An information statistics approach to data stream and communication complexity

Ziv Bar-Yossef* T. S. Jayram Ravi Kumar D. Sivakumar
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120.
Email: {ziv, jayram, ravi, siva}@almaden.ibm.com

September 4, 2003

Abstract

We present a new method for proving strong lower bounds in communication complexity. This method is based on the notion of the *conditional information complexity* of a function which is the minimum amount of information about the inputs that has to be revealed by a communication protocol for the function. While conditional information complexity is a lower bound on communication complexity, we show that it also admits a *direct sum theorem*. Direct sum decomposition reduces our task to that of proving conditional information complexity lower bounds for simple problems (such as the AND of two bits). For the latter, we develop novel techniques based on Hellinger distance and its generalizations.

Our paradigm leads to two main results:

(1) An improved lower bound for the multi-party set-disjointness problem in the general communication complexity model, and a nearly optimal lower bound in the one-way communication model. As a consequence, we show that for any real $k > 2$, approximating the k-th frequency moment in the data stream model requires essentially $\Omega(n^{1-2/k})$ space; this resolves a conjecture of Alon, Matias, and Szegedy [AMS99].

(2) A lower bound for the L_p approximation problem in the general communication model; this solves an open problem of Saks and Sun [SS02]. As a consequence, we show that for $p > 2$, approximating the L_p norm to within a factor of n^ϵ in the data stream model with constant number of passes requires $\Omega(n^{1-4\epsilon-2/p})$ space.

*Part of this work was done while the first author was a student at UC Berkeley, and a visitor at IBM. Supported by NSF ITR grant CCR-0121555.

1 Introduction

Alice and Bob are given a bit each and they wish to compute the AND of their bits by exchanging messages that reveal as little information about their bits as possible. In this paper we address problems of this kind, where we study the amount of information revealed in a communication protocol. Our investigations lead to a new lower bound method in communication complexity.

Communication complexity [Yao79] quantifies the amount of communication required among two or more players to compute a function, where each player holds only a portion of the function's input. This framework has been used to solve a variety of problems in diverse areas, ranging from circuit complexity and time-space tradeoffs to pseudorandomness—see [KN97]. Some recent applications of communication complexity arise in the areas of massive data set algorithms (see below) and in the design of combinatorial auctions [NS01].

A computation model that has been very useful for designing efficient algorithms for massive data sets is the *data stream* model. A data stream algorithm makes a few passes (usually one) over its input and is charged for the amount of read-write workspace it uses. Using randomization and approximation, space-efficient data stream algorithms have been developed for many problems [AMS99, FKSV02, GMMO00, Ind00, GGI$^+$02, AJKS02]. The data stream model generalizes the restrictive read-once oblivious branching program model for which strong lower bounds are known [Bry86, Weg87]; however, since data stream algorithms are allowed to be both probabilistic and approximate, proving space lower bounds for *natural* problems is challenging.

Communication complexity offers a framework in which one can obtain non-trivial space lower bounds for data stream algorithms. The relationship between communication complexity and the data stream model is natural—the workspace of the data stream algorithm corresponds to the amount of communication in a suitable communication protocol. Lower bounds for data stream algorithms have been shown both via generalization of existing methods (e.g., [AMS99]) and by the invention of new techniques (e.g., [SS02]).

1.1 Results

We develop a novel and powerful method for obtaining lower bounds for randomized communication complexity. We use this method to derive lower bounds for communication complexity problems arising in the data stream context.

(1) In the *multi-party set-disjointness* problem $\textsc{disj}_{n,t}$, there are t players and each player is given a subset of $[n]$ with the following promise: either the sets are pairwise disjoint (NO instances) or they have a unique common element but are otherwise disjoint (YES

instances). We show that the randomized communication complexity of this problem is $\Omega(n/t^2)$. Previously, Alon, Matias, and Szegedy[AMS99] had proved an $\Omega(n/t^4)$ bound, extending the $\Omega(n)$ bound for two-party set-disjointness [KS92, Raz92]. The best upper bound for this problem in the one-way communication model is $O(n/t)$ [CKS03]. In the one-way model (where each player sends exactly one message to the next player) we show a nearly optimal lower bound of $\Omega(n/t^{1+\epsilon})$ for arbitrarily small ϵ.

Our lower bound result in the one-way model implies the following: we obtain the first super-logarithmic (in fact, $n^{\Omega(1)}$) space lower bounds for approximating the k-th frequency moment F_k for any real $k > 2$ in the data stream model[1]. This resolves the conjecture of Alon, Matias, and Szegedy [AMS99], who showed an $\Omega(n^{1-5/k})$ lower bound for constant factor approximation of F_k, $k > 5$. We show that approximating F_k, $k > 2$, to within constant factors requires $\Omega(n^{1-(2+\gamma)/k})$ space, for any constant $\gamma > 0$. For $k > 2$, the best known space upper bound for F_k is $\tilde{O}(n^{1-1/k})$ [AMS99]. Since our lower bound is essentially optimal for the one-way model, closing this gap would require either a better algorithm or a different lower bound method for the frequency moment problem. Similarly, using the lower bound in the general communication model, we show that any data stream algorithm for approximating F_k that makes a *constant* number of passes requires $\Omega\left(n^{1-3/k}\right)$ space.

(2) In the L_∞ *promise problem*, Alice and Bob are given integers $\mathbf{x}, \mathbf{y} \in [0, m]^n$, respectively. The promise is that either $|\mathbf{x} - \mathbf{y}|_\infty \leq 1$ (YES instances) or $|\mathbf{x} - \mathbf{y}|_\infty \geq m$ (NO instances). We show that the randomized communication complexity of this problem is $\Omega(n/m^2)$. This solves the open problem of Saks and Sun [SS02], who showed this bound for the restricted one-way model.

A consequence of this result is a lower bound for approximating L_p distances for $p > 2$: approximating the L_p distance between n-dimensional vectors to within a factor of n^ϵ requires $\Omega(n^{1-4\epsilon-2/p})$ space in the data stream model for any constant number of passes over the input. This bound is optimal for $p = \infty$. The communication complexity lower bound of [SS02] gives a similar bound for the one-pass data stream model.

1.2 Methodology

Our method proceeds by first decomposing the original function into simpler "primitive" functions, together with an appropriate "composer" function. For example, the two-party set-disjointness function can be written in terms of n two-bit AND functions, one for each coordinate. By computing each AND function separately, we trivially obtain a protocol to compute disjointness. The direct sum question for communication protocols [KRW95] asks

[1]For a finite sequence $\mathbf{a} = a_1, a_2, \ldots$, where each element belongs to $[n]$, and for $j \in [n]$, let $f_j(\mathbf{a})$ denote the number of times j occurs in $\mathbf{a}$. The k-th frequency moment $F_k(\mathbf{a})$ is defined as $\sum_{j \in [n]} f_j^k(\mathbf{a})$.

whether there is a protocol with considerably less communication. We consider a related question, namely, the direct sum property for the information content of the transcripts of the protocol. We formalize this idea through the notion of *information cost* of a communication protocol, which measures the amount of information revealed by the transcript about the inputs. The *information complexity* of a function is the minimum information cost incurred by any protocol that computes the function; this measure is a lower bound on the communication complexity of a function. This concept was recently introduced by Chakrabarti, Shi, Wirth, and Yao [CSWY01] in the context of simultaneous messages communication complexity; it is also implicit in the works of Ablayev [Abl96] and Saks and Sun [SS02] (see also [BCKO93]). We give an appropriate generalization of information complexity for general communication models; the highlight of our generalization is that it admits a direct sum theorem. Thus, any correct protocol for disjointness must reveal in its transcript enough information to compute each of the constituent AND functions. This reduces our task to proving lower bounds for the AND function.

In carrying out an information complexity lower bound, we would like to create an input distribution that is intuitively hard for any communication protocol. It turns out that for many natural examples, these distributions necessarily have a non-product structure. This is one of the main obstacles to extending the direct sum methodology of [CSWY01] to general communication protocols; their work addresses the more restrictive case of simultaneous message protocols. In the proof technique of [SS02], the issue of such non-product distributions causes significant complications; they resolve this difficulty for the one-way model by using tools from information theory and Fourier analysis. We approach this problem by expressing the non-product distribution as a convex combination of product distributions; this approach has been previously considered for other problems such as the distributional complexity of set-disjointness [Raz92] and the parallel repetition theorem [Raz98]. The novelty of our method lies in extending the definition of information complexity to allow conditioning so that it admits a direct sum decomposition.

The direct sum theorem reduces our task to proving information complexity lower bounds for primitive (single coordinate) functions. Existing methods for communication complexity seem unsuitable for this task, since randomized protocols can use many bits of communication but reveal little information about their inputs. Our solution is based on considering probability distributions induced on transcripts, and relating these distributions via several statistical distance measures. In particular, the *Hellinger distance* [LY90], extensively studied in statistical decision theory, plays a crucial role in the proofs. We derive new properties of the Hellinger distance between distributions arising in communication complexity. In particular, we show that it satisfies a "cut-and-paste" property and an appropriate Pythagorean inequality; these are crucial to the proofs of the one-coordinate lower bounds.

Our result for the multi-party set-disjointness in the general communication complexity model is not tight. This is due to a limitation in our proof technique and can be attributed to the fact that the square of the Hellinger distance satisfies only a weak form of triangle inequality. This leads us to consider generalizations of the Hellinger distance, which, combined with the Markovian structure of one-way protocols, allows us to derive near-triangle inequalities. To the best of our knowledge, this is the first proof technique for *multi-party* one-way protocols—a model particularly relevant to data stream computations.

Related developments. By using the direct sum paradigm of this work, together with sharper analytical methods to obtain information complexity lower bounds for "primitive" functions, Chakrabarti, Khot, and Sun [CKS03] have obtained essentially optimal bounds for the communication complexity of the multi-party set-disjointness problem in the general and one-way communication models. Jayram [unpublished work, 2003] has shown that the information complexity methodology of this work yields lower bounds for distributional communication complexity as well. Jayram, Kumar, and Sivakumar [JKS03] have extended the methods of this paper to obtain new separations between nondeterministic/co-nondeterministic communication complexity and two-sided error randomized communication complexity. Jain, Radhakrishnan, and Sen [JRS03] have used the direct sum methodology to obtain quantum communication complexity lower bounds for set-disjointness.

Organization. Section 2 contains the preliminaries. In Section 3, we derive the lower bounds for data stream algorithms by applying the communication complexity lower bounds. In Section 4, we introduce the notions of information complexity and conditional information complexity. In Section 5, we present the direct sum theorem for conditional information complexity, and illustrate it via the set-disjointness problem in the two-party (general) communication complexity model. In Section 6, we describe the connection between communication complexity and "information statistics," a term that we coin to loosely describe the interplay between information theory and distances between probability distributions. As an illustration of our techniques, we prove an $\Omega(1)$ lower bound on the information complexity of the AND of two bits. Section 7 deals with the multi-party set-disjointness problem, and presents lower bounds in the general and one-way communication models. Section 8 contains the communication lower bound for the L_∞ promise problem. The Appendix contains results about various statistical notions of divergences between probability distributions that we use in the paper, including some technical lemmas that we prove.

2 Preliminaries

Communication complexity. In the two-party randomized communication complexity model [Yao79] two computationally all-powerful probabilistic players, Alice and Bob, are required to jointly compute a function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$. Alice is given $x \in \mathcal{X}$, Bob is given $y \in \mathcal{Y}$, and they exchange messages according to a shared protocol Π. For a fixed input pair (x, y), the random variable $\Pi(x, y)$ denotes the message transcript obtained when Alice and Bob follow the protocol Π on inputs x and y (the probability is over the coins of Alice and Bob). A protocol Π is called a *δ-error protocol for f*, if there exists a function Π_{out} such that for all input pairs (x, y), $\Pr[\Pi_{\text{out}}(\Pi(x, y)) = f(x, y)] \geq 1 - \delta$. The *communication cost* of Π, denoted by $|\Pi|$, is the maximum length of $\Pi(x, y)$ over all x, y, and over all random choices of Alice and Bob. The *δ-error randomized communication complexity of f*, denoted $R_\delta(f)$, is the cost of the best δ-error protocol for f.

Communication complexity can also deal with functions over a partial domain: $f : \mathcal{L} \to \mathcal{Z}$, $\mathcal{L} \subseteq \mathcal{X} \times \mathcal{Y}$. In this case, we will assume that any protocol for f is well-defined for *any* input pair (x, y), even if this pair does not belong to the domain $\mathcal{L}$. (This can be achieved by letting the players transmit the special symbol '*' and halt the protocol whenever they cannot continue executing the protocol.) Also, without loss of generality, we will assume the protocol always produces transcripts of the same length.

The model can be easily generalized to handle an arbitrary number of players t, who compute a function $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_t \to \mathcal{Z}$. Here, the i-th player is given $x_i \in \mathcal{X}_i$, and the players exchange messages according to some fixed protocol. A restricted model of communication is the *one-way communication model* [PS84, Abl96, KNR99], in which the i-th player sends exactly one message throughout the protocol to the $(i+1)$-st player (we define $t + 1 = 1$). We denote the *δ-error one-way communication complexity of f* by $R_\delta^{\text{1-way}}(f)$.

All our lower bounds will be proved in the following stronger model: all messages are written on a shared "blackboard," which is visible to all the players. In the one-way model, this is tantamount to saying that the players write their messages in turn, from player 1 to player t, where each message could depend on all previous messages written.

Notation. Throughout the paper we denote random variables in upper case, and vectors in boldface. For a random variable X and a distribution ν, we use $X \sim \nu$ to denote that X is distributed according to ν. Let $\mathbf{X} \sim \mu$ be a vector random variable. We say that μ is a *product* distribution if the components of $\mathbf{X}$ are mutually independent of each other. For example, the distribution $\mu = \nu^n$ obtained by taking n independent copies of ν is a product distribution. For a random variable $\Phi(z)$ on a set Ω, we write Φ_z to denote the distribution of $\Phi(z)$, i.e., $\Phi_z(\omega) = \Pr[\Phi(z) = \omega]$, for every $\omega \in \Omega$. We denote by $[n]$ the set $\{1, \ldots, n\}$, and by $[0, m]$ the set $\{0, \ldots, m\}$. All logarithms are to the base 2.

Information theory. Let μ be a distribution on a finite set Ω and let $X \sim \mu$. The *entropy* of X is defined by

$$\mathrm{H}(X) = \sum_{\omega \in \Omega} \mu(\omega) \log \frac{1}{\mu(\omega)}.$$

The *conditional entropy* of X given Y is

$$\mathrm{H}(X \mid Y) = \sum_{y} \mathrm{H}(X \mid Y = y) \Pr[Y = y],$$

where $\mathrm{H}(X \mid Y = y)$ is the entropy of the conditional distribution of X given the event $\{Y = y\}$. The *joint entropy* of two random variables X and Y is the entropy of their joint distribution and is denoted $\mathrm{H}(X, Y)$.

