COS597D: Information Theory in Computer Science November 21 and 23, 2011

Lecture 17-18

Lecturer: Mark Braverman *Scribe: Ankit Garg*

Last time, we defined the (internal) information cost of a protocol and a function.

$$IC(\pi,\mu) = I(\pi;X|Y) + I(\pi;Y|X)$$
$$IC_\mu^\epsilon(f) = \inf_{\substack{\pi \\ Pr_\mu(\pi(x,y)\neq f(x,y))\leq\epsilon}} IC(\pi,\mu)$$

The information cost is an interesting measure to study in itself. But what makes it even more interesting, is that it is useful in proving direct-sum results. i.e. the question:

$$D_{\mu^n}^\epsilon(f^n) \overset{?}{=} \Omega(nD_\mu^\epsilon(f))$$

Information cost seems to be the right quantity to study because of the following theorem, proved in [BR10]:

$$IC_\mu^\epsilon(f) = \lim_{n\to\infty}\frac{D_{\mu^n}^\epsilon(f^n)}{n}$$

We start by showing how information theoretic ideas can be useful in proving lower bounds in communication complexity.

1. Lower bound for Disjointness

Theorem 1. *The randomized communication complexity of non-disjointness $NONDISJ_n(X,Y) = \vee_{i=1}^n(x_i \wedge y_i)$ is $\Omega(n)$.*

Note that this will also imply that the randomized communication complexity of $DISJ_n$ is $\Omega(n)$.

Proof The theorem is from the 90's [KS92, Raz92] and the proofs were a little complicated, but information theoretic ideas helped simplify the proofs, at least conceptually [BYJKS04]. The main idea is that if we solve this problem using $o(n)$ bits of communication, then for some pair of bits, x_i, y_i, we convey very little amount of information, and hence we have no idea what $x_i \wedge y_i$ is. It requires a remarkably clever argument to formalize this. We prove $R_{1/10}(NONDISJ_n) > n/1000$. Define the distribution μ on pair of bits as

$$\mu = \begin{cases} 00 & \text{w.p. } 1/3 \\ 01 & \text{w.p. } 1/3 \\ 10 & \text{w.p. } 1/3 \end{cases}$$

Suppose that π_n is a protocol that computes $NONDISJ_n$ correctly w.p. $> 9/10$ *on all* inputs, and $CC(\pi_n) \leq n/1000$. The proof will go in two steps :

(1) Establish a protocol $\pi(x,y)$, $x,y \in \{0,1\}$, that computes $x\wedge y$ s.t. $IC(\pi,\mu)$ is small, and $\forall(x,y)$, $\pi(x,y) = x\wedge y$ w.p. $> 9/10$.

(2) Prove such a protocol cannot exist.

First we describe a protocol π for computing AND, that will convey very little information, but still compute AND *on all* inputs with high probability.

π : Input (x,y), output $x\wedge y$

(1) Alice and Bob publicly sample $i \in \{1,\ldots,n\}$ uniformly, and set $X_i = x, Y_i = y$.

(2) Alice and Bob publicly sample $X_1, \ldots, X_{i-1}, Y_{i+1}, \ldots, Y_n$, each 0 w.p. 2/3 and 1 w.p. 1/3.
(3) Alice and Bob privately (and conditionally) sample $X_{i+1}, \ldots, X_n, Y_1, \ldots, Y_{i-1}$ so that each pair (X_j, Y_j) is distributed according to μ.
(4) Run $\pi_n(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ and output the answer.

The sampling procedure might seem weird, but as we will see now, it almost leads to a miracle of some sort. Clearly this procedure will output $x \wedge y$ if the protocol π_n returns a correct answer on $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$. Since π_n returns a correct answer *on all* inputs w.p. $> 9/10$, hence $\forall (x, y)$, π_n will output $x \wedge y$ w.p. $> 9/10$.

Now, we will show that $IC(\pi, \mu)$ is small. Suppose that the input to π is distributed according to μ, and let X', Y' denote the random variables for the input. Let $X = X_1, \ldots, X_n$ and $Y = Y_1, \ldots, Y_n$. Note that (X, Y) is distributed according to μ^n. Also let I be the random variable for the index Alice and Bob sample in the first step. Then

$$\begin{aligned} I(\pi; Y'|X') &= I(\pi_n; Y'|I, X_1, \ldots, X_{I-1}, X', X_{I+1}, \ldots, X_n, Y_{I+1}, \ldots, Y_n) \text{ (What Alice learns)} \\ &= \frac{1}{n} \sum_{i=1}^{n} I(\pi_n; Y_i|X, Y_{i+1}, \ldots, Y_n) \text{ (removing the conditioning on } I) \\ &= \frac{1}{n} I(\pi_n; Y|X) \text{ (chain rule)} \end{aligned}$$

Similarly $I(\pi; X'|Y') = \frac{1}{n} I(\pi_n; X|Y)$, and hence $IC(\pi, \mu) = \frac{1}{n} IC(\pi_n, \mu^n) \leq 1/1000$. We have a long protocol for computing AND, and yet we transmit very little information. We still have to do some work to show that such a protocol cannot exist, but that is mostly mechanical. The miraculous sampling is essentially the heart of the proof.

