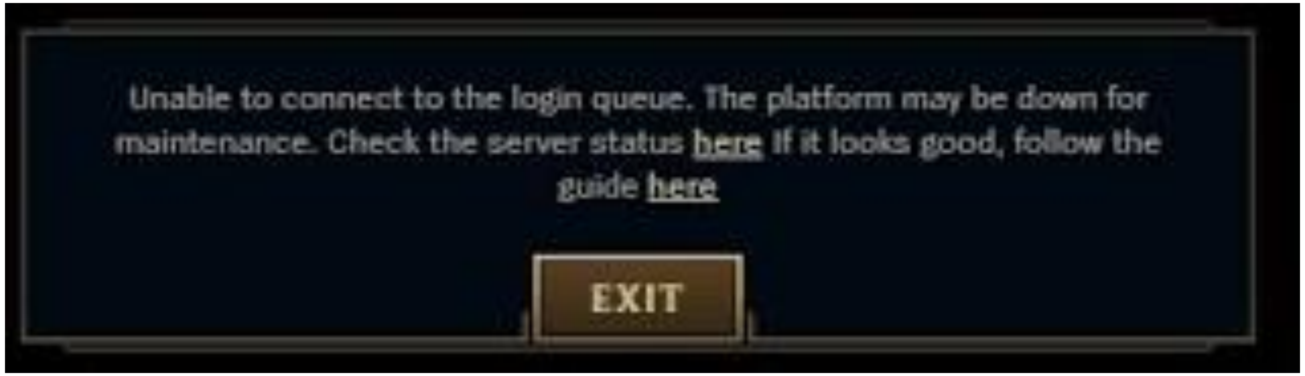
Insertion-Deletion Streams

- Version Control Systems: Insertion-deletion streams track changes made to files, enabling users to see what has been added (inserted) or removed (deleted) in each version

- Crucial for collaboration and managing software development projects, central to version control systems

Insertion-Deletion Streams

- Traffic Flow and Transportation Systems: Insertion-deletion streams are used to analyze traffic patterns and changes in transportation systems
- This helps in optimizing traffic flow, managing congestion, and improving transportation infrastructure



Frequent Items on Insertion-Deletion Streams

- Misra-Gries on Insertion-Deletion Streams
- “Increase f_1”
- “Increase f_3”
- “Increase f_2”
- “Increase f_2”
- “Decrease f_2”
- “Decrease f_2”
- “Decrease f_3”

CountMin

- Another algorithm for the (ε, k)-frequent items problem
- Can be used on insertion-deletion streams
- Can be easily parallelized across multiple servers

CountMin

- Initalization: Create b buckets of counters and use a random hash function $h: [n] \to [b]$
- Algorithm: For each update x_i, increment the counter $h(x_i)$

c_1	c_2	c_3	c_4
0	0	0	0

- At the end of the stream, output the counter $h(x_i)$ as the estimate for x_i

CountMin

f_1	f_2	f_3	f_4	f_5	f_6	f_7
0	0	0	0	0	0	0

1

c_1	c_2	c_3	c_4
0	0	0	0

CountMin

f_1	f_2	f_3	f_4	f_5	f_6	f_7
1	0	0	0	0	0	0

$h(x) = 3x + 2 \pmod 4$

1

c_1	c_2	c_3	c_4
0	0	0	0

CountMin

f_1	f_2	f_3	f_4	f_5	f_6	f_7
1	0	0	0	0	0	0

$$h(x) = 3x + 2 \pmod 4$$

1

c_1	c_2	c_3	c_4
0	0	0	0

CountMin

f_1	f_2	f_3	f_4	f_5	f_6	f_7
1	0	0	0	0	0	0

$h(x) = 3x + 2 \pmod 4$

1

c_1	c_2	c_3	c_4
1	0	0	0

CountMin

f_1	f_2	f_3	f_4	f_5	f_6	f_7
1	0	0	0	0	0	0

$$h(x) = 3x + 2 \pmod 4$$

3

c_1	c_2	c_3	c_4
1	0	0	0

CountMin

f_1	f_2	f_3	f_4	f_5	f_6	f_7
1	0	1	0	0	0	0

3

$h(x) = 3x + 2 \pmod 4$

c_1	c_2	c_3	c_4
1	0	1	0

CountMin

f_1	f_2	f_3	f_4	f_5	f_6	f_7
1	0	1	0	0	0	0

$h(x) = 3x + 2 \pmod 4$

2

c_1	c_2	c_3	c_4
1	0	1	0

CountMin

f_1	f_2	f_3	f_4	f_5	f_6	f_7
1	1	1	0	0	0	0

$$h(x) = 3x + 2 \pmod 4$$

2

c_1	c_2	c_3	c_4
1	0	1	1

CountMin

f_1	f_2	f_3	f_4	f_5	f_6	f_7
1	1	1	0	0	0	0

$h(x) = 3x + 2 \pmod 4$

1

c_1	c_2	c_3	c_4
1	0	1	1

CountMin

f_1	f_2	f_3	f_4	f_5	f_6	f_7
2	1	1	0	0	0	0

$$h(x) = 3x + 2 \pmod 4$$

1

c_1	c_2	c_3	c_4
2	0	1	1

CountMin

f_1	f_2	f_3	f_4	f_5	f_6	f_7
2	1	1	0	0	0	0

$h(x) = 3x + 2 \pmod 4$

$$5$$

c_1	c_2	c_3	c_4
2	0	1	1

CountMin

f_1	f_2	f_3	f_4	f_5	f_6	f_7
2	1	1	0	1	0	0

$$h(x) = 3x + 2 \pmod 4$$

5

c_1	c_2	c_3	c_4
3	0	1	1

CountMin

f_1	f_2	f_3	f_4	f_5	f_6	f_7
2	1	1	0	1	0	0

$$h(x) = 3x + 2 \pmod 4$$

- What is the estimation for f_4?
- What about f_3?
- What about f_5? What about f_1?

