Data stream lower bounds

Ravi Kumar
Yahoo! Research
ravikumar@yahoo-inc.com

Lecture 1

1a. CC lower bounds for set-disjointness [BJKS02]

Information cost and information complexity
Direct sum theorem
Bounds for one-bit problems

Two-player set-disjointness

- Universe = $[n] = \{1, 2, \ldots, n\}$
- Alice has $x \subseteq [n]$
- Bob has $y \subseteq [n]$
- YES: $x \cap y \neq \emptyset$
- NO: $x \cap y = \emptyset$

$DISJ(x, y) = \vee_{i=1,n} (x_i \wedge y_i)$

Alice x ⟺ Bob y

Number of bits exchanged to correctly compute DISJ(x,y)

d-cc(DISJ) $\geq \Omega(n)$

- Reduce from Equality (EQ)
  - Alice has $x \in \{0, 1\}^n$ Bob has $y \in \{0, 1\}^n$
  - YES if $x = y$ and NO if $x \neq y$
  - d-cc(EQ) $\geq \Omega(n)$
- Given instance of EQ on n/2 bits
  - Create $x' = x \cdot \neg x$ and $y' = y \cdot \neg y$
  - Run the c-bit protocol for DISJ on n bits
  - If $x = y$ then $x' \cap y' = \emptyset$
  - If $x \neq y$ then $x' \cap y' \neq \emptyset$
  - EQ on n/2 bits can be solved by a c-bit protocol

Randomized protocols

- The randomized complexity of DISJ
- Is there a δ-error protocol Π for DISJ such that
  - $\forall x, y \Pr[\Pi \text{ computes } DISJ(x, y)] \geq 1 - \delta$
  - $\max_{x, y, \$} \{ \text{transcript length of } \Pi(x, y) \} = o(n)$

Alice x, \$ ⟺ Bob y, \$

- Previous reduction doesn't give anything

Goal of this lecture

- Prove an $\Omega(n)$ lower bound for the randomized communication complexity of two-player DISJ [KS87, R90]

Intuition

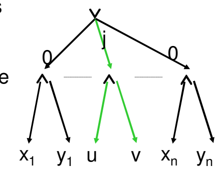
Why should r-cc(DISJ) be high?

$\mathrm{DISJ}(x, y) = \vee_{i=1,n} (x_i \wedge y_i) = \vee_{i=1,n} \mathrm{AND}(x_i, y_i)$

Have to look at these n one-bit $\wedge$-s before determining the output is 0

Ie, any correct protocol should implicitly solve n-instances of these one-bit $\wedge$-s

Ie, the transcript should contain "information" about each of the n pairs of inputs



Basic notation

- x = input to Alice, y = input to Bob, Π = protocol
- $\Pi(x, y)$ = message transcript
  - Distribution if Π is randomized
- Π is δ-error for f if $\forall x, y$

  $\Pr_{\$} [\Pi(x, y) = f(x, y)] \geq 1 - \delta$
- Communication cost = $\max_{x, y, \$} |\Pi(x, y)|$
- $R_\delta(f)$ = R(f) = communication cost of the best δ-error protocol for f

(Think of δ as small constant. We will drop δ hereafter.)

Overview of the proof

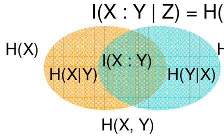
- Move from communication complexity to information complexity
- Prove a direct-sum theorem for information complexity of DISJ in terms of one-bit AND
- Prove a lower bound for information complexity of one-bit AND

Quick recap

- $X \sim \mu$, Entropy $H(X) = \sum \mu(\omega) \log 1/\mu(\omega)$
- Conditional entropy $H(X \mid Y) = E[H(X \mid Y{=}y)]$
- Mutual information

  $I(X : Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$
- Conditional mutual information

  $I(X : Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z)$



Sub-additivity

$H(X, Y \mid Z) \leq H(X \mid Z) + H(Y \mid Z)$

and equality iff $X \perp Y$ (indep.)

Information complexity

Measure of how much information a transcript reveals about its inputs

- If $(X, Y) \sim \mu$ is a distribution on inputs, information cost of Π wrt μ is

  $I(X, Y : \Pi(X, Y))$
- Information complexity of f wrt μ, denoted $IC_\mu(f)$, is the minimum information cost of a protocol for f wrt μ
- [CSWY01, A93, SS02]

Information cost vs communication

Let Π be a protocol for f. Then for any distribution μ

$$I(X, Y : \Pi(X, Y))$$
$$= H(\Pi(X, Y)) - H(\Pi(X, Y) \mid X, Y)$$
$$\le H(\Pi(X, Y))$$
$$\le \max_{X, Y} |\Pi(X, Y)|$$

