COMS 6998-2: Information Theory in Theoretical Computer Science Oct 17, 2017

Lecture 7 – Discrepancy, Information Cost and Hellinger Distance

Instructor: *Omri Weinstein* Scribes: *Kailash Meiyappan*

1 Introduction

In the last lecture, we started looking at randomized communication complexity. We looked at private and public randomness models and the minimax theorem.

The usual lower bound techniques for deterministic communication complexity don't work for randomized / distributional protocols. For example, looking for monochromatic rectangles no longer makes sense because we're allowed to have small errors. The goal of this lecture is to introduce tools to lower bound randomized communication complexity.

2 Discrepancy Lower Bound

Idea: Most rectangles are "roughly" monochromatic if we allow some small error ϵ. For a given function $f(x, y)$ over an input distribution μ, we define the discrepancy of a rectangle $R = S \times T \subseteq X \times Y$ as:

Definition 1.

$$Disc_\mu^f := \mu(R)|Pr_\mu(f = 1|R) - Pr_\mu(f = 0|R)|$$

Intuitively, the discrepancy can be thought of as the advantage or bias of deciding the function value for R. In particular, note that for a monochromatic rectangle, $Disc_\mu^f(R) = \mu(R)$. We now define the discrepancy of a function on a probability distribution μ as:

Definition 2.

$$Disc_\mu(f) = \max_{R=S\times T} Disc_\mu^f(R) := \delta$$

Intuitively, the discrepancy of the function is the largest advantage of any rectangle.

Theorem 3. *(Discrepancy Lower Bound):*
$\forall\ f\ \forall\ \mu,\ D_\mu^{\frac{1}{2}-\epsilon}(f) \geq \log\left(\frac{2\epsilon}{Disc_\mu(f)}\right)$

Note that it is trivial to achieve an advantage of $\frac{1}{2}$ if you just output the majority over the distribution μ.

Proof. Let Π be the protocol such that $|\Pi| := c = D_\mu^{\frac{1}{2}-\epsilon}(f)$.

$$Pr_{(x,y)\sim\mu}[\Pi(x, y) = f(x, y)] \geq \frac{1}{2} + \epsilon$$

$$Pr_{(x,y)\sim\mu}[\Pi(x, y) \neq f(x, y)] \leq \frac{1}{2} - \epsilon$$

So,

$$2\epsilon \leq Pr_{(x,y)\sim\mu}[\Pi(x,y) = f(x,y)] - Pr_{(x,y)\sim\mu}[\Pi(x,y) \neq f(x,y)]$$

We can break the difference in probabilities into a sum over each of the 2^c monochromatic rectangles that are formed by Π.

$$= \sum_{i=1}^{2^c} \mu(R_i)|Pr_\mu[\Pi = f|R_i] - Pr_\mu[\Pi \neq f|R_i]|$$

The term in the summation is precisely the discrepancy over the rectangle R_i!

$$= \sum_{i=1}^{2^c} Disc_\mu^f(R_i) \leq 2^c Disc_\mu(f)$$

And so,

$$c \geq \log \frac{2\epsilon}{Disc_\mu(f)}$$

□

Note that the last inequality in the proof is not a tight lower bound, and so this technique works better if the rectangles are roughly balanced in the discrepancy. If there is even one rectangle with an abnormally large discrepancy value, then the lower bound will not be robust because every rectangle will be treated as though it has a large discrepancy according to the inequality.

Example 4. *We show how to use this to lower bound inner product. Recall that*

$$IP_n(x,y) = \langle x,y\rangle \% 2 = \sum_i x_i y_i \mod 2$$

Claim: $D_{unif}^{\frac{1}{4}}(IP_n) \geq \Omega(n)$

Proof. There are three main ideas used in to prove this claim:

1. View IP_n as having an output in $\{[1,-1\}$, where 1 is the output 0 and -1 is the output 1. Intuitively, inner product becomes:
$$IP_n(x,y) = -1^{\langle x,y\rangle}$$
2. The Hadamard matrix H where the entry $H_{i,j} = \langle i,j\rangle$ in the $-1,1$ notation. In particular, $H^tH = 2^n I_{2^n}$.
3. Note that for a rectangle $R = S \times T$ in the Hadamard matrix,
$$Disc_\mu^{IP_n}(R) = \sum_{x,y\in R} -1^{f(x,y)}\mu(x,y) = \frac{1}{2^{2n}} \mathbb{1}_S^t H \mathbb{1}_T$$

