Communication Complexity

Lecturer: Dean Hirsch and Oliver Korten

1 Information Theory Basics

1.1 Entropy

Information theory is a means of quantifying the information contained in a random variable. The most basic concept in information theory is the "entropy" of a random variable. One way to understand the entropy of a random variable X is that it measures the amount of uncertainty one has, a-priori, about the value of a sample drawn from X. Alternatively, it can be viewed as the minimum cost of communicating the value of a random sample from X. Intuitively, when one has a great amount of uncertainty about the value of a draw from X, they will need to spend more bits describing which of the many possible values it took on.

The task to keep in mind is the following: there is a random variable X whose distribution is known to two parties Alice and Bob. Alice has access to a sampler for X, and starts to draw samples from it and then send the sampled value to Bob by communicating bits over some channel. If X always takes on the same value x with probability 1, Alice need not do any communication since Bob always knows the value she sampled. In this case the information content of a typical sample from X should be 0, and thus the entropy of X should be zero, which coincides with the fact that there is no uncertainty about the outcome of X. On the other hand, if X took on one of k values with equal probability, it seems that the best Alice can do is assign a distinct binary string of length $\lceil \log k \rceil$ to each possible value, and send the string associated to each value to Bob whenever it is sampled. In this case, the information content of a typical sample seems to be about $\lceil \log k \rceil$.

As discovered by Shannon, there is an explicit measure for probability distributions, known as "entropy," which is able to assign such a notion to all distributions in a way consistent with the above intuition. In particular, we define the entropy of X, or $H(X)$, as:

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)} = \mathbb{E}_{x \sim X} \log \frac{1}{p(x)}$$

where $p(x) \log \frac{1}{p(x)}$ is taken to be 0 when $p(x) = 0$, which is consistent with the limit.

A few things to note: we see that $H(X)$ is completely oblivious to the specific values X takes on, and only depends on its underlying probability distribution. We also see that its value is always nonnegative, consistent with the intuition that there is no such thing as "negative information" or "negative uncertainty." In particular, we see that a distribution has entropy 0 if and only if it assigns probability 1 to some sample (in this case there is no uncertainty; the outcome is fully determined). On the other end, we can upper bound the entropy of X as follows: if X only takes on values in some set U, then

$$H(X) \leq \log |U|$$

To prove this (and later results), we use Jensen's inequality, which says that for any concave function $\phi : \mathbb{R} \to \mathbb{R}$, and any real-values random variable X, we have $\phi(\mathbb{E}[X]) \geq \mathbb{E}[\phi(X)]$. We will typically apply this with $\phi = \log$, which can readily be seen to be concave. Using this we have:

$$\begin{aligned} H(X) &= \sum_{x \in U} p(x) \log \frac{1}{p(x)} \\ &= \mathbb{E}_{x \sim X} \left(\log \frac{1}{p(x)} \right) \\ &\leq \log \left(\mathbb{E}_{x \sim X} \frac{1}{p(x)} \right) \\ &= \log \left(\sum_{x \in U} p(x) \frac{1}{p(x)} \right) = \log |U| \end{aligned}$$

In particular, we see that if X is a distribution over n-bit strings, then $H(X) \leq n$.

1.1.1 Operational Interpretation of Entropy

As hinted in the begining of the previous section, entropy can be characterized alternatively in a fully operational sense. Let X be a random variable over a set U, and consider a scheme for encoding values of X by binary strings, call it $f : U \to \{0,1\}^*$. A natural desirable property of such a scheme is that it would allow us to uniquely specify a sequence of samples $x_1, \ldots, x_k$ from X by the concatenation of $f(x_1), \ldots, f(x_k)$; this is called a "uniquely decodable code." We then have the following:

Theorem 1. *Let $Q(X)$ be the minimum, over all uniquely decodable codes f for X, of*

$$\mathbb{E}_{x \sim X} |f(x)|$$

Then $H(X) \leq Q(X) \leq H(X) + 1$

We will prove this theorem for the special case of "prefix-free" codes: codes f such that for all $x \neq y$, $f(x)$ is not a prefix of $f(y)$. Such a code can readily be seen to be uniquely decodable.