Corollary: $IC_\mu(f) \le R(f)$

Choosing the distribution

- Hope is to show $IC_\mu(DISJ) \ge n \cdot IC_\nu(AND)$ by choosing an input distribution $(X, Y) \sim \mu$ carefully
- Product distributions ($X \perp Y$) are easier for direct sums
- But, we cannot hope to get $\Omega(n)$ bound if $X \perp Y$
- We have to use a non-product distribution
- We have to generalize the notion of information complexity to account for this

Conditional information complexity

- Key idea: Make X and Y conditionally independent
  - Define a random variable D such that $X \perp Y \mid D$
- If $((X, Y), D) \sim \mu$, then the conditional information cost of Π wrt μ is

$$I(X, Y : \Pi(X, Y) \mid D)$$

- Conditional information complexity of f wrt μ, denoted $IC_\mu(f \mid D)$, is the minimum conditional information cost of a protocol for f wrt μ
- Exercise: Show $IC_\mu(f \mid D) \le R(f)$
- Bonus: Show that $IC_\mu(f \mid D) \le IC_\mu(f)$

The distribution for DISJ

"Magic" random variable $M \in_R \{alice, bob\}$

- If M = alice, then $U = 0$, $V \in_R \{0, 1\}$
- If M = bob, then $U \in_R \{0, 1\}$, $V = 0$

One-bit distribution $\nu = ((U, V), M)$

- Note $U \perp V \mid M$

n-bit distribution on $((X, Y), D) \sim \mu = \nu \times \ldots \times \nu$

- $X \perp Y \mid D$
- Places mass only on the NO instances of DISJ

Direct sum theorem

Theorem: $IC_\mu(DISJ \mid D) \ge n \cdot IC_\nu(AND \mid M)$

Proof steps: Let Π be a protocol for DISJ and let $((X, Y), D) \sim \mu$

1. Decomposition step:

   $IC_\mu(DISJ \mid D) \ge \Sigma_j \; I(X_j, Y_j : \Pi(X, Y) \mid D)$
2. Reduction step: For each j

   $I(X_j, Y_j : \Pi(X, Y) \mid D) \ge IC_\nu(AND \mid M)$

Decomposition step

$$I(X,Y : \Pi(X,Y) \mid D)$$
$$= H(X,Y \mid D) - H(X, Y \mid D, \Pi(X,Y))$$
$$= (\Sigma_j H(X_j,Y_j \mid D)) - H(X, Y \mid D, \Pi(X,Y))$$
$$\ge \Sigma_j H(X_j,Y_j \mid D) - \Sigma_j H(X_j,Y_j \mid D, \Pi(X,Y))$$
$$= \Sigma_j I(X_j,Y_j : \Pi(X,Y) \mid D)$$
$$= \Sigma_j I(X_j,Y_j : \Pi(X,Y) \mid D_j, D_{-j})$$
$$= E_{D_{-j} = \Delta} \Sigma_j I(X_j,Y_j : \Pi(X,Y) \mid D_j, D_{-j} = \Delta)$$

Reduction step

Create a two-party protocol P for AND from Π

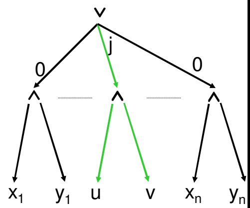
Given u, v, the protocol $P = \Pi_{j, \Delta}$ works as follows:

Alice, Bob create X,Y with $X_j = u$, $Y_j = v$, filling the other X_i's and Y_i's by using Δ

- Alice and Bob can fill in X_{-j} and Y_{-j} without any communication

Then they run $\Pi(X,Y)$ and output whatever Π outputs

- Since μ places mass only on NO instances of DISJ, AND is computed correctly



Reduction step, contd.

- Exercise: Show that

$(U, V, M, \Pi_{j, \Delta}) \equiv (X_j, Y_j, D_j, \Pi(X, Y) \mid D_{-j} = \Delta)$

- Thus each term in summation is conditional information cost wrt ν of a protocol P for AND, ie, is at least IC_ν(AND | M)

Lower bounding IC_ν(AND | M)

Assume P computes AND

Information cost of P wrt ν

$= I(U, V : P(U, V) \mid M)$

$= \frac{1}{2} (I(U, V : P(U, V) \mid M=alice)$

$+ I(U, V : P(U, V) \mid M=bob))$

$= \frac{1}{2} (I(Z : P(0, Z)) + I(Z : P(Z, 0)))$ $Z \in_R \{0, 1\}$

What is this quantity?

- Intuitively, if P(0, 0) and P(0, 1) are very different, then I(Z : P(0, Z)) must be large
- Conversely, if P(0, 0) and P(0, 1) are very similar, then I(Z : P(0, Z)) must be small

Thus, I(Z : P(0, Z)) measures some distance between the distributions of P(0, 0) and P(0, 1)

Formalizing …

- The Hellinger distance between two distributions P and Q

$h^2(P, Q) = 1 - \sum_\omega (P(\omega) Q(\omega))^{1/2}$

$= \sum_\omega (P(\omega) + Q(\omega))/2 - (P(\omega) Q(\omega))^{1/2}$

- Exercise: Show h is a metric

- Theorem: $I(Z : P(0, Z)) \geq h^2(P_{00}, P_{01})$ and $I(Z : P(Z, 0)) \geq h^2(P_{00}, P_{10})$

LB for IC_ν(AND | M), contd.

