On Information Complexity in the Broadcast Model

Mark Braverman*
Department of Computer Science
Princeton University
mbraverm@cs.princeton.edu

Rotem Oshman†
Department of Computer Science
Tel Aviv University
roshman@tau.ac.il

ABSTRACT

Information complexity is the extension of classical information theory to the *interactive* setting, where instead of one-way transmission we are interested in back-and-forth communication. This approach has been very influential in communication complexity, where it enables us to prove powerful lower bounds by quantifying the amount of information the participants in the computation must reveal about their inputs. In this paper we study information complexity in the classical broadcast model: k parties with private inputs wish to compute some function of their inputs, and they communicate by sending messages (one at a time) over a broadcast channel. We measure how much information the players reveal about their inputs to an external observer. This is called *external information cost*.

Using this approach, we prove a tight lower bound of $\Omega(n \log k + k)$ on the communication complexity of *set disjointness*, a fundamental problem in communication complexity. We also give a deterministic matching upper bound.

Next we study *compression*, a central question in information complexity: given a protocol with low information cost (but possibly high communication), can we *compress* the protocol so that its communication cost matches its information cost? In the two-player setting, it is known that every protocol can be compressed to roughly its external information cost. We show that for the multi-party case this is no longer true: there is a gap of at least $\Omega(k/\log k)$ between external information and communication. However, if we wish to compress *many* independent instances of the same protocol, then it is possible to do so with an amortized per-copy cost that approaches the information cost as the number of copies goes to infinity.

*Research supported in part by an NSF CAREER award (CCF-1149888), a Turing Centenary Fellowship, a Packard Fellowship in Science and Engineering, and the Simons Collaboration on Algorithms and Geometry.

†Research supported in part by the I-CORE Program of the Planning and Budgeting Committee and the Israel Science Foundation, Grant No. 4/11.

PODC'15, July 21–23, 2015, Donostia-San Sebastián, Spain.

ACM 978-1-4503-3617-8 /15/07 ...$15.00.
DOI: http://dx.doi.org/10.1145/2767386.2767425.

1. INTRODUCTION

Classical information theory studies the cost of transmitting information from a sender to a receiver, over a channel which may be subject to various types of noise. The traditional focus is on *amortized cost in the limit*: the sender has many messages $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ drawn independently from the same distribution, and we are interested in the amortized number of bits it must communicate per message (that is, the total communication required to send all n messages, divided by n), in the limit as n grows to infinity. This question was addressed by Shannon in his seminal work [27], where he showed that in the noiseless case the amortized per-copy cost is exactly $H(\boldsymbol{X})$, the entropy of the distribution from which the messages are drawn. Later, Huffman showed [20] that a *single* copy of $\boldsymbol{X}$ can be sent in $H(\boldsymbol{X}) + 1$ bits, so for (noiseless) one-way transmission, the single-shot cost and the cost-in-the-limit are essentially the same.

The emerging field of *information complexity* seeks to extend the ideas of classical information theory to an interactive setting, where the parties involved in the computation communicate back-and-forth in order to achieve some shared goal. The basic two-party setup has two players, Alice and Bob, who receive private inputs $\boldsymbol{X}, \boldsymbol{Y}$ respectively, and wish to compute some function $f(\boldsymbol{X}, \boldsymbol{Y})$ (or more generally, perform some task that depends on the inputs). We are interested in characterizing the amount of information the players must reveal about their inputs in order to compute f; this is called the *information cost* of f. In this paper we work with *external information cost* [7], which measures the amount of information revealed to an external observer who initially does not know the inputs to the protocol. (See [7] for alternative ways to measure information cost.)

Information cost turns out to be a natural lower bound on *communication cost*, the number of bits that the players need to send to compute f, so a lower bound on information cost translates immediately to a lower bound on communication. In addition, information cost has some nice properties, such as additivity, that can make it easier to work with than communication. Consequently, information lower bounds have been very useful in proving communication lower bounds (e.g., [2, 17, 5, 29]).

One can also ask whether, as is true for one-way transmission, any protocol can be *compressed* to its information cost. This question has not been fully resolved yet, but the answer will have deep implications in theoretical computer science [7, 3, 15]. Both single-shot compression [3] and amortized cost in the limit [7] are of great interest, and they appear to have very different costs [15]. In the two-party setting it is known that any protocol can be compressed almost all the way down to its external information cost, with a logarithmic dependence on the communication complexity of the original, uncompressed protocol [3].

The broadcast model. In this paper we study the classical broadcast model of communication (also called the *shared blackboard model* in communication complexity literature [22]). We have k players, each with a private input $X_i \in \{0,1\}^n$, and the players wish to compute a joint function $f(X_1, \ldots, X_k)$ of their inputs. The players communicate by "writing messages on a blackboard" that all players can read for free (i.e., they communicate by broadcasting messages to each other). We allow the players to use randomization when deciding what to write on the board. This model is widely studied in theoretical computer science, in areas ranging from streaming [1, 2, 17] to cryptography [16] and mechanism design [13]. Beyond being a fundamental model of communication complexity, the broadcast model can also be viewed as an abstract model of single-hop wireless networks, which abstracts away the details of contention management.

We study two fundamental questions in the broadcast model: the communication complexity of set disjointness, and the question of whether interactive protocols can be compressed.

Set disjointness in the broadcast model. Set disjointness is one of the most fundamental problems in communication complexity; it has been studied in the classical two-player model [21, 25], in the number-on-forehead model (e.g., [28]), and in many other settings (see [11] for a survey); it has applications in various areas ranging from streaming [1] to data structures [23], distributed computing [26], circuit lower bounds [10] and many other examples. In multi-party (number-in-hand) set disjointness, each of the k players receives a subset $X_i \subseteq [n]$, and the players need to determine whether $\bigcap_{i=1}^n X_i = \emptyset$; formally, we shall denote

$$\textsc{Disj}_{n,k}(X_1, \ldots, X_k) = \neg \bigvee_{j=1}^{n} \bigwedge_{i=1}^{k} X_i^j,$$

where X_i^j is coordinate j of player i's input. Set disjointness reduces to many natural problems, so a lower bound on its communication complexity implies lower bounds on many other problems as well.

It is known that for two players, set disjointness requires $\Omega(n)$ bits of communication [21, 25], and a trivial reduction shows that $\Omega(n)$ is also a lower bound for k players. It is not difficult to see that $\Omega(k)$ is also a lower bound, as intuitively, each player must speak at least once to solve the problem. As for upper bounds, there is a simple protocol with communication complexity $O(n \log n + k)$: the players go in order, with each player i writing on the board the coordinates j where $X_i^j = 0$, unless they already appear on the board. A player that has no new zero coordinates to contribute writes a single bit to indicate this. After all players have taken their turn, if there is some coordinate that does not appear on the board, then this coordinate is in the intersection; otherwise the intersection is empty.

A-priori, it is not obvious whether the "right answer" is $\Theta(n+k)$, $\Theta(n \log n + k)$, or somewhere in between. After all, in some cases where one might naively expect a logarithmic factor to arise, it does not: a famous example is the randomized protocol of Håstad and Wigderson [19], which solves two-player set disjointness under the promise that $|X| = |Y| = s$ in $O(s)$ bits, instead of the naive $O(s \log n)$. (In fact, two players can even *compute the exact intersection* of their sets X, Y using $O(s)$ bits when $|X| = |Y| = s$ [6, 8].)

In this paper we show that the randomized communication complexity of set disjointness with k players is actually $\Theta(n \log k + k)$. Intuitively, the factor $\log k$ arises because the protocol must "find", for each $j \in [n]$, some player i with $X_i^j = 0$, before it can declare that $\bigcap_{i=1}^k X_i = 0$. The index of such a player is "worth" $\log k$ bits of information, as this is how many bits we require to encode it. We are able to formalize this intuition using notions from information theory and prove a lower bound of $\Omega(n \log k + k)$ on the information cost of set disjointness

To our knowledge, ours is the first communication lower bound on a decision problem in the broadcast model that grows with the number of players. This serves as evidence that while reductions from two-party communication complexity are incredibly useful [12, 24], additional benefit can be derived from directly considering multi-party models. Although an additional factor of $\log k$ in the lower bound may not seem large, communication complexity lower bounds are often applied in distributed computing to prove lower bounds on the number of rounds required to solve some task, and in that context one divides by the number of bits that can be sent per round, which can end up being linear in the number of participants (see, e.g., [14]). So for instance, when $k = \Theta(n)$, the difference between a lower bound of $\Omega(n \log k)$ and a lower bound of $\Omega(n)$ on the total communication complexity can end up being the difference between a logarithmic round lower bound and a constant (i.e., trivial) bound.

