IBM Almaden Research Center
D. Sivakumar
Ravi Kumar
T.S. Jayram
Ziv Bar-Yossef

Information Statistics Approach to Data Stream and Communication Complexity
The Data Stream Model

- Generalization: $\ell$-pass data stream algorithms
- Main measure of complexity: space
- Algorithms are allowed to be randomized and approximate
- Input arrives in a one-way stream in arbitrary order

[HRR98, AMS96, FKS99]
Motivation

- Processing search engine query logs
- Web crawling
- Web Information Retrieval
- One-pass algorithms for large database relations
- Database
- Processing streams of IP packets
- Networking

Motivation
Algorithmic Results in Data Streams

1. Triangle counting [BK02]
2. Inversion counting [AK02]
3. Clustering [GM00]
4. Quantiles [MR98, MR99, GK01]
5. Histograms [GMP07, Mv00, GKS01, GKr00]
6. Distances and norms [FS99, FS00, 100]
7. Inversion counting [FM89, AMS96, GKS01, BKKS02, CKP02]
8. Frequency statistics

Algorithms Results in Data Streams
Example 1: Frequency Moments

\[ F_k(a_1; \ldots; a_n) = \sum_{j=1}^{m} f_k^j \]

where, \( a_1; \ldots; a_n \) \( \in \mathbb{M} \), and

\[ [u] \in \{ u_1; \ldots; u_n \} \]

\[ \bigcup_{u} \bigwedge_{u_1}^{u_n} = (u_1; \ldots; u_n) \]

Theorem 1 [This paper]

For any \( k \), requires polynomial space

Theorem [Alon, Matias, Szegedy, 96]

\[ F_0, F_1, F_2 \] can be approximated in logarithmic space

\[ F_k, \quad k > 5, \] needs polynomial space

\( F_k, \quad k > 5, \) requires polynomial space

Theorem 1 [This paper]

\( F_k, \) for any \( k > 2, \) requires polynomial space

\( F_k, \) for any \( k > 2, \) requires polynomial space

\[ \# = 0_F \]

\[ u = 1_F \]

\[ z < 3_F \]

\[ z < 3_F \]

where, \( a_1; \ldots; a_n \) \( \in \mathbb{M} \), and

\[ [u] \in \{ u_1; \ldots; u_n \} \]

\[ \bigcup_{u} \bigwedge_{u_1}^{u_n} = (u_1; \ldots; u_n) \]

Example 1: Frequency Moments
Example 2: \( \ell^p \)-Distance

\[
\| q - p \|_{\ell^p} \equiv \left( \sum_{j \in \mathbb{P}} (q_j - p_j)^p \right)^{\frac{1}{p}}
\]

\( p \) and \( q \) are given in arbitrary order.

\( d \)

\( \ell^p \)-Distance

\( p > 2 \) requires polynomial space, even for a constant number of passes over the input.

\( \ell^d \), for \( p > 2 \), requires polynomial space, even for a constant number of passes over the input.

Any one pass algorithm for \( \ell^d \), for \( p < d \), requires polynomial space \([Saks, Sun '02]\).

\[ \text{Theorem 2} \] \text{(This paper)}

\( \ell^d \), for \( p < d \), requires polynomial space, even for a constant number of passes over the input.

\[ \text{Theorem 2} \] \text{(This paper)}

\( \ell^d \), for \( p < d \), requires polynomial space, even for a constant number of passes over the input.
$R^g(f) = \text{Randomized communication complexity of } f \text{ with error } g$
Multi-Party Communication Complexity

(number in the hand)
Applications of Communication Complexity

- Circuit depth
  [Karchmer, Wigderson, Razborov, '90, Raz, Wigderson, '90]

- Time-space tradeoffs
  [Karchmer, Wigderson, Razborov, '98, Razborov, '90, Raz, Wigderson, '90]

- Data structure complexity
  [Miltersen, '95, Miltersen et al., '95]

- Decision tree complexity
  [Nisan, '93]

- Computation economics
  [Lengauer, '90]

- Pseudorandom generators for logarithmic space
  [Babai, Nisan, Szegedy, 89, Impagliazzo, Nisan, Wigderson, '94]

