Information Theory Methods in Communication Complexity

Ziv Bar-Yossef*
Computer Science Division
University of California at Berkeley
387 Soda Hall, Berkeley, CA 94720, USA
zivi@cs.berkeley.edu

T. S. Jayram Ravi Kumar D. Sivakumar
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120, USA
{jayram,ravi,siva}@almaden.ibm.com

Abstract

We use tools and techniques from information theory to study communication complexity problems in the one-way and simultaneous communication models. Our results include:

(1) A tight characterization of multi-party one-way communication complexity for product distributions in terms of VC-dimension and shatter coefficients;

(2) An equivalence of multi-party one-way and simultaneous communication models for product distributions;

(3) A suite of lower bounds for specific functions in the simultaneous communication model, most notably an optimal lower bound for the multi-party set disjointness problem of Alon et al. *[2] and for the generalized addressing function problem of Babai* et al. *[3] for arbitrary groups.*

Methodologically, our main contribution is rendering communication complexity problems in the framework of information theory. This allows us access to the powerful calculus of information theory and the use of fundamental principles such as Fano's inequality and the Maximum Likelihood Estimate Principle.

1 Introduction

Communication complexity [24] is one of the most versatile tools in complexity theory for proving lower bounds. It has applications in circuit complexity, VLSI theory, space-bounded computation, and distributed computing—see [16] for numerous examples. Some of the basic and reasonably well-understood lower bound techniques in this area include the fooling set argument, rank and related linear-algebraic methods, and the discrepancy set argument. It is perhaps fair to say that these can be classified as "first generation" methods—fundamental ideas that transcend the applications, but which ultimately need to be built upon to bring the theory to fruition.

In this work, we apply fundamental tools of information theory to address a wide variety of communication complexity problems.

Our paper has two distinct themes that are intertwined throughout: *methods* and *models*. From the methodological point of view, we apply a collection of basic ideas from information theory—Fano's inequality and the Maximum Likelihood Estimate Principle, to name two—to obtain crisp proofs of known and new communication complexity upper and lower bounds. In parallel, we advance the theory of special models of communication complexity—especially "one message per player" models, including the one-way and the simultaneous communication models (details shortly)—and obtain fairly tight equivalences between them. The two themes (methods and models) happen to blend well, as demonstrated by the efficacy of the information theory methods in obtaining a *variety* of results concerning these models—lower bounds, upper bounds, and simulations of one model by others.

METHODS

Information Theory [9] is concerned with reliable transmission of encoded information under the presence of noise. In this paper we utilize a particularly useful view of *distributional* communication complexity protocols (deterministic protocols that get an input chosen from a given distribution and are allowed to err with some fixed probability over the input distribution)[1] as decoding procedures: the encoded information is the function to be computed by the protocol; the encoding is the transcript of messages (plus local inputs) viewed by the player who announces the output; and the decoding procedure is what the player performs to compute the output.

*Part of this work was done while the author was visiting IBM Almaden Research Center. Supported by NSF Grant CCR-9820897.

[1] Recall that, by Yao's Lemma, distributional communication complexity is equivalent to randomized communication complexity with public randomness.

Having placed the communication complexity model into an information theory setting, a rich set of tools is at our disposal. Information theory provides a powerful *calculus* which allows one to perform sophisticated manipulations of *information*. Two results that are particularly useful to us are *Fano's inequality*—which provides a lower bound on the decoding error—and the *Maximum Likelihood Estimate Principle*—which gives an upper bound on the error of the Maximum Likelihood decoder. These two results imply upper and lower bounds on the error of distributional communication complexity protocols.

We should note right away the we are not the first to apply information theory to obtain communication complexity results (see, e.g., [21, 18, 1, 3, 19, 7, 22]). The main difference, however, is that all the above results use information theory as a *tool* in the proof (alongside other combinatorial tools), while we *reduce* the communication problem into an information theory problem, and then work in the information theory domain.

MODELS

In the *one-way communication* model [1, 15], a set of (two or more) players hold parts of an input to a function. The first player initiates the communication by sending a single message to the second player; the second sends one message to the third, and so forth; the last player announces the output of the protocol. This model arises very naturally in the context of proving lower bounds for algorithms in the data stream model [13, 2, 11], time-space tradeoffs [6], and data structure complexity [17]. Indeed, specific functions for which we study lower bound questions emerge from the data stream model, where data arrives in a stream, and an algorithm is efficient if it uses very little space. Since these algorithms are allowed to be randomized and approximate, surprising upper bounds have been obtained [2, 11], and proving strong lower bounds appears to be a fairly difficult challenge.

In the *simultaneous communication* model [24] a set of players hold parts of an input to a function, and each of them sends one message simultaneously to a *referee*. The referee, who does not see any of the inputs beforehand, has to announce the output of the protocol. This model can be simulated by the one-way model, and is significantly simpler to deal with. One of our goals is to obtain lower bound results in this model and to understand its weakness with respect to the one-way model.

SUMMARY OF RESULTS

Tight characterization of one-way communication complexity. We obtain a tight characterization of the distributional one-way communication complexity for *product distributions* in terms of the VC-dimension and shatter coefficients of the function matrix. Similar results were obtained for the two-party case by Kremer, Nisan and Ron [15] and by Ablayev [1]. Our bounds are sharper and for an arbitrary number of players. In deriving the upper bound in terms of VC-dimension, we provide a crisp proof of a weaker form of a major result from learning theory due to Blumer *et al.* [5]. The difference between the two is that, in our case, the hidden concept is chosen from a (known) distribution, and the learner is not expected to output a hypothesis, but only to predict the value of the concept on a new input. This weaker result is enough for the communication complexity upper bound, and it admits a clean information theory proof (modulo the use of Sauer's lemma, also a crucial ingredient in [5]). These results are described in Section 3.

One-way versus simultaneous communication complexity. We prove that for product distributions, the simultaneous communication model is as strong as the one-way model, for an arbitrary number of players. We also show that for non-product distributions, the one-way model (even for two players) is strictly stronger than the simultaneous model. This disproves a statement by [15]. These results are described in Section 4.

Specific lower bounds. We present a collection of lower bounds for specific problems in the one-way and simultaneous communication models; these illustrate the strength and flexibility of the methods we promote in this paper. These are discussed in Section 5.

We first consider the distributional complexity of set-disjointness with respect to a particular non-product distribution that was used by [21] to prove an $\Omega(n)$ lower bound in the general communication model. To illustrate the simplicity of our methods and their applicability for non-product distributions, we present an $\Omega(n)$ lower bound (that is also fairly tight with respect to the error) for this problem in the two-party one-way communication model.

We obtain a nearly optimal lower bound for an extension [2] of the [14, 21] distribution on the set disjointness problem for k players. Alon *et al.* [2] show an $\Omega(n/k^4)$ lower bound for this problem in the general communication model. We show an optimal $\Omega(n/k)$ lower bound for the same distribution for the simultaneous communication model. A lower bound of $\Omega(n/k)$ in the one-way model would yield strong space lower bounds for data stream algorithms for frequency moments.

Finally, we visit the simultaneous communication model in the "number on the forehead" model [8]. This model has connections to circuit complexity, and one of the important results here is that of Babai *et al.* [3], for the problem of the *generalized addressing function* with k players with respect to a group G (see definition in Section 5.3). For the deterministic case, they obtain lower bounds that apply to arbitrary groups; for the distributional case, they only consider the group $G = Z_2^{k-1}$ and obtain an $\Omega(n^{1/(k-1)}/(k-1))$ lower bound[2]. Using simple arguments, we obtain distri-

[2] Satya Lokam has informed us that they are now able to extend it to all

butional lower bounds that work for arbitrary groups, and which are analogous to their deterministic lower bounds. In addition, our methods are applicable to obtain distributional lower bounds for the "number in the hand" model.

2 Preliminaries

2.1 Communication Complexity Models

All variants of Yao's communication complexity model share the same high level framework: several players are to compute a function $f : \mathcal{X} \to \mathcal{Z}$. The input is partitioned into several (possibly overlapping) parts, and the parts are distributed to the players. The players communicate between themselves, and in the end of the communication one designated player announces the function's value. The main complexity measure of concern is the amount of data communicated between the players. The concrete models of communication complexity are specified by certain features of the game described above. In this paper we concentrate on four such features:

1. **Number of Players** (will be denoted by k).

2. **Input Partition** (will be denoted by $I_1, \ldots, I_k \subseteq [n]$, where n is the input length). Player j gets the projection of the input on the coordinates in I_j. An input partition is *disjoint*, if $I_1, \ldots, I_k$ are pairwise disjoint. Two common input partitions we consider are: (1) *Number-in-the-hand (NIH)* [24]—$I_j = \{j\}$, and (2) *Number-on-the-forehead (NOF)* [8]—$I_j = [n] \setminus \{j\}$.

