CS 2429 - Foundations of Communication Complexity

Information Complexity and Set Disjointness

Lecturer: Robert Robere

1 Information Theory

In this section we discuss the information-theoretic notions that we use later. Following much of the literature on information complexity the calligraphic letters $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$ will be used to denote sets (often supports of distributions), capital letters $X, Y, Z, \ldots$ to denote random variables, and lowercase letters $x, y, z, \ldots$ to actual values. Another useful convention is as follows: if we sample a value using a lower-case letter x from some distribution then we use the corresponding capital letter X to denote the associated random variable. With this convention in mind we will use $p(x)$ to denote $P[X = x]$, and other standard probabilistic constructions similarly (i.e. $p(x|y)$ for $\Pr[X = x|Y = y]$, etc.) Furthermore, all distributions that we consider are over finite sets.

The first information-theoretic quantity that we consider is the *entropy* of a (real-valued) random variable X, defined to be

$$H(X) := \mathbb{E}\left[\log \frac{1}{p(x)}\right] = \sum_x p(x) \log \frac{1}{p(x)},$$

where we take the convention that $0 \log 1/0 = 0$ (which can be justified for our application since $x \log 1/x \to 0$ as $x \to 0$ from the right). The entropy of a random variable X is a measure of our *uncertainty* of the value that X takes: if X has high entropy then we have essentially no ability to predict the value that X will take. Another, equivalent view says that the entropy measures[1] the *average number of bits* needed to store the result of sampling the random variable X. Under this interpretation, intuition suggests that the uniform distribution over a finite set $[n]$ should have maximal entropy, and indeed it does: if X samples from $[n]$ uniformly at random, then

$$H(X) = \sum_{i=1}^{n} n^{-1} \log(n^{-1}) = \log n,$$

which makes sense, since informally we should require $\lceil \log n \rceil$ bits to store a sample of X. If the distribution is skewed away from uniform, however, then the entropy decreases since we could get away with encoding the "uncommon" values with longer strings. For example, suppose that X is a random variable sampled from the following simple distribution on $[2^n + 1]$[2]: the value $2^n + 1$ appears with probability $1/2$, and for all $i \in [2^n + 1]$ other than $2^n + 1$, $p(i) = 1/2^{n+1}$. This is

[1] "measures" is an important qualifier here, as it is not exact

[2] we choose 2^n here instead of n to give an actual encoding that matches the entropy.

certainly a distribution since $2^n/2^{n+1} = 1/2$, and so calculating the entropy we get

$$H(X) = \sum_{i=1}^{2^n+1} p(i) \log p(i)^{-1} = \frac{1}{2} + \frac{1}{2^{n+1}} \log 2^{n+1} = \frac{1}{2} + \frac{n+1}{2^{n+1}}.$$

If we encode $2^n + 1$ as the string 0, and all other strings i as the binary expansion of i pre-pended with 1, we get that the average number of bits needed to store the random variable X is exactly the entropy $H(X)$.

If we consider a joint distribution (X, Y) for *independent* random variables then the entropy of the resulting random variable has the following nice property:

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

This suggests that there should perhaps exist a *conditional* entropy, and indeed there is: the following definition is the "right" one:

$$H(X|Y) = \mathop{\mathbb{E}}_{y}[H(X|Y = y)].$$

That is, the *conditional* entropy $H(X|Y)$ is the *average* entropy "left" in X after sampling Y. Much like entropy can be interpreted as the average number of bits needed to store the result of sampling X, we can interpret the conditional entropy as the average number of bits needed to store X if we already have a sample of Y. Conditional entropy generalizes the nice property above as you would expect:

$$H(X, Y) = H(X) + H(Y|X), \tag{1}$$

and this fact is known as the *chain rule*.

Now, in perfect analogy to Bayes rule, we can write the joint entropy in two different ways:

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

and re-arrange it to obtain

$$H(X) - H(X|Y) = H(Y) - H(Y|X).$$

This quantity is so important that it receives its own name: *mutual information*, and it is denoted by

$$I(X; Y) := H(X) - H(X|Y) = H(Y) - H(Y|X). \tag{2}$$

Intuitively, $I(X; Y)$ measures how much information is shared by X and Y: or, said another way, how much uncertainty is removed from X after we have learned the variable Y (on average). We can similarly define the *conditional mutual information*, denoted $I(X; Y|Z)$, where

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

which has a similar interpretation to the mutual information.

The next proposition records a number of properties of these information theoretic quantities:

Proposition 1 *Let W, X, Y, Z be random variables.*

1. *If X takes on at most s values then $H(X) \leq \log s$.*
2. *$I(X;Y) \geq 0$, or equivalently, $H(X|Y) \leq H(X)$ (conditioning can only reduce uncertainty).*
3. *$H(X,Y|Z) \leq H(X|Z) + H(Y|Z)$ (subadditivity of conditional entropy).*
4. *$I(XY;Z) = I(X;Z) + I(Y;Z|X)$ (chain rule for mutual information).*
5. *If Y and X are independent given Z and W then*

$$I(XY;Z|W) \geq I(X;Z|W) + I(Y;Z|W)$$

   (superadditivity of mutual information).