$I(U, V : P(U, V) \mid M)$

$= \frac{1}{2} (I(Z : P(0, Z)) + I(Z : P(Z, 0)))$

$Z \in_R \{0, 1\}$

$\geq \frac{1}{2} (h^2(P_{00}, P_{01}) + h^2(P_{00}, P_{10}))$

$\geq \frac{1}{4} (h(P_{00}, P_{01}) + h(P_{00}, P_{10}))^2$ [C-S]

$\geq \frac{1}{4} h^2(P_{01}, P_{10})$ [Triangle]

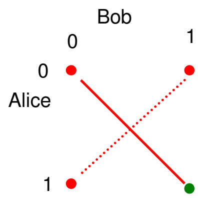
A point to ponder

$I(U, V : P(U, V) \mid M) \geq \frac{1}{4} h^2(P_{01}, P_{10})$

If P computes AND correctly, why should P_{01} be far from P_{10}

AND is 0 on both these inputs

The large distance is between P_{11} and P_{00}, P_{01}, P_{10}

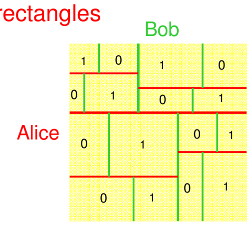


Aug 20-23, 2007 MADALGO, Aarhus 25

Rectangular property of d-cc

A deterministic communication protocol partitions the input matrix into monochromatic rectangles

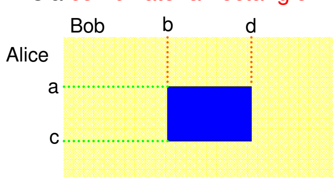
Alice and Bob send one bit in each round



Aug 20-23, 2007 MADALGO, Aarhus 26

Fundamental theorem of d-cc

P is a deterministic communication protocol, then the set of inputs with same transcript is a combinatorial rectangle



$P_{ab} = \tau = P_{cd}$

$\Rightarrow P_{ad} = \tau = P_{cb}$

Aug 20-23, 2007 MADALGO, Aarhus 27

Fundamental theorem of r-cc

P is a randomized communication protocol and T be the set of all transcripts

$\exists$ p: $T \times X \rightarrow \{0, 1\}$, q: $T \times Y \rightarrow \{0, 1\}$ such that

$$\Pr[P_{xy} = \tau] = p(\tau, x) \cdot q(\tau, y), \forall x, y, \tau$$

Exercise: Prove this. Hint: consider extended input = input + private random coins and apply the rectangular property

Aug 20-23, 2007 MADALGO, Aarhus 28

Cut-and-paste (X-) lemma

Lemma: $h^2(P_{ab}, P_{cd}) = h^2(P_{ad}, P_{cb})$

Proof:

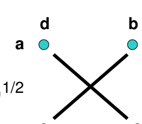
$1 - h^2(P_{ab}, P_{cd})$

$= \Sigma_\tau (\Pr[P_{ab} = \tau] \Pr[P_{cd} = \tau])^{1/2}$

$= \Sigma_\tau (p(\tau, a)\, q(\tau, b)\, p(\tau, c)\, q(\tau, d))^{1/2}$

$= \Sigma_\tau (\Pr[P_{ad} = \tau] \Pr[P_{cb} = \tau])^{1/2}$

$= 1 - h^2(P_{ad}, P_{cb})$



Aug 20-23, 2007 MADALGO, Aarhus 29

LB for IC_ν(AND | M), contd.

$I(U, V : P(U, V) \mid M) \geq \frac{1}{4} h^2(P_{01}, P_{10})$

$= \frac{1}{4} h^2(P_{00}, P_{11})$

How to relate $h(P_{00}, P_{11})$ to the error of P?

Via the total variation distance

$V(P, Q) = \frac{1}{2} \Sigma_\omega |P(\omega) - Q(\omega)|$

$= \max_{\Omega' \subseteq \Omega} |P(\Omega') - Q(\Omega')|$

- Bonus: Show $V(P, Q) \leq h(P, Q)\,(2 - h^2(P, Q))^{1/2}$

Aug 20-23, 2007 MADALGO, Aarhus 30

Variational distance

Lemma: If P is a δ-error protocol for AND, then $V(P_{00}, P_{11}) \geq 1 - 2\delta$