3. **Interaction Rules**: the set of rules that regulate the exchange of messages between the players (e.g., which player/s are allowed to send messages at any given moment, to whom a player can address a message, and which player announces the output). We consider two restricted models of interaction: (1) *one-way communication* [1, 15] (denoted $I_1 \to \cdots \to I_k$)—player j ($j = 1, \ldots, k-1$) sends exactly one message throughout the protocol, to player $j+1$, and (2) *simultaneous communication* [24] (denoted $I_1 || \cdots || I_k$): players $1, \ldots, k$ send exactly one message, to player $k+1$; player $k+1$ (called the "referee") gets no input and is supposed to announce the output.

4. **Complexity**: we consider two methods to account for the amount of communication used in a protocol: *total communication*—the sum of the lengths of messages exchanged, and *maximum communication* (denoted $\max$)—the length of the longest message.

All the complexity measures we consider are worst case ones—i.e., the complexity of a protocol is its maximum complexity over all possible inputs.

We consider two variants of communication complexity measures: (1) *randomized* (denoted $R_\delta(f)$)—the complexity of the best randomized protocol that computes f with error probability at most δ (the probability is over the random choices of the players), and (2) *μ-distributional* (denoted $D_{\mu,\delta}(f)$)—the complexity of the best deterministic protocol that computes f with error probability at most δ on inputs drawn according to the distribution μ (the probability is over μ). The *distributional* complexity of f (denoted $D_\delta(f)$), is $\max_\mu D_{\mu,\delta}(f)$, and the *rectangular distributional* complexity of f (denoted $D^{\Box}_\delta(f)$), is $\max_\mu D_{\mu,\delta}(f)$, where here the maximum is only over product (rectangular) distributions (a distribution μ on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_k$ is rectangular, if $\mu = \mu_1 \times \cdots \times \mu_k$, where each μ_j is a distribution over $\mathcal{X}_j$, and $\mu_1, \ldots, \mu_k$ are mutually independent). Yao's Lemma [25] states that $R_\delta(f) \geq D_\delta(f)$, and thus allows one to use distributional complexity lower bounds to prove randomized communication complexity lower bounds.

2.2 Tools from Information Theory

The *entropy* of a distribution P on a set Ω is $\mathrm{H}(P) = \sum_{\omega \in \Omega} P(\omega) \log_2 \frac{1}{P(\omega)}$. The entropy of a random variable is the entropy of the distribution it induces on its range. For a parameter $0 \leq p \leq 1$, $H_2(p)$ is the entropy of the Bernoulli random variable whose probability of success is p. The *conditional entropy* of random variable X given another random variable 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 random variables X and Y is denoted $\mathrm{H}(X, Y)$. Some basic properties of entropy are the following (proofs can be found in Chapter 2 of [9]):

Proposition 1 (Basic properties of entropy) *X, Y are random variables with ranges S_X, S_Y. $f : S_X \to \mathcal{Z}$ is a function.*
(1) $0 \leq \mathrm{H}(X) \leq \log|S_X|$. $\mathrm{H}(X) = \log|S_X|$ iff X is uniform on S_X.
(2) Conditioning reduces entropy: $\mathrm{H}(X \mid Y) \leq \mathrm{H}(X)$.
(3) $\mathrm{I}(X;Y) \geq 0$; equality iff X and Y are independent.
(4) Chain rule for entropy: $\mathrm{H}(X, Y) = \mathrm{H}(X) + \mathrm{H}(Y \mid X)$.
(5) Chain rule for mutual information: $\mathrm{I}(X, Y; Z) = \mathrm{I}(X;Z) + \mathrm{I}(X;Y \mid Z)$.
(6) Subadditivity: $\mathrm{H}(X, Y) \leq \mathrm{H}(X) + \mathrm{H}(Y)$; equality iff X and Y are independent.
(7) $\mathrm{H}(f(X) \mid X) = 0$.
(8) Data processing inequality: $\mathrm{I}(f(X); Y) \leq \mathrm{I}(X;Y)$.
(9) If X, Y are jointly independent of Z, then $\mathrm{H}(X \mid Y, Z) = \mathrm{H}(X \mid Y)$.

We next state the three major information theory results we are using in this paper. Fano's inequality gives a lower bound on prediction errors; the MLE Principle shows that the MLE almost achieves Fano's bound when predicting

groups.

Boolean values; and the entropy subset inequality generalizes the subadditivity of entropy.

Theorem 2 (Fano's Inequality [9]) *Let X and Y be two random variables over S_X and S_Y, and let $g : S_Y \to S_X$ be a prediction function, which given an observed value of Y guesses the value of X. Let δ be the prediction error: $\delta = \Pr(g(Y) \neq X)$, where the probability is over X and Y. Then, $H_2(\delta) + \delta \log(|S_X| - 1) \geq \mathrm{H}(X \mid Y)$. In particular, if X is Boolean (i.e., $|S_X| = 2$), then $H_2(\delta) \geq \mathrm{H}(X \mid Y)$.*

Theorem 3 (MLE Principle [10]) *Let X be a Boolean random variable we wish to predict, and let Y be an observed random variable. Then, $\Pr(\textsc{mle}(Y) \neq X) \leq \mathrm{H}(X \mid Y)$, where $\textsc{mle}(y)$ is the most likely x given $Y = y$.*

Theorem 4 (Entropy Subset Inequality [12]) *For n random variables $X_1, \ldots, X_n$ and a set $S \subseteq [n]$, define $X_S = \{X_i \mid i \in S\}$. Define for each $k = 1, \ldots, n$,*

$$H_k^{(n)} \stackrel{\text{def}}{=} \frac{1}{\binom{n}{k}} \sum_{S,|S|=k} \frac{H(X_S)}{k}.$$

Then, $H_1^{(n)} \geq \cdots \geq H_n^{(n)}$. The inequality holds also for conditional entropy.

3 One-Way Communication

In this section we study the randomized communication complexity of two-party and multi-party one-way protocols. The two-party variant was studied by [1] and [15], who give an almost complete characterization of the communication complexity for rectangular distributions (i.e., $D_\delta^{[],I_1 \to I_2}(f)$) in terms of the function matrix's VC-dimension. We present new proofs of this characterization using tools from information theory. Our characterization is in terms of *shatter coefficients*—a notion that generalizes VC-dimension; we show that this characterization sometimes yields tighter lower bounds. We then generalize the VC-dimension characterization to an arbitrary number of players. As a consequence, we prove that the two-party and the multi-party models with worst-case partition are equivalent for rectangular distributional complexity. We exhibit non-product distributions for which there exists an unbounded gap between the two models.

The *Vapnik-Chervonenkis (VC) dimension* [23] of a family of functions $\mathcal{H} = \{h : \mathcal{X} \to \{0,1\}\}$ is defined as follows. We think of the functions in $\mathcal{H}$ as binary strings of length $|\mathcal{X}|$ (the string corresponding to h is the truth table of h). A subset $S \subseteq \mathcal{X}$ is *shattered* by $\mathcal{H}$, if for every binary string $\alpha \in \{0,1\}^{|S|}$, there is a function $h \in \mathcal{H}$, such that $h|_S$ (i.e., the projection of the string h on the coordinates in S) is α. The *VC-dimension* of $\mathcal{H}$, $\textsc{vcd}(\mathcal{H})$, is the maximum size of a subset S that is shattered by $\mathcal{H}$. The *ℓ-th shatter coefficient of* $\mathcal{H}$, $\textsc{sc}(\mathcal{H}, \ell)$, is the largest number of patterns that one can obtain by considering all possible $h|_S$ where S varies over all subsets of size ℓ. Note that $\textsc{vcd}(\mathcal{H})$ is the maximum ℓ for which $\textsc{sc}(\mathcal{H}, \ell) = 2^\ell$.

Consider $f : \mathcal{X} \to \{0,1\}$ and a disjoint input partition I_1, I_2. We denote by $\mathcal{X}_1, \mathcal{X}_2$ the projections of $\mathcal{X}$ on I_1, I_2 respectively. For every possible Alice input $x_1 \in \mathcal{X}_1$, we define f_{x_1} to be the function induced by this input; that is, $f_{x_1}(x_2) = f(x_1, x_2)$. $f_{\mathcal{X}_1}$ is the family of functions $\{f_{x_1} \mid x_1 \in \mathcal{X}_1\}$. Kremer *et al.* [15] showed that for every $f : \mathcal{X} \to \{0,1\}$ and for every disjoint input partition I_1, I_2, $D_\delta^{[],I_1 \to I_2}(f) \geq \textsc{vcd}(f_{\mathcal{X}_1}) \cdot (1 - O(\delta))$. We use a simple entropy-based argument to prove a more general lower bound in terms of shatter coefficients:

Theorem 5 *For every function $f : \mathcal{X} \to \{0,1\}$, every $\ell \geq \textsc{vcd}(f_{\mathcal{X}_1})$, every disjoint input partition I_1, I_2, and every $\delta > 0$,*

$$D_\delta^{[],I_1 \to I_2}(f) \geq \log(\textsc{sc}(f_{\mathcal{X}_1}, \ell)) - \ell \cdot H_2(\delta)$$

Proof. Let M_f be the function matrix such that $M_f(x_1, x_2) = f(x_1, x_2)$. Let $\mathcal{X}_1' \subseteq \mathcal{X}_1$ and $\mathcal{X}_2' \subseteq \mathcal{X}_2$, where $|\mathcal{X}_2'| = \ell$ and $|\mathcal{X}_1'| = \textsc{sc}(f_{\mathcal{X}_1}, \ell)$, be such that M_f restricted to $\mathcal{X}_1' \times \mathcal{X}_2'$ consists of $\textsc{sc}(f_{\mathcal{X}_1}, \ell)$ distinct rows. In other words, $|\{f_{x_1}|_{\mathcal{X}_2'} \mid x_1 \in \mathcal{X}_1'\}| = \textsc{sc}(f_{\mathcal{X}_1}, \ell)$.