6. *If X and Z are conditionally independent given Y, then $I(X;Y|Z) \leq I(X;Y)$ (data-processing inequality)*

So you can see that they are algebraically well-behaved, and also strongly correspond to our intuition about how the quantities "should" work.

2 Communication Protocols and Information

Fundamentally, communication protocols are about an exchange of information: two parties wish to communicate the minimum amount of information about their inputs in order to solve the task at hand. It should therefore be no surprise that information theoretic tools are very useful in this setting. We focus on randomized and distributional communication complexity: for any $\delta > 0$ and any function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, let $R_\delta(f)$ be the randomized communication complexity of f over protocols that error on at most a δ-fraction of the inputs. Similarly, if μ is a distribution on $\mathcal{X} \times \mathcal{Y}$ let $R^\mu_\delta(f)$ denote the δ-error distributional communication complexity of f, where the inputs of the communication protocol are chosen from the distribution μ. If Π is a communication protocol for f then let $\Pi(x,y)$ denote the *transcript* between the two players on input x, y, which is the string containing the shared random bits used in the protocol and all messages exchanged between the two players. For any fixed input x, y the *communication complexity of the transcript* will be the length of $\Pi(x,y)$ — denoted $|\Pi(x,y)|$ — and we denote by $\mathsf{CC}(\Pi)$ the maximum communication complexity of Π over all inputs.

At the opening of this section it was stated that communication could be thought of as parties wanting to reveal the minimum amount of information[3] about their inputs in order to solve some pre-determined task. The next definition formalizes this in an information-theoretic sense.

Definition Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a function, μ a distribution on $\mathcal{X} \times \mathcal{Y}$ and $\delta > 0$. If Π is a protocol computing f with δ-error on the distribution μ, define the *external information complexity* of Π to be

$$IC^{ext}_{\mu,\delta}(\Pi) = I(XY;\Pi(X,Y)),$$

where the shared randomness of Π is part of the transcript (note that $\Pi(X,Y)$ is a random variable depending on μ and the shared randomness). Define the *internal information complexity* to be

$$IC_{\mu,\delta}(\Pi) = I(X;\Pi(X,Y)|Y) + I(Y;\Pi(X,Y)|X).$$

[3]Here things like zero-knowledge proofs is being swept under the rug.

The two definitions are necessary to capture a certain ambiguity in our statement before the definition: namely, who are the parties "revealing" their information too? The external information complexity measures the amount of information revealed about the inputs X, Y to an external observer who only observes the transcript of the protocol $\Pi(X, Y)$; on the other hand, the internal information complexity measures the amount of information revealed to each player about the other player's input: the first term measures how much information is revealed by the protocol $\Pi(X, Y)$ about the input X, knowing the value of Y, and vice-versa for the second term. Intuitively both of these measures should be lower bounds on the communication complexity of the protocol Π, and once again the intuition about information lines up with what we can prove formally:

Proposition 2 *Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a function, μ any distribution over $\mathcal{X} \times \mathcal{Y}$, and $\delta > 0$ some real number. Let Π be a deterministic protocol computing f with δ-error on inputs drawn from the distribution μ. Then $\mathsf{CC}(\Pi) \geq \mathsf{IC}^{ext}_{\mu,\delta}(\Pi) \geq \mathsf{IC}_{\mu,\delta}(\Pi)$.*

Proof The first inequality follows from a quick calculation

$$\mathsf{CC}(\Pi) \geq \max_{(x,y)} |\Pi(x, y)| \geq H(\Pi(X, Y)) \geq I(XY; \Pi(X, Y)) = \mathsf{IC}^{ext}_{\mu,\delta}(\Pi).$$

The second inequality requires a bit (but not much) more work. Intuitively, in each round only a single player can speak, and so only one player will learn any information about the other player's input (in the internal information sense). However an external observer will see all exchanged bits, and since he does not have either of the player's inputs he will learn some information in every round.

