CSCE 689: Special Topics in Modern Algorithms for Data Science

Lecture 10

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: The Streaming Model

- Input: Elements of an underlying data set S, which arrive sequentially
- Output: Evaluation (or approximation) of a given function
- Goal: Use space *sublinear* in the size m of the input S

1 0 1 1 1 0 0 1

Last Time: Reservoir Sampling

- Suppose we see a stream of elements from $[n]$. How do we uniformly sample one of the positions of the stream?
- [Vitter 1985]: Initialize $s = \perp$
- On the arrival of element i, replace s with x_i with probability $\frac{1}{i}$

47 72 81 10 14 33 51 29 54 9 36 46 10

Last Time: Reservoir Sampling

- Suppose we see a stream of elements from $[n]$. How do we uniformly sample one of the positions of the stream?
- [Vitter 1985]: Initialize $s = \perp$
- On the arrival of element i, replace s with x_i with probability $\frac{1}{i}$

47 72 81 10 14 33 51 29 54 9 36 46 10

Last Time: Frequent Items

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

f_1	f_2	f_3	f_4	f_5	f_6	f_7
10	0	1	1	2	0	9

- How many items can be returned? At most k coordinates with frequency at least $\frac{m}{k}$

- For $k = 20$, want items that are at least 5% of the stream

Last Time: Majority

- Goal: Given a set S of m elements from $[n]$ and a parameter $k = 2$, output the items from $[n]$ that have frequency at least $\frac{m}{2}$

- Find the item that forms the majority of the stream

Last Time: Majority

- Initialize item $V = 1$ with count $c = 0$
- For updates $1, \ldots, m$:
  - If $c = 0$, set $V = x_i$ and $c = 1$
  - Else if $V = x_i$, increment counter c by setting $c = c + 1$
  - Else if $V \neq x_i$, decrement counter c by setting $c = c - 1$

- Initialize $V = x_1$ and counter $c = 1$
- If x_1 is not majority, it must be deleted at some time T
- At time T, the stream will have consumed $\frac{T}{2}$ instances of x_1, preserving majority

Frequent Items

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

Frequent Items

- 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 item $V = 1$ with count $c = 0$
- For updates $1, \dots, m$:
  - If $c = 0$, set $V = x_i$
  - Else if $V = x_i$, increment counter c by setting $c = c + 1$
  - Else if $V \neq x_i$, decrement counter c by setting $c = c - 1$

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 $j \in [k]$

Misra Gries

- $n = 7, k = 3$
- $V_1 = \perp, c_1 = 0$
- $V_2 = \perp, c_2 = 0$
- $V_3 = \perp, c_3 = 0$

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

Misra Gries

- $V_1 = \perp, c_1 = 0$
- $V_2 = \perp, c_2 = 0$
- $V_3 = \perp, c_3 = 0$

3

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

Misra Gries

- $V_1 = 3, c_1 = 1$
- $V_2 = \perp, c_2 = 0$
- $V_3 = \perp, c_3 = 0$

3

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

Misra Gries

- $V_1 = 3, c_1 = 1$
- $V_2 = \perp, c_2 = 0$
- $V_3 = \perp, c_3 = 0$

$$1$$

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

Misra Gries

- $V_1 = 3, c_1 = 1$
- $V_2 = 1, c_2 = 1$
- $V_3 = \perp, c_3 = 0$

$$1$$

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

Misra Gries

- $V_1 = 3, c_1 = 1$
- $V_2 = 1, c_2 = 1$
- $V_3 = \perp, c_3 = 0$

2

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

Misra Gries

- $V_1 = 3, c_1 = 1$
- $V_2 = 1, c_2 = 1$
- $V_3 = 2, c_3 = 1$

$$2$$

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

Misra Gries

- $V_1 = 3, c_1 = 1$
- $V_2 = 1, c_2 = 1$
- $V_3 = 2, c_3 = 1$

$$1$$

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

Misra Gries

- $V_1 = 3, c_1 = 1$
- $V_2 = 1, c_2 = 2$
- $V_3 = 2, c_3 = 1$

1

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

Misra Gries

- $V_1 = 3, c_1 = 1$
- $V_2 = 1, c_2 = 2$
- $V_3 = 2, c_3 = 1$

$$4$$

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

Misra Gries

- $V_1 = 3, c_1 = 0$
- $V_2 = 1, c_2 = 1$
- $V_3 = 2, c_3 = 0$

$$4$$

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

Misra Gries

- $V_1 = 3, c_1 = 0$
- $V_2 = 1, c_2 = 1$
- $V_3 = 2, c_3 = 0$

$$2$$

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

Misra Gries

- $V_1 = 3, c_1 = 0$
- $V_2 = 1, c_2 = 1$
- $V_3 = 2, c_3 = 1$

2

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

Misra Gries

- $V_1 = 3, c_1 = 0$
- $V_2 = 1, c_2 = 1$
- $V_3 = 2, c_3 = 1$

1

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

Misra Gries

- $V_1 = 3, c_1 = 0$
- $V_2 = 1, c_2 = 2$
- $V_3 = 2, c_3 = 1$

1

f_1	f_2	f_3	f_4	f_5	f_6	f_7
3	2	1	1	0	0	0

Misra Gries

- $V_1 = 3, c_1 = 0$
- $V_2 = 1, c_2 = 2$
- $V_3 = 2, c_3 = 1$

5

f_1	f_2	f_3	f_4	f_5	f_6	f_7
3	2	1	1	0	0	0

Misra Gries

- $V_1 = 5, c_1 = 1$
- $V_2 = 1, c_2 = 2$
- $V_3 = 2, c_3 = 1$

5

f_1	f_2	f_3	f_4	f_5	f_6	f_7
3	2	1	1	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 1$
- $V_2 = 1, c_2 = 2$
- $V_3 = 2, c_3 = 1$

