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

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 the 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 $\Omega(n^{1-2/k})$ space; this resolves a conjecture of Alon, Matias, and Szegedy [3].

(2) A lower bound for the L_p approximation problem in the general communication model; this solves an open problem of Saks and Sun [23]. 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 [25] 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 [15]. Recently, communication complexity has seen a resurgence of interest, which can be traced to exciting developments in the areas of massive data set algorithms (see below) and the design of combinatorial auctions [18].

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 [3, 8, 10, 11, 9, 2]. This model generalizes the restrictive OBDD model [5, 24]; 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., [3]) and by inventing new techniques (e.g., [23]).

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, 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 $\Omega(n/t^2)$[1]. Previously, Alon, Matias, and Szegedy[3] had proved an $\Omega(n/t^4)$ bound, extending the $\Omega(n)$ bound for two-party set-disjointness [12, 21]. 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[2]. This resolves the conjecture of Alon, Matias, and Szegedy [3], 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 $\Omega(n^{1-2/k})$ space. For $k > 2$, the best known space upper bound for F_k is $\tilde{O}(n^{1-1/k})$ [3]. 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.

(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 [23], who showed this bound for the restricted one-way communication model (where Alice sends a single message to Bob).

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 [23] gives a similar bound for the one-pass data stream model.

[1]Following our work, Subhash Khot [private communication, August 2002] has improved the bound for the multi-party set-disjointness problem in the general communication model to $\Omega(n/t^{3/2})$.

[2]For a 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 $\sum_{j \in [n]} f_j^k(\mathbf{a})$.

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 [13] asks 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 [6]; it is also implicit in the works of Ablayev [1] and Saks and Sun [23]. 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 all natural examples, these distributions necessarily have a non-product structure. This is one of the main obstacles to extending the direct sum methodology of [6] for general communication protocols; their work addresses the more restrictive case of simultaneous message protocols. In the proof technique of [23], 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 [21] and the parallel repetition theorem [20]. 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* [16], 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 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.

Organization. Section 2 contains the preliminaries. In Section 3, we develop the notion of information complexity of functions. In Section 4, 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 5, 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 in Section 6 an $\Omega(1)$ lower bound on the information complexity of the AND of two bits, deriving a simple proof for the $\Omega(n)$ bound for two-party set disjointness. 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 lower bounds for data stream computations are derived as consequences of the results in Sections 7 and 8.

2. Preliminaries

Communication complexity. In the two-party randomized communication complexity model [25], 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, wlog., 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* [19, 1, 14], 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; sometimes, we also write $X \sim Y$, when X and Y are random variables, to denote that X has the same distribution as Y. Let $\boldsymbol{\mu}$ be a distribution on a Cartesian product of sets and let the vector random variable $\mathbf{X} \sim \boldsymbol{\mu}$. We say that $\boldsymbol{\mu}$ is a *product* distribution if the components of $\mathbf{X}$ are mutually independent of each other. For example, the distribution $\boldsymbol{\mu} \triangleq \nu^n$ 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 $\omega \in \Omega$. Let $[n] \triangleq \{1, \ldots, n\}$, and $[0, m] \triangleq \{0, \ldots, m\}$. All logarithms are to the base 2.

Information Theory. The *entropy* of a discrete distribution X on a set Ω is $\mathrm{H}(X) = \sum_{\omega \in \Omega} X(\omega) \log \frac{1}{X(\omega)}$. 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 *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 *joint entropy* of two distributions X and Y is denoted $\mathrm{H}(X, Y)$. We use several basic properties of entropy and mutual information in the paper, which may be found in Chapter 2 of [7]; see also the full version of this paper.

3. 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 4 and Section 5, resulting in a substantially simpler proof of an $\Omega(n)$ 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 [6, 1, 23]. 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 $\mathcal{L}_n \subseteq \mathcal{X}^n \times \mathcal{Y}^n$ of legal inputs and a function $f : \mathcal{L}_n \to \{0, 1\}$.

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

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 1 (Information cost of a protocol). Let Π be a randomized protocol whose inputs belong to $\mathcal{L}_n$. Let $\boldsymbol{\mu}$ be a distribution on $\mathcal{L}_n$, and suppose $(\mathbf{X}, \mathbf{Y}) \sim \boldsymbol{\mu}$. The *information cost of* Π *under* $\boldsymbol{\mu}$ is defined as $\mathrm{I}(\mathbf{X}, \mathbf{Y} \,;\, \Pi(\mathbf{X}, \mathbf{Y}))$.

Definition 2 (Information complexity of a function). The δ-error *information complexity* of f under a distribution $\boldsymbol{\mu}$, denoted $\mathrm{IC}_{\boldsymbol{\mu},\delta}(f)$, is defined as the minimum information cost of a δ-error protocol for f under $\boldsymbol{\mu}$.