For any $i \in \{1, \ldots, CC(\Pi)\}$ let $\Pi(x, y)_i$ be the ith bit exchanged by the protocol on input x, y, and define $\Pi(x, y)_{\leq i}, \Pi(x, y)_{<i}$ analogously. Below let us write Π instead of $\Pi(X, Y)$ for the sake of brevity. At step i only one player spoke: assume it is the X player w.l.o.g. Clearly the amount of information learned by this player about Y is 0 for this round:

$$I(Y; \Pi_i | \Pi_{<i}, X) = 0.$$

By the chain rule for mutual information, the amount of information learned by an external observer is at least the maximum amount of information learned by either party:

$$I(XY; \Pi_i | \Pi_{<i}) \geq \max\{I(X; \Pi_i | Y, \Pi_{<i}), I(Y; \Pi_i | X, \Pi_{<i})\}.$$

Combining these two facts proves the second inequality, after applying the chain rule to the transcript repeatedly:

$$\begin{aligned}
\mathsf{IC}^{ext}_{\mu,\delta}(\Pi) = I(XY; \Pi) &= \sum_{i=1}^{CC(\Pi)} I(XY; \Pi_i | \Pi_{<i}) \\
&\geq \sum_{i=1}^{CC(\Pi)} \max\{I(X; \Pi_i | Y, \Pi_{<i}), I(Y; \Pi_i | X, \Pi_{<i})\} \\
&\geq \sum_{i=1}^{CC(\Pi)} I(X; \Pi_i | Y, \Pi_{<i}) + I(Y; \Pi_i | X, \Pi_{<i}) \\
&= I(X; \Pi | Y) + I(Y; \Pi | X) = \mathsf{IC}_{\mu,\delta}(\Pi),
\end{aligned}$$

where the last inequality follows since one of the terms is 0 for each i.

Now, if f is a function let[4]

$$\mathsf{IC}_{\mu,\delta}(f) = \inf_{\text{protocols } \Pi \text{ computing } f} \mathsf{IC}_{\mu,\delta}(\Pi).$$

Then the previous proposition implies

$$R_\delta(f) \geq \mathsf{IC}_{\mu,\delta}(f)$$

after applying Yao's minimax principle.

So, with information complexity we get a new quantity that lower bounds communication complexity. What exactly can we get that was out of reach before? There are several answers to this, and they are each inter-related:

Direct Sum/Direct Product Theorems Informally, a direct-sum question is one of the following form: does the complexity of computing n copies of a function f scale as n times the complexity of computing a single copy of f? Due to its nice algebraic properties, information complexity admits a direct sum theorem quite easily: the information revealed by a protocol computing n copies of f is at least n times the information revealed by computing a single copy of f. This direct sum property of information can be transformed into a direct-sum for communication.

Communication Lower Bounds Due to the direct-sum property above, information complexity is quite useful for studying the communication complexity of *composed* functions: that is, functions of the form $h(x,y) = f(g(x_1,y_1), g(x_2,y_2), \ldots, g(x_n,y_n))$. A notable example of such a function is the disjointness problem: $DISJ(x,y) = \bigvee_{i=1}^n x_i \wedge y_i$. For example, the direct-sum property of disjointness reduces the problem of communication lower bounds for $DISJ$ to information complexity lower bounds on the 2-bit $\wedge$ function. Using information complexity tools Bar-Yossef et al [2] were able to give a simplified proof of the $\Omega(n)$ lower bound on the randomized communication complexity of disjointness. Later, this was improved by Braverman, Garg, Pankratov, and Weinstein [4] to give a tight bound on the complexity of disjointness of the following form: for every $\varepsilon > 0$ there is a $\delta > 0$ such that $\delta \to 0$ as $\varepsilon \to 0$ and

$$(C_{DISJ} - \delta)n \leq R_\varepsilon(DISJ) \leq C_{DISJ} \cdot n + o(n),$$

where $C_{DISJ} \approx 0.4827$.

Protocol Compression Given a protocol Π with communication complexity C and information complexity I, does there exist a protocol Π' computing the same function but with communication complexity $O(I)$? Or, informally, can we compress an arbitrary communication protocol down to its information content? It turns out that we can get *some* non-trivial compression (although not all the way down to $O(I)$), and this is used to proved the direct sum theorem for communication alluded to above.

We will not touch on protocol compression in this lecture, but we will give the full argument of [2] showing an $\Omega(n)$ lower bound on the randomized complexity of disjointness. But first, let us discuss direct-sum results.

[4] Yes, that should be an inf and not a min. There can be infinitely many protocol computing f with the same information complexity, and actually you can have an infinite sequence of protocols $\Pi_1, \Pi_2, \ldots$ such that the information complexity of the protocols converges to $\mathsf{IC}_{\mu,\delta}(f)$ in the limit.