From Cauchy Schwartz, we get:

$$\frac{1}{2^{2n}} \mathbb{1}_S^t H \mathbb{1}_T = \frac{1}{2^{2n}} \langle \mathbb{1}_S, H\mathbb{1}_T\rangle \leq \frac{1}{2^{2n}} ||\mathbb{1}||_2 ||H\mathbb{1}_T||_2 = \frac{1}{2^{2n}} \sqrt{|S|}\sqrt{\langle H\mathbb{1}_T, H\mathbb{1}_T\rangle}$$

$$= \frac{1}{2^{2n}}\sqrt{|S|}\sqrt{\mathbb{1}_T^t H^t H \mathbb{1}_T} = \frac{1}{2^{2n}}\sqrt{|S|}\sqrt{\mathbb{1}_T^t 2^n I_{2^n} \mathbb{1}_T} = \frac{1}{2^{2n}}\sqrt{|S||T|2^n} \leq \frac{1}{2^{2n}}\sqrt{2^n 2^n 2^n} = \frac{1}{2^{-n/2}}$$

And to complete the proof, from the theorem, we get:

$$D^{1/4}_{unif}(IP_n) \geq \Omega\left(\log\left(\frac{1}{Disc_\mu(IP_n)}\right)\right) = \log(2^{n/2}) = n/2$$

□

The discrepancy lower bound works well here, but that isn't always true. Consider set disjointness.

Example 5. $DISJ_n(x, y) = 1 \iff X \cap Y = \phi$
Claim: $\forall \mu, Disc_\mu(DISJ_n) \geq \Omega(1/n)$

Proof. Let A^ϕ be the "1-inputs" and $A^{\neq\phi}$ be the "0-inputs". If $|\mu(A^\phi) - \mu(A^{\neq\phi})| > \frac{1}{n}$, we are done, so let's assume without loss of generality that $|\mu(A^\phi) - \mu(A^{\neq\phi})| \leq \frac{1}{n}$
So, $\mu(A^{\neq\phi}) \geq \frac{1}{2} - \frac{1}{n}$
Consider the set $R_i := \{(x, y)|i \in x \cap y\}$

$$A^\phi = \bigcup_{i=1}^{n} R_i \Rightarrow \mu(\bigcup R_i) \geq \frac{1}{2} - \frac{1}{n} \Rightarrow \exists\, i_0 \; s.t. \; \mu(R_{i_0}) \geq \frac{1}{n}(\frac{1}{2} - \frac{1}{n}) = \Omega(1/n)$$

This completes the proof, because the discrepancy is equal to this probability, since the rectangle is constructed such that all entries in it are not disjoint (0-entries).

□

3 Information Complexity

It is clear that while the discrepancy technique is powerful in some cases, it does not solve the set disjointness problem. We would like to explore methods that could lead to better (ideally linear) lower bounds for disjointness. The idea of information complexity is this. Given a protocol Π, with inputs $(x, y) \sim \mu$. Instead of measuring the communication $|\Pi|$, we measure the actual information conveyed by the transcript of the protocol (since the communication could have been inefficient).

Definition 6. *Information Cost of a protocol* Π *for an input distribution* μ *is defined as:*

$$IC_\mu(\Pi) = I(\Pi; X|Y) + I(\Pi; Y|X)$$

Here, Π *is treated as a random variable over the public or private randomness R. Hence,* Π_{xy} *is still a random variable.*
We can write this in terms of the KL divergence as:

$$IC_\mu(\Pi) = E_{XY}\left[\mathbb{D}\left(\Pi_{x,y} \| \Pi_y\right)\right] + E_{XY}\left[\mathbb{D}\left(\Pi_{x,y} \| \Pi_x\right)\right]$$

where the notation Π_x *is* Π *conditioned on* x*.*

Example 7. *Suppose the protocol* Π *is for A and B to exchange inputs. The communication complexity* $||\Pi|| = 2n$*. However, depending on* μ*, the information cost could be very low. In particular, if* μ *was such that* $x = y$ *with high probability, then the information cost is 0.*

This definition of information cost is sometimes called "Internal" information cost, because it is conditioned on the internal input x and y. An alternative definition is the external information cost, which is the information learnt by a third player, who is an observer.

