CSE533: Information Theory in Computer Science October 25 and 27, 2010

Lectures 8 and 9 - Lower Bounds in Communication Complexity

Lecturer: Anup Rao *Scribes: Thach Nguyen and Abhay Jha*

1 Communication Complexity

Communication complexity studies how much information two or more parties must exchange in order to compute a function when each party only has partial knowledge of the input. In these two lectures, we use some of the tools we have learned to prove a lower bound on the randomized communication complexity of set disjointness, and show that multiple copies of a function must require more randomized communication than a single copy.

Communication protocols are a useful abstraction that can be applied to almost any model of computation. For example, a circuit with W wires naturally defines a communication protocol as follows: partition the gates and inputs of the circuit into two parts, with one part containing half the inputs and the other part containing the other half. Then the two parts of the circuit can be thought of as two communicating parties that send at most W messages in order to compute the output of the circuit from the inputs that they know. Thus, one can hope to give bounds on the value of W by bounding how much communication is needed to compute the function. Similarly, consider an algorithm that uses a small amount of memory to process a large stream of input data. Such an algorithm gives a communication protocol for the same computation where one party knows the first half of the inputs, the other knows the second half, and the amount of communication is equal to the memory used after reading the first half of the input. This connection can be used to give lower bounds on the memory usage of streaming computations.

Let us start by formally defining communication protocols (see [5] for more details). Two parties Alice and Bob wish to compute some function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$. Alice receives $x \in \mathcal{X}$, Bob receives $y \in \mathcal{Y}$ and they send messages back and forth to each other until both parties know $f(x, y)$. We are only interested in the number of bits exchanged between the two, thus we will allow Alice and Bob to have unbounded computational power. At any point, the protocol determines the active party and what he/she should send, which only depends on his/her input, the past messages, and possibly, any randomness used.

1.1 Deterministic Communication Protocols

A (deterministic) *protocol* π over domain $\mathcal{X} \times \mathcal{Y}$ with range $\mathcal{Z}$ is a binary tree where each internal node v is labeled either by a function $a_v : \mathcal{X} \to \{0, 1\}$ or by a function $b_v : \mathcal{Y} \to \{0, 1\}$, and each leaf is labeled with an element $z \in \mathcal{Z}$. The *value of the protocol* π on input (x, y), is the label of the leaf reached by starting from the root, and walking on the tree. At each internal node v labeled by a_v, the walk goes left if $a_v(x) = 0$ and right if $a_v(x) = 1$, and at each internal node labeled by b_v the walk goes left if $b_v(y) = 0$ and right if $b_v(y) = 1$. We say that π *computes* f if its value on input (x, y) is $f(x, y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

If the protocol reaches a node v labeled by a function a_v, it is Alice's turn to speak and the bit she should send is $a_v(x)$. Similarly, if v is labeled by a function b_v then it's Bob's turn to speak

and the bit he should send is $b_v(y)$. We write $\pi(x, y)$ to denote the concatenation of all messages transmitted in this way.

The *communication complexity of* π is the depth of the deepest leaf in π. The *deterministic communication complexity* of f, denoted by $D(f)$, is the minimum communication complexity of π, over all protocols π that compute f.

1.2 Randomized Communication Protocols

Here Alice and Bob are equipped with the power of randomization. In a *public coin randomized protocol* π, Alice and Bob have access to *the same* random string r, and each fixed r determines a deterministic protocol π_r, and on input x, y, Alice and Bob together sample r and execute the protocol π_r. We write $\pi(x, y)$ to denote the messages exchanged during the protocol.

We say that π computes the function f with if for all (x, y), π outputs the correct value $f(x, y)$ with probability at least $2/3$, where the probability is taken over r. The communication complexity of the protocol is the largest communication complexity of all the deterministic protocols involved. The *public coin randomized complexity* of a function f, denoted by $R(f)$, is defined as the minimum communication complexity of a protocol P that computes f.

In a *private coin randomized protocol* Alice and Bob are allowed to sampled random strings r_1, r_2 privately, and then run a deterministic protocol on the inputs $(x, r_1), (y, r_2)$. As above, the protocol computes f if for every x, y, it computes $f(x, y)$ with probability at least $2/3$, over the choice of r_1, r_2. The communication complexity is simply the communication complexity of π.

