New Separations Results for External Information

Mark Braverman * Dor Minzer †

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 obtain the following results:

- We prove an exponential separation between external and internal information complexity, which is the best possible; previously no separation was known.
- We prove a near-quadratic separation between amortized zero-error communication complexity and external information complexity for total functions, disproving a conjecture of [Bra12].
- We prove a matching upper showing that our separation result is tight.

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 cost $\mathrm{CC}_\mu[\pi]$ of π is the expected *number* of bits exchanged in π;
- the *external* information cost $\mathrm{I}_\mu^{\mathsf{external}}[\pi]$ of π is the expected amount of information *an external observer* learns about the inputs (x, y) by observing an execution of $\pi(x, y)$;
- the *internal* information cost $\mathrm{I}_\mu^{\mathsf{internal}}[\pi]$ of π is the expected amount of information *the protocol participants* learn about the inputs (x, y) by observing π.

For a given computational task — such as computing a boolean function $f(x, y)$ with a prescribed error ε — one can define the $\{$average case communication, external information, internal information$\}$ complexity of performing the task by taking the infimum of the corresponding cost over all protocols π that succeed at performing the task. Thus:

- the (average case) communication complexity $\mathrm{CC}_\mu[f, \varepsilon]$ is the smallest expected *number* of bits that need to be exchanged to compute f with error $\leqslant \varepsilon$;
- the *external* information complexity $\mathrm{IC}_\mu^{\mathsf{external}}[f, \varepsilon]$ is the smallest amount of information that must be revealed to an external observer by two parties who need to compute f with error $\leqslant \varepsilon$;

*Department of Computer Science, Princeton University. Research 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.

†Department of Mathematics, Massachusetts Institute of Technology.

[1] See Section 2 for rigorous definitions and further background.

- the *internal* information complexity $\mathrm{IC}_\mu^{\mathsf{internal}}[f,\varepsilon]$[2] is the smallest amount of information that must be revealed by the players to each other while computing f with error $\leqslant \varepsilon$.

It follows from basic information-theoretic calculations that for all tasks:

$$\mathrm{IC}_\mu^{\mathsf{internal}}[f,\varepsilon] \leqslant \mathrm{IC}_\mu^{\mathsf{external}}[f,\varepsilon] \leqslant \mathrm{CC}_\mu[f,\varepsilon]. \tag{1}$$

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 [BYCKO93], where a low $\mathrm{IC}_\mu^{\mathsf{internal}}[f,\varepsilon]$ would mean that information-theoretically secure two-party computation is possible (turns 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 [CSWY01, BYJKS04], 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 [Bra15, Wei15].

Starting with the works of Shannon in the 1940s, the main motivation for using information-theoretic quantities is that they *tensorize*[3]. 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,ε) the task of computing q independent copies of f, each with error $\leqslant \varepsilon$, then

$$\mathrm{IC}_{\mu^q}^{\mathsf{internal}}[f^q,\varepsilon] = q \cdot \mathrm{IC}_\mu^{\mathsf{internal}}[f,\varepsilon]. \tag{2}$$

Such a relationship was known to be false for CC, and was believed to also not hold for $\mathrm{IC}^{\mathsf{external}}$, although to the best of our knowledge only in the present paper we rule it out for all values of ε including $\varepsilon = 0$.

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

$$\mathrm{CC}_{\mu^q}[f^q,\varepsilon] \geqslant \mathrm{IC}_{\mu^q}^{\mathsf{internal}}[f^q,\varepsilon] = q \cdot \mathrm{IC}_\mu^{\mathsf{internal}}[f,\varepsilon]. \tag{3}$$

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 $\mathsf{Disjointness}$ [BYJKS04, BGPW13].

It turns out that, in fact, for $\varepsilon > 0$ (3) is tight [BR14]. For all f and $\varepsilon > 0$ the following holds:

$$\lim_{q\to\infty} \mathrm{CC}_{\mu^q}[f^q,\varepsilon]/q = \mathrm{IC}_\mu^{\mathsf{internal}}[f,\varepsilon]. \tag{4}$$

Equation (4) has given rise to two questions:

- **What is the relationship between $\mathrm{IC}_\mu^{\mathsf{internal}}[f,\varepsilon]$ and $\mathrm{CC}_\mu[f,\varepsilon]$?** 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 $\mathsf{EQ}_n(x,y)$ — the problem of determining whether two n-bit strings are equal is $O(\log 1/\varepsilon)$ independent of n when error ε is allowed, but increases to $n+1$ when no error is allowed.

[2] Sometimes called simply "information complexity".

[3] Also said to satisfy a "direct sum" property in the TCS literature.

Separation between information and communication. The first question was answered by Ganor, Kol, and Raz [GKR16], who showed an exponential separation between internal information complexity and communication complexity. Moreover, such separation is the best possible [Bra15]. In other words, there is a boolean function f such that $\mathrm{IC}_\mu^{\mathsf{internal}}[f,\varepsilon] = O(k)$, while $\mathrm{CC}_\mu[f,\varepsilon] = \Omega(2^k)$. In addition, exponential separation was shown between external information and communication complexity — at least for tasks [GKR19].

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 [GKR16] has an exponential external information complexity (and not just exponential communication complexity), and thus gives an example of an f such that

$$\mathrm{IC}_\mu^{\mathsf{external}}[f,\varepsilon] \geqslant 2^{\Omega(\mathrm{IC}_\mu^{\mathsf{internal}}[f,\varepsilon])}. \tag{5}$$

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:[4]

$$\lim_{q\to\infty} \mathrm{CC}_{\mu^q}[f^q,0]/q \tag{6}$$

A canonical example of a function where zero-error and vanishing-error communication costs diverge is the n-bit *Equality* function $\mathsf{EQ}_n(x,y) := \mathbf{1}_{x=y}$. It is known that the amortized communication complexity of $\mathsf{EQ_n}$ is $O(1)$ [FKNN95]. As a consequence of this result (which can also be seen directly [Bra15]), one gets for all μ,

$$\mathrm{IC}_\mu^{\mathsf{internal}}[\mathsf{EQ}_n,0] = O(1).$$

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

$$\lim_{q\to\infty} \mathrm{CC}_{\mu^q}[\mathsf{EQ}_n^q,0]/q = \Omega(n),$$

where μ 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 $\mathsf{EQ}_n^{-1}(1)$. Therefore (4) has no chance of holding when $\varepsilon = 0$. More precisely, half of the proof of (4) holds for $\varepsilon = 0$, yielding

$$\lim_{q\to\infty} \mathrm{CC}_{\mu^q}[f^q,0]/q \geqslant \mathrm{IC}_\mu^{\mathsf{internal}}[f,0], \tag{7}$$

but this inequality may be strict, as is indeed the case for $f = \mathsf{EQ}_n$.

