New Separations Results for External Information

Mark Braverman*  
mbraverm@gmail.com  
Department of Computer Science, Princeton University  
Princeton, NJ, USA

Dor Minzer  
minzer.dor@gmail.com  
Massachusetts Institute of Technology  
Cambridge, MA, USA

ABSTRACT
We obtain new separation results for the two-party external information complexity of Boolean functions. The external information complexity of a function \( f(x, y) \) is the minimum amount of information a two-party protocol computing \( f \) must reveal to an outside observer about the input. We prove an exponential separation between external and internal information complexity, which is the best possible; previously no separation was known. We use this result in order to then prove a near-quadratic separation between amortized zero-error communication complexity and external information complexity for total functions, disproving a conjecture of the first author. Finally, we prove a matching upper bound showing that our separation result is tight.

CCS CONCEPTS
• Theory of computation → Communication complexity.

KEYWORDS
Communication Complexity, Information Complexity

ACM Reference Format:
Mark Braverman and Dor Minzer. 2021. New Separations Results for External Information. In Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC ’21), June 21–25, 2021, Virtual, Italy. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3406325.3451044

1 INTRODUCTION
The main object of study in this paper is the external two-party information complexity of problems. For a two-party communication protocol \( \pi(x, y) \), with inputs distributed according to some \((x, y) \sim \mu\), one can define the following complexity measures:¹

- the (average case) communication complexity \( CC_\mu[\pi] \) of \( \pi \) is the expected number of bits exchanged in \( \pi \);
- the external information cost \( IC_\mu^\text{external}[\pi] \) of \( \pi \) is the expected amount of information an external observer learns about the inputs \((x, y)\) by observing an execution of \( \pi(x, y) \);

¹Research supported in part by the NSF Alan T. Waterman Award, Grant No. 1933331, a Packard Fellowship in Science and Engineering, and the Simons Collaboration on Algorithms and Geometry.

²See Section 2 for rigorous definitions and further background.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

STOC ’21, June 21–25, 2021, Virtual, Italy  
© 2021 Association for Computing Machinery.  
ACM ISBN 978-1-4503-8053-9/21/06 ...$15.00  
https://doi.org/10.1145/3406325.3451044

Within the theoretical computer science literature, notions of information complexity had been introduced at least twice. It was introduced once in the context of information-theoretic security [4], where a low \( IC_\mu^\text{internal}[f, \epsilon] \) would mean that information theoretically secure two-party computation is possible (it turn out to be impossible in most cases). It was introduced in a different set of works, starting with the use of external information cost, in [5, 12], in the context of proving communication complexity lower bounds by using information-theoretic reasoning to prove an information complexity lower bound, and then using (1) to deduce a communication complexity lower bound. More recent surveys on information complexity can be found in [8, 19].

Starting with the works of Shannon in the 1940s, the main motivation for using information-theoretic quantities is that they tensorize². For example, if \( X_1 \) and \( X_2 \) are two independent random variables, then their Shannon’s entropy satisfies \( H(X_1X_2) = H(X_1) + H(X_2) \). Turns out that internal information complexity satisfies a similar property (and, thus, is arguably the correct information-theoretic version of two-party communication complexity). For simplicity, denote by \((f^q, \epsilon)\) the task of computing \( q \) independent copies of \( f \), each with error \( \epsilon \), then

\[ IC_\mu^\text{internal}[f^q, \epsilon] = q \cdot IC_\mu^\text{internal}[f, \epsilon]. \]
Such a relationship was known to be false for CC, and was believed to also not hold for ICexternal, although to the best of our knowledge only in the present paper we rule it out for all values of \( \varepsilon \) including \( \varepsilon = 0 \).

Equations (1) and (2) together provide a blueprint for proving communication lower bound on computing multiple copies of a function:

\[
CC_\mu[f^q, \varepsilon] \geq IC^\text{internal}_\mu[f^q, \varepsilon] = q \cdot IC^\text{internal}_\mu[f, \varepsilon].
\]  

(3)

Thus, an information complexity lower bound on \( f \), implies a communication lower bound on multiple copies of \( f \). Moreover, a slight twist on (3) allows one to use similar reasoning to prove lower bounds on e.g. an OR of \( q \) copies of \( f \), to obtain tight bounds on the communication complexity of functions such as Disjointness [5, 9].

It turns out that, in fact, for \( \varepsilon > 0 \) (3) is tight [10]. For all \( f \) and \( \varepsilon > 0 \) the following holds:

\[
\lim_{q \to \infty} CC_\mu[f^q, \varepsilon]/q = IC^\text{internal}_\mu[f, \varepsilon].
\]  

(4)

Equation (4) has given rise to two questions:

- **What is the relationship between IC\( \text{internal}_\mu[f, \varepsilon] \) and CC\( _\mu[f, \varepsilon] \)?** How large can the gap in (1) be? This question is sometimes called the “interactive compression question”, and is equivalent to the direct sum question for two-party communication complexity.

- **What happens in (4) when \( \varepsilon = 0 \)—that is, when no error is allowed?** Note that, not coincidentally, in other communication settings allowing error drastically alters the communication complexity of problems. For example, the communication cost of EQ\( _\mu(x, y) \)—the problem of determining whether two \( n \)-bit strings are equal is \( \Omega(\log 1/\varepsilon) \) independent of \( n \) when error \( \varepsilon \) is allowed, but increases to \( n + 1 \) when no error is allowed.

**Separation between Information and Communication.** The first question was answered by Ganor, Kol, and Raz [15], who showed an exponential separation between internal information complexity and communication complexity. Moreover, such separation is the best possible [8]. In other words, there is a boolean function \( f \) such that IC\( \text{internal}_\mu[f, \varepsilon] = O(k) \), while CC\( _\mu[f, \varepsilon] = \Omega(2^k) \). In addition, exponential separation was shown between external information and communication complexity — at least for tasks [16].

Therefore, in the context of (1), at least for tasks, the second inequality was known to be strict (with the maximal possible exponential separation). A separation in the first inequality had been strongly suspected but never proven. In this paper (Theorem 1.2) we show that, in fact, the example from [15] has an exponential external information complexity (and not just exponential communication complexity), and thus gives an example of an \( f \) such that

\[
IC^\text{external}_\mu[f, \varepsilon] \geq 2^{\Omega(IC^\text{internal}_\mu[f, \varepsilon])}.
\]  

(5)

As will be discussed later, we need (5) in order to separate external information from zero-error amortized communication.

**Zero-error Amortized Communication.** A second mystery that remains in the wake of (4) is what happens with zero-error amortized communication? In other words, what can we say about the quantity:

\[
\lim_{q \to \infty} CC_\mu[f^q, 0]/q
\]  

(6)

A canonical example of a function where zero-error and vanishing-error communication costs diverge is the \( n \)-bit Equality function EQ\( _\mu(x, y) := 1_{x = y} \). It is known that the amortized communication complexity of EQ\( _n \) is \( O(1) \) [14]. As a consequence of this result (which can also be seen directly [8]), one gets for all \( \mu \),

\[
IC^\text{external}_\mu[EQ_n, 0] = O(1).
\]

(7)

On the other hand, it is not hard to see using fooling sets, that

\[
\lim_{q \to \infty} CC_\mu[EQ^q_n, 0]/q = \Omega(n),
\]

where \( \mu \) is the distribution \( \mu = \frac{1}{2}U_{(x, x)} + \frac{1}{2}U_{(x, y)} \) — a mixture of the uniform distribution and the uniform distribution on EQ\( _n^{-1}(1) \). Therefore (4) has no chance of holding when \( \varepsilon = 0 \). More precisely, half of the proof of (4) holds for \( \varepsilon = 0 \), yielding

\[
\lim_{q \to \infty} CC_\mu[f^q, 0]/q \geq IC^\text{internal}_\mu[f, 0].
\]  

(7)

but this inequality may be strict, as is indeed the case for \( f = EQ_n \).

An attempt to prove the \( \ll \) direction in (7) would involve trying to compress a low-internal information protocol for \( f^q \) into a low-communication one. Such compression procedures exist [10], but they inherently introduce errors (where with a tiny probability the message received doesn’t match the message sent). In contrast to internal information, there is a zero-error compression protocol for external information [17] leading to a variant of a converse to (7) where internal information is replaced with external information:

\[
\lim_{q \to \infty} CC_\mu[f^q, 0]/q \leq IC^\text{external}_\mu[f, 0].
\]  

(8)

We formally prove (8) for completeness purposes in Section ?? (Theorem ??). Inequality (8), along with the fact that inequality (7) is easily seen to not be tight led to the following conjecture [7]:

**Conjecture:** \( \lim_{q \to \infty} CC_\mu[f^q, 0]/q = \Theta(IC^\text{external}_\mu[f, 0]). \)  

(9)

There were several reasons to believe this conjecture. Translated to this language, a result of Ahlswede and Cai [2] shows that (9) holds (with a constant 1) when \( f \) is the 2-bit AND function for the hardest distribution \( \mu \) — the quantity on both sides is \( \log_2 3 \). A version of (9) in fact holds for one-sided non-deterministic communication/information, which we prove in Section ?? for completeness. Here the 1-superscript in \( IC^\text{external}_\mu \), stands for the complexity of proving that the value of \( f(x, y) = 1 \).