We remark that a *promise* version of set disjointness has received significant attention in the broadcast model ([2, 17] and others), due to its connections to streaming lower bounds [1]. Set disjointness and other problems have also recently been investigated in a *point-to-point* model of multi-party communication ([5, 29] and others). And finally, in [24], a technique called *symmetrization* is introduced to obtain multi-party lower bounds on several pointwise-Boolean functions, such as pointwise-OR, where the output is a vector Y such that $Y_i = \bigvee_j X_i^j$ for each $i \in [n]$. Although symmetrization can prove a lower bound of $\Omega(n \log k)$ on pointwise-OR, this technique seems too weak to prove a lower bound of $\Omega(n \log k)$ on set disjointness.

Compressing multi-party broadcast protocols. For two players, it is known that any interactive protocol can be compressed to roughly its external information cost [3], which is the amount of information revealed to an external observer about the players' inputs. We observe that for multiple parties this no longer holds: for the function $\bigwedge_{i=1}^k X_i$, the information cost under any distribution is bounded by $O(\log k)$, but we show that there is a distribution under which the communication cost is $\Omega(k)$. This shows that there is a gap of at least $\Omega(k/\log k)$ between external information and communication in the k-player broadcast model. However, while a *single* instance cannot be compressed, we do show that for many instances we can achieve compression: the *amortized* cost per-copy as the number of copies tends to infinity is bounded by the external information cost, generalizing the result of [7] for two players. We note that while [7] deals with *internal information*, which measures what the players learn about each others' inputs, here are we concerned with *external information*, which measures what an external observer learns. External information is generally an easier quantity to work with, and consequently we are able to simplify the protocol from [7].

Organization. The remainder of the paper is organized as follows. Before diving into the technical details, we give an informal overview of the set disjointness lower bound in Section 2. In Section 3 we review some basic notions from communication complexity and information theory, and in Section 4 we prove the two lower bounds on set disjointness, of $\Omega(n \log k)$ and $\Omega(k)$ respectively. A matching deterministic upper bound, with communication complexity $O(n \log k + k)$, is presented in Section 5. In Sec-

tion 6 we discuss the issue of compression, prove that there is a gap of $\Omega(k/\log k)$ between information and communication, and show that computation involving many independent instances of the same problem can be compressed such that in the limit as the number of instances goes to infinity, the per-instance cost tends to the information cost of the problem.

2. HIGH-LEVEL OVERVIEW OF THE SET DISJOINTNESS LOWER BOUND

Our lower bound follows the information-theoretic approach [9, 2]: in order to prove a lower bound on the communication complexity of set disjointness, we will show that any protocol that solves it with small (constant) error must reveal a lot of information about the players' inputs, in the information-theoretic sense: the entropy of the inputs (i.e., our uncertainty about them) is significantly reduced once the transcript of the protocol is observed. Each bit the protocol communicates can reveal at most one bit of information about the inputs, so we immediately obtain a lower bound on communication as well.

The precise notion of information cost we use in this paper was introduced in [2], and is called *conditional information cost*. We will work with a distribution $(\boldsymbol{X}, \boldsymbol{D}) \sim \mu$ on two variables: the input the players, $\boldsymbol{X} = \boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$, and an auxiliary variable $\boldsymbol{D}$. The essential property of the distribution is that conditioned on $\boldsymbol{D}$, the inputs $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ are independent, and this will be crucial to our proof. Conditional information cost measures the information an external observer learns about the inputs, conditioned on $\boldsymbol{D}$.

One of the advantages of the information-based approach is that it allows us to decompose the problem into many small problems, prove that each small problem is "a little bit hard", and obtain a lower bound for the overall problem that is equal to the sum of the bounds on the smaller problems. This is called *direct sum* [2]. In our case, as in [2], since $\textsc{Disj}_{n,k} = \neg \bigvee_{j=1}^{n} \bigwedge_{i=1}^{k} X_i^j$, we will break the problem up into n instances of one-bit AND. We prove a lower bound of $\Omega(\log k)$ on AND, and obtain a lower bound of $\Omega(n \log k)$ for disjointness.

In order to show that any protocol computing AND must reveal $\Omega(\log k)$ bits of information about the input, we argue that whenever the output of the protocol is 0, the protocol's transcript must "point" to some player i whose input is zero. The notion of "pointing to a player" is formalized using the *posterior probability* that $X_i = 0$ after the transcript is observed. Under our input distribution, the prior probability that $X_i = 0$ is only $O(1/k)$, but we will show that for "most" transcripts, there is a player i with a *constant* posterior probability of $X_i = 0$ given the transcript. Intuitively, since a-priori we had $X_i = 0$ with probability only $O(1/k)$, we are "very surprised" to learn that $X_i = 0$ with constant probability, and this "surprise" is worth $\Omega(\log k)$ bits of information. Viewed another way: a-priori, since each player receives 0 with probability roughly $1/k$, we expect a small (constant) number of players to receive 0, but we have no idea who they are. Once the transcript is revealed, we can "point" to a player that probably received 0, so our uncertainty about the input is greatly reduced. Since the identify of the player requires $\Omega(\log k)$ bits to write down, this is the amount of information we have learned about the input.

We proceed to review the definitions and notions required to make this outline formal.

3. PRELIMINARIES

Notation. We use bold-face letters to denote random variables. For variables $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_\ell$ with joint distribution μ, we let $\mu(\boldsymbol{A}_i)$ denote the marginal distribution of $\boldsymbol{A}_i$, and $\mu(\boldsymbol{A}_i | \boldsymbol{A}_j = a_j)$ denote the distribution of $\boldsymbol{A}_i$ conditioned on $\boldsymbol{A}_j = a_j$ (and similarly for more variables). For a string S, we let $|S|$ denote the length of S.

The broadcast model. In the *broadcast model*, also called the *shared blackboard model* [22], we have k players. Each player i receives a private input X_i, and the players wish to compute a joint function $f(X_1, \ldots, X_n)$ of their inputs. Communication is via a shared blackboard (or broadcast channel) that all players can read for free. At each point in the protocol, the current contents of the blackboard determine whose turn it is to speak next, and that player generates a message using its input, private and public randomness, and the contents of the blackboard, and writes it on the board. Then the next player speaks, and so on, until eventually the protocol halts and the players output some values (for which they are not charged).

Communication complexity. For a protocol Π, we define the *communication complexity of* Π, denoted $\mathrm{CC}(\Pi)$, as the worst-case number of bits that are written on the board in any execution of Π. The *communication complexity of a problem f with error ϵ*, denoted $\mathrm{CC}_\epsilon(f)$, is the minimum communication complexity of any protocol Π that solves f on any input with error probability at most ϵ over the protocol's randomness.

In Section 6 we shall also study *distributional* (or *average-case*) communication complexity. Given an input distribution μ, we define the *distributional complexity of f with error ϵ*, denoted $\mathrm{D}^\mu_\epsilon(f)$, to be the minimum communication complexity of a protocol that solves f with error probability at most ϵ when inputs are drawn from μ (i.e., the probability is over both the input and the protocol's randomness).

Information theory. We require the following notions from classical information theory.

The *entropy* of a random variable measures how much uncertainty we have about it, and corresponds to the number of bits required to encode the variable (in a lossless encoding).

Definition 1. The *entropy* of a random variable $\boldsymbol{X} \sim \mu$ with support $\mathcal{X}$ is given by

$$H(\boldsymbol{X}) = \sum_{x \in \mathcal{X}} \Pr_{\mu}[\boldsymbol{X} = x] \log \frac{1}{\Pr_{\mu}[\boldsymbol{X} = x]}.$$

We also define *conditional entropy*, which measure the average uncertainty that remains about a random variable $\boldsymbol{X}$, after we observe the value of another random variable $\boldsymbol{Y}$.

Definition 2. For two random variables $\boldsymbol{X}, \boldsymbol{Y}$ with joint distribution μ, the *conditional entropy of $\boldsymbol{X}$ given $\boldsymbol{Y}$* is

$$H(\boldsymbol{X} \mid \boldsymbol{Y}) = \mathop{\mathbb{E}}_{\boldsymbol{y} \sim \mu(\boldsymbol{Y})} H(\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}),$$

where $H(\boldsymbol{X} \mid \boldsymbol{Y} = a)$ denotes the entropy of $\boldsymbol{X}$ with respect to the distribution $\mu(\boldsymbol{X} \mid \boldsymbol{Y} = a)$.

To quantify the amount of information we have learned about a random variable $\boldsymbol{X}$ from observing another variable $\boldsymbol{Y}$, we use *mutual information*, which measures the "loss of uncertainty" about $\boldsymbol{X}$ when given $\boldsymbol{Y}$ (or, symmetrically, $\boldsymbol{Y}$ when given $\boldsymbol{X}$):

Definition 3. The *mutual information* between two random variables $\boldsymbol{X}, \boldsymbol{Y}$ is given by

$$\mathrm{I}(\boldsymbol{X}; \boldsymbol{Y}) = H(\boldsymbol{X}) - H(\boldsymbol{X} \mid \boldsymbol{Y}) = H(\boldsymbol{Y}) - H(\boldsymbol{Y} \mid \boldsymbol{X}).$$

The *conditional mutual information between* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *given* $\boldsymbol{Z}$ is

$$\begin{aligned}\mathrm{I}(\boldsymbol{X};\boldsymbol{Y} \mid \boldsymbol{Z}) &= H(\boldsymbol{X} \mid \boldsymbol{Z}) - H(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{Z}) \\ &= H(\boldsymbol{Y}, \boldsymbol{Z}) - H(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{Z}).\end{aligned}$$

For clarity, the distribution on the variables may sometimes be included as a subscript ("$\mathrm{I}_\mu(\boldsymbol{X};\boldsymbol{Y})$"), but we generally omit it when clear from the context.