Consider the product distribution $\mu = \mu_1 \times \mu_2$, for which μ_1 is uniform on $\mathcal{X}_1'$ and μ_2 is uniform on $\mathcal{X}_2'$. Let X_1 and X_2 denote random elements of $\mathcal{X}_1'$ and $\mathcal{X}_2'$ chosen according to μ_1 and μ_2 respectively. Consider a one-way protocol for f with complexity C and error at most δ in which $A = A(X_1)$ is the message sent by Alice to Bob. Bob estimates the output of the function given $A(X_1)$ and X_2. Therefore, by Fano's inequality (Theorem 2), $H_2(\delta) \geq \mathrm{H}(f(X_1, X_2) \mid A(X_1), X_2) = \sum_{x_2 \in \mathcal{X}_2'} \Pr(X_2 = x_2) \cdot \mathrm{H}(f(X_1, X_2) \mid A(X_1), X_2 = x_2) = \frac{1}{\ell} \sum_{x_2 \in \mathcal{X}_2'} \mathrm{H}(f(X_1, x_2) \mid A(X_1), X_2 = x_2)$.

Note that both $f(X_1, x_2)$ and $A(X_1)$ are random variables that depend only on X_1 since x_2 is a fixed element of $\mathcal{X}_2'$. Since X_1 and X_2 are independent, we can apply Proposition 1(9) to infer that $\mathrm{H}(f(X_1, x_2) \mid A(X_1), X_2 = x_2) = \mathrm{H}(f(X_1, x_2) \mid A(X_1))$. Using this fact in the equation above we obtain: $\ell \cdot H_2(\delta) \geq \sum_{x_2 \in \mathcal{X}_2'} \mathrm{H}(f(X_1, x_2) \mid A(X_1)) \geq \mathrm{H}(\{f(X_1, x_2)\}_{x_2 \in \mathcal{X}_2'} \mid A(X_1)) \geq \mathrm{H}(\{f(X_1, x_2)\}_{x_2 \in \mathcal{X}_2'}) - \mathrm{H}(A(X_1))$, where we applied the entropy inequalities Proposition 1, part (6) and part (4).

As X_1 ranges over the values in $\mathcal{X}_1'$ the (ordered) set $\{f(X_1, x_2)\}_{x_2 \in \mathcal{X}_2'}$ takes on $\textsc{sc}(f_{\mathcal{X}_1}, \ell)$ distinct values, each with equal probability. Therefore, $\mathrm{H}(\{f(X_1, x_2)\}_{x_2 \in \mathcal{X}_2'}) = \log \textsc{sc}(f_{\mathcal{X}_1}, \ell)$. Moreover, $\mathrm{H}(A(X_1)) \leq C$. Thus, $\ell \cdot H_2(\delta) \geq \log \textsc{sc}(f_{\mathcal{X}_1}, \ell) - C$ which yields the lower bound. $\square$

Note that by taking $\ell = \textsc{vcd}(f_{\mathcal{X}_1})$ we get exactly the lower bound of [15] as a special case.

We next show that for a restricted version of the addressing function $f : \mathcal{X}_1 \times \mathcal{X}_2 \to \{0,1\}, \mathcal{X}_1 = \{w \in \{0,1\}^n : |w| < 2\delta n\}, \mathcal{X}_2 = \{0,1\}^{\log n}, f(x,i) = x_i$, where $|w|$ is the Hamming weight of w, our bound yields a tighter dependence on δ. $\textsc{vcd}(f_{\mathcal{X}_1}) < 2\delta n$, because the pattern $1^{2\delta n}$ cannot be obtained by $f_{x_1}|_S$ for any $x_1 \in \mathcal{X}_1$ and any $S \subseteq [n]$. Therefore, the [15] lower bound is at most $2\delta n(1 - O(\delta))$. On the other hand, consider the n-th shatter coefficient of $f_{\mathcal{X}_1}$. Since all the rows of M_f are distinct, $\textsc{sc}(f_{\mathcal{X}_1}, n) = \sum_{i=0}^{2\delta n} \binom{n}{i} \geq 2^{nH_2(2\delta)}$. Therefore, our lower bound is at least $nH_2(2\delta) - nH_2(\delta) = n(H_2(2\delta) - H_2(\delta)) = \Omega(n\delta\log(1/\delta))$ for $\delta < 1/2$.

[15] also proved an upper bound on the one-way rectangular communication complexity in terms of VC-dimension, exploiting the famous result of Blumer *et al.* [5] connecting PAC learning and VC-dimension. We give a self-contained proof of this bound using entropy-based arguments and an application of Sauer's Lemma. Our result can be viewed as a "prediction version" of the classical learning theory result.

Theorem 6 ([15]) *For every function $f : \mathcal{X} \to \{0,1\}$, every disjoint input partition I_1, I_2 with $\textsc{vcd}(f_{\mathcal{X}_1}) \leq |I_2|/2$, every $\delta > 0$, and for every product distribution $\mu = \mu_1 \times \mu_2$ on $\mathcal{X}$,*

$$D^{I_1 \to I_2}_{\mu,\delta}(f) \;\leq\; \textsc{vcd}(f_{\mathcal{X}_1}) \cdot O((1/\delta)\log(1/\delta))$$

Proof. Let $2 \cdot \textsc{vcd}(f_{\mathcal{X}_1}) \leq m \leq |I_2|$ be a parameter to be fixed later. We will first describe a collection of deterministic protocols, each corresponding to a subset $\mathcal{X}'_2 \subseteq \mathcal{X}_2$ of size m. We will then argue that one of these protocols computes f on μ with error at most δ.

For each multiset $\mathcal{X}'_2$ consisting of m elements from $\mathcal{X}_2$, we define the following protocol. Upon receiving her input x_1, Alice sends to Bob $A(x_1, \mathcal{X}'_2) = \{f(x_1, x'_2)\}_{x'_2 \in \mathcal{X}'_2}$. (If $\mathcal{X}'_2$ contain duplicates, this protocol sends redundant information, but we ignore this fact in order to make the analysis simpler.) Bob finds the collection $\mathcal{X}'_1$ of all $x'_1 \in \mathcal{X}_1$ that are consistent with Alice's message (that is, $f(x'_1, x'_2) = f(x_1, x'_2)$ for all $x'_2 \in \mathcal{X}'_2$), and outputs the majority value (w.r.t. μ_1) of $\{f(x'_1, x_2)\}_{x'_1 \in \mathcal{X}'_1}$. Note that the communication of the protocol is exactly m.

Let X_1, X_2 be random inputs chosen according to μ_1, μ_2 respectively. Note that what Bob computes is the MLE of the random variable $f(X_1, X_2)$ given the observations $A(X_1, \mathcal{X}'_2)$ and X_2. Therefore, by the MLE Principle (Theorem 3), $\Pr(\textsc{mle}(A(X_1, \mathcal{X}'_2), X_2, \mathcal{X}'_2) \neq f(X_1, X_2)) \leq \mathrm{H}(f(X_1, X_2) \mid A(X_1, \mathcal{X}'_2), X_2)$.

Consider now choosing the set $\mathcal{X}'_2$ by picking m elements from $\mathcal{X}_2$ independently according to μ_2. Let S be the corresponding random variable. We thus obtain, $\Pr(\textsc{mle}(A(X_1, S), X_2, S) \neq f(X_1, X_2)) \leq \mathrm{H}(f(X_1, X_2) \mid A(X_1, S), X_2, S)$.

