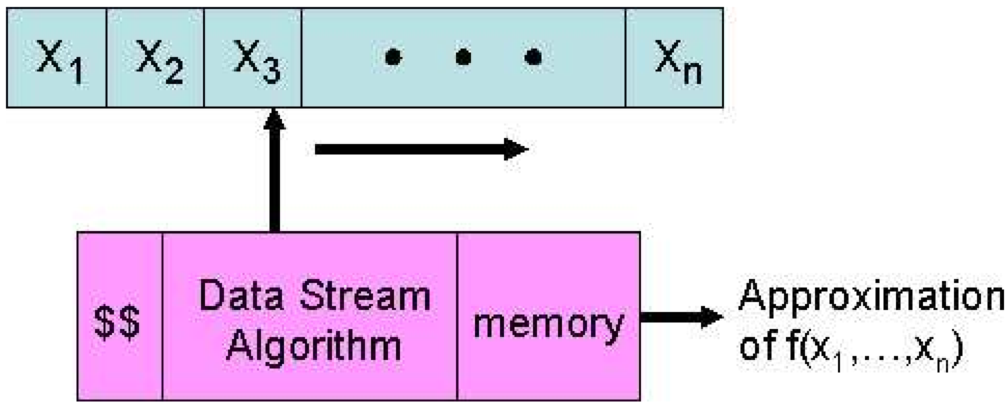
Information Statistics Approach to Data Stream and Communication Complexity

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

The Data Stream Model

[HRR98, AMS96, FKSV99]



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

Motivation

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

Algorithmic Results in Data Streams

- Frequency statistics
  [FM83, AMS96, GT01, BKS02, BJKST02, CCF02, KPS02]
- Distances and norms [FKSV99, FS00, I00]
- Histograms [GMP97, MVW00, GKS01, GGIKMS02]
- Quantiles [MRL98, MRL99, GK01]
- Clustering [GMMO00]
- Inversion counting [AJKS02]
- Triangle counting [BKS02]

Example 1: Frequency Moments

$$F_k(a_1, \ldots, a_n) = \sum_{j=1}^{m} f_j^k$$

where, $a_1, \ldots, a_n \in [m]$, and $f_j = |\{i \in [n] \mid a_i = j\}|$

- $F_0 = \#$ of distinct data items
- $F_1 = n$
- $F_k, k \geq 2 =$ measure of k-wise collision probability

Theorem [Alon, Matias, Szegedy '96]

- F_0, F_1, F_2 can be approximated in logarithmic space
- F_k, for $k > 5$, needs polynomial space

Theorem 1 [This paper]

F_k, for any $k > 2$, requires polynomial space

Example 2: L_p Distance

$$\mathrm{L}_p(\vec{a},\vec{b}) \stackrel{\text{def}}{=} ||\vec{a}-\vec{b}||_p$$

where the entries of $\vec{a},\vec{b} \in [m]^n$ are given in arbitrary order

Theorem

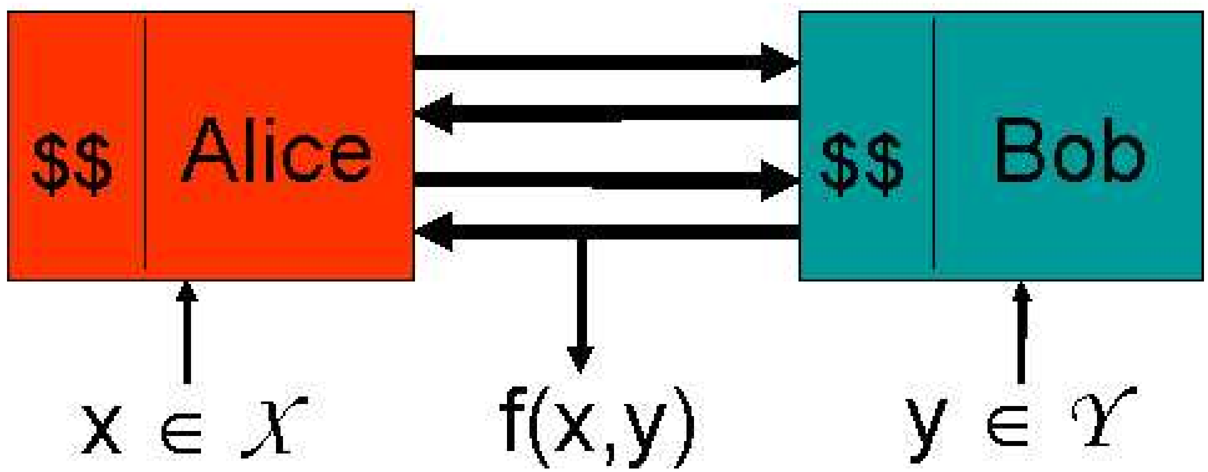
- L_p, for any $0 < p \leq 2$, can be approximated in poly-logarithmic space [Feigenbaum et al. '99, Fong, Strauss '00, Indyk '00]
- Any one pass algorithm for L_p, for $p > 2$, requires polynomial space [Saks, Sun '02]