3 Direct Sum for Information

Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be a function, let μ be a distribution over $\mathcal{X} \times \mathcal{Y}$, and let $\delta > 0$ be some error parameter. Earlier it was stated that information complexity admits a direct-sum property, and in this section we will make this formal. Let f^n be the problem of solving n independent copies of f: the input to f^n is a sequence of inputs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \sim \mu^n$, and the output is $f(X_i, Y_i)$ for each i. In [3] the following was proven[5]:

Theorem 3 (Theorem 4.3 in [3]) *Let f be a two-party function, μ any distribution over the inputs of f, and $\delta > 0$ an error parameter. Then*

$$\mathsf{IC}_{\mu^n,\delta}(f^n) = n \cdot \mathsf{IC}_{\mu,\delta}(f).$$

In lieu of proving this, we instead prove a consequence of the above theorem that follows a similar proof. Working in the restricted setting of the next theorem will allow us to directly argue about communication protocols and be finished, rather than having to worry about issues related to convergence and other analytic problems when considering $\mathsf{IC}_{\mu,\delta}(f)$. The next result is important to digest, so we give two separate arguments: the first is a simple argument when the distribution μ is restricted to be a product distribution, and the second is a more general argument from [5].

Theorem 4 *Let f be a two-party function, μ any distribution over the inputs of f, and $\delta > 0$ an error parameter. There exists a protocol τ solving f over the distribution μ with at most δ-error such that*

$$R_\delta(f^n) \geq n\mathsf{IC}_{\mu,\delta}(\tau)$$

and the communication complexity of τ is at most $R_\delta(f^n)$.

Proof Note again that our protocols may have both public *and* private randomness. Let Π be a deterministic protocol computing f^n on μ^n with δ-error that attains communication complexity $R_\delta^{\mu^n}(f^n)$. This proof is a subtle, so first let us assume that μ is a product distribution (i.e. when sampling (X, Y) according to μ the values of X and Y are independent). The obvious protocol that springs to mind is the following: given an input (X, Y) distributed according to μ, Alice and Bob sample a uniformly random $j \in [n]$ using shared randomness and $n - 1$ inputs of their own using private randomness:

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_{j-1}, Y_{j-1}), (X_{j+1}, Y_{j+1}), \ldots, (X_n, Y_n).$$

Note that they can do the second sampling step precisely because μ is assumed to be a product distribution. Then they simulate the protocol Π on the sampled inputs, with (X, Y) plugged into the jth spot. Call this protocol τ'.

Clearly $\mathsf{CC}(\tau') = \mathsf{CC}(\Pi)$. In the calculations of the remainder of the proof we let Π denote $\Pi((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$ for brevity. By Proposition 2 we have

$$\mathsf{CC}(\Pi) \geq I(X_1 X_2 \cdots X_n Y_1 Y_2 \cdots Y_n | \Pi) \geq \sum_{i=1}^{n} I(X_i Y_i; \Pi) = nI(X_J Y_J; \Pi | J),$$

[5] To keep things notationally simple we state a special case of this theorem. In [3] the same statement was proven for prior-free information complexity.

where the second-to-last inequality is just the super-additivity of mutual information (Proposition 1) and the last equality is a straightforward calculation (for uniformly random $J \in [n]$). (Note that super-additivity can be applied here since the coordinates are sampled independently of one another.) The value of J is independent of the values of X_J and Y_J, so we can write

$$I(X_J Y_J; \Pi | J) = I(X_J Y_J; \Pi | J) + I(X_J Y_J; J)$$

and applying the chain rule for mutual information yields

$$I(X_J Y_J; \Pi | J) = I(X_J Y_J; \Pi J) = I(XY; \Pi J).$$

Since Alice and Bob sampled J with their public randomness at the beginning of τ' the last term is just the external information complexity of τ'. Collecting the inequalities together and applying Proposition 2 we get

$$\mathsf{CC}(\Pi) \geq nI(X_J Y_J; \Pi | J) \geq nI(X_J Y_J; \Pi J) \geq nI^{ext}_{\mu,\delta}(\tau') \geq nI_{\mu,\delta}(\tau'),$$

which finishes the proof when μ is a product distribution.

When μ is non-product things get a bit more difficult. The problem is that in the protocol τ', Alice and Bob can no longer sample the inputs (X_i, Y_i) for $i \neq j$ on their own. To get around this we have to be a bit more tricky with public randomness (this proof is due to [5]).

The new protocol τ will be as follows. Alice receives an input x and Bob receives an input y. Alice and Bob sample $j \in [n]$ uniformly at random, like before. But now they use public randomness to sample $X_1, X_2, \ldots X_{j-1}$ and $Y_{j+1}, Y_{j+2}, \ldots, Y_n$ (say, by sampling n inputs (X, Y) from μ and throwing away the half of the sample that is not needed). Then, using *private* randomness, for each $i = j+1, j+2, \ldots, n$, Alice samples X_i from the distribution μ conditioned on the value of Y_i. Similarly, for each $i = 1, 2, \ldots, j-1$, Bob samples Y_i from μ conditioned on the value of X_i. Finally, they simulate the protocol Π on their inputs with (x, y) plugged into the jth index. The information complexity of this protocol is

