Streaming Algorithms

- Streaming Model of Computation
- Streaming Algorithms and DFA
- Stream: Motivation and Applications
- Synopsis: Sliding Window, Histogram, Wavelets
- Sampling from Stream: Reservoir Sampling
- Linear Sketch
- Count-Min Sketch
- AMS Sketch

Imdad ullah Khan

Synopsis: Linear Sketch

Synopsis: Linear Sketch

- Sample is a general purpose synopsis
- Process sample only – no advantage from observing the whole stream
- Sketches are specific to a particular purpose (query)
- Sketches (also histograms and wavelets) take advantage from the fact the processor see the whole stream (though can't remember all)

Linear Sketch

A linear sketch interprets a stream as defining the frequency vector

IP	Frequency
160.39.142.2	3
18.9.22.69	2
80.97.56.20	2

figure credit: Andoni@Columbia

$\mathcal{S}: \; a_1, a_2, a_3, a_4, \ldots, a_m$

$a_i \in [n]$

$$\mathbf{F}: \; \begin{array}{|c|c|c|c|c|} 1 & 2 & 3 & & n \\ \hline f_1 & f_2 & f_3 & \cdots \quad \cdots & f_n \\ \hline \end{array}$$

$f_j = |\{a_i \in \mathcal{S} : a_i = j\}|$ (frequency of j in $\mathcal{S}$)

$\mathcal{S}: \; 2, 5, 6, 7, 8, 2, 1, 2, 7, 5, 5, 4, 2, 8, 8, 9, 5, 6, 4, 4, 2, 5, 5$

$\mathbf{F}:$

1	2	3	4	5	6	7	8	9
1	5	0	3	6	2	2	3	1

Linear Sketch: Frequency Moments

Often we are interested in frequency moments of a stream

$\mathcal{S}: \; a_1, a_2, a_3, a_4, \ldots, a_m$

$a_i \in [n]$

$\mathbf{F}:$

1	2	3		n
f_1	f_2	f_3	$\ldots \quad \ldots$	f_n

$f_j = |\{a_i \in \mathcal{S} : a_i = j\}|$ (frequency of j in $\mathcal{S}$)

$\mathcal{S}: \; 2, 5, 6, 7, 8, 2, 1, 2, 7, 5, 5, 4, 2, 8, 8, 9, 5, 6, 4, 4, 2, 5, 5$

$\mathbf{F}:$

1	2	3	4	5	6	7	8	9
1	5	0	3	6	2	2	3	1

$F_0 := \sum_{i=1}^{n} f_i^0$ ▷ number of distinct elements

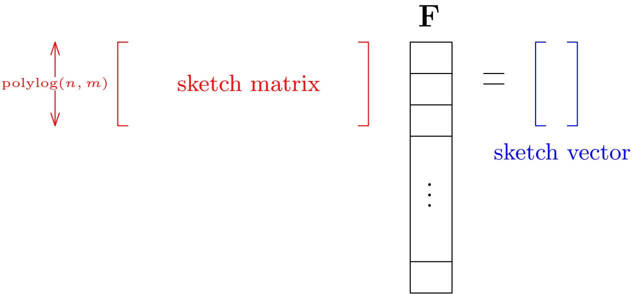
$F_1 := \sum_{i=1}^{n} f_i$ ▷ length of stream, m

$F_2 := \sum_{i=1}^{n} f_i^2$ ▷ second frequency moment

Synopsis: Linear Sketch

Linear sketch is a synopsis that can be computed as a linear transform of **F**

- Best suited for data streams, highly parallelizable
- Very good for our problems of computing norms of **F**
- Can be readily extended to variations of the basic stream model

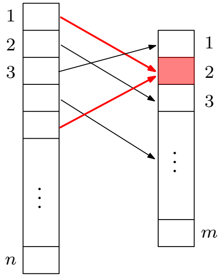


Linear Sketch: Hash Functions

Hash function is an efficient way to implement the Dictionary ADT

Hash functions map keys $A \subset U$ to m buckets labeled $\{0, 1, 2, \ldots, m-1\}$

$\triangleright$ A is not known in advance and $|A| = n$



Desired properties of hash functions

- Fewer collisions
- Small range (m)
- Small space complexity to store hash function
- Easy to evaluate hash value for any key

Linear Sketch: Universal Hash Functions

Universal hash functions have probabilistic guarantees on collision

A family $\mathcal{H}$ of hash functions of the form $h : U \mapsto [m]$ is **2-universal** if

for any distinct keys $x, y \in U$, $$\Pr_{h \in_R \mathcal{H}} \left[h(x) = h(y) \right] \leq \frac{1}{m}$$

▷ Source of randomness is picking h (at random) from the family

Linear Congruential Generators for $U = \mathbb{Z}$

- Pick a prime number $p > m$
- For any two integers a and b $(1 \leq a \leq p-1)$, $(0 \leq b \leq p-1)$
- A hash function $h_{a,b} : U \mapsto [m]$ is defined as

$$h_{a,b}(x) = (ax + b) \pmod{p} \pmod{m}$$

$\mathcal{H} := \{ h_{a,b} \ : \ 1 \leq a \leq p-1 \ , \ 0 \leq b \leq p-1 \}$ is 2-universal

Picking a random $h \in \mathcal{H}$ amounts to picking random a and b

Linear Sketch: Universal Hash Functions

Linear Congruential hash functions are 2-universal

A family $\mathcal{H}$ of hash functions of the form $h : U \mapsto [m]$ is 2-**universal** if

for any distinct keys $x, y \in U$, $$\Pr_{h \in_R \mathcal{H}} \big[h(x) = h(y)\big] \leq \frac{1}{m}$$

▷ Source of randomness is picking h (at random) from the family

Linear Congruential hash function of the form $h : \mathbb{Z} \mapsto [m]$

- Pick a prime number $p > m$
- For any two integers a and b $(1 \leq a \leq p-1)$, $(0 \leq b \leq p-1)$
- A hash function $h_{a,b} : U \mapsto [m]$ is defined as

$$h_{a,b}(x) = (ax + b) \pmod{p} \pmod{m}$$

$\mathcal{H} := \{h_{a,b} \; : \; 1 \leq a \leq p-1 \, , \; 0 \leq b \leq p-1\}$ is 2-universal

Picking a random $h \in \mathcal{H}$ amounts to picking random a and b