Proof: Let T be the set of transcripts where P outputs 0 as the answer

$P_{00}(T) \geq 1 - \delta$ and $P_{11}(T) \leq \delta$

$V(P_{00}, P_{11}) \geq P_{00}(T) - P_{01}(T) \geq 1 - 2\delta$

Corollary: $h^2(P_{00}, P_{11}) \geq 1 - 2\sqrt{\delta}$

Putting all together

Lower bound for $IC_\nu(AND \mid M)$

$I(U, V : P(U, V) \mid M) \geq \frac{1}{4} h^2(P_{01}, P_{10})$

$= \frac{1}{4} h^2(P_{00}, P_{11})$

$= \frac{1}{4} (1 - 2\sqrt{\delta})$

Combining with direct sum theorem

$R(DISJ) \geq IC_\mu(DISJ \mid D)$

$\geq n \cdot IC_\nu(AND \mid M)$

$\geq (n/4)(1 - 2\sqrt{\delta})$

1b. One-way bounds for set-disjointness [BJKS02]

Stronger bounds for set disjointness

One-way protocols

- Alice looks at her input x sends a message to Bob
- Bob looks at this message and his input y and outputs the answer
- It suffices to prove lower bounds in the one-way communication model

Alice x → Bob y

Fundamental theorem of 1-way cc

P is a randomized one-way communication protocol and T be the set of all transcripts

T_A is Alice's portion, T_B is Bob's portion

For each input x to Alice, y to Bob

$\exists\, p_x: T_A \to \{0, 1\}$, $M_y: T_A \times T_B \to \{0, 1\}$ such that for all transcripts (τ_A, τ_B)

$\Pr[P_{xy} = (\tau_A, \tau_B)] = p_x(\tau_A) \cdot M_y(\tau_A, \tau_B)$

p_x is a distribution, M_y is a transition matrix

Proof of this characterization

$\Pr[P_{x,y} = (\tau_A, \tau_B)]$

$= \Pr[A_x = \tau_A] \cdot \Pr[B_{y, \tau_A} = \tau_B \mid A_x = \tau_A]$

$= \Pr[A_x = \tau_A] \cdot \Pr[B_{y, \tau_A} = \tau_B]$

$= p_x(\tau_A) \cdot M_y(\tau_A, \tau_B)$

where the τ_A-th row of M_y describes the distribution of B_{y, τ_A}

Denote $(p^\circ M)(i, j) = p(i) \cdot M(i,j)$

Revisit one-bit proof for AND

$I(U, V : P(U, V) \mid M)$
$\geq \frac{1}{2} (h^2(P_{00}, P_{01}) + h^2(P_{00}, P_{10}))$
$\geq (\frac{1}{2}) (\frac{1}{2}) (h(P_{00}, P_{01}) + h(P_{00}, P_{10}))^2$ [C-S]
$\geq (\frac{1}{2}) (\frac{1}{2}) h^2(P_{01}, P_{10})$ [Triangle]
For one-way protocols, we have
$P_{00} = p_0{}^\circ M_0$, $P_{01} = p_0{}^\circ M_1$, $P_{10} = p_1{}^\circ M_1$
Can we get better bounds on
$h^2(p_0{}^\circ M_0, p_0{}^\circ M_1) + h^2(p_0{}^\circ M_0, p_1{}^\circ M_0)$

Improved bound

Lemma: $h^2(p^\circ M, q^\circ N) \leq (1+1/\sqrt{2}) (h^2(p^\circ M, q^\circ M) + h^2(p^\circ M, p^\circ N))$
Proof: Let C_i = i-th row of C, D_i = i-th row of D
$h^2(a^\circ C, b^\circ D) = 1 - \sum_{i \in \Omega, j \in \Gamma} (a_i C_{ij} b_i D_{ij})^{1/2}$
$= 1 - \sum_{i \in \Omega} (a_i b_i)^{1/2} \sum_{j \in \Gamma} (C_{ij}, D_{ij})^{1/2}$
$= 1 - \sum_{i \in \Omega} (a_i b_i)^{1/2} (1 - h^2(C_i, D_i))$
$= h^2(a, b) - \sum_{i \in \Omega} h^2(C_i, D_i) \cdot (a_i b_i)^{1/2}$

Improved bound, contd.

Let $\beta_i = h^2(M_i, N_i) \leq 1$
$h^2(p^\circ M, q^\circ N) \leq (1+1/\sqrt{2}) (h^2(p^\circ M, q^\circ M) + h^2(p^\circ M, p^\circ N)) \Leftrightarrow$
$h^2(p, q) + \sum_i (p_i q_i)^{1/2} \beta_i$
$\leq (1 + 1/\sqrt{2}) (h^2(p, q) + \sum_i p_i \beta_i) \Leftrightarrow$
$\sum_i \beta_i ((p_i q_i)^{1/2} - (1 + 1/\sqrt{2}) p_i)$
$\leq (1/\sqrt{2}) \sum_i ((p_i + q_i)/2 - (p_i q_i)^{1/2})$
Exercise: Prove this point-wise for each i

Using the improved bound

$I(U, V : P(U, V) \mid M)$
$\geq \frac{1}{2} (h^2(P_{00}, P_{01}) + h^2(P_{00}, P_{10}))$
$\geq (\frac{1}{2}) \cdot 0.586 \cdot h^2(P_{01}, P_{10})$ [Improved bound]

