Lower Bounds for Data Streams: A Survey

David Woodruff
IBM Almaden

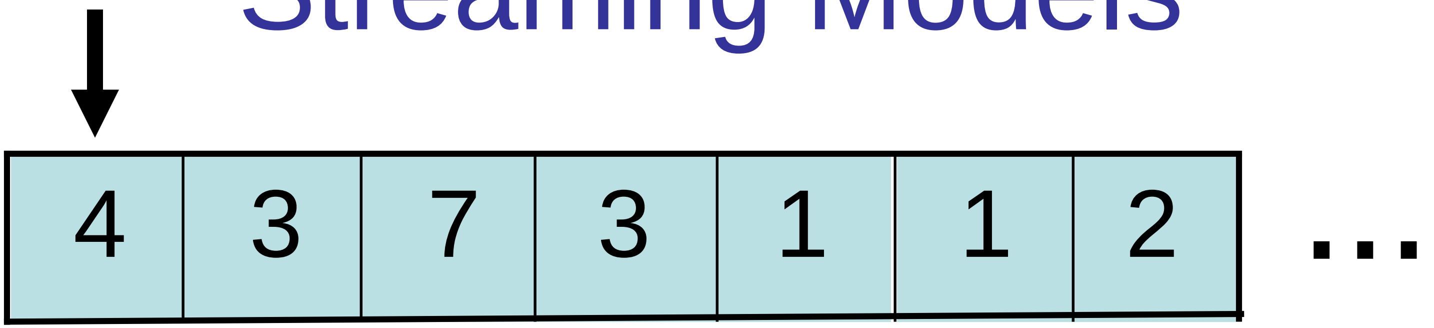
Outline

1. Streaming model and examples

2. Background on communication complexity for streaming
    1. Product distributions
    2. Non-product distributions

3. Open problems

Streaming Models



- Long sequence of items appear one-by-one
  - numbers, points, edges, …
  - (usually) adversarially ordered
  - one or a small number of passes over the stream

- Goal: approximate a function of the underlying stream
  - use small amount of space (in bits)

- Efficiency: usually necessary for algorithms to be both randomized and approximate

Example: Statistical Problems

- Sequence of updates to an underlying vector x
- Initially, $x = 0^n$
- t-th update (i, $Delta_t$) causes

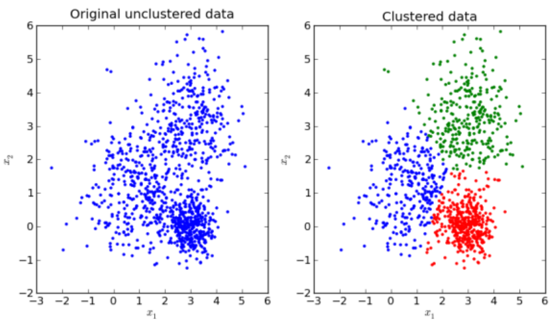
$$x_i \leftarrow x_i + Delta_t$$

- Approximate a function f(x)
  - Order-invariant function f
- If all $Delta_t > 0$, called the insertion model
- Otherwise, called the turnstile model
- Examples: $f(x) = |x|_p$, $f(x) = H(x/|x|_1)$, $|supp(x)|$

Example: Geometric Problems

- Sequence of points $p_1, \ldots, p_n$ in R^d

- Clustering problems
  - Family F of shapes (points, lines, subspaces)
  - Output: $\text{argmin}_{\{S \subset F, |S|=k\}} \text{sum}_i \, d(p_i, S)^z$
    - $d(p_i, S) = \min_{f \text{ in } S} d(p_i, f)^z$
    - k-median, k-means, PCA



- Distance problems
  - Typically points $p_1, \ldots, p_{2n}$ in R^2
  - Estimate minimum cost perfect matching
  - If n points are red, and n points are blue, estimate minimum cost bi-chromatic matching (EMD)

Example: String Processing

- Sequence of characters $\sigma_1, \sigma_2, \ldots, \sigma_n \in \Sigma$
- Often problem is not order-invariant
- Example: Longest Increasing Subsequence (LIS)
  - $\sigma_1, \sigma_2, \ldots, \sigma_n$ is a permutation of numbers from 1, 2, …, n
  - Find the longest length of a subsequence which is increasing