- Time-space tradeoffs

- Computational Economics
  [Chandra, Furst, Lipton, '83, Alon, Maass, '86, Babai, Nisan, Szegedy, '89]

- Circuit depth
  [Chandra, Furst, Lipton, '83, Alon, Maass, '86, Babai, Nisan, Szegedy, '89]

- Data stream space complexity
  [Nisan, Segal, Deng, Papadihal, Seta, '02]

- Circuit depth
  [Karchmer, Wigderson, Razborov, '90, Raz, Wigderson, '90]
One-Way Communication Complexity

Easy reduction:

Protocol for $g$-party $1$-way $t$-party communication

Reduction

Algorithm for $f$ data stream space

Lower bound for $g$

Lower bound for $f$
Example 1: Set-Disjointness

Input: \( S; T \)

\[ D_{ISJ}^n(S; T) = 1 \iff S \cap T = \emptyset. \]

Multi-party set-disjointness

Input: \( S_1, \ldots, S_t \)

\[ D_{DISJ}^n(S_1, \ldots, S_t) = 1 \iff [u] \supseteq S_1, \ldots, S_t, \quad [u] \subseteq S_1, \ldots, S_t. \]

Example 1: Set-Disjointness

Input: \( L \cup S \)

\[ D_{DISJ}^n(L, S) = 1 \iff [u] \supseteq L, S \]

\[ D_{DISJ}^n(L, S) = 1 \iff [u] \subseteq L, S. \]
Corollary

Using the reduction from $\text{DISJ}^{u+\frac{u}{\log u}}$ to $\text{DISJ}^u$ [AMS96]:

\begin{align*}
0 < \epsilon < 1 & \Rightarrow (\frac{t}{u})^0 \text{DISJ}^{u+\frac{u}{\log u}} \leq (\frac{t}{u})^0 \text{DISJ}^{u+\epsilon} \leq (\frac{t}{u})^0 \text{DISJ}^{u+\frac{u}{\log u}}.
\end{align*}

Theorem 3 [This Paper]

Example 1: Set-Disjointness (cont.)
stream model, even for a constant number of passes over the input.

Estimating \( \ell^1 \) to within \( u \) requires \( d \approx \frac{\alpha}{\sqrt{2} - 2} \) space in the data. 

Corollary

\[
(\frac{m}{u})^{\mathcal{U}} = \Theta\left(\ell^1, \frac{m}{u}\right)
\]

Theorem 4 [Saks, Sun '02]

\[
(\frac{m}{u})^{\Theta} = \Theta\left(\ell^1, \frac{m}{u}\right)
\]

Theorem 4 [This paper]

\[
(\frac{m}{u})^{\mathcal{U}} = \Theta\left(\ell^1, \frac{m}{u}\right)
\]

\[
(\frac{m}{u})^{\Theta} = \Theta\left(\ell^1, \frac{m}{u}\right)
\]

Corollary

\[
(\frac{m}{u})^{\mathcal{U}} = \Theta\left(\ell^1, \frac{m}{u}\right)
\]

\[
(\frac{m}{u})^{\Theta} = \Theta\left(\ell^1, \frac{m}{u}\right)
\]

\[
(\frac{m}{u})^{\mathcal{U}} = \Theta\left(\ell^1, \frac{m}{u}\right)
\]

\[
(\frac{m}{u})^{\Theta} = \Theta\left(\ell^1, \frac{m}{u}\right)
\]

Example 2: \( \ell^1 \) Promise Problem

\[
\forall \alpha, \forall \mathcal{U} \in [m] \text{ satisfying one of the following:}
\]

\[
\forall \alpha, \forall \mathcal{U} \in [m] \text{ satisfying one of the following:}
\]

\[
\forall \alpha, \forall \mathcal{U} \in [m] \text{ satisfying one of the following:}
\]
Outline of the Lower Bound Technique

1. Generalization of Information Complexity
2. A Direct Sum Theorem for Information Complexity
3. Statistical Lower Bounds on Information Complexity of “Primitive” Functions
The Direct Sum Question

Example

Is a protocol for \( \text{DISJ}_n \) the "sum" of \( n \) protocols for AND?

\[
\forall x, y \in \{0, 1\}^n \exists \{ f_i \} \subseteq \text{DISJ}_n^i \quad \text{such that} \quad (\forall i \in \{1, \ldots, n\}) f_i(x) \oplus f_i(y) = f(x) \oplus f(y)
\]

where \( \{ f_i \} \) are the characteristic vectors of the two sets.
Entropy and Mutual Information

\[ H(X) - H(Y) = H(Y|X) \]

Proposition [Subadditivity of entropy]

\[ H(X;Y) \leq H(X) + H(Y) \]

Equality iff \( X, Y \) are independent.

