CS 2429 - Foundations of Communication Complexity

Lecturer: Sergey Gorbunov

1 Introduction

In this lecture we will see how to use methods of (conditional) information complexity to prove lower bounds for communication complexity problems. The results presented in these notes are due to the following paper: *"An information statistics approach to data stream and communication complexity"* by Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar and D. Sivakumar [1]. Another exposition of these results can be found in the *"The story of set disjointness"* by A. Chattopadhyay, T. Pitassi [2].

We focus on the two-players set-disjointness problem. This problems is informally regarded as "complete" for the communication complexity since it leads to lower bounds for many other problems. We will show how to:

- use conditional information to argue communication complexity lower bounds,
- study communication complexity of a function by decomposing it into simpler primitives and studying its associated communication complexity,
- prove a lower bound of $\Omega(1)$ of an AND function,
- prove a lower bound of $\Omega(n)$ of a set-disjointness function.

2 Preliminaries

We begin by recalling the set-disjointness (DISJ) problem. For this lecture we will study the two-players case, which can be easily generalized. For $n \in \mathbb{N}$, in DISJ_n problem two players, each holding characteristic vectors x, y of subsets of $[n]$, try to determine if their sets intersect. That is,

$$\mathsf{DISJ}_n(x, y) \text{ outputs 1 if and only if } x \cap y \neq \emptyset$$

For a fixed input pair (x, y), let $\Pi(x, y)$ denote the random variable on the message transcript obtained by Alice and Bob (over the random coins of Alice and Bob). A protocol Π is δ-error is for all input pairs (x, y) it errors with probability at most δ. As usual, let $R_\delta(f)$ denote the cost of the best δ-error randomized protocol for a function f.

Information Theory. Let μ be a distribution over a finite set Q and $X \sim \mu$.

- The **entropy** of a random variable X is defined as:

$$H(X) = \sum_{q \in Q} \mu(q) \log \frac{1}{\mu(q)}$$

- The **conditional entropy** of X given Y is defined by:

$$H(X|Y) = E_y[H(X|Y = y)]$$

  where $H(X|Y = y)$ is the entropy of the conditional distribution of X given $Y = y$.

- The **joint entropy** of X, Y is:

$$H(X, Y) = H(X) + H(Y|X)$$

- The **mutual information** between X and Y is given by:

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

- The **conditional mutual information** between X and Y conditioned on Z is:

$$I(X; Y|Z) = E_z[I(X; Y \mid Z = z)]$$

Fix a set $\mathbf{K}$ of inputs and a functions $f : \mathbf{K} \to \{0, 1\}$. In this case, we view $\mathbf{K}$ as consisting of two inputs for two players. Intuitively, the information cost of a protocol is the amount of information the protocol transcript reveals about the function inputs.

Definition [Information Cost of a Protocol.] Let Π be a randomized protocol whose inputs belong to $\mathbf{K}$ and μ be a distribution on inputs in $\mathbf{K}$. Let $(\mathbf{X}, \mathbf{Y})$ be the random variable distributes according to μ. The information cost of Π with respect to μ is denoted by $I(\mathbf{X}, \mathbf{Y}; \Pi(\mathbf{X}, \mathbf{Y}))$.

Definition [Information Complexity of a Function.] The δ-error information complexity of a function f with respect to distribution on the inputs μ is denoted by $IC_{\mu,\delta}(f)$ and defined to be the minimum information cost of a δ-error protocol for f.

It is not hard to see that a complexity of the best randomized protocol protocol is at least the information complexity of a function. Intuitively, this holds since the randomized algorithm must communicate at least the amount of information needed to compute the function. More formally,

$$R_\delta(f) = |\Pi| \geq H(\Pi(\mathbf{X}, \mathbf{Y})) \geq I((\mathbf{X}, \mathbf{Y}); \Pi(\mathbf{X}, \mathbf{Y})) \geq IC_{\mu,\delta}(f)$$

From now one, when we use $\mathbf{X}$ to denote a vector random variable, and $\mathbf{X}_i$ its individual entries.

