2014 IEEE Annual Symposium on Foundations of Computer Science

Exponential Separation of Information and Communication

Anat Ganor*
Weizmann Institute of Science
Rehovot, Israel
Email: anat.ganor@weizmann.ac.il

Gillat Kol†
Institute for Advanced Study
Princeton, NJ
Email: gillat.kol@gmail.com

Ran Raz*‡
Weizmann Institute of Science
Rehovot, Israel, and
Institute for Advanced Study
Princeton, NJ
Email: ran.raz@weizmann.ac.il

***Abstract*—We show an exponential gap between communication complexity and information complexity, by giving an explicit example for a communication task (relation), with information complexity $\leq O(k)$, and distributional communication complexity $\geq 2^k$. This shows that a communication protocol cannot always be compressed to its internal information. By a result of Braverman [1], our gap is the largest possible. By a result of Braverman and Rao [2], our example shows a gap between communication complexity and amortized communication complexity, implying that a tight direct sum result for distributional communication complexity cannot hold.**

***Keywords*-communication complexity, amortized communication complexity, communication compression, direct sum, information complexity**

I. Introduction

Communication complexity is a central model in complexity theory that has been extensively studied in numerous works. In the two player distributional model, each player gets an input, where the inputs are sampled from a joint distribution that is known to both players. The players' goal is to solve a communication task that depends on both inputs. The players can use both common and private random strings and are allowed to err with some small probability. The players communicate in rounds, where in each round one of the players sends a message to the other player. The communication complexity of a protocol is the total number of bits communicated by the two players. The communication complexity of a communication task is the minimal number of bits that the players need to communicate in order to solve the task with high probability, where the minimum is taken over all protocols. For excellent surveys on communication complexity see [3], [4].

The information complexity model, first introduced by [5]–[7], studies the amount of information that the players need to reveal about their inputs in order to solve a communication task. The model was motivated by fundamental information theoretical questions of compressing communication, as well as by fascinating relations to communication complexity, and in particular to the direct sum problem in communication complexity, a problem that has a rich history, and has been studied in many works and various settings [5], [7]–[15] (and many other works). In this paper we will mainly be interested in internal information complexity (a.k.a, information complexity and information cost). Roughly speaking, the internal information complexity of a protocol is the number of information bits that the players learn about each other's input, when running the protocol. The information complexity of a communication task is the minimal number of information bits that the players learn about each other's input when solving the task, where the minimum is taken over all protocols.

Many recent works focused on the problem of compressing interactive communication protocols. Given a communication protocol with small information complexity, can the protocol be compressed so that the total number of bits communicated by the protocol is also small? There are several beautiful known results, showing how to compress communication protocols in several cases. Barak, Braverman, Chen and Rao showed how to compress any protocol with information complexity k and communication complexity c, to a protocol with communication complexity $\tilde{O}(\sqrt{ck})$ in the general case, and $\tilde{O}(k)$ in the case where the underlying distribution is a product distribution [7]. Braverman and Rao showed how to compress any one round (or small number of rounds) protocol with information complexity k to a protocol with communication complexity $O(k)$ [2]. Braverman showed how to compress any protocol with information complexity k to a protocol with communication complexity $2^{O(k)}$ [1] (see also [16], [17]). This last protocol is the most related to our work, as it gives a compression result that works in the general case and doesn't depend at

*Research supported by an Israel Science Foundation grant and by the I-CORE Program of the Planning and Budgeting Committee and the Israel Science Foundation.
†Research supported at the IAS by The Fund For Math and the Weizmann Institute of Science National Postdoctoral Award Program for Advancing Women in Science.
‡Research supported at the IAS by the The Fund For Math and The Simonyi Fund, and by NSF grants number CCF-0832797, DMS-0835373.

0272-5428/14 $31.00 © 2014 IEEE
DOI 10.1109/FOCS.2014.27

IEEE computer society

all on the communication complexity of the original protocol. Braverman also described a communication complexity task that has information complexity $O(k)$ and no known communication protocol with communication complexity smaller than 2^k [18]. However, there is no known lower bound on the communication complexity of that problem.

Another line of works shows that many of the known general techniques for proving lower bounds for randomized communication complexity also give lower bounds for information complexity [1], [16], [17].

In this work we show the first gap between information complexity and communication complexity of a communication task. We give an explicit example for a communication task (a relation), called the *bursting noise game*, parameterized by $k \in \mathbf{N}$ and played with an input distribution μ. We prove that the information complexity of the game is $O(k)$, while any communication protocol for solving this game, with communication complexity at most 2^k, almost always errs. By the above mentioned compression protocol of Braverman [1], our result gives the largest possible gap between information complexity and communication complexity.

*Theorem 1 (**Communication Lower Bound**):* Every randomized protocol (with shared randomness) for the bursting noise game with parameter k, that has communication complexity at most 2^k, errs with probability $\epsilon \geq 1 - 2^{-\Omega(k)}$ (over the input distribution μ).

*Theorem 2 (**Information Upper Bound**):* There exists a randomized protocol for the bursting noise game with parameter k, that has information cost $O(k)$ and errs with probability $\epsilon \leq 2^{-\Omega(k)}$ (over the input distribution μ).

We note that both the inputs and the outputs in our bursting noise game example are very long. Namely, the input length is triple exponential in k, and the output length is double exponential. The protocol that achieves information complexity $O(k)$ has communication complexity double exponential in k.

As mentioned above, information complexity is also related to the direct sum problem in communication complexity. Braverman and Rao showed that information complexity is equal to the amortized communication complexity, that is, the limit of the communication complexity needed to solve n tasks of the same type, divided by n [2] (see also [1], [18], [19]). Our result therefore shows a gap between distributional communication complexity and amortized distributional communication complexity, proving that tight direct sum results for the communication complexity of relations cannot hold.

Organization: The paper is organized as follows. In Section II we define the bursting noise game. In Section III we give general definitions and preliminaries. In Section IV we state the graph correlation lemma, a central tool that we will use in the lower bound proof. Section V gives an overview of our main result, the lower bound for the communication complexity of the bursting noise game (Theorem 1). Section VI gives a general tool that can be used to upper bound the information cost of a protocol, using the notion of a divergence cost of a tree. In Section VII we give a protocol for the bursting noise game with low information cost, thus proving the upper bound required by Theorem 2.

II. Bursting Noise Games

The *bursting noise game* is a communication game between two parties, called the *first player* and the *second player*. The game is specified by a parameter $k \in \mathbf{N}$, where $k > 2^{100}$. We set $c = 2^{4^k}$ and $w = 2^{100}k$.

The game is played on the binary tree $\mathcal{T}$ with $c \cdot w$ layers (the root is in layer 1 and the leaves are in layer $c \cdot w$), with edges directed from the root to the leaves. Denote the vertex set of $\mathcal{T}$ by V. Each player gets as input a bit for every vertex in the tree. Let x be the input given to the first player, and y be the input given to the second player, where $x, y \in \{0,1\}^V$. For a vertex $v \in V$, we denote by x_v and y_v the bits in x and y associated with v. The input pair (x, y) is selected according to a joint distribution μ on $\{0,1\}^V \times \{0,1\}^V$, defined below.

Denote by $\mathrm{Even}(\mathcal{T}) \subseteq V$ the set of non-leaf vertices in an even layer of $\mathcal{T}$ and by $\mathrm{Odd}(\mathcal{T}) \subseteq V$ the set of non-leaf vertices in an odd layer of $\mathcal{T}$. We think of the vertices in $\mathrm{Odd}(\mathcal{T})$ as "owned" by the first player and the vertices in $\mathrm{Even}(\mathcal{T})$ as "owned" by the second player. Let $v \in V$ be a non-leaf vertex. Let v_0 be the left child of v and v_1 be the right child of v. Let $b \in \{0,1\}$. We say that v_b is the *correct child* of v with respect to x, y, if either the first player owns v and $x_v = b$, or the second player owns v and $y_v = b$.