Proposition 3. *For any distribution $\boldsymbol{\mu}$ and error $\delta > 0$, $R_\delta(f) \geq \mathrm{IC}_{\boldsymbol{\mu},\delta}(f)$.*

Proof. Let Π denote the best δ-error protocol for f in terms of communication. Let $(\mathbf{X}, \mathbf{Y}) \sim \boldsymbol{\mu}$. Thus, $R_\delta(f) = |\Pi| \geq \mathrm{H}(\Pi(\mathbf{X}, \mathbf{Y})) \geq \mathrm{I}(\mathbf{X}, \mathbf{Y} \,;\, \Pi(\mathbf{X}, \mathbf{Y})) \geq \mathrm{IC}_{\boldsymbol{\mu},\delta}(f)$. $\square$

Suppose $\mathcal{L}_n = \mathcal{L}^n$, for some $\mathcal{L} \subseteq \mathcal{X} \times \mathcal{Y}$, 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.) If f depends (say, symmetrically) on the primitive in each coordinate, and if the distribution $\boldsymbol{\mu}$ on $\mathcal{L}_n$ is the product of independent copies of a distribution ν on $\mathcal{L}$, then any protocol for f must implicitly solve each instance of the primitive h. Thus, one can hope to show that $\mathrm{IC}_{\boldsymbol{\mu},\delta}(f) \geq n \cdot \mathrm{IC}_{\nu,\delta}(h)$, i.e. a direct sum property for information complexity.

The main technical obstacle to proving this result is that the distribution $\boldsymbol{\mu}$ is not necessarily a product distribution on $\mathcal{X}^n \times \mathcal{Y}^n$. This is because ν need not be a product distribution on $\mathcal{X} \times \mathcal{Y}$ (although $\boldsymbol{\mu}$ 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 [4]. To handle this, we will use the fact that $\boldsymbol{\mu}$ may be written as a convex combination $\boldsymbol{\mu} = \sum_{d \in K} \kappa_d \boldsymbol{\mu}_d$ of product distributions $\boldsymbol{\mu}_d$, where K is some index set. Such a decomposition, in general, is not unique, and we will choose one where the entropy of the collection of κ_d's, viewed as a distribution κ on the index set K, is as small as possible.

One way to realize $\boldsymbol{\mu}$ is as follows. If $D \sim \kappa$, then D "sets" $\boldsymbol{\mu} = \boldsymbol{\mu}_d$ with probability κ_d. Therefore, conditioned on D, $\boldsymbol{\mu}$ is a product distribution, that is, if $(\mathbf{X}, \mathbf{Y}) \sim \boldsymbol{\mu}$, then conditioned on D, $\mathbf{X}$ and $\mathbf{Y}$ are independent of each other. We will call $\boldsymbol{\mu}$ a *mixture* of distributions $\{\boldsymbol{\mu}_d\}_{d \in K}$ and say that κ *partitions* $\boldsymbol{\mu}$. In the above discussion, where ν is non-product and $\boldsymbol{\mu} = \nu^n$, we will first express ν as a mixture partitioned by some λ. Then, clearly $\boldsymbol{\kappa} = \lambda^n$ partitions $\boldsymbol{\mu}$ in a natural way. A useful consequence is that the coordinates $\{(X_j, Y_j)\}_{j \in [n]}$ are mutually independent of each other, and this continues to hold even when conditioned on $\mathbf{D} \sim \boldsymbol{\kappa}$.

For set-disjointness, we will use the non-product distribution ν given by $\nu(0, 0) = 1/2$, $\nu(0, 1) = \nu(1, 0) = 1/4$. Let λ, denoting the uniform distribution on $\{\mathrm{A}, \mathrm{B}\}$, partition ν as follows. Let $D \sim \lambda$. 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 $(X, Y) \sim \nu$.

Definition 4 (Conditional information cost). Let Π be a randomized protocol whose inputs belong to $\mathcal{L}_n$. Let $\boldsymbol{\mu}$ be a mixture of product distributions on $\mathcal{L}_n$, partitioned by κ. Suppose $(\mathbf{X}, \mathbf{Y}) \sim \boldsymbol{\mu}$ and $D \sim \kappa$. The *conditional information cost* of Π under $(\boldsymbol{\mu}, \kappa)$ is defined as $\mathrm{I}(\mathbf{X}, \mathbf{Y} \,;\, \Pi(\mathbf{X}, \mathbf{Y}) \mid D)$,

Definition 5 (Conditional information complexity). The δ-error *conditional information complexity* of f under $(\boldsymbol{\mu}, \kappa)$, denoted by $\mathrm{IC}_{\boldsymbol{\mu},\delta}(f \mid \kappa)$, is defined as the minimum conditional information cost of a δ-error protocol for f under $(\boldsymbol{\mu}, \kappa)$.

Proposition 6. *If κ partitions $\boldsymbol{\mu}$, then $\mathrm{IC}_{\boldsymbol{\mu},\delta}(f) \geq \mathrm{IC}_{\boldsymbol{\mu},\delta}(f \mid \kappa)$.*

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

Thus, by Proposition 3, lower bounds for conditional information complexity yield lower bounds for randomized communication complexity.