**Theorem 1.1.** Let \( \mu \) be a distribution with \( \text{supp}(\mu) \subseteq f^{-1}(1) \), then

\[
IC^\text{external}_\mu[f, 0] \leq \lim_{q \to \infty} CC^{\text{eq}1^\mu}_\mu[f^q, 0]/q.
\]

We should note that a quadratic upper bound on IC\( ^\text{external}_\mu[f, \varepsilon] \) in terms of amortized zero-error communication complexity (Theorem 1.5) does hold. Informally, this upper bound can be thought of as a consequence of Theorem 1.1 (along its co-nondeterministic counterpart) similarly to the \( D(f) \leq N^0(f) \cdot N^1(f) \) bound on deterministic communication complexity in terms of non-deterministic communication complexity.

---

4Here, zero-error means that the protocol has to always output the correct value of \( f(x, y) \), even if \( \mu(x, y) = 0 \).
It should be noted that in Conjecture (9) it is important that the zero-error of communication holds for all potential inputs to \( f \) (even when \( \mu \) is not full-support). In other words, correctness shouldn’t be predicated on a “promise” about the inputs \((x, y)\). In the promise setting, a counterexample has been constructed by Kol, Moran, Shpilka, and Yehudayoff [18]. In the context of the counterexample, Theorem 1.1 also doesn’t hold, which suggests a large gap between the promise and non-promise regimes.

Our main contribution is to disprove Conjecture (9). In light of (7), a prerequisite for disproving the conjecture is being able to separate internal information complexity from external information complexity along the lines of (5).

1.1 Main Results

We now state our main results formally. First, as alluded to before, we show an exponential separation between internal information and external information.

**Theorem 1.2.** For all \( \varepsilon > 0 \), for large enough \( k \), there is \( n \in \mathbb{N} \), a function \( f : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\} \) and an input distribution \( \mu \) satisfying:

1. \( \text{IC}_{\mu}^{\text{internal}}(f, \varepsilon) \leq O(k) \),
2. \( \text{IC}_{\mu}^{\text{external}}(f, \varepsilon) \geq \Omega(k) \).

Secondly, using Theorem 1.2 we disprove Conjecture 9. We show that even if one considers the external information of \( f \) for protocols with constant error, a near-quadratic gap between it and the amortized zero-error communication complexity is still possible:

**Theorem 1.3.** For large enough \( k \), there is \( n \in \mathbb{N} \), a function \( f : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\} \) and an input distribution \( \mu \) such that

1. \( \lim_{q \to \infty} \frac{1}{q} \text{CC}_{\mu \nu}(f^q, 0) \leq O(\sqrt{k} \log^2 k) \),
2. \( \text{IC}_{\mu}^{\text{external}}(f, 1/16) \geq \Omega(k) \).

If one insists on external information of protocols with zero-error, a much stronger separation result holds (and in fact quickly follows from Theorem 1.3):

**Corollary 1.4.** For large enough \( m \), there is \( n \in \mathbb{N} \), a function \( f : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\} \) and an input distribution \( \nu \) such that

1. \( \lim_{q \to \infty} \frac{1}{q} \text{CC}_{\nu \nu}(f^q, 0) \leq O(1) \),
2. \( \text{IC}_{\nu}^{\text{external}}(f, 0) \geq \Omega(m) \).

It is worth noting that the separation given in Theorem 1.3 is nearly tight, and in general the external information with constant error is at most the square of the amortized zero-error communication complexity:

**Theorem 1.5.** There exists an absolute constant \( C > 0 \), such that for any \( \varepsilon > 0 \), \( f : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\} \) and an input distribution \( \mu \), we have that

\[
\text{IC}_{\mu}^{\text{external}}(f, \varepsilon) \leq C \varepsilon^2 \left( \lim_{q \to \infty} \frac{1}{q} \text{CC}_{\mu \nu}(f^q, 0) \right)^2.
\]

1.2 Proof Overview and Discussion

**Separating External and Internal Information.** As discussed earlier, having a problem with a low internal information complexity but a high external information complexity appears essential for our main separation result. We actually prove that the Bursting Noise Function, which was introduced by [15] to separate internal information complexity from randomized communication complexity also separates internal information complexity from external information complexity. To that end, we extend the reach of the relative-discrepancy lower bound technique from [15] to apply to external information complexity.

We do this by introducing a property of protocols having “universally low external information”. We then show that (1) low external information cost protocols can be approximated with universally low external information protocols; and (2) universally low external information protocols (just like low-communication protocols in [15]) do very poorly when trying to compute functions with the appropriate relative discrepancy property.
With this intuition in mind, the starting point of our construction is a function \( h : \{0,1\}^m \to \{0,1\} \) and its AND-lifting \( h_{\ast} : \{0,1\}^{m \times m} \to \{0,1\} \) defined as \( h_{\ast}(x,y) = h(x \land y) \), with the following properties: (1) the external information of \( h_{\ast} \) is \( \Omega(m) \); (2) one can certify that \( h(z) = 0 \) or that \( h(z) = 1 \) using only \( C = \Theta(\sqrt{m}) \) bits. We then wish to construct a function \( H \) on 4-tuples, \((x,y,u,v)\), whose output on \((x,y,u,v)\) is \( h_{\ast}(x,y) \), and \((u,v)\) encodes a certificate for that on the support of our input distribution. Here, by "encodes" we mean that \((u,v)\) could be viewed as a sequence of \( C \) input tuples to \( f \), and that these bits encode a certificate to \( h_{\ast}(x,y) \). Thus, for the purposes of amortized communication complexity, we use the back-door \((u,v)\) and only have to pay communication proportional to the internal information of \( f \) (after retrieving the hint \( f(u,v) \) we still need to use \( O(C) \) communication to verify its veracity to ensure the final answer is never wrong). On the other hand, \( f(u,v) \) will look almost random to a low external information protocols, and hence is essentially useless for them, so the external information of our protocol must be the external information of \( h_{\ast} \), i.e. at least \( \Omega(m) \).

Discussion. Of the two questions raised in the beginning of Section 1, in this paper we have given the optimal separation between internal and external information complexity. On the other hand, the mystery of understanding amortized zero-error communication complexity has only deepened. We now know that it is different from external information complexity, and that the worst possible gap is in some sense quadratic.

This leaves the question of characterizing zero-error amortized communication complexity wide open.

Open Problem 1.6. Characterize the amortized zero-error communication complexity of functions. For simplicity, suppose \( \mu \) has full support. Characterize:

\[
\text{ZAC}(f, \mu) := \lim_{q \to \infty} \text{CC}_{\mu^q}(f^q, 0)/q.
\]

where \( \text{CC} \) is average-case distributional communication complexity with zero error.

Ideally, the characterization would be in terms of information-theoretic quantities pertaining to computing a single copy of \( f \). Such a characterization is sometimes called "single letter" characterization in the information theory literature. It is likely that understanding this quantity will lead to further communication complexity insights.

Organization. In Section 2, we recall some standard notions and tools that will be needed in our proofs. We prove Theorem 1.2 in Section 3. The proof of Theorems 1.3, 1.5 and Corollary 1.4 is deferred to the full version of the paper.

2 PRELIMINARIES

2.1 Information Theory

We begin with a few basic definitions from information theory. Throughout the paper, we only consider random variables with finite support.

Definition 2.1. Let \( X, Y \) be random variables with a finite support.

(1) The Shannon entropy of \( X \) is

\[
H[X] = \sum_x \Pr[X = x] \log \left( \frac{1}{\Pr[X = x]} \right).
\]

(2) The Shannon entropy of \( X \) conditioned on \( Y \) is \( H[X \mid Y] = \mathbb{E}_{y \sim Y} [H[X \mid Y = y]] \), where

\[
H[X \mid Y = y] = \sum_x \Pr[X = x \mid Y = y] \log \left( \frac{1}{\Pr[X = x \mid Y = y]} \right).
\]

Definition 2.2. Let \( X, Y, Z \) be random variables with a finite support.

(1) The mutual information between \( X \) and \( Y \) is \( I[X;Y] = H[X] - H[X \mid Y] \).

(2) The mutual information between \( X \) and \( Y \) conditioned on \( Z \) is \( I[X;Y \mid Z] = H[X \mid Z] - H[X \mid Y, Z] \).

Definition 2.3. Let \( X, Y \) be random variables with a finite support. The KL-divergence from \( Y \) to \( X \) is

\[
D_{KL}(X \mid Y) = \sum_{x,y} \Pr[X = x] \log \left( \frac{\Pr[X = x]}{\Pr[Y = y]} \right).
\]

We will need the following standard facts from information theory (for proofs, see [13] for example).