Proof. We start by proving the second inequality. We will assume without loss of generality that $U = [n]$, and that the elements of U are arranged in decreasing order of probability so that $1 \geq p(1) \geq \cdots \geq p(n) \geq 0$. We will give a prefix-free encoding which assigns each $x \in [n]$ a codeword of length $\ell_x := \lceil \log \frac{1}{p(x)} \rceil$. If this holds, then we have:

$$\begin{aligned} \mathbb{E}_{x \sim X} |f(x)| &= \mathbb{E}_{x \sim X} \ell_x \\ &= \mathbb{E}_{x \sim X} \lceil \log \frac{1}{p(x)} \rceil \\ &\leq \mathbb{E}_{x \sim X} \left(1 + \log \frac{1}{p(x)} \right) = H(X) + 1 \end{aligned}$$

completing the proof.

To construct such a code, we initialize a complete binary tree of depth n. Now, for each $x \in [n]$ in increasing order of x, we find a node of depth ℓ_x and delete all of its descendants so that this node becomes a leaf. We then give x the codeword specifying the root-to-leaf path of this node, and continue for the next value of x. So long as we can always find a node of the appropriate depth, each x will be given a codeword of the appropriate length, and no codeword will be a prefix of another since all codewords are leaves.

It suffices to show that after all the vertex deletions at step x, there is still a remaining vertex of depth ℓ_{x+1}. For $y < x$, the number of vertices of depth ℓ_x deleted at step y is exactly $2^{\ell_x - \ell_y}$. So the number of vertices of depth ℓ_x that are deleted before the x^{th} step is:

$$\sum_{y<x} 2^{\ell_x-\ell_y} = 2^{\ell_x} \sum_{y<x} 2^{-\ell_y} \leq 2^{\ell_x} \sum_{y<x} p(y) < 2^{\ell_x}$$

By definition there are 2^{ℓ_x} leaves of depth ℓ_x to begin with, so one must remain.

We know show the other inequality, that for any prefix free code f for X, its expected code length is at least $H(X)$. Again, assume that X lies over the universe $[n]$, and say that $x \in [n]$ is given a codeword of length ℓ_x by f. So then we have:

$$\begin{aligned}
\mathbb{E}_{x\sim X}\left(\ell_x\right) &= \mathbb{E}_{x\sim X}\left(\log\frac{1}{p(x)} - \log(2^{-\ell_x}/p(x))\right) \\
&= H(X) - \mathbb{E}_{x\sim X}\left(\log(2^{-\ell_x}/p(x))\right) \\
&\geq H(X) - \log\left(\mathbb{E}_{x\sim X}\left(2^{-\ell_x}/p(x)\right)\right) \\
&= H(X) - \log\left(\sum_x 2^{-\ell_x}\right)
\end{aligned}$$

So if we can show that $\sum\limits_x 2^{-\ell_x} \leq 1$ then we are done. Recall that f is a prefix free code, so we can think of f as assigning each element of $[n]$ a distinct node in a binary tree (the node arrived at from the root by going left/right in correspondence with the bits of its codeword), such that no assigned node is a descendent of another. Now, consider the random process which, starting at the root, moves left or right with probability $\frac{1}{2}$ until it reaches a codeword assigned by f. Then the probability of ever reaching some codeword is precisely $\sum\limits_x 2^{-\ell_x}$, and this sum is therefore at most one. □

1.1.2 Conditional Entropy

In what follows it will be useful to have a notion of conditional entropy, which quantifies the uncertainty that remains in some variable X after we know the value of a (potentially correlated) variable Y. The condition entropy of X given Y, denoted $H(X|Y)$, is defined as:

$$H(X|Y) = \mathbb{E}_{y\sim Y} H(X|Y = y)$$

In other words, its the expected entropy of X conditioned on a particular value of y, when y is sampled from Y.