5,3,0,7,10,8,2,13,15,9,2,20,2,3. LIS=6

Outline

1. Streaming model and examples

2. Background on communication complexity for streaming
   1. Product distributions
   2. Non-product distributions

3. Open problems

Communication Complexity

- Why are streaming problems hard?
- Don’t know what will be important in the future and can’t remember everything…
- How to formalize?
- Communication Complexity

Typical Communication Reduction

$a \in \{0,1\}^n$
Create stream s(a)

$b \in \{0,1\}^n$
Create stream s(b)

Lower Bound Technique

1. Run Streaming Alg on s(a), transmit state of Alg(s(a)) to Bob

2. Bob computes Alg(s(a), s(b))

3. If Bob solves g(a,b), space complexity of Alg at least the 1-way communication complexity of g

Example: Distinct Elements

- Give $a_1, \ldots, a_m$ in [n], how many *distinct* numbers are there?

- Index problem:
  - Alice has a bit string x in $\{0, 1\}^n$
  - Bob has an index i in [n]
  - Bob wants to know if $x_i = 1$

- Reduction:
  - $s(a) = i_1, \ldots, i_r$, where i_j appears if and only if $x_{i_j} = 1$
  - $s(b) = i$
  - If Alg(s(a), s(b)) = Alg(s(a))+1 then $x_i = 0$, otherwise $x_i = 1$

- Space complexity of Alg at least the 1-way communication complexity of Index

1-Way Communication of Index

- Alice has uniform $X \in \{0,1\}^n$
- Bob has uniform I in [n]
- Alice sends a (randomized) message M to Bob

- $I(M ; X) = \text{sum}_i \, I(M ; X_i \mid X_{< i})$
  $\geq \text{sum}_i \, I(M; X_i)$
  $= n - \text{sum}_i \, H(X_i \mid M)$

- By Fano's inequality, $H(X_i \mid M) < H(\delta)$ if Bob can predict X_i with probability $> 1- \delta$
- $CC_\delta(\text{Index}) > I(M ; X) \geq n(1-H(\delta))$
- Computing distinct elements requires $\Omega(n)$ space

Indexing is Universal for Product Distributions [Kremer, Nisan, Ron]

- If inputs drawn from a *product distribution*, then 1-way communication of a Boolean function is Θ(VC-dimension) of its communication matrix (up to δ dependence)

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

- Implies a reduction from Index is optimal
  - Entropy, linear algebra, spanners, norms, etc.
  - Not always obvious how to build a reduction, e.g., Gap-Hamming

Gap-Hamming Problem

$x \in \{0,1\}^n$ $\qquad$ $y \in \{0,1\}^n$

- Promise: Hamming distance satisfies $\Delta(x,y) > n/2 + \varepsilon n$ or $\Delta(x,y) < n/2 - \varepsilon n$
- Lower bound of $\Omega(\varepsilon^{-2})$ for randomized 1-way communication [Indyk, W], [W], [Jayram, Kumar, Sivakumar]
- Gives $\Omega(\varepsilon^{-2})$ bit lower bound for approximating number of distinct elements
- Same for 2-way communication [Chakrabarti, Regev]

Gap-Hamming From Index [JKS]

Public coin = $r^1, \ldots, r^t$, each in $\{0,1\}^t$

$t = \varepsilon^{-2}$

$x \in \{0,1\}^t$ $\rightarrow$ $a \in \{0,1\}^t$

$i \in [t]$ $\rightarrow$ $b \in \{0,1\}^t$

$a_k = \text{Majority}_{j \text{ such that } x_j = 1} \, r^k_j$

$b_k = r^k_i$

$E[\Delta(y,z)] = t/2 + x_i \cdot t^{1/2}$

Augmented Indexing

- Augmented-Index problem:
  - Alice has $x \in \{0, 1\}^n$
  - Bob has $i \in [n]$, and $x_1, \ldots, x_{i-1}$
  - Bob wants to learn x_i

- Similar proof shows $\Omega(n)$ bound
- $I(M ; X) = \text{sum}_i \, I(M ; X_i \mid X_{< i})$