We think of the $c \cdot w$ layers of the tree $\mathcal{T}$ as partitioned into c multi-layers, each consisting of w consecutive layers (e.g., the first multi-layer consists of layers 1 to w). We denote by i^* the first layer of the i^{th} multi-layer, that is, $i^* = (i-1)w + 1$.

For $s \leq t \in \mathbf{N}$, denote by $[s,t]$ the set $\{s, \ldots, t\}$ and by $[t]$ the set $\{1, \ldots, t\}$. Let $i \in [c]$ be a multi-layer. Denote $s = i^*$ and $t = s+w-1 = (i+1)^*-1$. Let $t' \in [(i+1)^*, cw]$, and let $v \in V$ be a vertex in layer t' of $\mathcal{T}$. For $j \in [s, t+1]$, let v_j be v's ancestor in layer j. We say that v is *typical* with respect to i, x, y, if the followings hold:

1) For at least 0.8-fraction of the indices $j \in [s,t] \cap \mathrm{Odd}(\mathcal{T})$, the vertex v_{j+1} is the correct child of v_j with respect to x, y.
2) For at least 0.8-fraction of the indices $j \in [s,t] \cap \mathrm{Even}(\mathcal{T})$, the vertex v_{j+1} is the correct child of v_j with respect to x, y.

Observe that in order to decide whether v is typical with respect to i, x, y, it suffices to know the bits that x, y assign to the vertices $v_s, \ldots, v_t$. When x, y are clear from the context, we omit x, y and say that v is typical with respect to multi-layer i.

Algorithm I: Sample (x, y) according to μ

1) Randomly select $i \in [c]$ (the noisy multi-layer).
2) Set every vertex in multi-layer i (layers $[i^*, i^* + w - 1]$) to be noisy.
3) If $i < c$: Let L be the set of all non-typical vertices in layer $i^* + w = (i+1)^*$ with respect to i, x, y (note that x, y were already defined on layers $[i^*, i^* + w - 1]$, and therefore the typical vertices are defined). For every $v \in L$, set all the vertices in the subtree rooted at v to be noisy.
4) Set all unset vertices in V to be non-noisy.

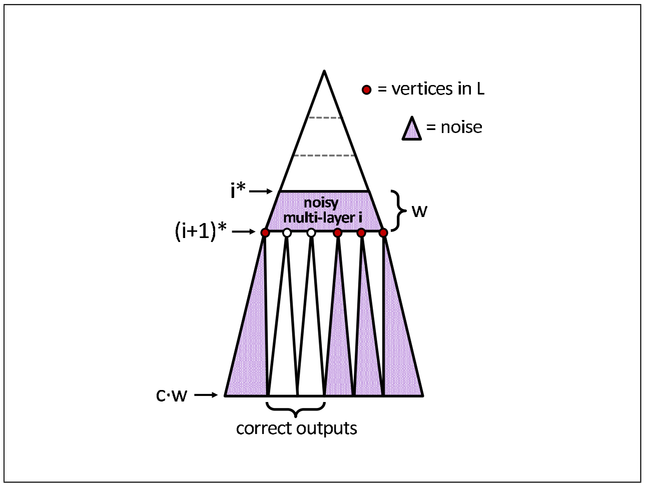

Figure 1: Illustration of Algorithm I

We next define the distribution μ on $\{0,1\}^V \times \{0,1\}^V$ by an algorithm for sampling an input pair (x, y) (Algorithm I below). In the algorithm, when we say "set v to be non-noisy", we mean "select $x_v \in \{0,1\}$ uniformly at random and set $y_v = x_v$". By "set v to be noisy", we mean "select $x_v \in \{0,1\}$ and $y_v \in \{0,1\}$ independently and uniformly at random". Figure 1 illustrates Algorithm I. The players' mutual goal is to output the same leaf $v \in V$, where v is typical with respect to i, x, y (that is, v is typical with respect to the noisy multi-layer; see Algorithm I).

For $i \in [c]$, we denote by μ_i the distribution μ conditioned on the event that the noisy multi-layer selected by Step 1 of the algorithm defining μ, is i. Note that $\mu = \frac{1}{c}\sum_{i \in [c]} \mu_i$.

Remark 1: Observe that it is not always possible to deduce i (i.e., the index of the noisy multi-layer used to construct the pair (x, y)) from the pair (x, y). Therefore, the bursting noise game does not induce a relation. Nevertheless, with extremely high probability, the first multi-layer on which x and y disagree is i. Thus, the game can be easily converted to a relation, by omitting the rare inputs (x, y) that agree on multi-layer i. Note that since the statistical distance between the two distributions is negligible, both our upper bound and lower bound trivially apply to the new game as well. For that reason, it will be helpful to think of the supports of the different μ_i's as if they were pairwise disjoint.

Remark 2: Observe that c is set to be double exponential in k. If c were set to be just exponential in k, a simple binary search algorithm would have been able to find the location of the noisy multi-layer, and thus solve the bursting noise game with communication complexity polynomial in k.

III. Definitions and Preliminaries

A. General Notation

Throughout the paper, all logarithms are taken with base 2, and we define $0\log(0) = 0$. For a set S, when we write "$x \in_R S$" we mean that x is selected uniformly at random from the set S. For a distribution τ, when we write "$x \leftarrow \tau$" we mean that x is selected according to the distribution τ. For Z that is either a random variable taking values in $\{0,1\}^V$ or an element in $\{0,1\}^V$, and a set $T \subseteq V$, we define Z_T to be the projection of Z to T.

B. Information Cost

*Definition 1 (**Information Cost**):* The *information cost* of a protocol π over random inputs (X, Y) that are drawn according to a joint distribution μ, is defined as

$$IC_\mu(\pi) = \mathbf{I}(\Pi; X|Y) + \mathbf{I}(\Pi; Y|X),$$

where Π is a random variable which is the transcript of the protocol π with respect to μ. That is, Π is the concatenation of all the messages exchanged during the execution of π. The ϵ information cost of a computational task f with respect to a distribution μ is defined as

$$IC_\mu(f, \epsilon) = \inf_\pi IC_\mu(\pi),$$

where the infimum ranges over all protocols π that solve f with error at most ϵ on inputs that are sampled according to μ.

C. Relative Entropy

*Definition 2 (**Relative Entropy**):* Let $\mu_1, \mu_2 : \Omega \to [0,1]$ be two distributions, where Ω is discrete (but not necessarily finite). The *relative entropy* between μ_1 and μ_2, denoted $\mathbf{D}(\mu_1 \| \mu_2)$, is defined as

$$\mathbf{D}(\mu_1 \| \mu_2) = \sum_{x \in \Omega} \mu_1(x) \log\left(\frac{\mu_1(x)}{\mu_2(x)}\right).$$

Proposition 3: Let $\mu_1, \mu_2 : \Omega \to [0,1]$ be two distributions. Then,

$$\mathbf{D}(\mu_1 \| \mu_2) \geq 0.$$

The following relation is called Pinsker's inequality, and it relates the relative entropy to the ℓ_1 distance.