4. 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 Π for $f : \mathcal{L}^n \to \{0, 1\}$, for some $\mathcal{L} \subseteq \mathcal{X} \times \mathcal{Y}$, and fix a distribution ν on $\mathcal{L}$ partitioned by λ. First, we show that conditional information cost of Π under $(\boldsymbol{\mu}, \boldsymbol{\kappa})$, where $\boldsymbol{\mu} = \nu^n$ and $\boldsymbol{\kappa} = \lambda^n$, must implicitly contain information about each of the input 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 $\boldsymbol{\mu}$, we then show that each coordinate-wise information quantity is in itself an upper bound on the information complexity of the primitive function. This will result in the direct sum theorem.

Lemma 7 (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 ν be a distribution on $\mathcal{L}$ partitioned by λ. Let $(\mathbf{X}, \mathbf{Y}) \sim \boldsymbol{\mu} = \nu^n$, and $\mathbf{D} \sim \boldsymbol{\kappa} = \lambda^n$ (which partitions $\boldsymbol{\mu}$). Then,* $\mathrm{I}(\mathbf{X}, \mathbf{Y} \,;\, \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D}) \geq \sum_j \mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \,;\, \Pi \mid \mathbf{D})$.

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 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 $\mathrm{I}(\mathbf{X}, \mathbf{Y} \,;\, \Pi \mid \mathbf{D}) \geq \sum_j \mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \,;\, \Pi \mid \mathbf{D})$. $\square$

Definition 8 (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, $\mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \,;\, \Pi \mid \mathbf{D}) \geq \mathrm{IC}_{\nu,\delta}(h \mid \lambda)$, 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 information cost with respect to ν is *exactly* $\mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \,;\, \Pi \mid \mathbf{D})$. To facilitate this, we will further restrict the distribution μ that we use to be a "collapsing distribution" of f.

Definition 9 (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 10 (Collapsing distribution). Let a function $f : \mathcal{L}^n \to \{0, 1\}$ be 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 $\boldsymbol{\mu}$ *collapsing* for f, if every $(\mathbf{x}, \mathbf{y})$ in the support of $\boldsymbol{\mu}$ 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 ν^n, and for every $j \in [n]$, $\bigvee_{i \neq j} (\mathbf{x}_i \wedge \mathbf{y}_i) = 0$. Therefore, for every $(u, v) \in \{0, 1\}^2$, $\textsc{disj}(\textsc{embed}(\mathbf{x}, j, u), \textsc{embed}(\mathbf{y}, j, v)) = u \wedge v$.

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 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 11 (Reduction lemma). *Let Π be a δ-error protocol for a decomposable function $f : \mathcal{L}^n \to \{0,1\}$ with primitive h. Let $\boldsymbol{\mu} = \nu^n$ be a collapsing distribution for f, and suppose λ partitions ν (and $\boldsymbol{\kappa} = \nu^n$ partitions $\boldsymbol{\mu}$). Let $(\mathbf{X}, \mathbf{Y}) \sim \boldsymbol{\mu}$ and $\mathbf{D} \sim \boldsymbol{\mu}$. Then for all j, $\mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \,;\, \Pi \mid \mathbf{D}) \geq \mathrm{IC}_{\nu,\delta}(h \mid \lambda)$.*

Proof Sketch. Let $\mathbf{D}_{-j}$ denote all except the j-th coordinate of $\mathbf{D}$. Since $\mathbf{D} = (\mathbf{D}_j, \mathbf{D}_{-j})$, we have $\mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \,;\, \Pi \mid \mathbf{D}) = \mathrm{E}_{\mathbf{d}}[\mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \,;\, \Pi \mid \mathbf{D}_j, \mathbf{D}_{-j} = \mathbf{d})]$, where $\mathbf{d}$ is indexed by $[n] \setminus \{j\}$. We will show that each term is the information cost of a δ-error protocol $\Pi_{j,\mathbf{d}}$ for h, which will prove the lemma.

The protocol $\Pi_{j,\mathbf{d}}$ has j and $\mathbf{d}$ "hardwired" into it. On input pair (u, v), Alice and Bob realize random variables $(\mathbf{U}_i, \mathbf{V}_i)$, for every $i \neq j$, by sampling (independently of each other using private coin tosses) from the product distribution $\nu_{\mathbf{d}_i}$; let (u_i, v_i) be the value produced thus. Next, Alice sets $u_j = u$, Bob sets $v_j = v$, and they simulate $\Pi(\mathbf{u}, \mathbf{v})$. Since ν^n is a collapsing distribution of f, we have $f(\mathbf{u}, \mathbf{v}) = h(u, v)$. Letting $\Pi_{j,\mathbf{d}}$ output whatever Π outputs, it follows that $\Pi_{j,\mathbf{d}}$ is a δ-error protocol for h.