Now, say a function f can be decomposed into n "simpler" primitives. Ideally, if we wanted to prove lower bounds on a function by proving lower bounds on simpler primitives, we would need something like a statement of the form $IC_{\mu,\delta} \geq n \cdot IC_{v,\delta}$. Unfortunately, we cannot claim a statement of this form since the distribution $\mu \sim (\mathbf{X}, \mathbf{Y})$ might not be a produce distribution (since $v \sim (\mathbf{X}_i, \mathbf{Y}_i)$ might not be a product distribution on its own). Instead, suppose $(\mathbf{X}_i, \mathbf{Y}_i) \sim v$, we define an "auxiliary" random variable $\mathbf{D}_i$ and let v' denote the joint distribution of $((\mathbf{X}_i, \mathbf{Y}_i), \mathbf{D}_i)$. We choose $\mathbf{D}_i$ such that conditioned on $\mathbf{D}_i$, $\mathbf{X}_i$ and $\mathbf{Y}_i$ are independent. Let $\mu = v^n$ and $\eta = (v')^n$. We call η a mixture of product distributions in this case. Note that if $((\mathbf{X}, \mathbf{Y}), \mathbf{D}) \sim \eta$ is a mixture of product distributions, then for all i, $(\mathbf{X}_i, \mathbf{Y}_i)$ are mutually independent even conditioned on $\mathbf{D}$.

Definition [Conditional Information Cost.] Let Π be a randomized protocol and let $((\mathbf{X}, \mathbf{Y}), \mathbf{D}) \sim \eta$ and that η is a mixture of product distribution. The conditional information cost of Π with respect to η is denoted by $I(\mathbf{X}, \mathbf{Y}; \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D})$.

Definition [Conditional Information Complexity.] The δ-error conditional information complexity of a function f with respect to η is denoted by $CIC_{\eta,\delta}(f)$ and defined to be the minimum conditional information cost of a δ-error protocol for f.

Lemma 1 *Let μ be a distribution on the inputs from the set $\mathbf{K}$ to a function f. If η is a mixture of produce distributions on $\mathbf{K} \times \mathbf{M}$, where the marginal distribution on $\mathbf{K}$ is μ, then*

$$IC_{\mu,\delta}(f) \geq CIC_{\eta,\delta}(f)$$

Proof Consider a protocol Π with information cost $IC_{\mu,\delta}$. Let $\eta \sim ((\mathbf{X}, \mathbf{Y}), \mathbf{D})$. By definition, $\mu \sim (\mathbf{X}, \mathbf{Y})$. Hence, since $\Pi(\mathbf{X}, \mathbf{Y})$ is conditionally independent of $\mathbf{D}$ given $\mathbf{X}, \mathbf{Y}$, then

$$IC_{\mu,\delta}(f) = I((\mathbf{X}, \mathbf{Y}); \Pi(\mathbf{X}, \mathbf{Y})) \geq I((\mathbf{X}, \mathbf{Y}); \Pi(\mathbf{X}, \mathbf{Y}) | \mathbf{D}) \geq CIC_{\eta,\delta}(f)$$

where the second inequality holds by the *data processing inequality* of information complexity.

Hence, our goal now will be to lower bound the conditional information complexity of a function on a mixture of product distributions.

3 Decomposing Disjointness Function

Clearly,

$$\mathsf{DISJ}_n(\mathbf{X}, \mathbf{Y}) = \vee_{i \in [n]}(\mathbf{X}_i \wedge \mathbf{Y}_i)$$

We now show how to define an "auxiliary" distribution $\mathbf{D}_i$ conditioned on which, each $\mathbf{X}_i$ and $\mathbf{Y}_i$ are independent. First, consider distribution $v \sim (\mathbf{X}_i, \mathbf{Y}_i)$ given by:

$$v(0,0) = 1/2 \quad v(0,1) = v(1,0) = 1/4$$

Let $\mathbf{D}_i$ denote a random variable on uniform distribution on $\{a, b\}$.

- If $\mathbf{D}_i = a$, then let $\mathbf{X}_i = 0$ and let $\mathbf{Y}_i$ be uniformly chosen from $\{0, 1\}$.
- If $\mathbf{D}_i = b$, then let $\mathbf{Y}_i = 0$ and let $\mathbf{X}_i$ be uniformly chosen from $\{0, 1\}$.

Denote $v' \sim ((\mathbf{X}_i, \mathbf{Y}_i), \mathbf{D}_i)$, and $\mu = v^n, \eta = (v')^n$. It is easy to see that conditioned on $\mathbf{D}_i$, $\mathbf{X}_i$ and $\mathbf{Y}_i$ are independent random variables. Therefore, the joint distribution $((\mathbf{X}_i, \mathbf{Y}_i), \mathbf{D}_i)$ (resp. $((\mathbf{X}, \mathbf{Y}), \mathbf{D})$) is a mixture of product distributions.