Theorem 2 [This paper]
L_p, for $p > 2$, requires polynomial space, even for a constant number of passes over the input.

Communication Complexity

[Yao '79]

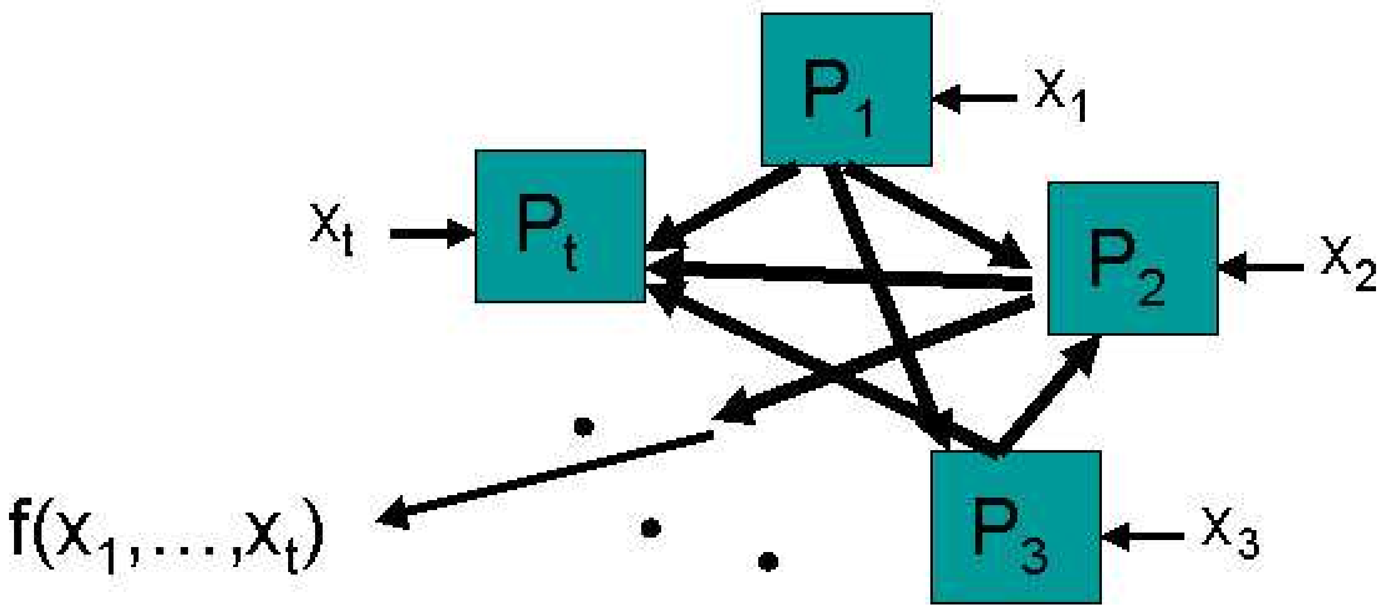
$$f: \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Z}$$



$R_\delta(f) =$ Randomized communication complexity of f with error δ

Multi-Party Communication Complexity
("number in the hand")

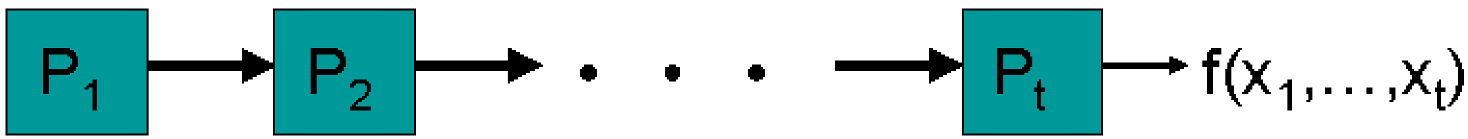
$$f: \mathcal{X}_1 \times \ldots \times \mathcal{X}_t \rightarrow \mathcal{Z}$$