An attempt to prove the $\leqslant$ 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 [BR14], 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 [HJMR10] leading to a variant of a converse to (7) where internal information is replaced with external information:

$$\lim_{q\to\infty} \mathrm{CC}_{\mu^q}[f^q,0]/q \leqslant \mathrm{IC}_\mu^{\mathsf{external}}[f,0]. \tag{8}$$

[4]Here, zero-error means that the protocol has to always output the correct value of $f(x,y)$, even if $\mu(x,y) = 0$.

We formally prove (8) for completeness purposes in Section A.1 (Theorem A.1). Inequality (8), along with the fact that inequality (7) is easily seen to not be tight led to the following conjecture [Bra12]:

$$\textbf{Conjecture}: \lim_{q\to\infty} \mathrm{CC}_{\mu^q}[f^q,0]/q = \Theta(\mathrm{IC}^{\mathsf{external}}_{\mu}[f,0]). \tag{9}$$

There were several reasons to believe this conjecture. Translated to this language, a result of Ahlswede and Cai [AC94] shows that (9) holds (with a constant 1) when f is the 2-bit AND function for the hardest distribution μ — 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 A.2 for completeness. Here the $\Box^1$ superscript stands for the complexity of proving that the value of $f(x,y)=1$.

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

$$\mathrm{IC}^{\mathsf{external},1}_{\mu}[f,0] \leqslant \lim_{q\to\infty} \mathrm{CC}^{1^q}_{\mu^q}[f^q,0]/q.$$

We should note that a quadratic upper bound on $\mathrm{IC}^{\mathsf{external}}_{\mu}[f,\varepsilon]$ 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 $D(f) \leqslant N^0(f)\cdot N^1(f)$ bound on deterministic communication complexity in terms of non-deterministic communication complexity.

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 μ 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 [KMSY16]. 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 $f\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ and an input distribution μ satisfying:*

1. $\mathrm{IC}^{\mathsf{internal}}_{\mu}[f,\varepsilon] \leqslant O(k)$,
2. $\mathrm{IC}^{\mathsf{external}}_{\mu}[f,\varepsilon] \geqslant 2^{\Omega(k)}$.

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 $f\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ and an input distribution μ such that*

1. $\lim_{q\to\infty} \frac{1}{q}\mathrm{CC}_{\mu^q}(f^q,0) \leqslant O(\sqrt{k}\log^2 k)$,

2. $\mathrm{IC}_{\mu}^{\mathsf{external}}[f, 1/16] \geqslant \Omega(k)$.

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 k, there is $n \in \mathbb{N}$, a function $f\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ and an input distribution μ such that*

1. $\lim_{q\to\infty} \frac{1}{q}\mathrm{CC}_{\mu^q}(f^q, 0) \leqslant O(1)$,
2. $\mathrm{IC}_{\mu}^{\mathsf{external}}[f, 0] \geqslant \Omega(k)$.

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 $\varepsilon > 0$, $f\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ and an input distribution μ, we have that*

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

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 [GKR16] 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 [GKR16] 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 [GKR16]) do very poorly when trying to compute functions with the appropriate relative discrepancy property.

Separating amortized zero-error communication from external information. It is actually surprisingly difficult to construct a candidate function for our main separation. We need a function with a low amortized zero-error communication, but a high external information complexity. Low amortized communication implies low internal information complexity. This means that inside the construction we should use a separation between internal and external information.

In addition, as seen in Theorem 1.1, non-deterministic zero-error amortized communication complexity appears to be connected to zero-error non-deterministic external information complexity. This connection is reinforced by Theorem 1.5. Therefore, the construction is likely to require a function featuring maximum possible (i.e. quadratic) separation between non-deterministic communication and deterministic communication complexity.

Indeed, our starting point is a boolean function whose query complexity exhibits a quadratic separation between deterministic and non-deterministic complexity. To "lift" this separation into the communication world, our construction uses an idea that is close in spirit to the cheat sheet lifting constructions [ABK16,

ABB$^+$16]. At a high level, we too want to give advantage (say, access to non-deterministic certificates) to certain class of protocols (in our case, protocols that solve many independent instances). Our construction however implements this high-level idea differently.

We hide extra information as an output of an auxiliary function f. Our chosen function f has low internal information complexity so as to be useful for constructing a protocol with low amortized communication complexity. This extra information allows us to evaluate our function on all but $o(1)$ fraction of the input tuples of the players in which they fail (and therefore importantly do not introduce errors), and for those one may use a trivial protocol (which would contribute at most $o(1)$ to the amortized communication anyway). The second property that we need is that this extra information would be useless for protocols with low external information attempting to solve only a single challenge. Indeed, our chosen function f will have high external information complexity, and our argument in the proof of Theorem 1.2 actually shows that low external information protocols are unable to gain even a slight advantage for computing f. In particular, the extra information we hide looks essentially random to them, and therefore does not offer any help.

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 $h_\wedge \colon \{0,1\}^m \times \{0,1\}^m \to \{0,1\}$ defined as $h_\wedge(x,y) = h(x \wedge y)$, with the following properties: (1) the external information of $h_\wedge$ is $\Omega(m)$; (2) one can certify that $h(z) = 0$ or that $h(z) = 1$ using only $C = O(\sqrt{m})$ bits. We then wish to construct a function H on 4-tuples, (x, y, u, v), whose output on (x, y, u, v) is $h_\wedge(x, y)$, 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 $h_\wedge(x, y)$. 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 $h_\wedge$, 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 μ has full support. Characterize:*

$$\mathrm{ZAC}(f, \mu) := \lim_{q \to \infty} \mathrm{CC}_{\mu^q}(f^q, 0)/q, \tag{10}$$

where 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 1.3 in Section 3, and Theorem 1.3 in Section 4. Finally, we prove Theorem 1.5 and Corollary 1.4 in Section 5.

2 Preliminaries

2.1 Information theory

We begin with a few basic definitions from information theory. Throughout the papers, 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* $\mathrm{H}[X] = \sum_x \Pr[X = x] \log\left(\frac{1}{\Pr[X=x]}\right)$.
2. *The Shannon entropy of X conditioned on Y is* $\mathrm{H}[X \mid Y] = \mathbb{E}_{y\sim Y}[\mathrm{H}[X \mid Y = y]]$*, where* $\mathrm{H}[X \mid Y = y] = \sum_x \Pr[X = x \mid Y = y] \log\left(\frac{1}{\Pr[X=x \mid Y=y]}\right)$.