We note that the number $2/3$ above (the success probability) can be made an arbitrary large constant just by running the protocol several times and taking the majority of the outcomes, so it is not so important what number we use here, as long as it is larger than $1/2$. Every private coin protocol can be simulated by a public coin protocol with no increase in the communication, simply by setting $r = r_1, r_2$ to be the public randomness.

1.3 Average Case Communication Complexity

We shall also consider the model where there is a distribution μ on the inputs X, Y. In this case, we say that a protocol computes f if it outputs the correct answer with probability at least $2/3$, where the probability is over the sample of X, Y (in addition to any randomness used). We write $D^{\mu}(f)$ to denote the communication complexity of computing f in this model.

It is an easy consequence of von Neumann's minimax theorem to prove Yao's minimax theorem:

Theorem 1. $R(f) = \max_{\mu} D^{\mu}(f)$.

The trivial direction is that $R(f) \geq \max_{\mu} D^{\mu}(f)$. The other direction can be proved by considering the zero-sum game where one player picks a distribution on inputs, and the other player picks a protocol of communication complexity at most t, and the payoff is the error encountered by the players. The minimax theorem for zero-sum games can then be used to prove the above theorem (we leave the rest of the proof as an exercise).

We will use this method to prove Theorem 2.

2 Communication Complexity of Set Disjointness

We are interested in the following function $\mathsf{DIST} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$. DIST takes two inputs $x, y \in \{0,1\}^n$ and viewing them as the indicator vectors of two subsets $x, y \subseteq [n]$, outputs whether those sets are disjoint or not. The trivial way to compute DIST (or for that matter, any function) takes $n+1$ bits. Kalyanasundaram and Schnitger showed that this is the best one can do, upto constant factors:

Theorem 2 (Kalyanasundaram and Schnitger [4])**.** $R(\mathsf{DIST}) = \Omega(n)$

The theorem was first proved by Kalyanasundaram and Schnitger. Before their work, the disjointness function stood out as a function that did not succumb to several standard methods of proving lower bounds in communication complexity. This theorem is now the starting point for many other lower bound results in complexity theory. The proof of the theorem we present here uses techniques refined and developed in a sequence of papers [8, 7, 1, 3, 6, 2].

To prove the theorem, we shall define a distribution μ over the inputs such that $D^{\mu}(\mathsf{DIST}) = \Omega(n)$. Let μ_0 be the distribution that is uniform on disjoint sets. Thus, in μ_0, (X_i, Y_i) is equally likely to be $(0,0), (0,1)$ or $(1,0)$, and the n pairs of variables are independent of each other. Let μ_1 be the distribution that is uniform over sets which intersect in exactly one coordinate. Thus, for a random coordinate j, $(X_j, Y_j) = (1,1)$, and the rest of the coordinates are again independent and equally likely to be $(0,0), (0,1)$ or $(1,0)$. Set $\mu = (\mu_0 + \mu_1)/2$ to be the convex combination of these two distributions. μ produces disjoint sets with probability exactly $1/2$.

Recall that the statistical distance between two distributions α, β sampling from a set U is defined to be $(1/2)\sum_{x \in U} |\alpha(x) - \beta(x)| = \max_{T \subset U}(\alpha(T) - \beta(T))$, namely the largest probability that a set T has of distinguishing one distribution from the other. We write $\alpha \overset{\epsilon}{\approx} \beta$ to assert that α is ϵ-close to β in statistical distance.

Given any deterministic protocol π, we write π_{μ}, π_{μ_0} and π_{μ_1} to denote the distributions of the messages in π when the inputs are sampled according to μ, μ_0, μ_1. We shall prove that any protocol with small communication cannot tell μ_0 apart from μ_1:

Theorem 3. *If the communication complexity of π is ϵn, then $\pi_{\mu_0} \overset{O(\epsilon^{1/4})}{\approx} \pi_{\mu_1}$.*

Theorem 3 implies Theorem 2, since it shows that for any protocol π with communication complexity ϵn (for small enough ϵ), π cannot compute DIST correctly on the distribution μ, since the probability that it concludes that the sets are intersecting is roughly the same whether the input comes from μ_0 or μ_1.