To complete the proof, we show that the conditional information cost of $\Pi_{j,\mathbf{d}}$ under (ν, λ),

$$\begin{aligned} &\mathrm{I}(U, V \,;\, \Pi_{j,\mathbf{d}}(U, V) \mid D) \\ &\quad = \mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \,;\, \Pi(\mathbf{X}, \mathbf{Y}) \mid \mathbf{D}_j, \mathbf{D}_{-j} = \mathbf{d}). \end{aligned} \tag{1}$$

Let $\Pi_{\text{trans}}(x, y, a, b)$ denote the fixed transcript produced on input (x, y) when the internal coin tosses of Alice and Bob are a and b, respectively. If (A, B) denotes the random variables corresponding to the private coin tosses, then $\Pi_{j,\mathbf{d}}(U, V) = \Pi_{\text{trans}}(\mathbf{U}, \mathbf{V}, A, B)$, where $\mathbf{U}_j = U$ and $\mathbf{V}_j = V$. Equation 1 follows if the joint distribution of the r.v.'s $(\mathbf{X}_j, \mathbf{Y}_j, \mathbf{D}_j, \Pi_{\text{trans}}(\mathbf{X}, \mathbf{Y}, A, B))$, conditioned on the event $\mathbf{D}_{-j} = \mathbf{d}$, is identical to the joint distribution of $(U, V, D, \Pi_{\text{trans}}(\mathbf{U}, \mathbf{V}, A, B))$. This can be verified easily and we omit the tedious probability statements. $\square$

Theorem 12 (Direct sum theorem). *Let $f : \mathcal{L}^n \to \{0,1\}$ be a decomposable function with primitive h. Let $\boldsymbol{\mu} = \nu^n$ be a collapsing distribution for f. Let λ partition ν so that $\boldsymbol{\kappa} = \lambda^n$ partitions ν^n. Then, $\mathrm{IC}_{\boldsymbol{\mu},\delta}(f \mid \boldsymbol{\kappa}) \geq n \cdot \mathrm{IC}_{\nu,\delta}(h \mid \lambda)$.*

Proof. Let Π be the optimal δ-error protocol for f in terms of conditional information cost. We have $\mathrm{IC}_{\boldsymbol{\mu},\delta}(f \mid \boldsymbol{\mu}) = \mathrm{I}(\mathbf{X}, \mathbf{Y} \,;\, \Pi \mid \mathbf{D})$, where $\mathbf{D} \sim \boldsymbol{\kappa}$. By the information cost decomposition lemma (Lemma 7), this is at least $\sum_j \mathrm{I}(\mathbf{X}_j, \mathbf{Y}_j \,;\, \Pi \mid \mathbf{D})$. By the reduction lemma (Lemma 11), this is at least $n \cdot \mathrm{IC}_{\nu,\delta}(h \mid \lambda)$. $\square$

Corollary 13 (of Prop. 3, Prop. 6, and Theorem 12). *With the notation and assumptions of Theorem 12, $R_\delta(f) \geq \mathrm{IC}_{\boldsymbol{\mu},\delta}(f) \geq \mathrm{IC}_{\boldsymbol{\mu},\delta}(f \mid \boldsymbol{\kappa}) \geq n \cdot \mathrm{IC}_{\nu,\delta}(h \mid \lambda)$.*

For set-disjointness, $R_\delta(\text{DISJ}) \geq \mathrm{IC}_{\boldsymbol{\mu},\delta}(\text{DISJ} \mid \boldsymbol{\kappa}) \geq n \cdot \mathrm{IC}_{\nu,\delta}(\text{AND} \mid \lambda)$. Thus it suffices to show an $\Omega(1)$ lower bound for the conditional information complexity of the 1-bit function AND!

5. Information statistics and communication complexity

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 shall develop some basic connections between communication complexity, statistical distance measures, and information theory. Some of the proofs in this section are omitted; see the full version of the paper.

Let Π be a randomized protocol for a function f. The proposition below points out that if two inputs (x, y) and (w, z) are assigned different values by f, then Π_{xy} and Π_{wz} must be "far" from each other. Recall that Π_{xy} denotes the distribution of the r.v. $\Pi(x, y)$.

Definition 14. The *total variation distance* between two distributions P and Q over a domain Ω is defined by $\mathrm{V}(P, Q) = \frac{1}{2} \sum_{\omega \in \Omega} |P(\omega) - Q(\omega)|$.

Proposition 15. *For any δ-error protocol Π for a function f, and for any two input pairs $(x, y), (w, z)$ for which $f(x, y) \neq f(w, z)$, $\mathrm{V}(\Pi_{xy}, \Pi_{wz}) \geq 1 - 2\delta$.*

The lemma below shows that the distributions on the transcripts of a randomized communication protocol have a rectangular structure. 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 16. *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 $\tau \in \mathcal{T}$,*

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

Next, we develop some useful consequences of the rectangular structure of the transcript distributions in randomized protocols. These properties may be considered to be the analogs of the following statement about deterministic communication protocols: if $\Pi(x, y) = \tau = \Pi(x', y')$, then $\Pi(x', y) = \tau = \Pi(x, y')$. The analogs are based on the notion of Hellinger distance, which, as the lemmas illustrate, turns out to be a natural choice in analyzing distributions of transcripts of communication protocols.