Definition 2.2. *Let X, Y, Z be random variables with a finite support.*

1. *The mutual information between X and Y is* $\mathrm{I}[X;Y] = \mathrm{H}[X] - \mathrm{H}[X|Y]$.
2. *The mutual information between X, Y conditioned on Z is* $\mathrm{I}[X;Y \mid Z] = \mathrm{H}[X \mid Z] - \mathrm{H}[X \mid Y, Z]$.

Definition 2.3. *Let X, Y be random variables with a finite support. The KL-divergence from Y to X is* $\mathrm{D}_{KL}(X \parallel Y) = \sum_{x,y} \Pr[X = x] \log\left(\frac{\Pr[X=x]}{\Pr[Y=y]}\right)$.

We will need the following standard facts from information theory (for proofs, see [Cov99] for example).

Fact 2.4. *Let X, Y, Z be random variables. Then*

$$\mathrm{I}[X, Y; Z] = \mathop{\mathbb{E}}_{(x,y)\sim(X,Y)}[\mathrm{D}_{KL}(Z|_{X=x,Y=y} \parallel Z)].$$

Fact 2.5. *Let $X, Y_1, \ldots, Y_n$ be random variables. Then*

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

Fact 2.6. *Let X, Y, Z be random variables. Then* $\mathrm{I}[X;Y \mid Z] \leqslant \mathrm{I}[X;Y,Z]$.

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

Fact 2.7. *Let $p, q \in [0,1]$ and suppose that $\frac{1}{3} \leqslant q \leqslant \frac{2}{3}$. Then*

$$2(p-q)^2 \leqslant \mathrm{D}_{KL}(\mathsf{B}(p) \parallel \mathsf{B}(q)) \leqslant \frac{9(p-q)^2}{2\ln 2}.$$

2.2 Communication complexity

Let $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ be a function, let μ be a distribution over its inputs and denote $(X,Y) \sim \mu$. Throughout, we denote by Π 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 π by $\mathsf{output}(\pi)$. Abusing notations, for inputs x, y, we denote by $\mathsf{output}(\Pi(x,y))$ the random variable which is the output of Π when ran on inputs x, y.

Definition 2.8. *The internal information of the protocol Π is defined as $\mathrm{I}^{\mathsf{internal}}_{\mu}[\Pi] = \mathrm{I}[\Pi; X|Y] + \mathrm{I}[\Pi; Y|X]$.*

For an error parameter $0 \leqslant \varepsilon < \frac{1}{2}$, we define the internal information cost of f on μ with error ε by

$$\mathrm{IC}^{\mathsf{internal}}_{\mu}[f, \varepsilon] = \inf_{\Pi : \Pr_{(x,y)\sim\mu}[f(x,x) \neq \mathsf{output}(\Pi(x,y))] \leqslant \varepsilon} \mathrm{I}^{\mathsf{internal}}_{\mu}[\Pi].$$

Definition 2.9. *The external information of the protocol Π is defined as $\mathrm{I}^{\mathsf{external}}_{\mu}[\Pi] = \mathrm{I}[\Pi; X, Y]$.*

For an error parameter $0 \leqslant \varepsilon < \frac{1}{2}$, we define the external information cost of f on μ with error ε by

$$\mathrm{IC}^{\mathsf{external}}_{\mu}[f, \varepsilon] = \inf_{\Pi : \Pr_{(x,y)\sim\mu}[f(x,y) \neq \Pi(x,y)] \leqslant \varepsilon} \mathrm{I}^{\mathsf{external}}_{\mu}[\Pi].$$

Fact 2.10. *For any function f and $\varepsilon > 0$ it holds that $\mathrm{IC}^{\mathsf{external}}_{\mu}[f, \varepsilon] \geqslant \mathrm{IC}^{\mathsf{internal}}_{\mu}[f, \varepsilon]$.*

We need to use the notion of smooth protocols, as defined in [BBCR13].

Definition 2.11. *A two-player protocol Π is called smooth if for every pair of inputs x, y, a step i in the protocol, and a possible transcript T up to the $(i-1)$-th step, it holds that the distribution of the next message M_i satisfies that $\frac{1}{3} \leqslant \Pr[M_i = 1 \mid x, y, T] \leqslant \frac{2}{3}$.*

An important fact that we will use, is that one can transform a given protocol Π into a smooth protocol Π' that has roughly the same error probability, whose information cost is the same as the original protocol Π. Such statement was proved in [BW15, Lemma 23] for internal information, and the same argument also works for external information. We thus have the following lemma.

Lemma 2.12. *Suppose $f\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ and that μ is a distribution over $\{0,1\}^n \times \{0,1\}^n$, and $\varepsilon, \varepsilon' > 0$. Then any protocol Π for (f, μ) with external information at most M and error at most ε' can be turned into a smooth protocol Π' for (f, μ) external information at most M and error at most $\varepsilon' + \varepsilon$.*

Fact 2.13. *Let P, Q be distributions over domain X. Then*

$$\sum_x P(x) \left| \log\left(\frac{P(x)}{Q(x)}\right) \right| \leqslant \mathrm{D}_{KL}(P \parallel Q) + 8.$$

Proof. Partition $A = \{x \mid P(x) \leqslant Q(X)\}$ into $A_j = \left\{x \mid 2^{-j-1}Q(x) < P(x) \leqslant 2^{-j}Q(x)\right\}$ where $j = 0, 1, \ldots$. Then the left hand side is

$$\sum_{x\in\bar{A}} P(x) \log\left(\frac{P(x)}{Q(x)}\right) + \sum_{j=0}^{\infty} \sum_{x\in A_j} P(x) \log\left(\frac{Q(x)}{P(x)}\right) = \mathrm{D}_{\mathrm{KL}}(P \parallel Q) + 2\sum_{j=0}^{\infty} \sum_{x\in A_j} P(x) \log\left(\frac{Q(x)}{P(x)}\right).$$

The proof is now concluded by noting that

$$\sum_{j=0}^{\infty} \sum_{x\in A_j} P(x) \log\left(\frac{Q(x)}{P(x)}\right) \leqslant \sum_{j=0}^{\infty} \sum_{x\in A_j} 2^{-j}Q(x)(j+1) \leqslant \sum_{j=0}^{\infty} 2^{-j}(j+1) = 4. \qquad \square$$

2.3 Probability

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

Fact 2.14. *Suppose that $(X_i)_{i=1,\ldots,m}$ is a martingale and $\mathbb{E}\left[X_m^2\right] \leqslant M$. Then for every $\varepsilon > 0$,*

$$\Pr\left[\exists i \text{ such that } |X_i| \geqslant \sqrt{M/\varepsilon}\right] \leqslant \varepsilon.$$

Proof. By Doob's martingale inequality, $\Pr\left[\max_i |X_i| \geqslant \sqrt{M/\varepsilon}\right] \leqslant \frac{\mathbb{E}[|X_m|]}{M/\varepsilon} \leqslant \frac{\sqrt{\mathbb{E}[X_m^2]}}{M/\varepsilon} \leqslant \varepsilon$. □

Definition 2.15. *Let X, Y be random variables over the same universe U. The statistical distance between X, Y is*

$$\mathsf{SD}(X,Y) = \frac{1}{2}\sum_{u\in U}\left|\Pr[X=u] - \Pr[Y=u]\right|.$$

Fact 2.16. *Let X, Y be discrete random variables over the same universe U. Then there is $A \subseteq U$ such that* $\mathsf{SD}(X,Y) = 1 - \sum_{x\in A}\Pr[X=x] - \sum_{x\notin A}\Pr[Y=x]$

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 [GKR16].