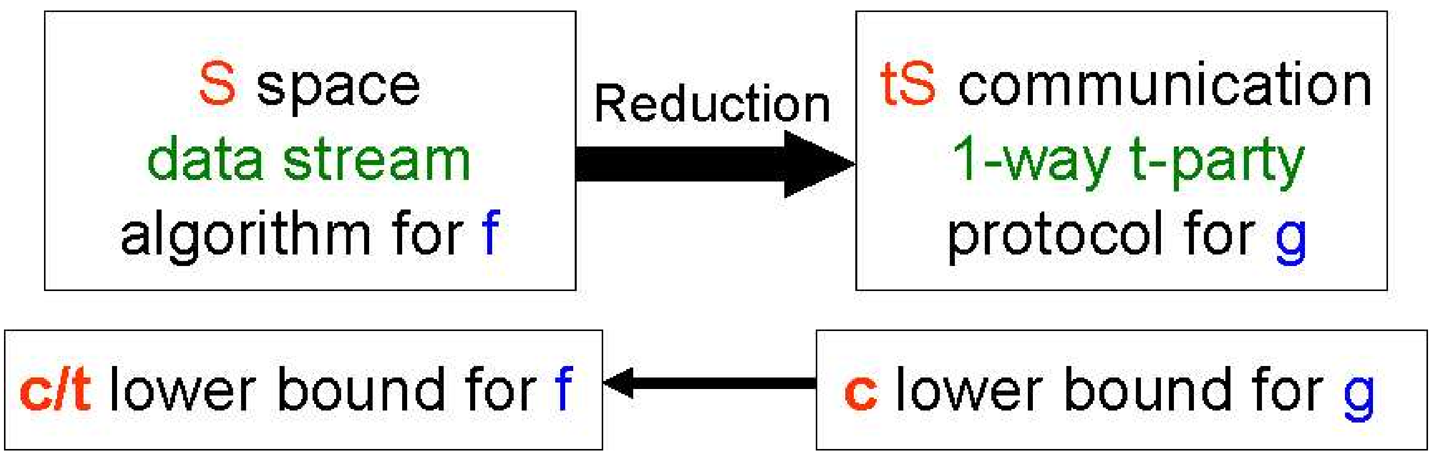
Applications of Communication Complexity

- Circuit depth [Karchmer,Wigderson '88, Razborov '90, Raz,Wigderson '90]
- Time-space tradeoffs
  [Chandra, Furst, Lipton '83, Alon, Maass '86, Babai, Nisan, Szegedy '89]
- VLSI theory [Lengauer '90]
- Decision tree complexity [Nisan '93]
- Data structure complexity [Miltersen '95, Miltersen et al. '95]
- Pseudorandom generators for logspace
  [Babai, Nisan, Szegedy '89, Impagliazzo, Nisan, Wigderson '94]
- Computational economics
  [Nisan, Segal '01, Deng, Papadimitriou, Safra '02]
- Data stream space complexity
  [Alon, Matias, Szegedy '96, Sask, Sun '02]

One-Way Communication Complexity



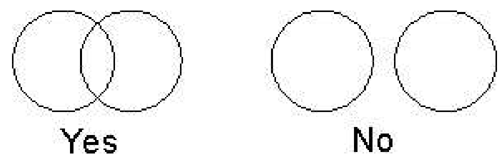
Easy reduction:



Example 1: Set-Disjointness

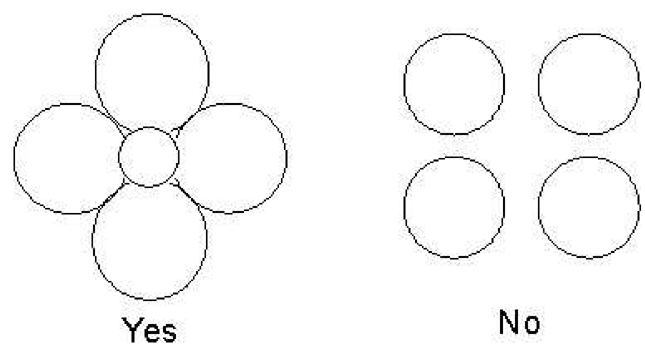
Set-disjointness

Input: $S, T \subseteq [n]$, $\quad \textsc{Disj}_{n,2}(S,T) = 1$ iff $S \cap T \neq \emptyset$.