Intuitively the theorem should hold because the parties do not know where the potential intersection is, and so do not know which coordinates to reveal information about. On average, they can only spend ϵ bits of communication per coordinate, so they must be ignoring a lot of coordinates, and whenever the intersection lies in one of the coordinates that they do not pay attention to, the messages in the protocol will carry no information about whether the inputs are intersecting or not. Actually turning this intuition into a proof is tricky. One immediate complication is that in μ all coordinates are *not* independent of each other, so transmitting information about one coordinate does reveal information about the other coordinates. On the other hand μ_0 is a product distribution, so let us try to start by applying our intuition to μ_0.

First Attempt Let X, Y be sampled according to μ_0. Let $\pi(X, Y)$ denote the messages transmitted in π for inputs X, Y. Recall that

Lemma 4. *If A and B are independent then*

$$I(AB \wedge C) \geq I(A \wedge C) + I(B \wedge C).$$

Since each of the coordinates in the inputs are independent of each other,

$$\epsilon n \geq I\left(\begin{array}{c} X_1, X_2 \ldots X_n \\ Y_1, Y_2 \ldots Y_n \end{array} \wedge \pi(X, Y)\right) \geq \sum_{i=1}^{n} I\left(\begin{array}{c} X_i \\ Y_i \end{array} \wedge \pi(X, Y)\right) \tag{1}$$

Thus, the messages only have ϵ information about the average coordinate, at least under the distribution μ_0. We also have the following lemma (that we do not prove here):

Lemma 5. $D(p||q) \geq |p - q|^2$

Note that there should not be any relationship in the reverse direction. If $q(x) = 0$ and $p(x) \neq 0$ for some $x \in X$ then $D(p||q)$ is ∞ while $|p - q|$ is always at most 1.

Lemma 5 allows us to show that if two random variables have small mutual information, then they are statistically close to being independent:

Lemma 6. *Let A and B be two random variables with distribution $p(a, b)$, then*

$$I(A \wedge B) \geq \mathop{\mathbb{E}}_{a}\left[|p(b|a) - p(b)|^2\right] \geq \left(\mathop{\mathbb{E}}_{a}\left[|p(b|a) - p(b)|\right]\right)^2.$$

Proof The second inequality follows from convexity. Here we prove the first inequality

$$\begin{aligned} I(A \wedge B) &= D(p(a, b)||p(a)p(b)) \\ &= \sum_{a,b} p(a, b) \log \frac{p(a, b)}{p(a)p(b)} \\ &= \sum_{a} p(a) D(p(b|a)||p(b)) \\ &\geq \sum_{a} p(a)|p(b|a) - b(b)|^2 \end{aligned}$$

The last inequality implies the lemma. ■

Consider the following protocol γ that takes as input bits a, b:

Protocol $\gamma(a, b)$

1. The parties publicly sample $j \in [n]$ uniformly at random.
2. For $i \neq j$, the parties publicly sample (X_i, Y_i) to be uniformly one of $(0, 0), (1, 0), (0, 1)$.
3. They set $X_j = a, Y_j = b$.
4. The parties execute $\pi(X, Y)$.

Let us write $\gamma(a,b)$ to denote the distribution of messages exchanged in γ on input a,b. Observe that if A,B are uniformly set to be bits not equal to $(1,1)$, $\gamma(A,B)$ has the same distribution as π_{μ_0}. On the other hand, $\gamma(1,1)$ has the same distribution as π_{μ_1}. By Equation (1), $I(AB \wedge \pi(X,Y)) \leq \epsilon$, so by Lemma 6, we can argue that the distribution of $\gamma(A,B)$ is essentially independent of (A,B): $\mathbb{E}_{A,B}\left[|p(\pi(x,y)|ab) - p(\pi(x,y))|\right] \leq \sqrt{\epsilon}$. Does this imply that $\gamma(1,1)$ has roughly the same distribution as $\gamma(A,B)$? One might hope that the following type of conjecture is true:

Conjecture 7. *If $\mathbb{E}_{A,B}\left[|p(\pi(x,y)|ab) - p(\pi(x,y))|\right]$ is small, then $|p(\pi(x,y)|a{=}1,b{=}1) - p(\pi(x,y))|$ must also be small.*

Indeed the conjecture is true when γ is forced to be a deterministic protocol — if $\gamma(0,0) = \gamma(0,1) = \gamma(1,0)$, the first bit of $\gamma(1,1)$ is the same as the first bit of $\gamma(1,0)$ or the first bit of $\gamma(0,1)$ (which are both the same), depending on who transmits the first bit. Similarly, the second bit must be the same, and so on: the entire communication of $\gamma(1,1)$ must be the same as in the other cases.