Definition 3.1. *For a function $f\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ and a distribution μ over $\{0,1\}^n \times \{0,1\}^n$, we say (f,μ) has (ε,δ) relative-discrepancy with respect to a distribution ρ, if for any rectangle $R = A \times B \subseteq \{0,1\}^n \times \{0,1\}^n$ for which $\rho(R) \geqslant \delta$, it holds that*

1. *$\mu(R \cap f^{-1}(0)) \geqslant (\frac{1}{2} - \varepsilon)\rho(R)$,*
2. *and $\mu(R \cap f^{-1}(1)) \geqslant (\frac{1}{2} - \varepsilon)\rho(R)$.*

In [GKR16], the authors show that if (f,μ) has strong relative-discrepancy, then $\mathrm{CC}_\mu(f, 1/2 - \varepsilon')$ must be high. More precisely, they show if (f,μ) has (ε,δ) discrepancy, then any protocol for f on the distribution μ that achieves advantage of ε', must communicate at least $\log\left(\frac{\varepsilon'-\varepsilon}{\delta}\right)$ bits. The main result of this section strengthens this assertion, as follows.

Theorem 3.2. *Let $M \in \mathbb{N}$, $\delta, \varepsilon > 0$ and let μ be a distribution over $\{0,1\}^n \times \{0,1\}^n$. Suppose $f\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ such that (f,μ) has (ε,δ) relative-discrepancy, and Π is a protocol for computing f such that $\mathrm{I}^{\mathsf{ext}}_\mu[\Pi] \leqslant M$, then*

$$\Pr_{(x,y)\sim\mu}[\Pi(x,y) = f(x,y)] \leqslant \frac{1}{2} + 2000\max\left(\varepsilon, \frac{M}{\log(1/\delta)}\right).$$

Counter positively, if (f,μ) has (ε,δ) relative-discrepancy and Π is a protocol for (f,μ) achieving advantage ε', then the external information of Π according to μ is at least $(\varepsilon' - 2000\varepsilon)\log(1/\delta)$. 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 3.2, we instantiate Theorem 3.2 with the "bursting noise function" from [GKR16], which we present next.

The bursting noise function. Let $k \in \mathbb{N}$ be thought of as large, and set $c = 2^{4^k}$. The bursting noise function, f_{burst} is a pointer chasing function on a tree of height c, however the input distribution μ is supported only on x, y that are very correlated. More precisely, for $b \in \{0,1\}$, we define the distribution μ_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 $p(u)$ denote the ancestor of u from layer i. If $p(u)$ is typical, we again take the bits to be uniform such as $x_u = b \oplus y_u$, and if $p(u)$ is atypical we take x_u, y_u as independently chosen bits.

For $(x,y) \in \mathsf{supp}(\mu_b)$, we define $f_{\mathsf{burst}}(x,y) = b$. We take $\mu = \frac{1}{2}\mu_0 + \frac{1}{2}\mu_1$. We remark that each one of x and y are n-bit Boolean strings where $n = \Theta(2^c) = \Theta(2^{2^{4^k}})$.

In [GKR16], the following two important properties are proved for the bursting noise function.

Lemma 3.3. $\mathrm{IC}_\mu^{\mathsf{internal}}[f, 2^{-k}] = O(k)$.

Lemma 3.4. *The pair (f,μ) has the (ε,δ) relative-discrepancy property with respect to ρ for $\varepsilon = 2^{-k}$ and $\delta = \varepsilon/2^{2^k}$.*

First, we quickly show that Theorem 1.2 follows from Theorem 3.2 and the above lemmas.

Proof of Theorem 1.2. Fix $\eta > 0$ a small constant, and choose k large enough. Using Lemma 3.4 together with 3.2 gives us that $\mathrm{IC}_\mu^{\mathsf{external}}[f,\eta] \geqslant 2^{\Omega(k)}$, whereas Lemma 3.3 gives us that $\mathrm{IC}_\mu^{\mathsf{internal}}[f,\eta] = O(k)$. □

The rest of this section is devoted to the proof of Lemma 3.2, and we begin by giving a proof outline.

Outline of the proof of Theorem 3.2. The proof has two components. Fix a function f and an input distribution μ. In the first step we show that any protocol Π for (f,μ) with low external information, can be converted into a protocol Π' such that (a) Π' has roughly the same error in computing (f,μ), and (b) Π' never reveals too much information about the player's input, with respect to any measure ρ; we refer to such protocols as having "universally low external information" (defined formally below). In the second step, we show that if (f,μ) has the (ε,δ) relative-discrepancy property, then a protocol Π with low universal external information can only have a small advantage of in computing (f,μ). 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 strong relative-discrepancy.

3.1 Universal external information

Suppose Π 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 point $i \in [m]$ in the protocol, a

possible exchange of messages $(a,b) \in \{0,1\}^m \times \{0,1\}^m$, and a pair of inputs $x, y \in \{0,1\}^k$, denote

$$P^x_{A,\Pi,i}(a,b) = \prod_{j<i} \Pr_{(X,Y)\sim\mu} [A_j = a_j \mid A_{<j} = a_{<j}, B_{<j} = b_{<j}, X = x],$$

$$P^y_{B,\Pi,i}(a,b) = \prod_{j<i} \Pr_{(X,Y)\sim\mu} [B_j = b_j \mid A_{\leqslant j} = a_{\leqslant j}, B_{<j} = b_{<j}, Y = y].$$

If this product runs through the whole protocol, i.e. $i = m+1$, we omit the subscript i and simply write $P^x_{A,\Pi}(a,b)$ and $P^y_{B,\Pi}(a,b)$.