Multi-party set-disjointness

Input: $S_1, \ldots, S_t \subseteq [n]$ are either pairwise disjoint or mutually intersecting, $\textsc{Disj}_{n,t}(S_1, \ldots, S_t) = 1$ iff $\bigcap_i S_i \neq \emptyset$.



Example 1: Set-Disjointness (cont.)

Theorem

$R_\delta(\textsc{Disj}_{n,2}) = \Omega(n)$ [Kalyanasundaram, Schnitger '87, Razborov '90]

$R_\delta(\textsc{Disj}_{n,t}) = \Omega(n/t^4)$ [Alon, Matias, Szegedy '96]

$R_\delta^{\text{sim}}(\textsc{Disj}_{n,t}) = \Theta(n/t)$ [B, Jayram, Kumar, Sivakumar '02]

Theorem 3 [This paper]

- $R_\delta(\textsc{Disj}_{n,t}) = \Omega(n/t^2)$
- $R_\delta^{\text{1-way}}(\textsc{Disj}_{n,t}) = \Omega(n/t^{1+\epsilon})$, for any $\epsilon > 0$

Using the reduction from $\textsc{Disj}_{n,n^{1/k}}$ to F_k [AMS96]:

Corollary

F_k requires $n^{1-2/k}$ space in the data stream model

Example 2: L_∞ Promise Problem

$\vec{a}, \vec{b}$ are two vectors in $[m]^n$ satisfying one of the following:

- $\mathrm{L}_\infty(\vec{a}, \vec{b}) \leq 1$ (YES instance)
- $\mathrm{L}_\infty(\vec{a}, \vec{b}) \geq m$ (NO instance)

Theorem [Saks, Sun '02]
$R_\delta^{\text{1-way}}(\mathrm{L}_\infty) = \Theta(n/m^2)$

Theorem 4 [This paper]
$R_\delta(\mathrm{L}_\infty) = \Omega(n/m^2)$

Corollary
Estimating L_p to within n^ϵ requires $n^{1-4\epsilon-2/p}$ space in the data stream model, even for a constant number of passes over the input.

Outline of the Lower Bound Technique

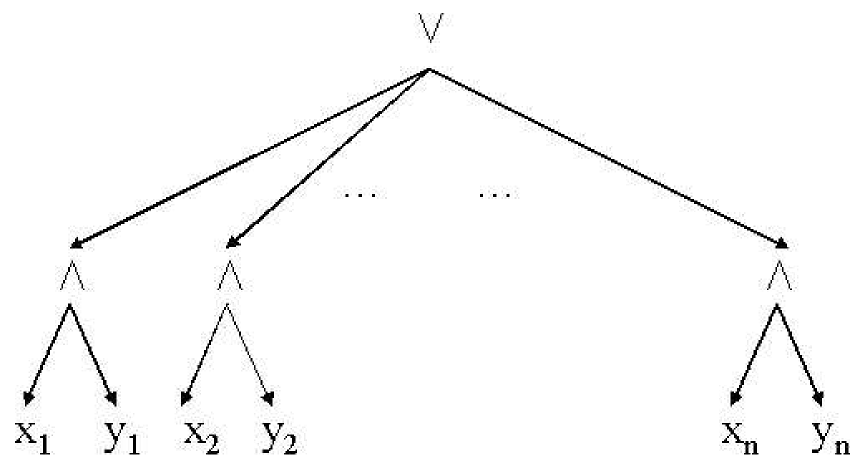
1. Generalization of information complexity
2. A direct sum theorem for information complexity
3. Statistical lower bounds on information complexity of “primitive” functions

The Direct Sum Question

Example

$$\mathrm{DISJ}_{n,2}(\vec{x},\vec{y}) \stackrel{\text{def}}{=} \bigvee_{i=1}^{n} (x_i \wedge y_i)$$

where $\vec{x}, \vec{y} \in \{0,1\}^n$ are the characteristic vectors of the two sets