1

f_1	f_2	f_3	f_4	f_5	f_6	f_7
3	2	1	1	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 1$
- $V_2 = 1, c_2 = 3$
- $V_3 = 2, c_3 = 1$

$$1$$

f_1	f_2	f_3	f_4	f_5	f_6	f_7
4	2	1	1	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 1$
- $V_2 = 1, c_2 = 3$
- $V_3 = 2, c_3 = 1$

4

f_1	f_2	f_3	f_4	f_5	f_6	f_7
4	2	1	1	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 0$
- $V_2 = 1, c_2 = 2$
- $V_3 = 2, c_3 = 0$

4

f_1	f_2	f_3	f_4	f_5	f_6	f_7
4	2	1	2	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 0$
- $V_2 = 1, c_2 = 2$
- $V_3 = 2, c_3 = 0$

3

f_1	f_2	f_3	f_4	f_5	f_6	f_7
4	2	1	2	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 0$
- $V_2 = 1, c_2 = 2$
- $V_3 = 3, c_3 = 0$

3

f_1	f_2	f_3	f_4	f_5	f_6	f_7
4	2	2	2	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 0$
- $V_2 = 1, c_2 = 2$
- $V_3 = 3, c_3 = 0$

$$1$$

f_1	f_2	f_3	f_4	f_5	f_6	f_7
4	2	2	2	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 0$
- $V_2 = 1, c_2 = 3$
- $V_3 = 3, c_3 = 0$

1

f_1	f_2	f_3	f_4	f_5	f_6	f_7
5	2	2	2	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 0$
- $V_2 = 1, c_2 = 3$
- $V_3 = 3, c_3 = 0$

3

f_1	f_2	f_3	f_4	f_5	f_6	f_7
5	2	2	2	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 0$
- $V_2 = 1, c_2 = 3$
- $V_3 = 3, c_3 = 1$

3

f_1	f_2	f_3	f_4	f_5	f_6	f_7
5	2	3	2	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 0$
- $V_2 = 1, c_2 = 3$
- $V_3 = 3, c_3 = 1$

$$1$$

f_1	f_2	f_3	f_4	f_5	f_6	f_7
5	2	3	2	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 0$
- $V_2 = 1, c_2 = 4$
- $V_3 = 3, c_3 = 1$

1

f_1	f_2	f_3	f_4	f_5	f_6	f_7
6	2	3	2	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 0$
- $V_2 = 1, c_2 = 4$
- $V_3 = 3, c_3 = 1$

3

f_1	f_2	f_3	f_4	f_5	f_6	f_7
6	2	3	2	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 0$
- $V_2 = 1, c_2 = 4$
- $V_3 = 3, c_3 = 2$

3

f_1	f_2	f_3	f_4	f_5	f_6	f_7
6	2	4	2	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 0$
- $V_2 = 1, c_2 = 4$
- $V_3 = 3, c_3 = 2$

3

f_1	f_2	f_3	f_4	f_5	f_6	f_7
6	2	4	2	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 0$
- $V_2 = 1, c_2 = 4$
- $V_3 = 3, c_3 = 3$

3

f_1	f_2	f_3	f_4	f_5	f_6	f_7
6	2	5	2	1	0	0

Misra Gries

- $V_1 = 5, c_1 = 0$
- $V_2 = 1, c_2 = 4$
- $V_3 = 3, c_3 = 3$

6

f_1	f_2	f_3	f_4	f_5	f_6	f_7
6	2	5	2	1	0	0

Misra Gries

- $V_1 = 6, c_1 = 1$
- $V_2 = 1, c_2 = 4$
- $V_3 = 3, c_3 = 3$

6

f_1	f_2	f_3	f_4	f_5	f_6	f_7
6	2	5	2	1	1	0

Misra Gries

- $V_1 = 6, c_1 = 1$
- $V_2 = 1, c_2 = 4$
- $V_3 = 3, c_3 = 3$

6

- Report 1, 3, and 6 as frequent items

f_1	f_2	f_3	f_4	f_5	f_6	f_7
6	2	5	2	1	1	0

Misra Gries

- Claim: At the end of the stream of length m, we report all items with frequency at least $\frac{m}{k}$

- Intuition: If there are k coordinates with frequency $\frac{m}{k}$, they will all be tracked and reported, since we have k counters
- If there are $\frac{k}{2}$ coordinates with frequency at least $\frac{m}{k}$, we still have $\frac{k}{2}$ counters for the remaining $\frac{m}{2}$, updates
- Will have at most $\frac{m}{k}$ decrement operations, which is small enough so that frequent items are still stored

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 $\frac{n}{k} - 1$

Misra Gries

- Intuition: Hard to decide whether coordinate has frequency $\frac{n}{k}$ or $\frac{n}{k} - 1$
- $x_1 = 2, x_2 = 5, x_3 = 4, x_4 = 7, x_5 = 1, x_6 = 9, \ldots$
- $\underbrace{x_{n-\frac{n}{k}+1} = \alpha, x_{n-\frac{n}{k}+2} = \alpha, \ldots, x_n = \alpha}_{\frac{n}{k} - 1 \text{ times}}$

(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 $(1-\varepsilon)\frac{m}{k}$

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

- 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 $j \in [k]$

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 $j \in [r]$

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

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

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

- Intuition: Have a lot of counters, so relatively few decrements

(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 $(1-\varepsilon)\frac{m}{k}$

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 $c_t \geq (1 - \varepsilon) \cdot \frac{m}{k}$

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

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