Theorem 2 *Let $v \sim (\mathbf{X}_i, \mathbf{Y}_i)$ and $v' \sim ((\mathbf{X}_i, \mathbf{Y}_i), \mathbf{D}_i)$ be a mixture of product distributions. Let $\mu = v^n$, $\eta = (v')^n$. Let Π be an δ-error randomized protocol for DISJ_n. Let AND denote the bitwise and of two bits. Then,*

$$I((\mathbf{X}_i, \mathbf{Y}_i), \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D}) \geq CIC_{v',\delta}(\mathsf{AND})$$

Proof Consider the protocol Π for DISJ_n. We will use it as a subroutine to show that there exists a protocol to solve AND function on input (b, b') with the same error and information cost. Let

$$\mathbf{D}_{-i} = \mathbf{D}_1, \ldots, \mathbf{D}_{i-1}, \mathbf{D}_{i+1}, \ldots, \mathbf{D}_n$$

Now, by definition,

$$I((\mathbf{X}_i, \mathbf{Y}_i); \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D}) = E_{\mathbf{d}}[I((\mathbf{X}_i, \mathbf{Y}_i); \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D}_i, \mathbf{D}_{-i} = \mathbf{d})]$$

- We derive a distribution of protocols $\Pi_{\mathbf{d}}$ using Π to solve AND function over v' distribution. Recall that $\eta = (v')^n$. Protocol $\Pi_{\mathbf{d}}$ has $\mathbf{d}$ hard-wired in it.
- Given bit b to Alice, she generates

  $$\mathbf{X}_{-i} = \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n$$

  conditioned on $\mathbf{D}_{-i} = \mathbf{d}$.
- Given bit b' to Bob, he generates

  $$\mathbf{Y}_{-i} = \mathbf{Y}_1, \ldots, \mathbf{Y}_{i-1}, \mathbf{Y}_{i+1}, \ldots, \mathbf{Y}_n$$

  conditioned on $\mathbf{D}_{-i} = \mathbf{d}$.
- Recall, that conditioned on $\mathbf{D}_{-i}$, $(\mathbf{X}_i, \mathbf{Y}_i)$ is a product distribution so Alice and Bob can generate their corresponding sequences independently.
- They put their input bits b, b' in the i'th positions of their sequences obtaining $\mathbf{X}, \mathbf{Y}$, respectively.
- Then, they run Π on $\mathbf{X}, \mathbf{Y}$ and output whatever it outputs.

It is not hard to see that the distribution produced is identical to $(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{D}_i, \Pi(\mathbf{X}_i, \mathbf{Y}_i))$. Hence, we obtain a protocol $\Pi_{\mathbf{d}}$ with the same error δ, establishing the claim by an averaging argument.

Lemma 3 *If $\eta \sim ((\mathbf{X}, \mathbf{Y}), \mathbf{D})$ is a mixture of product distributions, then*

$$I(\mathbf{X}, \mathbf{Y}; \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D}) \geq \sum_i I(\mathbf{X}_i, \mathbf{Y}_i; \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D})$$

Proof The proof follows directly by expanding the definitions and using the fact that $(\mathbf{X}_i, \mathbf{Y}_i)$ are independent conditioned on $\mathbf{D}$.

Combining Theorem 2 and Lemma 3, we obtain:

$$IC_{\mu,\delta} \geq CIC_{\eta,\delta}(\mathsf{DISJ}_n) \geq n \cdot CIC_{v',\delta}(\mathsf{AND})$$

Hence, it is left to lower bound the conditional information complexity of AND function.

4 $\Omega(1)$ Lower Bound of AND function

In this section, we will prove a lower bound $\Omega(1)$ on the AND function over $v \sim ((\mathbf{X}_i, \mathbf{Y}_i), \mathbf{D}_i)$ distribution. Consider an arbitrary Γ protocol for AND with δ-error. By definition,

$$I(\mathbf{X}_i, \mathbf{Y}_i; \Gamma \mid \mathbf{D}_i) =$$

$$\frac{1}{2} I((\mathbf{X}_i, \mathbf{Y}_i); \Gamma \mid \mathbf{D}_i = a) + \frac{1}{2} I((\mathbf{X}_i, \mathbf{Y}_i); \Gamma \mid \mathbf{D}_i = b) =$$

$$\frac{1}{2}[I(Z; \Gamma(0, Z)) + I(Z; \Gamma(Z, 0))]$$

Now intuitively, we do not expect the protocol to behave very differently on input $(1, 0)$ versus $(0, 1)$ since it must output 0 in both cases. However, we do expect it to behave differently on inputs $(0, 0)$ and $(1, 1)$. In order exploit this "intuition" we connect conditional information with Hellinger distance and show a special "cut-and-paste" property.