1.2 Mutual Information

In many situations we will have multiple jointly distributed random variables and it will be useful to measure the amount of information that is shared between them. For random variables X, Y, we will use $I(X : Y)$ ("mutual information between X, Y") to quantify this. Intuitively, $I(X : Y)$ should capture that amount of information we gain on X by knowing Y. Thus, when X and Y are independent, we should have that no information is gained, and thus $I(X : Y) = 0$. Alternatively, if $X = Y$, then either X or Y contains all the information about the other, and thus we should have $I(X : Y) = H(X) = H(Y)$. The formal definition is as follows, for jointly distributed variable X, Y

$$I(X : Y) = \mathbb{E}_{x,y\sim XY}\left(\log \frac{p(x,y)}{p(x)p(y)}\right)$$

Here, we use $p(x, y)$ to denote the probability of x, y under the joint distribution XY, while $p(x)$ and $p(y)$ denote the probability of x and y under the marginal distributions X and Y respectively. Note that $I(X : Y) = I(Y : X)$ by definition.

We now show the following alternative characterization:

$$I(X : Y) = H(X) + H(Y) - H(XY)$$

According to this formula, we can describe $I(X : Y)$ as the decrease in uncertainty obtained by considering X, Y as a joint distribution as opposed to considering their marginals independently. The proof of the equivalence of the two definitions is as follows:

$$I(X : Y) = \mathbb{E}_{x,y\sim XY}\left(\log \frac{p(x,y)}{p(x)p(y)}\right)$$
$$= \mathbb{E}_{x,y\sim XY}\left(\log \frac{1}{p(x)} + \log \frac{1}{p(y)} - \log \frac{1}{p(x,y)}\right)$$
$$= H(X) + H(Y) - H(XY)$$

Another important fact is that the mutual information is always non-negative. This can be obtained readily from Jensen's inequality:

$$\mathbb{E}_{x,y\sim XY} \log \frac{p(x,y)}{p(x)p(y)} = -\mathbb{E}_{x,y\sim XY} \log \frac{p(x)p(y)}{p(x,y)}$$
$$\geq -\log\left(\mathbb{E}_{x,y\sim XY} \frac{p(x)p(y)}{p(x,y)}\right) = -\log\left(\sum_{x,y} p(x)p(y)\right)$$
$$= -\log 1 = 0$$

As with entropy, there is also a notion of conditional mutual information. Analogously to entropy, we will define $I(X : Y|Z)$ to be the expectation over $z \sim Z$ of $I(X : Y|Z = z)$:

$$I(X : Y|Z) = \mathbb{E}_{z\sim Z}\left(I(X|Z = z : Y|Z = z)\right)$$

1.3 Key Properties

1.3.1 Chain Rules

Chain rules allow us to break down a joint distribution into a marginal and a conditional part. The basic chain rule of probabilities says:

$$P(x, y) = P(x)P(y|x)$$

This simple fact will allow us to similarly break down the various information measures into such a nice form. We start with the chain rule for entropy:

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

Intuitively, this says that the entropy of XY is equal to the entropy in X plus the entropy that remains in Y once you know X. The proof is as follows:

$$\begin{aligned}
H(X|Y) &= \sum_y p(y)H(X|y) \\
&= \sum_y p(y) \left(\sum_x p(x|y) \log \frac{1}{p(x|y)} \right) \\
&= \sum_{x,y} p(x,y) \log \frac{1}{p(x|y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(y)}{p(x,y)} \\
&= \sum_{x,y} p(x,y) \log \frac{1}{p(x,y)} - \sum_{x,y} p(x,y) \log \frac{1}{p(y)} \\
&= H(XY) - H(Y)
\end{aligned}$$

Recall that we showed before that:

$$I(X : Y) = H(X) + H(Y) - H(XY)$$

Combining this with the chain rule for entropy we get an alternative expression:

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

So the information between X and Y is the amount of uncertainty about X which you eliminate by knowing Y (and vice versa). For mutual information we have the following chain rule:

$$I(XY : Z) = I(X : Z) + I(Y : Z|X)$$

Its proof follows the same form as above.