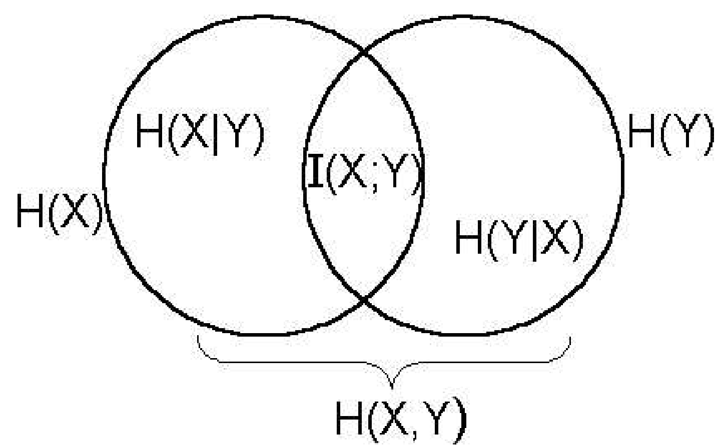


Is a protocol for $\mathrm{DISJ}_{n,2}$ the “sum” of n protocols for AND?

- Cannot use communication length
- Use information theory

Entropy and Mutual Information

- $H(X)$ – the amount of “uncertainty” in X
- $H(X \mid Y)$ – the amount of uncertainty left in X after knowing Y
- $I(X;\, Y) \stackrel{\text{def}}{=} H(X) - H(X \mid Y) \;=\; H(Y) - H(Y \mid X)$



Proposition [Subadditivity of entropy]
$H(X,Y) \;\le\; H(X) + H(Y)$. Equality iff X, Y are independent.

Information Complexity

[Chakrabarti, Shi, Wirth, Yao '01, Ablayev '93, Saks, Sun '02]

$(X,Y) \sim \mu$ – a distribution over inputs of f

Definition [Information Complexity]

$$IC_{\mu,\delta}(f) \stackrel{\text{def}}{=} \min_{\Pi:\ \text{error}(\Pi)\leq\delta} I\left(X,Y;\ \Pi(X,Y)\right)$$

"Amount of information a protocol that computes f has to reveal about its inputs"

Proposition For every μ, $\quad IC_{\mu,\delta}(f) \leq R_\delta(f)$

Conditional Information Complexity

μ is product if X, Y are independent.

- For direct sum, we need μ to be product.
- To create hard instances for some problems (e.g., set-disjointness), we need μ to be non-product.

Express μ as convex combination of product distributions $\sum_d \lambda_d \mu_d$

- Let D be a random variable such that $\Pr[D = d] = \lambda_d$
- Conditioned on $\{D = d\}$, X, Y are independent

Definition [Conditional Information Complexity]

$$IC_{\mu,\delta}(f \mid D) \stackrel{\text{def}}{=} \min_{\Pi \,:\, \text{error}(\Pi) \leq \delta} I\left(X, Y;\ \Pi(X, Y) \mid D\right)$$

Proposition For every μ, D, $\quad IC_{\mu,\delta}(f \mid D) \ \leq\ IC_{\mu,\delta}(f)$

Input Distribution for Set-Disjointness

$(X, Y) \sim \nu$: a distribution on $\{0, 1\}$ defined as follows:

- $D \in_R \{A, B\}$
- If $D = A$, let $X \in_R \{0, 1\}$ and $Y = 0$
- If $D = B$, let $X = 0$ and $Y \in_R \{0, 1\}$

$$\mu \stackrel{\text{def}}{=} \nu^n$$

- μ is non-product
- Conditioned on $\vec{D} = D^n$, μ is product
- μ is concentrated on 0's of $\textsc{Disj}_{n,2}$

Direct Sum for Information Complexity

Theorem $IC_{\mu,\delta}(\textsc{Disj}_{n,2} \mid \vec{D}) \ \geq \ n \cdot IC_{\nu,\delta}(\textsc{And} \mid D)$

Proof. Let Π be a protocol for $\textsc{Disj}_{n,2}$.

Let $(\vec{X}, \vec{Y}) \sim \mu$.