Definition 17 (Hellinger distance). The *Hellinger distance* between probability distributions P and Q on a space Ω is defined by $\mathrm{h}^2(P, Q) = 1 - \sum_{\omega \in \Omega} \sqrt{P(\omega)Q(\omega)}$. (Note that the Hellinger distance is defined by defining its square.)

Proposition 18. *For distributions P, Q on the same domain,* $\mathrm{V}(P, Q) \le \mathrm{h}(P, Q)\sqrt{2 - \mathrm{h}^2(P, Q)}$.

Lemma 19 (Cut & paste lemma). *For any randomized protocol Π and for any $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$,* $\mathrm{h}(\Pi_{xy}, \Pi_{x'y'}) = \mathrm{h}(\Pi_{xy'}, \Pi_{x'y})$.

Proof.

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

where (*) follows by applying Lemma 16. □

Lemma 20 (Pythagorean lemma). *For any randomized protocol Π and for any $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$,*
$\mathrm{h}^2(\Pi_{xy}, \Pi_{x'y}) + \mathrm{h}^2(\Pi_{xy'}, \Pi_{x'y'}) \le 2\,\mathrm{h}^2(\Pi_{xy}, \Pi_{x'y'})$.

Proof. Again using Lemma 16, we have

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

In (**) above, we used the AM–GM inequality. □

6. The information complexity of AND

The second phase of our lower bound paradigm focuses on proving a lower bound on the information that needs to be exchanged to compute the primitive function. In general, this depends on the specific primitive function and distribution; here we illustrate it for the AND function from set-disjointness; other applications appear in Sections 7 and 8.

We will show that for any randomized protocol P that correctly computes the AND function, an $\Omega(1)$ lower bound holds on $\mathrm{I}(U, V; P(U, V) \mid D)$ when the inputs (U, V) are distributed according to ν. We assume that for every input $(u, v) \in \{0, 1\}^2$, the protocol P computes $\mathrm{AND}(u, v)$ correctly with probability at least $1 - \delta$.

Let Z denote an r.v. distributed uniformly in $\{0, 1\}$. Using the definition of the distribution ν and expanding on values of D, we have

$$\begin{aligned}
&\mathrm{I}(U, V; P(U, V) \mid D) \\
&\quad = \frac{1}{2}[\mathrm{I}(U, V; P(U, V) \mid D = 0) \\
&\qquad + \mathrm{I}(U, V; P(U, V) \mid D = 1)] \\
&\quad = \frac{1}{2}[\mathrm{I}(Z; P(0, Z)) + \mathrm{I}(Z; P(Z, 0))] \qquad (2) \\
&\qquad \text{(since } (V \mid D = 0) \sim Z \text{ and } (U \mid D = 1) \sim Z\text{).}
\end{aligned}$$

Notice that the mutual information quantities in Equation (2) are of the form $\mathrm{I}(Z; \phi(Z))$, where $Z \in_R \{0, 1\}$ and $\phi(z)$ is a random variable, for each $z \in \{0, 1\}$. The following lemma provides an important passage from such quantities (and hence from information complexity) to the Hellinger distance, enabling the use of the machinery developed in Section 5. The proof proceeds via two notions from information theory—the Kullback–Leibler and Jensen–Shannon divergences (see [7, 17])—and is given in the full version of the paper.

Lemma 21. *Let $\phi(z_1)$ and $\phi(z_2)$ be two random variables. Let $Z \in_R \{z_1, z_2\}$. Then,* $\mathrm{I}(Z; \phi(Z)) \ge \mathrm{h}^2(\phi_{z_1}, \phi_{z_2})$.

Combining Equation 2 and Lemma 21, we obtain:

$$\begin{aligned}
&\mathrm{I}(U, V; P(U, V) \mid D) \\
&\quad \ge \frac{1}{2}(\mathrm{h}^2(P_{00}, P_{01}) + \mathrm{h}^2(P_{00}, P_{10})) \qquad \text{(Lemma 21)} \\
&\quad \ge \frac{1}{4}(\mathrm{h}(P_{00}, P_{01}) + \mathrm{h}(P_{00}, P_{10}))^2 \\
&\quad \ge \frac{1}{4}\,\mathrm{h}^2(P_{01}, P_{10}) \qquad \text{(triangle inequality)} \\
&\quad = \frac{1}{4}\,\mathrm{h}^2(P_{00}, P_{11}) \qquad \text{(Lemma 19)} \\
&\quad \ge \frac{1}{4}(1 - 2\sqrt{\delta}). \qquad \text{(Propositions 18 and 15)}
\end{aligned}$$

To sum up, we have shown:

Theorem 22. $R_\delta(\text{DISJ}) \geq \text{IC}_{\boldsymbol{\mu},\delta}(\text{DISJ}) \geq \text{IC}_{\boldsymbol{\mu},\delta}(\text{DISJ} \mid \boldsymbol{\kappa}) \geq n \cdot \text{IC}_{\nu,\delta}(\text{AND} \mid \lambda) \geq \frac{n}{4}(1 - 2\sqrt{\delta})$.