This improvement has implications for multi-player protocols
Using general Renyi-divergences, can improve this even more [BJKS02, CKS03]

Lecture 2

Bounds for distinct elements problem
Longest increasing subsequence
- Deterministic upper bounds
- Randomized exact lower bounds
- Deterministic lower bounds

2a. Bounds for distinct elements

Finding distinct elements

- Given $X = x_1, \ldots, x_n$ compute $F_0(X)$, the number of distinct elements in X, in the data stream model
  Assume $x_i \in [m]$
- (ε,δ)-approximation: Output $F'_0(X)$ such that with probability at least $1 - \delta$, $F'_0(X) = (1 \pm \varepsilon) F_0(X)$
- Zeroth frequency moment
- Assume $\log m = O(\log n)$; otherwise hash input
- Sampling needs lots of space
- Without randomization and approximation, this problem is uninteresting

Some previous work

- [FM85]: Assumed ideal hash functions
- [AMS99]: Pairwise independent hashing
  $(2+\varepsilon)$-approximation using $O(\log m)$ space
- [GT01]: Hashing-based
  ε-approximation using $O(1/\varepsilon^2 \log m)$ space
- [BKS03]: Hashing-based, range-summable
  ε-approximation using $O(1/\varepsilon^3 \log m)$ space
- [CDIM02]: Stable distributions
  ε-approximation using $O(1/\varepsilon^2 \log m)$ space

$\Omega(\log m)$ lower bound [AMS]

Reduction from set equality problem

Alice given X, Bob given Y, both m-bit vectors, and the question is if X = Y

- Randomized space bound of $\Omega(\log m)$

Given instance of equality, create $X' = \varphi(X)$, $Y' = \varphi(Y)$ where φ is error-correcting code

- If X = Y, then $F_0(X' \cup Y') = n'$
- If $X \neq Y$, then $F_0(X' \cup Y') \sim 2n'$

One-way $\Omega(1/\varepsilon)$ lower bound

Reduction from set-disjointness with special instances

Alice has bit vector X with $|X| = m/2$, Bob has bit vector Y with $|Y| = \varepsilon m$

- YES: $X \supset Y$
- NO: $X \cap Y = \emptyset$
- One-way lower bound [BJKS]: $\Omega(1/\varepsilon)$

Given disjointness, create $Z = (1, x_1) \ldots (m, x_m)\ (1, y_1) \ldots (m, y_m)$

- YES: If X contains Y, then $F_0(Z) = m/2$
- NO: If X and Y are disjoint, $F_0(Z) = m/2 + \varepsilon m = m/2(1 + 2\varepsilon)$

Gap-Hamming problem [IW]

Let $h(\cdot, \cdot)$ be Hamming distance

Alice given X, Bob given Y, both m-bit vectors

- YES: $h(X, Y) \geq m/2$
- NO: $h(X, Y) \leq m/2 - \sqrt{m}$

Gap-Hamming problem: distinguish the two cases in one-way or general communication model

Gap-Hamming captures F_0

- $Z = (1, x_1) \ldots (m, x_m)\ (1, y_1) \ldots (m, y_m)$
- $F_0(Z) = 2h(X,Y) + (m - h(X, Y)) = m + h(X,Y)$

- YES: if $h(X, Y) \geq m/2$ then $F_0(Z) \geq 3m/2$
- NO: if $h(X, Y) \leq m/2 - \sqrt{m}$ then $F_0(Z) \leq 3m/2 - \sqrt{m} = 3m/2(1 - 1/\sqrt{m})$

In this case, $\varepsilon \sim 1/\sqrt{m}$

Thus, $\Omega((\sqrt{m})^c)$ lower bound for gap-Hamming leads to $\Omega(1/\varepsilon^c)$ lower bound for F_0

Easy $\Omega(\sqrt{m})$ lower bound for gap-Hamming

Reduce from set-disjointness

Randomized lower bound of $\Omega(n)$ [KS, R] for a special input distribution

Universe partitioned into U_1, U_2, $\{i\}$

X = uniform set of size $n/4$ from $U_1 \cup \{i\}$

Y = uniform set of size $n/4$ from $U_2 \cup \{i\}$

- YES: X, Y such that $X \cap Y = \{i\}$
- NO: X, Y such that $X \cap Y = \varnothing$

$h(X, Y) = |X| + |Y| - 2\,|X \cap Y|$ and let $m = n^2$

Given X, Y, replace each 1 by n 1's, each 0 by n 0's to get X' and Y'

- YES: if $X \cap Y \neq \varnothing$, then $h(X',Y') = n^2/2 - 2n$
- NO: if $X \cap Y = \varnothing$ then $h(X',Y') = n^2/2$

One-pass $\Omega(m)$ lower bound for gap-hamming [IW, W]

- [IW, W] showed $\Omega(m)$ lower bound in the one-way model
  - Using VC-dimension and embedding
  - We will show a simpler proof of this result