Fact 2.4. Let \( X, Y, Z \) be random variables, and let \( P_{X,Y} \) be the joint distribution of \( X, Y \). Then

\[
I[X;Y;Z] = \mathbb{E}_{(x,y,z) \sim P_{X,Y,Z}} \left[ D_{KL}(Z \mid X=x, Y=y) \mid Z \right].
\]

Fact 2.5. Let \( X, Y_1, \ldots, Y_n \) be random variables. Then

\[
I[X;Y_1,\ldots,Y_n] = \sum_{i=1}^n I[X;Y_i \mid Y_{<i}].
\]

Fact 2.6. Let \( X, Y, Z \) be random variables. Then \( I[X;Y \mid Z] \leq I[X;Y;Z] \).

For \( p \in [0,1] \), we denote by \( B(p) \) a Bernoulli random variable with parameter \( p \).

Fact 2.7. Let \( p, q \in [0,1] \) and suppose that \( \frac{1}{2} \leq q \leq \frac{2}{3} \). Then

\[
2(p - q)^2 \leq D_{KL}(B(p) \mid B(q)) = \frac{9(q-p)^2}{2 \ln 2}.
\]

2.2 Communication Complexity

Let \( f : X \times Y \to \{0,1\} \) be a function, let \( \mu \) be a distribution over its inputs and denote \( (X,Y) \sim \mu \). Throughout, we denote by \( \Pi \) a two-player communication protocol, and by \( \Pi(X,Y) \) the distribution over transcripts of the protocol where the inputs are sampled according to the random variables \((X,Y)\). We denote the output of a specific transcript \( \pi \) by \( output(\pi) \). Abusing notations, for inputs \( x, y \), we denote by \( output(\Pi(x,y)) \) the random variable which is the output of \( \Pi \) when run on inputs \( x,y \).

Definition 2.8. The internal information of the protocol \( \Pi \) is defined as \( I_{\mu}^{\text{internal}}[\Pi] = I[\Pi;X|Y] + I[\Pi;Y|X] \).

For an error parameter \( 0 \leq \epsilon < \frac{1}{2} \), we define the internal information cost of \( f \) on \( \mu \) with error \( \epsilon \) by

\[
I_{\mu}^{\text{internal}}[f, \epsilon] = \inf_{\Pi:Pr[(x,y) \sim \mu]\left[output(\Pi(x,y)) \right] \leq \epsilon} I_{\mu}^{\text{internal}}[\Pi].
\]
The proof is now concluded by noting that
\[ \sum_{j=0}^{\infty} P(x) \log \frac{Q(x)}{P(x)} \leq \sum_{j=0}^{\infty} \sum_{x \in A_j} 2^{-j} Q(x)(j+1) \]
\[ \leq \sum_{j=0}^{\infty} 2^{-j}(j+1) = 4. \]

**2.3 Probability**

We will need the following immediate corollary of Doob’s martingale inequality.

**Fact 2.14.** Suppose \((X_i)_{i=1,\ldots,m}\) is a martingale and \(\mathbb{E}[X_m^2] \leq M\). Then for every \(\epsilon > 0\),
\[ \Pr \left[ \exists \ i \text{ such that } |X_i| \geq \sqrt{M/\epsilon} \right] \leq \epsilon. \]

**Proof.** By Doob’s martingale inequality,
\[ \Pr \left[ \max_i |X_i| \geq \sqrt{M/\epsilon} \right] \leq \mathbb{E} \left[ \frac{|X_m|}{M/\epsilon} \right] \leq \sqrt{\frac{\mathbb{E}[X_m^2]}{M/\epsilon}} \leq \epsilon. \]

**Definition 2.15.** Let \(X, Y\) be random variables over the same universe \(U\). The statistical distance between \(X, Y\) is
\[ SD(X, Y) = \frac{1}{2} \sum_{u \in U} |\Pr[X = u] - \Pr[Y = u]|. \]

**Fact 2.16.** Let \(X, Y\) be discrete random variables over the same universe \(U\). Then there is \(A \subseteq U\) such that \(SD(X, Y) = 1 - \sum_{x \in A} \Pr[X = x] - \sum_{x \not\in A} \Pr[Y = x]\).

**3 SEPARATING INTERNAL AND EXTERNAL INFORMATION COST**

In this section, we prove Theorem 1.2. A key notion of our proof will be the relative-discrepancy measure, introduced in [15].

**Definition 3.1.** For a function \(f: \{0,1\}^n \times \{0,1\}^m \rightarrow \{0,1\}\) and a distribution \(\mu\) over \(\{0,1\}^n \times \{0,1\}^m\), we say \((f, \mu)\) has \((\epsilon, \delta)\) relative-discrepancy with respect to a distribution \(\rho\), if for any rectangle \(R = A \times B \subseteq \{0,1\}^n \times \{0,1\}^m\) for which \(\rho(R) \geq \delta\), it holds that
\[ (1) \mu(R \cap f^{-1}(0)) \geq \left( \frac{1}{2} - \epsilon \right) \rho(R), \]
\[ (2) \mu(R \cap f^{-1}(1)) \leq \left( \frac{1}{2} + \epsilon - \delta \right) \rho(R). \]

**Definition 3.2.** For a function \(f: \{0,1\}^n \times \{0,1\}^n \rightarrow \{0,1\}\) and a distribution \(\mu\) over \(\{0,1\}^n \times \{0,1\}^n\), we say \((f, \mu)\) has \((\epsilon, \delta)\) relative-discrepancy if there is a distribution \(\rho\) such that \((f, \mu)\) has \((\epsilon, \delta)\) relative-discrepancy with respect to \(\rho\).

In [15], the authors show that if \((f, \mu)\) has strong relative discrepancy, then \(CC_{\mu}(f, 1/2 - \epsilon')\) must be high. More precisely, they show that if \((f, \mu)\) has \((\epsilon, \delta)\) relative-discrepancy, then any protocol for \(f\) on the distribution \(\mu\) that achieves advantage of \(\epsilon'\) must communicate at least \(\log \left( \frac{\epsilon' - \epsilon}{\delta} \right)\) bits. The main result of this section strengthens this assertion, as follows.

**Theorem 3.3.** Let \(M \in \mathbb{N}, \delta, \epsilon > 0\) and let \(\mu\) be a distribution over \(\{0,1\}^n \times \{0,1\}^m\). Suppose \(f: \{0,1\}^n \times \{0,1\}^m \rightarrow \{0,1\}\) such that \((f, \mu)\) has \((\epsilon, \delta)\) relative-discrepancy, and \(\Pi\) is a protocol for computing \(f\) such that \(\text{I.ext}[\Pi] \leq M\), then
\[ \Pr_{(x,y) \sim \mu}[\Pi(x, y) = f(x, y)] \leq \frac{1}{2} + 2000 \max \left( \epsilon, \frac{M}{\log(1/\delta)} \right). \]

Contrapositively, if \((f, \mu)\) has \((\epsilon, \delta)\) relative-discrepancy and \(\Pi\) is a protocol for \((f, \mu)\) achieving advantage \(\epsilon'\), then the external information of \(\Pi\) according to \(\mu\) is at least \((\epsilon' - 2000\epsilon)\log(1/\delta)\).
Therefore, the relative-discrepancy measure allows us to prove lower bounds on the external information cost of a function, which is always smaller than the communication complexity of a function.

To prove Theorem 1.2, we instantiate Theorem 3.3 with the “bursting noise function” from [15], which we present next.

**The Bursting Noise Function.** Let $k \in \mathbb{N}$ be thought of as large, and set $c = 2^{2^k}$. The bursting noise function, $f_{\text{burst}}$ is a pointer chasing function on a tree of height $c$, however the input distribution $\mu$ is supported only on $x, y$ that are very correlated. More precisely, for $b \in \{0, 1\}$, we define the distribution $\mu_b$ according to the following sampling procedure: we think of a complete binary tree of depth $c$, and for each vertex in the tree, each player has a bit in their input. We think of vertices from odd layers as being owned by Alice, and vertices from even layers as being owned by Bob. Partition the layers of the tree into $c/k$ multi-layers (a multi-layer consists of $k$ consecutive layers), and sample $i \in \{1, \ldots, c/k\}$ uniformly. For each multi-layer $j < i$, and for each vertex $u$ in multi-layer $j$, we choose $y_u \in \{0, 1\}$ uniformly, and set $x_u = b \oplus y_u$. In layer $i$, for each vertex $v$ in it, we choose $x_v, y_v \in \{0, 1\}$ independently and uniformly.

Next, we define the notion of a typical vertex. We say a vertex $p$ from layer $i$ is typical, if considering the part of the path from the root to $p$ that is inside multi-layer $i$, on at least $80\%$ of it, on at least $80\%$ of the odd locations on that path it agrees with $x$, and on at least $80\%$ of the even locations of the path it agrees with $y$.