However, the conjecture is not true for general randomized protocols. Consider the following protocol: the parties publicly sample a bit r, and the first party transmits the parity $a \oplus r$ to the second party. The second party then transmits $a \wedge b$. In this example, $\gamma(0,0) = \gamma(1,0) = \gamma(0,1)$, yet $\gamma(1,1)$ can be distinguished from the others with probability 1. The example shows that the public randomness can allow the parties to communicate even though this communication is not visible in the messages alone. We cannot (yet) rule out that this kind of thing happens in the γ we defined above. X_i, Y_i need to be correlated, so they may also be providing a secret communication channel when π is executed.

Still, one can show that the conjecture *is* true if γ is a private coin protocol:

Lemma 8. *If γ is a private coin protocol taking bits a,b as inputs, and $\gamma(0,1) \overset{\epsilon}{\approx} \gamma(1,0)$, then $\gamma(0,0) \overset{O(\sqrt{\epsilon})}{\approx} \gamma(1,1)$.*

The proof is based on the Hellinger distance between two distributions: $H(p,q) = \sqrt{1 - \sum_x \sqrt{p(x)q(x)}}$. The following lemmas (which we do not prove here) describe the relationship the Hellinger distance and the statistical distance:

Lemma 9. $\sqrt{2}H(p,q) \geq |p-q| \geq H^2(p,q)$

Proof of Lemma 8 We claim that $H(\gamma(1,1),\gamma(0,0)) = H(\gamma(1,0),\gamma(0,1))$. The proof is then immediate by applying Lemma 9.

The proof of the claim is very similar to the case when γ was deterministic.

Given any fixed sequence of messages $m = m_1, \ldots, m_k$, if the first party transmits the first bit in γ, then $\gamma(1,1)(m_1) = \gamma(1,0)(m_1)$, and $\gamma(0,0)(m_1) = \gamma(0,1)(m_1)$. If the second party speaks first, $\gamma(1,1)(m_1) = \gamma(0,1)(m_1)$, and $\gamma(0,0)(m_1) = \gamma(1,0)(m_1)$. In either case, $\gamma(1,1)(m_1) \cdot \gamma(0,0)(m_1) = \gamma(1,0)(m_1) \cdot \gamma(0,1)(m_1)$. Proceeding inductively, we can show that

$$\gamma(1,1)(m) \cdot \gamma(0,0)(m) = \gamma(1,0)(m) \cdot \gamma(0,1)(m),$$

which proves that $H(\gamma(1,1),\gamma(0,0)) = H(\gamma(1,0),\gamma(0,1))$. ■

Given Lemma 8, our strategy is going to be to try and break the dependence between X,Y to get the contradiction we need. To do this, we shall introduce some additional random variables.

Let $V = V_1, V_2, \ldots, V_n$ be uniformly chosen bits. Define

$$T_i = \begin{cases} X_i & V_i = 0 \\ Y_i & \text{else} \end{cases}$$

$$W = W_1, \ldots, W_n = \begin{pmatrix} T_1 T_2 \ldots T_n \\ V_1 V_2 \ldots V_n \end{pmatrix}$$

We write W_{-i} to denote all coordinates of W except for the i'th one.