Reduction from indexing [JKS]

Alice has n-bit vector T with $|T| = n/2$ and Bob has index i; assume $n/2$ is odd

Using public randomness, Alice and Bob pick a random n-bit ± 1 vector r

Alice computes $x = \mathrm{sign}(\langle T, r \rangle)$

Bob computes $y = \mathrm{sign}(r_i)$

Now look at the correlation between random variables x and y

Analyzing the correlation

Let $s = \sum_{j \in T} r_j$

$n/2$ odd implies $\Pr[s < 0] = \Pr[s > 0] = 1/2$

- NO: If i is not in T, then x is independent of y so $\Pr[x = y] = \Pr[\mathrm{sign}(s) = \mathrm{sign}(r_i)] = 1/2$
- YES: If $i \in T$, then let $s = s' + r_i$
  $\Pr[s' = 0] = \eta = c/\sqrt{n}$
  $\Pr[s' < 0] = \Pr[s' > 0] = (1 - \eta)/2$
  $\Pr[x = y] = \Pr[s' = 0] + \Pr[\mathrm{sign}(s') = \mathrm{sign}(r_i) \mid s' \neq 0]$
  $= \eta + (1 - \eta)/2 = (1 + c/\sqrt{n})/2$

Amplifying the gap

- We have random variables x and y with the property that
  - NO: $\Pr[x = y] = 1/2$
  - YES: $\Pr[x = y] = 1/2 + c'/\sqrt{n}$
- Repeat with different independent random vectors $r^1, r^2, \ldots, r^t$ to get t-bit vectors X and Y
  - Chernoff shows that if $t = O(n)$ then whp we have
    - NO: $h(X, Y) \geq (1/2 - c_1)n$
    - YES: $h(X, Y) \leq (1/2 - c_1)n - c_2\sqrt{n}$

A geometric interpretation

Exercise: Normalize the two vectors in Euclidean space. The inner product is either ½ or ½ - $\sqrt{n}$. Show a reduction using Goemans-Williamson random hyperplane approach

2b. Bounds for ordering problems [GJKK07, GG07, EJ07]

LIS problem and deterministic upper bound

Lower bound for exact computation

Deterministic lower bound for approximation

Sortedness of a sequence

- Given sequence $\sigma = \sigma(1), \ldots, \sigma(n)$, each $\sigma(i) \in [m]$, how sorted is σ?
  - Kendall distance
    $\#\{ (i, j) \mid (i < j \text{ and } \sigma(i) > \sigma(j)) \text{ or } (i > j \text{ and } \sigma(i) < \sigma(j)) \}$
  - Edit distance (ED)
    Minimal number of inserts/deletes in σ to make it sorted
  - Longest increasing subsequence (LIS)
    Max subsequence $i_1 \leq \ldots \leq i_k$ such that $\sigma(i_1) \leq \ldots \leq \sigma(i_k)$
    $LIS(\sigma) = n - ED(\sigma)$
  - Transpositions, reversals, …

Longest increasing subsequence

Given sequence $\sigma = \sigma(1), \ldots, \sigma(n)$, each $\sigma(i) \in [m]$, LIS(σ) is the maximum subsequence $i_1 \leq \ldots \leq i_k$ such that $\sigma(i_1) \leq \ldots \leq \sigma(i_k)$

Eg, σ = 3 2 4 6 7 1 8 5

LIS(σ) = 3 2 4 6 7 1 8 5

= 3 4 6 7 8

LIS algorithm: Patience sort

Let $P_\sigma(i)$ = smallest letter $a \in [m]$ such that there is an increasing sequence of length i in σ ending at a

Algorithm is to maintain this table and update as σ "arrives"

- Data stream algorithm
- O(n) space

Patience sort

$P(1) = \ldots = P(n) = \infty$

for j = 1, …, n

  read $\sigma(j)$

  find largest i such that $P(i) < \sigma(j)$

  set $P(i+1) = \sigma(j)$

output largest i such that $P(i) \neq \infty$

2-player protocol for LIS

Alice σ_1 ⇒ Bob σ_2

Theorem: $O(1/\varepsilon \log m + \log n)$ bits to approximate $LIS(\sigma = \sigma_1 \cdot \sigma_2)$ to within $1 \pm \varepsilon$

Proof: Alice runs Patience Sort on σ_1 and computes $k_1 = LIS(\sigma_1)$

She sends $\langle i, P_{\sigma_1}(i) \rangle$ for $i = \{\varepsilon k_1, 2\varepsilon k_1, \ldots, k_1\}$

Bob finds best extension of these sequences by σ_2 and outputs the longest, k_2

$k_2 \geq (1-\varepsilon)$ LIS(σ)

Let LIS = $\pi_1 \cdot \pi_2$, $|\pi_1| = \lambda_1$, $|\pi_2| = \lambda_2$

$\pi_1(\lambda_1) = a < b = \pi_2(1)$

Let λ'_1 be multiple of εk_1 s.t. $\lambda_1 - \varepsilon k_1 \leq \lambda'_1 \leq \lambda_1$