Intuitively, the mutual information between two variables $\boldsymbol{A}, \boldsymbol{B}$ should also correspond to how much the distribution of $\boldsymbol{A}$ is distorted from its prior distribution when the value of $\boldsymbol{B}$ is revealed. This relationship is captured using *Kullback-Leibler divergence* (also known as *relative entropy*) to measure the distance (or difference) between two distributions.

Definition 4. Given two distributions μ_1, μ_2 with support $\mathcal{X}$, the *KL divergence* of μ_1 from μ_2 is

$$\mathsf{D}\left(\frac{\mu_1}{\mu_2}\right) = \sum_{x \in \mathcal{X}} \mu_1(x) \log \frac{\mu_1(x)}{\mu_2(x)}.$$

KL divergence is not a metric: it is non-negative and it equals zero only when $\mu_1 = \mu_2$, but it is not symmetric. It is helpful to think of μ_1 as the "true" or posterior distribution, and of μ_2 as our prior belief.

Mutual information and KL divergence are related as follows:

$$\begin{aligned}\mathrm{I}(\boldsymbol{X};\boldsymbol{Y}) &= \mathop{\mathbb{E}}_{\boldsymbol{y} \sim \mu(\boldsymbol{Y})} \mathsf{D}\left(\frac{\mu(\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y})}{\mu(\boldsymbol{X})}\right) \\ &= \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mu(\boldsymbol{X})} \mathsf{D}\left(\frac{\mu(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x})}{\mu(\boldsymbol{Y})}\right), \qquad (1)\end{aligned}$$

and similarly for conditional mutual information. In words: the information $\boldsymbol{Y}$ conveys about $\boldsymbol{X}$ is the expected divergence between the posterior distribution of $\boldsymbol{X}$ when the value of $\boldsymbol{Y}$ is revealed, and the prior distribution of $\boldsymbol{X}$ when $\boldsymbol{Y}$ is not known (or vice-versa).

External information cost. Given a protocol Π, we use $\boldsymbol{\Pi}$ to denote the random variable representing the transcript of Π. (The randomness arises both from the input distribution and the protocol's own private and public randomness.) *External information cost* is defined, by extension of the two-party case [9, 2], as follows:

Definition 5. Given a protocol Π and distribution μ, we define the *external information cost of* Π *under* μ by

$$\underset{\mu}{\mathrm{IC}}(\Pi) := \mathrm{I}(\boldsymbol{\Pi}; \boldsymbol{X}),$$

where $\boldsymbol{X} \sim \mu$ is the joint input to the protocol.

For a problem f, distribution μ and error parameter $\epsilon \in (0,1)$, the *external information cost of* f *under* μ *with error* ϵ is defined to be

$$\underset{\mu}{\mathrm{IC}}(f, \epsilon) := \inf_{\Pi} \underset{\mu}{\mathrm{IC}}(\Pi),$$

where the infimum is taken over all protocols that solve f with worst-case error probability at most ϵ (for any input, over the protocol's randomness).

Observe that $\mathrm{IC}_\mu(\Pi) := \mathrm{I}(\boldsymbol{\Pi}; \boldsymbol{X}) \le H(\boldsymbol{\Pi}) \le |\boldsymbol{\Pi}|$, so any lower bound on the information cost immediately translates to a communication lower bound.

For our lower bound we use the notion of *conditional information cost*, introduced in [2]. It generalizes the usual information cost by introducing conditioning on an auxiliary variable.

Definition 6. Let Π be a randomized protocol, and let μ be a distribution on $\mathcal{X} \times \mathcal{D}$ for some $\mathcal{X}, \mathcal{D}$. The *conditional information cost* of Π with respect to μ is given by

$$\underset{\mu}{\mathrm{CIC}}(\Pi) = \underset{(\boldsymbol{X},\boldsymbol{D}) \sim \mu}{\mathrm{I}}(\boldsymbol{\Pi}; \boldsymbol{X} \mid \boldsymbol{D}).$$

For a problem f and an error parameter $\epsilon \in (0,1)$, we define

$$\underset{\mu}{\mathrm{CIC}}(f, \epsilon) = \inf_{\Pi} \underset{\mu}{\mathrm{CIC}}(\Pi),$$

where the infimum is taken over all protocols that solve P with worst-case error ϵ on any input.

As with external information, conditional information is a lower bound on communication complexity.

Now we are ready to prove our main lower bound.

4. SET DISJOINTNESS LOWER BOUND

Following the approach of [2], in order to prove a lower bound of $\Omega(n \log k)$ on the communication complexity of $\textsc{Disj}_{n,k}$, we decompose the disjointness problem into n copies of k-player $\textsc{And}_k$ on a single bit, and argue that the cost adds up linearly.

LEMMA 1 ([2]). *Let* μ *be a distribution on* $\{0,1\}^k \times \mathcal{D}$*, such that*

(1) For any (X, D) *in the support of* μ*,* $\bigwedge_{i=1}^k X_i = 0$*,*[1] *and*

(2) For any $d \in \mathcal{D}$*, when we draw* $(\boldsymbol{X}, \boldsymbol{D}) \sim \mu$ *conditioned on* $\boldsymbol{D} = d$*, the variables* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ *are independent.*

Then $\mathrm{CIC}_{\mu^n}(\textsc{Disj}_{n,k}, \epsilon) \ge n \cdot \mathrm{CIC}_\mu(\textsc{And}_k, \epsilon)$.

Therefore, to obtain a lower bound of $\Omega(n \log k)$ on $\textsc{Disj}_{n,k}$, it is sufficient to prove a lower bound of $\Omega(\log k)$ on $\textsc{And}_k$, and this is is what we shall do next. We will separately show that $\Omega(k)$ is also a lower bound on the communication complexity of $\textsc{And}_k$, but this is an easy argument that does not require information theory.

4.1 Lower Bound for Single-Bit AND

For this lower bound we must exhibit a "hard distribution" μ on inputs $\{0,1\}^k$, which satisfies the conditions of Lemma 1, and on which $\textsc{And}_k$ has information cost $\Omega(\log k)$. Our considerations in choosing the distribution μ are as follows:

- To satisfy condition (1) of Lemma 1, the distribution must always assign zero to at least one player. We ensure this by choosing one uniformly random player $\boldsymbol{Z} \in [k]$ and assigning it zero. We will satisfy condition (2) by designing our distribution such that conditioned on the value of $\boldsymbol{Z}$, the players' inputs will be independent. Our analysis will be conditioned on $\boldsymbol{Z}$ (that is, $\boldsymbol{Z}$ will serve as our "auxiliary variable" previously denoted by $\boldsymbol{D}$).
- For the remaining players, intuitively we wish to have as few zeroes as possible, to make it hard for the protocol to find a player that received zero. One might even be tempted to assign one to all players other than $\boldsymbol{Z}$, but since $\mathrm{CIC}(\Pi) = \mathrm{I}(\boldsymbol{\Pi}; \boldsymbol{X} \mid \boldsymbol{Z}) \le H(\boldsymbol{X}|\boldsymbol{Z})$, we cannot do this—we must ensure that the input has residual entropy at least $\Omega(\log k)$ even conditioned on $\boldsymbol{Z}$. (If all players other than $\boldsymbol{Z}$ receive 1, the residual entropy would be 0.)

[1] This requirement may seem odd: if we fix the output of $\textsc{And}_k$ to zero, how can the distribution be hard for the protocol? The answer is that the protocol must be correct w.h.p. on *any* input, even inputs that are not in the support. The distribution is used only to analyze the *information cost* of the protocol, not its correctness.

- Ideally, then, one would choose a single player other than $\boldsymbol{Z}$, assign it zero, and assign everyone else one. But this would not be a product distribution, even conditioned on $\boldsymbol{Z}$.

- We can come close to the ideal above by assigning each player other than $\boldsymbol{Z}$ a zero with probability $1/k$, independent of the other players.

Formally, the distribution μ is defined as follows: first we select a uniformly random player $\boldsymbol{Z} \in [k]$; we set $\boldsymbol{X}_{\boldsymbol{Z}} = 0$. For each $i \neq \boldsymbol{Z}$ we have $\Pr[\boldsymbol{X}_i = 0 \mid \boldsymbol{Z}] = 1/k$ independent of all the other inputs.