*Proposition 4 (**Pinsker's Inequality**):* Let $\mu_1, \mu_2 : \Omega \to [0,1]$ be two distributions. Then,

$$2\ln(2) \cdot \mathbf{D}(\mu_1 \| \mu_2) \geq \|\mu_1 - \mu_2\|^2,$$

where

$$\begin{aligned} \|\mu_1 - \mu_2\| &= \sum_{x \in \Omega} |\mu_1(x) - \mu_2(x)| \\ &= 2 \max_{E \subseteq \Omega} \{\mu_1(E) - \mu_2(E)\}. \end{aligned}$$

D. Information

*Definition 3 (**Information**):* Let $\mu : \Omega \to [0,1]$ be a distribution and let $\mathcal{U}$ be the uniform distribution over Ω. The *information* of μ, denoted $\mathbf{I}(\mu)$, is defined by

$$\begin{aligned} \mathbf{I}(\mu) &= \mathbf{D}(\mu \parallel \mathcal{U}) \\ &= \sum_{x \in \mathrm{supp}(\mu)} \mu(x) \log \left(\frac{\mu(x)}{\frac{1}{|\Omega|}} \right) \\ &= \sum_{x \in \mathrm{supp}(\mu)} \mu(x) \log (|\Omega| \mu(x)). \end{aligned}$$

Equivalently,

$$\mathbf{I}(\mu) = \log(|\Omega|) - \mathbf{H}(\mu),$$

where $\mathbf{H}(\mu)$ denotes the Shannon entropy of μ.

For a random variable X taking values in Ω, with distribution $P_X : \Omega \to [0,1]$, we define $\mathbf{I}(X) = \mathbf{I}(P_X)$.

E. Shearer-Like Inequality for Information

The following version of Shearer's inequality [20], [21] is due to [22].

*Lemma 5 (**Shearer's Inequality**):* Let $X_1, \ldots, X_M$ be M random variables. Let $X = (X_1, \ldots, X_M)$. Let $T = \{T_i\}_{i \in I}$ be a collection of subsets of $[M]$, such that each element of $[M]$ appears in at least K members of T. For $A \subseteq [M]$, let $X_A = \{X_j : j \in A\}$. Then,

$$\sum_{i \in I} \mathbf{H}[X_{T_i}] \geq K \cdot \mathbf{H}[X].$$

We state here the following "Shearer-like" inequality for information. A variant of this lemma was proved in [23].

*Lemma 6 (**Shearer-Like Inequality for Information**):* Let $X_1, \ldots, X_M$ be M random variables, taking values in $\Omega_1, \ldots, \Omega_M$, respectively. Let $X = (X_1, \ldots, X_M)$ be a random variable, taking values in $\Omega_1 \times \cdots \times \Omega_M$. Let $T = \{T_i\}_{i \in I}$ be a collection of subsets of $[M]$, such that each element of $[M]$ appears in at most $\frac{1}{K}$ fraction of the members of T. For $A \subseteq [M]$, let $X_A = \{X_j : j \in A\}$. Then,

$$K \cdot \mathop{\mathbf{E}}_{i \in_R I}[\mathbf{I}(X_{T_i})] \leq \mathbf{I}(X).$$

The next lemma generalizes Lemma 6, and gives a Shearer-like inequality for relative entropy. A variant of this lemma was proved in [23]. The lemma will not be used in the paper, but we include it here as it may be useful in this context.

*Lemma 7 (**Shearer-Like Inequality for Relative Entropy**):* Let $P, Q : \Omega_1 \times \cdots \times \Omega_M \to [0,1]$ be two distributions, such that Q is a product distribution, i.e., for every $j \in [M]$, there exists $Q_j : \Omega_j \to [0,1]$, such that $Q(x_1, \ldots, x_M) = \prod_{j \in [M]} Q_j(x_j)$. Let $T = \{T_i\}_{i \in I}$ be a collection of subsets of $[M]$, such that each element of $[M]$ appears in at most $\frac{1}{K}$ fraction of the members of T. For $A \subseteq [M]$, let P_A and Q_A be the marginal distributions of A in the distributions P and Q (respectively). Then,

$$K \cdot \mathop{\mathbf{E}}_{i \in_R I}[\mathbf{D}(P_{T_i} \| Q_{T_i})] \leq \mathbf{D}(P \| Q).$$

IV. The Graph Correlation Lemma

*Lemma 8 (**Graph Correlation Lemma**):* [1] Let $G = (U \cup W, E)$ be a bipartite (multi)-graph with sets of vertices U, W and (multi)-set of edges E, such that, G is bi-regular and $|U| = |W|$. Let $M > T > k \in \mathbf{N}$ be such that, $T \leq 2^{-20k} M$, and $k \geq 4$. For every $(u, w) \in E$, let $T(u, w) \subset [M]$ be a set of size T, such that, for every $u \in U$, each element of $[M]$ appears in at most 2^{-20k} fraction of the sets in $\{T(u,w)\}_{(u,w) \in E}$, and for every $w \in W$, each element of $[M]$ appears in at most 2^{-20k} fraction of the sets in $\{T(u,w)\}_{(u,w) \in E}$.

Let Σ be a finite set. For every $u \in U$, let $X^u \in \Sigma^M$ be a random variable, such that, $\mathbf{I}(X^u) \leq 2^{4k}$, and for every $w \in W$, let $Y^w \in \Sigma^M$ be a random variable, such that, $\mathbf{I}(Y^w) \leq 2^{4k}$, and such that, for every $u \in U$ and $w \in W$, the random variables X^u and Y^w are mutually independent.

For $(u, w) \in E$, denote

$$\mu(u, w) = \frac{\Pr_{X^u, Y^w}[X^u_{T(u,w)} = Y^w_{T(u,w)}]}{|\Sigma|^{-T}}.$$

Let

$$\mathcal{D} = \{(u, w) \in E : \mu(u, w) \leq 1 - 2^{-4k}\}.$$

Then,

$$\frac{|\mathcal{D}|}{|E|} \leq 2^{-4k}.$$

V. Overview of the Lower Bound Proof

Due to space limitations, we give an overview of the proof of Theorem 1.

Rectangle Partition: We fix the random strings for the protocol so that we have a deterministic protocol. We show that if the protocol communicates at most 2^k bits, it errs with probability $1 - 2^{-\Omega(k)}$ on inputs sampled according to μ. We will show that for almost all $i \in [c]$, the protocol errs with probability $1 - 2^{-\Omega(k)}$ on inputs sampled according to μ_i, that is, the distribution μ conditioned on the event

[1] Many variants of this lemma can be proven. In particular, a similar argument can be used to prove a similar statement with sets $T(u, w)$ that are not of the same size. We state the lemma here for sets $T(u, w)$ of the same size T, for convenience of notation.

that the noisy multi-layer selected by Step 1 of Algorithm I defining μ, is i. Note that the distribution μ_i is uniformly distributed over $\mathrm{supp}(\mu_i)$, and that for every pair of inputs $(x, y) \in \mathrm{supp}(\mu_i)$, the projection of x and y on the first $i-1$ multi-layers is the same.

As mentioned above, it will be helpful to think of the supports of the different μ_i's as if they were pairwise disjoint (this property holds if we remove a μ_i-negligible set of inputs from the support of each μ_i).

Let $\{R^1, \ldots, R^m\}$ be the rectangle partition induced by the protocol, where $R^t = A^t \times B^t$, and $m \leq 2^{2^k}$. For $i \in [c]$ and an assignment z to the first $i-1$ multi-layers, we denote by $R^{t,z} = A^{t,z} \times B^{t,z}$, the rectangle of all pairs of inputs $(x, y) \in R^t$, such that the projection of both x, y on the first $i-1$ multi-layers is equal to z. Let $X^{t,z}$ be a random variable uniformly distributed over $A^{t,z}$. Let $Y^{t,z}$ be a random variable uniformly distributed over $B^{t,z}$. We denote by $X_i^{t,z}, Y_i^{t,z}$ the projections of $X^{t,z}, Y^{t,z}$, respectively, on multi-layer i.

For fixed i, z, we define $\rho^{i,z}$ to be a probability distribution that selects a rectangle in $\{R^{1,z}, \ldots, R^{m,z}\}$ according to its relative size. That is, $\rho^{i,z}$ is defined as follows: Randomly select x, y, such that the projection of both x and y on the first $i-1$ multi-layers is z. Select t to be the index of the unique rectangle $R^{t,z}$ containing (x, y).