The *mutual information* between X and Y is $\mathrm{I}(X \,;\, Y) = \mathrm{H}(X) - \mathrm{H}(X \mid Y) = \mathrm{H}(Y) - \mathrm{H}(Y \mid X)$. The *conditional mutual information* between X and Y conditioned on Z is $\mathrm{I}(X \,;\, Y \mid Z) = \mathrm{H}(X \mid Z) - \mathrm{H}(X \mid Y, Z)$. Equivalently, it can be defined as

$$\mathrm{I}(X \,;\, Y \mid Z) = \sum_{z} \mathrm{I}(X \,;\, Y \mid Z = z) \Pr[Z = z],$$

where $\mathrm{I}(X \,;\, Y \mid Z = z)$ is the mutual information between the conditional distributions of X and Y given the event $\{Z = z\}$.

We use several basic properties of entropy and mutual information in the paper, which we summarize below (proofs can be found in Chapter 2 of [CT91]).

Proposition 2.1 (Basic properties of entropy). *Let X, Y be random variables.*

1. *If X takes on at most s values, then $0 \leq \mathrm{H}(X) \leq \log s$.*
2. $\mathrm{I}(X \,;\, Y) \geq 0$.
3. *Subadditivity:* $\mathrm{H}(X, Y) \leq \mathrm{H}(X) + \mathrm{H}(Y)$*; equality if and only if X and Y are independent.*
4. *Subadditivity of conditional entropy:* $\mathrm{H}(X, Y \mid Z) \leq \mathrm{H}(X \mid Z) + \mathrm{H}(Y \mid Z)$*; equality if and only if X and Y are independent conditioned on Z.*
5. *Data processing inequality: if random variables X and Z are conditionally independent given Y, then* $\mathrm{I}(X \,;\, Y \mid Z) \leq \mathrm{I}(X \,;\, Y)$.

3 Data stream lower bounds

3.1 Frequency moments

Given a finite sequence of integers $\mathbf{a} = \mathbf{a}_1, \mathbf{a}_2, \ldots, \in [n]$, the frequency of $j \in [n]$ is $f_j = |\{i \mid \mathbf{a}_i = j\}|$. For $k \geq 0$, the k-th frequency moment $F_k(\mathbf{a})$ is defined as $\sum_{j=1}^{n} f_j^k$ (taking $0^0 = 1$).

For $k = 2$, Alon, Matias, and Szegedy [AMS99] presented a data stream algorithm that estimates F_2 to within a multiplicative error of $1 \pm \epsilon$ using space which is logarithmic in n and polynomial in $1/\epsilon$. For $k > 2$ their algorithms use space $\tilde{O}(n^{1-1/k})$ (and polynomial in $1/\epsilon$). They also showed that approximating F_k to within constant factors requires space $\Omega(n^{1-5/k})$ in the data stream model. This implies that for $k > 5$, approximating F_k requires polynomial space.

We show that approximating F_k requires space $\Omega(n^{1-(2+\gamma)/k})$ for arbitrarily small $\gamma > 0$. This shows that for any $k > 2$, approximating F_k requires polynomial space, affirming a conjecture of Alon, Matias, and Szegedy. In order to prove the space lower bound we will adapt the reduction of [AMS99] to our case.

Theorem 3.1. *For any $k > 2$ and $\gamma > 0$, any (one-pass) data stream algorithm that approximates F_k to within a constant factor with probability at least $3/4$ requires $\Omega\left(n^{1-(2+\gamma)/k}\right)$ space. For the same problem, any data stream algorithm that makes a constant number of passes requires $\Omega\left(n^{1-3/k}\right)$ space.*

Proof. Let $\mathcal{A}$ be an s-space data stream algorithm that approximates F_k to within $1 \pm \epsilon$ multiplicative error with confidence $1 - \delta$, where $0 < \delta < 1/4$. We use $\mathcal{A}$ to construct a δ-error one-way protocol for $\text{DISJ}_{n,t}$, where $t = ((1 + 3\epsilon)n)^{1/k}$.

Recall that the inputs of $\text{DISJ}_{n,t}$ are t subsets $S_1, \ldots, S_t \subseteq [n]$ with the following promise:

NO instances: for $i \neq j$, $S_i \cap S_j = \emptyset$;

YES instances: there exists $x \in [n]$ such that for all $i \neq j$, $S_i \cap S_j = \{x\}$.

The sets translate into a data stream in the following way: first all the elements of S_1, then all the elements of S_2, and so forth.

The protocol for $\text{DISJ}_{n,t}$ simulates the algorithm $\mathcal{A}$ as follows: the first player starts the execution by running $\mathcal{A}$ on the elements of S_1. When $\mathcal{A}$ has finished processing all elements of S_1, she transmits the content of the memory of $\mathcal{A}$ ($O(s)$ bits) to the second player. The second player resumes the execution of $\mathcal{A}$ on her part of the stream (the elements of S_2) and sends the memory of $\mathcal{A}$ to the third player. At the end of the execution, Player t obtains B, the output of $\mathcal{A}$. If $B \leq (1 + \epsilon)n$, then Player t sends to Player $t + 1$ the bit "0" (meaning the sets are disjoint) and otherwise, she sends the bit "1" (meaning the sets intersect).

Clearly, the protocol is one-way. We next prove that the bit Player t sends to Player $t + 1$ is indeed $\text{DISJ}_{n,t}$ with probability at least $1 - \delta$. If the input sets are disjoint, then each element has a frequency of at most one in the stream, and therefore F_k is at most n. On the other hand, if the sets are uniquely intersecting, then there is at least one element whose frequency is t, and therefore F_k is at least $t^k = (1+3\epsilon)n$. Since $\mathcal{A}$ produces an answer B that, with probability at least $1-\delta$, is in the interval $((1-\epsilon)F_k, (1+\epsilon)F_k)$, it follows that if the sets are disjoint, with probability $1-\delta$, $B \leq n(1+\epsilon)$, and if the sets are uniquely intersecting, then with probability $1-\delta$, $B \geq (1-\epsilon)(1+3\epsilon)n > (1+\epsilon)n$. Thus, our protocol is correct on any input with probability at least $1-\delta$.

We next derive a lower bound on s. Note that the protocol uses $O(s(t-1)+1) = O(st)$ bits of communication. By Theorem 7.1, part (2), this communication is at least $\Omega(n/t^{1+\gamma}) = \Omega(n^{1-(1+\gamma)/k})$. Therefore, $s = \Omega(n^{1-(2+\gamma)/k})$.

The proof for a constant number of passes is similar. The main difference is that now we use an ℓ-pass s-space data stream algorithm $\mathcal{A}$ for F_k to construct a t-player multi-round protocol for $\text{DISJ}_{n,t}$. In the end of each pass, the last player sends the content of the memory back to the first player. Thus the total communication is at most ℓst. Here we use the lower bound for the general communication complexity of $\text{DISJ}_{n,t}$ (Theorem 7.1, part (1)) to derive the data stream space lower bound. □

3.2 L_p distances

Theorem 3.2. *For any $p > 0$ (including $p = \infty$) and for ϵ such that $0 < \epsilon < \frac{1}{4} - \frac{1}{2p}$, any data stream algorithm that makes a constant number of passes over its input and approximates the L_p distance between two vectors in $[0,m]^n$ to within a factor of n^ϵ with probability at least $3/4$ requires $\Omega(n^{1-4\epsilon-2/p})$ space.*

Proof. Consider first the problem of approximating the L_∞ distance between two vectors in the communication complexity model. That is, Alice is given $\mathbf{x} \in [0,m]^n$ and Bob is given $\mathbf{y} \in [0,m]^n$, and they are required to find a value B s.t. $(1/n^\epsilon)\|\mathbf{x}-\mathbf{y}\|_\infty \leq B \leq n^\epsilon\|\mathbf{x}-\mathbf{y}\|_\infty$. Clearly, any protocol to solve this problem is immediately a protocol to solve the L_∞ promise problem for any $m > n^{2\epsilon}$: distinguishing between the cases $\|\mathbf{x}-\mathbf{y}\|_\infty \leq 1$ and $\|\mathbf{x}-\mathbf{y}\|_\infty = m$. Therefore, by Theorem 8.1, this problem requires $\Omega(n^{1-4\epsilon})$ communication.

We now translate this bound to the communication complexity of approximating the L_p distance. Using the relationship between norms, we have that

$$\|\mathbf{x}-\mathbf{y}\|_\infty \leq \|\mathbf{x}-\mathbf{y}\|_p \leq n^{1/p}\|\mathbf{x}-\mathbf{y}\|_\infty,$$

or equivalently, the quantity $n^{-1/(2p)}\|\mathbf{x}-\mathbf{y}\|_p$ approximates $\|\mathbf{x}-\mathbf{y}\|_\infty$ to within a (multiplicative) factor of $n^{1/(2p)}$. Thus, approximating the L_p norm to within a factor of n^ϵ implies an $n^{\epsilon+1/(2p)}$-approximation to L_∞. Using the lower bound for approximating the L_∞ distance, we obtain an $\Omega(n^{1-4\epsilon-2/p})$ communication lower bound for approximating the L_p distance to within a factor of n^ϵ.

Suppose there exists an s-space data stream algorithm with a constant number of passes that approximates the L_p distance to within a factor of n^ϵ with confidence 3/4. Similar to the proof of Theorem 3.1, this yields a communication complexity protocol that approximates the L_p distance with the same approximation factor and the same confidence, and whose communication cost is $O(s)$. Thus, $s = \Omega(n^{1-4\epsilon-2/p})$. □

4 Information complexity

In this section we define the fundamental notions of information measures associated with communication protocols alluded to in the introduction. As the main illustration of our definitions and techniques, we consider the two-party set-disjointness problem. We will continue the illustration in Section 5 and Section 6, resulting in a simple proof of the $\Omega(n)$ lower bound for the set-disjointness problem.

Our lower bound method is built on an information-theoretic measure of communication complexity, called *information complexity*, defined with respect to a given distribution over the inputs to the function; our definitions generalize similar notions that were considered previously [CSWY01, BCKO93, Abl96, SS02]. The discussion that follows is in the framework of two-party communication complexity; the generalization to an arbitrary number of players is straightforward.

Fix a set $\mathcal{K} \subseteq \mathcal{X}^n \times \mathcal{Y}^n$ of legal inputs and a function $f : \mathcal{K} \to \{0, 1\}$.

> *In the set-disjointness problem, Alice and Bob hold, respectively, the characteristic vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *of two subsets of* $[n]$. $\textsc{disj}(\mathbf{x}, \mathbf{y})$ *is defined to be* 1 *if and only if* $\mathbf{x} \cap \mathbf{y} \neq \emptyset$.

Informally, information cost is the amount of information one can learn about the inputs from the transcript of messages in a protocol on these inputs. Formally it is defined as follows:

Definition 4.1 (Information cost of a protocol). Let Π be a randomized protocol whose inputs belong to $\mathcal{K}$. Let μ be a distribution on $\mathcal{K}$, and suppose $(\mathbf{X}, \mathbf{Y}) \sim \mu$. The *information cost of* Π *with respect to* μ is defined as $\mathrm{I}(\mathbf{X}, \mathbf{Y} \,;\, \Pi(\mathbf{X}, \mathbf{Y}))$.

Definition 4.2 (Information complexity of a function). The δ-error *information complexity* of f with respect to a distribution μ, denoted $\mathrm{IC}_{\mu,\delta}(f)$, is defined as the minimum information cost of a δ-error protocol for f with respect to μ.

Proposition 4.3. *For any distribution* μ *and error* $\delta > 0$, $R_\delta(f) \geq \mathrm{IC}_{\mu,\delta}(f)$.

Proof. Let Π denote the best δ-error protocol for f in terms of communication. Let $(\mathbf{X},\mathbf{Y}) \sim \mu$. If $|\Pi|$ denotes the length of the longest transcript produced by the protocol Π (on any input), then we have:

$$R_\delta(f) = |\Pi| \geq \mathrm{H}(\Pi(\mathbf{X},\mathbf{Y})) \geq \mathrm{I}(\mathbf{X},\mathbf{Y} \; ; \; \Pi(\mathbf{X},\mathbf{Y})) \geq \mathrm{IC}_{\mu,\delta}(f). \qquad \square$$

Suppose $\mathcal{L} \subseteq \mathcal{X} \times \mathcal{Y}$, and suppose $f : \mathcal{L}^n \to \{0,1\}$ can be expressed in terms of a simpler "primitive" $h : \mathcal{L} \to \{0,1\}$ applied to each coordinate of the input pair $(\mathbf{x},\mathbf{y})$. (This notion will formalized later; as an example, note that $\textsc{disj}(\mathbf{x},\mathbf{y}) = \bigvee_{i\in[n]}(\mathbf{x}_i \wedge \mathbf{y}_i)$, where the primitive h is the AND of two bits.) If f depends symmetrically on the primitive in each coordinate, then we expect that any protocol for f must implicitly solve each instance of the primitive h. Further, if the distribution μ on $\mathcal{L}^n$ is the product of independent copies of a distribution ν on $\mathcal{L}$, then one can hope to show that $\mathrm{IC}_{\mu,\delta}(f) \geq n \cdot \mathrm{IC}_{\nu,\delta}(h)$—a direct sum property for information complexity.

The main technical obstacle to proving this result is that the distribution μ is not necessarily a product distribution, i.e. if $(\mathbf{X},\mathbf{Y}) \sim \mu$, then $\mathbf{X}$ and $\mathbf{Y}$ may not be independent. This is because ν need not be a product distribution on $\mathcal{X} \times \mathcal{Y}$ (although μ is the product of n copies of ν). In fact, for set-disjointness, it becomes essential to consider non-product distributions to obtain an $\Omega(n)$ lower bound [BFS86]. To handle this, we proceed as follows. Let T denote an auxiliary random variable with domain $\mathcal{T}$, and let η denote the joint distribution of $((\mathbf{X},\mathbf{Y}),T)$. The choice of T will be made such that conditioned on T, $\mathbf{X}$ and $\mathbf{Y}$ are independent. In this case, we say that η is a *mixture of product distributions*.