Then, we can bound

$$\begin{aligned} \epsilon n &\geq I(XY \wedge \pi \mid W) \\ &\geq \sum_{i=1}^{n} I(X_i Y_i \wedge \pi \mid W) \\ \Rightarrow \epsilon &\geq (1/n) \sum_{i=1}^{n} \mathop{\mathbb{E}}_{W_{-i}} \left[\left(I(X_i Y_i \wedge \pi \mid W_{-i} X_i) + I(X_i Y_i \wedge \pi \mid W_{-i} Y_i) \right) / 2 \right] \end{aligned} \tag{2}$$

In particular, Equation 6 implies that

$$2\epsilon \geq (1/n) \sum_{i=1}^{n} \mathop{\mathbb{E}}_{W_{-i}} \left[I(X_i Y_i \wedge \pi \mid W_{-i} Y_i) \right]$$

Then by Lemma 6, we get

$$\begin{aligned} 2\epsilon &\geq (1/n) \sum_{i=1}^{n} \mathop{\mathbb{E}}_{W_{-i}} \left[\mathop{\mathbb{E}}_{Y_i, \pi} \left[|p(x_i y_i | \pi W_{-i} Y_i) - p(x_i y_i | W_{-i} Y_i)|^2 \right] \right] \\ \Rightarrow 2\epsilon &\geq \left((1/n) \sum_{i=1}^{n} \mathop{\mathbb{E}}_{W_{-i}, Y_i, \pi} \left[|p(x_i y_i | \pi W_{-i} Y_i) - p(x_i y_i | W_{-i}, Y_i)| \right] \right)^2, \end{aligned}$$

where the last inequality follows from the convexity of the square function.

In other words, we have proved that for an average coordinate, the distribution of X_i, Y_i is essentially independent of the messages π. Let i be a uniformly random coordinate. Then we have showed that:

$$\begin{aligned} p(i x_i y_i w_{-i} \pi) &\overset{\sqrt{2\epsilon}}{\approx} p(i y_i w_{-i} \pi) \cdot p(x_i | i y_i w_{-i}) \\ &= p(i y_i w_{-i}) \cdot p(\pi | i y_i w_{-i}) \cdot p(x_i | i y_i w_{-i}) \\ \Rightarrow p(i x_i y_i w_{-i} \pi) &\overset{\sqrt{2\epsilon}}{\approx} p(i x_i y_i w_{-i}) \cdot p(\pi | i y_i w_{-i}) \end{aligned} \tag{3}$$

Now consider the following protocol γ:

Protocol $\gamma(a,b)$

1. The parties publicly sample $i \in [n]$ uniformly at random.
2. The parties publicly sample W_{-i}.
3. They set $X_i = a, Y_i = b$.
4. They each privately sample the rest of X, Y conditioned on the inputs that they already sampled.
5. The parties execute $\pi(X,Y)$.

As before, let A, B be uniform bits that are not both 1. Then by Equation 3,

$$p(iabw_{-i}\gamma) \overset{\sqrt{2\epsilon}}{\approx} p(iabw_{-i}) \cdot p(\gamma|ibw_{-i})$$

In particular, since $\Pr[(A,B)=(0,0)] = 1/3$, we get

$$p(iw_{-i}\gamma(0,0)) \overset{3\sqrt{2\epsilon}}{\approx} p(iw_{-i}) \cdot p(\gamma|iw_{-i}, b{=}0)$$

Repeating the same argument shows that

$$p(iw_{-i}\gamma(1,0)) \overset{3\sqrt{2\epsilon}}{\approx} p(iw_{-i}) \cdot p(\gamma|iw_{-i}, b{=}0)$$

We can conclude that

$$p(iw_{-i}\gamma(1,0)) \overset{6\sqrt{2\epsilon}}{\approx} p(iw_{-i}\gamma(0,0)) \overset{6\sqrt{2\epsilon}}{\approx} p(iw_{-i}\gamma(0,1)) \tag{4}$$

Note that the messages in $\gamma(1,1)$ are distributed exactly as in π_{μ_1}, and $\gamma(A,B)$ is distributed exactly as in π_{μ_0}, which by Equation 4 is close to each of $\gamma(1,0), \gamma(0,0), \gamma(0,1)$.