Under the distribution μ described above, it is somewhat likely (constant probability) that besides the special player $\boldsymbol{Z}$, exactly one other player receives zero. In our analysis we condition on this event, which makes the distribution symmetric: a uniformly random pair of players receive zero, and the others receive one. Notice that conditioned on exactly two players receiving zero and on $\boldsymbol{Z}$, the identity of the other player that received zero is uniform in $[k] \setminus \{\boldsymbol{Z}\}$, so it is worth roughly $\log k$ bits of information. Our proof argues that the protocol must *find* some player that received zero, and since, due to the symmetry, it cannot tell which is the special player $\boldsymbol{Z}$, it finds the *other* player that got zero with probability $1/2$. This will show that the conditional information cost of the protocol is $\Omega(\log k)$.

Posterior probabilities. What does it mean for the protocol to "find a player that received zero"? We formalize this in terms of the *posterior probability distribution* of the input $\boldsymbol{X}$ given the transcript and $\boldsymbol{Z}$. "Finding a zero" will mean that for some $i \in [k]$, the posterior probability of $\boldsymbol{X}_i = 0$ given the transcript is *constant* (whereas the prior was only $O(1/k)$).

Recall that the conditional information cost can be re-phrased in terms of the KL-divergence between the posterior and prior of $\boldsymbol{X}$:

$$\mathrm{I}(\boldsymbol{\Pi}; \boldsymbol{X} \mid \boldsymbol{Z}) = \mathop{\mathbb{E}}_{\boldsymbol{\ell},\boldsymbol{z}} \mathsf{D}\left(\frac{\mu(\boldsymbol{X} \mid \boldsymbol{\Pi} = \boldsymbol{\ell}, \boldsymbol{Z} = \boldsymbol{z})}{\mu(\boldsymbol{X} \mid \boldsymbol{Z} = \boldsymbol{z})}\right).$$

It is more convenient to work with the distributions of the individual inputs $\boldsymbol{X}_i$ rather than the distribution of the entire input $\boldsymbol{X}$. And this will be sufficient to prove the lower bound:

LEMMA 2.

$$\mathrm{I}(\boldsymbol{\Pi}; \boldsymbol{X} \mid \boldsymbol{Z}) \geq \sum_{i=1}^{k} \mathop{\mathbb{E}}_{\boldsymbol{\ell},\boldsymbol{z}} \mathsf{D}\left(\frac{\mu(\boldsymbol{X}_i \mid \boldsymbol{\Pi} = \boldsymbol{\ell}, \boldsymbol{Z} = \boldsymbol{z})}{\mu(\boldsymbol{X}_i \mid \boldsymbol{Z} = \boldsymbol{z})}\right).$$

In other words, it is enough to show that for "many" transcripts ℓ there is *some* player i whose posterior distribution given $\boldsymbol{\Pi} = \ell$ is very different from his prior distribution (all conditioned on $\boldsymbol{Z}$). In general, when $i = \boldsymbol{Z}$, the divergence is zero: we know in advance that $\boldsymbol{X}_i = 0$, so the posterior and the prior are the same. So in fact we are interested in bounding

$$\mathrm{I}(\boldsymbol{\Pi}; \boldsymbol{X} | \boldsymbol{Z}) \geq \frac{1}{k} \sum_{i=1}^{k} \sum_{z \neq i} \mathop{\mathbb{E}}_{\boldsymbol{\ell} \sim \pi | \boldsymbol{Z} = z} \mathsf{D}\left(\frac{\mu(\boldsymbol{X}_i \mid \boldsymbol{\Pi} = \boldsymbol{\ell}, \boldsymbol{Z} = z)}{\mu(\boldsymbol{X}_i \mid \boldsymbol{Z} = z)}\right). \tag{2}$$

As we said, we will show that for many transcript ℓ, some player has "shown his hand" and revealed that its input was probably 0: for some $i \neq z$, the posterior probability of $\boldsymbol{X}_i = 0$ given $\boldsymbol{\Pi} = \ell, \boldsymbol{Z} = z$ is constant. Since the prior is only $O(1/k)$, this gives us a divergence of $\Omega(\log k)$: specifically, if the posterior probability of 0 is p, then

$$\begin{aligned}
&\mathsf{D}\left(\frac{\mu(\boldsymbol{X}_i \mid \boldsymbol{\Pi} = \ell, \boldsymbol{Z} = z)}{\mu(\boldsymbol{X}_i \mid \boldsymbol{Z} = z)}\right) \qquad (3)\\
&= \sum_{b=0,1} \left(\Pr[\boldsymbol{X}_i = b \mid \boldsymbol{\Pi} = \ell, \boldsymbol{Z} = z] \cdot \right.\\
&\qquad \left. \cdot \log \frac{\Pr[\boldsymbol{X}_i = b \mid \boldsymbol{\Pi} = \ell, \boldsymbol{Z} = z]}{\Pr[\boldsymbol{X}_i = b \mid \boldsymbol{Z} = z]} \right)\\
&= p \log \frac{p}{\frac{1}{k}} + (1-p) \log \frac{1-p}{1-\frac{1}{k}} \geq p \log k - H(p)\\
&\geq p \log(k) - 1. \qquad (4)
\end{aligned}$$

(As usual, we assume the convention that $0 \log 0 = 0$.) Plugging (4) into (2), with constant p, yields the lower bound. But first we must show that indeed many transcripts "point" to a player that received zero.

Finding zeroes. In order to analyze the posterior probabilities, we examine the structure of the protocol. As usual for this kind of proof, we use the fact that for any transcript, the probability of getting this transcript can be broken up into the product of functions that each depend only on the input to a single player:

LEMMA 3 ([2]). *For any transcript ℓ, there exist functions $\{q^{\ell}_{i,b}\}_{i \in [k], b \in \{0,1\}}$ s.t. $\Pr[\boldsymbol{\Pi}(X) = \ell] = \prod_{i=1}^{k} q^{\ell}_{i,X_i}$.*

This is proven by an easy induction on rounds: at each point in the protocol, some player i is next to speak. If the transcript so far is ℓ, the probability of getting a particular extension ℓm depends only on player i's random choices and its input X_i. Therefore we set $q^{\ell m}_{i,X_i} = q^{\ell}_{i,X_i} \cdot \Pr[\text{player } i \text{ sends } m \text{ on inputs } X_i \text{ after seeing } \ell]$, and for each $j \neq i$ we set $q^{\ell m}_{j,X_j} = q^{\ell}_{j,X_j}$.

The probability that a particular transcript ℓ is generated when $\boldsymbol{X}_i = b$ is proportional to $q^{\ell}_{i,b}$. Therefore, the posterior probability that $\boldsymbol{X}_i = 0$ given transcript ℓ is related to the ratio between $q^{\ell}_{i,0}$ and $q^{\ell}_{i,1}$: intuitively, the larger $q^{\ell}_{i,0}$ is with respect to $q^{\ell}_{i,1}$, the more we are inclined to believe that $\boldsymbol{X}_i = 0$ if we observe transcript ℓ. Indeed, using Bayes' rule we can show:

LEMMA 4. *Suppose that $q_{i,1} \neq 0$, and let $\alpha^{\ell}_i := q^{\ell}_{i,0}/q^{\ell}_{i,1}$. For any transcript ℓ we have*

$$\begin{aligned}
\Pr[\boldsymbol{X}_i = 0 \mid \boldsymbol{\Pi} = \ell, \boldsymbol{Z} \neq i] &= \frac{q^{\ell}_{i,0}}{q^{\ell}_{i,0} + (k-1) q_{i,1}}\\
&= \frac{\alpha^{\ell}_i}{\alpha^{\ell}_i + k - 1} \geq \frac{\alpha^{\ell}_i}{\alpha^{\ell}_i + k}. \qquad (5)
\end{aligned}$$

When $q_{i,1} = 0$, the posterior is of course 1.

Notice that when $\alpha^{\ell}_i = \Omega(k)$ this posterior probability becomes constant. The factor of $\Omega(k)$ is needed to overcome the prior probability that $\boldsymbol{X}_i = 0$, which is only $1/k$.

"Good transcripts". Our task now is to show that for "many" transcripts ℓ we indeed have $\alpha^{\ell}_i = \Omega(k)$ for some $i \in [k]$. Not all transcripts satisfy this property: for example, the protocol can decide with probability ϵ to throw its hands up and do nothing. However, we will show that transcripts that do not point to a player that received zero contribute to the error of the protocol, and therefore the total probability of getting such a transcript cannot be large.

We focus our attention on inputs that have exactly two zeroes, and our "good transcripts" will also be chosen with respect to their

behavior on these inputs. Let $\mathcal{X}_c$ denote the set of inputs with c zeroes for $c \in [k]$. Note that the correct answer for AND_k is 1 on inputs in $\mathcal{X}_0$ (there is only one such input, 1^k) and 0 on inputs in $\mathcal{X}_c$ for $c \geq 1$.