Using Proposition 1, part (4), we have:

$$\begin{aligned} &\mathrm{H}(f(X_1, X_2) \mid A(X_1, S), X_2, S) = \\ &\quad \mathrm{H}(f(X_1, X_2), A(X_1, S) \mid X_2, S) - \mathrm{H}(A(X_1, S) \mid X_2, S) \end{aligned} \tag{1}$$

By the definition of A and S, $A(X_1, S) = \{f(X_1, X_2^j)\}_{j=1}^m$, where $X_2^1, \ldots, X_2^m$ are chosen according to μ_2. Therefore, the distribution of $f(X_1, X_2)$ is identical to the distribution of $f(X_1, X_2^j)$, $j = 1, \ldots, m$. We can thus rewrite the joint distribution $(f(X_1, X_2), A(X_1, S))$ as $\{f(X_1, X_2^j)\}_{j=0}^m$, where $X_2^0 = X_2$. Using the entropy subset inequality (Theorem 4), we now obtain:

$$\begin{aligned} &\mathrm{H}(f(X_1, X_2), A(X_1, S) \mid X_2, S) \\ &\quad = \mathrm{H}(\{f(X_1, X_2^j)\}_{j=0}^m \mid X_2, S) \\ &\quad \leq \frac{m+1}{m+1} \cdot \sum_{T \subseteq [m+1], |T| = m} \frac{1}{m} \cdot \mathrm{H}(\{f(X_1, X_2^j)\}_{j \in T} \mid X_2, S) \\ &\quad = \frac{1}{m} \cdot (m+1) \cdot \mathrm{H}(\{f(X_1, X_2^j)\}_{j=1}^m \mid X_2, S) \\ &\quad = \left(1 + \frac{1}{m}\right) \cdot \mathrm{H}(A(X_1, S) \mid X_2, S) \end{aligned}$$

Combining with Equation 1, we obtain:

$$\Pr(\textsc{mle}(A(X_1, S), X_2, S) \neq f(X_1, X_2)) \leq \frac{1}{m}\mathrm{H}(A(X_1, S) \mid X_2, S)$$

Since conditioning reduces entropy (Proposition 1, part (2)), then:

$$\begin{aligned} \frac{1}{m}\mathrm{H}(A(X_1, S) \mid X_2, S) &\leq \frac{1}{m}\mathrm{H}(A(X_1, S) \mid S) \\ &= \frac{1}{m}\sum_{\mathcal{X}'_2} \mathrm{H}(A(X_1, \mathcal{X}'_2)) \Pr(S = \mathcal{X}'_2) \end{aligned} \tag{2}$$

Consider any choice $\mathcal{X}'_2$ for S. Since the entropy is always at most the logarithm of the size of the range (Proposition 1, part (1)), then:

$$\begin{aligned} \frac{1}{m}\mathrm{H}(A(X_1, \mathcal{X}'_2)) &= \frac{1}{m}\mathrm{H}(\{f(X_1, x'_2)\}_{x'_2 \in \mathcal{X}'_2}) = \frac{1}{m}\mathrm{H}(f_{X_1}|_{\mathcal{X}'_2}) \\ &\leq \frac{1}{m}\log|f_{\mathcal{X}_1}|_{\mathcal{X}'_2}| \end{aligned}$$

where $f_{\mathcal{X}_1}|_{\mathcal{X}'_2}$ is the family of functions f_{x_1}, $x_1 \in \mathcal{X}_1$, restricted to the columns in $\mathcal{X}'_2$. We now use Sauer's Lemma to bound the size of this family:

Lemma 7 (Sauer) *For any family $\mathcal{F} \subseteq \{0,1\}^n$, $|\mathcal{F}| \leq \sum_{i=0}^d \binom{n}{i} \leq (en/d)^d$, where $d = \textsc{vcd}(\mathcal{F})$.*

We thus obtain:

$$\frac{1}{m} \cdot \mathrm{H}(A(X_1, \mathcal{X}'_2)) \leq \frac{\textsc{vcd}(f_{\mathcal{X}_1}|_{\mathcal{X}'_2})}{m} \cdot \log\frac{em}{\textsc{vcd}(f_{\mathcal{X}_1}|_{\mathcal{X}'_2})}.$$

Since the function $x \log(1/x)$ is monotone increasing for $x \leq 1/2$, $\mathrm{VCD}(f_{\mathcal{X}_1}|_{\mathcal{X}_2'}) \leq \mathrm{VCD}(f_{\mathcal{X}_1})$, and $\mathrm{VCD}(f_{\mathcal{X}_1})/m \leq 1/2$, then

$$\frac{1}{m}\mathrm{H}(A(X_1, \mathcal{X}_2')) \leq \frac{\mathrm{VCD}(f_{\mathcal{X}_1})}{m} \cdot \log \frac{em}{\mathrm{VCD}(f_{\mathcal{X}_1})}$$

By choosing $m = O(\mathrm{VCD}(f_{\mathcal{X}_1})/\delta \cdot \log(1/\delta))$, we have $(1/m)\mathrm{H}(A(X_1, \mathcal{X}_2')) \leq \delta$. This holds for any choice of $\mathcal{X}_2'$, and therefore, substituting in Equation 2,

$$\Pr(\mathrm{MLE}(A(X_1, S), X_2, S) \neq f(X_1, X_2)) \leq \delta$$

In particular, there is a choice $\mathcal{X}_2'$ for S, for which

$$\Pr(\mathrm{MLE}(A(X_1, \mathcal{X}_2'), X_2, \mathcal{X}_2') \neq f(X_1, X_2)) \leq \delta$$

The protocol $(A(\cdot, \mathcal{X}_2'), \mathrm{MLE}(\cdot, \cdot, \mathcal{X}_2'))$ is the desired deterministic protocol for f. □

We next generalize the VC-dimension characterization of the two-party model to an arbitrary number of players. We state the results in terms of the maximum communication. Similar statements hold for the total communication, because the two measures differ by a factor of at most k in the one-way model.

Let $f : \mathcal{X} \to \{0, 1\}$ be any function. We associate with every disjoint input partition $I_1, \ldots, I_k$ for k players $k-1$ disjoint input partitions for two players: for $j = 1, \ldots, k-1$, $J_1^j = \bigcup_{i=1}^{j} I_i$ and $J_2^j = \bigcup_{i=j+1}^{k} I_i$. For each j, we define $\mathcal{X}_j, \mathcal{X}_1^j$, and $\mathcal{X}_2^j$ to be the projections of $\mathcal{X}$ on the coordinates in I_j, J_1^j, and J_2^j respectively. We define for every $x_1^j \in \mathcal{X}_1^j$ the function $f_{x_1^j}(x_2^j) = f(x_1^j, x_2^j)$. $f_{\mathcal{X}_1^j}$ is the family of functions $\{f_{x_1^j} \mid x_1^j \in \mathcal{X}_1^j\}$.

We first note that the multi-party one-way model is not stronger than the two-party model:

Proposition 8 *For every function $f : \mathcal{X} \to \mathcal{Z}$, every disjoint input partition $I_1, \ldots, I_k$, and every $\delta > 0$,*

$$R_\delta^{\max, I_1 \to \cdots \to I_k}(f) \;\geq\; \max_j R_\delta^{J_1^j \to J_2^j}(f)$$

The proof is straightforward; the two players can simulate the k-player protocol as follows: Alice simulates the first j players; it then transmits what the j-th player would have sent to the $j+1$-st player to Bob; Bob completes the computation by simulating the last $k - j$ players.

The result also holds for the μ-distributional complexity for any input distribution μ. It therefore follows from Theorem 5 that:

Corollary 9 *For every function $f : \mathcal{X} \to \{0, 1\}$, every disjoint input partition $I_1, \ldots, I_k$, every $\ell \geq \max_j \mathrm{VCD}(f_{\mathcal{X}_1^j})$, and every $\delta > 0$,*

$$D_\delta^{[], \max, I_1 \to \cdots \to I_k}(f) \geq \max_j (\log(\mathrm{SC}(f_{\mathcal{X}_1^j}, \ell)) - \ell \cdot H_2(\delta))$$

We next prove that this bound is tight modulo the dependence on δ and k:

Theorem 10 *For every function $f : \mathcal{X} \to \{0, 1\}$, every disjoint input partition $I_1, \ldots, I_k$, every product distribution μ, and every $\delta > 0$,*

$$D_{\mu,\delta}^{\max, I_1 \to \cdots \to I_k}(f) \;\leq\; \max_j \mathrm{VCD}(f_{\mathcal{X}_1^j}) \cdot O(\frac{k}{\delta} \cdot \log \frac{k}{\delta})$$

Proof. We will construct a protocol Q for k players that uses $k-1$ copies $P_1, \ldots, P_{k-1}$ of the protocol for two players by Kremer *et al.* [15]. Their protocol (which is slightly different from the protocol we described in the proof of Theorem 6) uses a fixed set $\mathcal{X}_2' \subseteq \mathcal{X}_2$. Given her input $x_1 \in \mathcal{X}_1$, Alice sends to Bob $f_{x_1}(x_2')$ for all $x_2' \in \mathcal{X}_2'$. Bob finds an arbitrary input x_1' that is consistent with Alice's message (i.e., $f_{x_1'}(x_2') = f_{x_1}(x_2')$ for all $x_2' \in \mathcal{X}_2'$) and outputs $f_{x_1'}(x_2)$. Using the connection between PAC learning and VC-dimension [5], Kremer *et al.* prove that for any product distribution μ, there exists such a set $\mathcal{X}_2'$ of size $\mathrm{VCD}(f_{\mathcal{X}_1}) \cdot O((1/\delta) \cdot \log(1/\delta))$, such that the protocol errs with probability at most δ on μ.