1. Decomposition step:

$$I\left(\vec{X}, \vec{Y};\ \Pi(\vec{X}, \vec{Y}) \mid \vec{D}\right) \ \geq \ \sum_j I\left(X_j, Y_j;\ \Pi(\vec{X}, \vec{Y}) \mid \vec{D}\right)$$

2. Reduction step:

$$I\left(X_j, Y_j;\ \Pi(\vec{X}, \vec{Y}) \mid \vec{D}\right) \ \geq \ IC_{\nu,\delta}(\textsc{And} \mid D)$$

Proof of Decomposition Step

$$
\begin{aligned}
& I\left(\vec{X},\vec{Y};\ \Pi(\vec{X},\vec{Y}) \mid \vec{D}\right) \\
= \quad & H\left(\vec{X},\vec{Y} \mid \vec{D}\right) - H\left(\vec{X},\vec{Y} \mid \Pi(\vec{X},\vec{Y}),\vec{D}\right) \\
\geq \quad & \textstyle\sum_j H\left(X_j,Y_j \mid \vec{D}\right) - \sum_j H\left(X_j,Y_j \mid \Pi(\vec{X},\vec{Y}),\vec{D}\right)
\end{aligned}
$$

(by independence of $\{X_j, Y_j\}$,

and subadditivity of entropy)

$$
= \quad \textstyle\sum_j I\left(X_j,Y_j;\ \Pi(\vec{X},\vec{Y}) \mid \vec{D}\right)
$$

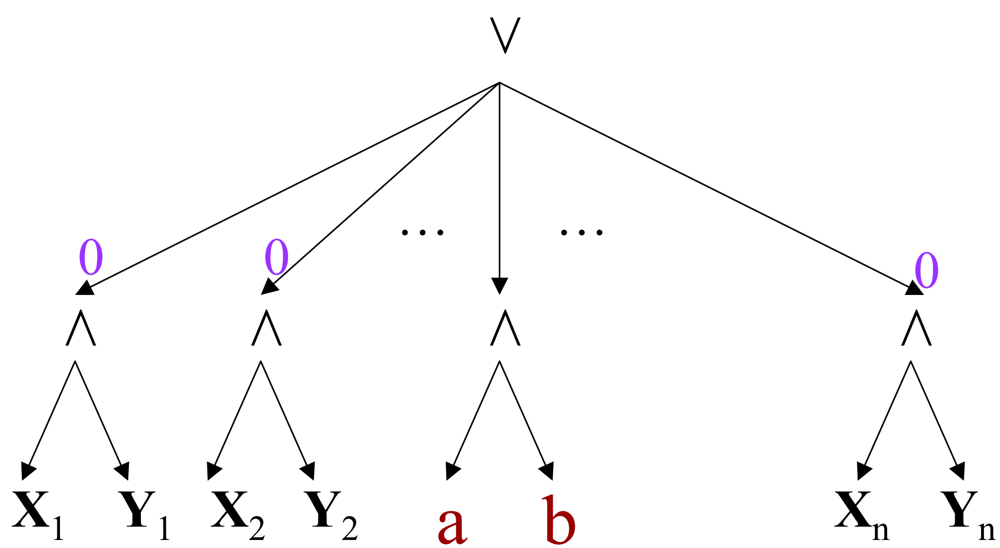
Proof of Reduction Step

$$I\left(X_j, Y_j;\ \Pi(\vec{X}, \vec{Y}) \mid \vec{D}\right)\ \geq\ IC_{\nu,\delta}(\textsc{And} \mid D)$$

1. Expand $I\left(X_j, Y_j;\ \Pi(\vec{X}, \vec{Y}) \mid \vec{D}\right)$ over all $\vec{D}_{-j} = \vec{d}_{-j}$:

   $\sum_{\vec{d}_{-j}} \Pr\left(\vec{D}_{-j} = \vec{d}_{-j}\right) \cdot I\left(X_j, Y_j;\ \Pi(\vec{X}, \vec{Y}) \mid D_j, \vec{D}_{-j} = \vec{d}_{-j}\right)$

2. Create a protocol for computing $\textsc{And}(a, b)$



Lower Bound on $IC_{\nu,\delta}(\textsc{And} \mid D)$

Recall the distribution ν:

- $D \in_R \{A, B\}$
- If $D = A$, let $X \in_R \{0,1\}$ and $Y = 0$
- If $D = B$, let $X = 0$ and $Y \in_R \{0,1\}$

Suppose P is a protocol for $\textsc{And}$. Then,

$$I\left(X, Y;\ P(X,Y) \mid D\right)$$

$$= \tfrac{1}{2} \cdot \left[I\left(X;\ P(X,Y) \mid D = A\right) + I\left(Y;\ P(X,Y) \mid D = B\right)\right]$$

$$= \tfrac{1}{2} \cdot \left[I\left(U;\ P(U,0)\right) + I\left(U;\ P(0,U)\right)\right] \qquad (U \in_R \{0,1\})$$