Bounding the Information on the Noisy Multi-Layer: The main intuition of the proof is that since c is significantly larger than 2^k, the protocol cannot make progress on all multi-layers $i \in [c]$ simultaneously. We first show that for a random $i \in [c]$, a random z, and a random rectangle $R^{t,z}$, chosen according to $\rho^{i,z}$, very little information is known about $X_i^{t,z}$ and $Y_i^{t,z}$.

Formally, we prove that

$$\mathop{\mathbf{E}}_{i} \mathop{\mathbf{E}}_{z} \mathop{\mathbf{E}}_{t \leftarrow \rho^{i,z}} \left[\mathbf{I}\left(X_i^{t,z}\right)\right] \leq \frac{m}{c}, \tag{1}$$

and similarly,

$$\mathop{\mathbf{E}}_{i} \mathop{\mathbf{E}}_{z} \mathop{\mathbf{E}}_{t \leftarrow \rho^{i,z}} \left[\mathbf{I}\left(Y_i^{t,z}\right)\right] \leq \frac{m}{c}. \tag{2}$$

The proof doesn't follow by a trivial application of super-additivity of information. That's because choosing i, z at random and t according to $\rho^{i,z}$ and then choosing a random variable X to be uniformly distributed on $A^{t,z}$, gives a random variable X with distribution that may be very far from uniform. Moreover, the probability that X is in the set A^t, associated with a rectangle R^t, may be very far from the probability that a uniformly distributed input is in A^t. Nevertheless, we are still able to prove this using the fact that we have a bound of m on the total number of times that an input x appears in the cover $\{A^1, \ldots, A^m\}$.

We fix $\gamma = 2^{-k/4}$, and we fix i, z, t, such that,

1) $\mathbf{I}\left(X_i^{t,z}\right) \leq \frac{1}{\gamma} \cdot \frac{m}{c}$
2) $\mathbf{I}\left(Y_i^{t,z}\right) \leq \frac{1}{\gamma} \cdot \frac{m}{c}$
3) The rectangle $R^{t,z}$ is not too small.

By equations (1) and (2), and by Markov's inequality, we know that when we choose i, z uniformly at random, and t according to $\rho^{i,z}$, the triplet (i, z, t) satisfies all three conditions with high probability. Therefore, we ignore triplets (i, z, t) that do not satisfy all three conditions.

Unique Answer Rectangles: In the rectangle $R^{t,z}$, the answer of each of the two players in the protocol may not be unique, as the answer of each player may also depend on the input that she gets. Nevertheless, using the fact that if the two players answer differently then the protocol errs, we are able to subdivide the rectangle $R^{t,z}$ into $\mathrm{poly}(1/\gamma)$ sub-rectangles $R^{t,s,z}$, such that in each rectangle $R^{t,s,z}$ the answer is unique, except for a bad set of rectangles whose total size is negligible compared to the size of $R^{t,z}$. When subdividing $R^{t,z}$, we also need to change the answers given by the two players on each rectangle, but we are able to do that without adding errors to the protocol.

We ignore rectangles $R^{t,s,z}$ where the answer of the protocol is not unique, as their total size is small, and only consider rectangles $R^{t,s,z} = A^{t,s,z} \times B^{t,s,z}$ where the answer is unique. Let $X^{t,s,z}$ be a random variable uniformly distributed over $A^{t,s,z}$. Let $Y^{t,s,z}$ be a random variable uniformly distributed over $B^{t,s,z}$. For the rectangles $R^{t,s,z}$ we no longer have the strong bounds $\mathbf{I}\left(X_i^{t,z}\right) \leq \frac{1}{\gamma} \cdot \frac{m}{c}$, and $\mathbf{I}\left(Y_i^{t,z}\right) \leq \frac{1}{\gamma} \cdot \frac{m}{c}$, but rather the weaker bounds

$$\mathbf{I}\left(X_i^{t,s,z}\right) \leq O\left(\log\left(1/\gamma\right)\right),$$

and

$$\mathbf{I}\left(Y_i^{t,s,z}\right) \leq O\left(\log\left(1/\gamma\right)\right).$$

How the Proof Works: Fix i, z, t, s. In the rectangle $R^{t,s,z}$ the answer is unique, denote that answer by $\omega^{t,s,z}$. We define $\Lambda^{t,s,z}$ to be the set of input pairs $(x, y) \in \mathrm{supp}(\mu_i)$, such that $\omega^{t,s,z}$ is not a correct answer for the input (x, y). Let P_i be the probability for a uniformly distributed pair of inputs (x, y), that have the same projection on the first $i-1$ multi-layers, to be in $\mathrm{supp}(\mu_i)$. We prove that

$$\Pr\left[(X^{t,s,z}, Y^{t,s,z}) \in \Lambda^{t,s,z}\right] \geq \left(1 - 2^{-\Omega(k)}\right) P_i. \tag{3}$$

Summing over all possibilities for t, s, z, this implies, for almost all $i \in [c]$, that the protocol errs on μ_i with probability $1 - 2^{-\Omega(k)}$, which concludes the proof.

In what follows, we outline the proof of (3).

***The Graph** G:* We define the complete bipartite graph $G = (U \cup W, E)$, where $U = W$ is the set of all possible assignments for multi-layer i (for one player), and $E = U \times W$.

Let M be the number of vertices in layer $(i+1)^*$ of the tree $\mathcal{T}$. We identify the set $[M]$ with the set of vertices in layer $(i+1)^*$. Let $u \in U, w \in W$. We define $T(u, w) \subset [M]$ to be the set of all vertices in layer $(i+1)^*$ that are set to be non-noisy for inputs u, w, by Algorithm I defining μ, when the noisy multi-layer is i. Observe that u and w determine for every vertex in layer $(i+1)^*$ if it is noisy or not. Note that by a symmetry argument, $T(u,w)$ is of the same size T for every u,w.

Let $\mathcal{E} \subseteq E$ be the set of all $(u,w) \in E$ for which the output $\omega^{t,s,z}$ is correct for inputs $(x,y) \in \operatorname{supp}(\mu_i)$, where $x_i = u$ and $y_i = w$. Note that if the noisy multi-layer is i, then u and w determine the correctness of $\omega^{t,s,z}$. It holds that

$$|\mathcal{E}| \leq 2^{-20k}|E|,$$

as for any fixed u and every $v \in [M]$, at most a fraction of 2^{-20k} of the sets $\{T(u,w)\}_{(u,w)\in E}$ contain v, and the output $\omega^{t,s,z}$ is correct only if it has an ancestor in $T(u,w)$.

Let Σ be the set of all possible boolean assignments to the vertices of a subtree of $\mathcal{T}$ rooted at layer $(i+1)^*$.

Denote $X := X^{t,s,z}$ and $Y := Y^{t,s,z}$. For $u \in U$, we define the random variable X^u, over the domain $\Sigma^{[M]}$, to be the conditional variable $(X_{>i}|X_i = u)$, that is, X^u has the distribution of $X_{>i}$ conditioned on the event $X_i = u$, where $X_{>i}$ denotes the projection of X to all multi-layers after multi-layer i. Similarly, for $w \in W$, we define the random variable Y^w, over the domain $\Sigma^{[M]}$, to be $(Y_{>i}|Y_i = w)$, that is, Y^w has the distribution of $Y_{>i}$ conditioned on the event $Y_i = w$.