Definition 3.5. *With the above notations, we say a protocol Π 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 $\eta(a,b) = \eta_A(a,b)\eta_B(a,b)$ 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*

$$2^{-M} \leqslant \frac{P^x_{A,\Pi}(a,b)}{\eta_A(a,b)} \leqslant 2^M, \qquad 2^{-M} \leqslant \frac{P^y_{B,\Pi}(a,b)}{\eta_B(a,b)} \leqslant 2^M. \tag{11}$$

Informally, a protocol has low universal external information, if for all possible transcript $\pi = (a,b)$, no input of Alice (or Bob) makes π 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 3.6. *Suppose that a protocol Π has universal external information at most M. Then for any distribution ρ over (x,y), we have that $\mathrm{I}^{\mathsf{external}}_\rho[\Pi] \leqslant 2M$.*

Proof. Let η_A, η_B and $\eta = \eta_A \cdot \eta_B$ be from Definition 3.5. First, we argue that for all inputs x, y it holds that $\mathrm{D}_{\mathrm{KL}}\left(\Pi|_{X=x,Y=y} \parallel \eta\right) \leqslant 2M$. Indeed, by definition

$$\mathrm{D}_{\mathrm{KL}}\left(\Pi|_{X=x,Y=y} \parallel \eta\right) = \sum_\pi \Pr\left[\Pi = \pi \mid X = x, Y = y\right] \log\left(\frac{\Pr\left[\Pi = \pi \mid X = x, Y = y\right]}{\eta(\pi)}\right).$$

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

$$\frac{\Pr\left[\Pi = \pi \mid X = x, Y = y\right]}{\eta(\pi)} \leqslant 2^{2M},$$

and plugging that in above yields that $\mathrm{D}_{\mathrm{KL}}\left(\Pi|_{X=x,Y=y} \parallel \eta\right) \leqslant 2M$.

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

$$\mathrm{I}^{\mathsf{external}}_\rho[\Pi] = \mathrm{I}_\rho[X,Y;\Pi] = \mathop{\mathbb{E}}_{(x,y)\sim\rho}\left[\mathrm{D}_{\mathrm{KL}}\left(\Pi|_{X=x,Y=y} \parallel \Pi\right)\right],$$

and therefore

$$\begin{aligned}
\mathop{\mathbb{E}}_{(x,y)\sim\rho}\left[\mathrm{D}_{\mathrm{KL}}\left(\Pi|_{X=x,Y=y}\parallel\eta\right)\right]-\mathrm{I}_{\rho}^{\mathsf{external}}[\Pi] &= \mathop{\mathbb{E}}_{(x,y)\sim\rho}\left[\mathrm{D}_{\mathrm{KL}}\left(\Pi|_{X=x,Y=y}\parallel\eta\right)-\mathrm{D}_{\mathrm{KL}}\left(\Pi|_{X=x,Y=y}\parallel\Pi\right)\right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\rho}\left[\sum_{\pi}\Pr\left[\Pi(x,y)=\pi\right]\log\left(\frac{\Pr_{X,Y\sim\rho}\left[\Pi=\pi\right]}{\eta(\pi)}\right)\right] \\
&= \sum_{\pi}\Pr_{X,Y\sim\rho}\left[\Pi=\pi\right]\log\left(\frac{\Pr_{X,Y\sim\rho}\left[\Pi=\pi\right]}{\eta(\pi)}\right) \\
&= \mathrm{D}_{\mathrm{KL}}\left(\Pi\parallel\eta\right)\geqslant 0.\square
\end{aligned}$$

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 3.7. *Let $\varepsilon,\varepsilon'>0$ and let μ be distributions over $\{0,1\}^n\times\{0,1\}^n$. Suppose $f\colon\{0,1\}^n\times\{0,1\}^n\to\{0,1\}$ is a function, and Π is a protocol (f,μ) that has error at most ε' and $\mathrm{I}_{\mu}^{\mathsf{external}}[\Pi]\leqslant M$. Then there is a protocol Π' for (f,μ) such that*

1. *The error of Π' on (f,μ) is at most $\varepsilon'+40\varepsilon$.*
2. *The universal external information of Π' is at most $2M/\varepsilon+1$.*

We begin by explaining the idea in behind the design of Π'. In Π' we will simulate Π, except that each player will also measure how many bits of external information they have leaked so far (this is possible to do since it only depends on 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 answer according to their answer distribution at that point in Π, conditioned on the transcript so far (but not on their specific input).

Let us now be more precise. Fix a protocol Π such that $I_{\mu}^{\mathsf{ext}}[\Pi]\leqslant M$, and recall the definitions of $P^x_{A,\Pi,i}(a,b)$ and $P^y_{B,\Pi,i}(a,b)$ above. We will also need to define their averaged counterparts, i.e.

$$\eta_{A,i}(a,b)\stackrel{def}{=}P_{A,\Pi,i}(a,b)=\prod_{j<i}\Pr_{(X,Y)\sim\mu}\left[A_j=a_j\mid A_{<j}=a_{<j},B_{<j}=b_{<j}\right],$$

$$\eta_{B,i}(a,b)\stackrel{def}{=}P_{B,\Pi,i}(a,b)=\prod_{j<i}\Pr_{(X,Y)\sim\mu}\left[B_j=b_j\mid A_{\leqslant j}=a_{\leqslant j},B_{<j}=b_{<j}\right],$$

(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 Π' we construct has low universal external information). We note that $P_{A,\Pi,i},P^x_{A,\Pi,i}$ only depend on the $i-1$-prefixes of a and b, and $P_{B,\Pi,i},P^y_{B,\Pi,i}$ only depends on the i-prefix of a and $i-1$-prefix of b. We will therefore sometimes abuse notations and drop the rest of a,b from the notation. We also note that $P^x_{A,\Pi,i},P^y_{A,\Pi,i}$ depend only on x and y (and not on μ). 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 $S_{A,i}(a_{<i}, b_{<i}, x) = \frac{P^x_{A,i}(a_{<i},b_{<i})}{\eta_{A,i}(a,b)}$ and $S_{B,i}(a_{\leqslant i}, b_{<i}, y) = \frac{P^y_{B,i}(a_{\leqslant i},b_{<i})}{\eta_{B,i}(a,b)}$. Intuitively, these quantities measure how much more/ less likely a given exchange of messages (a, b) is, when knowing x and y respectively, compared to only knowing that $(x, y) \sim \mu$. 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 $S_{A,m}(a, b, x) \geqslant 2^M$ (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 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 (and analogously Bob) can compute $S_{A,i}(a_{<i}, b_{<i}, x)$ (analogously $S_{B,i}(a_{\leqslant i}, b_{<i}, y)$) 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 $2^{2M/\varepsilon}$. Indeed, this is what our protocol Π' does.