For the rest of the layers, for each vertex $u$, let $\rho(u)$ denote the ancestor of $u$ from layer $i$. If $\rho(u)$ is typical, we again take the bits to be uniform such as $x_u = b \oplus y_u$, and if $\rho(u)$ is atypical we take $x_u, y_u$ as independently chosen bits. For $(x, y) \in \text{supp}(\mu_b)$, we define $f_{\text{burst}}(x, y) = b$. We take $\mu = \frac{1}{2} j_0 + \frac{1}{2} \mu_1$. We remark that each one of $x$ and $y$ are $n$-bit Boolean strings where $n = \Theta(2^k) = \Theta(2^{2^k})$.

In [15], the following two important properties are proved for the bursting noise function.

**Lemma 3.4.** $IC^\mu_{\text{int}}[f, 2^{-k}] = O(k)$.

**Lemma 3.5.** The pair $(f, \mu)$ has the $(\epsilon, \delta)$ relative-discrepancy property with respect to $\rho$ for $\epsilon = 2^{-k}$ and $\delta = \epsilon/2^{2^k}$.

First, we quickly show that Theorem 1.2 follows from Theorem 3.3 and the above lemmas.

**Proof of Theorem 1.2.** Fix $\eta > 0$ a small constant, and choose $k$ large enough. Using Lemma 3.3 together with 3.5 gives us that $IC^\mu_{\text{int}}[f, \eta] \geq 2\Omega(k)$, whereas Lemma 3.4 gives $IC^\mu_{\text{int}}[f, \eta] = O(k)$. \hfill $\square$

The rest of this section is devoted to the proof of Lemma 3.3, and we begin by giving a proof outline.

**Outline of the Proof of Theorem 3.3.** The proof has two components. Fix a function $f$ and an input distribution $\mu$. In the first step we show that any protocol $\Pi$ for $(f, \mu)$ with low external information, can be converted into a protocol $\Pi’$ such that (a) $\Pi’$ has roughly the same error in computing $(f, \mu)$, and (b) $\Pi’$ never reveals too much information about the player’s input, with respect to any measure $\rho$; we refer to such protocols as having “universally low external information” (defined formally below). In the second step, we show that if $(f, \mu)$ has the $(\epsilon, \delta)$ relative-discrepancy property, then a protocol $\Pi$ with low universal external information can only have a small advantage of in computing $(f, \mu)$. Quantitative issues aside, it is clear that one can combine steps (a) and (b) above to prove that a low external information protocol cannot have a significant advantage in computing a function that has low relative-discrepancy.

### 3.1 Universal External Information

Suppose $\Pi$ is a protocol between Alice and Bob. Suppose Alice speaks first, and denote her messages by $A = (A_1, \ldots, A_m)$, and Bob’s messages by $B = (B_1, \ldots, B_m)$. For each input $\pi \in [m]$ in the protocol, a possible exchange of messages $(a_\pi, b_\pi) \in \{0, 1\}^m \times \{0, 1\}^m$, and a pair of inputs $x, y \in \{0, 1\}^k$, denote

$$P^\Pi_{\text{A,}\pi}(a_\pi, b_\pi) = \prod_{j \leq m} \Pr_{y \sim \rho\pi} \left[ A_j = a_j \mid A_{\leq j} = a_{\leq j}, B_{< j} = b_{< j}, X = x \right],$$

$$P^\Pi_{\text{B,}\pi}(a_\pi, b_\pi) = \prod_{j \leq m} \Pr_{x \sim \mu\pi} \left[ B_j = b_j \mid A_{\leq j} = a_{\leq j}, B_{< j} = b_{< j}, Y = y \right].$$

If this product runs through the whole protocol, i.e. $i = m + 1$, we omit the subscript $i$ and simply write $P^\Pi_{\text{A,}\pi}(a_\pi, b_\pi)$ and $P^\Pi_{\text{B,}\pi}(a_\pi, b_\pi)$.

**Definition 3.6.** With the above notations, we say a protocol $\Pi$ has universal external information at most $M$, if there are non-negative functions $\eta_A(a, b)$ and $\eta_B(a, b)$ over transcripts, such that the following holds.

1. The function $\eta_A(a, b) = \eta_A(a, b) \theta_B(a, b)$ is a probability distribution.

2. For any $(x, y) \in \{0, 1\}^k \times \{0, 1\}^k$ and $(a, b) \in \{0, 1\}^m \times \{0, 1\}^m$ it holds that

$$2^{-M} \leq \frac{P^\Pi_{\text{A,}\pi}(a_\pi, b_\pi)}{\eta_A(a_\pi, b_\pi)} \leq 2^M, \quad 2^{-M} \leq \frac{P^\Pi_{\text{B,}\pi}(a_\pi, b_\pi)}{\eta_B(a_\pi, b_\pi)} \leq 2^M. \quad (11)$$

Informally, a protocol has low universal external information, if for all possible transcript $\pi = (a, b)$, no input of Alice (or Bob) makes $\pi$ much more likely from their point of view. We remark that having low universal external information is a very strong property. For example, it implies that the external information of the protocol is low with respect to any distribution.

**Lemma 3.7.** Suppose that a protocol $\Pi$ has universal external information at most $M$. Then for any distribution $\rho$ over $(x, y)$, we have that $\eta_{\text{ext}}^{\rho}(\Pi) \leq 2M$.

**Proof.** Let $\eta_A, \eta_B$ and $\eta = \eta_A \oplus \eta_B$ be from Definition 3.6. First, we argue that for all inputs $x, y$ it holds that $DKL \left( \Pi_{x=x, y=y} \parallel \eta \right) \leq 2M$. Indeed, by definition

$$DKL \left( \Pi_{x=x, y=y} \parallel \eta \right) = \sum_\pi \Pr \left[ \Pi = \pi \mid X = x, Y = y \right] \log \left( \frac{Pr \left[ \Pi = \pi \mid X = x, Y = y \right]}{\eta(\pi)} \right).$$

Noting that $Pr \left[ \Pi = \pi \mid X = x, Y = y \right] = P^\Pi_{\text{A,}\pi}(\pi)P^\Pi_{\text{B,}\pi}(\pi)$, we get from the universal external information property that

$$\Pr \left[ \Pi = \pi \mid X = x, Y = y \right] \leq 2^{-M}. \quad \eta(\pi)$$
and plugging that in above yields that \( \text{DKL} \left( \Pi |_{X=x, Y=y} \parallel \eta \right) \leq 2M \).

The statement will thus follow if we show that \( \text{D}_{\text{ext}}[\Pi] \leq \mathbb{E}_{(x, y) \sim \rho} \left( \text{DKL} \left( \Pi |_{X=x, Y=y} \parallel \eta \right) \right) \). Indeed, using the definition of external information and Fact 2.4 we get that

\[
\text{D}_{\text{ext}}[\Pi] = \mathbb{E}_{(x, y) \sim \rho} \left( \text{DKL} \left( \Pi |_{X=x, Y=y} \parallel \eta \right) \right),
\]

and therefore

\[
\begin{align*}
\mathbb{E}_{(x, y) \sim \rho} \left( \text{DKL} \left( \Pi |_{X=x, Y=y} \parallel \eta \right) \right) &= \mathbb{E}_{(x, y) \sim \rho} \left( \text{DKL} \left( \Pi |_{X=x, Y=y} \parallel \eta \right) - \text{D}_{\text{ext}}[\Pi] \right) \\
&= \mathbb{E}_{(x, y) \sim \rho} \left( \sum_{\pi} \Pr \left[ \Pi(x, y) = \pi \right] \log \frac{\Pr_{X,Y \sim \rho} \left[ \Pi = \pi \right]}{\eta(\pi)} \right) \\
&= \mathbb{E}_{(x, y) \sim \rho} \left( \sum_{\pi} \Pr \left[ \Pi(x, y) = \pi \right] \log \frac{\Pr_{X,Y \sim \rho} \left[ \Pi = \pi \right]}{\eta(\pi)} \right) \\
&= \text{DKL}(\Pi \parallel \eta) \geq 0.
\end{align*}
\]

\[\square\]

3.2 Step (a): Fixing External Information Leakage

Our goal in this section is to prove the following lemma, asserting that a low external information protocol may be converted into a protocol with low universal external information with only small additional error.

**Lemma 3.8.** Let \( \epsilon, \epsilon' > 0 \) and let \( \mu \) be distributions over \( \{0, 1\}^n \times \{0, 1\}^n \). Suppose \( f : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\} \) is a function, and \( \Pi \) is a protocol for \((f, \mu)\) that has error at most \( \epsilon' \) and \( \text{D}_{\text{ext}}[\Pi] \leq M \). Then there is a protocol \( \Pi' \) for \((f, \mu)\) such that

1. The error of \( \Pi' \) on \((f, \mu)\) is at most \( \epsilon' + 4\epsilon \).
2. The universal external information of \( \Pi' \) is at most \( 2M/\epsilon + 1 \).