$P_{\sigma_1}(\lambda'_1) = a' \leq \pi_1(\lambda'_1) \leq \pi_1(\lambda_1) = a$

Bob extends this sequence to get

$k_2 \geq \lambda'_1 + \lambda_2$

$\geq \lambda_1 - \varepsilon k_1 + \lambda_2$

$= \text{LIS}(\sigma) - \varepsilon k_1$

$\geq (1 - \varepsilon)\, \text{LIS}(\sigma)$

t-player protocol/algorithm

Alice σ_1 ⇒ Bob σ_2 ⇒ Carol σ_3 ⇒ Tom σ_t

- Players computed $Q_\sigma \approx P_\sigma$
- If k_j is the longest sequence detected by the j-th player, he sends $Q_\sigma(i)$ for $i \in \{\varepsilon/(t-1)\, k_j, 2\varepsilon/(t-1), \ldots, k_j\}$
- Need to make sure $|Q_\sigma|$ remains small (cleanup)
- Communication = $(t/\varepsilon) \log m$
- Input to each player = n/t
- If $t = \sqrt{\varepsilon n}$ then gets a one-pass data stream algorithm with space $O(\sqrt{n})$

Randomized lb for exact LIS

- Reduction from one-bit AND(x, y)
- Alice applies $\sigma'_A(\cdot)$ to x and Bob applies $\sigma'_B(\cdot)$ to y

x	y	$\sigma'_A(x)$	$\sigma'_B(y)$	LIS($\sigma'_A(x) \cdot \sigma'_B(y)$)
0	0	4	1	1
0	1	4	3	1
1	0	2	1	1
1	1	2	3	2

LB for exact LIS, contd.

- Reduction from DISJ
- $\sigma_A(i, x_i) = 4(i-1) + \sigma'_A(x_i)$
- $\sigma_B(i, y_i) = 4(i-1) + \sigma'_B(y_i)$

x	y	$\sigma'_A(x)$	$\sigma'_B(y)$	LIS($\sigma'_A(x) \cdot \sigma'_B(y)$)
0	0	4	1	1
0	1	4	3	1
1	0	2	1	1
1	1	2	3	2

LB for exact LIS, contd.

Let $\sigma = \sigma_A(x) \cdot \sigma_B(y)$

Theorem: If $x \cap y \neq \emptyset$, then LIS(σ) = n+1, else LIS(σ) = n

Proof: If $x \cap y = \emptyset$, then any increasing sequence can have only one element from [4(i-1)+1, 4(i-1)+4]. So, LIS(σ) = n

If $i \in x \cap y \neq \emptyset$, then the following is an (n+1) long increasing subsequence

$\sigma(1), \ldots, \sigma(i), \sigma(n+i), \ldots, \sigma(2n)$

- Exercise: Make σ a permutation. Hint: use 8

Det. lb for approx. LIS [GG07, EJ07]

Goal: Any deterministic algorithm for approximating LIS to within $(1 \pm \varepsilon)$ needs $\Omega(\sqrt{n})$ space

Overview of proof [EJ07]

- Define a primitive function h
- Prove $\Omega(t)$ lb for computing h in the t-player model
  - Probabilistic construction of a fooling set
- Define composite function g = OR of t separate copies of h
- Prove $\Omega(t^2)$ lb for computing g in the t-player model
  - “Direct-sum” theorem for fooling sets
- Reduce g to approximating LIS, set $t = \sqrt{n}$

Fooling sets, recap

- A fooling set $S \subseteq \{0,1\}^n \times \{0,1\}^n$ is such that
  - $\forall (x, y) \in S, f(x, y) = 0$
  - $\forall (x_1, y_1) \neq (x_2, y_2) \in S,$ $f(x_1, y_2) = 1$ or $f(x_2, y_1) = 1$

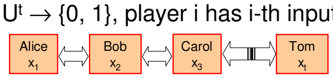
Theorem: Let S be a fooling set. Then d-cc(f) $\geq \log |S|$

EQ: $S = \{ (x, x) : x \in \{0, 1\}^n \}$

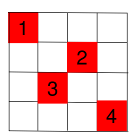
DISJ: $S = \{ (x, \neg x) : x \in \{0, 1\}^n \}$

Basic definitions

- U = Universe of input to each player
- f: $U^t \rightarrow \{0, 1\}$, player i has i-th input



- $M \subseteq U^t$ (think of a matrix with t columns)
- span(M) = $\{ y \in U^t : \forall i \in [t],$ y_i is present in the j-th column of M $\}$



General fooling sets

S is a k-fooling set if
- $f(x) = 1$ for each $x \in S$
- $\forall S' \subseteq S, |S'| = k, \exists y \in \text{span}(S')$ s.t. $f(y) = 0$

Theorem: Let S be a k-fooling set. Then d-cc(f) $\geq \log |S|/k$

Proof: We need a new transcript for every $|S|/k$ subsets.