Note that this sampling procedure can be generalized to prove that $IC(f, \mu, \epsilon) \leq \frac{IC(f^n, \mu^n, \epsilon)}{n}$. And, if it is required that the error for each copy individually is $\leq \epsilon$, then equality holds, because the trivial protocol that runs the "single-copy-protocol" for each copy of f^n indivdually has information cost $nIC(f, \mu, \epsilon)$.

We now show that there cannot exist a protocol π s.t. $IC(\pi, \mu) \leq 1/1000$ and $\forall (x, y)$, $\pi(x, y) = x \wedge y$ w.p. $> 9/10$. Hellinger distance can be used to make this part simpler, but to emphasize that this part is just mechanical, we prove this by elementary means. Recall that for two distributions P, Q, the divergence is defined as

$$D(P||Q) = \sum_x P(x) \log \left(\frac{P(x)}{Q(x)} \right)$$

The information between two random variables is defined as

$$I(X; Y) = \mathbb{E}_x D[Y|_x || Y]$$

Consider the random variables $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}$, where π_{xy} denotes the transcript of π on (x, y) as input. Also let $\pi_{0?} = \frac{\pi_{00} + \pi_{01}}{2}$. Similarly define $\pi_{?0}$. Let (X, Y) be distributed according to μ. Now

$$\begin{aligned} \frac{1}{1000} &\geq I(\pi; X|Y) \\ &= \frac{2}{3} I(\pi; X|Y = 0) + \frac{1}{3} I(\pi; X|Y = 1) \\ &= \frac{2}{3} I(\pi; X|Y = 0) \text{ (if } Y = 1\text{, then } X \text{ is fixed)} \\ &= \frac{2}{3} (\frac{1}{2} D(\pi_{00}||\pi_{?0}) + \frac{1}{2} D(\pi_{10}||\pi_{?0})) \end{aligned}$$

Hence $D(\pi_{00}||\pi_{?0}) \leq \frac{3}{1000}$ and $D(\pi_{10}||\pi_{?0}) \leq \frac{3}{1000}$. Now, recall that for distributions P, Q, $||P-Q||_1 \leq \sqrt{2 \ln 2\, D(P||Q)}$. Thus $||\pi_{00} - \pi_{?0}||_1 < 0.065$ and $||\pi_{10} - \pi_{?0}||_1 < 0.065$. Thus $||\pi_{00} - \pi_{10}||_1 < 0.13$. Also $||\pi_{xy} - \pi_{11}||_1 \geq 1.6$, if one of x, y is 0. This is because, if the protocol is not wrong on both (x, y) and $(1,1)$, then the outputs are different, and since it is not wrong on both w.p. ≥ 0.8. Therefore the statistical distance is atleast 0.8, and thus $||\pi_{xy} - \pi_{11}||_1 \geq 1.6$.

Now denote by $\pi_{xy}(z)$, the probability that given x, y, the protocol has transcript z. We can write $\pi_{xy}(z) = P_x(z)Q_y(z)$, where $P_x(z)$ is defined as follows (let Z denote the path from the root to the leaf consistent with z)

$$P_x(z) = \prod_{\substack{v \in Z \\ \text{Alice owns } v}} Pr[\text{Alice's move at } v \text{ consistent with } z| \text{ reaching } v]$$

$Q_y(z)$ is defined similarly. Note that here we have crucially used that π is a protocol. This implies

$$\pi_{11}(z) \geq min(\pi_{01}(z) + \pi_{10}(z) - \pi_{00}(z), \pi_{01}(z), \pi_{10}(z))$$

Indeed, this is true if $P_1(z) \geq P_0(z)$ or $Q_1(z) \geq Q_0(z)$. So, we can assume that $P_1(z) \leq P_0(z)$ and $Q_1(z) \leq Q_0(z)$. Then

$$(P_0(z) - P_1(z))(Q_0(z) - Q_1(z)) \geq 0 \implies \pi_{11}(z) \geq \pi_{01}(z) + \pi_{10}(z) - \pi_{00}(z)$$