P_j will not be a protocol for computing the function f but rather a protocol for a function $f^j : \mathcal{X} \to \{0, 1\}$ to be defined shortly. P_j computes f^j with the input partition J_1^j, J_2^j; that is, in P_j Alice receives an input $x_1^j \in \mathcal{X}_1^j$ and Bob receives an input $x_2^j \in \mathcal{X}_2^j$. Let us denote by $A_j(\cdot)$ the function Alice computes on her input to produce the message to Bob, and by $B_j(\cdot)$ the function Bob applies on the message from Alice to find a consistent input x_1'. P_j will err with probability at most $\delta/(k-1)$ on inputs drawn from μ, meaning that $\Pr_\mu(f^j(B_j(A_j(x_1^j)), x_2^j) \neq f^j(x_1^j, x_2^j)) \leq \delta/(k-1)$.

$f^1, \ldots, f^{k-1}$ are defined recursively: $f^1 = f$, and $f^j(x_1^j, x_2^j) = f^{j-1}(B_{j-1}(A_{j-1}(x_1^{j-1})), x_2^{j-1})$ for $j = 2, \ldots, k-1$. That is, f^j is the output of P_{j-1} when trying to compute f^{j-1} (note that since P_{j-1} is allowed to err, f^j is not necessarily identical to f^{j-1}). Denote by f^k the output of the last protocol P_{k-1}.

The protocol Q works as follows: the first player simulates Alice's part in P_1, and sends $m_1 = A_1(x_1)$ to the second player. The j-th player ($j = 2, \ldots, k-1$), upon receiving a message m_{j-1} from player $j-1$ and his own input x_j, works in two phases: first he simulates Bob's part in P_{j-1}, and computes the function $B_{j-1}(m_{j-1})$; he then simulates Alice's part in P_j, and computes $m_j = A_j(B_{j-1}(m_{j-1}) \circ x_j)$; he sends m_j to player $j+1$. The last player, upon receiving the message m_{k-1} and his input part x_k, simulates Bob's part in P_{k-1}, and outputs $f^{k-1}(B_{k-1}(m_{k-1}), x_k)$.

The protocol Q induces a mapping from any input $x \in \mathcal{X}$ to a sequence of $k-1$ inputs $x^1, \ldots, x^{k-1} \in \mathcal{X}$. x^j is the input on which Q runs the simulation of P_j; in this simulation Alice receives x_1^j (the projection of x^j on J_1^j) and

Bob receives x_2^j (the projection of x^j on J_2^j). $x^1,\ldots,x^{k-1}$ are defined recursively: $x^1 = x$ and for $j = 2,\ldots,k-1$, $x_1^j = B_{j-1}(A_{j-1}(x_1^{j-1})) \circ x_j, x_2^j = x|_{J_2^j}$. We also define $x_1^k = B_{k-1}(A_{k-1}(x_1^{k-1})) \circ x_k$ and $x_2^k = \emptyset$. What Q outputs is $f^k(x^k)$.

We first prove that Q errs with probability at most δ on inputs x drawn according to μ.

$$
\begin{aligned}
&\Pr_\mu(Q(x) \neq f(x)) = \\
&= \Pr_\mu(f^k(x^k) \neq f(x)) = \Pr_\mu(f^k(x_1^k, x_2^k) \neq f^1(x_1^1, x_2^1)) \\
&\leq \sum_{j=1}^{k-1} \Pr_\mu(f^{j+1}(x_1^{j+1}, x_2^{j+1}) \neq f^j(x_1^j, x_2^j))
\end{aligned}
$$

For each $j = 1,\ldots,k-1$,

$$
\begin{aligned}
&f^{j+1}(x_1^{j+1}, x_2^{j+1}) = \\
&= f^{j+1}(B_j(A_j(x_1^j)) \circ x_{j+1}, x_2^{j+1}) \\
&= f^j(B_j(A_j(B_j(A_j(x_1^j)))), x_2^j) = f^j(B_j(A_j(x_1^j)), x_2^j)
\end{aligned}
$$

Now, since P_j computes f^j with error at most $\delta/(k-1)$ on μ, then $\Pr_\mu(f^j(B_j(A_j(x_1^j)), x_2^j) \neq f^j(x_1^j, x_2^j)) \leq \frac{\delta}{k-1}$. Thus, $\Pr_\mu(Q(x) \neq f(x)) \leq \sum_{j=1}^{k-1} \frac{\delta}{k-1} = \delta$.

We next bound the complexity of the protocol Q. The crucial observation is that the family $f^j_{\mathcal{X}_1^{j-1}}$ is a subset of the family $f^{j-1}_{\mathcal{X}_1^{j-1}}$, because for every $x_1^{j-1} \in \mathcal{X}_1^{j-1}$, $f^j_{x_1^{j-1}} = f^{j-1}_{B_{j-1}(A_{j-1}(x_1^{j-1}))} \in f^{j-1}_{\mathcal{X}_1^{j-1}}$. Furthermore, since $\mathcal{X}$ is a Cartesian product $\mathcal{X}_1 \times \cdots \times \mathcal{X}_k$, then also $f^j_{\mathcal{X}_1^i} \subseteq f^{j-1}_{\mathcal{X}_1^i}$ for all $i \geq j-1$. It follows by induction that $f^j_{\mathcal{X}_1^j} \subseteq f^1_{\mathcal{X}_1^j} = f_{\mathcal{X}_1^j}$ for all j, deriving $\text{VCD}(f^j_{\mathcal{X}_1^j}) \leq \text{VCD}(f_{\mathcal{X}_1^j})$. It follows now from Theorem 6, that the complexity of P_j is at most $\text{VCD}(f_{\mathcal{X}_1^j}) \cdot O((k/\delta)\log(k/\delta))$. Therefore, maximum communication of Q is at most $\max_j \text{VCD}(f_{\mathcal{X}_1^j}) \cdot O((k/\delta)\log(k/\delta))$. □

Combining Theorem 10 and Theorem 5 we can prove that the connection between two-party and multi-party one-way protocols illustrated in Proposition 8 is tight for rectangular distributional complexity modulo the dependence on the error δ and on the number of players k:

Corollary 11 *For every function $f : \mathcal{X} \to \{0,1\}$, every disjoint input partition $I_1,\ldots,I_k$, and every $\delta > 0$,*

$$
\begin{aligned}
&D_\delta^{[],\max,I_1\to\cdots\to I_k}(f) \leq \\
&\qquad \max_j D_\delta^{[],J_1^j\to J_2^j}(f) \cdot O(\frac{k}{\delta} \cdot \log\frac{k}{\delta} \cdot \frac{1}{1-H_2(\delta)})
\end{aligned}
$$

We now show that for non-product distributions there can be an unbounded gap between the two-party and the multi-party models. The function we consider is the following multi-party version of the disjointness function (cf. [2]): $\text{DISJ} : (\{0,1\}^n)^k \to \{0,1\}$ where $n = k(2m-1)+1$ for some parameter m. Each player gets a binary string in $\{0,1\}^n$ which we associate with a subset $S \subseteq [n]$. μ is a distribution on $(\{0,1\}^n)^k$ which is uniform on all the collections of sets $S_1,\ldots,S_k$ that satisfy the following:

(1) each S_i is of size m

(2) $S_1,\ldots,S_k$ are either uniquely intersecting (i.e., all the sets share exactly one element and are otherwise pairwise disjoint) or pairwise disjoint.

Alon, Matias, and Szegedy [2] proved a lower bound for the general communication complexity of this problem: $D_{\mu,\delta}^{I_1,\ldots,I_k}(f) \geq \Omega(n/k^4)$ for any small enough constant δ. In particular, the one-way communication complexity of the function is also $\Omega(n/k^4)$. The two-party communication complexity, however, is $O(1)$ for $k \geq 3$, because in every disjoint input partition (J_1^j, J_2^j), one of the players gets at least two sets and can thus immediately determine whether they are disjoint or intersecting.

4 Simultaneous Communication vs. One-Way Communication

In this section we compare the simultaneous and one-way communication models. It is well-known [15] that the simultaneous model is not stronger than the one-way model: every simultaneous protocol can be simulated by a one-way protocol with the same complexity. Kremer *et al.* [15] claimed to prove that the two models are in fact equivalent for two players with respect to the general distributional complexity . We disprove this statement by showing that a version of the Generalized Addressing Function (GAF) exhibits an exponential separation between the two models. We then prove that the two models are equivalent with respect to the *rectangular* distributional complexity for an arbitrary number of players.

