Theory of Computation Complexity

Tutorial 3

LIU yang

Outline

- Data streams
- Counting Distinct Elements
- Sketching

Data Streams

- A data stream is a massive sequence of data.
  - Too large to be stored.
- Examples:
  - World Wide Web
  - Internet traffic logs
  - Financial transactions
  - Database transactions
- Technique:
  - Stream algorithms
  - Sketching
  - …

Data Stream Model

- Items in the stream are presented sequentially
- Single pass over the data: $i_1, i_2, \ldots, i_n$
  - n is known.
- Bounded storage
  - Typically sublinear, n^{α} or log n.
- Processing time per element must be fast
  - Allow randomness.

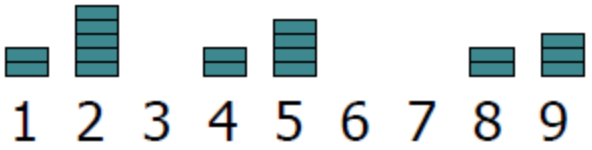
E.g. Counting Distinct Elements

- Stream elements: numbers from {1,…,m}
- Goal: Count the number of distinct elements(DE) in the stream
  - Given a T > 0, output YES, if DE>(1+ϵ)T; No if DE < (1- ϵ)T, with high probability.

Vector Interpretation

Stream: 8 2 1 9 1 9 2 4 4 9 4 2 5 4 2 5 8 5 2 5

Vector X:



- Initially, $x=0$
- Insertion of i is interpreted as

$$x_i = x_i + 1$$

- Want to estimate $DE(x)$

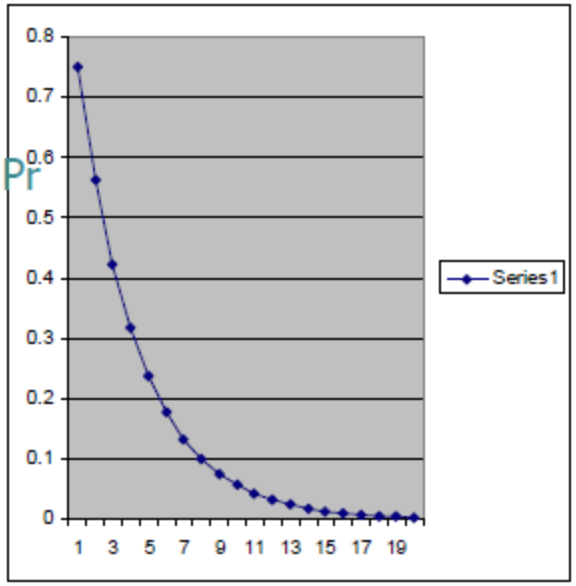
**This slide is from lecture notes of Prof. Piotr Indyk for the course Sketching, Streaming and Sub-linear Space Algorithms*

Estimating DE(x)

Vector X:

Set S: + + + (T=4)

- Choose a random set S of coordinates
  - For each i, we have $Pr[i \in S]=1/T$
- Maintain $Sum_S(x) = \Sigma_{i \in S} x_i$
- Estimation algorithm A:
  - YES, if $Sum_S(x)>0$
  - NO, if $Sum_S(x)=0$
- Analysis:
  - $Pr=Pr[Sum_S(x)=0] = (1-1/T)^{DE}$
  - For T large enough: $(1-1/T)^{DE} \approx e^{-DE/T}$
  - Using calculus, for ε small enough:
    - If $DE> (1+\varepsilon)T$, then $Pr \approx e^{-(1+\varepsilon)} < 1/e - \varepsilon/3$
    - if $DE< (1-\varepsilon)T$, then $Pr \approx e^{-(1-\varepsilon)} > 1/e + \varepsilon/3$



**This slide is from lecture notes of Prof. Piotr Indyk for the course Sketching, Streaming and Sub-linear Space Algorithms*

Sketching

- Just like public-coin SMP model
  - Alice and Bob send messages to Referee
  - Referee announces the output
- Streams algorithms can be used for sketching
  - E.g., counting distinct elements
- A simple glimpse
  - Equality test in public-con SMP

Summery

- Data Streams
- Counting Distinct Elements
- Sketching

Thanks