Application of the Graph Correlation Lemma: By the definition of the distribution μ_i, the left hand side of (3) is equal to

$$\sum_{\substack{(u,w)\in E\\(u,w)\notin\mathcal{E}}} \Pr\left[X_i = u\right]\cdot\Pr\left[Y_i = w\right]\cdot\Pr\left[X^u_{T(u,w)} = Y^w_{T(u,w)}\right], \tag{4}$$

where $X^u_{T(u,w)}$ and $Y^w_{T(u,w)}$ are the projections of X^u, Y^w, respectively, to coordinates in $T(u,w)$. This is true because a pair (x,y) is in $\operatorname{supp}(\mu_i)$ if and only if x,y agree on all the subtrees rooted at vertices in layer $(i+1)^*$ that are set to be non-noisy for inputs x_i, y_i, by Algorithm I defining μ, when the noisy multi-layer is i.

Our graph correlation lemma (Lemma 8), that may be interesting in its own right, gives a general way to bound such expressions by

$$\begin{aligned} &\geq \left(1 - 2^{-\Omega(k)}\right)|\Sigma|^{-T} \\ &\quad\cdot \sum_{\substack{(u,w)\in E\\(u,w)\notin(\mathcal{E}\cup\mathcal{D})}} \Pr\left[X_i = u\right]\cdot\Pr\left[Y_i = w\right], \end{aligned} \tag{5}$$

where $\mathcal{D} \subset E$ is a small set, compared to the size of E, and $|\Sigma|^{-T}$ is a normalization factor that would have been equal to $\Pr[X^u_{T(u,w)} = Y^w_{T(u,w)}]$ if X^u, Y^w were uniformly distributed (independent) random variables.

Thus, using Lemma 8, we are able to bound the left hand side of (3), which is an expression that depends on the variables X, Y, by the expression in (5) that depends only on the projections of these variables to multi-layer i.

We still need to bound from below the expression

$$\sum_{\substack{(u,w)\in E\\(u,w)\notin(\mathcal{E}\cup\mathcal{D})}} \Pr\left[X_i = u\right]\cdot\Pr\left[Y_i = w\right]. \tag{6}$$

Since $\mathcal{E}\cup\mathcal{D}$ is a small set (compared to the size of E), we will first ignore the set $\mathcal{E}\cup\mathcal{D}$, and observe that

$$\begin{aligned} &\sum_{(u,w)\in E} \Pr\left[X_i = u\right]\cdot\Pr\left[Y_i = w\right] \\ &= \sum_{u\in U}\Pr\left[X_i = u\right]\cdot\sum_{w\in W}\Pr\left[Y_i = w\right] = 1. \end{aligned} \tag{7}$$

It remains to show that

$$\sum_{(u,w)\in\mathcal{E}\cup\mathcal{D}} \Pr\left[X_i = u\right]\cdot\Pr\left[Y_i = w\right],$$

is negligible.

Bounding the Sum over the Bad Sets: We use the fact that $R^{t,s,z} \subseteq R^{t,z}$, to bound the last sum by

$$\frac{|R^{t,z}|}{|R^{t,s,z}|}\sum_{(u,w)\in\mathcal{E}\cup\mathcal{D}} \Pr\left[X^{t,z}_i = u\right]\cdot\Pr\left[Y^{t,z}_i = w\right].$$

Since $\mathbf{I}\left(X^{t,z}_i\right) \leq \frac{1}{\gamma}\cdot\frac{m}{c}$, and $\mathbf{I}\left(Y^{t,z}_i\right) \leq \frac{1}{\gamma}\cdot\frac{m}{c}$, we know that the distributions of $X^{t,z}_i$ and $Y^{t,z}_i$ are extremely close to uniform, and hence the sum in the last expression is negligible. Using also the fact that $\frac{|R^{t,z}|}{|R^{t,s,z}|} \leq \operatorname{poly}(1/\gamma)$, we get that the entire expression is negligible.

A difficulty that we ignored in the discussion so far, is that the graph correlation lemma (Lemma 8) requires random variables X^u, Y^w with bounded information for all u, w, while we have variables with bounded information for almost all u, w. To fix that, we just replace every X^u or Y^w that has large information, with a uniformly distributed random variable. This works since G is the complete graph.

Proof of the Graph Correlation Lemma and Shearer's Inequality: To bound expressions such as the expression in (4), we show that if $\Pr[X^u_{T(u,w)} = Y^w_{T(u,w)}]$ is significantly smaller than what is obtained by uniformly distributed variables, then either $\mathbf{I}(X^u_{T(u,w)})$ or $\mathbf{I}(Y^w_{T(u,w)})$ are non negligible (or both). We use this to show that for some u (or some w) we have that $\mathbf{I}(X^u)$ (or $\mathbf{I}(Y^w)$) are large, deriving a contradiction.

Our proof relies on a variant of Shearer's inequality [20], [21] that follows easily by Radhakrishnan's beautiful information theoretical proof [22] (see Lemmas 6 and 7 and [23]).

VI. Bounding Information Cost by Tree Divergence Cost

In this section we give a general tool that can be used to upper bound the information cost of a protocol π, using the notion of a divergence cost of a tree. This notion is implicit in [7] and was formally defined in [2].

Let π be a communication protocol between two players. We assume that the first player has the private input x and the second player has the private input y, where (x, y) were chosen according to some joint distribution μ. In this section, we assume without loss of generality that π does not use public randomness (but may use private randomness), as for the purpose of upper bounding the information cost, the public randomness can always be replaced by private randomness. We also assume, without loss of generality, that the players take alternating turns sending bits to each other. That is, in odd rounds, the first player sends a bit to the second player, and in even rounds the second player sends a bit to the first player (if this is not the case, we can add dummy rounds that do not change the information cost).

We denote by $\mathcal{T}_\pi$ the binary tree associated with the communication protocol π. That is, every vertex v of $\mathcal{T}_\pi$ corresponds to a possible transcript of π, and the two edges going out of v are labeled by 0 and 1, corresponding to the next bit to be transmitted. We think of the first player as owning the vertices in odd layers of $\mathcal{T}_\pi$ (where the root is in layer 1), and of the second player as owning the vertices in even layers of $\mathcal{T}_\pi$. When the protocol π reaches a non-leaf vertex v, the player who owns v sends a bit to the other player.

Every input pair (x, y) for the protocol π induces a distribution $P_v = (p_v, 1 - p_v)$ for every non-leaf vertex v of the tree $\mathcal{T}_\pi$, where p_v is the probability that the next bit transmitted by the protocol π on the vertex v and inputs x, y is 0. We think of P_v as a distribution over the two children of the vertex v. Observe that the player who owns v knows P_v. Given the binary tree $\mathcal{T}_\pi$ and the distributions P_v for every non-leaf vertex v of $\mathcal{T}_\pi$, where for each v the player who owns v knows P_v, we can assume without loss of generality that the protocol π operates as follows: Starting from the root until reaching a leaf, at every vertex v, the player who owns v samples a bit according to P_v and sends this bit to the other player. Both players continue to the child of v that is indicated by the communicated bit.

Assume that for every non-leaf vertex v of $\mathcal{T}_\pi$, we have an additional distribution $Q_v = (q_v, 1 - q_v)$ that is known to the player who doesn't own v. We think of every P_v as the "correct" distribution over the two children of v. This distribution is known to the player who owns v. We think of Q_v as an estimation of P_v, based on the knowledge of the player who doesn't own v. For the rest of the section, we think of $\mathcal{T}_\pi$ as the tree $\mathcal{T}_\pi$ together with the distributions P_v and Q_v, for every non-leaf vertex v in the tree $\mathcal{T}_\pi$.

To upper bound the information cost of a protocol π it is convenient to use the notion of divergence cost of a tree [2], [7].