Now it is easy to check this in turn implies that if $\pi_{00}(z) \geq \pi_{11}(z)$, then

$$\pi_{00}(z) - \pi_{11}(z) \leq |\pi_{10}(z) - \pi_{00}(z)| + |\pi_{01}(z) - \pi_{00}(z)|$$

Thus

$$\sum_{z:\pi_{00}(z) \geq \pi_{11}(z)} \pi_{00}(z) - \pi_{11}(z) \leq \sum_z |\pi_{10}(z) - \pi_{00}(z)| + \sum_z |\pi_{01}(z) - \pi_{10}(z)| \leq 0.26$$

Thus $||\pi_{00} - \pi_{11}||_1 \leq 0.52$, which is a contradiction. ■

2. A Direct Sum result

As we mentioned, the sampling procedure can be generalized to prove that $IC(f, \mu, \epsilon) \leq \frac{IC(f^n, \mu^n, \epsilon)}{n}$, and since $IC(f^n, \mu^n, \epsilon) \leq D^{\epsilon}_{\mu^n}(f)$, to prove direct sum results, we just need to prove $D^{\epsilon}_{\mu}(f) = O(IC(f, \mu, \epsilon))$. This would prove $D^{\epsilon}_{\mu^n}(f^n) = \Omega(nD^{\epsilon}_{\mu}(f))$. This is not known. We prove a weaker result.

We prove the weak direct sum result by proving a theorem about compression of protocols, that is, we study the question, whether a low information protocol for a problem implies the existence of a protocol with low communication.

2.1. Compression.

Theorem 2. [BBCR10, Bra11] *Given a protocol with internal information cost, $IC(\pi, \mu) = I$, external information cost,$I(\Pi, XY) = I^{ext}$, and $|\pi| = C$, we can simulate π using protocols π'(adding a small amount of error) s.t.*

(1) $|\pi'| \leq O(\sqrt{CI} polylog(C))$
(2) $|\pi'| \leq O(I^{ext} polylog(C))$
(3) $|\pi'| \leq 2^{O(I)}$

Note that the protocols in parts 1,2 and 3 are different.

This gives us the following theorem :

Theorem 3. *For every $\alpha > 0$, $D^{\mu^n}_{\rho}(f^n).polylog(D^{\mu^n}_{\rho}(f^n)/\alpha) \geq \Omega(\alpha\sqrt{n}D^{\mu}_{\rho+\alpha}(f))$*

Proof We will not worry about the technicalities regarding the error α we introduce. Let π be a protocol that computes f^n with error ϵ(for each copy individually) w.r.t. μ^n, and let $|\pi| = C_n$. Then by the embedding argument used before, we get a protocol π' computing f s.t. $|\pi'| \leq C_n$ and $IC(\pi', \mu) \leq C_n/n$. Now we can compress this to get a protocol π'' for f with $|\pi''| \leq O(\sqrt{C_n^2/n}\, polylog(C_n))$, thus the theorem. The error α is because we introduce some extra error while compressing. ■

Note that we will always add some error while compressing because information is an average case quantitity, while communication is worst case. For some, pairs (x, y), we might run into a long branch in the protocol tree, and we might have to cut that branch in order to keep the protocol cost small.

We now prove part (1) of Theorem 3. We omit some technical details. The reader is referred to [3] for the full proof.

Proof Compressing each message separately is not a good idea, because it might be that the protocol releases information at a very low and uniform rate. So somehow, we have to simulate many rounds with small amount of communication.

$$IC(\pi, \mu) = I(\pi; X|Y) + I(\pi; Y|X) = \mathbb{E}_{xy} D(\pi_{xy}||\pi_y) + \mathbb{E}_{xy} D(\pi_{xy}||\pi_x)$$

π_{xy} denotes the random variable for the transcript of the protocol π if Alice's input is x and Bob's input is y. π_x denotes the random variable for the transcript if Alice's input is x and y is distributed according to μ conditioned on Alice's input being x. Similarly define π_y.

Now consider the case when $IC(\pi, \mu) = 0$. Thus $\pi_{xy} = \pi_x = \pi_y$. Then Alice and Bob don't need to communicate as they can sample a path distributed according to $\pi_{xy}(= \pi_x = \pi_y)$ using public randomness. How to sample the same path is explained below (note that if we can sample the same bit at each node in the protocol tree, then we are fine).

Suppose $0 \leq p, q \leq 1$. Alice knows the distribution B_p of a bit, and Bob has an estimate B_q. Then they can sample the same bit (with Alice's bit's distribution B_p, and Bob's B_q) with error $|p - q|$ using shared randomness. Using shared randomness, they sample a uniformly random number u between 0 and 1. Then Alice selects 1 if $u < p$, otherwise 0. Bob selects 1 if $u < q$, else 0. This trick is known as Holenstein's Lemma [Hol07].