Let π_2 be the distribution of transcripts conditioned on the input being in $\mathcal{X}_2$:

$$\pi_2(\ell) = \sum_{X \in \mathcal{X}_2} \left[\mu(X|_{\mathcal{X}_2}) \prod_{i=1}^{k} q^{\ell}_{i,X_i} \right].$$

Let L be the set of transcripts satisfying the following constraints: for each $\ell \in L$, the output of the protocol is 0, and also

$$\pi_2(\ell) \geq C \cdot \prod_{i=1}^{k} q^{\ell}_{i,1},$$

where C is some large constant whose value will be chosen later. In other words, L is the set of transcripts with output 0 which "strongly prefer" inputs with two zeroes over 1^k. (In particular, these transcripts do not contribute much to the error of the protocol.)

To show that L has large mass under π_2, let us partition the complement of L into two sets B_0, B_1 based on the output value on each transcript. Neither set can be large:

- The transcripts in B_1 cannot have large mass under π_2 because they yield the wrong output (1) on all inputs in $\mathcal{X}_2$:

$$\begin{aligned}
\pi_2(B_1) &= \sum_{\ell \in B_1} \sum_{X \in \mathcal{X}_2} \mu(X|_{\mathcal{X}_2}) \Pr[\boldsymbol{\Pi} = \ell \mid \boldsymbol{X} = X] \\
&= \frac{1}{\mu(\mathcal{X}_2)} \sum_{X \in \mathcal{X}_2} \mu(X) \sum_{\ell \in B_1} \Pr[\boldsymbol{\Pi} = \ell \mid \boldsymbol{X} = X] \\
&\leq \frac{1}{\mu(\mathcal{X}_2)} \sum_{X \in \mathcal{X}_{\geq 1}} \mu(X) \sum_{\ell \in B_1} \Pr[\boldsymbol{\Pi} = \ell \mid \boldsymbol{X} = X] \\
&\leq \frac{1}{\mu(\mathcal{X}_2)} \Pr_{\mu}[\Pi \text{ outputs } 1] \leq \frac{\delta}{\mu(\mathcal{X}_2)}.
\end{aligned}$$

- B_0 contains transcripts on which the output is 0, but

$$\pi_2(\ell) < C \cdot \prod_{i=1}^{k} q^{\ell}_{i,1}.$$

  These transcripts contribute to the error of the protocol on 1^k at least in proportion to their mass under π_2, and therefore they also cannot have large mass:

$$\begin{aligned}
\Pr\left[\boldsymbol{\Pi}(1^k) \text{ outputs } 0\right] &\geq \sum_{\ell \in B_0} \Pr\left[\boldsymbol{\Pi}(1^k) = \ell\right] \\
&= \sum_{\ell \in B_0} \prod_{i=1}^{k} q^{\ell}_{i,1} \\
&> \sum_{\ell \in B_0} \pi_2(\ell)/C = \pi_2(B_0)/C,
\end{aligned}$$

  and therefore $\pi_2(B_0) < C \cdot \delta$.

Together we have that assuming $C \cdot \delta < 1/200$ and $\delta/\mu(\mathcal{X}_2) < 1/200$ (both constant requirements), $\pi_2(B_0 \cup B_1) < 1/100$. Note that we can choose C arbitrarily large, and compensate by assuming a smaller error probability δ.

For any transcript $\ell \in L$, we have

$$\pi_2(\ell) = \sum_{X \in \mathcal{X}_2} \mu(X|_{\mathcal{X}_2}) \prod_{i=1}^{k} q^{\ell}_{i,X_i} \geq C \prod_{i=1}^{k} q^{\ell}_{i,1}.$$

Given membership in $\mathcal{X}_2$, all two-zero inputs are equally likely, so $\mu(X|_{\mathcal{X}_2}) = \frac{1}{\binom{k}{2}}$ for any $X \in \mathcal{X}_2$. Note that for each $X \in \mathcal{X}_2$ we can write $\prod_{i=1}^{k} q^{\ell}_{i,X_i} = q^{\ell}_{j_1,0} q^{\ell}_{j_2,0} \cdot \prod_{i \neq j_1, j_2} q^{\ell}_{i,1}$, where $j_1 \neq j_2$ are the two indices where X has zero. Dividing both sides by $\prod_{i=1}^{k} q^{\ell}_{i,1}$ (and renaming the indices for convenience), we obtain

$$\frac{1}{\binom{k}{2}} \sum_{i<j} \alpha^{\ell}_i \alpha^{\ell}_j \geq C.$$

Because $\sum_{i<j} a_i a_i \leq (\sum_i a_i)^2$ for any sequence $a_1, \ldots, a_N$, we get that

$$\sum_i \alpha^{\ell}_i \geq \sqrt{\frac{k(k-1)}{2} \cdot C} \geq \sqrt{\frac{k^2}{4} \cdot C} = \frac{\sqrt{C}}{2} k. \tag{6}$$

That is, the *sum* of the coefficients is linear for each $\ell \in L$. However, we need to show that for many transcripts the *maximum* coefficient is linear, and this does not necessarily hold for all of L; for example, it could be that for half the values of i we have $\alpha^{\ell}_i = \Omega(1)$, instead of having one value of i for which $\alpha^{\ell}_i = \Omega(k)$. To eliminate such transcripts, we further restrict our attention to transcripts that are somewhat likely to appear when there are exactly two zeroes in the input, not more. For such transcripts, there are only two players that can be "pointed to", so only two coefficients can be large.

Formally, we focus our attention on the subset $L' \subseteq L$ of transcripts ℓ satisfying

$$\Pr[\boldsymbol{\Pi} = \ell \mid \boldsymbol{X} \in \mathcal{X}_2] \geq \frac{1}{2} \Pr[\boldsymbol{\Pi} = \ell \mid \boldsymbol{X} \in \mathcal{X}_3],$$

that is, transcripts that "like" inputs with two zeroes not much less than inputs with three zeroes. We have $\pi_2(L') \geq \pi_2(L) - 1/2$, because

$$\begin{aligned}
\pi_2(L \setminus L') &= \sum_{\ell \in L \setminus L'} \Pr[\boldsymbol{\Pi} = \ell \mid \boldsymbol{X} \in \mathcal{X}_2] \\
&\leq \frac{1}{2} \sum_{\ell \in L \setminus L'} \Pr[\boldsymbol{\Pi} = \ell \mid \boldsymbol{X} \in \mathcal{X}_3] \leq \frac{1}{2}.
\end{aligned}$$

Now fix $\ell \in L'$, and let us show that for some $i \in [k]$, the posterior probability of $X_i = 0$ given ℓ is $\Omega(k)$ times the posterior probability of 1. If there is some i for which $q^{\ell}_{i,1} = 0$, then we are done, so assume that this is not the case. Because $\ell \in L'$,

$$\sum_{X \in \mathcal{X}_2} \mu(X|_{\mathcal{X}_2}) \prod_{i=1}^{k} q^{\ell}_{i,X_i} \geq \frac{1}{2} \sum_{X \in \mathcal{X}_3} \mu(X|_{\mathcal{X}_3}) \prod_{i=1}^{k} q^{\ell}_{i,X_i},$$

that is,

$$\frac{1}{\binom{k}{2}} \sum_{i<j} \alpha^{\ell}_i \alpha^{\ell}_j \geq \frac{1}{2\binom{k}{3}} \sum_{i<j<m} \alpha^{\ell}_i \alpha^{\ell}_j \alpha^{\ell}_m. \tag{7}$$

Now assume for the sake of contradiction that for a constant C' whose value will be fixed later, we have $\alpha^{\ell}_i < C'k$ for each $i \in [k]$. Then

$$\begin{aligned}
6 \sum_{i<j<m} \alpha^{\ell}_i \alpha^{\ell}_j \alpha^{\ell}_m &= \left(\sum_i \alpha^{\ell}_i \right)^3 - 3 \sum_{i \neq j} \left(\alpha^{\ell}_i \right)^2 \alpha^{\ell}_j - \sum_i \alpha^3_i \\
&> \left(\sum_i \alpha^{\ell}_i \right)^3 - 3C'k \sum_{i \neq j} \alpha^{\ell}_i \alpha^{\ell}_j - (C')^2 k^2 \sum_i \alpha_i \\
&\geq \left(\sum_i \alpha^{\ell}_i \right)^3 - 3C'k \left(\sum_i \alpha^{\ell}_i \right)^2 - (C')^2 k^2 \sum_i \alpha_i.
\end{aligned}$$

From (6) we know that $k \le 2\sum_i \alpha_i^\ell/(\sqrt{C})$, so

$$6\sum_{i<j<m}\alpha_i^\ell\alpha_j^\ell\alpha_m^\ell \ge \left(\sum_i \alpha_i^\ell\right)^3 - \\ - 6\frac{C'}{\sqrt{C}}\left(\sum_i \alpha_i^\ell\right)^3 - 4\frac{(C')^2}{C}\left(\sum_i \alpha_i\right)^3.$$

If we choose, e.g., $C' < \sqrt{C}/100$, we get that

$$6\sum_{i<j<m}\alpha_i^\ell\alpha_j^\ell\alpha_m^\ell \ge \frac{1}{2}\left(\sum_i \alpha_i^\ell\right)^3 \overset{(6)}{\ge} \frac{1}{4}\sqrt{C}k\left(\sum_i \alpha_i^\ell\right)^2,$$

and therefore

$$\frac{1}{2\binom{k}{3}}\sum_{i<j<m}\alpha_i^\ell\alpha_j^\ell\alpha_m^\ell \ge \frac{\sqrt{C}}{16k^2}\left(\sum_i \alpha_i^\ell\right)^2.$$

This gives us a lower bound on the right-hand side in (7). The left-hand side is bounded from above by

$$\frac{1}{\binom{k}{2}}\sum_{i<j}\alpha_i^\ell\alpha_j^\ell \le \frac{4}{k^2}\left(\sum_i \alpha_i^\ell\right)^2.$$

If we choose C sufficiently large we obtain a contradiction to (7). Note that the value of C' is constrained only by the value of C, so by increasing C (which requires only assuming a smaller error probability δ for the protocol) we can obtain an arbitrarily large lower bound $\max_i \alpha_i^\ell \ge C'k$.