In the above discussion, suppose ν is non-product and $\mu = \nu^n$. Let $(X,Y) \sim \nu$. We will define a random variable D such that X and Y are independent, conditioned on D. Let ζ denote the joint distribution of $((X,Y),D)$. It is clear that $\eta = \zeta^n$ is a mixture of product distributions, whose marginal distribution on $\mathcal{L}^n$ is $\nu^n = \mu$. A useful consequence is that if $((\mathbf{X},\mathbf{Y}),\mathbf{D}) \sim \eta$, then the coordinates $\{(\mathbf{X}_j,\mathbf{Y}_j)\}_{j\in[n]}$ are mutually independent of each other, and this continues to hold even when conditioned on $\mathbf{D}$.

For set-disjointness, we will use the non-product distribution ν on the inputs given by $\nu(0,0) = 1/2$, $\nu(0,1) = \nu(1,0) = 1/4$. Let D denote a random variable with uniform distribution on $\{\mathrm{A},\mathrm{B}\}$. If $D = \mathrm{A}$, then let $X = 0$ and let Y be a uniform element of $\{0,1\}$; if $D = \mathrm{B}$, then let $Y = 0$ and let X be a uniform element of $\{0,1\}$. It is clear that conditioned on D, X and Y are independent, and $(X,Y) \sim \nu$. Therefore, the joint distribution ζ of $((X,Y),D)$ is a mixture of product distributions.

Definition 4.4 (Conditional information cost). Let Π be a randomized protocol whose inputs belong to $\kappa \subseteq \mathcal{X}^n \times \mathcal{Y}^n$. Suppose $((\mathbf{X},\mathbf{Y}),T) \sim \eta$, and that η is a mixture of product distributions on $\kappa \times \mathcal{T}$. The *conditional information cost of* Π *with respect to* η is defined as $\mathrm{I}(\mathbf{X},\mathbf{Y} \; ; \; \Pi(\mathbf{X},\mathbf{Y}) \mid T)$.

Definition 4.5 (Conditional information complexity). The δ-error *conditional information complexity of f with respect to η*, denoted by $\mathrm{CIC}_{\eta,\delta}(f)$, is defined as the minimum conditional information cost of a δ-error protocol for f with respect to η.

Proposition 4.6. *Let μ be a distribution on κ, the set of inputs to f. If η is a mixture of product distributions on $\kappa \times \mathcal{T}$ such that the marginal distribution on κ is μ, then $\mathrm{IC}_{\mu,\delta}(f) \geq \mathrm{CIC}_{\eta,\delta}(f)$.*

Proof. Let Π be a protocol whose information cost equals $\mathrm{IC}_{\mu,\delta}(f)$. Let $((\mathbf{X}, \mathbf{Y}), T) \sim \eta$. Note that $(\mathbf{X}, \mathbf{Y}) \sim \mu$. Since $\Pi(\mathbf{X}, \mathbf{Y})$ is conditionally independent of T given $\mathbf{X}, \mathbf{Y}$ (because the private coins of Π are independent of T), the data processing inequality implies: $\mathrm{IC}_{\mu,\delta}(f) = \mathrm{I}(\mathbf{X}, \mathbf{Y} \,;\, \Pi(\mathbf{X}, \mathbf{Y})) \geq \mathrm{I}(\mathbf{X}, \mathbf{Y} \,;\, \Pi(\mathbf{X}, \mathbf{Y}) \mid T) \geq \mathrm{CIC}_{\eta,\delta}(f)$. $\square$

Corollary 4.7 (of Propositions 4.3 and 4.6). *Let $f : \mathcal{K} \to \{0, 1\}$, and let η be any distribution on $\mathcal{K} \times \mathcal{T}$ for some set $\mathcal{T}$. Then $R_\delta(f) \geq \mathrm{CIC}_{\eta,\delta}(f)$.*

Remarks. In general, the choice of the random variable T in expressing η as a mixture of product distributions is not unique. We will choose one where the entropy of T is not too large. By a more precise application of the data processing inequality, it can also be seen that the difference between $\mathrm{IC}_{\mu,\delta}(f)$ and $\mathrm{CIC}_{\eta,\delta}(f)$ is at most $\mathrm{H}(T)$; thus the degradation in the lower bound is not much as long as T has small entropy.

5 A direct sum theorem for conditional information complexity

We now turn to the development of the direct sum theorem for the conditional information complexity of decomposable functions. Let Π be a δ-error protocol for $f : \mathcal{L}^n \to \{0, 1\}$, for some $\mathcal{L} \subseteq \mathcal{X} \times \mathcal{Y}$. Let ζ be a mixture of product distributions on $\mathcal{L} \times \mathcal{D}$, let $\eta = \zeta^n$, and suppose $((\mathbf{X}, \mathbf{Y}), \mathbf{D}) \sim \eta$. First, we show that the conditional information cost of the protocol Π with respect to η can be decomposed into information about each of the coordinates. This reduces our task to proving lower bounds for the coordinate-wise information-theoretic quantities. Next, we formalize the notion of decomposing a function into primitive functions. By imposing a further restriction on the input distribution, we then show that each coordinate-wise information quantity itself is lower bounded by the conditional information complexity of the primitive function. This will result in the direct sum theorem.

Lemma 5.1 (Information cost decomposition lemma). *Let Π be a protocol whose inputs belong to $\mathcal{L}^n$, for some $\mathcal{L} \subseteq \mathcal{X} \times \mathcal{Y}$. Let ζ be a mixture of product distributions* *on $\mathcal{L} \times \mathcal{D}$, let $\eta = \zeta^n$, and suppose $((\mathbf{X}, \mathbf{Y}), \mathbf{D}) \sim \eta$. Then,* $\mathrm{I}(\mathbf{X}, \mathbf{Y} \ ; \ \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D}) \geq \sum_j \mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \ ; \ \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D})$.

Proof. Abbreviating $\Pi(\mathbf{X}, \mathbf{Y})$ by Π, note that by definition, $\mathrm{I}(\mathbf{X}, \mathbf{Y} \ ; \ \Pi \mid \mathbf{D}) = \mathrm{H}(\mathbf{X}, \mathbf{Y} \mid \mathbf{D}) - \mathrm{H}(\mathbf{X}, \mathbf{Y} \mid \Pi, \mathbf{D})$. Now, observe that $\mathrm{H}(\mathbf{X}, \mathbf{Y} \mid \mathbf{D}) = \sum_j \mathrm{H}(\mathbf{X}_j, \mathbf{Y}_j \mid \mathbf{D})$, since the pairs $(\mathbf{X}_j, \mathbf{Y}_j)$, $j \in [n]$, are independent of each other conditioned on $\mathbf{D}$. By the subadditivity of conditional entropy, $\mathrm{H}(\mathbf{X}, \mathbf{Y} \mid \Pi, \mathbf{D}) \leq \sum_j \mathrm{H}(\mathbf{X}_j, \mathbf{Y}_j \mid \Pi, \mathbf{D})$. Thus $\mathrm{I}(\mathbf{X}, \mathbf{Y} \ ; \ \Pi \mid \mathbf{D}) \geq \sum_j \mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \ ; \ \Pi \mid \mathbf{D})$. □

Definition 5.2 (Decomposable functions). $f : \mathcal{L}^n \to \{0,1\}$ is *g-decomposable with primitive h* if it can be written as $f(\mathbf{x}, \mathbf{y}) = g(h(\mathbf{x}_1, \mathbf{y}_1), \ldots, h(\mathbf{x}_n, \mathbf{y}_n))$, for some functions $h : \mathcal{L} \to \{0,1\}$ and $g : \{0,1\}^n \to \{0,1\}$. Sometimes we simply write *f is decomposable with primitive h.*

It is easy to see that set-disjointness is OR*-decomposable with primitive* AND*:* $\textsc{disj}(\mathbf{x}, \mathbf{y}) = \bigvee_{i \in [n]} (\mathbf{x}_i \wedge \mathbf{y}_i)$. *Here $\mathcal{L} = \{0,1\}^2$, $h =$* AND*, $g =$* OR.

Other examples of decomposable functions are the following.

(1) *Inner product*: Again $\mathcal{L} = \{0,1\}^2$ and h is the AND of two bits; g is the XOR of n bits.

(2) *L_∞ promise problem*: Here $\mathcal{L} = [0, m]^2$, for some m, $h(x, y) = 1$ if $|x - y| \geq m$ and 0 if $|x - y| \leq 1$; g is the OR of n bits.

Now, we would like to lower bound the information about each coordinate by the conditional information complexity of h, that is, $\mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \ ; \ \Pi \mid \mathbf{D}) \geq \mathrm{CIC}_{\zeta,\delta}(h)$, for each j. We achieve this by presenting, for each j, a family of protocols for h that use a protocol Π for f as a subroutine, and whose average conditional information cost with respect to ζ is *exactly* $\mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \ ; \ \Pi \mid \mathbf{D})$. To facilitate this, we will further restrict the input distribution that we use to be a "collapsing distribution" for f.

Definition 5.3 (Embedding). For a vector $\mathbf{w} \in \mathcal{L}^n$, $j \in [n]$, and $u \in \mathcal{L}$, we define $\textsc{embed}(\mathbf{w}, j, u)$ to be the n-dimensional vector over $\mathcal{L}$, whose i-th component, $1 \leq i \leq n$, is defined as follows: $\textsc{embed}(\mathbf{w}, j, u)[i] = \mathbf{w}_i$ if $i \neq j$, and $\textsc{embed}(\mathbf{w}, j, u)[j] = u$. (In other words, we replace the j-th component of $\mathbf{w}$ by u, and leave the rest intact.)

Definition 5.4 (Collapsing distribution). Suppose $f : \mathcal{L}^n \to \{0,1\}$ is g-decomposable with primitive $h : \mathcal{L} \to \{0,1\}$. We call $(\mathbf{x}, \mathbf{y}) \in \mathcal{L}^n$ a *collapsing* input for f, if for every $j \in [n]$, $f(\textsc{embed}(\mathbf{x}, j, u), \textsc{embed}(\mathbf{y}, j, v)) = h(u, v)$. We call a distribution μ on $\mathcal{L}^n$ *collapsing* for f, if every $(\mathbf{x}, \mathbf{y})$ in the support of μ is a collapsing input.

Since our distribution ν for set-disjointness never places any mass on the pair $(1,1)$, it follows that for every $(\mathbf{x}, \mathbf{y})$ in the support of $\mu = \nu^n$, and for every $j \in [n]$, $\bigvee_{i \neq j} (\mathbf{x}_i \wedge \mathbf{y}_i) = 0$. Therefore, for every $(u, v) \in \{0,1\}^2$, $\textsc{disj}(\textsc{embed}(\mathbf{x}, j, u), \textsc{embed}(\mathbf{y}, j, v)) = u \wedge v$, implying that μ is a collapsing distribution for DISJ.

Informally, a collapsing input $(\mathbf{x}, \mathbf{y})$ projects f to h in each coordinate. By fixing one such $(\mathbf{x}, \mathbf{y})$, any protocol Π for f can be used to derive n different protocols for h: the j-th protocol is obtained by simply running Π on $(\text{EMBED}(\mathbf{x}, j, u), \text{EMBED}(\mathbf{y}, j, v))$, where (u, v) is the input to the protocol. Clearly, each of these protocols has the same error as Π. A collapsing distribution allows us to argue that Π is in fact the "sum" of n protocols for h.

Lemma 5.5 (Reduction lemma). *Let Π be a δ-error protocol for a decomposable function $f : \mathcal{L}^n \to \{0,1\}$ with primitive h. Let ζ be a mixture of product distributions on $\mathcal{L} \times \mathcal{D}$, let $\eta = \zeta^n$, and suppose $((\mathbf{X}, \mathbf{Y}), \mathbf{D}) \sim \eta$. If the distribution of $(\mathbf{X}, \mathbf{Y})$ is a collapsing distribution for f, then for all $j \in [n]$, $\mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \,;\, \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D}) \geq \mathrm{CIC}_{\zeta,\delta}(h)$.*

Proof. Let $\mathbf{D}_{-j}$ stand for $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_{j-1}, \mathbf{D}_{j+1}, \ldots, \mathbf{D}_n$. Since $\mathbf{D} = (\mathbf{D}_j, \mathbf{D}_{-j})$, we have $\mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \,;\, \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D}) = \mathrm{E}_{\mathbf{d}}[\mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \,;\, \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D}_j, \mathbf{D}_{-j} = \mathbf{d})]$, where $\mathbf{d}$ is indexed by $[n] \setminus \{j\}$. We will show that each term is the conditional information cost with respect to ζ of a δ-error protocol $P_{j,\mathbf{d}}$ for h, which will prove the lemma.

Notation. If $((X, Y), D) \sim \zeta$, then let ν denote the distribution of (X, Y), and for $d \in \mathcal{D}$, let ν_d denote the distribution of (X, Y), conditioned on the event $\{D = d\}$. Note that ν_d is a product distribution. Also, note that ν^n is the distribution of $(\mathbf{X}, \mathbf{Y})$, and it is a collapsing distribution for f.

The protocol $P_{j,\mathbf{d}}$ has j and $\mathbf{d}$ "hardwired" into it. Suppose (u, v) is the input to $P_{j,\mathbf{d}}$. In this protocol, Alice and Bob will simulate $\Pi(\mathbf{x}', \mathbf{y}')$, where $\mathbf{x}'$ and $\mathbf{y}'$ are values, respectively, of random variables $\mathbf{X}' = \mathbf{X}'(u, j, \mathbf{d})$ and $\mathbf{Y}' = \mathbf{Y}'(v, j, \mathbf{d})$, defined as follows. The j-th coordinates of $\mathbf{X}'$ and $\mathbf{Y}'$ will be constants, defined by $\mathbf{X}'_j = u$ and $\mathbf{Y}'_j = v$; and for $i \neq j$, $(\mathbf{X}'_i, \mathbf{Y}'_i) \sim \nu_{\mathbf{d}_i}$. Note that since $\nu_{\mathbf{d}_i}$ is a product distribution, Alice can produce $\mathbf{x}'_i$ and Bob can produce $\mathbf{y}'_i$ independently of each other using private coin tosses. Now, Alice and Bob simulate $\Pi(\mathbf{x}', \mathbf{y}')$ and output whatever it outputs.

Define $\mathbf{x}$ and $\mathbf{y}$ as follows. For $i \neq j$, $\mathbf{x}_i = \mathbf{x}'_i$ and $\mathbf{y}_i = \mathbf{y}'_i$, and $(\mathbf{x}_j, \mathbf{y}_j)$ is some value of the random variable $(X, Y) \sim \nu$. Since $(\mathbf{x}_i, \mathbf{y}_i)$, for $i \neq j$, are also values in the support of ν, it follows that $(\mathbf{x}, \mathbf{y})$ is in the support of ν^n, which is a collapsing distribution for f. This implies that $(\mathbf{x}, \mathbf{y})$ is a collapsing input for f, so $f(\mathbf{x}', \mathbf{y}') = f(\text{EMBED}(\mathbf{x}, j, u), \text{EMBED}(\mathbf{y}, j, v)) = h(u, v)$. It follows that $P_{j,\mathbf{d}}$ is a δ-error protocol for h.