*Definition 4 (**Divergence Cost** [2], [7]):* Consider a binary tree $\mathcal{T}$, whose root is r, and distributions $P_v = (p_v, 1 - p_v), Q_v = (q_v, 1 - q_v)$ for every non-leaf vertex v in the tree. We think of P_v and Q_v as distributions over the two children of the vertex v. We define the *divergence cost* of the tree $\mathcal{T}$ recursively, as follows. $\mathbf{D}(\mathcal{T}) = 0$ if the tree has depth 0, otherwise,

$$\mathbf{D}(\mathcal{T}) = \mathbf{D}(P_r \| Q_r) + \mathop{\mathbf{E}}_{v \sim P_r}[\mathbf{D}(\mathcal{T}_v)], \tag{8}$$

where for every vertex v, $\mathcal{T}_v$ is the subtree of $\mathcal{T}$ whose root is v.

An equivalent definition of the divergence cost of $\mathcal{T}$ is obtained by following the recursion in (8) and is given by the following equation:

$$\mathbf{D}(\mathcal{T}) = \sum_{v \in V} \tilde{p}_v \cdot \mathbf{D}(P_v \| Q_v), \tag{9}$$

where V is the vertex set of $\mathcal{T}$, and for a vertex $v \in V$, $\tilde{p}_v$ is the probability to reach v by following the distributions P_v, starting from the root. Formally, if v is the root of the tree $\mathcal{T}$, then $\tilde{p}_v = 1$, otherwise,

$$\tilde{p}_v = \begin{cases} \tilde{p}_u \cdot p_u & \text{if } v \text{ is the left-hand child of } u \\ \tilde{p}_u \cdot (1 - p_u) & \text{if } v \text{ is the right-hand child of } u. \end{cases}$$

Let X be the input to the first player and Y be the input to the second player. In the protocol π, the players use two private random strings and no public randomness. Denote the private random string of the first player by R_1, and the private random string of the second player by R_2. For a layer d of $\mathcal{T}_\pi$, let Π_d be the vertex in layer d that the players reach during the execution of the protocol π, when the inputs are (X, Y) and the private random strings are R_1 and R_2 (if π ends before layer d, then Π_d is undefined).

Let the tree $\mathcal{T}'_\pi$ be the same as $\mathcal{T}_\pi$, except that every distribution Q_v, for every non-leaf vertex v in $\mathcal{T}_\pi$, is replaced with the distribution $Q'_v = (q'_v, 1 - q'_v)$, where q'_v is defined as follows: Let d be the layer of v. If v is owned by the first player, q'_v is the function of v, y and r_2, defined as

$$q'_v = \mathop{\mathbf{E}}_{X, R_1}[p_v | Y = y, R_2 = r_2, \Pi_d = v].$$

If v is owned by the second player, q'_v is the function of v, x and r_1, defined as

$$q'_v = \mathop{\mathbf{E}}_{Y, R_2}[p_v | X = x, R_1 = r_1, \Pi_d = v].$$

We think of Q'_v as the best estimation of the correct distribution P_v, based on the knowledge of the player who doesn't own v, whereas Q_v is some estimation. Intuitively, $\mathbf{D}(P_v \| Q_v)$ is the information that the player who doesn't own v learns on P_v from the bit sent during the protocol at the vertex v, assuming that she expects this bit to be distributed according to Q_v, whereas $\mathbf{D}(P_v \| Q'_v)$ is the information that she learns based on the best possible estimation of P_v. Therefore, intuitively, the divergence cost of $\mathcal{T}'_\pi$ is at most the divergence cost of $\mathcal{T}_\pi$, in expectation. This is formulated in the following lemma.

Observe that the protocol π induces the distributions P_v (known to the player who owns v) and Q'_v (known to the player who doesn't own v), while the distribution Q_v may be any distribution known to the player who doesn't own v.

The following lemma relates the information cost of π to the expected divergence cost of $\mathcal{T}_\pi$.

Lemma 9: For every protocol π and distributions Q_v known to the player who doesn't own v, as above, it holds that

$$IC_\mu(\pi) = \mathbf{E}[\mathbf{D}(\mathcal{T}'_\pi)] \leq \mathbf{E}[\mathbf{D}(\mathcal{T}_\pi)],$$

where the expectation is over the sampling of the inputs according to μ and over the randomness.

Proof: It was shown in [2] (see Lemma 5.3 therein) that $IC_\mu(\pi) = \mathbf{E}[\mathbf{D}(\mathcal{T}'_\pi)]$. We will prove that $\mathbf{E}[\mathbf{D}(\mathcal{T}'_\pi)] \leq \mathbf{E}[\mathbf{D}(\mathcal{T}_\pi)]$. By (9),

$$\begin{aligned}&\mathop{\mathbf{E}}_{X,Y,R_1,R_2}[\mathbf{D}(\mathcal{T}_\pi) - \mathbf{D}(\mathcal{T}'_\pi)]\\ &= \mathop{\mathbf{E}}_{X,Y,R_1,R_2}\left[\sum_v \tilde{p_v}\left(\mathbf{D}(P_v\|Q_v) - \mathbf{D}(P_v\|Q'_v)\right)\right],\end{aligned}$$

where $\tilde{p}_v$ is as in Definition 4. We separate the sum on the vertices to layers and work on each layer separately. Fix a layer d in the tree. Let L_d be the set of vertices in layer d. To simplify notation, let A denote (X, R_1), let B denote (Y, R_2), and let V denote Π_d. Then,

$$\begin{aligned}&\mathop{\mathbf{E}}_{X,Y,R_1,R_2}\left[\sum_{v\in L_d} \tilde{p_v}\left(\mathbf{D}(P_v\|Q_v) - \mathbf{D}(P_v\|Q'_v)\right)\right]\\ &= \mathop{\mathbf{E}}_{A,B,V}\left[\mathbf{D}(P_V\|Q_V) - \mathbf{D}(P_V\|Q'_V)\right].\end{aligned}$$

(Recall that V is undefined when the protocol ends before layer d. In that case, for simplicity, we think of P_V, Q_V and Q'_V as all being equal, and hence $\mathbf{D}(P_V\|Q_V) = \mathbf{D}(P_V\|Q'_V) = 0$). By the definition of relative entropy,

$$\begin{aligned}&\mathop{\mathbf{E}}_{A,B,V}\left[\mathbf{D}(P_V\|Q_V) - \mathbf{D}(P_V\|Q'_V)\right]\\ &= \mathop{\mathbf{E}}_{A,B,V}\Bigg[p_V\left(\log\left(\frac{p_V}{q_V}\right) - \log\left(\frac{p_V}{q'_V}\right)\right)\\ &\quad + (1-p_V)\left(\log\left(\frac{1-p_V}{1-q_V}\right) - \log\left(\frac{1-p_V}{1-q'_V}\right)\right)\Bigg]\\ &= \mathop{\mathbf{E}}_{A,B,V}\left[p_V\log\left(\frac{q'_V}{q_V}\right) + (1-p_V)\log\left(\frac{1-q'_V}{1-q_V}\right)\right].\end{aligned} \tag{10}$$

Assume that the first player owns the vertices in layer d. The case that the second player owns the vertices in layer d is analogous. Consider the first summand in (10). It holds that,

$$\begin{aligned}&\mathop{\mathbf{E}}_{A,B,V}\left[p_V\log\left(\frac{q'_V}{q_V}\right)\right]\\ &= \mathop{\mathbf{E}}_{B,V}\left[\mathop{\mathbf{E}}_{A}\left[\left(p_V\log\left(\frac{q'_V}{q_V}\right)\right)\middle| B,V\right]\right].\end{aligned}$$