  $= n - \text{sum}_i \, H(X_i \mid M, X_{< i})$

- By Fano's inequality, $H(X_i \mid M, X_{< i}) < H(\delta)$ if Bob can predict X_i with probability $> 1- \delta$ from $M, X_{< i}$
- $CC_\delta(\text{Augmented-Index}) > I(M ; X) \geq n(1-H(\delta))$
- Surprisingly powerful implications

Indexing with Low Error

- Index Problem with 1/3 error probability and 0 error probability both have $\Theta(n)$ communication

- In some applications want lower bounds in terms of error probability

- Indexing on Large Alphabets:
  - Alice has $x \in \{0,1\}^{n/\delta}$ with $wt(x) = n$, Bob has $i \in [n/\delta]$
  - Bob wants to decide if $x_i = 1$ with error probability δ
  - [Jayram, W] 1-way communication is $\Omega(n \log(1/\delta))$

Compressed Sensing

- Compute a sketch S·x with a small number of rows (also known as measurements)
  - S is oblivious to x

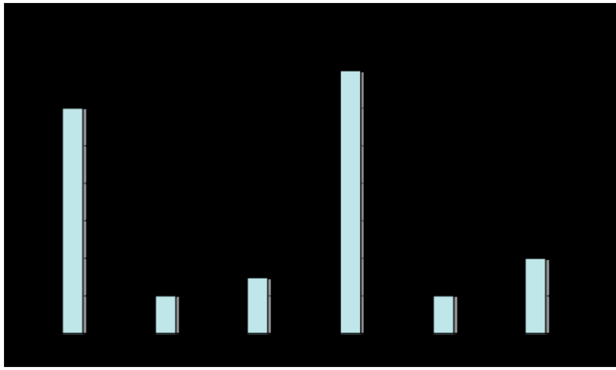
- For all x, with constant probability over S, from S·x, we can output x' which approximates x:

$$|x'-x|_2 \le (1+\varepsilon)\ |x-x_k|_2$$

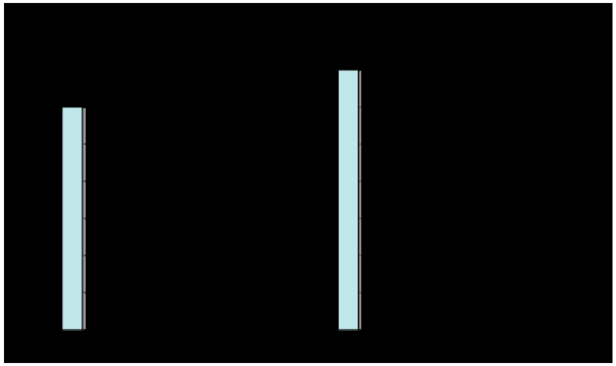
where x_k is an optimal k-sparse approximation to x (x_k is a "top-k" version of x)

- Optimal lower bound on number of rows of S via reduction from Augmented-Indexing

  - Bob's partial knowledge about x is crucial in the reduction



x



x_2

Recognizing Languages

-2-way communication tradeoff for Augemented Indexing: if Alice sends $n/2^b$ bits then Bob sends $\Omega(b)$ bits
[Chakrabarti, Cormode, Kondapally, McGregor]

- Streaming lower bounds for recognizing DYCK(2) [Magniez, Mathieu, Nayak]

((([])()[])) ∈ DYCK(2) ([([]])[])) ∉ DYCK(2)

- Multi-pass $\Omega(n^{1/2})$ space lower bound for length-n streams

- Interestingly, one forward pass plus one backward pass allows for an O~(log n) bits of space

Outline

1. Streaming model and examples

2. Background on communication complexity for streaming
   1. Product distributions
   2. Non-product distributions

3. Open problems

Non-Product Distributions

- Needed for stronger lower bounds
- Example: approximate $|x|_\infty$ up to a multiplicative factor of B in a stream
  - Lower bounds for heavy hitters, p-norms, etc.