Definition 8 (External IC)**.** *Another information measure which makes sense at certain contexts is the* external *information cost of a protocol,* $\mathsf{IC}^{\mathsf{ext}}{}_{\pi,\mu} := I(\Pi; XY)$*, which captures what an* external *observer learns on average about both player's inputs by observing the transcript of* π*. This quantity will be of minor interest in this survey (though it playes a central role in many applications). The external information cost of a protocol is always at least as large as its (internal) information cost, since intuitively an external observer is "more ignorant" to begin with. We remark that when* μ *is a* product *distribution, then* $\mathsf{IC}^{\mathsf{ext}}{}_{\pi,\mu} = \mathsf{IC}_{\mu}(\pi)$*.*

One can now define the *information complexity* of a function f with respect to μ and error ϵ as the least amount of information the players need to reveal to each other in order to compute f with error at most ϵ:

Definition 9. *The Information Complexity* of f with respect to μ (and error ϵ) is

$$\mathsf{IC}^{\epsilon}_{\mu}(f) := \inf_{\pi:\, \Pr_{\mu}[\pi(x,y)\neq f(x,y)]\leq\epsilon} \mathsf{IC}_{\mu}(\pi).$$

Note that we cannot must use the infimum and not the minimum because in order to minimize the information cost, we may take infinite rounds of communication.
What is the relationship between the information and communication complexity of f? This question and its (major) implications within TCS will occupy us for a full lecture in 2-3 weeks from now. There is one relationship that is easy to show:

$$\forall\mu\forall f, IC^{\epsilon}_{\mu}(\Pi) \leq D^{\epsilon}_{\mu}(f)$$

This is true because one bit of communication can only convey at most one bit of information. Hard: Is it true that

$$D^{\epsilon}_{\mu}(f) \leq O(IC^{\epsilon}_{\mu}(f))?$$

In other words, can we compress a communication protocol to only convey number of bits equal to the information learnt?

We saw in the beginning of this course that the answer to this question was affirmative (in fact, an equality) for *one-way* protocols where only Alice speaks (Shannon's noiseless coding them and the Slepian-Wolf amortized them, Huffman one-shot). It turns out that, at least in the amortized sense, the answer is "YES" (we'll see this later on in the course). The answer to the *one-shot* problem, i.e., "is there an interactive analogue of the Huffman code?", is one of the most fascinating open problems in communication complexity, and we will devote an entire lecture to this problem and its major implications within complexity theory. But for the purpose of the next 2 lectures, we will study IC as an (elegant) tool for proving LBs on distributional (i.e., randomized) CC.

The role of private randomness in information complexity. A subtle but vital issue when dealing with information complexity, is understanding the role of private vs. public randomness. In public-coin communication complexity, one often ignores the usage of private coins in a protocol, as they can always be simulated by public coins. When dealing with *information complexity*, the situation is somewhat the opposite: Public coins are essentially a redundant resource (as it can be easily shown via the chain rule that $\mathsf{IC}_\mu(\pi) = \mathbb{E}_R[\mathsf{IC}_\mu^{\pi_R}]$), while the usage of private coins is crucial for minimizing the information cost, and fixing these coins is prohibitive (once again, for communication purposes in the distributional model, one may always fix the entire randomness of the protocol, via the averaging principle). To illustrate this point, consider the simple example where in the protocol π, Alice sends Bob her 1-bit input $X \sim Ber(1/2)$, XORed with some random bit Z. If Z is private, Alice's message clearly reveals 0 bits of information to Bob about X. However, for any fixing of Z, this message would reveal an entire bit(!). The general intuition is that a protocol with low information cost would reveal information about the player's inputs in a "careful manner", and the usage of private coins serves to "conceal" parts of their inputs. Indeed, it was recently shown that the restriction to public coins may cause an exponential blowup in the information revealed compared to private-coin protocols ([1, 2]).

Lemma 10 (Additivity of Information Complexity)**.** $\mathsf{IC}^{\varepsilon}_{\mu^n}(f^n) = n \cdot \mathsf{IC}^{\varepsilon}_{\mu}(f)$.

Proof. The ($\leq$) direction of the lemma is easy, and follows from a simple argument that applies the single-copy optimal protocol independently to each copy of f^n, with independent randomness. We leave the simple analysis of this protocol as an exercise to the reader.

The ($\geq$) direction is the main challenge. Will will prove it in a contra-positive fashion: Let Π be an ε-error protocol for f^n, such that $\mathsf{IC}^{\Pi}_{\mu^n} = I$ (recall that here ε denotes the per-copy error of Π in computing $f(x_i, y_i)$). We shall use Π to produce a *single-copy* protocol for f whose information cost is $\leq I/n$, which would complete the proof. The guiding intuition for this is that Π should reveal I/n bits of information about an average coordinate.