Let $((U, V), D) \sim \zeta$. The conditional information cost of $P_{j,\mathbf{d}}$ with respect to ζ equals $\mathrm{I}(U, V \,;\, P_{j,\mathbf{d}}(U, V) \mid D)$. We will show that the joint distribution of $(U, V, D, P_{j,\mathbf{d}}(U, V))$ is identical to that of $(\mathbf{X}_j, \mathbf{Y}_j, \mathbf{D}_j, \Pi(\mathbf{X}, \mathbf{Y}))$ conditioned on the event $\{\mathbf{D}_{-j} = \mathbf{d}\}$. This will imply the following, which completes the proof.

$$\mathrm{I}(U, V \,;\, P_{j,\mathbf{d}}(U, V) \mid D) = \mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \,;\, \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D}_j, \mathbf{D}_{-j} = \mathbf{d}).$$

It is easy to see that for any values u, v, and d,

$$\begin{aligned}
&\Pr[U = u, V = v, D = d] \\
&\quad = \Pr[\mathbf{X}_j = u, \mathbf{Y}_j = v, \mathbf{D}_j = d] \\
&\quad = \Pr[\mathbf{X}_j = u, \mathbf{Y}_j = v, \mathbf{D}_j = d \mid \mathbf{D}_{-j} = \mathbf{d}] \qquad \text{(by independence of } \mathbf{X}_j, \mathbf{Y}_j, \text{ and } \mathbf{D}_j \text{ from } \mathbf{D}_{-j})
\end{aligned}$$

Furthermore, for any transcript τ,

$$\begin{aligned}
&\Pr[P_{j,\mathbf{d}}(U, V) = \tau \mid U = u, V = v, D = d] \\
&\quad = \Pr[P_{j,\mathbf{d}}(u, v) = \tau \mid U = u, V = v, D = d] \\
&\quad = \Pr[P_{j,\mathbf{d}}(u, v) = \tau] \qquad \text{(by independence of } P_{j,\mathbf{d}}(u,v) \text{ from } (U, V, D)) \\
&\quad = \Pr[\Pi(\mathbf{X}'(u, j, \mathbf{d}), \mathbf{Y}'(v, j, \mathbf{d})) = \tau].
\end{aligned}$$

Notice that the distribution of $(\mathbf{X}'(u, j, \mathbf{d}), \mathbf{Y}'(v, j, \mathbf{d}))$ is identical to the distribution of $(\mathbf{X}, \mathbf{Y})$ conditioned on the event $\{\mathbf{X}_j = u, \mathbf{Y}_j = v, \mathbf{D}_{-j} = \mathbf{d}\}$. Therefore, we have

$$\begin{aligned}
&\Pr[P_{j,\mathbf{d}}(U, V) = \tau \mid U = u, V = v, D = d] \\
&\quad = \Pr[\Pi(\mathbf{X}, \mathbf{Y}) = \tau \mid \mathbf{X}_j = u, \mathbf{Y}_j = v, \mathbf{D}_{-j} = \mathbf{d}] \\
&\quad = \Pr[\Pi(\mathbf{X}, \mathbf{Y}) = \tau \mid \mathbf{X}_j = u, \mathbf{Y}_j = v, \mathbf{D}_j = d, \mathbf{D}_{-j} = \mathbf{d}].
\end{aligned}$$

The last equality uses the independence of $\Pi(\mathbf{X}, \mathbf{Y})$ from $\mathbf{D}_j$, conditioned on the events $\{\mathbf{X}_j = u\}$ and $\{\mathbf{Y}_j = v\}$. ☐

Theorem 5.6 (Direct sum theorem). *Let $f : \mathcal{L}^n \to \{0, 1\}$ be a decomposable function with primitive h. Let ζ be a mixture of product distributions on $\mathcal{L} \times \mathcal{D}$, let $\eta = \zeta^n$, and suppose $((\mathbf{X}, \mathbf{Y}), \mathbf{D}) \sim \eta$. If the distribution of $(\mathbf{X}, \mathbf{Y})$ is a collapsing distribution for f, then* $\mathrm{CIC}_{\eta,\delta}(f) \geq n \cdot \mathrm{CIC}_{\zeta,\delta}(h)$.

Proof. Let Π be the optimal δ-error protocol for f in terms of conditional information cost with respect to η. If $((\mathbf{X}, \mathbf{Y}), \mathbf{D}) \sim \eta$, then we have $\mathrm{CIC}_{\eta,\delta}(f) = \mathrm{I}(\mathbf{X}, \mathbf{Y} \, ; \, \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D})$. By the information cost decomposition lemma (Lemma 5.1), this is at least $\sum_j \mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \, ; \, \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D})$. By the reduction lemma (Lemma 5.5), this is at least $n \cdot \mathrm{CIC}_{\zeta,\delta}(h)$. ☐

Corollary 5.7 (of Corollary 4.7, and Theorem 5.6). *With the notation and assumptions of Theorem 5.6,* $R_\delta(f) \geq \mathrm{CIC}_{\eta,\delta}(f) \geq n \cdot \mathrm{CIC}_{\zeta,\delta}(h)$.

For set-disjointness, $R_\delta(\textsc{disj}) \geq n \cdot \mathrm{IC}_{\zeta,\delta}(\textsc{and})$. Thus it suffices to show an $\Omega(1)$ lower bound for the conditional information complexity of the 1-bit function AND *with respect to ζ.*

6 Information complexity lower bound for primitives

The direct sum theorem of the foregoing section effectively recasts the task of proving randomized communication complexity lower bounds for many functions. Namely, the goal now is to prove conditional information complexity lower bounds for "primitive functions", where the communicating parties are given inputs from a small domain, and wish to check a fairly simple predicate. In this section, we illustrate how we accomplish this by proving an $\Omega(1)$ lower bound for the conditional information complexity of the AND function with respect to the distribution ζ. In doing so, we develop some basic connections between communication complexity, statistical distance measures, and information theory; these connections will be later used in the proofs of our main results on multi-party set-disjointness and the L_∞ problem. To aid the exposition, we state and use various Lemmas and Propositions; their proofs are collected in Section 6.1 and Appendix A.

We will show that for any randomized protocol P that correctly computes the AND function, an $\Omega(1)$ lower bound holds on $\mathrm{I}(U, V; P(U, V) \mid D)$, where $((U, V), D) \sim \zeta$. Recall that we have $\Pr[D = 0] = \Pr[D = 1] = 1/2$. We assume that for every input $(u, v) \in \{0, 1\}^2$, the protocol P computes $\textsc{and}(u, v)$ correctly with probability at least $1 - \delta$.

Let Z denote a random variable distributed uniformly in $\{0, 1\}$. Using the definition of the distribution ζ and expanding on values of D, we have

$$\begin{aligned} \mathrm{I}(U, V; P(U, V) \mid D) &= \frac{1}{2}[\mathrm{I}(U, V; P(U, V) \mid D = 0) + \mathrm{I}(U, V; P(U, V) \mid D = 1)] \\ &= \frac{1}{2}[\mathrm{I}(Z; P(0, Z)) + \mathrm{I}(Z; P(Z, 0))] \end{aligned} \tag{1}$$

In the last equality, we use the following facts. Conditioned on the event $\{D = 0\}$, U is identically 0, and V is distributed uniformly in $\{0, 1\}$; similarly, conditioned on the event $\{D = 1\}$, V is identically 0, and U is distributed uniformly in $\{0, 1\}$.

Notice that the mutual information quantities in Equation (1) are of the form $\mathrm{I}(Z; \Phi(Z))$, where Z is uniformly distributed in $\{0, 1\}$, and $\Phi(z)$ is a random variable, for each $z \in \{0, 1\}$. The next lemma provides an important passage from such quantities (and hence from information complexity) to metrics on probability distributions. The advantage of working with a metric is that it allows us the use of the triangle inequality when needed; furthermore, as will be evident from Lemmas 6.3 and 6.4 later, *Hellinger distance* turns out to be a natural choice in analyzing distributions of transcripts of communication protocols.

Definition 6.1 (Hellinger distance). The *Hellinger distance* between probability distributions P and Q on a domain Ω is defined by

$$\mathrm{h}^2(P, Q) = 1 - \sum_{\omega \in \Omega} \sqrt{P(\omega) Q(\omega)} = \sum_{\omega \in \Omega} \left(\frac{P(\omega) + Q(\omega)}{2} - \sqrt{P(\omega) Q(\omega)} \right).$$

(*Note.* The above equation defines the square of the Hellinger distance.)

For the discussion below, recall our notation that for a random variable $\Phi(z)$ on a set Ω, we write Φ_z to denote the distribution of $\Phi(z)$. The following lemma is proved in Appendix A as Lemma A.7.

Lemma 6.2. *Let $\Phi(z_1)$ and $\Phi(z_2)$ be two random variables. Let Z denote a random variable with uniform distribution in $\{z_1, z_2\}$. Suppose $\Phi(z)$ is independent of Z for each $z \in \{z_1, z_2\}$. Then,* $\mathrm{I}(Z; \Phi(Z)) \geq \mathrm{h}^2(\Phi_{z_1}, \Phi_{z_2})$.

Combining Equation (1) and Lemma 6.2, we obtain:

$$\begin{aligned} \mathrm{I}(U, V; P(U, V) \mid D) &\geq \frac{1}{2}(\mathrm{h}^2(P_{00}, P_{01}) + \mathrm{h}^2(P_{00}, P_{10})) && \text{(Lemma 6.2)} \\ &\geq \frac{1}{4}(\mathrm{h}(P_{00}, P_{01}) + \mathrm{h}(P_{00}, P_{10}))^2 && \text{(Cauchy–Schwarz)} \\ &\geq \frac{1}{4}\,\mathrm{h}^2(P_{01}, P_{10}). && \text{(Triangle inequality)} \end{aligned}$$

At this point, we have shown that the conditional information cost of P with respect to ζ is bounded from below by $\mathrm{h}^2(P_{01}, P_{10})$. This leads us to the task of lower bounding the Hellinger distance between P_{01} and P_{10}. Of the four distributions P_{00}, P_{01}, P_{10}, and P_{11} on the set of possible transcripts of P, it is natural to expect P_{11} to be quite different from the rest since $\textsc{and}(1, 1) = 1$, while the value of $\textsc{and}$ on the other three input pairs is 0. Given that $\textsc{and}(0, 1)$ and $\textsc{and}(1, 0)$ are both 0, it is not clear why these two distributions (on the set of possible transcripts of P) should be far apart. This is where the "rectangular" nature of the transcripts of communication protocols comes in. We will show that the transcript distributions on various inputs satisfy two important properties, which may be considered to be analogs of the following statement about deterministic communication protocols: if $\Pi(x, y) = \tau = \Pi(x', y')$, then $\Pi(x', y) = \tau = \Pi(x, y')$.

Lemma 6.3 (Cut and paste lemma). *For any randomized protocol Π and for any $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$,* $\mathrm{h}(\Pi_{xy}, \Pi_{x'y'}) = \mathrm{h}(\Pi_{xy'}, \Pi_{x'y})$.

Lemma 6.4 (Pythagorean lemma). *For any randomized protocol Π and for any $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$,* $\mathrm{h}^2(\Pi_{xy}, \Pi_{x'y}) + \mathrm{h}^2(\Pi_{xy'}, \Pi_{x'y'}) \leq 2\,\mathrm{h}^2(\Pi_{xy}, \Pi_{x'y'})$.

Note. Lemma 6.4 is not used in the lower bound for $\textsc{and}$; it is used only in Section 8.

Lemma 6.3 implies that $\mathrm{h}^2(P_{01}, P_{10}) = \mathrm{h}^2(P_{00}, P_{11})$, so we have:

$$\begin{aligned} \mathrm{I}(U, V; P(U, V) \mid D) &\geq \frac{1}{4}\,\mathrm{h}^2(P_{01}, P_{10}) \\ &= \frac{1}{4}\,\mathrm{h}^2(P_{00}, P_{11}). && \text{(Lemma 6.3)} \end{aligned}$$

The final point is that since $\textsc{and}(0,0) \neq \textsc{and}(1,1)$, we expect the distributions P_{00} and P_{11} to be far from each other.

Lemma 6.5. *For any δ-error protocol Π for a function f, and for any two input pairs (x,y) and (x',y') for which $f(x,y) \neq f(x',y')$, $\mathrm{h}^2(\Pi_{xy}, \Pi_{x'y'}) \geq 1 - 2\sqrt{\delta}$.*

We now have:

$$\begin{aligned}
\mathrm{CIC}_{\zeta,\delta}(\textsc{and}) &\geq \mathrm{I}(U,V; P(U,V) \mid D) \\
&\geq \frac{1}{4}\,\mathrm{h}^2(P_{00}, P_{11}) \\
&\geq \frac{1}{4}(1 - 2\sqrt{\delta}). \qquad\qquad \text{(Lemma 6.5)}
\end{aligned}$$

To sum up, we have shown:

Theorem 6.6. $R_\delta(\textsc{disj}) \geq n \cdot \mathrm{CIC}_{\zeta,\delta}(\textsc{and}) \geq \frac{n}{4}(1 - 2\sqrt{\delta})$.

6.1 Statistical structure of randomized communication protocols

We begin with a lemma that formulates the rectangular structure of the distributions on the transcripts of a randomized communication protocol. This is a probabilistic analog of the fundamental lemma of communication complexity—the set of inputs that have the same transcript in a deterministic communication protocol is a combinatorial rectangle.