By the definition of q'_V, for fixed B, V, it holds that $q'_V = \mathbf{E}_A\left[p_V|B,V\right]$. Since q'_V and q_V are functions of B and V, when we condition on B and V, q'_V and q_V are fixed. Therefore, conditioned on B and V, the term $\log\left(\frac{q'_V}{q_V}\right)$ is independent of A. We get that,

$$\begin{aligned}&\mathop{\mathbf{E}}_{B,V}\left[\mathop{\mathbf{E}}_{A}\left[\left(p_V\log\left(\frac{q'_V}{q_V}\right)\right)\middle| B,V\right]\right]\\ &= \mathop{\mathbf{E}}_{B,V}\left[\mathop{\mathbf{E}}_{A}\left[p_V|B,V\right]\log\left(\frac{q'_V}{q_V}\right)\right]\\ &= \mathop{\mathbf{E}}_{B,V}\left[q'_V\log\left(\frac{q'_V}{q_V}\right)\right].\end{aligned}$$

In the same way, we get that the second summand in (10) is

$$\begin{aligned}&\mathop{\mathbf{E}}_{A,B,V}\left[(1-p_V)\log\left(\frac{1-q'_V}{1-q_V}\right)\right]\\ &= \mathop{\mathbf{E}}_{B,V}\left[(1-q'_V)\log\left(\frac{1-q'_V}{1-q_V}\right)\right].\end{aligned}$$

Put together it holds that,

$$\begin{aligned}&\mathop{\mathbf{E}}_{A,B,V}\left[\mathbf{D}(P_V\|Q_V) - \mathbf{D}(P_V\|Q'_V)\right]\\ &= \mathop{\mathbf{E}}_{B,V}\left[\mathbf{D}(Q'_V\|Q_V)\right] \geq 0,\end{aligned}$$

since the divergence is non-negative. This is true for every layer d in the tree. Therefore, summing over all layers, we get that

$$\mathop{\mathbf{E}}_{A,B}[\mathbf{D}(\mathcal{T}'_\pi)] \leq \mathop{\mathbf{E}}_{A,B}[\mathbf{D}(\mathcal{T}_\pi)].$$

■

VII. Information Upper Bound

In this section we prove Theorem 2. Let $(x, y) \in \text{supp}(\mu)$ be an input pair to the bursting noise game. Consider the following protocol π' for the bursting noise game. Starting from the root until reaching a leaf, at every vertex v, if the first player owns v, she sends the bit x_v with probability 0.9, and the bit $1 - x_v$ with probability 0.1. Similarly, if the second player owns v, she sends the bit y_v with probability 0.9, and the bit $1 - y_v$ with probability 0.1. Both players continue to the child of v that is indicated by the communicated bit. When they reach a leaf they output that leaf. By the Chernoff bound, the probability that the players output a leaf that is not typical with respect to the noisy multi-layer is at most $2^{-\Omega(w)}$. That is, the error probability of π' is exponentially small in k.

The information cost of the protocol π' is too large. The reason is that if the protocol reaches a non-typical vertex at the end of the noisy multi-layer (with respect to the noisy multi-layer), an event that occurs with probability exponentially small in k, then the rest of the protocol reveals to each player $\Omega\left((c - i)w\right)$ bits of information about the input of the other player, in expectation (as all the vertices below a non-typical vertex are noisy), and note that $\Omega\left((c-i)w\right)$ is double exponentially large (for almost all i). Thus, in expectation, the information revealed to each player about the input of the other player is double exponential in k.

For that reason, we consider a variant of the protocol π', called π. Informally speaking, the protocol π operates like π' but aborts if too much information about the inputs is revealed. Recall that in every round of the protocol π', the players are at a vertex v of $\mathcal{T}$ and the player who owns v sends a bit b_v indicating one of v's children. In the new protocol π, after receiving that bit, the receiving party sends a bit a_v indicating whether they should abort the protocol, where $a_v = 1$ stands for abort and $a_v = 0$ stands for continue. If a bit $a_v = 1$, indicating an abort, was sent, the protocol terminates and both players output an arbitrary leaf of the tree $\mathcal{T}$. It remains to specify how the receiving party, without loss of generality the second player, decides whether to abort or continue, that is, how she determines the value of a_v.

To determine whether to abort, the second player considers the last $\ell = 2^{100k}$ vertices $v_1, \ldots, v_\ell$, reached by the protocol and owned by the first player, and the corresponding bits $b_{v_1}, \ldots, b_{v_\ell}$ that were sent by the first player (if less than ℓ bits were sent by the first player so far, then the second player does not abort). For every $j \in [\ell]$, the second player compares b_{v_j} and y_{v_j}. The second player decides to abort and sends $a_v = 1$ if and only if less than 0.8ℓ of these pairs are equal (otherwise the second player sends $a_v = 0$).

The following claim shows that the probability that π aborts is exponentially small in k. If π does not abort, it gives the same output as π'. We conclude that the error probability of π is exponentially small in k.

Claim 10: Let $(x, y) \in \operatorname{supp}(\mu)$ be an input pair to the bursting noise game. The protocol π aborts with probability at most 2^{-10k} on the input (x, y).

Proof: Fix $(x, y) \in \operatorname{supp}(\mu_i)$ for some $i \in [c]$. Let E be the event that the protocol π reaches a non-typical vertex after multi-layer i (with respect to multi-layer i). By the Chernoff bound, the event E occurs with probability at most 2^{-100k}, as $w = 2^{100}k$. Let A be the event that the protocol π aborts. Assume that E does not occur. By the Chernoff bound, the probability of aborting after each round is at most $2^{-2^{50k}}$, as $\ell = 2^{100k}$ and since if E does not occur then x_v and y_v can only differ for at most w vertices reached by the protocol π. By the union bound, the probability of abort (conditioned on $\neg E$) is at most $cw \cdot 2^{-2^{50k}} < 2^{-100k}$. Therefore, $\Pr[A] \leq \Pr[E] + \Pr[A|\neg E] \leq 2 \cdot 2^{-100k}$. ■

To upper bound the information cost of the protocol π we will use Lemma 9. We denote by $\mathcal{T}_\pi$ the binary tree associated with the communication protocol π, as in Section VI. That is, every vertex v of $\mathcal{T}_\pi$ corresponds to a possible transcript of π, and the two edges going out of v are labeled by 0 and 1, corresponding to the next bit to be transmitted. The non-leaf vertices of the tree $\mathcal{T}_\pi$ have the following structure: Every non-leaf vertex v in an odd layer of $\mathcal{T}_\pi$ corresponds to a non-leaf vertex of $\mathcal{T}$, the binary tree on which the bursting noise game is played. Since the correspondence is one-to-one, we refer to the vertex in $\mathcal{T}$ corresponding to v also as v. The next bit to be transmitted by π on the vertex v is b_v. For a non-leaf vertex v in an even layer of $\mathcal{T}_\pi$, the next bit to be transmitted by π on the vertex v is a_v.

As explained in Section VI, every input pair $(x, y) \in \operatorname{supp}(\mu)$ to the bursting noise game, induces a distribution $P_v = (p_v, 1-p_v)$ for every non-leaf vertex v of the tree $\mathcal{T}_\pi$, where p_v is the probability that the next bit transmitted by the protocol π on the vertex v and inputs x, y is 0. Namely, if v is in an odd layer of $\mathcal{T}_\pi$ (and recall that in this case we think of v as both a vertex of $\mathcal{T}_\pi$ and of $\mathcal{T}$), the distribution P_v is the following: In the case that the first player owns v in $\mathcal{T}$, if $x_v = 0$ then $P_v = (0.9, 0.1)$, and if $x_v = 1$ then $P_v = (0.1, 0.9)$. In the case that the second player owns v, if $y_v = 0$ then $P_v = (0.9, 0.1)$, and if $y_v = 1$ then $P_v = (0.1, 0.9)$. If v is in an even layer of $\mathcal{T}_\pi$ then P_v is $P_v = (0, 1)$ if the player sending a_v decides to abort, and $P_v = (1, 0)$ if she decides to continue (note that given x, y, v, this decision is deterministic).