All together, we have now shown the following:

LEMMA 5. *Fix a constant probability $p \in (0,1)$, and let $c > 0$ be a constant such that $c/(c+1) \ge p$. Then there exist constants $\delta, p \in (0,1)$ such that for any k, any protocol that solves AND_k with error at most δ has a set L of transcripts such that $\pi_2(L) \ge p$, and for each $\ell \in L$ there is a player $i = i(\ell)$ with $\alpha_i^\ell \ge ck$.*

Notice that by (5), for the player $i(\ell)$ from the lemma we have $\Pr\left[\boldsymbol{X}_{i(\ell)} = 0 \mid \boldsymbol{\Pi} = \ell, \boldsymbol{Z} \ne i\right] \ge \alpha_{i(\ell)}^\ell/(\alpha_{i(\ell)}^\ell + k) \ge (ck)/(ck+k) \ge p$, i.e., the posterior is constant. From (4), assuming $k \ge 2^{2/p}$, for any $z \ne i(\ell)$,

$$\mathsf{D}\left(\frac{\mu(\boldsymbol{X}_{i(\ell)} \mid \boldsymbol{\Pi} = \ell, \boldsymbol{Z} = z)}{\mu(\boldsymbol{X}_{i(\ell)} \mid \boldsymbol{Z} = z)}\right) \ge p\log k - 1 \ge (p/2)\log k, \tag{8}$$

so we are "very surprised" by the posterior distribution of $\boldsymbol{X}_{i(\ell)}$ given $\boldsymbol{\Pi}$. Now recall that this is lower bound on the divergence is what we needed to show: by Lemma 2, together with some technical details that are omitted here, we obtain $\mathrm{I}_\mu(\boldsymbol{\Pi}; \boldsymbol{X} \mid \boldsymbol{Z}) \ge \Omega(\log k)$, completing the proof.

THEOREM 1. *For the distribution μ we described and a sufficiently small δ, $\mathrm{CIC}_\mu(\mathrm{AND}_k, \delta) \ge \Omega(\log k)$.*

COROLLARY 1. *By Lemma 1 we have $\mathrm{CIC}_{\mu^n}(\mathrm{DISJ}_{n,k}, \delta) \ge \Omega(n\log k)$, and hence $\mathrm{CC}_\delta(\mathrm{DISJ}_{n,k}) \ge \Omega(n\log k)$.*

We now show that computing AND_k requires $\Omega(k)$ bits of communication, which also implies a lower bound of $\Omega(k)$ on $\mathrm{DISJ}_{n,k}$. This is an easy argument which does not require information complexity.

LEMMA 6. *For any constant error parameter $\epsilon \in (0, 1/3)$ we have $\mathrm{CC}_\epsilon(\mathrm{AND}_k) = \Omega(k)$.*

PROOF. By the easy direction of Yao's minimax principle [30], it is sufficient to show that for some input distribution μ we have $\mathrm{D}^\mu_\epsilon(\mathrm{AND}_k) = \Omega(k)$, that is, any deterministic protocol that errs with probability at most ϵ on inputs drawn from μ uses $\Omega(k)$ bits of communication.

Fix $\epsilon' > \epsilon$ such that $\epsilon/(1-\epsilon') < 1/2$. Consider the following input distribution μ: with probability $\epsilon' > \epsilon$, all players receive 1, and with probability $1-\epsilon'$, one random player receives 0 and the other players receive 1.

Fix a deterministic protocol Π for AND_k. Let $p_1, \ldots, p_\ell$ be the order in which players speak when the input is 1^k, and assume for the sake of contradiction that $\ell < (1 - \epsilon/(1-\epsilon'))\cdot k$.

If $\Pi(1^k) = 0$, then Π's error under μ is $\epsilon' > \epsilon$, so we can assume that $\Pi(1^k) = 1$. Let $\mathcal{E}$ be the event that the input is not 1^k, but all of the players $p_1, \ldots, p_\ell$ receive 1. We have $\Pr_\mu[\mathcal{E}] = (1-\epsilon')\cdot(1-\ell/k) > \epsilon$. But when $\mathcal{E}$ occurs, the transcript of Π is identical to its transcript on 1^k, as Π is deterministic and all players that speak when the input is 1^k still receive 1 (in particular, the order of players that speak also remains the same under $\mathcal{E}$). Therefore with probability $> \epsilon$ the protocol outputs the wrong answer. □

5. MATCHING UPPER BOUND FOR SET DISJOINTNESS

We now describe a deterministic protocol for $\mathrm{DISJ}_{n,k}$ with communication complexity $O(n\log k + k)$, which, in light of the lower bounds above, is optimal even for randomized algorithms. The protocol improves upon the naive version described in the introduction: the players attempt to convince themselves that $\bigcap_{i=1}^k X_i = \emptyset$ by writing on the board the indices of coordinates $i \in [n]$ where their input is zero; these coordinates cannot be in the intersection. To reduce the amount of communication, players never write on the board a coordinate that already appears on the board, and we also "pack together" coordinates: instead of writing them one-by-one, which requires $\Theta(\log n)$ bits per coordinate, we write batches of many coordinates simultaneously, encoding them as a set. This reduces the amortized cost per coordinate to $\Theta(\log k)$.

The protocol runs in *cycles*, where in each cycle some prefix of the players $1, \ldots, k$ each speak exactly once, in order, and the remaining players do not speak. Let Z_i be the set of coordinates that do not appear on the board at the beginning of cycle i, and let $z_i := |Z_i|$. Notice that if the input sets are indeed disjoint, then by the pigeonhole principle, at least one player has at least z_i/k zero coordinates that do not appear on the board ("new zeroes").

Suppose that at the beginning of cycle i we still have $z_i \ge k^2$. When it is the turn of player j to speak, if player j has at least z_i/k new zeroes, then it chooses z_i/k of them and writes them on the board, encoded as a subset of Z_i. The number of possible subsets is $\binom{z_i}{z_i/k} \le ((z_i e)/(z_i/k))^{z_i/k} = (ek)^{z_i/k}$, so encoding one subset requires $(z_i/k)\log(ek)$ bits (i.e., the amortized cost per coordinate written is $\log(ek)$). If player j does not have z_i/k new zeroes to contribute, it writes a single bit on the board indicating this ("pass"), and we move on to the next player.

When at the beginning of some cycle i we have $z_i < k^2$, each player simply writes all its new zeroes on the board in the naive encoding (as elements of Z_i). This requires $O(\log k)$ bits per coordinate.

The protocol ends when one of the following occurs: if at any point all coordinates appear on the board, then the players halt and output "disjoint". Otherwise, if a complete cycle goes by in which all players pass, then the players halt and output "non-disjoint". Also, if we reach a cycle i with $z_i < k^2$, and at the end of the cycle

not all coordinates appear on the board, then the players output "non-disjoint".

THEOREM 2. *The protocol describes above solves* $\text{DISJ}_{n,k}$ *in* $O(n \log k + k)$ *bits of communication.*

6. ON INTERACTIVE COMPRESSION FOR THE BROADCAST MODEL

The issue of *compression* is at the heart of information theory: can we condense messages (or, in the interactive case, protocols) to their information content? For the case of one-way transmission we know that the answer is yes, both for the amortized per-message cost in the limit as the number of messages (called the *block size*) goes to infinity [27] and also for just a single message [20].

In the interactive setting, it is known that a single instance of any protocol can be compressed to nearly its external information cost [3], while for amortized compression with an infinite block size (i.e., as the number of copies goes to infinity) we can achieve even better compression, to a measure called the *internal* information cost [7].[2] Let us discuss each of these goals in the k-party broadcast model.