We begin by explaining the idea in behind the design of \( \Pi' \). In \( \Pi' \) we will simulate \( \Pi \), except that each player will also measure their information, they will forget about their input and instead sample \( x \) and \( y \) from the distribution \( \Pi \). For each message of each player, they will consider the transcript up to that point, and their input. In case this number of bits has exceeded a certain threshold, the player changes their behaviour and enters a "strike" in which they will act in a way that does not reveal any additional external information regarding his/her input. Strictly speaking, once a player determines they have leaked too much information, they will forget about their input and instead sample their message according to their message distribution at that point in \( \Pi \), conditioned on the transcript so far (but not on the specific input).

Let us now be more precise. Fix a protocol \( \Pi \) such that \( \text{D}_{\text{ext}}[\Pi] \leq M \), and recall the definitions of \( \text{P}_{A,i}^X, \text{P}_{B,i}^Y \) above. We will also need to define their averaged counterparts, i.e.

\[
\eta_{A,i}(a, b) \overset{\text{def}}{=} \text{P}_{A,i}(a, b), \quad \eta_{B,i}(a, b) \overset{\text{def}}{=} \text{P}_{B,i}(a, b).
\]

(The \( \eta_{A,i}, \eta_{B,i} \) notations is not a coincidence, and we will use these functions to exhibit the fact that the protocol \( \Pi \) we construct has low universal external information). We note that \( P_{A,i,1}^{X}, P_{B,i,1}^{Y} \) only depend on the \( i-1 \)-prefixes of \( a \) and \( b \), and \( P_{B,i,1}^{Y}, P_{B,i,1}^{Y} \) only depend on the \( i-1 \)-prefixes of \( a \) and \( b \). We will therefore sometimes abuse notations and drop the rest of \( a \) and \( b \) from the notation. We also note that \( \text{P}_{A,i}^{X}, \text{P}_{B,i}^{Y} \) only depend on \( x \) and \( y \) (and not on \( \mu \)). This is because, at each point in time, a player’s message only depends on their input, and the messages they received from the other player so far.

With these notations, we may consider for each \( a, b \), the likelihood ratios \( S_{A,i}(a_{<i}, b_{<i}, x) = \frac{P_{A,i}^{X}(a_{<i}, b_{<i}, x)}{\eta_{A,i}(a_{<i}, b_{<i}, x)} \) and \( S_{B,i}(a_{<i}, b_{<i}, y) = \frac{P_{B,i}^{Y}(a_{<i}, b_{<i}, y)}{\eta_{B,i}(a_{<i}, b_{<i}, y)} \). Intuitively, these quantities measure how much more/less likely a given exchange of messages (of \( a, b \)) is, when knowing \( x \) and \( y \) respectively, compared to only knowing that \( (x, y) = \mu \). Thus, we may expect an external observer to learn many bits of information in case the protocol was executed and the resulting exchange of messages \( a, b \) has high likelihood ratios, say \( S_{A,B,i}(a, b, x) \geq 2^M \) (in which case we expect an external observer to learn \( \approx M \) bits of information). This turns out to be true, and actually with slightly more work, one can show that the same holds if the likelihood ratios become large at some earlier point in the protocol, \( i < m \).

With this intuition in mind, and noting that Alice (respectively Bob) can compute \( S_{A,i}(a_{<i}, b_{<i}, x) \) (respectively \( S_{B,i}(a_{<i}, b_{<i}, y) \)) it makes sense that the players should alter their behaviour if at some point in their protocol, their likelihood ratio gets too high – say, larger than \( 2^{2M/\epsilon} \). Indeed, this is what our protocol \( \Pi' \) does.

The Protocol \( \Pi' \). We simulate the protocol \( \Pi \), with a small change in the beginning of each player’s turn. Consider a player intending to send their \( i \)th message – say Alice. First, Alice computes \( S_{A,i}(a_{<i}, b_{<i}, x) \) where \( \pi_A \) are the messages of Alice so far, and \( \pi_B \) are the messages of Bob so far. If this quantity is larger than \( 2^{2M/\epsilon} \), or at most \( 2^{-2M/\epsilon} \), Alice moves into “strike mode”, and otherwise proceeds as usual according to the protocol \( \Pi \). Upon entering “strike mode”, Alice will sample her subsequent messages only conditioned on the transcript of the protocol up to that point without taking her input \( x \) into consideration. I.e., to send her \( j \)th message, for \( j \geq i \), Alice considers the transcript of the protocol thus far, \( a_{<j}, b_{<j} \), and the distribution \( A_j(X, Y) \) where \( A_j \) is the next message according to \( \Pi \) and \( b_{<j} \) is at most \( 2^{-2M/\epsilon} \), sampled from the next message according to \( \Pi \) and \( b_{<j} \).

In the remainder of this section, we argue that the probability that a player ever enters “strike mode” in \( \Pi' \) is small, and so \( \Pi' \) retains roughly the same error as \( \Pi \) on \((f, \mu)\). We then show that \( \Pi' \) has universal external information at most \( 2M/\epsilon + 1 \). We remark that for technical reasons, we will need to assume that our original protocol \( \Pi \) is smooth (as in Definition 2.11). Thankfully, by Lemma 2.12, we may indeed do so while only slightly increasing the error of the protocol.

3.2.1 The Error of \( \Pi' \) on \( \mu \) is Comparable to the Error of \( \Pi \). In this section we prove the following lemma.

**Lemma 3.9.** The probability that at least one of the players enters “strike mode” in the protocol \( \Pi' \) when ran on \( \mu \) is at most \( 38\epsilon \).

First, by Lemma 2.12 we may assume henceforth that the protocol \( \Pi \) is smooth and has error at most \( \epsilon' + \epsilon \). Thus, once we prove
Lemma 3.9 it will follow that the error of $\Pi'$ is at most $\epsilon' + 39\epsilon$. The rest of this section is therefore devoted to the proof of Lemma 3.9. By Fact 2.4 and the definition of KL-divergence

$$M \geq I_\mu[X, Y; \Pi] = \mathbb{E}_{(x, y) \sim \mu} \left[ \sum_{a, b} \Pr[\Pi(x, y) = (a, b)] \log \left( \frac{\Pr[\Pi(x, y) = (a, b)]}{\Pr[\Pi(x, y) = (b, a)]} \right) \right].$$

For $a, b \in \{0, 1\}^m$, we define

$$p_{A, \Pi, i}(a, b) = \Pr_{(X, Y) \sim \mu} \left[ A_i = a_i \mid A_1 < a_1, B_1 < b_1 \right],$$

and similarly we define for Bob

$$p_{B, \Pi, i}(a, b) = \Pr_{(X, Y) \sim \mu} \left[ B_i = b_i \mid A_1 < a_1, B_1 < b_1 \right].$$

Also, let

$$p_{A, \Pi, i}^X(a, b) = \Pr_{(X, Y) \sim \mu} \left[ A_i = a_i \mid A_1 < a_1, B_1 < b_1, X = x \right]$$

and

$$p_{B, \Pi, i}^Y(a, b) = \Pr_{(X, Y) \sim \mu} \left[ B_i = b_i \mid A_1 < a_1, B_1 < b_1, Y = y \right].$$

We remark that $p_{A, \Pi, i}(a, b), p_{B, \Pi, i}(a, b)$ depend only on the $i$-prefix of $a$ and the $i - 1$ prefix of $b$, and $p_{A, \Pi, i}^X(a, b), p_{B, \Pi, i}^Y(a, b)$ depend on the $i$-prefixes of both $a$ and $b$. Thus, abusing notations, we sometimes plug in strings of length $i$ into $p_{B, \Pi, i}(a, b)$ and so on.

With these notations, we have

$$\Pr[\Pi = (a, b)] = \prod_{i=1}^m p_{A, \Pi, i}(a, b),$$

and for every fixed $x, y$ it holds that

$$\Pr[\Pi(x, y) = (a, b)] = \prod_{i=1}^m p_{A, \Pi, i}(a_i, b_i) \prod_{i=1}^m p_{B, \Pi, i}(a_i, b_i).$$

Thus, (12) gives us that

$$\sum_{i=1}^m \mathbb{E}_{a, b} \Pr[\Pi = (a, b)] \left( \log \frac{p_{A, \Pi, i}(a, b)}{p_{A, \Pi, i}(a, b)} + \log \frac{p_{B, \Pi, i}(a, b)}{p_{B, \Pi, i}(a, b)} \right)$$

is at most $M$. We consider the two terms on the left hand side separately, i.e. define

$$(I) = \sum_{i=1}^m \mathbb{E}_{(x, y) \sim \mu} \sum_{a, b} \Pr[\Pi(x, y) = (a, b)] \log \left( \frac{p_{A, \Pi, i}(a, b)}{p_{A, \Pi, i}(a, b)} \right),$$

$$(II) = \sum_{i=1}^m \mathbb{E}_{(x, y) \sim \mu} \sum_{a, b} \Pr[\Pi(x, y) = (a, b)] \log \left( \frac{p_{B, \Pi, i}(a, b)}{p_{B, \Pi, i}(a, b)} \right).$$