$$\mathsf{IC}_{\mu,\delta}(\tau) = I(X; \tau | Y) + I(Y; \tau | X), \tag{3}$$

and note that

$$\tau = J X_1 X_2 \ldots X_{j-1} Y_{j+1} Y_{j+2} \ldots Y_n \Pi.$$

We upper bound each component of (3) individually. For the first:

$$\begin{aligned} I(X; \tau | Y) &= I(X; J X_1 X_2 \ldots X_{J-1} Y_{J+1} Y_{J+2} \ldots Y_n \Pi | Y) \\ &\leq I(X; J X_1 X_2 \ldots X_{J-1} Y_1 \ldots Y_n \Pi | Y), \end{aligned}$$

where the inequality follows since we can only reveal strictly more information by showing the rest of Bob's sampled inputs. By the definition of the sampling in τ, X is conditionally independent of everything on the right of the mutual information except for Π, so

$$\begin{aligned} I(X; J X_1 \ldots X_{J-1} Y_1 \ldots Y_n \Pi | Y) &= I(X; J X_1 \ldots X_{J-1} Y_1 \ldots Y_n | Y) + I(X; \Pi | J X_1 \ldots X_{J-1} Y_1 \ldots Y_n) \\ &= I(X; \Pi | J X_1 \ldots X_{J-1} Y_1 \ldots Y_n). \end{aligned}$$

The rest of the argument is essentially identical to the case of the product distribution above. Since $X = X_J$, expanding the expectation of J yields

$$\begin{aligned} I(X_J; \Pi | J X_1 \dots X_{J-1} Y_1 \dots Y_n) &= \frac{1}{n} \sum_{j=1}^{n} I(X_j; \Pi | X_1 \dots X_{j-1} Y_1 \dots Y_n) \\ &\leq \frac{1}{n} I(X_1 \dots X_n; \Pi | Y_1 \dots Y_n), \end{aligned}$$

which is one of the terms in the information complexity of $\mathsf{IC}_{\mu^n,\delta}(\Pi)$. Applying the argument symmetrically to $I(Y; \tau | X)$ yields

$$\mathsf{IC}_{\mu,\delta}(\tau) \leq \frac{1}{n} \mathsf{IC}_{\mu^n,\delta}(\Pi) \leq \frac{1}{n} R_\delta(f^n).$$

There are a few remarks to be made about the above proof. Notice that we are lower bounding the *randomized* communication complexity and not the distributional communication complexity. This is a subtle point, and will be important later when we give a lower bound on the communication complexity of set disjointness.

In the product case we were able to get away by proving that $R_\delta^{\mu^n}(f) \geq n \mathsf{IC}_{\mu,\delta}^{ext}(\tau)$, and then use the fact that the external information greater than the internal information. This is symptomatic of a deeper phenomenon: namely, that if μ is a product distribution then $\mathsf{IC}_{\mu,\delta}^{ext}(\tau) = \mathsf{IC}_{\mu,\delta}(\tau)$.

It is important to digest this argument before proceeding: in the next section we will modify it to get a "direct-sum result" for the complexity of disjointness.

4 Lower Bounds for Disjointness

The first result is a new proof of the $\Omega(n)$ lower bound for disjointness, which is originally due to Bar Yossef et al [2]. We closely follow their presentation.

Recall that the disjointness function $DISJ(x, y)$ is defined to be

$$DISJ(x, y) = \bigvee_{i=1}^{n} (x_i \wedge y_i). \tag{4}$$

Our next goal is to get a linear lower bound on the randomized communication complexity of disjointness by reducing to (a variant of) the direct sum result proven in the previous section. Following an essentially identical argument as in Theorem 4, one can show that for *any* distribution μ over $\{0, 1\}$, the amount of information revealed in any randomized protocol computing disjointness on inputs drawn from μ^n is at least n times the amount of information revealed in any protocol computing the 1-bit AND function over μ. Then we can exhibit a hard distribution μ, over which any randomized protocol must reveal at least $\Omega(1)$ bits of information when computing AND.

Before we get into the details, however, we should discuss the choice of our prior distribution μ. It is known [1] that if μ is a product distribution then the distributional complexity of $DISJ$ is $O(\sqrt{n} \log n)$, and so we will need to choose μ to be a non-product distribution (and we will therefore need the full strength of Theorem 4). With this in mind, consider the following distribution μ: flip an unbiased coin, and if it comes up Heads set Alice's input to 0 and Bob's input to a uniformly random bit, and if it comes up Tails set Bob's input to 0 and Alice's input to a uniformly random bit. We will see later that any randomized protocol that computes AND correctly with probability $1 - \delta$ on *all* inputs must reveal $\Omega(1)$ bits of information on inputs sampled from μ.

