Algorithms for Data Streams

MIT

Piotr Indyk

Plan for this week

Time	Sunday August	Monday August 20th	Tuesday August 21th	Wednesday August 22th	Thursday August 23th	Friday August
08:30		Registration				
09:00		Lecture: Piotr Indyk Introduction. Norm/Count estimation	Lecture: T.S. Jayram Intro to lower bounds, communication complexity	Lecture: T.S. Jayram Lower bounds	Lecture: Ravi Kumar Lower bounds	
10:30		Coffe break	Coffe break	Coffe break	Coffe break	
11:00		Lecture: Sudipto Guha Metric data. Clustering. Graph data.	Lecture: Ravi Kumar Lower bounds	Lecture: Ravi Kumar Recitations.	Lecture: Martin Strauss Heavy hitters. Wavelets and histograms.	MADALGO Inauguration
12:30		Lunch break	Lunch break	Lunch break	Lunch break	
14:00		Lecture: Piotr Indyk Geometric data. Clustering. MST.	Lecture: Sudipto Guha Random order streams	Excursion	Lecture: Martin Strauss Compressed sensing. LP-based decoding.	
15:00		Coffee break	Coffee break		Coffee break	
15:30		Lecture Recitations	Lecture Recitations		Lecture Recitations/Conclusions	
16:30		Student poster session				
17:00	Registration and welcome reception					
18:00			Summer School dinner			
20:00						

Data Streams

- A data stream is a (massive) sequence of data
  - Too large to store (on disk, memory, cache, etc.)
- Examples:
  - Network traffic (source/destination)
  - Sensor networks
  - Satellite data feed, etc.
- Approaches:
  - Ignore it
  - Develop algorithms for dealing with such data

Plan For This Lecture

- Introduce the data stream model(s)
- Basic algorithms
  - Estimating number of distinct elements in a stream
  - Frequency moments and norms

Basic Data Stream Model

- Single pass over the data: $i_1, i_2, \ldots, i_n$
  - Typically, we assume n is known
- Bounded storage (typically n^α or $\log^c n$)
  - Units of storage: bits, words or „elements" (e.g., points, nodes/edges)
- Fast processing time per element
  - Randomness OK (in fact, almost always necessary)

8 2 1 9 1 9 2 4 6 3 9 4 2 3 4 2 3 8 5 2 5 6 ...

Example: Counting Distinct Elements

- Stream elements: numbers from $\{1 \ldots m\}$
- Goal: estimate the number of distinct elements DE in the stream
  - Up to $1 \pm \varepsilon$
  - With probability $1-P$
- Simpler goal: for a given $T>0$, provide an algorithm which, with probability $1-P$:
  - Answers YES, if $DE > (1+\varepsilon)T$
  - Answers NO, if $DE < (1-\varepsilon)T$
- Run, in parallel, the algorithm with

  $T=1,\ 1+\varepsilon,\ (1+\varepsilon)^2, \ldots, n$
  - Total space multiplied by $\log_{1+\varepsilon} n \approx \log(n)/\varepsilon$

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: 
1 2 3 4 5 6 7 8 9

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

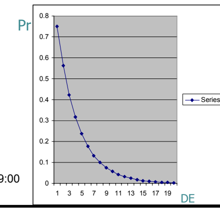
  $x_i = x_i + 1$
- Want to estimate DE(x)

Estimating DE(x)

Vector X: 
1 2 3 4 5 6 7 8 9

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:
  - $Pr=Pr[Sum_S(x)=0] = (1-1/T)^{DE}$
  - Using calculus (for T, $1/\varepsilon$ large enough):
    - If $DE> (1+\varepsilon)T$, then $Pr<1/e - \varepsilon/3$
    - if $DE< (1-\varepsilon)T$, then $Pr>1/e + \varepsilon/3$



Estimating DE(x) ctd.