Lemma 6.7. *(1) Let Π be a two-player randomized communication protocol with input set $\mathcal{L} \subseteq \mathcal{X} \times \mathcal{Y}$; let $\mathcal{T}$ denote the set of possible transcripts of Π. There exist mappings $q_1 : \mathcal{T} \times \mathcal{X} \to \mathbf{R}$, $q_2 : \mathcal{T} \times \mathcal{Y} \to \mathbf{R}$ such that for every $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and for every transcript $\tau \in \mathcal{T}$,*

$$\Pr[\Pi(x,y) = \tau] = q_1(\tau; x) \cdot q_2(\tau; y).$$

(2) Let Π be a t-player randomized communication protocol with input set $\mathcal{L} \subseteq \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_t$; let $\mathcal{T}$ denote the set of possible transcripts of Π. Let A, B be a partition of the set of players into two non-empty sets; denote by $\mathcal{X}_A$ and $\mathcal{X}_B$ the projections of $\mathcal{X}$ to the coordinates in A and in B, respectively. Then, there exist mappings $q_A : \mathcal{T} \times \mathcal{X}_A \to \mathbf{R}$, $q_B : \mathcal{T} \times \mathcal{X}_B \to \mathbf{R}$, such that for every $\mathbf{y} \in \mathcal{X}_A, \mathbf{z} \in \mathcal{X}_B$, and for every transcript $\tau \in \mathcal{T}$,

$$\Pr[\Pi(\mathbf{y},\mathbf{z}) = \tau] = q_A(\tau; \mathbf{y}) \cdot q_B(\tau; \mathbf{z}).$$

Proof. We first prove part (1). Recall that by our convention, Π is well-defined for every pair $(x,y) \in \mathcal{X} \times \mathcal{Y}$, regardless of whether it is a legal input (i.e., belongs to $\mathcal{L} \subseteq \mathcal{X} \times \mathcal{Y}$) or not.

In the proof we use the following "rectangle" property of *deterministic* communication complexity protocols (cf. [KN97], Chapter 1): for any possible transcript τ of a deterministic communication protocol with input sets $\mathcal{X}$ and $\mathcal{Y}$, the set of pairs on which the protocol's transcript equals τ is a combinatorial rectangle; that is, a set of the form $\mathcal{A} \times \mathcal{B}$ where $\mathcal{A} \subseteq \mathcal{X}$ and $\mathcal{B} \subseteq \mathcal{Y}$.

In order to apply this property to randomized protocols, we note that a randomized protocol can be viewed as a deterministic protocol if we augment the inputs of Alice and Bob with their private random strings. If a and b denote, respectively, the private coin tosses of Alice and Bob, under this view, the ("extended") input of Alice is (x, a) and that of Bob is (y, b).

For $\tau \in \mathcal{T}$, let $\mathcal{A}(\tau) \times \mathcal{B}(\tau)$ be the combinatorial rectangle that corresponds to the transcript τ in the (extended, deterministic) protocol Π. That is, for all $(\xi, \alpha) \in \mathcal{A}(\tau)$ and for all $(\eta, \beta) \in \mathcal{B}(\tau)$ (and only for such pairs), $\Pi((\xi, \alpha), (\eta, \beta)) = \tau$. For each $x \in \mathcal{X}$, define $\mathcal{A}(\tau, x) \subseteq \mathcal{A}(\tau)$ by $\mathcal{A}(\tau, x) = \{(\xi, \alpha) \in \mathcal{A}(\tau) \mid \xi = x\}$, and define $\mathcal{X}(x)$ to be the set of all pairs of the form (x, α). Similarly, define $\mathcal{B}(\tau, y)$ and $\mathcal{Y}(y)$ for each $y \in \mathcal{Y}$. Finally define $q_1(\tau; x) = |\mathcal{A}(\tau, x)|/|\mathcal{X}(x)|$ and $q_2(\tau; y) = |\mathcal{B}(\tau, y)|/|\mathcal{Y}(y)|$.

Note that on input x, y, Alice chooses a pair (x, a) from $\mathcal{X}(x)$ uniformly at random, and Bob chooses a pair (y, b) from $\mathcal{Y}(y)$ uniformly at random. For any $\tau \in \mathcal{T}$, the transcript of Π would be τ if and only if $(x, a) \in \mathcal{A}(\tau, x)$ and $(y, b) \in \mathcal{B}(\tau, y)$. Since the choices of a and b are independent, it follows that $\Pr[\Pi(x, y) = \tau] = q_1(\tau; x) \cdot q_2(\tau; y)$.

The proof for part (2) is by a straightforward reduction to part (1), obtained by letting Alice and Bob simulate the messages sent by the players in A and B, respectively. $\square$

We also formulate a special Markovian property for one-way protocols, which will be used in the proof for the multi-party set-disjointness in Section 7.

Lemma 6.8 (Markov property of one-way protocols). *Let Π be a t-player one-way randomized communication protocol with input set $\mathcal{L} \subseteq \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_t$; let $\mathcal{T}$ denote the set of possible transcripts of Π. Let $A = [1, k]$ and $B = [k+1, t]$ $(1 \leq k < t)$ be a partition of the set of players. Denote by $\mathcal{X}_A$ and $\mathcal{X}_B$ the projections of $\mathcal{X}$ to the coordinates in A and in B, respectively; similarly, denote by $\mathcal{T}_A$ and $\mathcal{T}_B$ the projections of $\mathcal{T}$ to the set of messages sent by players in A and in B, respectively. Then, for each assignment $\mathbf{y} \in \mathcal{X}_A$ there exists a distribution $p_{\mathbf{y}}$ on $\mathcal{T}_A$ and for each assignment $\mathbf{z} \in \mathcal{X}_B$ there exists a probability transition matrix $M_{\mathbf{z}}$ on $\mathcal{T}_A \times \mathcal{T}_B$, such that for every transcript $\tau = (\tau_A, \tau_B)$, where $\tau_A \in \mathcal{T}_A$, $\tau_B \in \mathcal{T}_B$,*

$$\Pr[\Pi(\mathbf{y}, \mathbf{z}) = \tau] = p_{\mathbf{y}}(\tau_A) \cdot M_{\mathbf{z}}(\tau_A, \tau_B).$$

Proof. Since Π is a one-way protocol, for any transcript $\tau = (\tau_A, \tau_B)$, τ_A depends only on the inputs and private coins of players in A; τ_B depends only on τ_A and the inputs and private coins of players in B. Thus, we can write $\Pi(\mathbf{y}, \mathbf{z}) = (\Pi_A(\mathbf{y}), \Pi_B(\mathbf{z}, \Pi_A(\mathbf{y})))$, where Π_A and Π_B are the messages sent by players in A and in B, respectively. Therefore,

$$\Pr[\Pi(\mathbf{y}, \mathbf{z}) = (\tau_A, \tau_B)] = \Pr[\Pi_A(\mathbf{y}) = \tau_A] \cdot \Pr[\Pi_B(\mathbf{z}, \tau_A) = \tau_B \mid \Pi_A(\mathbf{y}) = \tau_A].$$

Define $p_{\mathbf{y}}$ to be the distribution of $\Pi_A(\mathbf{y})$. Since the coins of players in A and players in B are independent, it follows that $\Pi_A(\mathbf{y})$ and $\Pi_B(\mathbf{z}, \tau_A)$ are independent. We obtain: $\Pr[\Pi_B(\mathbf{z}, \tau_A) = \tau_B \mid \Pi_A(\mathbf{y}) = \tau_A] = \Pr[\Pi_B(\mathbf{z}, \tau_A) = \tau_B]$. Define $M_{\mathbf{z}}$ to be the matrix whose τ_A-th row describes the distribution of $\Pi_B(\mathbf{z}, \tau_A)$. The lemma follows. □

Remark. Extending the above lemma to general protocols Π, it can be shown that for all inputs $(\mathbf{y}, \mathbf{z})$, there exist a column-stochastic matrix $M_{\mathbf{y}}$ and a row-stochastic matrix $M_{\mathbf{z}}$ such that $\Pr[\Pi(\mathbf{y}, \mathbf{z}) = \tau] = M_{\mathbf{y}}(\tau_A, \tau_B) \cdot M_{\mathbf{z}}(\tau_A, \tau_B)$. This is a slightly stronger form of Lemma 6.7.

6.1.1 Proofs of Lemmas 6.3, 6.4, and 6.5

Let Π denote a δ-error randomized protocol for a function f on $\mathcal{X} \times \mathcal{Y}$. Let $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$. We first prove the cut and paste lemma, that is, $\mathrm{h}(\Pi_{xy}, \Pi_{x'y'}) = \mathrm{h}(\Pi_{xy'}, \Pi_{x'y})$.

Proof of Lemma 6.3.

$$\begin{aligned}
&1 - \mathrm{h}^2(\Pi_{xy}, \Pi_{x'y'}) \\
&\quad = \sum_\tau \sqrt{\Pr[\Pi(x, y) = \tau] \cdot \Pr[\Pi(x', y') = \tau]} \\
&\quad = \sum_\tau \sqrt{q_1(\tau; x) \cdot q_2(\tau; y) \cdot q_1(\tau; x') \cdot q_2(\tau; y')} && \text{(Lemma 6.7)} \\
&\quad = \sum_\tau \sqrt{\Pr[\Pi(x, y') = \tau] \cdot \Pr[\Pi(x', y) = \tau]} \\
&\quad = 1 - \mathrm{h}^2(\Pi_{xy'}, \Pi_{x'y}).
\end{aligned}$$
□

Next we prove the Pythagorean lemma, that is, $\mathrm{h}^2(\Pi_{xy}, \Pi_{x'y}) + \mathrm{h}^2(\Pi_{xy'}, \Pi_{x'y'}) \le 2\,\mathrm{h}^2(\Pi_{xy}, \Pi_{x'y'})$.

Proof of Lemma 6.4. Using Lemma 6.7, we have

$$
\begin{aligned}
&\frac{1}{2}\left[\left(1-\mathrm{h}^2(\Pi_{xy},\Pi_{x'y})\right)+\left(1-\mathrm{h}^2(\Pi_{xy'},\Pi_{x'y'})\right)\right] \\
&\quad = \frac{1}{2}\sum_{\tau} \sqrt{q_1(\tau;x)\cdot q_2(\tau;y)\cdot q_1(\tau;x')\cdot q_2(\tau;y)}+\sqrt{q_1(\tau;x)\cdot q_2(\tau;y')\cdot q_1(\tau;x')\cdot q_2(\tau;y')} \\
&\quad = \sum_{\tau} \frac{q_2(\tau;y)+q_2(\tau;y')}{2}\sqrt{q_1(\tau;x)\cdot q_1(\tau;x')} \\
&\quad \geq \sum_{\tau} \sqrt{q_2(\tau;y)\cdot q_2(\tau;y')}\sqrt{q_1(\tau;x)\cdot q_1(\tau;x')} \qquad \text{(AM–GM inequality)} \\
&\quad = 1-\mathrm{h}^2(\Pi_{xy},\Pi_{x'y'}).
\end{aligned}
$$

□

Finally, we prove Lemma 6.5, that is, $\mathrm{h}^2(\Pi_{xy},\Pi_{x'y'}) \geq 1-2\sqrt{\delta}$ if $f(x,y)\neq f(x',y')$. The proof uses the well-known total variation distance between distributions, and its connection to the Hellinger distance (proved in Section A).

Definition 6.9 (Total variation distance). The *total variation distance* between probability distributions P and Q on a domain Ω is defined by

$$\mathrm{V}(P,Q)=\max_{\Omega'\subseteq\Omega}(P(\Omega')-Q(\Omega'))=\frac{1}{2}\sum_{\omega\in\Omega}|P(\omega)-Q(\omega)|.$$

Proposition 6.10. *If P and Q are distributions on the same domain, then* $\mathrm{V}(P,Q)\leq \mathrm{h}(P,Q)\sqrt{2-\mathrm{h}^2(P,Q)}$.

Proof of Lemma 6.5. Let $\mathcal{T}$ be the set of all transcripts τ on which Π outputs $f(x,y)$ (i.e., $\Pi_{\mathrm{out}}(\tau)=f(x,y)$). Since Π outputs $f(x,y)$ with probability at least $1-\delta$ on (x,y), we have $\Pr[\Pi(x,y)\in\mathcal{T}]\geq 1-\delta$; similarly, since Π outputs $f(x,y)$ with probability at most δ on (x',y'), we have $\Pr[\Pi(x',y')\in\mathcal{T}]\leq\delta$. It follows that $\mathrm{V}(\Pi_{xy},\Pi_{x'y'})\geq 1-2\delta$. The lemma follows by an application of Proposition 6.10. □

7 Multi-party set-disjointness

Let $\textsc{disj}_{n,t}(\mathbf{x_1},\ldots,\mathbf{x_t})=\bigvee_{j=1}^{n}\bigwedge_{i=1}^{t}x_{i,j}$, where the $\mathbf{x_i}$'s are n-bit vectors. Thus, $\textsc{disj}_{n,t}$ is $\textsc{or}$-decomposable, and the induced "primitive" functions are all $\textsc{and}_t$—the t-bit $\textsc{and}$. The legal inputs for $\textsc{and}_t$ are the all-zero $\mathbf{0}$, the all-one $\mathbf{1}$, and the standard unit vectors $\mathbf{e}_i$ with 1 in the i-th position[2].

[2] The definition of $\textsc{disj}_{n,t}$ also requires that $\mathbf{1}$ be assigned to at most one coordinate; this can be handled via a simple modification to the direct sum paradigm and will not be described here.

Theorem 7.1. *For any* $0 < \delta < 1/4$*, and any* $0 < \epsilon < 1$*,*

$$\begin{aligned}
&(1) \quad R_\delta(\textsc{disj}_{n,t}) \;\geq\; \frac{n}{t^2} \cdot (1 - 2\sqrt{\delta}),\\
&(2) \quad R_\delta^{\text{1-way}}(\textsc{disj}_{n,t}) \;\geq\; \frac{n}{t^{1+\epsilon}} \cdot \frac{\epsilon^2 \cdot \ln^2 2}{8} \cdot (1 - 2\sqrt{\delta}).
\end{aligned}$$

Proof. We will employ the direct sum paradigm and define an input distribution for $\textsc{disj}_{n,t}$ by defining the input distribution for $\textsc{and}_t$.

We will define random variables $(\mathbf{U}, D)$ in $\{0,1\}^t \times [t]$, with distribution ζ, as follows. The random variable D has uniform distribution on $[t]$; conditioned on the event $\{D = i\}$, $\mathbf{U}$ is uniformly distributed in $\{\mathbf{0}, \mathbf{e}_i\}$. If ν denotes the distribution of $\mathbf{U}$, it is clear that ν^n is a collapsing distribution for $\textsc{disj}_{n,t}$. Thus, all we need to prove is a lower bound on the conditional information complexity of $\textsc{and}_t$ with respect to ζ.