We now take a moment to reinterpret these two quantities. Define $Z_{x, a, b, i}, Z_{y, a, b, i} : \{0, 1\} \to \mathbb{R}$ as

$$Z_{x, a, b, A, i}(c) = \log \left( \frac{p_{A, \Pi, i}(a_i, b_i)}{p_{A, \Pi, i}(a_i, b_i)} \right),$$

$$Z_{y, a, b, B, i}(c) = \log \left( \frac{p_{B, \Pi, i}(a_i, b_i)}{p_{B, \Pi, i}(a_i, b_i)} \right).$$

Thus, the above two terms can be rewritten as

$$(I) = \sum_{i=1}^m \mathbb{E}_{(x, y) \sim \mu} \sum_{a, b} \Pr[\Pi(x, y) = (a, b)] Z_{x, a, b, A, i}(c),$$

$$(II) = \sum_{i=1}^m \mathbb{E}_{(x, y) \sim \mu} \sum_{a, b} \Pr[\Pi(x, y) = (a, b)] Z_{y, a, b, B, i}(c).$$

Here, $c$ is to be thought of as the next message of the respective player, conditioned on their input and the transcript so far. Note that the distribution of $Z_{x, a, b, A, i}$ only depends on the $(i - 1)$ prefix of $a, b$ and the distribution of $Z_{x, a, b, B, i}$ only depends on the i-prefix of $a$ and $(i - 1)$-prefix of $b$. Let $E_{x, a, b, A, i}, E_{y, a, b, B, i}$ be their expectations, respectively, i.e.

$$E_{x, a, b, A, i} = \sum_{c \in \{0, 1\}} p_{A, \Pi, i}(a_i, b_i) Z_{x, a, b, A, i}(c),$$

$$E_{y, a, b, B, i} = \sum_{c \in \{0, 1\}} p_{B, \Pi, i}(a_i, b_i) Z_{y, a, b, B, i}(c).$$

With these notations, we note that (14) and (15) translate to

$$(I) = \sum_{i=1}^m \mathbb{E}_{(x, y) \sim \mu} \sum_{a_i < a_i} \mathbb{E}_{b_i \in [B_1 \cup B_2]} [E_{x, a, b, A, i}],$$

$$E_{y, a, b, B, i} = \sum_{c \in \{0, 1\}} p_{B, \Pi, i}(a_i, b_i) Z_{y, a, b, B, i}(c).$$

Analyzing the Probability to Enter "strike mode". With the notations we have set, we have that

$$\log(S_{A, i}(a, b, x)) = \sum_{j < i} Z_{x, a, b, A, i}(a_i),$$

and similarly for Bob. Let $\mathcal{A}$ be the event that

$$\sum_{j < i} Z_{x, a, b, A, i}(a_i) \geq \frac{2M}{a - B}$$

for some $a \in [m]$, and $\mathcal{B}$ be the event that

$$\sum_{j < i} Z_{y, a, b, B, i}(a_i) \geq \frac{2M}{a - B}$$

for some $b \in [m]$. Then, and so probability that one of the players in $\Pi$ enters "strike mode" is $Pr_{(x, y) \sim \mu} [\mathcal{A} \cup \mathcal{B}]$, and we bound the probability of $\mathcal{A}, \mathcal{B}$ separately each by $19\epsilon$. Lemma 3.9 thus follows from the union bound. We focus on upper bounding the probability of $\mathcal{A}$, and the argument for Bob is analogous.

Claim 3.10. For each $x, y, a, b$ and $i$, we have that

$$E_{x, a, b, A, i} \geq 2(p_{A, \Pi, i}(a_i < 1, b_i < 1) - p_{A, \Pi, i}(a_i < 1, b_i < 1))^2,$$

$$E_{y, a, b, B, i} \geq 2(p_{B, \Pi, i}(a_i < 1, b_i < 1) - p_{B, \Pi, i}(a_i < 1, b_i < 1))^2.$$

Proof. We show the argument for $E_{x, a, b, A, i}$ and the argument for $E_{x, a, b, B, i}$ is identical. Note that

$$E_{x, a, b, A, i} = DKL(A_i | x_1 \cup A_{<i} = a_1, b_{<i} = b_{<i} \parallel A_i | A_{<i} = a_{<i}, B_{<i} = b_{<i}).$$

Since $\Pi$ is a smooth protocol, we may use Fact 2.7 and conclude that

$$E_{x, a, b, A, i} \geq 2(p_{A, \Pi, i}(a_i < 1, b_i < 1) - p_{A, \Pi, i}(a_i < 1, b_i < 1))^2.$$

We note that Claim 3.10 combined with (13) and (16) immediately implies:
**Corollary 3.11.** We have that
\[
M \geq (l)
\]
\[
\geq 2 \sum_{i=1}^{m} \mathbb{E}_{(x,y) \sim (x,y)} \left[ \mathbb{E}_{a_{<i}, b_{<i} \sim (a_{<i}, b_{<i})} \left[ (p_{A,\Pi,i}^x(a_{<i}, 1, b_{<i}) - p_{A,\Pi,i}(a_{<i}, 1, b_{<i}))^2 \right] \right].
\]

Consider a random choice of \((x, y) \sim \mu, a \sim A(x, y)\) and \(b \sim B(x, y)\). Note that the sequence \(Q_i^A = X_{a,b,A,i}(a_i) - X_{a,b,A,i}\) forms the sum-martingale \(G_i^A = \sum_{j \leq i} Q_i^A\), and similarly the sequence \(Q_i^B = X_{a,b,b,i}(b_i) - X_{a,b,b,i}\) forms the sum martingale \(G_i^B = \sum_{j \leq i} Q_i^B\), both with respect to the natural filtration defined by the transcript of the protocol at each step. The following claim upper bounds the expectation of the square of these sum-martingales in the end.

**Claim 3.12.** \(\mathbb{E} \left[ \left| G_m^A \right|^2 \right] \leq 18l \leq 18M.\)

**Proof.** Using the martingale property we have that
\[
\mathbb{E} \left[ \left| G_m^A \right|^2 \right] = \sum_{i=1}^{m} \mathbb{E} \left[ \left| G_i^A \right|^2 \right] \leq \sum_{i=1}^{m} \mathbb{E} \left[ X_{a,b,A,i}(a_i) \right]^2.
\]

Note that fixing \(x, a_{<i}, b_{<i}\) we have that \(\mathbb{E}_{a_{<i}} \left[ X_{a,b,A,i}(a_i) \right]^2\) is equal to
\[
\sum_{c \in \{0,1\}} \mathbb{P}_{a_{<i}, b_{<i}, c} \left( X_{a,b,A,i}(a_i) \right)^2 \log^2 \left( \frac{\mathbb{P}_{a_{<i}, b_{<i}, c} \left( X_{a,b,A,i}(a_i) \right)}{\mathbb{P}_{a_{<i}, b_{<i}, c} \left( X_{a,b,A,i}(a_i) \right)} \right).
\]
Using the smoothness of \(\Pi\), we have that \(p_{A,\Pi,i}^x(a_{<i}, c, b_{<i})\) is at least \(\frac{1}{2}\) for all \(c \in \{0,1\}\), and since \(|\log(z)| \leq 2|z-1|\) for all \(z \geq 1/2\) we get that \(\mathbb{E}_{a_{<i}} \left[ X_{a,b,A,i}(a_i) \right]^2\) is at most
\[
4 \sum_{c \in \{0,1\}} \mathbb{P}_{a_{<i}, b_{<i}, c} \left( X_{a,b,A,i}(a_i) \right)^2 \log^2 \left( \frac{\mathbb{P}_{a_{<i}, b_{<i}, c} \left( X_{a,b,A,i}(a_i) \right)}{\mathbb{P}_{a_{<i}, b_{<i}, c} \left( X_{a,b,A,i}(a_i) \right)} \right).
\]
By the smoothness of \(\Pi\) we have \(p_{A,\Pi,i}(a_{<i}, c, b_{<i}) \geq 1/3\) for all \(c \in \{0,1\}\), so the above inequality implies that \(\mathbb{E}_{a_{<i}} \left[ X_{a,b,A,i}(a_i) \right]^2 \leq 36 \left( \mathbb{P}_{A,\Pi,i}(a_{<i}, 1, b_{<i}) \right)^2 \). The claim now follows by combining this with Corollary 3.11.

We are now ready to prove Lemma 3.9.