\( I(X;Y) \) def = \( H(X) - H(X|Y) \)

\( H(X) \geq H(Y) \geq H(Y|X) \)

\( (X|Y)H - (Y|X)H = (Y|X)H - (X)H = (X|Y)H \)

\( I(X;Y) \)

Entropy and Mutual Information
**Definition [Information Complexity]**

\[ (f) \geq \min_{\text{error}(\Pi)} (\lambda X, Y) I(X; Y) \]

\[ (f) \geq \min_{\text{error}(\Pi)} (\lambda X, Y) I(X; Y) \]

Proposition: For every protocol that computes \( f \) has to reveal about its inputs, the amount of information a protocol that computes \( f \) has to reveal about its inputs.

\[ (f) \geq \min_{\text{error}(\Pi)} (\lambda X, Y) I(X; Y) \]

\[ (f) \geq \min_{\text{error}(\Pi)} (\lambda X, Y) I(X; Y) \]

Chakrabarti, Shi, Wirth, Yao 01, Ablayev, 93, Saks, Sun 02.

Information Complexity
Conditional Information Complexity

\[ (f)^{g \cdot r \cdot f} I \geq (D | f)^{g \cdot r \cdot f} I \]

Proposition: For every \( D \), \( I^D \)

\[ (D | (\lambda X) I (\lambda X)) I \geq \min_{\Pi} \text{error}(\Pi) = (D | f)^{g \cdot r \cdot f} I \]

[Definition: Conditional Information Complexity]

Let \( D \) be a random variable such that \( p_{\{p = \alpha\}} \). Let \( p \) be a random variable. Express as convex combination of product distributions. Express \( p_{\{p = \alpha\}} \) as non-product.

Express as convex combination of product distributions. For every \( D \), \( I^D \). For direct sum, we need \( p_{\{p = \alpha\}} \) to be product. For set-disjointness, we need \( p_{\{p = \alpha\}} \) to be non-product.

Conditional Information Complexity
\text{Input Distribution for Set-Disjointness}

\begin{align*}
\text{is concentrated on 0's of } \text{DISJ}_n \quad &\bullet \\
\text{Conditioned on } D = B, \text{ is product} \quad &\bullet \\
\text{Conditioned on } D \neq B, \text{ is non-product} \quad &\bullet \\
\end{align*}

\begin{align*}
\{1\} \; \forall n \in \mathbb{N} \quad &\text{Let } X = 0 = X \quad &\bullet \\
\{0\} \; \forall n \in \mathbb{N} \quad &\text{Let } A = A \quad &\bullet \\
\{B \forall n \in \mathbb{N} \quad &\bullet \\
\end{align*}

\text{Input Distribution defined as follows:} \{0\} \; \forall \sim (X, X)
Theorem

Let \( \mathcal{P} \) be a protocol for \( \mathcal{D}_{\text{ISJ}}^{n; 2j} \).

1. Decomposition step:

\[
\mathcal{D} \mid (\lambda x \eta \gamma) \Rightarrow \mathcal{D} \mid (\lambda x \eta \gamma x) I \\
\]

2. Reduction step:

\[
\mathcal{D} \mid (\lambda x \eta \gamma x) I \Rightarrow \mathcal{D} \mid (\lambda x \eta \gamma x) I \\
\]

Proof. Let II be a protocol for \( \mathcal{D}_{\text{ISJ}}^{n; 2} \).

\[
(\mathcal{D} \mid (\lambda x \eta \gamma) \text{AND}) I C \cdot u \Rightarrow (\mathcal{D} \mid \text{DISJ}^{n; 2} \text{AND}) I C
\]