$\text{Gap}_\infty(x,y)$ Problem

$x \in \{-B, \ldots, B\}^n$

$y \in \{-B, \ldots, B\}^n$

- Promise: $|x-y|_\infty \leq 1$ or $|x-y|_\infty \geq B$
- Hard distribution non-product
- $\Omega(n/B^2)$ 2-way lower bound [Saks, Sun] [Bar-Yossef, Jayram, Kumar, Sivakumar]

Direct Sums

- $Gap_\infty(x,y)$ doesn't have a hard product distribution, but has a hard distribution $\mu = \lambda^n$ in which the coordinate pairs $(x_1, y_1), \ldots, (x_n, y_n)$ are independent
  - w.pr. 1-1/n, (x_i, y_i) random subject to $|x_i - y_i| \leq 1$
  - w.pr. 1/n, (x_i, y_i) random subject to $|x_i - y_i| \geq B$
- Direct Sum: solving $Gap_\infty(x,y)$ requires solving n single-coordinate sub-problems f
- In f, Alice and Bob have $J,K \in \{-M, \ldots, M\}$, and want to decide if $|J-K| \leq 1$ or $|J-K| \geq B$

Direct Sum Theorem

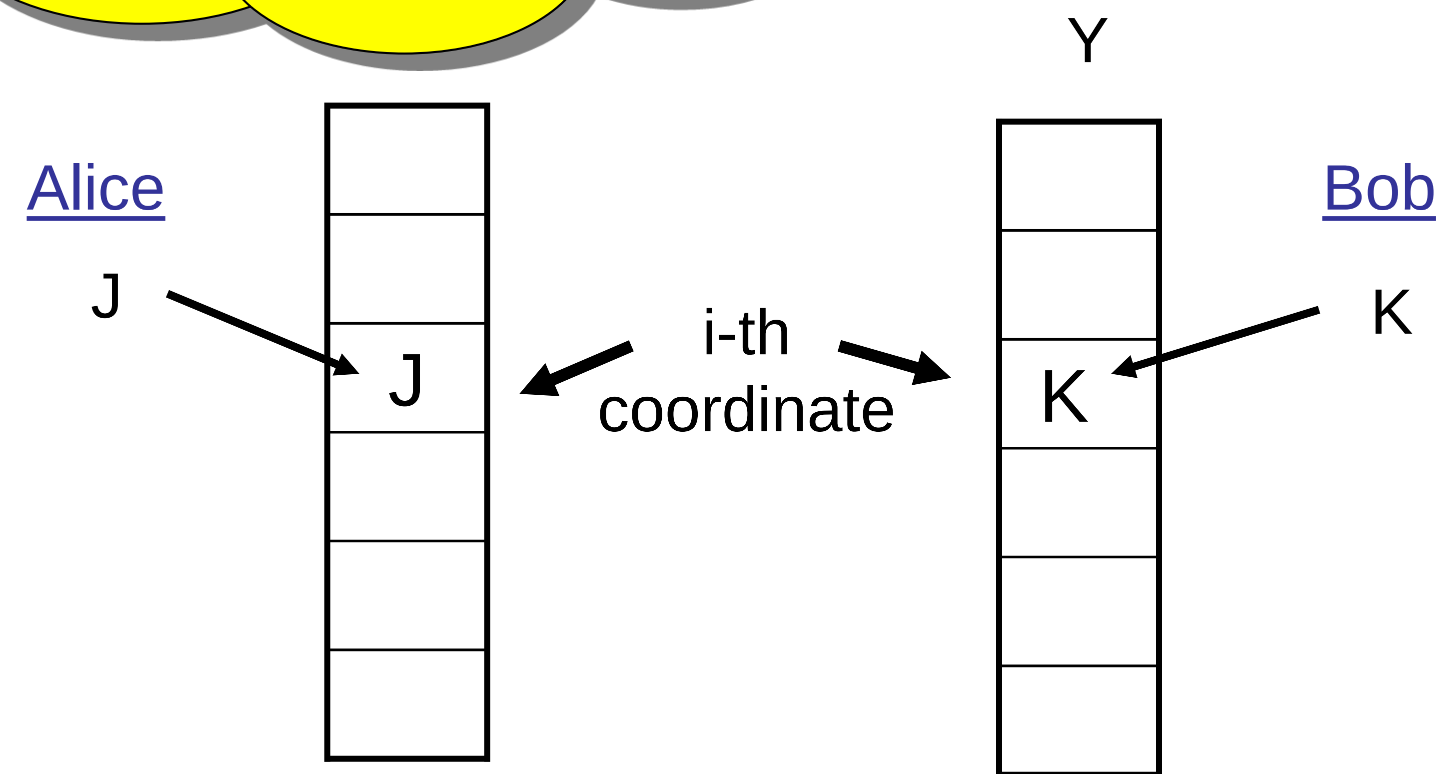
- π is the transcript between Alice and Bob
- For X, Y $\sim$ μ, $I(\pi ; X, Y) = H(X,Y) - H(X,Y \mid \pi)$ is the (external) information cost
- [BJKS]: ?!?!?!?! the protocol has to be correct on every input, so why not measure $I(\pi ; X, Y)$ when (X,Y) satisfy $|X-Y|_\infty \leq 1$?
  - Is $I(\pi ; X, Y)$ large?
- Redefine $\mu = \lambda^n$, where $(X_i, Y_i) \sim \lambda$ is random subject to $|X_i-Y_i| \leq 1$
- $IC(f) = \inf_\psi I(\psi ; A, B)$, where ψ ranges over all 2/3-correct protocols for f, and $A,B \sim \lambda$