This is crucial, since if Bob has a good estimate of the bit to be transmitted by Alice at some node, then Alice doesn't need to transmit that bit and they can sample it using public randomness. Of course, Alice and Bob don't know when their estimates are close. So we still need to do some more work.

Now we describe the protocol :

(1) Using their estimates π_x and π_y, Alice and Bob sample a path (by applying the sampling trick at each node).

(2) Using hashing, they check if they reach the same leaf, if not, then obtain the first disagreement, fix it (listen to the owner of the node), and continue the sampling from there.

We find the first disagreement using binary search and use $\log(CC(\pi))$ hash samples at each step, so we take $O(\log^2(CC(\pi)))$ communication to find and fix one mistake (this step can be improved to $O(\log CC(\pi))$ communication using a more careful construction). So the expected communication complexity is $O(\log^2(CC(\pi))) * \mathbb{E}_{\pi(X,Y)}[\#\, mistakes\, on\, \pi(X, Y)]$(because finally we are sampling a path according to $\pi(X, Y)$). Now the expected number of mistakes is in some sort, proportional to the internal information cost, as whenever there is a mistake, say on Alice's node, then Bob's estimate is way off, hence Bob will learn a lot of information when the bit is transmited on this node. To bound the expected number of Bob's mistakes on Alice's nodes,

$$\begin{aligned}
\#\text{Bob's mistakes} &= \mathbb{E}_{xy}||\pi_1|_{xy} - \pi_1|_y||_1 + \mathbb{E}_{xy}||\pi_3|_{xy} - \pi_3|_y||_3 + \ldots \\
&\leq O(\mathbb{E}_{xy}\sqrt{D(\pi_1|_{xy}||\pi_1|_y)} + \mathbb{E}_{xy}\sqrt{D(\pi_3|_{xy}||\pi_3|_y)} + \ldots) \;\; (||p-q||_1 \leq O(\sqrt{D(p||q)})) \\
&\leq O(\sqrt{\mathbb{E}_{xy}D(\pi_1|_{xy}||\pi_1|_y)} + \sqrt{\mathbb{E}_{xy}D(\pi_3|_{xy}||\pi_3|_y)} + \ldots) \;\; (\text{concavity of } \sqrt{z}) \\
&\leq O(\sqrt{CC(\pi)(\mathbb{E}_{xy}D(\pi_1|_{xy}||\pi_1|_y) + \mathbb{E}_{xy}D(\pi_3|_{xy}||\pi_3|_y) + \ldots)}) \;\; (\text{Cauchy Schwarz inequality})
\end{aligned}$$

Now

$$\begin{aligned}
I(\pi;X|Y) &= I(\pi_1\pi_2\ldots\pi_C;X|Y) \\
&= I(\pi_1;X|Y) + I(\pi_2;X|Y\pi_1) + \ldots + I(\pi_C;X|Y\pi_1\ldots\pi_{C-1}) \\
&= I(\pi_1;X|Y) + I(\pi_3;X|Y\pi_1\pi_2) + \ldots \;\; (I(\pi_{2r};X|Y\pi_1\ldots\pi_{2r-1}) = 0) \\
&= \mathbb{E}_{xy}D(\pi_1|_{xy}||\pi_1|_y) + \mathbb{E}_{xy}D(\pi_3|_{xy}||\pi_3|_y) + \ldots
\end{aligned}$$

Thus expected # Bob's mistakes $\leq O(\sqrt{CI})$ and similarly expected # Alice's mistakes $\leq O(\sqrt{CI})$. Hence we can compress to $O(\sqrt{CI}polylog(C))$ ■

It might be tempting to apply this compression recursively, but Alice and Bob are conveying a lot more information to each other in the new protocol (they are conveying where their estimates differ), and it is not clear how to bound the information cost of this new protocol.

References

[BBCR10] Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to compress interactive communication. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing*, 2010.

[BR10] Mark Braverman and Anup Rao. Information equals amortized communication. *CoRR*, abs/1106.3595, 2010.

[Bra11] Mark Braverman. Interactive information complexity. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:123, 2011.

[BYJKS04] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004.

[Hol07] Thomas Holenstein. Parallel repetition: Simplifications and the no-signaling case. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, 2007.

[KS92] Bala Kalyanasundaram and Georg Schnitger. The probabilistic communication complexity of set intersection. *SIAM Journal on Discrete Mathematics*, 5(4):545–557, November 1992.

[Raz92] Alexander Razborov. On the distributed complexity of disjointness. *TCS: Theoretical Computer Science*, 106, 1992.