**Proof of Lemma 3.9.** Consider a random choice of \((x, y) \sim \mu, a \sim A(x, y)\) and \(b \sim B(x, y)\). Let \(W_1\) be the event that \(\sum_{i=1}^{m} X_{a,b,A,i} \geq M/\epsilon\) and let \(W_2\) be the event that \(\left| G_i^A \right| \geq \sqrt{M/\epsilon}\) for some \(i\). By Markov’s inequality we have that
\[
\mathbb{P} \left[ W_1 \right] \leq \frac{(l)}{M/\epsilon} \leq \epsilon,
\]
where we used Corollary 3.11. For \(W_2\), using Fact 2.14 and Claim 3.12 gives that
\[
\mathbb{P} \left[ W_2 \right] \leq \frac{\mathbb{E} \left[ \left| G_m^A \right|^2 \right]}{M/\epsilon} \leq 18\epsilon.
\]

Note that \(\sum_{j<i} Z_{x,a,b,A,i}(a_j) \geq (\epsilon)_{k-1} \frac{\sum_{i=1}^{m} X_{a,b,A,i}}{M/\epsilon}\) so if none of \(W_1, W_2\) hold, then
\[
\sum_{j<i} Z_{x,a,b,A,i}(a_j) \leq M/\epsilon + \sqrt{M/\epsilon} \leq 2M/\epsilon
\]
and Alice never enters “strike mode” in the duration on the execution of \(\Pi'\). Therefore, the probability Alice enters into “strike mode” on an execution of \(\Pi'\) is at most \(\mathbb{P} \left[ W_1 \right] + \mathbb{P} \left[ W_2 \right] \leq 19\epsilon\). The same goes for Bob, and Lemma 3.9 follows from the union bound.

3.2.2. The Protocol \(\Pi'\) Has Universal External Information at most \(2M/\epsilon + 1\). Let \(A = \eta_{A,m+1}\) and \(B = \eta_{B,m+1}\). It is easy to see that the \(\eta(a, b) = \eta_{A}(a, b)\eta_{B}(a, b)\) is a distribution. We show that it holds that \(\Pi'\) has low universal external information.

Suppose towards contradiction that this is not the case. Then there are inputs \(x, y\) and possible transcripts \(a, b\) and a step \(i\) such that (11) fails – suppose without loss of generality that \(\frac{p_{A,\Pi}(a, b)}{\eta_{A}(a, b)} > 2^{2M/\epsilon + 1}\).

We claim that on an execution of \(\Pi'\) on \(x, y\) that yields the transcript \((a, b)\), it must be the case that Alice entered “strike mode”. Otherwise, we have that
\[
\eta_{A,m} X_{a,b,A,i}(a_i) = \mathbb{P} \left[ A_m = a_m \text{ in } \Pi \mid a_{<m}, b_{<m}, x \right] \mathbb{P} \left[ A_m = a_m \text{ in } \Pi \mid a_{<m}, b_{<m}, x \right]
\]
Since Alice did not enter “strike mode”, the first fraction is between \(2^{-2M/\epsilon}\) and \(2^{2M/\epsilon}\), and since \(\Pi\) is smooth, the second fraction is between 1/2 and 2. It follows that (11) holds for \(2M/\epsilon + 1\) for Alice, in contradiction.

Let \(i\) be the step in the protocol in which Alice decided to enter “strike mode”. By the minimality of \(i\) and smoothness of \(\Pi\) we conclude that \(S_{A,i}(a_{<i}, b_{<i}, x)\) is equal to
\[
S_{A,i-1}(a_{<i-1}, b_{<i-1}, x) \mathbb{P} \left[ A_{i-1} = a_{i-1} \text{ in } \Pi \mid a_{<i-1}, b_{<i-1}, x \right] \mathbb{P} \left[ A_{i-1} = a_{i-1} \text{ in } \Pi \mid a_{<i-1}, b_{<i-1}, x \right]
\]
which is at most \(2^{2M/\epsilon} \cdot 2\). Now, note that by the behaviour of Alice in strike mode it follows that
\[
p_{A,\Pi}(a, b) = \mathbb{P}_X \left[ A_k = a_k \text{ in } \Pi \mid a_{<k}, b_{<k} \right] \mathbb{P}_X \left[ A_k = a_k \text{ in } \Pi \mid a_{<k}, b_{<k} \right]
\]
and clearly
\[
\eta_{A,m}(a, b) = \mathbb{P}_X \left[ A_k = a_k \text{ in } \Pi \mid a_{<k}, b_{<k} \right] \mathbb{P}_X \left[ A_k = a_k \text{ in } \Pi \mid a_{<k}, b_{<k} \right]
\]
so
\[
\frac{p_{A,\Pi}(a, b)}{\eta_{A,m}(a, b)} = S_{A,i}(a_{<i}, b_{<i}, x) \leq 2^{2M/\epsilon + 1}\]
which is a contradiction.

3.3. Step (b): Strong Relative Discrepancy Implies High Universal External Information

Next, we prove the following lemma, asserting that a protocol with low universal external information cannot compute functions that have low relative-discrepancy.
LEMMA 3.13. Let $\delta, \epsilon, \epsilon' > 0$ and $M \in \mathbb{N}$. Let $\mu$ be a distribution over $\{0,1\}^n \times \{0,1\}^n$, and suppose that $f : \{0,1\}^n \times \{0,1\}^n \rightarrow \{0,1\}$ is such that $(f, \mu)$ has $(\epsilon, \delta)$ relative-discrepancy. If $\Pi$ is a protocol for $(f, \mu)$ whose universal external information at most $M$, then for $(X, Y) \sim \mu$,

$$SD(\Pi(X, Y)|f(X, Y)=0, \Pi(X, Y)|f(X, Y)=1) \leq 20 \left( \epsilon + \epsilon' + \frac{2M}{\epsilon^2} \delta \right).$$

Proof. By Definition 3.6, there are functions $\eta_0(a, b), \eta_1(a, b) \geq 0$ such that $\eta(a, b) = \eta_0(a, b)\eta_1(a, b)$ is a distribution, and inequality (11) holds for all $x, y, a, b$. For each possible transcript $\pi$, we partition $x, y$ into rectangles according to the ratios $\frac{P^X_\pi(x)}{\eta_0(\pi)}$ and $\frac{P^Y_\pi(y)}{\eta_1(\pi)}$. Namely, for $-M/\epsilon' \leq i, j < M/\epsilon'$ we denote

$$R^X_\pi[i] = \left\{ x \mid 1 + \epsilon'y' \leq \frac{P^X_\pi(x)}{\eta_0(\pi)} \leq 1 + \epsilon'y' + 1 \right\},$$

$$R^Y_\pi[j] = \left\{ y \mid 1 + \epsilon'y' \leq \frac{P^Y_\pi(y)}{\eta_1(\pi)} \leq 1 + \epsilon'y' + 1 \right\},$$

and define $R_{\pi}[i, j] = R^X_\pi[i] \times R^Y_\pi[j]$. We note that the number of rectangles is at most $(2M/\epsilon')^2$, and that they partition the entire domain.

Let $\rho$ be a distribution from Definition 3.2 exhibiting the fact that $(f, \mu)$ has $(\epsilon, \delta)$ relative-discrepancy. We say a rectangle is heavy if $\rho(R_{\pi}[i, j]) \geq \delta$ and otherwise we say it is light. For future reference, note that the total $\rho$-weight on light rectangles is at most $(2M/\epsilon')^2 \delta$.

Fix $\pi$ and let $H_\pi = \{(i, j) \mid R_{\pi}[i, j] \text{ is heavy}\}$. Then for all $b \in \{0,1\}$ we have

$$\Pr_{(x,y)\sim\mu}[\Pi(x, y) = \pi, f(x, y) = b] = \sum_{(x,y) \in \Pi(x, y) = \pi, f(x, y) = b} \mu(x,y)P^X_\pi(x)P^Y_\pi(y) \geq \sum_{(i,j) \in H_\pi} \sum_{f(x,y)=b} \mu(x,y)P^X_\pi(x)P^Y_\pi(y) \geq \sum_{(i,j) \in H_\pi} \sum_{f(x,y)=b} (1 + \epsilon'y')^j \mu(x,y)\eta_0(\pi)\eta_1(\pi),$$

where in the last inequality we used the definition of $R_{\pi}[i, j]$. Thus, we get that

$$\Pr_{(x,y)\sim\mu}[\Pi(x, y) = \pi, f(x, y) = b] \geq \eta(\pi) \sum_{(i,j) \in H_\pi} (1 + \epsilon'y')^j \rho(R_{\pi}[i, j] \cap f^{-1}(b)).$$

By the relative-discrepancy property we have that $\rho(R_{\pi}[i, j] \cap f^{-1}(b)) \geq \left( \frac{1}{2} - \epsilon \right) \rho(R_{\pi}[i, j])$, and so

$$\Pr_{(x,y)\sim\mu}[\Pi(x, y) = \pi, f(x, y) = b] \geq \left( \frac{1}{2} - \epsilon \right) \eta(\pi) \sum_{(i,j) \in H_\pi} (1 + \epsilon'y')^j \rho(R_{\pi}[i, j]).$$

and we analyze the last sum. Note that

$$\sum_{(i,j) \in H_\pi} (1 + \epsilon'y')^j \rho(R_{\pi}[i, j]) = \sum_{(i,j) \in H_\pi, (x,y) \in R_{\pi}[i,j]} (1 + \epsilon'y')^j \rho(x,y) \geq \sum_{(i,j) \in H_\pi, (x,y) \in R_{\pi}[i,j]} (1 + \epsilon'y')^j \frac{P^X_\pi(x)P^Y_\pi(y)}{\eta_0(\pi)\eta_1(\pi)} \rho(x,y) = \sum_{(i,j) \in H_\pi, (x,y) \in R_{\pi}[i,j]} (1 + \epsilon'y')^2 \frac{1}{\eta_0(\pi)} \rho(x,y)P^X_\pi(x)P^Y_\pi(y) = \frac{(1 + \epsilon'y')^2}{\eta_0(\pi)} \Pr_{(X,Y)\sim\rho}[\Pi(X,Y) = \pi, (X,Y) \text{ in a heavy rectangle}] = \frac{(1 + \epsilon'y')^2}{\eta_0(\pi)} \rho(\pi) = \frac{(1 + \epsilon'y')^2}{\eta_0(\pi)} \Pr_{(X,Y)\sim\rho}[\Pi(X,Y) = \pi, (X,Y) \text{ in light rectangle}].$$