The protocol Π'. We simulate the protocol Π, with a small change in the beginning of each player's turn. Consider a player intending to send their ith message – say Alice. First, Alice computes $S_{A,i}(\pi_A, \pi_B, x)$ (where π_A are the messages of Alice so far, and π_B are the messages of Bob so far). If this quantity is larger than $2^{2M/\varepsilon}$, or at most $2^{-2M/\varepsilon}$, Alice moves into *"strike mode"*, and otherwise proceeds as usual according to the protocol Π. 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 jth message, for $j \geqslant i$, Alice considers the transcript of the protocol thus far, $a_{<j}, b_{<j}$, and the distribution $A_j(X, Y) \mid A_{<j}(X, Y) = a_{<j}, B_{<j}(X, Y) = a_{<j}$ where $(X, Y) \sim \mu$, and samples her next message according to it. Bob implements an analogous check during his turns.

In the remainder of this section, we argue that the probability that a player ever enters "strike mode" in Π' is small, and so Π' retains roughly the same error as Π on (f, μ). We then show that Π' has universal external information at most $2M/\varepsilon + 1$. We remark that for technical reasons, we will need to assume that our original protocol 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 Π' on μ is comparable to the error of Π

In this section we prove the following lemma.

Lemma 3.8. *The probability that at least one of the players enters "strike mode" in the protocol Π' when ran on μ is at most 38ε.*

First, by Lemma 2.12 we may assume henceforth that the protocol Π is smooth and has error at most $\varepsilon' + \varepsilon$. Thus, once we prove Lemma 3.8 it will follow that the error of Π' is at most $\varepsilon' + 39\varepsilon$. The rest of this section is therefore devoted to the proof of Lemma 3.8.

By Fact 2.4 and the definition of KL-divergence

$$\begin{aligned} M \geqslant \mathrm{I}_\mu[X, Y; \Pi] &= \mathop{\mathbb{E}}_{(x,y)\sim\mu} \left[\mathrm{D}_{\mathrm{KL}}\left(\Pi|_{X=x,Y=y} \parallel \Pi\right)\right] \\ &= \mathop{\mathbb{E}}_{(x,y)\sim\mu} \left[\sum_{a,b} \Pr\left[\Pi(x, y) = (a, b)\right] \log\left(\frac{\Pr\left[\Pi(x, y) = (a, b)\right]}{\Pr\left[\Pi = (a, b)\right]}\right)\right]. \end{aligned} \tag{12}$$

We set up some notations. For $a, b \in \{0,1\}^m$, let $p^A_{a,b,i}(a_i) = \Pr_{(X,Y)\sim\mu}[A_i = a_i \mid A_{<i} = a_{<i}, B_{<i} = b_{<i}]$, and similarly we define for Bob $p^B_{a,b,i}(b_i) = \Pr_{(X,Y)\sim\mu}[B_i = b_i \mid A_{\leqslant i} = a_{\leqslant i}, B_{<i} = b_{<i}]$. Also, define

$$p^A_{a,b,i}(x, a_i) = \Pr_{(X,Y)\sim\mu}[A_i = a_i \mid A_{<i} = a_{<i}, B_{<i} = b_{<i}, X = x]$$

and

$$p^B_{a,b,i}(y, a_i) = \Pr_{(X,Y)\sim\mu}[B_i = b_i \mid A_{\leqslant i} = a_{\leqslant i}, B_{<i} = b_{<i}, Y = y].$$

With these notations, we have $\Pr[\Pi = (a,b)] = \prod_{i=1}^m p^A_{a,b,i}(a_i) \prod_{i=1}^m p^B_{a,b,i}(b_i)$, and for every fixed x, y it holds that $\Pr[\Pi(x,y) = (a,b)] = \prod_{i=1}^m p^A_{a,b,i}(a_i, x) \prod_{i=1}^m p^B_{a,b,i}(b_i, y)$. Thus, (12) gives us that

$$\sum_{i=1}^m \mathop{\mathbb{E}}_{(x,y)\sim\mu}\left[\sum_{a,b}\Pr[\Pi(x,y) = (a,b)]\left(\log\left(\frac{p^A_{a,b,i}(x,a_i)}{p^A_{a,b,i}(a_i)}\right) + \log\left(\frac{p^B_{a,b,i}(x,b_i)}{p^B_{a,b,i}(b_i)}\right)\right)\right] \leqslant M. \tag{13}$$

We consider the two terms on the left hand side separately, i.e. define

$$(I) = \sum_{i=1}^m \mathop{\mathbb{E}}_{(x,y)\sim\mu}\left[\sum_{a,b}\Pr[\Pi(x,y) = (a,b)]\log\left(\frac{p^A_{a,b,i}(x,a_i)}{p^A_{a,b,i}(a_i)}\right)\right], \tag{14}$$

$$(II) = \sum_{i=1}^m \mathop{\mathbb{E}}_{(x,y)\sim\mu}\left[\sum_{a,b}\Pr[\Pi(x,y) = (a,b)]\log\left(\frac{p^B_{a,b,i}(x,b_i)}{p^B_{a,b,i}(b_i)}\right)\right]. \tag{15}$$

We now take a moment to reinterpret these two quantities. Define the random variables $Z_{x,a,b,A,i}(c) = \log\left(\frac{p^A_{a,b,i}(x,c)}{p^A_{a,b,i}(c)}\right)$, $Z_{y,a,b,B,i}(c) = \log\left(\frac{p^B_{a,b,i}(y,c)}{p^B_{a,b,i}(c)}\right)$, where c is 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.

$$E_{x,a,b,A,i} = \sum_{c\in\{0,1\}} p^A_{a,b,i}(x,c) Z_{x,a,b,A,i}(c), \qquad E_{x,a,b,B,i} = \sum_{c\in\{0,1\}} p^B_{a,b,i}(y,c) Z_{x,a,b,B,i}(c).$$

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

$$(I) = \sum_{i=1}^m \mathop{\mathbb{E}}_{(x,y)\sim\mu}\left[\mathop{\mathbb{E}}_{\substack{a_{<i}\sim A_{<i}(x,y)\\ b_{<i}\sim B_{<i}(x,y)}}[E_{x,a,b,A,i}]\right], \qquad (II) = \sum_{i=1}^m \mathop{\mathbb{E}}_{(x,y)\sim\mu}\left[\mathop{\mathbb{E}}_{\substack{a_{\leqslant i}\sim A_{\leqslant i}(x,y)\\ b_{<i}\sim B_{<i}(x,y)}}[E_{y,a,b,B,i}]\right]. \tag{16}$$

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, and so probability that one of the players in Π' enters "strike mode" is

$$\Pr_{\substack{(x,y)\sim\mu \\ a\sim A, b\sim B}}\left[\exists i\in[m] \text{ such that } \left|\sum_{j<i} Z_{x,a,b,A,i}(a_i)\right| \geqslant \frac{2M}{\varepsilon} \text{ or } \left|\sum_{j<i} Z_{x,a,b,B,i}(b_i)\right| \geqslant \frac{2M}{\varepsilon}\right].$$

The rest of the proof is dedicated to upper bound the probability that $\left|\sum_{j\leqslant i} Z_{x,a,b,A,i}(a_i)\right| \geqslant \frac{2M}{\varepsilon}$, and the probability that $\left|\sum_{j<i} Z_{x,a,b,B,i}(b_i)\right| \geqslant \frac{2M}{\varepsilon}$, each by 19ε. Lemma 3.8 thus follows from the union bound. We focus on upper bounding the probability for Alice, and the argument for Bob is analogous.