Definition Let ν be the distribution over $\{0,1\}^2 \times \{\mathtt{h},\mathtt{t}\}$ defined by the following sampling procedure: first choose D from $\{\mathtt{h},\mathtt{t}\}$ uniformly at random. If $D = \mathtt{h}$ then set $X = 0$ and choose Y uniformly at random from $\{0,1\}$, and return $((X,Y),D)$. If $D = \mathtt{t}$ then set $Y = 0$ and choose X uniformly at random from $\{0,1\}$, and return $((X,Y),D)$. Notice that the marginal distribution of (X,Y) defined by ν is exactly the distribution μ defined above, and moreover that X and Y are conditionally independent given D.

4.1 Reduction to the Information Complexity of AND

The direct sum result in Section 3 is nice, but it does not directly apply (as stated) to $DISJ$ since $DISJ$ is the OR of n ANDs, rather then computing n copies of AND in parallel. However, it is possible to show that as long as the distribution μ on the ANDs has a special property (it is *collapsing*, using the term coined in [2]) then the direct sum reduction will still go through.

Definition Let σ be a distribution on $\{0,1\}^n \times \{0,1\}^n$. For any $i \in [n], x \in \{0,1\}^n$, and $a \in \{0,1\}$ let $x^{i \leftarrow a}$ be the string obtained from x by replacing the ith coordinate in x with a. We say that σ is *collapsing* if

$$DISJ(x^{i\leftarrow a}, y^{i\leftarrow b}) = a \wedge b,$$

for all (x,y) in the support of σ, all $i \in [n]$, and all $(a,b) \in \{0,1\}^2$.

It is easy to see that the distribution μ defined above is collapsing since it only places mass on 0s of $DISJ$. The next lemma completes the reduction from $DISJ$ to the 1-bit AND over μ.

Lemma 5 *Let $((X,Y),D)$ be sampled according to ν, and let $\delta > 0$. There exists a randomized protocol τ computing AND with probability $1-\delta$ on all inputs such that*

$$R_\delta(DISJ) \geq nI(XY;\tau(X,Y)|D).$$

Proof Let π be a randomized protocol with communication complexity $R_\delta(DISJ)$ that outputs the correct value with probability $1-\delta$ on all inputs. Consider the following protocol τ.

1. Alice and Bob receive input (x,y).
2. Using public randomness, Alice and Bob sample $j \in [n]$ uniformly at random, and $n-1$ variables $d_1,\ldots,d_{j-1},d_{j+1},\ldots,d_n$ from $\{\mathtt{h},\mathtt{t}\}$ uniformly at random.
3. Using private randomness, for each $i \neq j$ Alice samples x_i from ν conditioned on the value of d_i, and sets $x_j = x$.
4. Using private randomness, for each $i \neq j$ Bob samples y_i from ν conditioned on the value of d_i, and sets $y_j = y$.
5. Alice and Bob simulate the protocol π on $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$.

Since μ is a collapsing distribution it follows that τ is a protocol for computing AND.

Let $((\mathbf{X},\mathbf{Y}),\mathbf{D})$ be distributed according to ν^n. The transcript $\Pi(\mathbf{X},\mathbf{Y})$ and $\mathbf{D}$ are conditionally independent given $\mathbf{XY}$, so the data-processing inequality (Proposition 1) combined with Proposition 2 implies that

$$R_\delta(DISJ) \geq I(\mathbf{XY};\Pi) \geq I(\mathbf{XY};\Pi|\mathbf{D}). \tag{5}$$

Since the n coordinates are sampled independently, the super-additivity of mutual information yields

$$I(\mathbf{X}, \mathbf{Y}; \Pi|\mathbf{D}) \geq \sum_{j=1}^{n} I(X_j Y_j; \Pi|\mathbf{D}) = nI(X_J Y_J; \Pi|\mathbf{D}J) \tag{6}$$

where J is sampled uniformly at random from $[n]$.

Let $\mathbf{D}^{-J}$ be the random variable $D_1, \ldots, D_{j-1}, D_{j+1}, \ldots, D_n$, and note that X_J, Y_J, D^{-J}, and J are all independent of one another, even conditioned on D_J. This independence implies that $I(X_J Y_J; \mathbf{D}^{-J} J|D_J) = 0$, and so

$$I(X_J Y_J; \Pi|\mathbf{D}J) = I(X_J Y_J; \Pi|\mathbf{D}J) + I(X_J Y_J; \mathbf{D}^{-J} J|D_J) = I(X_J Y_J; \Pi J \mathbf{D}^{-J}|D_J).$$

But $\Pi J \mathbf{D}^{-J}$ is just the transcript of τ on X_J, Y_J, and since X_J, Y_J are distributed according to μ we can write

$$I(X_J Y_J; \Pi J \mathbf{D}^{-J}|D_J) = I(XY; \tau|D). \tag{7}$$

Combining (5), (6), and (7) yields

$$R_\delta(DISJ) \geq nI(XY; \tau|D).$$

4.2 A Lower Bound on the Information Complexity of AND

The next lemma is the target of this section:

Lemma 6 *Let $((X, Y), D)$ be distributed according to ν (cf. Definition 4). Let τ be any randomized protocol computing AND correctly with probability at least $2/3$ on all inputs. Then*

$$I(XY; \tau(X, Y)|D) \geq \frac{2}{9}.$$

Before we begin discussing the proof of this lemma, we will first need to quantify the *distance* between two probability distributions.