7. Multi-party set-disjointness

Let $\text{DISJ}_{n,t}(\mathbf{x_1}, \ldots, \mathbf{x_t}) = \bigvee_{j=1}^n \bigwedge_{i=1}^t x_{i,j}$, where the $\mathbf{x_i}$'s are n-bit vectors. Thus, $\text{DISJ}_{n,t}$ is OR-decomposable, and the induced "primitive" functions are all AND_t—the t-bit AND. The legal inputs for 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[3].

Theorem 23. *For any $\delta < 1/4$, and any $0 < \epsilon < 1$, set*
(1) $R_\delta(\text{DISJ}_{n,t}) \geq (1 - 2\sqrt{\delta}) \cdot \frac{n}{t^2}$.
(2) $R_\delta^{\textit{1-way}}(\text{DISJ}_{n,t}) \geq \frac{\epsilon^2 \cdot \ln^2 2}{8} \cdot (1 - 2\sqrt{\delta}) \cdot \frac{n}{t^{1+\epsilon}}$.

Proof. We will employ the direct sum paradigm. We define a distribution ν on inputs of AND_t as follows: let λ be the uniform distribution on $[t]$, and let $D \sim \lambda$. Conditioned on $D = i$, ν is uniform on $\{\mathbf{0}, \mathbf{e_i}\}$. It follows that $\boldsymbol{\mu} = \nu^n$ is a collapsing distribution for $\text{DISJ}_{n,t}$. Thus, all we need to prove is a lower bound on the conditional information complexity of AND_t under (ν, λ).

Let Π be any δ-error protocol that computes AND_t; to keep the notation simple we will suppress any reference to the underlying private randomness used in Π. We denote by $\mathbf{U}$ a random input for AND_t chosen according to ν. The conditional information cost is now given by

$$\mathrm{I}(\mathbf{U} \,;\, \Pi(\mathbf{U}) \mid D) = \frac{1}{t} \sum_i \mathrm{I}(\mathbf{U} \,;\, \Pi(\mathbf{U}) \mid D = i). \quad (3)$$

Notice that conditioned on $D = i$, $\mathbf{U}$ is uniformly distributed in $\{\mathbf{0}, \mathbf{e_i}\}$, so Lemma 21 allows us passage to the Hellinger distance. Our lower bounds for general communication are essentially obtained this way; for the one-way model, we are able to obtain stronger bounds by employing a generalization of the Hellinger distance. For $0 < \alpha < 1$, the *Rényi divergence of order* α [22] between two distributions P and Q is defined by $\mathrm{D}_\alpha(P, Q) = 1 - \sum_\omega P(\omega)^\alpha Q(\omega)^{1-\alpha}$. When $\alpha = 1/2$, $\mathrm{D}_{1/2}(P, Q) = \mathrm{h}^2(P, Q)$, the squared Hellinger distance.

For $\alpha \geq 1/2$, let $c_1 = c_1(\alpha) = 1/(2\alpha)$ and $c_2 = c_2(\alpha) = 2(1 - \alpha)$. Analogous to Lemma 21, we have:

$\mathrm{I}(Z; \phi(Z)) \geq c_1 \, \mathrm{D}_\alpha(\phi_{z_1}, \phi_{z_2})$,

which implies that $\mathrm{I}(\mathbf{U}; \Pi(\mathbf{U}) \mid D = i) \geq c_1 \, \mathrm{D}_\alpha(\Pi_\mathbf{0}, \Pi_\mathbf{e_i})$. Also, we have for distributions P and Q,

$\mathrm{D}_\alpha(P, Q) \geq c_2 \cdot \mathrm{h}^2(P, Q)$.

Suppose we can show for some constant $c = c_{\alpha,t}$ that depends only on α and t that

$$\sum_i \mathrm{D}_\alpha(\Pi_\mathbf{0}, \Pi_\mathbf{e_i}) \geq c \cdot \mathrm{D}_\alpha(\Pi_\mathbf{0}, \Pi_\mathbf{1}). \quad (4)$$

[3]The definition of $\text{DISJ}_{n,t}$ 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.

Combining Equation (4) with the inequalities for D_α, and substituting in Equation (3), it follows that

$$\begin{aligned} \mathrm{I}(\mathbf{U} \,;\, \Pi(\mathbf{U}) \mid D) &\geq (c_1 c_2 / t)\, c \;\mathrm{h}^2(\Pi_\mathbf{0}, \Pi_\mathbf{1}) \\ &\geq (c \cdot c_1 \cdot c_2 / t)\, (1 - 2\sqrt{\delta}). \quad (5) \end{aligned}$$

An appropriate choice of α will then complete the proof.

We will now prove Eqn. (4) via an inductive argument. The generalization of Lemma 16 to multi-parties shows that for every i and every input a assigned to player i, there a function $q_i(\tau; a)$ on transcripts with range $[0, 1]$, such that, for any input $\mathbf{x}$, $\Pr[\Pi(\mathbf{x}) = \tau] = \prod_i q_i(\tau; x_i)$.