Let Π be any δ-error protocol that computes $\textsc{and}_t$; to keep the notation simple we will suppress any reference to the private randomness used in Π. The conditional information cost of Π with respect to ζ is now given by

$$\mathrm{I}(\mathbf{U} \,;\, \Pi(\mathbf{U}) \mid D) = \frac{1}{t}\sum_i \mathrm{I}(\mathbf{U} \,;\, \Pi(\mathbf{U}) \mid D = i). \tag{2}$$

Notice that conditioned on the event $\{D = i\}$, $\mathbf{U}$ is uniformly distributed in $\{\mathbf{0}, \mathbf{e}_i\}$, so Lemma 6.2 allows us passage to the Hellinger distance. Thus we have

$$\mathrm{I}(\mathbf{U} \,;\, \Pi(\mathbf{U}) \mid D) \geq \frac{1}{t}\sum_{i=1}^{t} \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{e}_i}). \tag{3}$$

We will provide lower bounds on the RHS of Equation (3) in terms of $\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{1}})$. By Lemma 6.5, we know that $\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{1}}) \geq 1 - 2\sqrt{\delta}$. Part (1) of the Theorem follows from Lemma 7.2, and part (2) of the Theorem follows from Lemma 7.3. $\square$

Lemma 7.2. *Let* Π *be a randomized* t*-party communication protocol with inputs from* $\{0,1\}^t$*. Then* $\sum_{i=1}^t \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{e}_i}) \;\geq\; (1/t)\, \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{1}})$*.*

Proof. The lemma is proved by a tree-induction argument. For simplicity of exposition, we assume that t is a power of 2. Let T be a complete binary tree of height $\log t$. We will label the nodes of the tree as follows. The leaves are labeled 1 through t; each internal node is labeled by the interval formed by the leaves in the sub-tree below the node. Using this labeling, we uniquely identify the node of T labeled $[a,b]$ with the t-bit input $\mathbf{e}_{[a,b]}$, which is the characteristic vector of the integer interval $[a,b] \subseteq [t]$. It is easy to see that the root is identified with the input $\mathbf{1}$ and the t leaves of the tree are identified with $\mathbf{e}_1, \ldots, \mathbf{e}_t$.

The inductive step is to prove the following: for any internal node u in T whose children are v and w, $\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_u) \leq 2 \cdot (\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_v) + \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_w))$.

Suppose $u = \mathbf{e}_{[a,b]}$, for some a, b, so that $v = \mathbf{e}_{[a,c]}$, and $w = \mathbf{e}_{[c+1,b]}$, where $c = \lfloor \frac{a+b}{2} \rfloor$. Let A denote the set of players $[1, c]$ and B denote the set of players $[c+1, t]$. Let $\mathbf{y}$ be the projection of $\mathbf{0}$ on the coordinates in A and let $\mathbf{y}'$ be the projection of u on the coordinates in A. Similarly, let $\mathbf{z}, \mathbf{z}'$ be the projections of $\mathbf{0}$ and u on the coordinates in B, respectively. Note that $v = \mathbf{y}'\mathbf{z}$ and $w = \mathbf{y}\mathbf{z}'$.

The key step in the proof is an analog of the cut and paste lemma (Lemma 6.3), applied to t-player protocols, implying that

$$\mathrm{h}(\Pi_{\mathbf{0}}, \Pi_u) = \mathrm{h}(\Pi_{\mathbf{yz}}, \Pi_{\mathbf{y'z'}}) = \mathrm{h}(\Pi_{\mathbf{yz'}}, \Pi_{\mathbf{y'z}}) = \mathrm{h}(\Pi_w, \Pi_v). \tag{4}$$

The correctness of Equation 4 can be verified analogously to the proof of Lemma 6.3, using part (2) of Lemma 6.7.

By the triangle inequality, $\mathrm{h}(\Pi_v, \Pi_w) \leq \mathrm{h}(\Pi_{\mathbf{0}}, \Pi_v) + \mathrm{h}(\Pi_{\mathbf{0}}, \Pi_w)$, which by the Cauchy–Schwarz inequality is at most $2 \cdot (\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_v) + \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_w))$. Substituting in Equation 4, we obtain $\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_u) \leq 2 \cdot (\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_v) + \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_w))$. The lemma follows. $\square$

Lemma 7.3. *Let Π be a randomized t-party one-way communication protocol with inputs from $\{0,1\}^t$. Then, for any $0 < \epsilon < 1$,*

$$\sum_{i=1}^{t} \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{e}_i}) \geq \frac{(\ln^2 2)\epsilon^2}{8\ t^{\epsilon}} \cdot \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{1}}).$$

The main idea in the proof of Lemma 7.3 is to exploit the Markovian structure of transcript distributions that arise in one-way protocols, captured by Lemma 6.8. To obtain the bound in the lemma, we use Rényi divergences, which are generalizations of the Hellinger distance. This makes the proof technically tedious, and therefore we defer it to the Appendix. Here we prove a weaker version of the lemma, which still conveys the main ideas needed to obtain the stronger bound. This weaker version yields a lower bound of $(n/t^{1+c}) \cdot (1 - 2\sqrt{\delta})$ on $R_{\delta}^{\text{1-way}}(\textsc{disj}_{n,t})$, where $c \approx 0.77155$.

Lemma 7.4. *Let Π be a randomized t-party one-way communication protocol with inputs from $\{0,1\}^t$. Then, $\sum_{i=1}^{t} \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{e}_i}) \geq (1/t^c)\, \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{1}})$, where $c = \log_2(1 + \frac{1}{\sqrt{2}}) \approx 0.77155$.*

Proof. The proof is similar to that of Lemma 7.2, where the inductive step is now the following: for any internal node u in T whose children are v and w, $\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_u) \leq (1 + 1/\sqrt{2})\,(\mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_v) + \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_w))$.

Suppose $u = \mathbf{e}_{[a,b]}$, $v = \mathbf{e}_{[a,c]}$, and $w = \mathbf{e}_{[c+1,b]}$, where $c = \lfloor \frac{a+b}{2} \rfloor$. Define the sets of players A, B and the input assignments $\mathbf{y}, \mathbf{y}', \mathbf{z}, \mathbf{z}'$ as before. Recall that $\mathbf{0} = \mathbf{yz}$, $u = \mathbf{y}'\mathbf{z}'$,

$v = \mathbf{y}'\mathbf{z}$, and $w = \mathbf{y}\mathbf{z}'$. A crucial step is to rewrite $\Pi_{\mathbf{0}}, \Pi_u, \Pi_v$, and Π_w by applying the Markov property of one-way protocols (Lemma 6.8).

Notation. For a probability vector p on Ω and a probability transition matrix M on $\Omega \times \Gamma$, let $p \circ M$ denote the distribution on $\Omega \times \Gamma$ where $(p \circ M)(i, j) = p(i) \cdot M(i, j)$.

Applying Lemma 6.8 to $\Pi_{\mathbf{0}}, \Pi_u, \Pi_v$, and Π_w, we have

$$\Pi_{\mathbf{0}} = \Pi_{\mathbf{yz}} = p_{\mathbf{y}} \circ M_{\mathbf{z}} \qquad \Pi_u = \Pi_{\mathbf{y}'\mathbf{z}'} = p_{\mathbf{y}'} \circ M_{\mathbf{z}'}$$

$$\Pi_v = \Pi_{\mathbf{y}'\mathbf{z}} = p_{\mathbf{y}'} \circ M_{\mathbf{z}} \qquad \Pi_w = \Pi_{\mathbf{yz}'} = p_{\mathbf{y}} \circ M_{\mathbf{z}'},$$

where $p_{\mathbf{y}}$ and $p_{\mathbf{y}'}$ are probability vectors, and $M_{\mathbf{z}}$ and $M_{\mathbf{z}'}$ are probability transition matrices. To complete the proof, we will show

$$\mathrm{h}^2(p_{\mathbf{y}} \circ M_{\mathbf{z}}, p_{\mathbf{y}'} \circ M_{\mathbf{z}'}) \le \left(1 + \frac{1}{\sqrt{2}}\right) \cdot (\mathrm{h}^2(p_{\mathbf{y}} \circ M_{\mathbf{z}}, p_{\mathbf{y}'} \circ M_{\mathbf{z}}) + \mathrm{h}^2(p_{\mathbf{y}} \circ M_{\mathbf{z}}, p_{\mathbf{y}} \circ M_{\mathbf{z}'})).$$

This follows from the lemma below, which is a general property of the Hellinger distance. □

Lemma 7.5. *Let p, q be probability distributions on Ω, and let M, N be probability transition matrices on $\Omega \times \Gamma$, for some Ω and Γ. Then*

$$\mathrm{h}^2(p \circ M, q \circ N) \le \left(1 + \frac{1}{\sqrt{2}}\right) \cdot (\mathrm{h}^2(p \circ M, q \circ M) + \mathrm{h}^2(p \circ M, p \circ N)).$$

Proof. Let a, b be any two probability distributions on Ω, and C, D be any two probability transition matrices on $\Omega \times \Gamma$. Let C_i and D_i denote the i-th row of C and D, respectively (note that the rows of C and D are distributions). We have:

$$\begin{aligned} \mathrm{h}^2(a \circ C, b \circ D) &= 1 - \sum_{i \in \Omega, j \in \Gamma} \sqrt{a_i C_{ij} b_i D_{ij}} = 1 - \sum_{i \in \Omega} \sqrt{a_i b_i} \sum_{j \in \Gamma} \sqrt{C_{ij} D_{ij}} \\ &= 1 - \sum_{i \in \Omega} \sqrt{a_i b_i} (1 - \mathrm{h}^2(C_i, D_i)) = \mathrm{h}^2(a, b) + \sum_{i \in \Omega} \mathrm{h}^2(C_i, D_i) \sqrt{a_i b_i} \end{aligned}$$

Define β_i to be the squared Hellinger distance between the i-th row of M and the i-th row of N. Using the above observation, we can write the three (squared) Hellinger distances as follows: $\mathrm{h}^2(p \circ M, q \circ N) = \mathrm{h}^2(p, q) + \sum_{i \in \Omega} \beta_i \sqrt{p_i q_i}$, $\mathrm{h}^2(p \circ M, q \circ M) = \mathrm{h}^2(p, q)$, and $\mathrm{h}^2(p \circ M, p \circ N) = \sum_{i \in \Omega} p_i \beta_i$.

Set $\gamma = 1/\sqrt{2}$. After minor rearrangement, it suffices to prove:

$$\sum_{i \in \Omega} \beta_i (\sqrt{p_i q_i} - (1 + \gamma) p_i) \le \gamma \, \mathrm{h}^2(p, q) = \gamma \left(\sum_{i \in \Omega} \left(\frac{p_i + q_i}{2} \right) - \sqrt{p_i q_i} \right).$$

We will prove the inequality pointwise, that is, for each $i \in \Omega$. Since $\beta_i \leq 1$ and since the i-th term in the right hand side is always non-negative, it is enough to show

$$\sqrt{p_i q_i} - (1+\gamma)p_i \leq \gamma\left(\left(\frac{p_i+q_i}{2}\right) - \sqrt{p_i q_i}\right).$$

This is equivalent to showing $p_i(1+3\gamma/2)+q_i(\gamma/2)-(1+\gamma)\sqrt{p_i q_i} \geq 0$, which is true since the LHS is the square of the quantity $(\sqrt{p_i(1+3\gamma/2)} - \sqrt{q_i(\gamma/2)})$ (recall that $\gamma = 1/\sqrt{2}$). ☐

8 L_∞ distance

In the L_∞ promise problem, Alice and Bob are given, respectively, two n-dimensional vectors, $\mathbf{x}$ and $\mathbf{y}$ from $[0,m]^n$ with the following promise: either $|\mathbf{x}_i - \mathbf{y}_i| \leq 1$ for all i, or for some i, $|\mathbf{x}_i - \mathbf{y}_i| \geq m$. The function $L_\infty(\mathbf{x},\mathbf{y}) = 1$ if and only if the latter case holds.

Theorem 8.1. *For* $0 < \delta < 1/4$,

$$R_\delta(L_\infty) \;\geq\; \frac{n}{4m^2}\cdot(1-2\sqrt{\delta}).$$

Proof. Note that L_∞ is OR-decomposable, since $L_\infty(\mathbf{x},\mathbf{y}) = \bigvee_j \textsc{dist}(\mathbf{x}_j,\mathbf{y}_j)$, where $\textsc{dist}(x,y) = 1$, if $|x-y| \geq m$ and $\textsc{dist}(x,y) = 0$ if $|x-y| \leq 1$.

We will once again use the direct sum paradigm. Define the random variable $((X,Y),D)$ with values in $[0,m]^2 \times ([0,m]\times\{0,1\})$, with distribution ζ, as follows. The random variable D is uniformly distributed in $([0,m]\times\{0,1\}) \setminus \{(0,1),(m,0)\}$. If $D=(d,0)$, then $X=d$ and Y is uniformly distributed in $\{d,d+1\}$; if $D=(d,1)$, then $Y=d$ and X is uniformly distributed in $\{d-1,d\}$. It is easy to see that X and Y are independent, conditioned on D. Let ν denote the distribution of (X,Y). Since $\textsc{dist}(x,y)=0$ for all values (x,y) of (X,Y), it follows ν^n is a collapsing distribution for L_∞. The theorem follows by applying Lemma 8.2 given below. ☐

Lemma 8.2. *For any* $0 < \delta < 1/4$,

$$\mathrm{CIC}_{\zeta,\delta}(\textsc{dist}) \;\geq\; \frac{1}{4m^2}\cdot(1-2\sqrt{\delta}).$$

Proof. Let Π be any δ-error protocol for $\textsc{dist}$ whose conditional information cost with respect to ζ is $\mathrm{CIC}_{\zeta,\delta}(\textsc{dist})$, and let U_d denote a random variable with uniform distribution in $\{d,d+1\}$. By expanding on values of D, it can be shown that

$$\mathrm{CIC}_{\zeta,\delta}(\textsc{dist}) = \frac{1}{2m}\left(\sum_{d=0}^{m-1} \mathrm{I}(U_d\,;\,\Pi(d,U_d)) + \sum_{d=1}^{m} \mathrm{I}(U_{d-1}\,;\,\Pi(U_{d-1},d))\right).$$

Therefore,

$$
\begin{aligned}
&\mathrm{CIC}_{\zeta,\delta}(\textsc{dist}) \\
&\geq \frac{1}{2m}\left(\sum_{d=0}^{m-1} \mathrm{h}^2(\Pi_{dd},\Pi_{d,d+1}) + \sum_{d=1}^{m} \mathrm{h}^2(\Pi_{d-1,d},\Pi_{dd})\right) && \text{(by Lemma 6.2)} \\
&\geq \frac{1}{4m^2}\left(\sum_{d=0}^{m-1} \mathrm{h}(\Pi_{dd},\Pi_{d,d+1}) + \sum_{d=1}^{m} \mathrm{h}(\Pi_{d-1,d},\Pi_{dd})\right)^2 && \text{(Cauchy–Schwarz)} \\
&\geq \frac{1}{4m^2}\ \mathrm{h}^2(\Pi_{00},\Pi_{mm}). && \text{(Triangle inequality)}
\end{aligned}
$$

We cannot directly bound $\mathrm{h}^2(\Pi_{00},\Pi_{mm})$ from below, because DIST is 0 on both inputs. However, by Lemma 6.4, we have that $\mathrm{h}^2(\Pi_{00},\Pi_{mm}) \geq \frac{1}{2}(\mathrm{h}^2(\Pi_{00},\Pi_{m0}) + \mathrm{h}^2(\Pi_{0m},\Pi_{mm}))$, which, by Lemma 6.5, is at least $1 - 2\sqrt{\delta}$. $\square$

Acknowledgement

We thank the anonymous referees for their valuable comments.