By Lemma 9, these equations (together with the convexity of the square function, imply that

$$\begin{aligned}
&|\gamma(1,1) - \gamma(0,0)|^2 \\
&\leq \left(\mathop{\mathbb{E}}_{i,W_{-i}} [|p(\gamma(1,1)|iW_{-i}) - p(\gamma(0,0)|iW_{-i})|] \right)^2 \\
&\leq \mathop{\mathbb{E}}_{i,W_{-i}} \left[|p(\gamma(1,1)|iW_{-i}) - p(\gamma(0,0)|iW_{-i})|^2 \right] \\
&\leq O\left(\mathop{\mathbb{E}}_{i,W_{-i}} [|p(\gamma(1,0)|iW_{-i}) - p(\gamma(0,1)|iW_{-i})|] \right) \\
&\leq O(\sqrt{\epsilon})
\end{aligned}$$

This proves Theorem 3.

3 Direct Sums and Protocol Compression

Given a function $f(x, y)$, define

$$f^n \left(\begin{array}{c} X_1 X_2 \dots X_n \\ Y_1 Y_2 \dots Y_n \end{array} \right) = f(X_1, Y_1) f(X_2, Y_2) \dots f(X_n, Y_n),$$

namely the function that outputs the concatenation of the outputs from each f.

It is easy to see that $\log(n) \cdot n \cdot R(f) \geq R(f^n)$ — to compute n copies, we can use the best protocol for computing one copy, n times. The log term is needed to reduce the error in each copy to guarantee that all the answers are correct with probability $1 - O(1/n)$. Can we lower bound $R(f^n)$ by $n \cdot R(f)$? It turns out that some kind of relationship in the other direction is true, as was proved by Barak, Braverman, Chen and Rao:

Theorem 10 ([2])**.**

$$R(f^n) \geq \Omega \left(\frac{\sqrt{n} \cdot R(f)}{\log(R(f^n))} \right)$$

To prove the theorem, it suffices to prove a distributional version. Let μ^n denote the product distribution obtained by taking n independent copies of the distribution μ. Then,

Theorem 11 ([2])**.**

$$D^{\mu^n}(f^n) \geq \Omega \left(\frac{\sqrt{n} \cdot D^{\mu}(f)}{\log(D^{\mu^n}(f^n))} \right)$$

Theorem 10 follows from Theorem 11 — by Yao's min-max theorem, there is a distribution μ for which $D^{\mu}(f) = R(f)$, then by Theorem 11, $D^{\mu^n}(f^n) \geq \Omega \left(\frac{\sqrt{n} \cdot R(f)}{\log(R(f^n))} \right)$, which implies that $R(f^n) \geq \Omega \left(\frac{\sqrt{n} \cdot R(f)}{\log(R(f^n))} \right)$.

Theorem 11 is proved is by reduction. We start with a protocol realizing $D^{\mu^n}(f)$, and use it to get a protocol with low communication complexity ($O(D^{\mu^n}(f^n) \cdot \log(D^{\mu^n}(f))/\sqrt{n})$) that computes f over inputs drawn from μ. This would prove that $D^{\mu^n}(f) \cdot \log(D^{\mu^n}(f))/\sqrt{n} \geq \Omega(D^{\mu}(f))$ as required.

First Attempt Let π be a protocol realizing $D^{\mu^n}(f^n)$, and let

$$\begin{array}{c} X^n \\ Y^n \end{array} = \begin{array}{ccccc} X_1 & X_2 & \cdots & X_n \\ Y_1 & Y_2 & & Y_n \end{array}$$

be sampled according to μ^n. Then, since each coordinate is independent, we have

$$D^{\mu^n}(f^n) \geq I(X^n Y^n \wedge \pi) \geq \sum_{i=1}^{n} I \left(\begin{array}{c} X_i \\ Y_i \end{array} \wedge \pi \right) \tag{5}$$

Thus, the average coordinate has only $D^{\mu^n}(f^n)/n$ information about the messages in π. Consider the following protocol for f:

Protocol $\gamma(x,y)$:

1. Publicly sample a coordinate $i \in [n]$ uniformly.
2. For $j \neq i$, publicly sample X_j, Y_j according to μ.
3. Set $(X_i, Y_i) = (x, y)$.
4. Run π.