1.3.2 Inequalities

The only remaining results we need for the next section are the following inequalities:

$$H(XY) \leq H(X) + H(Y)$$

which we refer to as the "subadditivity of entropy," and

$$H(X|Y) \leq H(X)$$

which says that conditioning on a variable can never increase its entropy. The first inequality follows from the previously proven identity $I(X : Y) = H(X) + H(Y) - H(XY)$, and the fact from before that mutual information is always nonnegative. Now, combining this "subadditivity of entropy" with the chain rule for entropy, we get:

$$H(Y) + H(X|Y) = H(XY) \leq H(Y) + H(X)$$

so in particular $H(X|Y) \leq H(X)$, establishing the second inequality.

2 Disjointess Lower Bound Introduction

The lower bound of $\Omega(n)$ on the randomized communication complexity of Disjointness is perhaps the most impactful result in the entire area. The first linear lower bound is due to Schnitger and Kalyanasundaram [2], and a simpler proof was obtained by Razborov [1]. It has consequences in circuit complexity, property testing, algorithmic game theory, extension complexity, data structures, etc. All known proofs of the tight lower bound use information theory one way or another, but the ideas we've developed in our study of information complexity give a particularly simple and intuitively satisfying proof. The presentation here follows the proof of Bar-Yossef, Jayram, Kumar, and Sivakumar, "An information statistics approach to data stream and communication complexity" [5].

It will be more convenient to work with the complement of the Disjointness function, which we'll call the Set-Intersection function. Define

$$\mathrm{INT}_n(x, y) = \bigvee_{i=1}^{n} (x_i \wedge y_i).$$

The idea of the proof will be to show that for some distribution ζ over $\{0,1\}^2$, a protocol for computing INT_n over inputs drawn from $\zeta^{\otimes n}$ will have to reveal n times as much information as a protocol for computing AND_2 over inputs drawn from ζ. We will then exhibit a distribution ζ over $\{0,1\}^2$ which requires $\Omega(1)$ bits of information to compute AND_2.

Definition 1 (Repeated drawing)**.** *For any distribution ζ, we denote by $\zeta^{\otimes n}$ the distribution obtained by randomly drawing n independent times from ζ and concatenating the results.*

If the distribution ζ is over tuples (as will be the case in the following), say drawing from ζ outputs a tuple (a, b, d), then we will denote the output of $\zeta^{\otimes n}$ as (A, B, D) where $A = (a_1, \ldots, a_n)$, $B = (b_1, \ldots, b_n)$, $D = (d_1, \ldots, d_n)$.

3 From INT_n to AND_2

Define the distribution ζ over $\{0,1\}^2 \times \{a,b\}$ as follows. Let D be uniformly random from $\{a,b\}$. If $D = a$, let $A = 0$ and B be uniform over $\{0,1\}$. If $D = b$, let $B = 0$ and A be uniform over $\{0,1\}$. Then A and B are independent conditioned on D.

The marginal distribution over the bits (A,B) is such that $(0,0)$ has probability $\frac{1}{2}$, and $(0,1)$ and $(1,0)$ both have probability $\frac{1}{4}$. The output $(1,1)$ is impossible.

Crucially, the marginal distribution over A, B obtained using ζ is such that A AND B is always 0. When drawing several random outputs from ζ to receive *strings* A and B (i.e. drawing from $\zeta^{\otimes n}$ as in Definition 1), the marginal distribution on A, B is such that they are disjoint, and therefore there exists a trivial upper bound for this specific distribution over inputs. However, we do not aim to prove a distributional lower bound. Rather, we aim to argue that enough information has to be communicated on the A and B drawn from this distribution, under the assumption that the protocol has low error on *any* input.

We begin by studying $I(AB; \Pi(A,B)|\vec{D})$, where $((A,B),\vec{D}) \sim \zeta^{\otimes n}$ (where $\zeta^{\otimes n}$ is as in Definition 1) and $\Pi(A,B)$ is the transcript of the communication between Alice and Bob. That is, this is the amount of information a bystander looking at the communication learns about Alice and Bob's inputs, given he already knows $\vec{D}$.