c_1	c_2	c_3	c_4
3	0	1	1

CountMin

- Given a set S of m elements from $[n]$, let $\widehat{f_i}$ be the estimated frequency for f_i

- Claim: We always have $\widehat{f_i} \geq f_i$

- Suppose $h(i) = a$ so that $c_a = \widehat{f_i}$
- Note that c_a counts the number f_j of occurrences of any j with $h(j) = a = h(i)$, including f_i itself

CountMin

- Suppose $h(i) = a$ so that $c_a = \widehat{f_i}$
- Note that c_a counts the number f_j of occurrences of any j with $h(j) = a = h(i)$, including f_i itself
- $c_a = \sum_{j:h(j)=a} f_a \geq f_i$ since $h(i) = a$

- $c_a = f_i + \sum_{j \neq i, \text{ with } j:h(j)=a} f_j$

CountMin Error Analysis

- $c_a = f_i + \sum_{j \neq i,\ \text{with}\ j:h(j)=a} f_j$
- What is the expected error for f_i?

CountMin Error Analysis

- $c_a = f_i + \sum_{j \neq i, \text{ with } j : h(j) = a} f_j$
- What is the expected error for f_i?
- $\mathrm{E}\left[\left|\sum_{j \neq i, \text{ with } j : h(j) = a} f_j\right|\right] \leq \sum_{j \neq i} \mathrm{E}\left[\left|f_j\right| \cdot I_{h(j) = h(i)}\right]$

CountMin Error Analysis

- $c_a = f_i + \sum_{j \neq i,\text{ with } j:h(j)=a} f_j$
- What is the expected error for f_i?
- $\mathrm{E}\left[\left|\sum_{j \neq i,\text{ with } j:h(j)=a} f_j\right|\right] \leq \sum_{j \neq i} \mathrm{E}\left[\left|f_j\right| \cdot I_{h(j)=h(i)}\right]$
  $= \sum_{j \neq i} \mathrm{E}\left[I_{h(j)=h(i)}\right] \cdot \left|f_j\right|$

CountMin Error Analysis

- $c_a = f_i + \sum_{j \neq i,\ \text{with } j: h(j)=a} f_j$
- What is the expected error for f_i?
- $\mathrm{E}\left[\left|\sum_{j \neq i,\ \text{with } j: h(j)=a} f_j\right|\right] \leq \sum_{j \neq i} \mathrm{E}\left[\left|f_j\right| \cdot I_{h(j)=h(i)}\right]$

$$= \sum_{j \neq i} \mathrm{E}\left[I_{h(j)=h(i)}\right] \cdot \left|f_j\right|$$

$$= \sum_{j \neq i} \Pr[h(j) = h(i)] \cdot \left|f_j\right|$$

CountMin Error Analysis

- $c_a = f_i + \sum_{j \neq i, \text{ with } j : h(j) = a} f_j$
- What is the expected error for f_i?
- $$\begin{aligned} \mathrm{E}\left[\left|\sum_{j \neq i, \text{ with } j : h(j) = a} f_j\right|\right] &\leq \Sigma_{j \neq i} \mathrm{E}\left[\left|f_j\right| \cdot I_{h(j) = h(i)}\right] \\ &= \Sigma_{j \neq i} \mathrm{E}\left[I_{h(j) = h(i)}\right] \cdot \left|f_j\right| \\ &= \Sigma_{j \neq i} \Pr[h(j) = h(i)] \cdot \left|f_j\right| \\ &= \Sigma_{j \neq i} \frac{1}{b} \cdot \left|f_j\right| = \frac{\|f\|_1}{b} \end{aligned}$$

CountMin Error Analysis

- $c_a = f_i + \sum_{j \neq i, \text{ with } j:h(j)=a} f_j$
- What is the expected error for f_i?
- $\mathrm{E}\left[\left|\sum_{j \neq i, \text{ with } j:h(j)=a} f_j\right|\right] \leq \sum_{j \neq i} \mathrm{E}\left[\left|f_j\right| \cdot I_{h(j)=h(i)}\right]$

$$= \sum_{j \neq i} \mathrm{E}\left[I_{h(j)=h(i)}\right] \cdot \left|f_j\right|$$

$$= \sum_{j \neq i} \Pr[h(j) = h(i)] \cdot \left|f_j\right|$$

$$= \sum_{j \neq i} \frac{1}{b} \cdot \left|f_j\right| = \frac{\|f\|_1}{b}$$

- Set $b = \frac{9k}{\varepsilon}$, then the expected error is at most $\frac{\varepsilon \|f\|_1}{9k}$

CountMin Error Analysis

- Set $b = \frac{9k}{\varepsilon}$, then the expected error is at most $\frac{\varepsilon \|f\|_1}{9k}$

- By Markov's inequality, the error for f_i is at most $\frac{\varepsilon \|f\|_1}{3k}$ with probability at least $\frac{2}{3}$

- How to ensure accuracy for all $i \in [n]$?

CountMin Error Analysis

- By Markov's inequality, the error for f_i is at most $\frac{\varepsilon\|f\|_1}{3k}$ with probability at least $\frac{2}{3}$

- How to ensure accuracy for all $i \in [n]$?

- Repeat $\ell := O(\log n)$ times to get estimates $e_1, \dots, e_\ell$ for each $i \in [n]$ and set $\widehat{f_i} = \min(\mathrm{e}_1, \dots, \mathrm{e}_\ell)$