Single-shot compression. In [3] it is shown that a two-party protocol Π with communication complexity C and external information cost I can be simulated (compressed) by a protocol with communication $O(I \cdot \log C)$. (External information for two players is given by Definition 6 for $k = 2$.) However, we observe that we cannot achieve such compression for the general multiparty case in the blackboard model: we already noted that the function AND_k requires $\Omega(k)$ bits of communication under some distribution. We now observe that $\text{IC}_\mu(\text{AND}_k) \leq O(\log k)$ for any distribution μ: we can solve AND_k by a protocol Π, where the players go in order and write their input on the board, until we find a player whose input was 0 (in which case we halt), or have seen that all players had 1. The entropy of the transcript of Π is $O(\log k)$, because we can encode it by writing the index of the first player that wrote 0, or $\perp$ if there is no such player. Therefore $\text{IC}_\mu(\text{AND}_k) \leq \text{I}(\boldsymbol{\Pi}; \boldsymbol{X}) \leq H(\boldsymbol{\Pi}) \leq O(\log k)$.

This demonstrates that the AND_k function *cannot be compressed* to its external information cost: there is a gap of $\Omega(k/\log(k))$ between its communication and information complexity. It is not clear whether or not this gap is the largest possible, but it seems reasonable that any function can be compressed to no more than k times its external information cost, with a possible polylogarithmic dependence on the communication complexity (as in [3]).

Amortized compression. Let $T(f^n, \epsilon)$ be the task of computing f on n independent inputs with marginal error ϵ on each instance. It is not difficult to extend the compression protocol from [7] (which is designed for a different notion of information cost, *internal* information [4]) to the external information of protocols in the k-party broadcast model, obtaining the following result:

THEOREM 3. *For any function* f*, error parameter* $\epsilon \in (0, 1)$ *and distribution* μ*,*

$$\lim_{n\to\infty} \frac{\text{D}_{\mu^n}(T(f^n, \epsilon))}{n} \leq \text{IC}_\mu(f, \epsilon).$$

(For two players, external information is bounded from below by internal information, the notion used in [7], so our result does not improve on theirs for the two-party case. However, the notion used in [7] does not extend to the multiparty broadcast model for $k > 2$.)

We give a high-level overview of the compression scheme. This is a simplified version of the protocol from [7]. A similar but somewhat less explicit construction is given in [18].

Fix a protocol Π, and assume for simplicity that it has a fixed number of rounds and the players speak in an order that is fixed in advance (our result does not require these assumptions). Let $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_r$ be random messages representing the messages of the protocol, where each message $\boldsymbol{M}_j$ is sent by player i_j. The randomness is over both the input distribution and the private and public randomness of the protocol.

Using the chain rule for mutual information, the information cost of Π can be seen to accumulate over rounds as follows:

$$\text{IC}(\Pi) = \text{I}(\boldsymbol{\Pi}; \boldsymbol{X}) = \sum_{j=1}^{r} \text{I}(\boldsymbol{M}_j; \boldsymbol{X} \mid \boldsymbol{M}_{<j}),$$

in other words, the information we learn from observing $\boldsymbol{\Pi}$ is the information we learn from the first message, plus the information we learn from the second message given what we already learned from the first, and so on. Moreover, since message $\boldsymbol{M}_j$ is sent by player i_j and depends only on i_j's input and the transcript so far, $\boldsymbol{M}_j$ conveys no information about the other players' inputs given $\boldsymbol{M}_{<j}$, so in fact

$$\begin{aligned}\text{IC}(\Pi) &= \sum_{j=1}^{r} \text{I}(\boldsymbol{M}_j; \boldsymbol{X}_{i_j} \mid \boldsymbol{M}_{<j}) \\ &\overset{(1)}{=} \sum_{j=1}^{r} \mathop{\mathbb{E}}_{\boldsymbol{m}_{<j}, \boldsymbol{x}_{i_j}} \text{D}\left(\frac{\mu(\boldsymbol{M}_j \mid \boldsymbol{X}_{i_j} = \boldsymbol{x}_{i_j}, \boldsymbol{M}_{<j} = \boldsymbol{m}_{<j})}{\mu(\boldsymbol{M}_j \mid \boldsymbol{M}_{<j} = \boldsymbol{m}_{<j})}\right).\end{aligned}$$

We can interpret this as follows: in each round j, player i_j generates message $\boldsymbol{M}_j$ from a distribution η that depends on the transcript so far, $\boldsymbol{M}_{<j}$, and on its input $\boldsymbol{X}_{i_j}$. An external observer who has seen $\boldsymbol{M}_{<j}$ but does not know $\boldsymbol{X}$ can try to predict the next message $\boldsymbol{M}_j$; let ν be the distribution representing the external observer's prediction.[3] The information revealed about $\boldsymbol{X}_{i_j}$ in round j corresponds to the observer's ability to predict $\boldsymbol{M}_j$: the closer ν is to η, the less "surprised" the observer is when it sees $\boldsymbol{M}_j$, and the less information it learns. Therefore, in protocols that do not reveal a lot of information, the observer knows a prior distribution ν that is already fairly close to the true distribution η of the next message, and we can use this fact to *compress* the next message. Note that all players are able to compute the prior ν of the external observer (it depends only on $\boldsymbol{M}_{i_j}$). However, only player i_j knows the true distribution η.

Before proceeding it will be useful to review an elementary form of *rejection sampling*. Suppose we have a distribution η over some domain U, and we wish to sample from η, but we are only able to sample uniform random variables. One way to sample from η is to generate an infinite sequence of two-dimensional points $(\boldsymbol{x}_1, \boldsymbol{p}_1), (\boldsymbol{x}_2, \boldsymbol{p}_2), \ldots$ uniformly distributed in $U \times [0, 1]$, select

[2] *Compressing a protocol* Π *to some cost* T means constructing another protocol Θ, whose communication complexity is T, such that given inputs $\boldsymbol{X}, \boldsymbol{Y}$, the players can use Θ to sample from a distribution close to the distribution of Π's transcript $\boldsymbol{\Pi}(\boldsymbol{X}, \boldsymbol{Y})$. The players are not charged for writing out the transcript of Π that they sampled, only for the communication they exchange in order to agree on the sample. This is generalized to multiple players in the natural way.

[3] Formally, to sample from ν, one samples $\boldsymbol{X}_{i_j}$ from its distribution given $\boldsymbol{M}_{<j}$ and then computes $\boldsymbol{M}_j$.

the first point i that falls under the curve of η — that is, satisfies $\eta(\boldsymbol{x}_i) > \boldsymbol{p}_i$ — and output $\boldsymbol{x}_i$. It is not difficult to see that this generates the correct distribution η, since given $\boldsymbol{x}_i = x$, the probability that point i falls under the curve of η is exactly $\eta(x)$. Moreover, the probability for any given point to fall under the curve is $1/|U|$, so the expected number of points we need before finding a good point is U.

In the compressed protocol we simulate round j as follows. We let U be the set of possible next messages (the domain of $\boldsymbol{M}_j$), and we publicly sample points $(\boldsymbol{x}_1, \boldsymbol{p}_1), (\boldsymbol{x}_2, \boldsymbol{p}_2), \ldots$ uniformly distributed in $U \times [0,1]$ using public randomness. Player i_j, who knows the true distribution η of the next message, selects the first point that falls under the curve of η. The other players do not know η, but they do know some prior ν, which is also known to player i_j. If ν is close to η, then the other players already have some idea of which points player i_j might have selected. Instead of writing the message on the board as it would in the original protocol, player i_j writes some information on the board to help the other players learn exactly which point it selected, using its knowledge of both η and ν. The closer ν is to the true distribution η, the less player i_j will need to write on the board.

Let $a_\ell = (x_\ell, p_\ell)$ be the point selected by player i_j. First, player i_j writes on the board the value $\lceil i/|U| \rceil$. This requires a small number of bits, because $\boldsymbol{i}$ is a binomial random variable with expectation $|U|$. Let P be the set of points with indices i' such that $\lceil i'/|U| \rceil = \lceil i/|U| \rceil$; all players restrict their attention to points in the set P, whose size is $|U|$.

Next player i_j writes the log-ratio $s = \lceil \log(\eta(x_\ell)/\nu(x_\ell)) \rceil$.[4] This requires roughly $\log\log(\eta(x_\ell)/\nu(x_\ell))$ bits, using a variable-length encoding. All other players now scale up their prior ν by 2^s, and eliminate all points that do not fall under the scaled curve $2^s \cdot \nu$ from consideration, since we know that $p_\ell < \eta(x_\ell) \le 2^s \cdot \nu(x_\ell)$. Let $P' \subseteq P$ be the remaining set (see Figure 1).

This is not sufficient to narrow the candidate pool down to a single point; indeed, the expected number of points in P' is 2^s, because there are $|U|$ points in P and each one falls under the curve $2^s \cdot \nu$ with probability $2^s/|U|$. However, player i_j knows the set P', and it now simply writes the index of the point it selected inside the set P' (that is, if $P' = \{a_{t_1}, \ldots, a_{t_m}\}$ then player i_j writes the index c such that $t_c = \ell$). Since we expect that $|P'| \approx 2^s$, this requires roughly s bits. Now all players know which point was selected by player i_j, and we can continue on to the next round.