For every non-leaf vertex v of $\mathcal{T}_\pi$, we define an additional distribution $Q_v = (q_v, 1 - q_v)$ (depending on the input (x, y)). We think of every P_v as the "correct" distribution over the two children of v. This distribution is known to the player who sends the next bit on the vertex v. We think of Q_v as an estimation of P_v, based on the knowledge of the player who doesn't send the next bit. For a vertex v in an odd layer of $\mathcal{T}_\pi$ (and recall that in this case we think of v as both a vertex of $\mathcal{T}_\pi$ and of $\mathcal{T}$), the distribution Q_v is the following: In the case that the first player owns v in $\mathcal{T}$, if $y_v = 0$ then $Q_v = (0.9, 0.1)$, and if $y_v = 1$ then $Q_v = (0.1, 0.9)$. In the case that the second player owns v, if $x_v = 0$ then $Q_v = (0.9, 0.1)$, and if $x_v = 1$ then $Q_v = (0.1, 0.9)$. If v is in an even layer of $\mathcal{T}_\pi$ then $Q_v = (1 - \frac{1}{cw}, \frac{1}{cw})$.

For the rest of the section, we think of $\mathcal{T}_\pi$ as the tree $\mathcal{T}_\pi$ together with the distributions P_v and Q_v, for every vertex v in the tree $\mathcal{T}_\pi$.

Proposition 11: It holds that

$$\mathbf{D}(\mathcal{T}_\pi) = O(k).$$

Proof: Fix $(x, y) \in \operatorname{supp}(\mu_i)$ for some $i \in [c]$. By (9),

$$\mathbf{D}(\mathcal{T}_\pi) = \sum_v \tilde{p}_v \cdot \mathbf{D}(P_v \| Q_v),$$

where $\tilde{p}_v$ is the probability that the protocol π reaches the vertex v on input (x, y). We will bound the last sum separately for vertices v in odd layers and for vertices v in even layers.

We first sum over vertices in even layers. For every vertex v in an even layer of $\mathcal{T}_\pi$, if $P_v = (0, 1)$ (protocol aborts) we have $\mathbf{D}(P_v \| Q_v) = \log(cw)$, and if $P_v = (1, 0)$ (protocol continues) we have $\mathbf{D}(P_v \| Q_v) = \log\left(\frac{1}{1-1/cw}\right) = \log\left(1 + \frac{1}{cw-1}\right) < \frac{2}{cw}$. By Claim 10, the probability that π aborts is at most 2^{-10k}. Therefore, the sum in (9) taken over vertices in even layers is at most $cw \cdot \frac{2}{cw} + 2^{-10k} \cdot \log(cw) \leq 3$, as for each of the cw even layers, the probability of reaching a vertex in this layer is at most 1.

We next sum over vertices in odd layers. Recall that each such vertex corresponds to a vertex in $\mathcal{T}$. Let v be a vertex in an odd layer of $\mathcal{T}_\pi$. If v corresponds to a non-noisy vertex in $\mathcal{T}$ we have $\mathbf{D}(P_v \| Q_v) = 0$, and if v corresponds to a noisy vertex in $\mathcal{T}$ we have $\mathbf{D}(P_v \| Q_v) \leq 4$. Recall that i is the noisy multi-layer. Then,

1) The vertices above multi-layer i in $\mathcal{T}$ add nothing to the divergence cost.
2) Multi-layer i of $\mathcal{T}$ adds $O(w)$ to the divergence cost.
3) If $i < c$: Let v be the vertex in layer $i^* + w$ of $\mathcal{T}$ that the players reach during the execution of the protocol π. If v is a typical vertex with respect to multi-layer i, the vertices below v add nothing to the divergence cost. If v is a non-typical vertex, the protocol aborts after at most 4ℓ rounds in expectation. Since the probability that v is a non-typical vertex with respect to multi-layer i is at most 2^{-1000k} (as $w = 2^{100}k$), the expected divergence cost added by this case is at most $2^{-1000k} \cdot 4\ell \cdot 4 \leq 1$.

Together, the total divergence cost is $O(w) = O(k)$, as claimed. ■

By Proposition 11 and Lemma 9 we get that $IC_\mu(\pi) \leq O(k)$.

Acknowledgements

We thank Andy Drucker for a suggestion that lead to a substantial simplification of our original proof.

References

[1] M. Braverman, "Interactive information complexity," in *STOC*, 2012, pp. 505–524.

[2] M. Braverman and A. Rao, "Information equals amortized communication," in *FOCS*, 2011, pp. 748–757.

[3] E. Kushilevitz and N. Nisan, "Communication complexity," *Cambridge University Press*, 1997.

[4] T. Lee and A. Shraibman, "Lower bounds in communication complexity," *Foundations and Trends in Theoretical Computer Science*, vol. 3, no. 4, pp. 263–398, 2009.

[5] A. Chakrabarti, Y. Shi, A. Wirth, and A. C.-C. Yao, "Informational complexity and the direct sum problem for simultaneous message complexity," in *FOCS*, 2001, pp. 270–278.

[6] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar, "An information statistics approach to data stream and communication complexity," *J. Comput. Syst. Sci.*, vol. 68, no. 4, pp. 702–732, 2004.

[7] B. Barak, M. Braverman, X. Chen, and A. Rao, "How to compress interactive communication," in *STOC*, 2010, pp. 67–76.

[8] T. Feder, E. Kushilevitz, M. Naor, and N. Nisan, "Amortized communication complexity," *SIAM J. Comput.*, vol. 24, no. 4, pp. 736–750, 1995.

[9] R. Jain, J. Radhakrishnan, and P. Sen, "A direct sum theorem in communication complexity via message compression," in *ICALP*, 2003, pp. 300–315.

[10] P. Harsha, R. Jain, D. A. McAllester, and J. Radhakrishnan, "The communication complexity of correlation," in *IEEE Conference on Computational Complexity*, 2007, pp. 10–23.

[11] H. Klauck, "A strong direct product theorem for disjointness," in *STOC*, 2010, pp. 77–86.

[12] R. Jain, "New strong direct product results in communication complexity," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 18, p. 24, 2011.

[13] R. Jain, A. Pereszlényi, and P. Yao, "A direct product theorem for the two-party bounded-round public-coin communication complexity," in *FOCS*, 2012, pp. 167–176.

[14] M. Braverman, A. Rao, O. Weinstein, and A. Yehudayoff, "Direct products in communication complexity," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 19, p. 143, 2012.

[15] ——, "Direct product via round-preserving compression," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 20, p. 35, 2013.

[16] M. Braverman and O. Weinstein, "A discrepancy lower bound for information complexity," in *APPROX-RANDOM*, 2012, pp. 459–470.

[17] I. Kerenidis, S. Laplante, V. Lerays, J. Roland, and D. Xiao, "Lower bounds on information complexity via zero-communication protocols and applications," in *FOCS*, 2012, pp. 500–509.

[18] M. Braverman, "A hard-to-compress interactive task?" *In 51th Annual Allerton Conference on Communication, Control, and Computing*, 2013.

[19] ——, "Coding for interactive computation: progress and challenges," *In 50th Annual Allerton Conference on Communication, Control, and Computing*, 2012.

[20] F. R. K. Chung, R. L. Graham, P. Frankl, and J. B. Shearer, "Some intersection theorems for ordered sets and graphs," *J. Comb. Theory, Ser. A*, vol. 43, no. 1, pp. 23–37, 1986.

[21] J. Kahn, "An entropy approach to the hard-core model on bipartite graphs," *Combinatorics, Probability and Computing*, vol. 10, pp. 219–237, 5 2001.

[22] J. Radhakrishnan, "Entropy and counting," *IIT Kharagpur Golden Jubilee Volume*, p. 125, 2003.

[23] M. M. Madiman and P. Tetali, "Information inequalities for joint distributions, with interpretations and applications," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2699–2713, 2010.