- We have Algorithm A:
  - If $DE> (1+\varepsilon)T$, then $Pr<1/e-\varepsilon/3$
  - if $DE< (1-\varepsilon)T$, then $Pr>1/e+\varepsilon/3$
- Algorithm B:
  - Select sets $S_1 \ldots S_k$, $k=O(\log(1/P)/\varepsilon^2)$
  - Let Z = number of $Sum_{Sj}(x)$ that are equal to 0
  - By Chernoff bound (define), with probability $>1-P$
    - If $DE> (1+\varepsilon)T$, then $Z<k/e$
    - if $DE< (1-\varepsilon)T$, then $Z>k/e$
- Total space: $O(\log(n)/\varepsilon \;\; \log (1/P)/\varepsilon^2)$ numbers in range $0 \ldots n$
- Homework: remove the $1/\varepsilon$ factor [Flajolet-Martin'85]

Interlude – Chernoff bound

- Let $Z_1 \ldots Z_k$ be i.i.d. Bernoulli variables, with $Pr[Z_j=1]=p$
- Let $Z=\Sigma_j Z_j$
- For any $1>\varepsilon>0$, we have

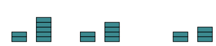
  $Pr[\ |E[Z]-Z| > \varepsilon E[Z]\ ] \leq 2\exp(\ -\varepsilon^2 E[Z]/3\ )$

Comments

- Implementing S:
  - Choose a hash function $h: \{1..m\} \rightarrow \{1..T\}$
  - Define $S=\{i:\ h(i)=1\}$
- Implementing h
  - Pseudorandom generators
- Better algorithms known:
  - Theory: $O(\ \log(1/\varepsilon)/\varepsilon^2 + \log n)$ bits

    [Bar-Yossef-Jayram-Kumar-Sivakumar-Trevisan'02]
  - Practice: need 128 bytes for all works of Shakespeare , $\varepsilon \approx 10\%$ [Durand-Flajolet'03]

More comments

Vector X: 
1 2 3 4 5 6 7 8 9

- The algorithm uses "linear sketches"

  $Sum_{Sj}(x)=\Sigma_{i \in Sj}\, x_i$
- Can implement decrements $x_i=x_i-1$
  - I.e., the stream can contain deletions of elements (as long as $x \geq 0$)
  - Other names: dynamic model, turnstile model

More General Problem

- What other functions of a vector x can we maintain in small space ?
- L_p norms:

$$||x||_p = (\sum_i |x_i|^p)^{1/p}$$

  - We also have $||x||_\infty = \max_i |x_i|$
  - … and $||x||_0 = DE(x)$, since $||x||_p^p = \sum_i |x_i|^p \to DE(x)$ as $p \to 0$

- How much space do you need to estimate $||x||_p$ (for const. ε) ?
- Theorem:
  - For $p \in [0,2]$: polylog n space suffices
  - For $p>2$: $n^{1-2/p}$ polylog n space suffices and is necessary

[Alon-Matias-Szegedy'96, Feigenbaum-Kannan-Strauss-Viswanathan'99, Indyk'00, Coppersmith-Kumar'04, Ganguly'04, Bar-Yossef-Jayram-Kumar-Sivakumar'02'03, Saks-Sun'03, Indyk-Woodruff'05]

Interlude: Normal Distribution

- Normal distribution:
  - Range: $(-\infty, \infty)$
  - Density: $f(x)=e^{-x^2/2} / (2\pi)^{1/2}$
  - Mean=0, Variance=1
- Basic facts:
  - If X and Y independent r.v. with normal distribution, then $X+Y$ has normal distribution
  - $Var(cX)=c^2 Var(X)$
  - If X,Y independent, then $Var(X+Y)=Var(X)+Var(Y)$