Primitive function h

$x = x_1 \ldots x_t$, where $x_i \in [t] \cup \{ 0 \}$

x can be viewed as a subset of [t] and non-zero elements are in increasing order

- NO: h(x) = 0 if there are no consecutive non-zero elements in x
  - LIS(x) $\leq t/2 + 1$
- YES: h(x) = 1 if LIS(x) $\geq \alpha t$ for some $\alpha > ½$

h restricted to only above inputs

A large fooling set for h

- Explicitly fooling set seems hard
- Probabilistic method to show one exists

Intuition: Pick random subsets and hope they form a fooling set

Theorem: For large k, there is a k-fooling set for f of size c^t for some $c > 1$

Method: Pick M random subsets of [t] by picking each element with probability p

For p = 1/k, everything works!

Number of good subsets

Subset is good if it has no two consecutive elements of [t]

Lemma: Pr[subset good] ≥ $(1 - p^2)^t$

Proof. g(i) = Pr[subset good for [i]]

$g(i) = (1-p)\, g(i-1) + p(1-p)\, g(i-2)$

Solving, $g(t) = q^t$, $q = ½ ((1-p)+\sqrt{(1-p)(1+3p)})$

Exercise: Show $q > 1 - p^2$

Corollary: E[good subsets] = $(1 - p^2)^t M$

Covering αt by k subsets

$J_1, \ldots, J_k$ be random subsets

Consider $J = J_1 \cup \ldots \cup J_k$

Pr[element ∈ J] = $1 - (1 - p)^k = \gamma$

$E[|J|] = \gamma t$

With $p = 1/k$, $\varepsilon \in (0, ½-1/e)$, $\alpha = (\gamma-\varepsilon) > ½$

$\Pr[|J| < (\gamma-\varepsilon)t] \le \exp(-\varepsilon^2\gamma t/2) = \delta$

Finishing the existence

If (M choose k)·δ < 1 then by union bound every k collection of random subsets cover at least αt elements and with positive probability, there are $(1 - p^2)^t M$ good subsets, which is

$$(1 - p^2)^t (1/\delta)^{1/k} = c^t$$

where

$c = (1 - 1/k^2) \cdot \exp((\varepsilon^2 \gamma)/(2k))$

c > 1 if k is sufficiently large

Corollary: d-cc(h) ≥ Ω(t)

Composite function g

$h_1, \ldots, h_t$ be primitive, with universe of h_i is [(i-1)t + 1, it]

Inputs to g

B = t × t matrix, where i-th row B_i is an input for function h_i

$g(B) = \vee_v h_i(B_i)$

s = sequence formed by concatenating columns of B

Lemma: LIS(s) ≤ 2t

Proof. s can only go right or down

0	1	2	0
5	0	0	8
9	10	0	0
0	14	0	16

Reduction to LIS

Theorem: d-cc(LIS) ≥ d-cc(g)

Proof. If g(B) = 0, then ∀i, $h_i(B_i) = 0$. On each row, when going right, we skip two cells. LIS(s) ≤ 3/2 t

If g(B) = 1, ∃i, $h_i(B_i) = 1$. Go along first column to i-th row, along i-th row, and along last column. LIS(s) ≥ (1+α) t

Suppose i-th player gets i-th column in B. A streaming algorithm to (1+ε)-approximate LIS can distinguish the above gap and hence can yield a protocol for g

Lower bound for d-cc(g)

Theorem: For large k, there is a k^t-fooling set for g of size c^{t^2} for some c > 1

Corollary: d-cc(g) ≥ $\Omega(t^2)$

Corollary: d-cc(LIS) ≥ $\Omega(\sqrt{n})$

Intuition: Build a fooling set for g using fooling set for h

Proof: "Direct-sum" property

- Let F_i be k-fooling set for h_i
- $F = (F_1 \times \ldots \times F_t)$
- $|F| = |F_i|^t = c^{t^2}$
- Argue that F is a k^t-fooling set for g
- If $B = (B_1, \ldots, B_t) \in F$, then $g(B) = \vee h(B_i) = 0$ since F_i is a fooling set for h_i
- Conversely, let $F' \subseteq F$, $|F'| = k^t$
- Let H_i be projection of i-th column of F'
- $\exists$ j, $|H_j| \geq k$

Proof, contd.

- $W \subseteq F'$, $|W| = k$, cover H_j
- Each element of W is a $t \times t$ matrix
- $W = W_1, \ldots, W_k$
- For $B \in$ span(W), i-th column of B is picked from one of the i-th columns of one of W_r's (columns are the input to the players)
- H_j = union of j-th rows of $W_1, \ldots, W_k$
- H_j is a fooling set for $h_j \Rightarrow h_j(B_j) = 1$
- $\Rightarrow g(B) = \vee h_i(B_i) = 1$

Thank you!

ravikumar@yahoo-inc.com