References

[Abl96] F. Ablayev. Lower bounds for one-way probabilistic communication complexity and their application to space complexity. *Theoretical Computer Science*, 157(2):139–159, 1996.

[AJKS02] M. Ajtai, T. S. Jayram, R. Kumar, and D. Sivakumar. Approximate counting of inversions in a data stream. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pages 370–379, 2002.

[AMS99] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.

[BCKO93] R. Bar-Yehuda, B. Chor, E. Kushilevitz, and A. Orlitsky. Privacy, additional information, and communication. *IEEE Transactions on Information Theory*, 39(6):1930–1943, 1993.

[BFS86] L. Babai, P. Frankl, and J. Simon. Complexity classes in communication complexity theory (preliminary version). In *Proceedings of the 27th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 337–347, 1986.

[Bry86] R. E. Bryant. Graph-based algorithms for Boolean function manipulations. *IEEE Transactions on Computers*, 35:677–691, 1986.

[CKS03] A. Chakrabarti, S. Khot, and X. Sun. Near-optimal lower bounds on the multi-party communication complexity of set-disjointness. In *Proceedings of the 18th Annual IEEE Conference on Computational Complexity (CCC)*, 2003.

[CSWY01] A. Chakrabarti, Y. Shi, A. Wirth, and A. C-C. Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In *Proceedings of the 42nd IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 270–278, 2001.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, Inc., 1991.

[FKSV02] J. Feigenbaum, S. Kannan, M. Strauss, and M. Viswanathan. An approximate L^1-difference algorithm for massive data streams. *SIAM Journal on Computing*, 32:131–151, 2002.

[GGI+02] A. Gilbert, S. Guha, P. Indyk, Y. Kotidis, S. Muthukrishnan, and M. Strauss. Fast, small-space algorithms for approximate histogram maintenance. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pages 389–398, 2002.

[GMMO00] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams. In *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 359–366, 2000.

[Ind00] P. Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computations. In *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 189–197, 2000.

[JKS03] T.S. Jayram, R. Kumar, and D. Sivakumar. Two applications of information complexity. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing (STOC)*, pages 673–682, 2003.

[JRS03] R. Jain, J. Radhakrishnan, and P. Sen. A lower bound for bounded round quantum communication complexity of set disjointness. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2003. To appear.

[KN97] E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[KNR99] I. Kremer, N. Nisan, and D. Ron. On randomized one-round communication complexity. *Computational Complexity*, 8(1):21–49, 1999.

[KRW95] Mauricio Karchmer, Ran Raz, and Avi Wigderson. Super-logarithmic depth lower bounds via the direct sum in communication complexity. *Computational Complexity*, 5(3/4):191–204, 1995.

[KS92] B. Kalyanasundaram and G. Schnitger. The probabilistic communication complexity of set intersection. *SIAM Journal on Discrete Math*, 5(5):545–557, 1992.

[Lin91] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[LY90] L. Le Cam and G. L. Yang. *Asymptotics in Statistics—Some Basic Concepts*, pages 24–30. Springer–Verlag, 1990.

[NS01] N. Nisan and I. Segal. The communication complexity of efficient allocation problems. In *DIMACS workshop on Computational Issues in Game Theory and Mechanism Design*, 2001.

[PS84] C. H. Papadimitriou and M. Sipser. Communication complexity. *Journal of Computer and System Sciences*, 28(2):260–269, 1984.

[Raz92] A. A. Razborov. On the distributional complexity of disjointness. *Theoretical Computer Science*, 106(2):385–390, 1992.

[Raz98] R. Raz. A parallel repetition theorem. *SIAM Journal on Computing*, 27(3):763–803, 1998.

[Rén60] A. Rényi. On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, pages 547–561, 1960.

[SS02] M. Saks and X. Sun. Space lower bounds for distance approximation in the data stream model. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pages 360–369, 2002.

[Weg87] I. Wegener. *The Complexity of Boolean Functions*. Wiley–Teubner Series in Computer Science. John Wiley & Sons, 1987.

[Yao79] A. C-C. Yao. Some complexity questions related to distributive computing. In *Proceedings of the 11th Annual ACM Symposium on Theory of Computing (STOC)*, pages 209–213, 1979.

A Measures of information and statistical differences

Definition A.1 (Statistical distance measures). Let P and Q be two distributions on the same probability space Ω. The *total variation distance* V, the *Hellinger distance* h, the *Kullback–Leibler divergence* KL, the *Jensen–Shannon divergence* $\overline{\mathrm{D}}$, and the *Rényi divergence* D_α $(0 < \alpha < 1)$ between P and Q are defined as follows:

$$
\begin{aligned}
\mathrm{V}(P,Q) &= \tfrac{1}{2}\textstyle\sum_{\omega\in\Omega}|P(\omega)-Q(\omega)| = \max_{\Omega'\subseteq\Omega}|P(\Omega')-Q(\Omega')|,\\
\mathrm{h}(P,Q) &= (1-\textstyle\sum_{\omega\in\Omega}\sqrt{P(\omega)Q(\omega)})^{\frac{1}{2}} = (\tfrac{1}{2}\textstyle\sum_{\omega\in\Omega}(\sqrt{P(\omega)}-\sqrt{Q(\omega)})^2)^{\frac{1}{2}},\\
\mathrm{KL}(P\parallel Q) &= \textstyle\sum_{\omega\in\Omega}P(\omega)\log\frac{P(\omega)}{Q(\omega)},\\
\overline{\mathrm{D}}(P,Q) &= \tfrac{1}{2}\left(\mathrm{KL}(P\parallel \tfrac{P+Q}{2})+\mathrm{KL}(Q\parallel\tfrac{P+Q}{2})\right),\\
\mathrm{D}_\alpha(P,Q) &= 1-\textstyle\sum_{\omega\in\Omega}P(\omega)^\alpha Q(\omega)^{1-\alpha}.
\end{aligned}
$$

While $\mathrm{V}(\cdot,\cdot)$ and $\mathrm{h}(\cdot,\cdot)$ are metrics, $\mathrm{KL}(\cdot\parallel\cdot)$, $\overline{\mathrm{D}}(\cdot,\cdot)$, and $\mathrm{D}_\alpha(\cdot,\cdot)$ are not. However, they are always non-negative and equal 0 if and only if $P=Q$. The Rényi divergence is a generalization of the Hellinger distance: $\mathrm{D}_{\frac{1}{2}}(P,Q)=\mathrm{h}^2(P,Q)$.

Proposition A.2 (Proposition 6.10 restated; [LY90]). *If P and Q are distributions on the same domain, then* $\mathrm{V}(P,Q) \le \mathrm{h}(P,Q)\sqrt{2-\mathrm{h}^2(P,Q)}$.

Proposition A.3.

$$\forall\alpha<\beta,\quad \frac{\alpha}{\beta}\,\mathrm{D}_\beta(P,Q) \;\le\; \mathrm{D}_\alpha(P,Q) \;\le\; \frac{1-\alpha}{1-\beta}\,\mathrm{D}_\beta(P,Q).$$

Proof. We use Hölder's inequality (for vectors $\mathbf{u},\mathbf{v}$, and for p,q that satisfy $1/p+1/q=1$, $|\langle\mathbf{u},\mathbf{v}\rangle|\le\|\mathbf{u}\|_p\cdot\|\mathbf{v}\|_q$) with $p=\beta/\alpha$ and $q=\beta/(\beta-\alpha)$:

$$
\begin{aligned}
&1-\mathrm{D}_\alpha(P,Q)\\
&\quad= \sum_\omega P(\omega)^\alpha Q(\omega)^{1-\alpha} = \sum_\omega P(\omega)^\alpha Q(\omega)^{\alpha/\beta-\alpha}\cdot Q(\omega)^{1-\alpha/\beta}\\
&\quad\le \left(\sum_\omega\left(P(\omega)^\alpha Q(\omega)^{\alpha/\beta-\alpha}\right)^{\beta/\alpha}\right)^{\alpha/\beta}\cdot\left(\sum_\omega\left(Q(\omega)^{1-\alpha/\beta}\right)^{\beta/(\beta-\alpha)}\right)^{(\beta-\alpha)/\beta}\\
&\quad= \left(\sum_\omega P(\omega)^\beta Q(\omega)^{1-\beta}\right)^{\alpha/\beta}\cdot\left(\sum_\omega Q(\omega)\right)^{(\beta-\alpha)/\beta}\\
&\quad= (1-\mathrm{D}_\beta(P,Q))^{\alpha/\beta}.
\end{aligned}
$$

We now use the following simple analytic claim:

Claim A.4. *For any* $0 \le \epsilon, \delta \le 1$ *(excluding the case* $\epsilon = 1$ *and* $\delta = 0$*),* $(1-\epsilon)^\delta \le 1 - \delta\epsilon$*.*

Proof. The cases $\delta = 0, 1$ are trivial. So assume $\delta \in (0,1)$ and consider the function $f(\epsilon) = 1 - \delta\epsilon - (1-\epsilon)^\delta$. We need to show f is non-negative in the interval $[0,1]$. Taking the derivative of f, we have: $f'(\epsilon) = \delta(1/(1-\epsilon)^{1-\delta} - 1)$; since $1-\epsilon \le 1$ and $1-\delta > 0$, $f'(\epsilon) \ge 0$. Therefore, f is non-decreasing in the interval $[0,1]$, implying its minimum is obtained at $\epsilon = 0$. Since $f(0) = 0$, we have that $f(\epsilon) \ge 0$ for all $\epsilon \in [0,1]$. □

Since both $\mathrm{D}_\beta(P,Q)$ and α/β are in the interval $[0,1]$ (and $\alpha/\beta > 0$), we obtain the left inequality:

$$1 - \mathrm{D}_\alpha(P,Q) \;\le\; (1 - \mathrm{D}_\beta(P,Q))^{\alpha/\beta} \;\le\; 1 - \frac{\alpha}{\beta} \cdot \mathrm{D}_\beta(P,Q).$$

For the other direction, note that $\mathrm{D}_\beta(P,Q) = \mathrm{D}_{1-\beta}(Q,P)$, by definition. Therefore, using the first direction,

$$\mathrm{D}_\beta(P,Q) \;=\; \mathrm{D}_{1-\beta}(Q,P) \;\ge\; \frac{1-\beta}{1-\alpha}\,\mathrm{D}_{1-\alpha}(Q,P) \;=\; \frac{1-\beta}{1-\alpha}\,\mathrm{D}_\alpha(P,Q).$$ □

Proposition A.5 ([Lin91]). *For distributions* P *and* Q *on the same domain,* $\overline{\mathrm{D}}(P,Q) \ge \mathrm{h}^2(P,Q)$*.*

The next proposition is used crucially in all our proofs to rephrase mutual information quantities in terms of the Jensen–Shannon divergence, which then allows us, via Proposition A.5, the use of the Hellinger distance or the Rényi divergences.

Proposition A.6. *Let* $\Phi(z_1)$ *and* $\Phi(z_2)$ *be two random variables. Let* Z *denote a random variable with uniform distribution in* $\{z_1, z_2\}$*. Suppose* $\Phi(z)$ *is independent of* Z *for each* $z \in \{z_1, z_2\}$*. Then,* $\mathrm{I}(Z;\Phi(Z)) = \overline{\mathrm{D}}(\Phi_{z_1}, \Phi_{z_2})$*.*

Proof. The mutual information between two random variables X and Y can be written as follows (cf. [CT91]):

$$I(X\,;\,Y) = \sum_{x\in\mathcal{X}} \Pr[X=x] \sum_{y\in\mathcal{Y}} \Pr[Y=y \mid X=x] \cdot \log \frac{\Pr[Y=y \mid X=x]}{\Pr[Y=y]},$$

where $\mathcal{X}$ and $\mathcal{Y}$ denote the supports of the distributions of X and Y, respectively.

Let μ denote the distribution of Y, and for any $x \in \mathcal{X}$, let μ_x denote the distribution of Y conditioned on the event $\{X = x\}$. Then the above equation can be rewritten using KL-divergence:

$$I(X\,;\,Y) = \sum_{x\in\mathcal{X}} \Pr[X=x] \cdot \mathrm{KL}(\mu_x \parallel \mu) \tag{5}$$

For the proof, we set $X = Z$ and $Y = \Phi(Z)$. For each $z \in \{z_1, z_2\}$, $\Phi(z)$ is independent of Z; therefore, conditioned on the event $\{Z = z\}$, the distribution of $\Phi(Z)$ equals Φ_z. Moreover, because Z is uniformly distributed in $\{z_1, z_2\}$, we have $\Phi(Z) \sim (\Phi_{z_1} + \Phi_{z_2})/2$. By Equation 5,

$$I(Z\,;\,\Phi(Z)) = \sum_{z=z_1,z_2} \Pr[Z = z] \cdot \mathrm{KL}(\Phi_z \parallel \frac{\Phi_{z_1} + \Phi_{z_2}}{2}) = \overline{\mathrm{D}}(\Phi_{z_1}, \Phi_{z_2}). \qquad \square$$

Finally, we state the lemma that we use in the proofs of information complexity lower bounds of primitive functions; the lemma follows directly from Propositions A.6 and A.5.