We first note the following generalization of a result of [15] showing that one-way communication is not weaker than simultaneous communication to an arbitrary number of players. The proof works by a simple simulation (we use the notation $I_{-j} = [n] \setminus I_j$):

Proposition 12 *For every function f and for every disjoint input partition $I_1,\ldots,I_k$, $D_\delta^{I_1||\cdots||I_k}(f) \geq \sum_{j=1}^k D_\delta^{I_j\to I_{-j}}(f)$.*

Kremer *et al.* claimed to prove that the other direction is also true for two players: that is, for every Boolean function f and for every input partition I_1, I_2, $D_\delta^{I_1||I_2}(f) \leq O(D_\delta^{I_1\to I_2}(f) + D_\delta^{I_2\to I_1}(f))$. We now present a counterexample to this statement. Consider the following modification of the Generalized Addressing Function: G :

$\{0,1\}^{2n+2\log n} \to \{0,1\}$, $G(f,x,g,y) = f(x \oplus y)$ if $f = g$ and 0 otherwise (f and g are Boolean functions on $\log n$ bits and x, y are inputs to these functions). Consider the input partition $I_1 = \{f, x\}$ and $I_2 = \{g, y\}$. We prove the following separation between one-way protocols and simultaneous protocols for G:

Proposition 13 *For every small enough constant $0 < \delta < 1$, (1) $R_\delta^{I_1 \to I_2}(G) = R_\delta^{I_2 \to I_1}(G) = O(\log n)$ and (2) $D_\delta^{I_1 || I_2}(G) \geq \Omega(\sqrt{n})$.*

Proof. The upper bound is easy: Alice and Bob use the $O(\log n)$ communication protocol for equality to check whether $f = g$, and Alice also sends x to Bob. Bob outputs 1 iff $f = g$ and $g(x \oplus y) = 1$.

The lower bound relies on the following simultaneous communication complexity lower bound [4, 20] (see also Section 5.3) for the Generalized Addressing Function (Alice gets f, x, Bob gets f, y, and the the referee outputs $f(x \oplus y)$): $D_{\mu',\delta}(\text{GAF}) \geq \Omega(\sqrt{n})$, where μ' is the uniform distribution on (f, x, y), and δ is a small enough constant. Consider now the (non-product) distribution μ on inputs of G that is uniform on inputs (f, x, g, y) for which $f = g$. Observe that any δ-error simultaneous protocol for G on μ is in fact also a δ-error simultaneous protocol for GAF on μ'. The lower bound follows. □

We next prove that Proposition 12 is tight with respect to rectangular distributions:

Theorem 14 *For every Boolean function $f : \mathcal{X} \to \{0,1\}$ and for every disjoint input partition $I_1, \ldots, I_k$, $D_\delta^{[],I_1||\cdots||I_k}(f) \leq \sum_{j=1}^k D_{\delta_j}^{[],I_j \to I_{-j}}(f)$, where $\delta \geq \sum_{j=1}^k H_2(\delta_j)$.*

Proof. Let $\mu = \mu_1 \times \cdots \times \mu_k$ be any product distribution on $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$. We will show that $D_{\mu,\delta}^{I_1||\cdots||I_k}(f) \leq \sum_{j=1}^k D_{\mu,\delta_j}^{I_j \to I_{-j}}(f)$. This would imply the desired upper bound.

For each $j = 1, \ldots, k$, let P_j be any δ_j-error one-way protocol for f on the distribution μ with input partition $I_j \to I_{-j}$. We denote by A_j the function Alice applies on her input in P_j to compute the message to be sent to Bob. Let X_j be a random variable whose range is $\mathcal{X}_j$ and that distributes according to μ_j. X_j represents the input to the j-th player in the simultaneous protocol. Since μ is a product distribution, $X_1, \ldots, X_k$ are mutually independent.

The simultaneous protocol works as follows: player j, given her input $x_j \in \mathcal{X}_j$, computes $A_j(x_j)$ and sends it to the referee. The referee computes the MLE of the random variable $f(X_1, \ldots, X_k)$ given the observations $A_j(X_j) = A_j(x_j), j = 1, \ldots, k$. Therefore, by Theorem 3, the error probability of our protocol is:

$\Pr(\text{MLE}(A_1(X_1), \ldots, A_k(X_k)) \neq f(X_1, \ldots, X_k)) \leq \mathrm{H}(f(X_1, \ldots, X_k) \mid A_1(X_1), \ldots, A_k(X_k))$

Define for each j, $X^{-j} \stackrel{\text{def}}{=} (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k)$. X^{-j} is the random variable representing Bob's input in the protocol P_j. In this protocol Bob predicts $f(X_1, \ldots, X_k)$ based on the message $A_j(X_j)$ he receives from Alice and based on his own input X^{-j}. Therefore, by Fano's inequality (Theorem 2), $H_2(\delta_j) \geq \mathrm{H}(f(X_1, \ldots, X_k) \mid A_j(X_j), X^{-j})$, and the theorem now follows from the following entropy inequality:

Claim 15 $\mathrm{H}(f(X_1, \ldots, X_k) \mid A_1(X_1), \ldots, A_k(X_k)) \leq \sum_{j=1}^k \mathrm{H}(f(X_1, \ldots, X_k) \mid A_j(X_j), X^{-j}))$

Proof. For simplicity of notation, define $F = f(X_1, \ldots, X_k)$, $A_j = A_j(X_j)$, $A = (A_1, \ldots, A_k)$, and $A^{-j} = A \setminus \{A_j\}$. Our goal is thus to prove: $\mathrm{H}(F \mid A) \leq \sum_{j=1}^k \mathrm{H}(F \mid X^{-j}, A_j)$.

We first note that $\mathrm{H}(F \mid X^{-j}, A_j) = \mathrm{H}(F \mid X^{-j}, A^{-j}, A_j) = \mathrm{H}(F \mid X^{-j}, A)$, because A^{-j} is a function of X^{-j}. Therefore, we can rewrite the right-hand-side of what we need to prove as: $\sum_{j=1}^k \mathrm{H}(F \mid X^{-j}, A)$.

We apply the entropy subset inequality (Theorem 4) to get: $(k-1)\mathrm{H}(X \mid F, A) \leq \sum_{j=1}^k \mathrm{H}(X^{-j} \mid F, A)$.

The mutual independence of X_j's implies that $g(X_j) \stackrel{\text{def}}{=} (X_j, A_j)$ are mutually independent as well. It follows that $\mathrm{H}(X \mid A) = \sum_{i=1}^k \mathrm{H}(X_i \mid A_i)$ and $\mathrm{H}(X^{-j} \mid A) = \sum_{i \neq j} \mathrm{H}(X_i \mid A_i)$. Therefore, $(k-1)\mathrm{H}(X \mid A) = \sum_{j=1}^k \mathrm{H}(X^{-j} \mid A)$.

Combined with the above, we finish the proof: $\sum_{j \in [k]} [\mathrm{H}(X^{-j} \mid A) - \mathrm{H}(X^{-j} \mid F, A)] \leq (k-1)[\mathrm{H}(X \mid A) - \mathrm{H}(X \mid F, A)]$. Therefore, $\sum_{j \in [k]} \mathrm{I}(F; X^{-j} \mid A) \leq (k-1)\mathrm{I}(F; X \mid A)$. It follows, $\sum_{j \in [k]} [\mathrm{H}(F \mid A) - \mathrm{H}(F \mid X^{-j}, A)] \leq (k-1)[\mathrm{H}(F \mid A) - \mathrm{H}(F \mid A, X)]$. Hence, $\mathrm{H}(F \mid A) \leq \sum_{j \in [k]} \mathrm{H}(F \mid X^{-j}, A)$, because $\mathrm{H}(F \mid A, X) = 0$. □

□

5 Specific Lower Bounds

In this section, we present a collection of lower bounds for specific problems in various communication models. In Section 5.1 we give a simple proof of the two-party lower bound for set-disjointness of [14, 21] for the one-way communication model. This is primarily to demonstrate the simplicity of our methods. In Section 5.2 we prove a new optimal lower bound for the generalized multi-party set-disjointness problem in the simultaneous communication model. Finally, in Section 5.3, we prove lower bounds for the multi-party generalized addressing function [3] for arbitrary groups in the simultaneous communication model, and consider both the NIH and the NOF variants of the problem.

5.1 One-way Two-party Set-Disjointness

In the two-party set-disjointness problem, Alice and Bob are given two sets $X_1, X_2 \subseteq [n]$ and they are to decide whether X_1, X_2 are disjoint. We prove a distributional communication complexity lower bound with respect to a specific distribution μ. The distribution μ, introduced by [14, 21], is defined as follows: Let $n = 4m + 1$. With probability 3/4, μ chooses a uniform pair (X_1, X_2) of disjoint sets of size m, and with probability 1/4, μ chooses a uniform pair (X_1, X_2) of uniquely intersecting sets (that is, $|X_1 \cap X_2| = 1$) of size m.

Theorem 16 *For every $\delta > 0$,*

$$D_{\mu,\delta}^{I_1 \to I_2}(\textsc{disj}) \geq (1 - 2H_2(\delta)) \cdot \frac{n}{2} - O(\log n)$$

We omit the proof for lack of space.