Is $I(\pi ; X, Y) = \Omega(n) \cdot IC(f)$?

The Embedding Step

Suppose Alice and Bob could fill in the remaining coordinates j of X, Y so that $(X_j, Y_j) \sim \lambda$

Then we get a correct protocol for f!



Conditional Information Cost

- $(X_j, Y_j) \sim \lambda$ is not a product distribution

- [BJKS] Define $D = ((P_1, V_1) \ldots, (P_n, V_n))$:
  - P_j uniform in {Alice, Bob}
  - V_j uniform $\{-B+1, \ldots, B-1\}$
  - If P_j = Alice, then $X_j = V_j$ and Y_j is uniform in $\{V_j, V_j-1, V_j+1\}$
  - If P_j = Bob, then $Y_j = V_j$ and X_j is uniform in $\{V_j, V_j-1, V_j+1\}$

X and Y are independent conditioned on D!

- $I(\pi ; X, Y \mid D) = \Omega(n) \cdot IC(f \mid (P,V))$

- $IC(f) = \inf_\psi I(\psi ; A, B \mid (P,V))$, where ψ ranges over all 2/3-correct protocols for f, and $A,B \sim \lambda$

Primitive Problem

- Need to lower bound IC(f | (P,V))
- For fixed P = Alice and V = v, this is I(ψ ; K) where K is uniform over v, v+1
- Basic information theory: $I(\psi ; K) \geq D_{JS}(\psi_{v,v} , \psi_{v, v+1})$
- $IC(f \mid (P,V)) \geq E_v [D_{JS}(\psi_{v,v} , \psi_{v, v+1}) + D_{JS}(\psi_{v,v} , \psi_{v+1, v})]$

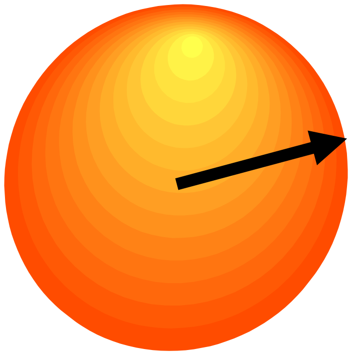
Forget about distributions, let's move to unit vectors!

tance

- This proof just needs the triangle inequality of Euclidean distance

- Other properties sometimes useful, such as short diagonals [Jayram, W]

• tor



• $D_{JS}(\psi_{v,v} , \psi_{v, v+1}) \geq |S(\psi_{v,v}) - S(\psi_{v,v+1})|_2^2$

(*) $IC(f \mid (P,V)) \geq E_v [|S(\psi_{v,v}) - S(\psi_{v,v+1})|_2^2 + |S(\psi_{v,v}) - S(\psi_{v+1,v})|_2^2]$