If X, Y are distributed according to μ, note that $\gamma(X,Y)$ is distributed exactly according to $\pi(X^n, Y^n)$. Thus, $I(XY \wedge \gamma) \leq D^{\mu^n}(f^n)/n$. One might hope that we can now try to compress the communication in γ so that the total length of messages in γ is close to the information in them. Unfortunately, this is an impossible task! The counterexample is the same as before, namely the public randomness can be used to communicate secretly. Consider the following protocol, that operates on n-bit strings a, b:

Protocol $\gamma'(a,b)$:

1. Publicly sample a random n-bit string r.
2. The first party transmits the bitwise parity of her input with r to the second party.
3. the second party transmits $f(a,b)$.

Call the first message of the protocol M_1 and the second M_2. Then observe that M_1 has no information about the inputs. Thus for any distribution on inputs, $I(AB \wedge \gamma') = I(AB \wedge M_1) + I(AB \wedge M_2 | M_1) \leq 1$. And yet this protocol computes f. Thus, the mutual information between the messages and the inputs is not a good measure of how much information the protocol reveals.

As in the proof of the lower bound for the disjointness function, the fix is to try and break the dependence with the public randomness. Define $V = V_1, V_2, \ldots, V_n$ be uniformly chosen bits. Define

$$T_i = \begin{cases} X_i & V_i = 0 \\ Y_i & \text{else} \end{cases}$$

$$W = W_1, \ldots, W_n = \begin{pmatrix} T_1 T_2 \ldots T_n \\ V_1 V_2 \ldots V_n \end{pmatrix}$$

We write W_{-i} to denote all coordinates of W except for the i'th one.

We conclude that

$$\begin{aligned} D^{\mu^n}(f) &\geq I(X^n Y^n \wedge \pi \mid W) \\ &\geq \sum_{i=1}^{n} I(X_i Y_i \wedge \pi \mid W) \\ \Rightarrow D^{\mu^n}(f)/n &\geq (1/n) \sum_{i=1}^{n} \mathop{\mathbb{E}}_{W_{-i}} \left[\left(I(X_i Y_i \wedge \pi \mid W_{-i} X_i) + I(X_i Y_i \wedge \pi \mid W_{-i} Y_i) \right)/2 \right] \end{aligned} \tag{6}$$

Then consider the following protocol for computing f:

Protocol $\gamma(x,y)$:

1. Publicly sample a coordinate $i \in [n]$ uniformly.
2. Publicly sample W_{-i} according.
3. Set $(X_i, Y_i) = (x, y)$.
4. Privately sample X^n, Y^n conditioned on all the publicly sampled variables.
5. Run π.

Then we have argued that in γ, if R denotes the public randomness, we have that $2D^{\mu^n}(f)/n \geq I(XY \wedge \gamma|RY) + I(XY \wedge \gamma|RX)$. Most of the work in [2] then goes into showing that any protocol with small information in this sense can be simulated with small communication. Intuitively the measure of information that's being bounded here is the amount of information that the each party learns about the other parties input, based on everything that they see (Note that $I(XY \wedge \gamma|RY) = I(X \wedge \gamma R|Y)$).

They prove:

Theorem 12 ([2]). *If π is a protocol on inputs XY with communication C, then π can be simulated with expected communication* $\log C \cdot \sqrt{C \cdot (I(XY \wedge \pi|RX) + I(XY \wedge \pi|RY))}$.

Applying the compression to the protocol γ above gives a protocol for f with communication $\log D^{\mu^n}(f) \cdot \sqrt{D^{\mu^n}(f) \cdot D^{\mu^n}(f)/n} = D^{\mu^n}(f) \cdot \log D^{\mu^n}(f)/\sqrt{n}$.

References

[1] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci*, 68(4):702–732, 2004.

[2] Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to compress interactive communication. In Leonard J. Schulman, editor, *STOC*, pages 67–76. ACM, 2010.

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

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

[5] Eyal Kushilevitz and Noam Nisan. *Communication complexity*. Cambridge University Press, Cambridge, 1997.

[6] Anup Rao. Parallel repetition in projection games and a concentration bound. In Cynthia Dwork, editor, *STOC*, pages 1–10. ACM, 2008.

[7] Ran Raz. A parallel repetition theorem. *SIAM Journal on Computing*, 27(3):763–803, June 1998.

[8] Alexander A. Razborov. On the distributional complexity of disjointness. *Theor. Comput. Sci*, 106(2):385–390, 1992.