5.2 Multi-Party Set Disjointness in Simultaneous Messages

In this section we consider the multi-party version of the disjointness function due to [2] (see Section 3). We show an (almost) optimal lower bound on the distributional complexity of this problem in the simultaneous communication model.

Recall that in the k-party disjointness problem, $\textsc{disj}_k$, k players are given each a set X_i from a universe $[n]$, and they have to distinguish between two extreme cases: either the sets are pairwise disjoint, or they are uniquely intersecting. We first note an upper bound for this problem in the simultaneous model:

Proposition 17 *For every $\delta > 0$, $R_\delta^{I_1||\cdots||I_k}(\textsc{disj}_k) = O(\frac{n}{k} \cdot \log n \cdot \log \frac{1}{\delta})$.*

Proof. The protocol works as follows: player i picks a random subset S_i of his input X_i of size $3(n/k^2)\ln(1/\delta)$, and sends it to the referee. The referee declares the sets to be disjoint if and only if $S_1, \ldots, S_k$ are pairwise disjoint. Clearly, the total communication of this protocol is $O((n/k)\log n \log(1/\delta))$.

If $X_1, \ldots, X_k$ are pairwise disjoint, this protocol is correct with probability 1. Consider then the case that the sets are uniquely intersecting. Let m_i be the size of X_i. Clearly, $\sum_{i=1}^k m_i \leq n + (k-1)$, and therefore an averaging argument implies that at least half of the sets are of size at most $2n/k + 2 \leq 3n/k$. Call these players "good". The protocol is erroneous if and only if less than two players choose the intersection point. This happens only if the first $k/2 - 1$ good players do not pick the intersection point. A good player does not choose the intersection point with probability at most $(1 - k/(3n))^{3n/k^2 \ln(1/\delta)} \leq \delta^{1/k}$, implying the error probability of the protocol is at most δ. □

Alon, Matias, and Szegedy [2] showed that for a certain distribution μ', the μ'-distributional communication complexity (for arbitrary, interactive, communication) of this problem is $\Omega(n/k^4)$. We next improve this bound to $\Omega(n/k)$ (which is almost optimal, modulo logarithmic factors) for the simultaneous model. Our lower bound holds for a distribution μ which is a slight variation of μ' (albeit, the proof can be tweaked to work for μ' as well). Note that this implies also an almost-optimal $\Omega(n/k^2)$ lower bound for the maximum communication, because the two complexity measures can differ by at most a factor of k.

The distribution μ is defined by the following process. First, let $(P_1, P_2, \ldots, P_k, I)$ be a uniformly chosen partition of the universe $[n], n = (2m-1)k + 1$, for which $|P_j| = 2m - 1$, $j = 1, \ldots, k$, and $|I| = 1$. Next, choose $X_j \subseteq P_j \cup I$, $j \in [k]$, as follows: With equal probability, either $I \in X_j$, for all j, or $I \notin X_j$, for all j. Then, for each j, choose independently for each $i \in P_j$ to include i in X_j with probability $1/2$.

Theorem 18 *For every $\delta > 0$,*

$$D_{\mu,\delta}^{I_1||\cdots||I_k}(\textsc{disj}_k) \geq (\frac{n-1}{k} + 1) \cdot (1 - H_2(\delta)).$$

Proof. In the proof we use the following notation: given a collection of random variables $Z = (Z_1, \ldots, Z_k)$, define for each j, $Z^{<j} \stackrel{\text{def}}{=} (Z_1, \ldots, Z_{j-1})$ and $Z^{-j} \stackrel{\text{def}}{=} Z \setminus \{Z_j\}$.

Let X_j denote a random input of the j-th player chosen according to μ. Consider a δ-error simultaneous protocol for $\textsc{disj}_k$ in which $A_j = A_j(X_j)$, for $j \in [k]$, is the message sent by player j to the referee. Note that $\mathrm{H}(A_j)$ is at most the communication of player j in the protocol, and therefore to prove the theorem, we will show that $\sum_j \mathrm{H}(A_j) \geq 2m \cdot (1 - H_2(\delta))$.

Let $F \stackrel{\text{def}}{=} \textsc{disj}_k(X_1, \ldots, X_k)$, $P \stackrel{\text{def}}{=} (P_1, \ldots, P_k)$, and $A \stackrel{\text{def}}{=} (A_1, \ldots, A_k)$. Using Fano's inequality (Theorem 2) and the chain rule for mutual information (Proposition 1(5)),

$$\begin{aligned} H_2(\delta) &\geq \mathrm{H}(F \mid A) \geq \mathrm{H}(F \mid A, P) \\ &= \mathrm{H}(F \mid P) - \mathrm{I}(F; A \mid P) \\ &= \mathrm{H}(F \mid P) - \sum_{j \in [k]} \mathrm{I}(F; A_j \mid A^{<j}, P) \end{aligned} \tag{3}$$

Our goal now is to upper bound each term in the above sum separately. Fix a $j \in [k]$. We have

$$\begin{aligned} &\mathrm{I}(F; A_j \mid A^{<j}, P) = \\ &= \mathrm{H}(A_j \mid A^{<j}, P) - \mathrm{H}(A_j \mid F, A^{<j}, P) \\ &\leq \mathrm{H}(A_j \mid P) - \mathrm{H}(A_j \mid F, A^{<j}, P) \end{aligned} \tag{4}$$

because conditioning reduces entropy. For the second term, we use the crucial observation that, given P and F, the

X_k's are mutually independent random variables. Therefore, $\mathrm{I}(X_j; X^{<j} \mid F, P) = 0$. By the data processing inequality for mutual information, $\mathrm{I}(A_j; A^{<j} \mid F, P) = 0$ as well, that is, $\mathrm{H}(A_j \mid F, A^{<j}, P) = \mathrm{H}(A_j \mid F, P)$. Substituting this fact in Equation 4 above:

$$
\begin{aligned}
\mathrm{I}(F; A_j \mid A^{<j}, P) &\leq \mathrm{H}(A_j \mid P) - \mathrm{H}(A_j \mid F, P) \quad (5)\\
= \quad & \mathrm{I}(F; A_j \mid P) = \mathrm{H}(F \mid P) - \mathrm{H}(F \mid A_j, P)
\end{aligned}
$$

Note that $\mathrm{H}(F \mid P) = 1$ because given P, F is uniform on $\{0,1\}$. Consider the second term in the equation above. Since P_j is fixed given P_k, $j \neq k$, and I,

$$
\begin{aligned}
\mathrm{H}(F \mid A_j, P) &= \mathrm{H}(F \mid A_j, I, P^{-j}) = \quad (6)\\
&= \sum_p \Pr(P^{-j} = p) \cdot \mathrm{H}(F \mid A_j, I, P^{-j} = p)
\end{aligned}
$$

where the sum ranges over all valid assignments p to P^{-j}. Consider a fixed p in the sum above, and for brevity, let E denote the event "$P^{-j} = p$". Let σ denote the complement of the union of the sets in p. This set has $2m$ elements, so let $\sigma = \{1, 2, \ldots 2m\}$, without loss of generality. Given E, I is chosen uniformly from σ, which then fixes P_j as well. Thus, X_j is a subset of σ given E. Let Z_i, for $i \in [2m]$, denote the indicator random variable for whether $i \in X_j$. Given E, the Z_i's completely determine X_j and vice versa. Moreover, given E and "$I = i$", $F = Z_i$. Therefore,

$$
\begin{aligned}
&\mathrm{H}(F \mid A_j, I, E) =\\
&= \sum_{i \in [2m]} \Pr(I = i \mid E) \cdot \mathrm{H}(F \mid A_j, I = i, E) \quad (7)\\
&= \frac{\sum_{i \in [2m]} \mathrm{H}(Z_i \mid A_j, I = i, E)}{2m}
\end{aligned}
$$

Now, it can be verified that given E, the random variables X_j and I are independent. Therefore both Z_i and $A_j = A_j(X_j)$, which depend only on X_j, are also independent of I given E. It follows that $\mathrm{H}(Z_i \mid A_j, I = i, E) = \mathrm{H}(Z_i \mid A_j, E)$. Substituting this into Equation 7 above,

$$
\begin{aligned}
\mathrm{H}(F \mid A_j, I, E) &= \frac{\sum_{i \in [2m]} \mathrm{H}(Z_i \mid A_j, E)}{2m}\\
&\geq \frac{\mathrm{H}(Z_1, \ldots, Z_{2m} \mid A_j, E)}{2m} = \frac{\mathrm{H}(X_j \mid A_j, E)}{2m}
\end{aligned}
$$

using the entropy inequality Proposition 1, part (6). Using this bound in Equation 6, we obtain

$$
\begin{aligned}
\mathrm{H}(F \mid A_j, P) &\geq \frac{\mathrm{H}(X_j \mid A_j, P^{-j})}{2m} \geq \frac{\mathrm{H}(X_j \mid A_j, P)}{2m}\\
&\geq \frac{\mathrm{H}(X_j \mid P) - \mathrm{H}(A_j \mid P)}{2m} \geq 1 - \frac{\mathrm{H}(A_j)}{2m}
\end{aligned}
$$

because given P, X_j is a random subset of $P_j \cup I$. Combining this bound with Equations 5 and 3, and again using $\mathrm{H}(F \mid P) = 1$, proves the theorem. □

5.3 Generalized Addressing Function

In this section we prove lower bounds in the simultaneous model for the k-party *generalized addressing function* (GAF_{k+1}) [3], both in the NIH and the NOF models[3]. GAF_{k+1} is defined as follows. Let $F : G \to \{0,1\}$ where G is a group of size n. (Think of F as a table of size n indexed by the elements of the group.) In the NIH model, player i, for $i = 1, \ldots, k$, holds F and $X_i \in G$; in the NOF model player i holds F and all $X_j \in G$ for $j \neq i$. In both models the referee holds $X_1, \ldots, X_k$, but not F[4] and is required to output $F(X_1 \cdots X_k)$.