To formalize this intuition, let $(x, y) \sim \mu$, and denote $\overline{X} := X_1 \dots X_n$, $X_{\leq i} := X_1 \dots X_i$ and $X_{-i} := X_1 \dots X_{i-1}, X_{i+1}, \dots, X_n$, and similarly for $\overline{Y}, Y_{\leq i}, Y_{-i}$. A natural idea is for Alice and Bob to "embed" their respective inputs (x, y) to a (publicly chosen) random coordinate $i \in [n]$ of Π, and execute Π. However, Π is defined over n input copies, so in order to execute it, the players need to somehow "fill in" the rest $(n-1)$ coordinates, each according to μ. How should this step be done? The first attempt is for Alice and Bob to try and complete X_{-i}, Y_{-i} privately. This approach fails if μ is a non-product distribution, since there's no way the players can sample X and Y privately, such that $(X, Y) \sim \mu$ if μ correlates the inputs. The other extreme – sampling X_{-i}, Y_{-i} using public randomness only – would resolve the aforementioned correctness issue, but might leak too much information: An instructive example to consider is where, in the first message of Π, Alice sends Bob the XOR of the n bits of her uniform input X: $M = X_1 \oplus X_2 \oplus \dots \oplus X_n$. Conditioned on X_{-i}, Y_{-i}, M reveals 1 bit of information about X_i to Bob, while we want to argue that in this case, only $1/n$ bits are revealed about X_i. So this approach reveals too much information.

It turns out that the "right" way of breaking the dependence across the coordinates is to use a combination of public and private randomness. Let us define, for each $i \in [n]$, the public random variable

$$R_i := X_{<i}, Y_{>i}.$$

Note that given R_i, Alice can complete all her missing inputs $X_{>i}$ *privately* according to μ, and Bob

Figure 1:

can do the same for $Y_{<i}$. Let us denote by $\theta(x, y, i, R_i)$ the protocol transcript produced by running $\Pi(X_1, ..., X_{i-1}, x, X_{i+1}, ..., X_n\ ,\ Y_1, ..., Y_{i-1}, y, Y_{i+1}, ..., Y_n)$ and outputting its answer on the i'th coordinate. Let $\Theta(x, y)$ be the protocol obtained by running $\theta(x, y, i, R_i)$ on a uniformly selected $i \in [n]$.

By definition, Π computes f^n with a *per-copy* error of ε, and thus in particular $\Theta(x, y) = f(x, y)$ with probability $\geq 1 - \varepsilon$. To analyze the information cost of Θ, we write:

$$\begin{aligned} I(\Theta; x|y) &= \mathbb{E}_{i,R_i}[I(\theta; x|y, R_i)] = \sum_{i=1}^{n} \frac{1}{n} \cdot I(\Pi; X_i \mid Y_i, R_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} I(\Pi; X_i \mid Y_i, X_{<i}Y_{>i}) = \frac{1}{n} \sum_{i=1}^{n} I(\Pi; X_i \mid X_{<i}Y_{\geq i}) \\ &\leq \frac{1}{n} \sum_{i=1}^{n} I(\Pi; X_i \mid X_{<i}\overline{Y}) = \frac{1}{n} \cdot I(\Pi; \overline{X} \mid \overline{Y}), \end{aligned}$$

where the inequality follows from Lemma 11, since $I(Y_{<i}; X_i|X_{<i}) = 0$ by construction, and the last transition is by the chain rule for mutual information. By symmetry of construction, an analogous argument shows that $I(\Theta; y|x) \leq I(\Pi; \overline{Y} \mid \overline{X})/n$, and combining these facts gives

$$\mathsf{IC}_{\mu}^{\Theta} \leq \frac{1}{n} \left(I(\Pi; \overline{X} \mid \overline{Y}) + I(\Pi; \overline{Y} \mid \overline{X}) \right) = \frac{I}{n}. \tag{1}$$

□

Lemma 11 (Conditioning on independent variables does not decrease information)**.** *Let A, B, C, D be four random variables in the same probability space. If A and D are conditionally independent given C, then it holds that $I(A; B|C) \leq I(A; B|CD)$.*

Proof. We apply the chain rule for mutual information twice. On one hand, we have $I(A; BD|C) = I(A; B|C) + I(A; D|CB) \geq I(A; B|C)$ since mutual information is nonnegative. On the other hand, $I(A; BD|C) = I(A; D|C) + I(A; B|CD) = I(A; B|CD)$ since $I(A; D|C) = 0$ by the independence assumption on A and D. Combining both equations completes the proof. □

References

[1] Balthazar Bauer, Shay Moran, and Amir Yehudayoff. "Internal Compression of Protocols to Entropy". In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2015, August 24-26, 2015, Princeton, NJ, USA*. 2015, pp. 481–496.

DOI: 10.4230/LIPIcs.APPROX-RANDOM.2015.481. URL: https://doi.org/10.4230/LIPIcs.APPROX-RANDOM.2015.481.

[2] Anat Ganor, Gillat Kol, and Ran Raz. "Exponential Separation of Information and Communication for Boolean Functions". In: *J. ACM* 63.5 (2016), 46:1–46:31. URL: http://dl.acm.org/citation.cfm?id=2907939.