More formally, the sampling procedure can be summarized as follows:

LEMMA 7. *Suppose that all players are given a distribution η over some universe $\mathcal{U}$, and in addition, player i is given a distribution ν over $\mathcal{U}$. Fix $\epsilon \in (0,1)$. Then there is a protocol Θ, at the end of which player i outputs an element $\boldsymbol{X} \sim \eta$, and the other players all output the same element $\boldsymbol{Y}$, with the property that for each $x \in \mathcal{U}$, $\Pr[\boldsymbol{X} = \boldsymbol{Y} \mid \boldsymbol{X} = x] \ge 1 - \epsilon$. The expected communication of Θ is $\mathsf{D}(\nu \parallel \eta) + O(\log \mathsf{D}(\nu \parallel \eta) + \log(1/\epsilon))$.*

As mentioned above, this result is proven slightly differently in [18].

Using this lemma repeatedly, we can sample an entire transcript of Π using communication $\mathrm{IC}(\Pi) + O(\log \mathrm{IC}(\Pi)) + O(r \cdot \log(1/\epsilon))$, where r is the number of rounds in Π. (The cost is obtained by summing over the costs of the individual rounds as bounded in Lemma 7, using the log-sum inequality to move the sum inside the log in the second term.)

Now suppose we are given n independent inputs drawn from μ, and we solve $T(f^n, \delta)$ by running n independent instances of Π in parallel: first we execute the first round of Π on each of the instances, then the next round, and so on. The resulting protocol has the same number of rounds r that Π has, and since the copies are independent, its information cost is $n \cdot \mathrm{IC}(\Pi)$. (It is crucial that when we execute Π in parallel on the different instances so that the number of rounds does not increase.) We can compress the n-fold protocol using the scheme outlined above, reducing its communication complexity to

$$C := n \cdot \mathrm{IC}(\Pi) + r \cdot O(\log(n \cdot \mathrm{IC}(\Pi)) + \log(1/\epsilon)),$$

and $C/n \to \mathrm{IC}(\Pi)$ as n tends to infinity, as desired. (This informal presentation glosses over many details, including how to handle the additional error incurred during compression. They will appear in the full version of the paper.)

It is not clear whether the compression result of Theorem 3 is tight in general, but it is tight for product distributions.

THEOREM 4. *For any function f, error parameter $\epsilon \in (0,1)$ and product distribution μ,*

$$\lim_{n \to \infty} \frac{\mathrm{D}_{\mu^n}(T(f^n, \epsilon))}{n} = \mathrm{IC}_\mu(f, \epsilon).$$

PROOF SKETCH. Since $\mathrm{D}_\eta(g, \epsilon) \ge \mathrm{IC}_\eta(g, \epsilon)$ for any task g and distribution η, to prove the theorem it suffices to show that $\mathrm{IC}_{\mu^n}(T(f^n, \epsilon)) \ge n \cdot \mathrm{IC}_\mu(T(f^n, \epsilon))$ when μ is a product distribution. (The other direction is given by Theorem 3.) This is very similar to the direct sum result from [2], given in Lemma 1: indeed, it is even simpler to show than the statement in Lemma 1, because here we compare the information cost of actually solving n independent copies to the cost of solving one copy, whereas in Lemma 1 we compared the cost of computing the *disjunction* on n copies to the cost of solving a single copy. To show that $\mathrm{IC}_{\mu^n}(T(f^n, \epsilon)) \ge n \cdot \mathrm{IC}_\mu(T(f^n, \epsilon))$ for a product distribution μ we can apply the same proof used to show Lemma 1 in [2], using an "empty variable" as our auxiliary variable $\boldsymbol{D}$. □

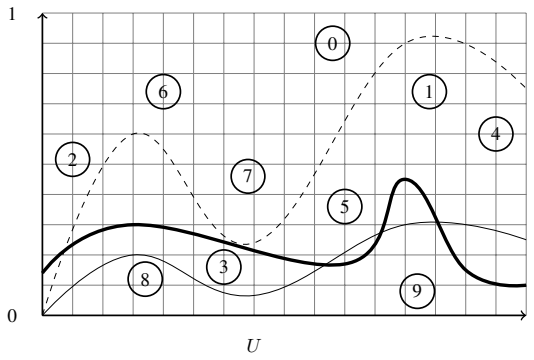

Figure 1: Illustration for the sampling procedure. The thick curve is the true distribution η, the thin curve is the prior ν, and the dashed curve is the scaled prior $2^s \cdot \nu$. In this example, player i_j will select point 3 to output, and the candidates that will be considered by the other players are $P' = \{1, 3, 4, 5, 8, 9\}$. Player i_j will send '2' to indicate that the second point in P', point 3, should be selected.

[4] It may happen that $\nu(x_\ell) < \eta(x_\ell)$ and the logarithm is negative, but this will not be a problem.

Characterizing the exact cost of one-shot compression and of compression with infinite block length in the broadcast model remain open questions.

7. REFERENCES

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

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

[3] Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to compress interactive communication. *SIAM J. Comput.*, 42(3):1327–1363, 2013.

[4] Mark Braverman. Interactive information complexity. In *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012*, pages 505–524, 2012.

[5] Mark Braverman, Faith Ellen, Rotem Oshman, Toniann Pitassi, and Vinod Vaikuntanathan. A tight bound for set disjointness in the message-passing model. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 668–677, 2013.

[6] Mark Braverman, Ankit Garg, Denis Pankratov, and Omri Weinstein. From information to exact communication. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 151–160. ACM, 2013.

[7] Mark Braverman and Anup Rao. Information equals amortized communication. In *Proceedings of the 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, FOCS '11, pages 748–757, 2011.

[8] Joshua Brody, Amit Chakrabarti, Ranganath Kondapally, David P. Woodruff, and Grigory Yaroslavtsev. Beyond set disjointness: The communication complexity of finding the intersection. In *Proceedings of the 2014 ACM Symposium on Principles of Distributed Computing*, PODC '14, pages 106–113, 2014.

[9] Amit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew Chi-Chih Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001*, pages 270–278, 2001.

[10] Ashok K. Chandra, Merrick L. Furst, and Richard J. Lipton. Multi-party protocols. In *Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing*, STOC '83, pages 94–99, 1983.

[11] Arkadev Chattopadhyay and Toniann Pitassi. The story of set disjointness. *SIGACT News*, 41(3):59–85, 2010.

[12] Atish Das Sarma, Stephan Holzer, Liah Kor, Amos Korman, Danupon Nanongkai, Gopal Pandurangan, David Peleg, and Roger Wattenhofer. Distributed verification and hardness of distributed approximation. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, pages 363–372, 2011.

[13] Shahar Dobzinski, Noam Nisan, and Sigal Oren. Economic efficiency requires interaction. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 233–242, 2014.

[14] Andrew Drucker, Fabian Kuhn, and Rotem Oshman. On the power of the congested clique model. In *ACM Symposium on Principles of Distributed Computing, PODC '14, Paris, France, July 15-18, 2014*, pages 367–376, 2014.

[15] Anat Ganor, Gillat Kol, and Ran Raz. Exponential separation of information and communication. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pages 176–185, 2014.

[16] Oded Goldreich and A Warning. Secure multi-party computation. unpublished manuscript, 1998.

[17] André Gronemeier. Asymptotically optimal lower bounds on the NIH-multi-party information complexity of the AND-function and disjointness. In *Proc. 26th Symp. on Theor. Aspects of Comp. Sc. (STACS)*, pages 505–516, 2009.

[18] Prahladh Harsha, Rahul Jain, David McAllester, and Jaikumar Radhakrishnan. The communication complexity of correlation. In *Computational Complexity, 2007. CCC'07. Twenty-Second Annual IEEE Conference on*, pages 10–23. IEEE, 2007.

[19] Johan Håstad and Avi Wigderson. The randomized communication complexity of set disjointness. *Theory of Computing*, 3(11):211–219, 2007.

[20] D.A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

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

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

[23] Mihai Patrascu. Unifying the landscape of cell-probe lower bounds. *SIAM J. Comput.*, 40(3):827–847, 2011.

[24] Jeff M. Phillips, Elad Verbin, and Qin Zhang. Lower bounds for number-in-hand multiparty communication complexity, made easy. SODA '12, pages 486–501, 2012.

[25] A. A. Razborov. On the distributional complexity of disjointness. *Theor. Comput. Sci.*, 106:385–390, 1992.

[26] Atish Das Sarma, Stephan Holzer, Liah Kor, Amos Korman, Danupon Nanongkai, Gopal Pandurangan, David Peleg, and Roger Wattenhofer. Distributed verification and hardness of distributed approximation. *SIAM J. Comput.*, 41(5):1235–1265, 2012.

[27] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[28] Alexander A. Sherstov. Communication lower bounds using directional derivatives. In *Proc. 45th Symp. on Theory of Comp. (STOC)*, pages 921–930, 2013.

[29] David P. Woodruff and Qin Zhang. An optimal lower bound for distinct elements in the message passing model. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 718–733, 2014.

[30] Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. SFCS '77, pages 222–227, 1977.