Direct Sum for Information Complexity
\[
\left( p \mid (\mathcal{A},X) \Pi \colon \mathcal{X}, \mathcal{X} \right) \mathcal{I} \leq \frac{1}{2}
\]

(by independence of \( f \),
and subadditivity of entropy)

\[
\left( p \mid (\mathcal{A},X) \Pi \colon \mathcal{X}, \mathcal{X} \right) \mathcal{H} \leq \left( p \mid \mathcal{X}, \mathcal{X} \right) \mathcal{H} \leq
\]

\[
\left( p \mid (\mathcal{A},X) \Pi \colon \mathcal{X}, \mathcal{X} \right) \mathcal{H} - \left( p \mid \mathcal{X}, \mathcal{X} \right) \mathcal{H} =
\]

ProofofDecompositionStep
1. Expand

2. Create a protocol for computing AND

\[
(q, p) \rightarrow (q, p)
\]

Expanding over all

\[
(A | (A, X) \Pi ^{\ell} \lambda X) I
\]

Proof of Reduction Step
Recall the distribution:

\[ \{ \{1,0\}^{\mathcal{B}} \in \mathcal{A} \} \cdot \frac{\zeta}{1} = \]

\[ \left[ (\mathcal{B} = A \mid (\lambda, X, \mathcal{D} \{^\lambda\}) I + (\mathcal{V} = A \mid (\lambda, X, \mathcal{D} \{^\lambda\}) I \right] \cdot \frac{\zeta}{1} = \]

\[ (\mathcal{B} \mid (\lambda, X, \mathcal{D} \{^\lambda\}) I \]

Suppose \( \mathcal{P} \) is a protocol for \( \text{AND} \). Then,

\[ \{ \{1,0\}^{\mathcal{B}} \in \mathcal{X} \mid \lambda \} \quad \text{if} \quad \mathcal{B} = \mathcal{A} \mid \lambda \quad \text{and} \quad \mathcal{X} = \{ \{0,1\}^{\mathcal{B}} \in \mathcal{A} \mid \lambda \} \quad \text{if} \quad \mathcal{B} = \mathcal{A} \mid \lambda \quad \text{and} \quad \mathcal{X} = \{ \{0,1\}^{\mathcal{B}} \in \mathcal{A} \mid \lambda \} \quad \text{if} \quad \mathcal{B} \in \mathcal{A} \mid \lambda \]

Recall the distribution:

\[ (\mathcal{B} \mid \text{IC}_{\mathcal{D}}(\text{AND}) (\mathcal{D}) ) \]
\[ ((\mathbb{I}, 0) \mathcal{D}_r (0, \mathbb{I}) \mathcal{D}_r) S I \cdot \frac{\gamma}{\tau} \leq \frac{\gamma}{\tau} \]

\[ [((\mathbb{I}, 0) \mathcal{D}_r (0, 0) \mathcal{D}_r) S I + ((0, \mathbb{I}) \mathcal{D}_r (0, 0) \mathcal{D}_r) S I] \cdot \frac{\gamma}{\tau} = \]

\[ [((\mathbb{I}, 0) \mathcal{D}_r \mathcal{D}_r I + (0, \mathbb{I}) \mathcal{D}_r \mathcal{D}_r I) \cdot \frac{\gamma}{\tau} = \]

\[ (\mathcal{D} \mid (\lambda X) \mathcal{D}_r \mathcal{D}_r \lambda X) I \]

\[ (\mathbb{I}, \mathbb{I}) I \overset{\text{def}}{=} (\mathbb{I}, 0, \mathbb{I}) \mathcal{D}_r \]

\[ \{0, 1\} \cup \mathbb{I} \]

\[ \frac{\gamma}{\tau} \]

\[ \mathcal{D}_r \]

\[ \text{two distributions} \]

Jensen-Shannon Divergence
If \( P \) computes \( \text{AND} \), why should \( P(0, 1) \) be far from \( P(1, 0) \)?

The large distance is on the other diagonal.

\( \text{AND} \) is 0 on both of these inputs.

\( \text{AND} \) computes \( \text{AND} \), why should \( P(0, 1) \) be far from \( P(1, 0) \)?

A Point to Ponder
Let \( \Pi \) be a deterministic protocol.

Fix a transcript \( \tau \).