Lemma A.7 (Lemma 6.2 restated). *Let $\Phi(z_1)$ and $\Phi(z_2)$ be two random variables. Let Z denote a random variable with uniform distribution in $\{z_1, z_2\}$. Suppose $\Phi(z)$ is independent of Z for each $z \in \{z_1, z_2\}$. Then, $\mathrm{I}(Z; \Phi(Z)) \geq \mathrm{h}^2(\Phi_{z_1}, \Phi_{z_2})$.*

B Proof of Lemma 7.3

Lemma B.1 (Lemma 7.3 restated). *Let Π be a randomized t-party one-way communication protocol with inputs from $\{0,1\}^t$. Then, for any $0 < \epsilon < 1$,*

$$\sum_{i=1}^{t} \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{e}_i}) \geq \frac{(\ln^2 2)\epsilon^2}{8\ t^{\epsilon}} \cdot \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{1}}).$$

Proof. In the proof we employ Rényi divergences D_α [Rén60] (see Appendix A for the definition) and as we remarked earlier, this proof will be a generalization of the proof of Lemma 7.4. By Proposition A.3, we have for $1/2 \leq \alpha < 1$ and distributions P and Q on the same domain,

$$\frac{1}{2\alpha}\,\mathrm{D}_\alpha(P,Q) \leq \mathrm{h}^2(P,Q) \leq \frac{1}{2(1-\alpha)}\,\mathrm{D}_\alpha(P,Q). \tag{6}$$

We fix $\alpha = \alpha(\epsilon)$ to be chosen later. Using Equation 6, we have:

$$\sum_{i=1}^{t} \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{e}_i}) \;\geq\; \frac{1}{2\alpha} \cdot \sum_{i=1}^{t} \mathrm{D}_\alpha(\Pi_{\mathbf{0}}, \Pi_{\mathbf{e}_i}), \tag{7}$$

$$\mathrm{D}_\alpha(\Pi_{\mathbf{0}}, \Pi_{\mathbf{1}}) \;\geq\; 2(1-\alpha) \cdot \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{1}}). \tag{8}$$

It would thus suffice to prove the following counterpart of Lemma 7.2 for the Rényi divergence.

Lemma B.2. *For any one-way protocol Π, for any $0 < \epsilon < 1$, if $\alpha = 1 - \gamma^2/(4(1+\gamma))$, where $\gamma = 2^\epsilon - 1$, then $\sum_{i=1}^{t} \mathrm{D}_\alpha(\Pi_{\mathbf{0}}, \Pi_{\mathbf{e}_i}) \geq (1/t^\epsilon)\,\mathrm{D}_\alpha(\Pi_{\mathbf{0}}, \Pi_{\mathbf{1}})$.*

Assuming Lemma B.2, we will complete the proof of Lemma 7.3. By Equations 7 and 8 and using Lemma B.2,

$$\sum_{i=1}^{t} \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{e}_i}) \;\geq\; \frac{1-\alpha}{\alpha} \cdot \frac{1}{t^{\epsilon}} \cdot \mathrm{h}^2(\Pi_{\mathbf{0}}, \Pi_{\mathbf{1}}).$$

By our choice of α,

$$\frac{1-\alpha}{\alpha} \geq 1-\alpha = \frac{\gamma^2}{4(1+\gamma)} \geq \frac{\gamma^2}{8}.$$

Since $\gamma = 2^{\epsilon} - 1 \geq \epsilon \ln 2$, we have $\gamma^2/8 \geq (\epsilon^2 \ln^2 2)/8$, and Lemma 7.3 follows. $\square$

Proof of Lemma B.2. The proof goes along the same lines of the proof of Lemma 7.2, with Hellinger distance replaced by the Rényi divergence. The inductive step is the following. Let u be any internal node in T and let v and w be its left child and right child, respectively. Then, $\mathrm{D}_\alpha(\Pi_{\mathbf{0}}, \Pi_u) \leq (1+\gamma) \cdot (\mathrm{D}_\alpha(\Pi_{\mathbf{0}}, \Pi_v) + \mathrm{D}_\alpha(\Pi_{\mathbf{0}}, \Pi_w))$.

Similar to the proof of Lemma 7.2, suppose $u = \mathbf{e}_{[a,b]}$, $v = \mathbf{e}_{[a,c]}$, and $w = \mathbf{e}_{[c+1,b]}$, where $c = \lfloor \frac{a+b}{2} \rfloor$. Define the sets of players A, B and the input assignments $\mathbf{y}, \mathbf{y}', \mathbf{z}, \mathbf{z}'$ as before. Recall that $\mathbf{0} = \mathbf{yz}$, $u = \mathbf{y}'\mathbf{z}'$, $v = \mathbf{y}'\mathbf{z}$, and $w = \mathbf{y}\mathbf{z}'$.

For a probability vector p on Ω and a probability transition matrix M on $\Omega \times \Gamma$, let $p \circ M$ denote the distribution on $\Omega \times \Gamma$ where $(p \circ M)(i,j) = p(i) \cdot M(i,j)$. Applying Lemma 6.8, we have $\Pi_{\mathbf{0}} = \Pi_{\mathbf{yz}} = p_{\mathbf{y}} \circ M_{\mathbf{z}}$, $\Pi_u = \Pi_{\mathbf{y}'\mathbf{z}'} = p_{\mathbf{y}'} \circ M_{\mathbf{z}'}$, $\Pi_v = \Pi_{\mathbf{y}'\mathbf{z}} = p_{\mathbf{y}'} \circ M_{\mathbf{z}}$, and $\Pi_w = \Pi_{\mathbf{y}\mathbf{z}'} = p_{\mathbf{y}} \circ M_{\mathbf{z}'}$. The lemma now follows from the following property of the Rényi divergence, whose proof uses convexity and analytical arguments. $\square$

Lemma B.3. *Let p, q be probability distributions on Ω, and let M, N be probability transition matrices on $\Omega \times \Gamma$, for some Ω and Γ. For any $\gamma > 0$, if $\alpha \geq 1 - \gamma^2/(4(1+\gamma))$, then*

$$\mathrm{D}_\alpha(p \circ M, q \circ N) \leq (1+\gamma) \cdot (\mathrm{D}_\alpha(p \circ M, q \circ M) + \mathrm{D}_\alpha(p \circ M, p \circ N)).$$

Proof of Lemma B.3. We define β_i to be the Rényi α-divergence between the i-th row of M and the i-th row of N. Similar to the proof of Lemma 7.4, we can rewrite the three Rényi divergences as: $\mathrm{D}_\alpha(p \circ M, q \circ N) = \mathrm{D}_\alpha(p,q) + \sum_{i \in \Omega} p_i^{\alpha} q_i^{1-\alpha} \beta_i$, $\mathrm{D}_\alpha(p \circ M, q \circ M) = \mathrm{D}_\alpha(p,q)$, and $\mathrm{D}_\alpha(p \circ M, p \circ N) = \sum_{i \in \Omega} p_i \beta_i$. Thus, what we need to prove is:

$$\begin{aligned}
& \mathrm{D}_\alpha(p,q) + \sum_{i \in \Omega} p_i^{\alpha} q_i^{1-\alpha} \beta_i \;\leq\; (1+\gamma) \cdot \left(\mathrm{D}_\alpha(p,q) + \sum_{i \in \Omega} p_i \beta_i \right) \\
\iff\;& \sum_{i \in \Omega} p_i^{\alpha} q_i^{1-\alpha} \beta_i \;\leq\; \gamma \cdot \mathrm{D}_\alpha(p,q) + (1+\gamma) \left(\sum_{i \in \Omega} p_i \beta_i \right) \\
\iff\;& \sum_{i \in \Omega} \beta_i \left(p_i^{\alpha} q_i^{1-\alpha} - (1+\gamma) p_i \right) \;\leq\; \gamma \cdot \mathrm{D}_\alpha(p,q).
\end{aligned}$$

Let us denote by Ω_1 the set of all $i \in \Omega$, for which $p_i^\alpha q_i^{1-\alpha} \geq (1+\gamma)p_i$. Let $\Omega_2 = \Omega \setminus \Omega_1$. Since $\beta_i \leq 1$, then

$$\sum_{i \in \Omega} \beta_i \left(p_i^\alpha q_i^{1-\alpha} - (1+\gamma)p_i\right) \leq \sum_{i \in \Omega_1} p_i^\alpha q_i^{1-\alpha} - (1+\gamma)p_i.$$

Thus, it suffices to prove:

$$\sum_{i \in \Omega_1} \left(p_i^\alpha q_i^{1-\alpha} - (1+\gamma)p_i\right) \;\leq\; \gamma \cdot \mathrm{D}_\alpha(p,q).$$

Substituting $\mathrm{D}_\alpha(p,q) = 1 - \sum_{i \in \Omega} p_i^\alpha q_i^{1-\alpha}$ in the RHS of the above inequality and rearranging the terms, we need to show that

$$\sum_{i \in \Omega_1} (1+\gamma)p_i^\alpha q_i^{1-\alpha} + \sum_{i \in \Omega_2} \gamma p_i^\alpha q_i^{1-\alpha} - \sum_{i \in \Omega_1} (1+\gamma)p_i \;\leq\; \gamma. \tag{9}$$

We note the following convexity property of the function $f(x,y) = x^\alpha y^{1-\alpha}$:

Claim B.4. *For any non-negative numbers $x_1, \ldots, x_n, y_1, \ldots, y_n$,*

$$\sum_{i=1}^n x_i^\alpha y_i^{1-\alpha} \;\leq\; \left(\sum_{i=1}^n x_i\right)^\alpha \cdot \left(\sum_{i=1}^n y_i\right)^{1-\alpha}.$$

The proof follows directly from an application of Hölder's inequality.

Define $z = \sum_{i \in \Omega_1} p_i$ and $w = \sum_{i \in \Omega_1} q_i$. Applying Claim B.4 in Equation 9, it suffices to prove the following:

$$(1+\gamma) \cdot z^\alpha w^{1-\alpha} + \gamma \cdot (1-z)^\alpha (1-w)^{1-\alpha} - (1+\gamma)z - \gamma \leq 0. \tag{10}$$

This inequality is shown to be true using analytic tools in Lemma B.5 below. This completes the proof. ◻

Lemma B.5. *Let $0 \leq z, w \leq 1$ be real numbers, and γ be any non-negative real number. Then,*

$$(1+\gamma) \cdot z^\alpha w^{1-\alpha} + \gamma \cdot (1-z)^\alpha (1-w)^{1-\alpha} - (1+\gamma)z - \gamma \leq 0,$$

provided $\alpha \geq 1 - \gamma^2/(4(1+\gamma))$.

Proof. Define $f_\alpha(z,w)$ to be the left-hand-side of (10). For any given value of z we will maximize $f_\alpha(z,w)$ as a function of w and show that this maximum is less than 0, if α satisfies the bound given in the statement of the lemma. For simplicity of notation, we denote: $a = (1+\gamma)z^\alpha$, $b = \gamma(1-z)^\alpha$ and $\delta = 1-\alpha$. We thus have: $f_{\alpha,z}(w) = aw^\delta + b(1-w)^\delta - (1+\gamma)z - \gamma$.

$$\frac{df_{\alpha,z}}{dw} \;=\; a\delta w^{\delta-1} - b\delta(1-w)^{\delta-1}.$$

Thus, the extremal point is at:

$$w^* \;=\; \frac{a^{1/(1-\delta)}}{a^{1/(1-\delta)} + b^{1/(1-\delta)}}.$$

This point is a maximum in the interval $[0,1]$, since

$$\frac{d^2 f_{\alpha,z}}{dw^2} \;=\; a\delta(\delta-1)w^{\delta-2} + b\delta(\delta-1)(1-w)^{\delta-2} < 0.$$

Thus the value at the maximum point is:

$$\begin{aligned}
f_{\alpha,z}(w^*) &= \frac{a^{1/(1-\delta)}}{\left(a^{1/(1-\delta)} + b^{1/(1-\delta)}\right)^{\delta}} + \frac{b^{1/(1-\delta)}}{\left(a^{1/(1-\delta)} + b^{1/(1-\delta)}\right)^{\delta}} - (1+\gamma)z - \gamma \\
&= \left(a^{1/(1-\delta)} + b^{1/(1-\delta)}\right)^{1-\delta} - (1+\gamma)z - \gamma \\
&= \left((1+\gamma)^{1/\alpha} z + \gamma^{1/\alpha}(1-z)\right)^{\alpha} - (1+\gamma)z - \gamma.
\end{aligned}$$

We want this maximum to be non-positive for every $z \in [0,1]$. That is,

$$\begin{aligned}
& \left((1+\gamma)^{1/\alpha} z + \gamma^{1/\alpha}(1-z)\right)^{\alpha} \;\le\; (1+\gamma)z + \gamma \\
\iff\; & ((1+\gamma)z + \gamma)^{1/\alpha} - (1+\gamma)^{1/\alpha} z - \gamma^{1/\alpha}(1-z) \;\ge\; 0.
\end{aligned} \tag{11}$$

Let $g_\alpha(z)$ be the left-hand-side of (11), and for simplicity of notation, let $\ell = 1/\alpha$. We would like to show that for an appropriate choice of α, $g_\alpha(z) \ge 0$ for all $z \in [0,1]$. Note that $g_\alpha(0) = 0$. Thus, it suffices to show that g is non-decreasing in the interval $[0,1]$.

$$g'(z) \;=\; \ell(1+\gamma)\left((1+\gamma)z + \gamma\right)^{\ell-1} - (1+\gamma)^{\ell} + \gamma^{\ell} \;\ge\; \ell(1+\gamma)\gamma^{\ell-1} - (1+\gamma)^{\ell} + \gamma^{\ell},$$

where the last inequality follows from the fact $z \ge 0$. Thus g would be non-decreasing if:

$$\ell(1+\gamma)\gamma^{\ell-1} - (1+\gamma)^{\ell} + \gamma^{\ell} \;\ge\; 0 \;\iff\; \ell\left(\frac{\gamma}{1+\gamma}\right)^{\ell-1} - 1 + \left(\frac{\gamma}{1+\gamma}\right)^{\ell} \;\ge\; 0.$$

Write $\eta = \gamma/(1+\gamma)$. Note that $0 < \eta < 1$. We thus need to prove:

$$\begin{aligned}
\eta^{\ell} + \ell\eta^{\ell-1} - 1 \;\ge\; 0 \;&\iff\; \eta^{\ell-1}(\eta + \ell) - 1 \;\ge\; 0 \\
&\Longleftarrow\; \eta^{\ell-1}(1+\eta) - 1 \;\ge\; 0 \;\iff\; \eta^{\ell-1} \;\ge\; \frac{1}{1+\eta}.
\end{aligned}$$

Since $\eta < 1$, $1/(1+\eta) \le e^{-\eta/2}$. Thus it suffices that:

$$\eta^{\ell-1} \;\ge\; e^{-\eta/2} \;\iff\; \ell - 1 \;\le\; \frac{\eta}{2\ln(1/\eta)}.$$

Therefore, we need $\alpha = 1/\ell$ to satisfy

$$\alpha \;\geq\; \frac{1}{1 + \frac{\eta}{2\ln(1/\eta)}}.$$

Thus, it suffices that

$$\alpha \;\geq\; 1 - \frac{\eta}{4\ln(1/\eta)} \;=\; 1 - \frac{\gamma}{4(1+\gamma)\ln((1+\gamma)/\gamma)}.$$

And for the last inequality to hold it suffices that

$$\alpha \geq 1 - \frac{\gamma^2}{4(1+\gamma)}.$$

$\square$