Sketching

- We will use a linear sketch Ax, where A is a $k*n$ matrix
  - Increments/decrements come "for free"
  - $k=C\log(1/P)/\varepsilon^2$
- Each entry of A has normal distribution
- A^i is a row of A
- What can we say about Ax ?
- Consider $(Ax)_i=A^i *x = a*x=\sum_i a_i x_i$
- Each term $a_i x_i$
  - Has normal distribution
  - With variance x_i^2
- Thus, $(Ax)_i$ has normal distribution, with variance $\sum_i x_i^2=||x||_2^2$

Estimation - intuition

- From previous slide: $(Ax)_i$ has normal distribution, with variance $\sum_i x_i^2=||x||_2^2$
- Consider a random variable

$$Z=\text{median}[\ |(Ax)_1|, \dots , |(Ax)_k|]$$

- Intuitively, for large enough k, Z should be "close" to the median* of $||x||_2 |a|$, where a has normal distribution
- Then we could use an estimator

$$E=Z/\text{median}\ (a)$$

*M is the median of a random var a if $\Pr[a>M]=1/2$

Formally

- Lemma 1: Let $Z_1 \dots Z_k$ be i.i.d. real random variables chosen from any distribution having continuous c.d.f. F and median M
  - I.e., $F(t)=\Pr[Z_i <t]$ and $F(M)=1/2$

Define $Z=\text{median}[Z_1,\dots,Z_k]$. Then, for some absolute const. C

$$\Pr[F(Z) \in (1/2-\varepsilon,1/2-\varepsilon)] \geq 1-e^{-C\varepsilon^2 k}$$

- Proof:
  - Consider events E_i: $F(Z_i)<1/2-\varepsilon$
  - We have $p=\Pr[E_i]=1/2-\varepsilon$
  - By Chernoff bound, the probability that at least $k/2$ of these events hold is at most $e^{-C\varepsilon^2 k}$
  - If less than $k/2$ of these events hold, then $F(Z)\geq 1/2-\varepsilon$
  - Therefore, $\Pr[F(Z)< 1/2-\varepsilon]$ is at most $e^{-C\varepsilon^2 k}$
  - The other case can be dealt with in an analogous manner

Formally, ctd.

- Lemma 2: Let F be c.d.f of a random variable $||x||_2|a|$, a normal.

  If for some z we have

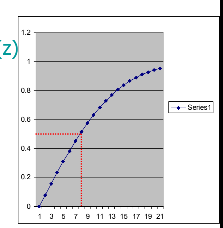
$$F(z) \in (1/2-\varepsilon,1/2-\varepsilon)$$

  then, for some abs. const. C'

$$z = ||x||_2\ [\text{median}(a) \pm C' \varepsilon]$$

- Proof: Use calculus.

F(z)



10z

Johnson-Lindenstrauss Lemma

- We used an estimator

  $Z=\text{median}[\ |(Ax)_1|, \ldots , |(Ax)_k|]/\text{Scale}$
- Instead, we could have used

  $Z=[\ |(Ax)_1|^2 + \ldots +|(Ax)_k|^2\ ]^{1/2} /\text{Scale}$
- Johnson-Lindenstrauss: the latter estimator works
- Proof similar to the proof of the Chernoff bound

Recap

- Total space: $O(\log (1/P)/\varepsilon^2)$ real numbers
  - Not including the random bits
- Can discretize the numbers so that they have $O(\log n)$ bits of precision
- In fact, a very similar algorithm works if the entries of A are Bernoulli random variables [Alon-Matias-Szegedy'96]

Other norms

- Key property of normal distribution: if X, Y, Z independent, then

  $aX+bY$ is distributed as $(a^2+b^2)^{1/2}Z$
- This is possible to achieve for 2 replaced by any $p \in (0,2]$ using "p-stable distributions"
- The median estimator and the proofs go through, albeit the constant C' (previous slide) depends on p in an unclear way
- Geometric mean estimator [Li'06] gives an explicit dependence on p

Summary

- Streaming model
  - Insertions-only vs. insertions+deletions
- Maintaining L_p norm under updates
  - Polylogarithmic space for $p \leq 2$