We now upper bound the last probability. By conditioning we have that

$$\Pr_{(X,Y)\sim\rho}[\Pi(X,Y) = \pi, (X,Y) \text{ in a light rectangle of } \pi] \leq 2^{2M} \eta(\pi) \sum_{(i,j) \in H_\pi, (x,y) \in R_{\pi}[i,j]} \rho(x,y),$$

and by our earlier notations we have $\rho(x,y) = P^X_\pi(x)P^Y_\pi(y) \leq 2^{2M} \eta_0(\pi)\eta_1(\pi) = 2^{2M} \eta(\pi)$, so

$$\Pr_{(X,Y)\sim\rho}[\Pi(X,Y) = \pi, (X,Y) \text{ in a light rectangle of } \pi] \leq 2^{2M} \eta(\pi) \sum_{(i,j) \in H_\pi, (x,y) \in R_{\pi}[i,j]} \rho(x,y),$$

which is at most $2^{2M} \eta(\pi) \cdot (2M/\epsilon')^2 \leq 2^{2M} \eta(\pi) \cdot 2^{2M} \eta(\pi) = 2^{4M} \eta(\pi)$. Plugging this into (18), and then (18) into (17) yields that

$$\Pr_{(x,y)\sim\mu}[\Pi(x, y) = \pi, f(x, y) = b] \geq \left( \frac{1}{2} - \epsilon \right) \left( 1 + \epsilon'y' \right)^2 \left( \rho(\pi) - \frac{2^{2M} \delta \rho(\pi)}{\epsilon^2} \right).$$

(19)

We note that summing this up over all $\pi$, we get that

$$\Pr_{(x,y)\sim\mu}[f(x, y) = b] \geq \left( \frac{1}{2} - \epsilon \right) \left( 1 + \epsilon'y' \right)^2 (1 - \frac{2^{4M} \delta}{\epsilon^2} \cdot \rho(\pi)) \geq \frac{1}{2} - \epsilon - \epsilon' - \frac{2^{4M} \delta}{\epsilon^2}.$$ 

(20)

We can now bound the statistical distance between $\Pi(X, Y)|f(X, Y)=1$ and $\Pi(X, Y)|f(X, Y)=0$ where $(X, Y) \sim \mu$. By Fact 2.16, there is $A \subseteq \text{Supp}(\Pi)$ such that this statistical distance is equal to

$$1 - \sum_{\pi \in A} \Pr[\Pi(X, Y) = \pi \mid f(X, Y) = 1] - \sum_{\pi \notin A} \Pr[\Pi(X, Y) = \pi \mid f(X, Y) = 0].$$

(21)
Denote $p = \Pr_{(X, Y) \sim \mu}[f(X, Y) = 1]$, and note that by (21) we have that $|p - 1/2| \leq \varepsilon + \varepsilon' + \frac{24M}{\varepsilon^2} \delta$. We get from the above that the statistical distance is equal to

$$1 - \frac{1}{\rho} \sum_{\pi \in A} \Pr[f(X, Y) = 1, \Pi(X, Y) = \pi] - \frac{1}{1 - p} \sum_{\pi \notin A} \Pr[f(X, Y) = 0, \Pi(X, Y) = \pi],$$

which is at most

$$1 - \sum_{\pi \in A} \Pr[f(X, Y) = 1, \Pi(X, Y) = \pi] - \sum_{\pi \notin A} \Pr[f(X, Y) = 0, \Pi(X, Y) = \pi] + 8(\varepsilon + \varepsilon' + \frac{24M}{\varepsilon^2} \delta), \tag{22}$$

and it is enough to lower bound the two sums. Define $b_\pi = 1$ if $\pi \in A$, and $b_\pi = 0$ if $\pi \notin A$. Then together the two sums can be written as

$$\sum_{\pi \in A} \Pr(f(X, Y) = b_\pi, \Pi(X, Y) = \pi) \geq \sum_{\pi \in A} \left(\frac{1}{2} - \varepsilon\right)(1 + \varepsilon')^{-2} \left(\rho(\pi) - \frac{24M}{\varepsilon^2} \delta \eta(\pi)\right),$$

where we used (19). Since the sum of $\rho(\pi)$, as well as the sum of $\eta(\pi)$, is 1, we get that the last expression is equal to $\left(\frac{1}{2} - \varepsilon\right)(1 + \varepsilon')^{-2}(1 - \frac{24M}{\varepsilon^2} \delta) \geq \frac{1}{2} - \varepsilon - \varepsilon' - \frac{24M}{\varepsilon^2} \delta$, and plugging this into (22) yields the result. \(\square\)

REFERENCES

[1] Scott Aaronson, Shalev Ben-David, and Robin Kothari. 2016. Separations in query complexity using cheat sheets. In Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016. 863–876. https://doi.org/10.1145/2893518.2893644

[2] Rudolf Ahlswede and Ning Cai. 1994. On communication complexity of vector-valued functions. IEEE Transactions on Information Theory 40, 6 (1994), 2062–2067.

[3] Anurag Anshu, Aleksandrs Belovs, Shalev Ben-David, Mika Göös, Rahul Jain, Robin Kothari, Troy Lee, and Miklos Santha. 2016. Separations in Communication Complexity Using Cheat Sheets and Information Complexity. In IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA. 555–564. https://doi.org/10.1109/FOCS.2016.66

[4] Reuven Bar-Yehuda, Benny Chor, Eyal Kushilevitz, and Alon Orlovtsky. 1993. Privacy, additional information and communication. IEEE Transactions on Information Theory 39, 6 (1993), 1930–1943.

[5] Ziv Bar-Yossef, Thathachar S Jayram, Ravi Kumar, and D Sivakumar. 2004. An information statistics approach to data stream and communication complexity. J. Comput. System Sci. 68, 4 (2004), 792–793.

[6] Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. 2013. How to Compress Interactive Communication. SIAM J. Comput. 42, 3 (2013), 1327–1363. https://doi.org/10.1137/100811969

[7] Mark Braverman. 2012. Coding for interactive computation: progress and challenges. In 50th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2012.

[8] Mark Braverman. 2015. Interactive information complexity. SIAM J. Comput. 44, 6 (2015), 1698–1739.

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

[10] Mark Braverman and Anup Rao. 2014. Information Equals Amortized Communication. IEEE Trans. Inf. Theory 60, 10 (2014), 6058–6069. https://doi.org/10.1109/TIT.2014.2347282

[11] Mark Braverman and Omri Weinstein. 2015. An Interactive Information Odometer and Applications. In Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015. 341–350. https://doi.org/10.1145/2746539.2746548

[12] Ankit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew Yao. 2001. Informational complexity and the direct sum problem for simultaneous message complexity. In Proceedings 42nd IEEE Symposium on Foundations of Computer Science. IEEE, 270–278.

[13] Thomas M Cover. 1999. Elements of information theory. John Wiley & Sons.

[14] Tomas Feder, Eyal Kushilevitz, Moni Naor, and Noam Nisan. 1995. Amortized compression. SIAM J. Comput. 24, 4 (1995), 736–750.

[15] Anat Ganor, Gillat Kol, and Ran Raz. 2016. Exponential Separation of Information and Communication for Boolean Functions. J. ACM 63, 5 (2016), 46:1–46:31. https://doi.org/10.1145/2907939

[16] Anat Ganor, Gillat Kol, and Ran Raz. 2019. Exponential separation of communication from external information. SIAM J. Comput. 0 (2019), STOC’16–236.

[17] Prabhodh Harsha, Rahul Jain, David A. Mccalister, and Jakumarr Radatharshnan. 2010. The communication complexity of correlation. IEEE Trans. Inf. Theory 56, 1 (2010), 438–449. https://doi.org/10.1109/TIT.2009.2034824

[18] Gillat Kol, Shay Moran, Amir Shpilka, and Amir Yehudayoff. 2016. Direct Sum Fails for Zero-Error Average Communication. Algorithmica 76, 3 (2016), 782–795. https://doi.org/10.1007/s00453-016-0144-9

[19] Omri Weinstein. 2015. Information complexity and the quest for interactive compression. ACM SIGACT News 46, 2 (2015), 41–64.