It will be helpful to consider the contribution of groups of players. To formalize this, we introduce some notation. For two functions R and S on the same domain let RS denote the function defined by $RS(\omega) = R(\omega) \cdot S(\omega)$, for all ω. We call the pair (R, S) *rectangular*, if RS is a probability distribution. For any set $A \subseteq [t]$, and an input assignment $\mathbf{y}$ to A, the contribution of A on input $\mathbf{y}$ to a transcript τ is defined by $q_A(\tau; \mathbf{y}) = \prod_{i \in A} q_i(\tau; y_i)$. Consider a partition of $[t]$ into sets A and B. For any input $\mathbf{x}$ whose restrictions to A and B are $\mathbf{y}$ and $\mathbf{z}$, respectively, the above property implies that $(q_A(\cdot; \mathbf{y}), q_B(\cdot; \mathbf{z}))$ is a rectangular pair on the set of transcripts.

Let P_A, Q_A, P_B, Q_B be any four functions on the same domain and same range $[0, 1]$ such that each of the pairs (P_A, P_B), (P_A, Q_B), (Q_A, P_A), and (Q_A, P_B) is rectangular. Suppose that for any $\alpha \geq 1/2$, we can show that there exists a constant β depending only on α such that

$$\begin{aligned} &\mathrm{D}_\alpha(P_A P_B, Q_A Q_B) \\ &\quad \leq \beta \cdot (\mathrm{D}_\alpha(P_A P_B, Q_A P_B) \\ &\qquad + \mathrm{D}_\alpha(P_A P_B, P_A Q_B)). \quad (6) \end{aligned}$$

Using Equation (6), we can prove Equation (4) as follows. Assume without loss of generality that t is a power of two. Let $r = 2^h$ for some $h \geq 0$. We will prove by induction on h that for any $0 \leq q \leq t - r$,

$$\mathrm{D}_\alpha(\Pi_\mathbf{0}, \Pi_{0^q 1^r 0^{t-q-r}}) \leq \beta^h \sum_{i=q+1}^{q+r} \mathrm{D}_\alpha(\Pi_\mathbf{0}, \Pi_\mathbf{e_i}).$$

For $h = 0$, this is an equality, so assume that $h > 0$. Let $A = [q + r/2]$ and $B = [t] \setminus A$ be a partition of $[t]$. Let $\mathbf{a} = 0^{q+r/2}$ and $\mathbf{a}' = 0^q 1^{r/2}$ be two input assignments to A. Let $\mathbf{b} = 0^{t-q-r/2}$ and $\mathbf{b}' = 1^{r/2} 0^{t-q-r}$ be two input assignments to B. Setting $P_A = q_A(\cdot; \mathbf{a})$, $Q_A = q_A(\cdot; \mathbf{a}')$, $P_B = q_B(\cdot; \mathbf{b})$, and $Q_B = q_B(\cdot; \mathbf{b}')$, we can apply Eqn. (6). A minor calculation then proves the induction step. Finally, setting $h = \log t$ and $q = 0$ proves Eqn. (4) with $c = c_{\alpha,t} = 1/t^{\log \beta}$.

It remains to prove Eqn. (6) and to choose appropriate values for α. Ideally, we will need β to be as close to 1 as possible. For general communication, a simple argument shows that $\mathrm{h}^2 = \mathrm{D}_{1/2}$ satisfies Eqn. (6) with $\beta = 2$.

Plugging these values in Equation (5) will then imply Theorem 23, part(1). For one-way communication protocols, we obtain an improved result by employing their special structure.

Observe that in the induction argument, the only partitions (A, B) of $[t]$ that arise are of the form $A = [k]$ and $B = [t] \setminus A$, for some $0 \leq k < t$. Write any transcript τ as (τ_A, τ_B), where τ_A (respectively, τ_B) is that part of τ produced by players in A (respectively, B). The one-way protocol satisfies a crucial Markovian property in the manner described below.

For any assignment y to A, notice that $q_A(\tau_A, \tau_B; y)$ does not depend on τ_B; thus we can define a function $p_{A,y}$ on "partial" transcripts τ_A such that $p_{A,y}(\tau_A) = q_A(\tau_A, \tau_B; y)$, for all τ_B. Furthermore, it can be verified that $p_{A,y}$ is a probability distribution. Similarly, for every input z to B, and every τ_A, it can be shown that $\sum_{\tau_B:\exists\tau=(\tau_A,\tau_B)} q_B(\tau_A, \tau_B; z) = 1$. Thus, the matrix $M_{B,z}$ whose (τ_A, τ_B)-th entry is $q_B(\tau_A, \tau_B; z)$ is a probability transition matrix.

For a probability vector p on Ω and a probability transition matrix M on $\Omega \times \Gamma$, let $p \circ M$ denote the joint distribution on $\Omega \times \Gamma$ where $p \circ M(i, j) = p(i) \cdot M(i, j)$. Then, Eqn. (6) reduces to the following lemma; the proof uses convexity and analytical arguments and is described in the full paper.