Jensen-Shannon Divergence

- Q_0, Q_1: two distributions
- $U \in_R \{0,1\}$
- Define $\quad JS\,(Q_0, Q_1) \stackrel{\text{def}}{=} I\,(U;\ Q_U)$

$$I\,(X,Y;\ P(X,Y) \mid D)$$

$$= \ \tfrac{1}{2} \cdot [I\,(U;\ P(U,0)) + I\,(U;\ P(0,U))]$$

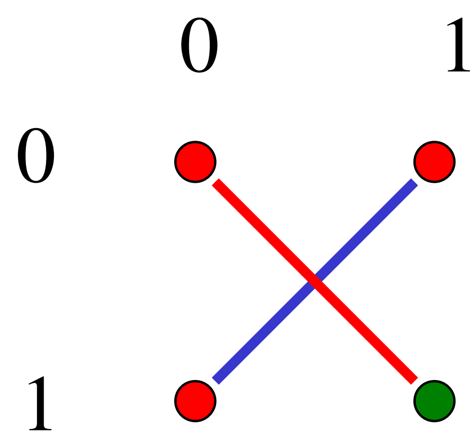
$$= \ \tfrac{1}{2} \cdot [JS\,(P(0,0), P(1,0)) \ + \ JS\,(P(0,0), P(0,1))]$$

$$\stackrel{?}{\geq} \ \tfrac{1}{2} \cdot JS\,(P(1,0), P(0,1))$$

A Point to Ponder

If P computes $\textsc{And}$, why should $P(0,1)$ be far from $P(1,0)$?

- $\textsc{And}$ is 0 on both of these inputs
- The large distance is on the other diagonal!



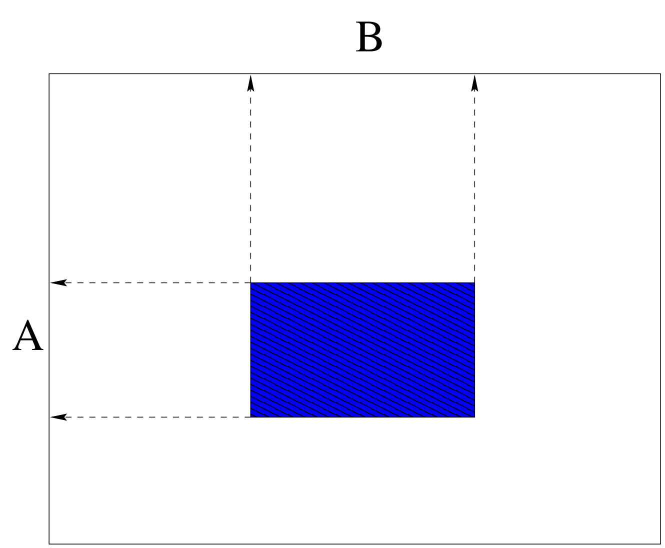
Rectangle Property of Communication Complexity

Let Π be a deterministic protocol.

Fix a transcript τ

$\Pi^{-1}(\tau) = \{(x, y) : \Pi(x, y) = \tau\}$.

Then, $\Pi^{-1}(\tau)$ a combinatorial rectangle:



A Probabilistic Analog

Let Π be a randomized protocol.

Fix a transcript τ

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

$$\Pr[\Pi(x,y) = \tau] = p_\tau(x) \cdot q_\tau(y), \qquad \forall x, y$$

Hellinger Distance

Let P and Q be two probability distributions

$$h^2\left(P,Q\right) = 1 - \sum_\alpha \sqrt{P(\alpha)Q(\alpha)}$$

- $JS\left(P,Q\right) \geq h^2\left(P,Q\right)$
- $h\left(\cdot,\cdot\right)$ is a metric