The motivation is twofold. First, we can analyze this distribution, using the fact that A, B are disjoint under this distribution, and independent given $\vec{D}$. Second, the amount of information is upper bounded by the entropy in the communication, $H(\Pi(A,B))$, which is upper bounded by the log of the support of $\Pi(A,B)$ as with all random variables. Thus, this lower bounds the number of bits communicated:

$$|\text{communication}| \le I(AB; \Pi(A,B)|\vec{D}) \tag{1}$$

Lemma 2 (Information Cost Decomposition)**.** *Let ζ be the distribution defined above, and let $((A,B),\vec{D}) \sim \zeta^{\otimes n}$. Then for any protocol* Π*,*

$$I(AB; \Pi(A,B)|\vec{D}) \ge \sum_{i=1}^{n} I(A_iB_i; \Pi(A,B)|\vec{D}).$$

Proof. Abbreviating $\Pi = \Pi(A,B)$, we have

$$\begin{aligned} I(AB;\Pi\vec{D}) &= H(AB|\vec{D}) - H(AB|\Pi\vec{D}) \\ &= \sum_{i=1}^{n} H(A_iB_i|\vec{D}) - H(AB|\Pi\vec{D}) && \text{since the } A_i, B_i\text{'s are independent given } D \\ &\ge \sum_{i=1}^{n} H(A_iB_i|\vec{D}) - \sum_{i=1}^{n} H(A_iB_i|\Pi\vec{D}) && \text{by subadditivity of entropy} \\ &= \sum_{i=1}^{n} I(A_iB_i;\Pi|\vec{D}). \end{aligned}$$

□

Lemma 3 (Reduction Lemma)**.** *Let* Π *compute* INT_n *with probability at least* $1-\varepsilon$ *on every input. Let ζ be the distribution defined above and let $((A,B),\vec{D}) \sim \zeta^{\otimes n}$ and $((U,V),D) \sim \zeta$. Then for* *every $i \in [n]$,*

$$I(A_iB_i; \Pi(A,B)|\vec{D}) \geq \inf_P I(UV; P(U,V)|D),$$

where the infimum is taken over protocols P computing AND_2 *with probability at least $1-\varepsilon$ on every input.*

Proof. Let Π be a protocol computing INT_n with probability at least $1-\varepsilon$ on every input. Fix an index i. By the definition of conditional mutual information,

$$I(A_iB_i; \Pi(A,B)|\vec{D}) = \mathbb{E}_{\vec{d}\sim\{a,b\}^{n-1}}[I(A_iB_i; \Pi(A,B)|D_i, \vec{D}_{-i} = \vec{d})].$$

So it suffices to show that for every fixed $\vec{d} = (d_1, \ldots, d_{i-1}, d_{i+1}, \ldots, d_n) \in \{a,b\}$, there is a protocol for AND_2 with information cost $I(A_iB_i; \Pi(A,B)|D_i, \vec{D}_{-i} = \vec{d})$.

We now describe such a protocol $P_{i,\vec{d}}$ for AND_2. On input u, v, Alice and Bob set $x_i = u, y_i = v$ and every other (x_j, y_j) to a sample from ζ conditioned on d_j (using private randomness). They then run the protocol $\Pi(x,y)$ and output its result. This protocol computes the AND_2 function since $u = v = 1$ iff x and y are intersecting inputs.

Next, we analyze the conditional information cost of $P_{i,\vec{d}}$. One can check by inspection that the joint distribution of $(U, V, D, P_{i,\vec{d}}(U,V))$ is equal to that of $(A_i, B_i, D_i, \Pi(A,B))$ conditioned on $\vec{D}_{-i} = \vec{d}$.