Then, \( \Pi_{\tau} \) is a combinatorial rectangle:

\[
\{ \tau = (f', x) : (f', x) \} = \Pi_{\tau}
\]

Let \( \Pi \) be a deterministic protocol.

Rectangle Property of Communication Complexity
A Probabilistic Analog

Let $\Pi$ be a randomized protocol.

Fix a transcript $\tau$

Then, there exist functions $p_{\tau} : \mathcal{X} \to [0, 1]$ and $q_{\tau} : \mathcal{Y} \to [0, 1]$ such that

$$\Pr[\Pi(x, y) = \tau] = p_{\tau}(x) \cdot q_{\tau}(y), \quad \forall x, y$$
Let $P$ and $Q$ be two probability distributions. The Hellinger distance is defined as:

$$((0,1)P \cdot (1,0)Q) \frac{1}{2} \geq \left[(((0,1)P \cdot (0,0)Q) \frac{1}{2} + (((1,0)P \cdot (0,0)Q) \frac{1}{2}) \right. \cdot \left. \frac{1}{2} \geq \left[(((0,1)P \cdot (0,0)Q) JS + (((1,0)P \cdot (0,0)Q) JS \right] \cdot \frac{1}{2} = (D \mid \lambda X) P \mid \lambda X) \frac{1}{2}$$

is a metric.

$(\delta ' d) \frac{1}{2} \leq (\delta ' d) JS \frac{1}{2}$

$(\nu) (\delta (\nu) d)^{\nu Z} - 1 = (\delta ' d) \frac{1}{2}$

Hellinger Distance
\[ ((\hat{h} \cdot x) \downarrow (\hat{h} \cdot x)) \parallel Y - 1 = \]

\[ (\bot = (\hat{h} \cdot x) \downarrow \bot \cdot (\bot = (\hat{h} \cdot x) \downarrow \bot) \downarrow A \parallel \bot = \]

\[ ((\hat{h} \cdot x) \downarrow A \cdot (\hat{h} \cdot x) \downarrow A \cdot (\hat{h} \cdot x) \downarrow A \parallel \bot = \]

\[ ((\hat{h} \cdot x) \downarrow x) \parallel Y - 1 \]

A Cut & Paste Lemma
\[(X;Y;P(X;Y)|D) \leq (P(0;1);P(1;0)\text{ (Correctness of } P)) \leq (D | (\lambda X)P(\lambda X)) I\]

Therefore:

\[R^g(D\text{ISJ}_n;2j\sim D\text{ISJ}_n;2j~D) = (D | (\lambda X)P(\lambda X)) I\]
divergences
- Uses generalizations of the Hellinger distance – Renyi!
- Exploits the Markovian structure of one-way protocols.

\[ \mathcal{U}(1/\varepsilon) \] bound for one-way communication:

\[ \mathcal{U}(1/t+\varepsilon) \] bound for general communication:

\[ U(1/t) \] bound for one-way communication:

\[ U(1/t) \] bound for general communication:

Need a lower bound for \( t \)-bit AND.

Same direct sum argument

\( t \)-party set-disjointness
Need a lower bound for the difference problem: for \( q, \in \mathbb{N} \),

\[
\frac{1}{z} \geq ((m,m) \mathcal{I},(0,0) \mathcal{I})_z^y + ((0,m) \mathcal{I},(0,0) \mathcal{I})_z^y
\]

A Pythagorean lemma:

\[
|q - v| \geq |q - v| \text{ or } 1 \text{ or } \mathbb{N}^{\mathbb{Z}/m^2}
\]

Decide whether \( m, \in \mathbb{N} \).

Same direct sum argument

\( \ast \) Promise Problem
Conclusions

A powerful lower bound methodology in communication complexity

Several novel ideas introduced:
- Conditional information complexity
- Crisp connections between statistical distance measures and reduction to proving lower bounds for "simple" functions
- Conditional information complexity

Method gives strong results even for promise problems particularly useful for data stream lower bounds

Conclusions
Subsequent Work

- Distributional communication complexity lower bounds \( \Omega(n^{t^3/2}) \) bound for \( t \)-party set-disjointness [Khot '02].
- Generalization to AND-OR trees of depth 3 [Jayram, Kumar, Sivakumar '02].