Definition Let P, Q be two finite probability distributions with the same support $\mathcal{S}$. Then the *Hellinger distance* $h(P, Q)$ between the two distributions is defined to be

$$h(P, Q) = \left(1 - \sum_{s \in \mathcal{S}} \sqrt{P(s)Q(s)}\right)^{1/2}.$$

Let $h^2(P, Q)$ denote the square of the Hellinger distance.

If we imagine the distributions P, Q as real-valued vectors indexed by elements in $\mathcal{S}$, then the Hellinger distance is just the normalized ℓ_2 distance between the roots of the two vectors:

$$h(P, Q) = \frac{1}{\sqrt{2}} ||\sqrt{P} - \sqrt{Q}||_2, \tag{8}$$

where the square-roots are taken component-wise. This fact will be useful later.

The proof of Lemma 6 follows first by reducing the lower bound on mutual information to a lower bound on Hellinger distance. Let us first manipulate the term $I(XY; \tau(X,Y)|D)$ a bit to see this. By directly expanding the expectation implicit in the conditional mutual information, we get

$$\begin{aligned} I(XY;\tau(X,Y)|D) &= \frac{1}{2}(I(XY;\tau(X,Y)|D=0) + I(XY;\tau(X,Y)|D=1)) \\ &= \frac{1}{2}(I(Z;\tau(0,Z)) + I(Z;\tau(Z,0)), \end{aligned} \tag{9}$$

where Z is a uniformly random bit on $\{0,1\}$. For $xy \in \{0,1\}^2$ let τ_{xy} represent the distribution of transcripts of τ on input (x,y). The following is a technical lemma which lower bounds the mutual information of the terms above by the Hellinger distance of their corresponding distributions. The proof is technical (and requires introducing more information theoretic terminology), so it is omitted.

Lemma 7 *Let Z be a uniformly random bit and let τ be any randomized protocol that computes AND correctly with error δ on all inputs. Then*

$$I(Z;\tau(0,Z)) \geq h^2(\tau_{00},\tau_{01}) \quad \text{and} \quad I(Z;\tau(Z,0)) \geq h^2(\tau_{00},\tau_{10}).$$

Proof Omitted. See [2] for a full proof (this lemma follows directly from Lemma 6.2 in that paper, which is proven as Lemma A.7 in their appendix).

Applying this lemma to (9) yields

$$\begin{aligned} I(XY;\tau(X,Y)|D) &\geq \frac{1}{2}(I(Z;\tau(0,Z)) + I(Z;\tau(Z,0)) \\ &\geq \frac{1}{2}(h^2(\tau_{00},\tau_{01}) + h^2(\tau_{00},\tau_{10})) \\ &\geq \frac{1}{4}(h(\tau_{00},\tau_{01}) + h(\tau_{00},\tau_{10}))^2 \\ &\geq \frac{1}{4}h^2(\tau_{10},\tau_{01}), \end{aligned} \tag{10}$$

where the third inequality is Cauchy-Schwarz and the final inequality is the triangle inequality (note that both of these can be applied thanks to (8)). We have reduced the problem of lower bounding this conditional mutual information term to lower-bounding the Hellinger distance of the two "diagonal" 0-distributions of τ. Intuitively, of course, this makes sense: if the distribution of transcripts produced by 00 is "close" to the distribution of transcripts produced by 01 and 10, then we expect that the two distributions of transcripts $01, 10$ are also close.

To finish off the proof of Lemma 6 we will need to exploit the fact that τ_{xy} is a probability distribution over transcripts, and the set of inputs producing that produce a particular transcript is a combinatorial rectangle. The next lemma is a probabilistic interpretation of this fact:

Lemma 8 (Cut-and-Paste Lemma) *Let Π be any randomized protocol computing a 2-party function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$. Let $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$. Then*

$$h(\Pi_{xy}, \Pi_{x'y'}) = h(\Pi_{x'y}, \Pi_{xy'}).$$

Proof Let T be a possible transcript of the protocol Π. If Π was a deterministic protocol, then the collection of inputs that reach T would form a rectangle (a set of the form $\mathcal{A} \times \mathcal{B}$ for some $\mathcal{A} \subseteq \mathcal{X}, \mathcal{B} \subseteq \mathcal{Y}$). However, now that Π is probabilistic this is not true in exactly the same sense, since on an input (u, v) the transcript $\Pi(u, v)$ may depend on the private random strings of Alice and Bob which do not appear in the transcript. So, let $(u, a), (v, b)$ be an extension of the input (u, v) with some strings of private random bits a, b. If we consider Π to depend on $(u, a), (v, b)$ then the rectangle property can be recovered: the set of inputs that reaches any fixed transcript T is a rectangle on the set of extended inputs.