Hence

$$I(UV; P_{j,\vec{d}}(U,V)|D) = I(A_iB_i; \Pi(A,B)|D_i, \vec{D}_{-i} = \vec{d}).$$

□

Theorem 4. *Let ζ be the distribution defined above and let $((A,B),D) \sim \zeta^{\otimes n}$. Then*

$$\mathbf{BPP}^{pub}_{\varepsilon}(\mathrm{INT}_n) \geq n \cdot \inf_P I(AB; P(A,B)|D)$$

where the infimum is taken over protocols P computing AND_2 *with probability at least $1-\varepsilon$.*

Proof. Follows by combining the Reduction Lemma and the Information Cost Decomposition, together with Equation 1. □

4 Why Not a Product Distribution?

Alice and Bob's inputs when sampled from $\zeta^{\otimes n}$ are independent conditioned on D, but they are not fully independent. Why couldn't we use a fully independent distribution to prove a lower bound for Set-Intersection? It turns out that Set-Intersection is easy for product distributions: Babai, Frankl, and Simon [4] showed that for every product distribution μ, there is a distributional protocol for Set-Intersection showing that $D_\mu(\mathrm{INT}_n) = O(\sqrt{n}\log n)$. But this may still not be so convincing, since there is a distributional protocol with respect to $\zeta^{\otimes n}$ as well: all inputs in its support are non-intersecting, so nothing needs to be communicated. The magic in this proof lies in the fact that any protocol for computing INT_n on *every* input with high probability must reveal a lot of information when run on the distribution $\zeta^{\otimes n}$, even though that distribution isn't even supported on yes inputs.

To see why this argument wouldn't work for when Alice and Bob are fully independent, we argue that for any product distribution there is a protocol which succeeds on every input but has low expected communication when run on that distribution. Specifically, let's fix a product distribution on (A, B) and consider the following protocol for Set-Intersection. For some parameter $\varepsilon > 0$, Alice finds a coordinate (if one exists) such that $H(A_i) \geq \varepsilon$ and $H(B_i) \geq \varepsilon$. If one is found, the parties exchange the coordinate, output 1 if both coordinates are 1, condition on the values seen and repeat. When they run out of high-entropy coordinates, then the remaining entropy in A and B must be at most $\varepsilon \cdot n$, so they can just transmit their sets with roughly this many bits of communication.

How many rounds should we expect this protocol to last? If $H(A_i) \geq \varepsilon$, then $\Pr[A_i = 1] \geq \Omega(\varepsilon/\log(1/\varepsilon))$. So the probability of finding an intersection is at least about ε^2. Thus with high probability, the protocol will not last for more than about $1/\varepsilon^2$ rounds when actually run on the distribution (A, B). Setting $\varepsilon \approx n^{-1/3}$ gives expected total communication roughly $n^{2/3}$.

5 Information Complexity of AND

Our goal is now to prove a lower bound on the information complexity of any protocol computing AND_2. In this section, we'll give the proof as a consequence of a sequence of lemmas, and give the proofs of those lemmas in the following section.

Theorem 5. *Let P be a protocol which computes $\mathrm{AND}(u, v)$ with probability at least $1 - \varepsilon$. Let $((U, V), D) \sim \zeta$. Then*

$$I(UV; P(U, V)|D) \geq \frac{1}{4}(1 - 2\sqrt{\varepsilon}).$$

To begin analyzing this, we expand the quantity on the right to make it easier to work with. Let $Z \in \{0, 1\}$ be uniformly random. Then by definition

$$\begin{aligned} I(UV; P(U, V)|D) &= \frac{1}{2} I(UV; P(U, V)|D = a) + \frac{1}{2} I(UV; P(U, V)|D = b) \\ &= \frac{1}{2} I(Z; P(0, Z)) + \frac{1}{2} I(Z; P(Z, 0)). \end{aligned}$$

For $(u, v) \in \{0, 1\}^2$, let p_{uv} denote the distribution of transcripts when running $P(u, v)$. The next step is to relate these mutual information quantities to distances between distributions. The distance which will be convenient for us to use is the Hellinger distance.