• Because ψ is a protocol,
  – (Cut-and-paste): $|S(\psi_{a,b}) - S(\psi_{c,d})|_2^2 = |S(\psi_{a,d}) - S(\psi_{b,c})|_2^2$]
  – (Correctness): $|S(\psi_{0,0}) - S(\psi_{0,B})|_2^2 = \Omega(1)$

• Minimizing (*) subject to these properties, $IC(f \mid (P,V)) = \Omega(1/B^2)$

Direct Sum Wrapup

- $\Omega(n/B^2)$ bound for $Gap_\infty(x,y)$
- Similar argument gives $\Omega(n)$ bound for disjointness [BJKS]
- [MYW] Sometimes can “beat” a direct sum: solving all n copies simultaneously with constant probability as hard as solving each copy with probability 1-1/n
  - E.g., 1-way communication complexity of Equality
- Direct sums are nice, but often a problem can’t be split into simpler smaller problems, e.g., no known embedding step in gap-Hamming

Outline

1. Streaming model and examples

2. Background on communication complexity for streaming
    1. Product distributions
    2. Non-product distributions

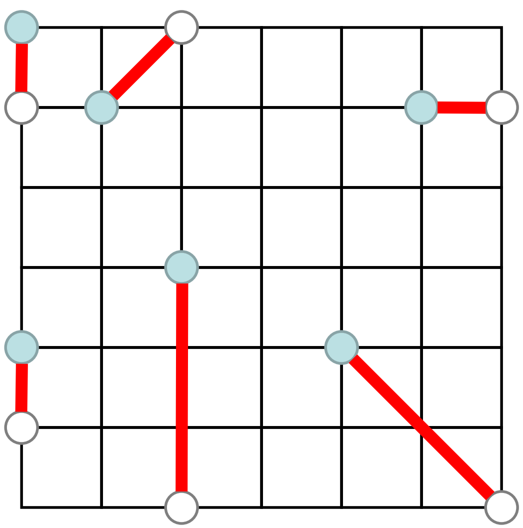
3. Open problems

Earthmover Distance

- For multisets A, B of points in $[\Delta]^2$, $|A|=|B|=N$,

$$\mathrm{EMD}(A,B) = \min_{\pi:A\to B} \sum_{a\in A} \|a - \pi(a)\|$$

i.e., min cost of perfect matching between A and B



$\mathrm{EMD}(\bullet, \circ) = 6 + 3\sqrt{2}$

Upper bound:

O(1/γ)-approximation using Δ^{γ} bits of space, for any $\gamma > 0$

Lower bound:

$\log \Delta$ bits, even for $(1+\varepsilon)$-approx.

Can we close this huge gap?

Longest Increasing Subsequence

- Permutation of 1, 2, …, n given one number at a time
- Find the longest length of an increasing subsequence

  - 5,3,0,7,10,8,2,13,15,9,2,20,2,3.   LIS=6

- For finding the exact length, $\tilde{\Theta}(|LIS|)$ is optimal for randomized algorithms
- For finding a (1+ε)-approximation, $\tilde{\Theta}(n^{1/2})$ is optimal for deterministic algorithms
- For randomized algorithms we know nothing!

Is polylog(n) bits of space possible for (1+ε)-approximation?

Matchings

- Given a sequence of edges $e_1, \ldots, e_m$, output an approximate maximum matching in O~(n) bits of space

- Greedy algorithm gives a ½-approximation

- [Kapralov] no 1-1/e approximation is possible in O~(n) bits of space

Is there anything better than the trivial greedy algorithm?

- Suppose we allow edge deletions, so we have a sequence of insertions and deletions to edges that have already appeared

Can one obtain a $\Omega(1)$-approximation in $o(n^2)$ bits of space?

Matrix Norms

- Let A be an n x n matrix of integers of magnitude at most poly(n)
- Suppose you see the entries of A one-by-one in a stream in an arbitrary order

How much space is needed to estimate the operator norm $|A|_2 = \sup_x |Ax|_2/|x|_2$ up to a factor of 2?

[Li, Nguyen, W], [Regev]: if the entries of A are real numbers and $L: R^{n^2} \rightarrow R^k$ is a linear map chosen independent of A, then $k = \Omega(n^2)$ to estimate $|A|_2$ up to a factor of 2

- *Can we even rule out linear maps in the discrete case?*