Following [3], suppose G can be written as a product $G = G_1 \cdot G_2 \cdot \ldots \cdot G_k$ of k groups. Let $H = G_1 \times \cdots \times G_k$ denote the Cartesian product of the G_i's. For each $g \in G$, we pick $(g_1, \ldots, g_k) \in H$ such that $g_1 g_2 \cdots g_k = g$ and identify it with g, i.e., $g \cong (g_1, \ldots, g_k)$. Let $\hat{H}_i = G_1 \times \cdots \times G_{i-1} \times G_{i+1} \times \cdots \times G_k$, and let $\rho = \max_i |\hat{H}_i|$. Babai *et al.* [3] show that the deterministic communication complexity of GAF_{k+1} is at least $|G|/(\rho k)$. We exhibit a simple proof of an analogous result for the more general distributional case.

Define the distribution μ that picks $g \in G$ and $F \in \{0,1\}^n$ uniformly at random, independently. Suppose $g \cong (g_1, g_2, \ldots, g_k)$. Set $X_i = g_i$. Note that this is *not* necessarily a product distribution on the H_i's. Still, we can massage the information theory inequalities to obtain the distributional complexity result.

Theorem 19 *In the NOF model, for every $\delta > 0$,*

$$
D_{\mu,\delta}^{\max, I_1 || \cdots || I_k}(\mathrm{GAF}_{k+1}) \geq \frac{|G|}{\rho k} \cdot (1 - H_2(\delta)).
$$

Proof. For a collection of random variables $Z = (Z_1, \ldots, Z_k)$, we use the notation $Z^{<j}$ and Z^{-j}, similar to the notation in the proof of Theorem 18.

Let $A = (A_1, \ldots, A_k)$ be such that $A_j = A_j(F, X^{-j})$ is the protocol used by player j in a δ-error protocol. Using Fano's inequality, and $\Pr(g \in G) = 1/n$,

$$
\begin{aligned}
H_2(\delta) &\geq \mathrm{H}(\mathrm{GAF}_{k+1}(F, X) \mid A, X) \quad (8)\\
&= \frac{1}{n} \sum_g \mathrm{H}(\mathrm{GAF}_{k+1}(F, X) \mid A, X = g)
\end{aligned}
$$

where $g \cong (g_1, \ldots, g_k)$. For any $g \in G$ in the sum above, $\mathrm{H}(\mathrm{GAF}_{k+1}(F, X) \mid A, X = g) = \mathrm{H}(\mathrm{GAF}_{k+1}(F, g) \mid \{A_j(F, g^{-j}) : j \in [k]\}, X = g)$. Now, observe that we can drop the event $X = g$ because F is independent of X, and so is $\mathrm{GAF}_{k+1}(F, g)$ for a fixed g. Moreover, replacing

[3]We use GAF_{k+1} to be consistent with [3] who consider the referee as an additional player.

[4]This is a slight variation from the standard definition of the simultaneous model.

the collection $\{A_j(F, g^{-j}) : j \in [k]\}$ by a larger collection $S \stackrel{\text{def}}{=} \{A_j(F, h_j) : j \in [k], h_j \in \hat{H}_j\}$ cannot increase the entropy. Therefore, $\sum_g \mathrm{H}(\text{GAF}_{k+1}(F, X) \mid A, X = g) \geq \sum_g \mathrm{H}(\text{GAF}_{k+1}(F, g) \mid S) \geq \mathrm{H}(\{\text{GAF}_{k+1}(F, g) : g \in G\} \mid S) \geq \mathrm{H}(\{\text{GAF}_{k+1}(F, g) : g \in G\}) - \mathrm{H}(S)$, by the entropy inequalities. Note that $\mathrm{H}(\{\text{GAF}_{k+1}(F, g) : g \in G\}) = \mathrm{H}(F) = n$, and $\mathrm{H}(S) \leq \rho k C$, where C is the maximum communication used by any player. Substituting this bound in Equation 8, proves the theorem. $\square$

A similar proof also leads to an analogous theorem for the NIH model (proof omitted):

Theorem 20 *In the NIH model, for every $\delta > 0$,*

$$D_{\mu,\delta}^{\max, I_1||\cdots||I_k}(\text{GAF}_{k+1}) \geq \frac{|G|}{k \max_i |G_i|} \cdot (1 - H_2(\delta)).$$

6 Discussion and open problems

The most glaring open problem is to extend our method to arbitrary round communication complexity, going beyond the one-way and simultaneous message models. It seems that obtaining the result of [14, 21] would be a great first step. In the one-way model, obtaining the analogue of Theorem 18 would be very interesting, and would give new lower bounds for data stream algorithms, settling some open problems from [2]. For the classical $\mathbf{Z}_2$ inner product problem, our methods can be applied, but they yield poor bounds compared to the discrepancy method. Can this be rectified? It seems that for symmetric functions, simultaneous message models are as powerful as one-way models, even for non-product distributions, but formalizing this issue seems subtle, and a proof seems considerably difficult. Lastly, we have not addressed communication complexity in the private coin model.

Acknowledgments

We thank Christos Papadimitriou for helpful discussions.

References

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

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

[3] L. Babai, A. GáL, P. Kimmel, and S. V. Lokam. Simultaneous messages vs. communication. Technical Report TR-96-23, University of Chicago, 1996.

[4] L. Babai, P. Kimmel, and S. V. Lokam. Simultaneous messages vs. communication. In *Proc. 12th STACS*, volume 900 of *LNCS*, pages 361–372, 1995.

[5] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *JACM*, 36(4):929–965, 1989.

[6] A. Borodin. Time space tradeoffs (getting closer to the barrier?). In *Proc. 4th ISAAC*, volume 762 of *LNCS*, pages 209–220, 1993.

[7] A. Chakrabarti, Y. Shi, A. Wirth, and A. C.-C. Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In *Proc. 42nd FOCS*, pages 270–278, 2001.

[8] A. K. Chandra, M. L. Furst, and R. J. Lipton. Multi-party protocols. In *Proc. 15th STOC*, pages 94–99, 1983.

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

[10] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Patterm Recognition*. Springer-Verlag, 1996.

[11] J. Feigenbaum, S. Kannan, M. Strauss, and M. Viswanathan. An approximate L^1-difference algorithm for massive data streams. In *Proc. 40th FOCS*, pages 501–511, 1999.

[12] T. S. Han. Non-negative entropy measures of multivariate symmetric correlations. *Info. & Control*, 36:133–156, 1978.

[13] M. Henzinger, P. Raghavan, and S. Rajagopalan. Computing on data streams. In *DIMACS series in Discr. Math. & and Theor. Comp. Sc.*, volume 50, pages 107–118, 1999.

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

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

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

[17] P. B. Miltersen, N. Nisan, S. Safra, and A. Wigderson. On data structures and asymmetric communication. *JCSS*, 57(1):37–49, 1998.

[18] N. Nisan and A. Wigderson. Rounds in communication complexity revisited. *SICOMP*, 22:211–219, 1993.

[19] S. J. Ponzio, J. Radhakrishnan, and S. Venkatesh. The communication complexity of pointer chasing Applications of entropy and sampling. In *Proc. 31st STOC*, pages 602–611, 1999.

[20] P. Pudlák, V. Rödl, and J. Sgall. Boolean circuits, tensor ranks and communication complexity. *SICOMP*, 26(3):605–633, 1997.

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

[22] M. Saks and X. Sun. Space lower bounds for distance approximation in the data stream model. In *Proc. 34th STOC*, 2002. To appear.

[23] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theor. of Prob. & its Appl.*, XVI(2):264–280, 1971.

[24] A. C.-C. Yao. Some complexity questions related to distributed computing. In *Proc. 11th STOC*, pages 209–213, 1979.

[25] A. C.-C. Yao. Lower bounds by probabilistic arguments. In *Proc. 24th FOCS*, pages 420–428, 1983.