Definition 2. *The squared Hellinger distance between two probability distributions p, q over domain X is defined by*

$$h^2(p, q) = 1 - \sum_{x \in X} \sqrt{p(x)q(x)} = \frac{1}{2} \sum_{x \in X} (\sqrt{p(x)} - \sqrt{q(x)})^2.$$

Lemma 6 (Hellinger Lower Bound)**.** *Let P be a protocol computing AND with probability at least $1 - \varepsilon$. Then for $Z \in \{0, 1\}$ uniform,*

$$I(Z; P(0, Z)) \geq h^2(p_{00}, p_{01}) \qquad \text{and} \qquad I(Z; P(Z, 0)) \geq h^2(p_{00}, p_{10}).$$

Continuing our calculation,

$$\begin{aligned}
I(UV;P(U,V)|D) &= \frac{1}{2}I(Z;P(0,Z)) + \frac{1}{2}I(Z;P(Z,0)) \\
&\geq \frac{1}{2}h^2(p_{00},p_{01}) + \frac{1}{2}h^2(p_{00},p_{10}) \\
&\geq \frac{1}{4}(h(p_{00},p_{01}) + h(p_{00},p_{10}))^2 && \text{Cauchy-Schwarz} \\
&\geq \frac{1}{4}h^2(p_{01},p_{10}) && \text{Triangle Inequality}
\end{aligned}$$

At this point, it is not clear why we should expect the distance between p_{01} and p_{10} to be large, as both are 0-inputs to the AND function. This is the point where we exploit the fact that these are distributions over transcripts, and that the set of inputs resulting in any given transcript is a rectangle:

Lemma 7 (Cut-and-Paste Lemma)**.** *Let P be a randomized protocol over $X \times Y$. Then for every $x, x' \in X$ and every $y, y' \in Y$, we have $h(P_{xy}, P_{x',y'}) = h(P_{x,y'}, P_{x',y})$.*

Applying the Cut-and-Paste Lemma lets us conclude that

$$I(UV;P(U,V)|D) \geq \frac{1}{4}h^2(p_{00},p_{11}).$$

And now, we should expect to be done, because any accurate protocol for AND must induce very different distributions on $(0,0) \in \text{AND}^{-1}(0)$ and $(1,1) \in \text{AND}^{-1}(1)$. Indeed, we have

Lemma 8 (Distinguishing Lemma)**.** *If P computes a function f with probability $2/3$ on every input, and (x,y) and (x',y') are inputs such that $f(x,y) \neq f(x',y')$, then*

$$h^2(p_{xy}, p_{x'y'}) \geq 1 - 2\sqrt{\varepsilon}.$$

This allows us to conclude the proof of Theorem 5.

6 Deferred Proofs

"Proof" of Hellinger Lower Bound Lemma 6. The statement is true as given, but the proof is more complicated and requires introducing a few more information-theoretic quantities. We will prove the weaker statement that if $Z \in \{0,1\}$ is uniform and if P is a randomized function on one bit, then

$$I(Z;P(Z)) \geq \frac{\log e}{2}h^2(p,q),$$

where p is the distribution of $P(Z)$ and q_0 and q_1 are the distributions of $P(0)$ and $P(1)$ respectively.

For any pair of distributions q, p, we have

$$\begin{aligned}
D(q\|p) &= -\sum_T q(T)\log\frac{p(T)}{q(T)} \\
&\geq \sum_T q(T)\cdot(2\log e)\left(1-\sqrt{\frac{p(T)}{q(T)}}\right) && \text{since} \ln y \leq y-1 \\
&= 2(\log e)h^2(q,p)
\end{aligned}$$

For any random variables X, Y with joint distribution p, we may write

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) \\ &= \sum_{y} p(y) \sum_{x} p(x|y) \log\left(\frac{p(x|y)}{p(x)}\right) \\ &= \mathbb{E}_{y\sim Y} D(p(\cdot|y)\|p). \end{aligned}$$

Hence

$$\begin{aligned} I(Z;P(Z)) &= \mathbb{E}_{z\sim Z} D(q_z\|p) \\ &= \frac{1}{2}(D(q_0\|p) + D(q_1\|p)) \\ &\geq \log e \cdot (h^2(q_0,p) + h^2(q_1,p)) \\ &\geq \frac{\log e}{2}(h(q_0,p) + h(q_1,p))^2 && \text{by Cauchy-Schwarz} \\ &\geq \frac{\log e}{2} h^2(q_0,q_1) && \text{by Triangle Inequality.} \end{aligned}$$

□

Proof of Cut-and-Paste Lemma 7. Let P be a private coin protocol and let p_{xy} denote the distribution on transcripts when P is run on (x,y). Recall that for any P, we can, for every transcript T, decompose $\Pr[P(x,y)=T] = q_A(x,T)\cdot q_B(y,T)$ for some functions q_A, q_B.