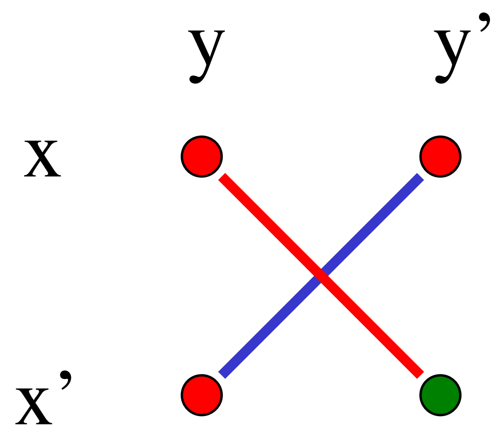
$$I\left(X,Y;\ P(X,Y) \mid D\right)$$

$$= \ \frac{1}{2} \cdot \left[JS\left(P(0,0),P(0,1)\right) \ + \ JS\left(P(0,0),P(1,0)\right)\right]$$

$$\geq \ \frac{1}{2} \cdot \left[h^2\left(P(0,0),P(0,1)\right) + h^2\left(P(0,0),P(1,0)\right)\right]$$

$$\geq \ \frac{1}{4} \cdot h^2\left(P(0,1),P(1,0)\right)$$

A Cut & Paste Lemma



$$
\begin{aligned}
&1 - h^2\left(P(x,y), P(x',y')\right)\\
&= \sum_\tau \sqrt{\Pr\left(P(x,y)=\tau\right)\cdot \Pr\left(P(x',y')=\tau\right)}\\
&= \sum_\tau \sqrt{p_\tau(x)\cdot q_\tau(y)\cdot p_\tau(x')\cdot q_\tau(y')}\\
&= \sum_\tau \sqrt{\Pr\left(P(x,y')=\tau\right)\cdot \Pr\left(P(x',y)=\tau\right)}\\
&= 1 - h^2\left(P(x,y'), P(x',y)\right)
\end{aligned}
$$

Summary of Proof

$I\,(X,Y;\ P(X,Y) \mid D)$

$\geq\ \frac{1}{4}\, h^2\,(P(0,1), P(1,0))$

$=\ \frac{1}{4}\, h^2\,(P(0,0), P(1,1))$ (Cut & paste)

$\geq\ \frac{1}{4}\,(1 - 2\sqrt{\delta})$ (Correctness of P)

Therefore:

$$\begin{aligned} R_\delta(\textsc{Disj}_{n,2}) &\geq IC_{\mu,\delta}(\textsc{Disj}_{n,2} \mid \vec{D}) \geq n \cdot IC_\nu(\textsc{And} \mid D) \\ &\geq n \cdot \frac{1}{4}\,(1 - 2\sqrt{\delta}) \end{aligned}$$

t-Party Set-Disjointness

- Same direct sum argument
- Need a lower bound for t-bit $\textsc{And}$.

$\Omega(1/t^2)$ bound for general communication:

- ν: pick $D \in_R [t]$ and then set $\vec{X} \in_R \{e_D, \vec{0}\}$
- t repeated applications of triangle inequality + cut & paste.

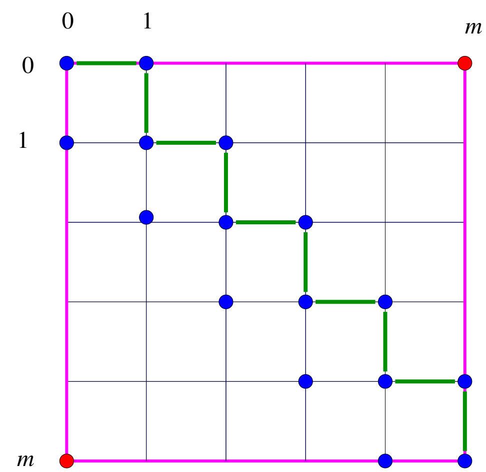
$\Omega(1/t^{1+\epsilon})$ bound for one-way communication:

- Exploits the Markovian structure of one-way protocols.
- Uses generalizations of the Hellinger distance – Rényi divergences

L_∞ Promise Problem

- Same direct sum argument
- Need a lower bound for the difference problem: for $a, b \in [m]$, decide whether $|a-b| \leq 1$ or $|a-b| \geq m$.

$\Omega(1/m^2)$ lower bound:



A Pythagorean lemma:

$$h^2\left(P(0,0), P(m,m)\right) \geq \frac{1}{2}[h^2\left(P(0,0), P(0,m)\right) + h^2\left(P(m,0), P(m,m)\right)]$$

Conclusions

A powerful lower bound methodology in communication complexity

Method gives strong results even for promise problems

- particularly useful for data stream lower bounds

Several novel ideas introduced:

- Conditional information complexity
- Reduction to proving lower bounds for “simple” functions
- Crisp connections between statistical distance measures and communication complexity

Subsequent Work

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

Thank You!