Claim 3.9. *For each x, y, a, b and i, we have that $E_{x,a,b,A,i}, E_{x,a,b,B,i}$ are non-negative and furthermore*

$$E_{x,a,b,A,i} \geqslant 2(p^A_{a,b,i}(x,1) - p^A_{a,b,i}(1))^2, \qquad E_{x,a,b,B,i} \geqslant 2(p^B_{a,b,i}(x,1) - p^B_{a,b,i}(1))^2.$$

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

$$E_{x,a,b,A,i} = \mathrm{D}_{\mathrm{KL}}\left(A_i|_{X=x,A_{<i}=a_{<i},B_{<i}=b_{<i}} \parallel A_i|_{A_{<i}=a_{<i},B_{<i}=b_{<i}}\right),$$

from which the non-negativity is clear. Since Π is a smooth protocol, we may use Fact 2.7 and conclude that

$$E_{x,a,b,A,i} \geqslant 2(p^A_{a,b,i}(x,1) - p^A_{a,b,i}(1))^2. \qquad \square$$

We note that Claim 3.9 combined with (13) and (16) immediately implies:

Corollary 3.10. *We have that*

$$M \geqslant (I) \geqslant 2\sum_{i=1}^{m} \mathop{\mathbb{E}}_{(x,y)\sim\mu}\left[\mathop{\mathbb{E}}_{\substack{a_{<i}\sim A_{<i}(x,y) \\ b_{<i}\sim B_{<i}(x,y)}}\left[(p^A_{a,b,i}(x,1) - p^A_{a,b,i}(1))^2\right]\right].$$

Consider a random choice of $(x,y) \sim \mu$, $a \sim A(x,y)$ and $b \sim B(x,y)$. Note that the sequence $Q^A_i = Z_{x,a,b,A,i}(a_i) - E_{x,a,b,A,i}$ forms the sum-martingale $G^A_i = \sum_{j\leqslant i} Q^{A_i}$, and the similarly the sequence $Q^B_i = Z_{x,a,b,B,i}(b_i) - E_{x,a,b,B,i}$ forms the sum martingale $G^B_i = \sum_{j\leqslant i} Q^{B_i}$, 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 3.11. $\mathbb{E}\left[(G^A_m)^2\right] \leqslant 18(I) \leqslant 18M$.

Proof. Using the martingale property we have that

$$\mathbb{E}\left[(G^A_m)^2\right] = \sum_{i=1}^{m} \mathbb{E}\left[(Q^{A_i})^2\right] \leqslant \sum_{i=1}^{m} \mathbb{E}\left[Z_{x,a,b,A,i}(a_i)^2\right].$$

Note that fixing $x, a_{<i}, b_{<i}$ we have that

$$\mathbb{E}\left[Z_{x,a,b,A,i}(a_i)^2\right] = \sum_{c\in\{0,1\}} p^A_{a_{<i},b_{<i},i}(x,c)\log^2\left(\frac{p^A_{a_{<i},b_{<i},i}(x,c)}{p^A_{a_{<i},b_{<i},i}(c)}\right).$$

Using the smoothness of Π, we have that $\frac{p^A_{a_{<i},b_{<i},i}(x,c)}{p^A_{a_{<i},b_{<i},i}(c)}$ is at least $\frac{1}{2}$ for all $c\in\{0,1\}$, and since $|\log(z)| \leqslant 2|z-1|$ for all $z \geqslant 1/2$ we get that

$$\mathbb{E}\left[Z_{x,a,b,A,i}(a_i)^2\right] \leqslant 4\sum_{c\in\{0,1\}} p^A_{a_{<i},b_{<i},i}(x,c)\left(\frac{p^A_{a_{<i},b_{<i},i}(x,c) - p^A_{a_{<i},b_{<i},i}(c)}{p^A_{a_{<i},b_{<i},i}(c)}\right)^2.$$

By the smoothness of Π we have $p^A_{a_{<i},b_{<i},i}(c) \geqslant 1/3$ for all $c\in\{0,1\}$, so the above inequality implies that $\mathbb{E}\left[Z_{x,a,b,A,i}(a_i)^2\right] \leqslant 36\left(p^A_{a_{<i},b_{<i},i}(x,1) - p^A_{a_{<i},b_{<i},i}(1)\right)^2$. The claim now follows by combining this with Corollary 3.10. □

We are now ready to prove Lemma 3.8.

Proof of Lemma 3.8. 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 $\sum_{i=1}^{m} E_{x,a,b,A,i} \geqslant M/\varepsilon$ and let W_2 be the event that $\left|G^A_i\right| \geqslant \sqrt{M/\varepsilon}$ for some i. By Markov's inequality we have that

$$\Pr[W_1] \leqslant \frac{(I)}{M/\varepsilon} \leqslant \varepsilon,$$

where we used Corollary 3.10. For W_2, using Fact 2.14 and Claim 3.11 gives that

$$\Pr[W_2] \leqslant \frac{\mathbb{E}\left[(G^A_m)^2\right]}{M/\varepsilon} \leqslant 18\varepsilon.$$

Note that $\left|\sum_{j<k} Z_{x,a,b,A,i}(a_i)\right| = \left|G^A_{k-1}\right| + \sum_{i=1}^{m} E_{x,a,b,A,i}$, so if none of W_1, W_2 hold, then $\left|\sum_{j<i} Z_{x,a,b,A,i}(a_i)\right| \leqslant M/\varepsilon + \sqrt{M/\varepsilon} \leqslant 2M/\varepsilon$ and Alice never enters "strike mode" in the duration on the execution of Π'. Therefore, the probability Alice enters into "strike mode" on an execution of Π' is at most $\Pr[W_1] + \Pr[W_2] \leqslant 19\varepsilon$. The same goes for Bob, and Lemma 3.8 follows from the union bound. □

3.2.2 The protocol Π' has universal external information at most $2M/\varepsilon + 1$

Let $\eta_A = \eta_{A,m+1}$ and $\eta_B = \eta_{B,m+1}$. 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 exhibits that Π' 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 $\frac{P^x_{A,\Pi'}(a,b)}{\eta_A(a,b)} > 2^{2M/\varepsilon+1}$.