Then

$$\begin{aligned} 1 - h^2(p_{xy}, p_{x'y'}) &= \sum_{T} \sqrt{\Pr[P(x,y)=T]\cdot\Pr[P(x',y')=T]} \\ &= \sum_{T} \sqrt{q_A(x,T)q_B(y,T)q_A(x',T)q_B(y',T)} \\ &= \sum_{T} \sqrt{\Pr[P(x',y)=T]\cdot\Pr[P(x,y')=T]} \\ &= 1 - h^2(p_{x'y}, p_{x'y}). \end{aligned}$$

□

Proof of Distinguishing Lemma 8. We'll actually begin by showing that p_{xy} and $p_{x',y'}$ are far in total variation distance, where we recall that the total variation distance between two distributions p,q over $\mathcal{T}$ is

$$TV(p,q) = \max_{S\subseteq\mathcal{T}}(p(S) - q(S)) = \frac{1}{2}\sum_{T\in\mathcal{T}} |p(T) - q(T)|.$$

Let S be the set of transcripts on which P outputs $f(x,y)$. Then $p_{xy}(S) \geq 1-\varepsilon$ and $p_{x'y'}(S) \leq \varepsilon$. Hence $TV(p_{xy}, p_{x'y'}) \geq 1 - 2\varepsilon$.

Now we relate total variation distance to Hellinger distance as follows:

$$\begin{aligned}
TV^2(p,q) &= \frac{1}{4}\left(\sum_T |p(T)-q(T)|\right)^2 \\
&= \frac{1}{4}\left(\sum_T (\sqrt{p(T)}+\sqrt{q(T)})(\sqrt{p(T)}-\sqrt{q(T)})\right)^2 \\
&\le \frac{1}{4}\left(\sum_T (\sqrt{p(T)}+\sqrt{q(T)})^2\right)\left(\sum_T (\sqrt{p(T)}-\sqrt{q(T)})^2\right) && \text{Cauchy-Schwarz} \\
&\le \frac{1}{2}h^2(p,q)\cdot\left(2+2\sum_T \sqrt{p(T)}\sqrt{q(T)}\right) \\
&= h^2(p,q)(2-h^2(p,q)).
\end{aligned}$$

This allows us to conclude that $h^2(p,q) \ge 1-2\sqrt{\varepsilon}$. □

7 Acknowledgement

A large portion of the presentation and notes are based on Mark Bun's lecture notes [3]. We thank Mark for generously sharing with us the source code for those notes.

References

[1] A.A. Razborov *On the Distributional Complexity of Disjointness*, Theoretical Computer Science, 106(2), 1992, pp. 385-390. 6

[2] Schnitger, G., and Kalyanasundaram, B., *The probabilistic communication complexity of set intersection*, Proceedings of Second Annual Conference on Structure in Complexity Theory, 1987, pp. 16-19. 6

[3] Mark Bun, *Lecture Notes 8: Disjointness Lower Bound*, https://cs-people.bu.edu/mbun/courses/591_F19/notes/lec8.pdf 12

[4] L. Babai, P. Frankl and J. Simon, *Complexity classes in communication complexity theory*, 27th Annual Symposium on Foundations of Computer Science (sfcs 1986), 1986, pp. 337-347, url: https://ieeexplore.ieee.org/document/4568225. 8

[5] Ziv Bar-Yossef, T.S. Jayram, Ravi Kumar, D. Sivakumar, *An information statistics approach to data stream and communication complexity* , https://www.sciencedirect.com/science/article/pii/S0022000003001855 6