Lemma 24. *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 $\epsilon > 0$, if $\alpha > 1 - \epsilon^2/(4(1+\epsilon))$, then*

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

We now prove Theorem 23, part(2), as follows. Given ϵ, we choose ϵ' such that $t^{\log(1+\epsilon')} \leq t^\epsilon$. Therefore, $\epsilon' = 2^\epsilon - 1$ will do. We choose $\alpha = 1 - \epsilon'^2/(4(1+\epsilon'))$, and substitute in Equation (5). □

As consequences of Theorem 23, parts (1) and (2), we obtain space lower bounds for computing frequency moments in the single-pass and constant-pass data stream models. The details of the reduction from data stream lower bounds to communication lower bounds are given in the full version of the paper.

Theorem 25. *For any k and $\gamma > 0$, any (one-pass) data stream algorithm that approximates $F_k(n, m)$ to within a constant factor with probability $> 3/4$ requires $\Omega\left(m^{1-(2+\gamma)/k}\right)$ space. For the same problem, any data stream algorithm that makes constant number of passes requires $\Omega\left(m^{1-3/k}\right)$ space.*

8. L_∞ and L_p distances

In the L_∞ promise problem, Alice and Bob are, respectively, given 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 $\mathrm{L}_\infty(\mathbf{x}, \mathbf{y}) = 1$ iff the latter case holds.

Theorem 26. *For $\delta < 1/4$,*
$R_\delta(\mathrm{L}_\infty) \geq \frac{1-2\sqrt{\delta}}{4} \cdot \frac{n}{m^2}$.

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

We thus use the direct sum paradigm with the input distribution $\boldsymbol{\mu} = \nu^n$. The random variable (X, Y) distributed according to ν is defined as follows. Let λ be the uniform distribution on $([0, m] \times \{0, 1\}) \setminus \{(0, 1), (m, 0)\}$, and let $D \sim \lambda$. If $D = (d, 0)$, then $X = d$ and $Y \in_R \{d, d+1\}$; if $D = (d, 1)$, then $Y = d$ and $X \in_R \{d-1, d\}$. It is easy to see that λ partitions ν. Furthermore, since $\mathrm{DIST}(x, y) = 0$ for all (x, y) generated according to ν, it follows μ is a collapsing distribution for L_∞. The theorem follows by applying Lemma 27 given below. □

Lemma 27. *For any $\delta < 1/4$,*
$\mathrm{IC}_{\nu,\delta}(\mathrm{DIST} \mid \lambda) \geq \frac{1-2\sqrt{\delta}}{4} \cdot \frac{1}{m^2}$.

Proof. Let Π be any δ-error protocol for DIST, and let $U_d \in_R \{d, d+1\}$. By expanding on values of D, it can be shown that

$$\mathrm{IC}_{\nu,\delta}(\Pi \mid \lambda) = \frac{1}{2m}\left(\sum_{d=0}^{m-1} \mathrm{I}(U_d \,;\, \Pi(d, U_d)) + \sum_{d=1}^{m} \mathrm{I}(U_{d-1} \,;\, \Pi(U_{d-1}, d))\right).$$

Therefore,

$$\begin{aligned}
&\mathrm{IC}_{\nu,\delta}(\Pi \mid \lambda) \\
&\geq \frac{1}{2m}\left(\sum_{d=0}^{m-1} \mathrm{h}^2(\Pi_{dd}, \Pi_{d,d+1}) + \sum_{d=1}^{m} \mathrm{h}^2(\Pi_{d-1,d}, \Pi_{dd})\right) \\
&\quad \text{(by Lemma 21)} \\
&\geq \frac{1}{4m^2}\left(\sum_{d=0}^{m-1} \mathrm{h}(\Pi_{dd}, \Pi_{d,d+1}) + \sum_{d=1}^{m} \mathrm{h}(\Pi_{d-1,d}, \Pi_{dd})\right)^2 \\
&\geq \frac{1}{4m^2}\ \mathrm{h}^2(\Pi_{00}, \Pi_{mm}) \qquad \text{(Triangle inequality)}
\end{aligned}$$

We cannot directly bound $\mathrm{h}^2(\Pi_{00}, \Pi_{mm})$ from below, because DIST is 0 on both inputs. However, by Lemma 20, 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 Propositions 15 and 18, is at least $1 - 2\sqrt{\delta}$. □

A consequence of Theorem 26 is a space lower bound for computing L_p distances in the constant-pass data stream model. Once again, the details of the reduction from data stream lower bounds to communication lower bounds are given in the full version of the paper.

Theorem 28. *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 $> 3/4$ requires $\Omega(n^{1-4\epsilon-2/p})$ space.*

References

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

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

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

[8] J. Feigenbaum, S. Kannan, M. Strauss, and M. Viswanathan. An approximate L^1-difference algorithm for massive data streams. In *Proceedings of the 40th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 501–511, 1999.

[9] 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.

[10] 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.

[11] 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.

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

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

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

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

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

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

[18] 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.

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

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

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

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

[23] 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.

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

[25] 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.