For any distribution V over a discrete domain $Q = \{q_1, \ldots, q_n\}$ can be viewed as a unit vector in $\mathbb{R}^Q$ whose i'th coordinate is $\sqrt{V(q_i)}$. The Hellinger distance is just $1/\sqrt{2}$ of the Euclidean distance between distributions viewed as vectors.

Definition [Hellinger Distance.] Let V and W be two probability distributions over a domain Q. The square of Hellinger distance is defined as:

$$h^2(V, W) = 1 - \sum_{q \in Q} \sqrt{V(q) \cdot W(q)}$$

We summarize a few properties connecting Hellinger distance and information. We refer the reader to the paper [1] for the proofs.

- (Hellinger distance and Information) Let $\psi(z_1), \psi(z_2)$ be two random variables and let Z denote a random variable with uniform distribution over $\{z_1, z_2\}$. Then,

$$I(Z; \psi(Z)) \geq h^2(\psi(z_1), \psi(z_2))$$

- (Cut-and-Paste Property) For any randomized protocol Γ and for any two pairs of inputs $(\mathbf{X}_i, \mathbf{Y}_i)$ and $(\mathbf{X}'_i, \mathbf{Y}'_i)$,

$$h(\Gamma(\mathbf{X}_i, \mathbf{Y}_i), \Gamma(\mathbf{X}'_i, \mathbf{Y}'_i)) = h(\Gamma(\mathbf{X}_i, \mathbf{Y}'_i), \Gamma(\mathbf{X}'_i, \mathbf{Y}_i))$$

  Cut-and-Paste property can be seen an an extension to the following property of deterministic protocols Γ'. If transcript induced by Γ' on (x, y) is identical to the transcript on (x', y'), then the transcripts induced by Γ' on inputs (x', y) and (x, y') must also be identical. For randomized algorithms, equality is replaced by the Hellinger distance over the protocol transcripts.

- (Soundness) For any δ-error protocol Γ for a function f, and for any two pairs of inputs $(\mathbf{X}_i, \mathbf{Y}_i)$ and $(\mathbf{X}'_i, \mathbf{Y}'_i)$ such that $f(\mathbf{X}_i, \mathbf{Y}_i) \neq f(\mathbf{X}'_i, \mathbf{Y}'_i)$,

$$h^2(\Gamma(\mathbf{X}_i, \mathbf{Y}_i), \Gamma(\mathbf{X}'_i, \mathbf{Y}'_i)) \geq 1 - 2\sqrt{\delta}$$

Putting it together, we obtain that for any protocol Γ computing AND with δ-error:

$$\begin{aligned}
I((\mathbf{X}_i, \mathbf{Y}_i); \Gamma(\mathbf{X}_i, \mathbf{Y}_i) \mid \mathbf{D}_i) &\geq \tfrac{1}{2}[I(Z; \Gamma(0, Z)) + I(Z; \Gamma(Z, 0))] \\
&\geq (1/2)[h^2(\Gamma(0,1), \Gamma(0,0)) + h^2(\Gamma(0,0), \Gamma(1,0))] \\
&\geq (1/4)[h(\Gamma(0,0), \Gamma(0,1)) + h(\Gamma(0,0), \Gamma(1,0))]^2 && \text{(Cauchy-Schwarz)} \\
&\geq (1/4)[h^2(\Gamma(0,1), \Gamma(1,0))] && \text{(Triangle Inequality)} \\
&= (1/4)[h^2(\Gamma(0,0), \Gamma(1,1))] && \text{(Cut-and-Paste)} \\
&= (1/4)[1 - 2\sqrt{\delta}] && \text{(Soundness)}
\end{aligned}$$

Hence, $CIC_{v',\delta}(\mathsf{AND}) = \Omega(1)$. As a result, we obtain $\Omega(n)$ lower bound on DISJ_n function.

References

[1] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar, *An information statistics approach to data stream and communication complexity*, in Proceedings of the 43rd Symposium on Foundations of Computer Science, FOCS '02, Washington, DC, USA, 2002, IEEE Computer Society, pp. 209–218.

[2] A. Chattopadhyay and T. Pitassi, *The story of set disjointness*, SIGACT News, 41 (2010), pp. 59–85.