Let T be any transcript of Π and let $\mathcal{A} \times \mathcal{B}$ be the rectangle of extended inputs (i.e. inputs of the form $(x, a), (y, b)$ for $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and (a, b) some strings of random bits) that reach T. Let $x \in \mathcal{X}$, and let $q_1(x, T)$ be the probability that that $(x, a) \in \mathcal{A}$ for a uniformly random string of coin flips a. Similarly, let $q_2(y, T)$ be the probability that $(y, b) \in \mathcal{B}$ for a uniformly random string of coin flips b. It follows from the discussion above that

$$\Pr[\Pi(x, y) = T] = q_1(x, T)q_2(y, T), \tag{11}$$

which is a "probabilistic rectangle property".

From (11) we can prove the lemma by a direct calculation. Namely,

$$\begin{aligned}
1 - h(\Pi_{xy}, \Pi_{x'y'}) &= \sum_T \sqrt{\Pr[\Pi_{xy} = T] \Pr[\Pi_{x'y'} = T]} \\
&= \sum_T \sqrt{q_1(x, T)q_2(y, T)q_1(x', T), q_2(y', T)} \\
&= \sum_T \sqrt{q_1(x, T)q_2(y', T)q_1(x', T), q_2(y, T)} \\
&= \sum_T \sqrt{\Pr[\Pi_{xy'} = T] \Pr[\Pi_{x'y} = T]} \\
&= 1 - h(\Pi_{xy'}, \Pi_{x'y}).
\end{aligned}$$

Applying the cut-and-paste lemma to (10) yields

$$I(XY; \tau(X, Y)|D) \geq \frac{1}{4} h^2(\tau_{00}, \tau_{11}). \tag{12}$$

And now we are certainly done (at least intuitively): since the protocol τ correctly computes AND with high probability, it simply cannot be the case that the distributions τ_{00} and τ_{11} are "close", since τ_{11} is the distribution of transcripts of 1s of the function.

In fact it is quite easy to prove that τ_{00} and τ_{11} are "distant" under a different form of distance between probability distributions. For two finite probability distributions P, Q with the same support $\mathcal{S}$ define the *total variational distance* V to be

$$V(P, Q) := \max_{s \in \mathcal{S}} |P(s) - Q(s)|.$$

In fact, as the Hellinger distance is just the ℓ_2 distance of the probability distributions, the total variational distance can be re-written as the ℓ_1 distance of the probability distributions:

$$V(P, Q) = \frac{1}{2} ||P - Q||_1. \tag{13}$$

Since P, Q are vectors satisfying $||P||_1 = ||Q||_1 = 1$ we can embed the ℓ_1 norm into the ℓ_2 norm with distortion at most $\sqrt{2}$, so

$$V(P,Q) \leq \sqrt{2}h(P,Q). \tag{14}$$

Finally, the probability that $\tau_{11} = \tau_{00}$ is at least $1 - 2\delta$ over the coin-flips of the protocol. This implies that $V(P,Q) \geq 1 - 2\delta$. Combining this fact with (12) and (14) implies

$$I(XY; \tau(X,Y)|D) \geq \frac{1}{4}h^2(\tau_{00}, \tau_{11}) \geq \frac{1}{2}V(P,Q)^2 \geq \frac{(1-1/3)^2}{2} = \frac{2}{9},$$

proving Lemma 6 for $c = 2/9$.

Combining Lemma 6 and Lemma 5 yields

$$R_{1/3}(DISJ) \geq \Omega(n).$$

References

[1] L. Babai, P. Frankl, and J. Simon, *Complexity classes in communication complexity theory (preliminary version)*, in 27th Annual Symposium on Foundations of Computer Science, Toronto, Canada, 27-29 October 1986, IEEE Computer Society, 1986, pp. 337–347.

[2] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar, *An information statistics approach to data stream and communication complexity*, J. Comput. Syst. Sci., 68 (2004), pp. 702–732.

[3] M. Braverman, *Interactive information complexity*, in Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012, H. J. Karloff and T. Pitassi, eds., ACM, 2012, pp. 505–524.

[4] M. Braverman, A. Garg, D. Pankratov, and O. Weinstein, *From information to exact communication*, in Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013, D. Boneh, T. Roughgarden, and J. Feigenbaum, eds., ACM, 2013, pp. 151–160.

[5] M. Braverman and A. Rao, *Towards coding for maximum errors in interactive communication*, in Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011, L. Fortnow and S. P. Vadhan, eds., ACM, 2011, pp. 159–166.