We claim that on an execution of Π' on x, y that yields the transcript (a,b), it must be the case that Alice entered "strike mode". Otherwise, we have that:

$$\frac{P^x_{A,\Pi'}(a,b)}{\eta_A(a,b)} = \frac{P^x_{A,\Pi,m}(a,b)}{\eta_{A,m}(a,b)} \cdot \frac{\Pr\left[A_m = a_m \text{ in } \Pi \mid a_{<m}, b_{<m}, x\right]}{\Pr\left[A_m = a_m \text{ in } \Pi \mid a_{<m}, b_{<m}\right]}.$$

Since Alice did not enter "strike mode", the first fraction is between $2^{-2M/\varepsilon}$ and $2^{2M/\varepsilon}$, and since Π is smooth, the second fraction is between $1/2$ and 2. It follows that (11) holds for $2M/\varepsilon + 1$ 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 Π we conclude that

$$S_{A,i}(a_{<i}, b_{<i}, x) = S_{A,i-1}(a_{<i-1}, b_{<i-1}, x)\frac{\Pr\left[A_{i-1} = a_{i-1} \text{ in } \Pi \mid a_{<i-1}, b_{<i-1}, x\right]}{\Pr\left[A_{i-1} = a_{i-1} \text{ in } \Pi \mid a_{<i-1}, b_{<i-1}\right]} \leqslant 2^{2M/\varepsilon} \cdot 2,$$

Now, note that by the behaviour of Alice in strike mode it follows that

$$P^x_{A,\Pi'}(a,b) = P^x_{A,\Pi,i}(a,b) \cdot \prod_{k=i}^{m} \Pr_{(X,Y)\sim\mu}\left[A_k = a_k \text{ in } \Pi \mid a_{<k}, b_{<k}\right],$$

and clearly

$$\eta_{A,m+1}(a,b) = \eta_{A,i}(a,b) \prod_{k=i}^{m} \Pr_{(X,Y)\sim\mu}\left[A_k = a_k \text{ in } \Pi \mid a_{<k}, b_{<k}\right],$$

so

$$\frac{P^x_{A,\Pi'}(a,b)}{\eta_{A,m}(a,b)} = S_{A,i}(a_{<i}, b_{<i}, x) \leqslant 2^{2M/\varepsilon+1},$$

and contradiction. □

3.3 Step (b): strongs 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 strong relative-discrepancy.

Lemma 3.12. *Let $\delta, \varepsilon, \varepsilon' > 0$ and $M \in \mathbb{N}$. Let μ be a distribution over $\{0,1\}^n \times \{0,1\}^n$, and suppose that $f\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ is such that (f,μ) has (ε,δ) relative-discrepancy. If Π is a protocol for (f,μ) whose universal external information at most M, then for $(X,Y) \sim \mu$,*

$$\mathsf{SD}(\Pi(X,Y)|_{f(X,Y)=0}, \Pi(X,Y)|_{f(X,Y)=1}) \leqslant 20\left(\varepsilon + \varepsilon' + \frac{2^{4M}}{\varepsilon'^2}\delta\right).$$

Proof. By Definition 3.5, there are functions $\eta_A(a,b), \eta_B(a,b) \geqslant 0$ such that $\eta(a,b) = \eta_A(a,b)\eta_B(a,b)$ is a distribution, and inequality (11) holds for all x, y, a, b. For each possible transcript π, we partition x, y into rectangles by according to the ratios $\frac{P^x_A(\pi)}{\eta_B(\pi)}$ and $\frac{P^y_B(\pi)}{\eta_B(\pi)}$. Namely, for $-M/\varepsilon' \leqslant i, j < M/\varepsilon'$ we denote

$$R^X_\pi[i] = \left\{x \mid (1+\varepsilon')^i \leqslant \frac{P^x_A(\pi)}{\eta_A(\pi)} \leqslant (1+\varepsilon')^{i+1}\right\}, \quad R^Y_\pi[j] = \left\{y \mid (1+\varepsilon')^j \leqslant \frac{P^y_B(\pi)}{\eta_B(\pi)} \leqslant (1+\varepsilon')^{j+1}\right\},$$

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 $(2M/\varepsilon')^2$, and that they cover the entire domain.

Let ρ be a distribution from Definition 3.1 exhibiting the fact that (f,μ) has (ε,δ) relative-discrepancy. We say a rectangle is heavy if $\rho(R_\pi[i,j]) \geqslant \delta$ and otherwise we say it is light. For future reference, note that the total ρ-weight on light rectangles is at most $(2M/\varepsilon')^2\delta$.

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

$$\begin{aligned}\Pr_{(x,y)\sim\mu}[\Pi(x,y)=\pi, f(x,y)=b] &= \sum_{\substack{(x,y)\\ f(x,y)=b}} \mu(x,y)P_A^x(\pi)P_B^x(\pi) \\ &\geqslant \sum_{(i,j)\in H_\pi}\sum_{\substack{(x,y)\in R_\pi[i,j]\\ f(x,y)=b}} \mu(x,y)P_A^x(\pi)P_B^x(\pi) \\ &\geqslant \sum_{(i,j)\in H_\pi}\sum_{\substack{(x,y)\in R_\pi[i,j]\\ f(x,y)=b}} (1+\varepsilon')^{i+j}\mu(x,y)\eta_A(\pi)\eta_B(\pi),\end{aligned}$$

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

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

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

$$\Pr_{(x,y)\sim\mu}[\Pi(x,y)=\pi, f(x,y)=b] \geqslant \left(\frac{1}{2}-\varepsilon\right)\eta(\pi)\sum_{(i,j)\in H_\pi}(1+\varepsilon')^{i+j}\rho(R_\pi[i,j]), \tag{17}$$

and we analyze the last sum. Note that

$$\begin{aligned}\sum_{(i,j)\in H_\pi}(1+\varepsilon')^{i+j}\rho(R_\pi[i,j]) &= \sum_{(i,j)\in H_\pi}\sum_{(x,y)\in R_\pi[i,j]}(1+\varepsilon')^{i+j}\rho(x,y) \\ &\geqslant \sum_{(i,j)\in H_\pi}\sum_{(x,y)\in R_\pi[i,j]}(1+\varepsilon')^{-2}\frac{P_A^x(\pi)P_B^y(\pi)}{\eta_A(\pi)\eta_B(\pi)}\rho(x,y) \\ &= \sum_{(i,j)\in H_\pi}\sum_{(x,y)\in R_\pi[i,j]}(1+\varepsilon')^{-2}\frac{1}{\eta(\pi)}\rho(x,y)P_A^x(\pi)P_B^y(\pi) \\ &= \frac{(1+\varepsilon')^{-2}}{\eta(\pi)}\Pr_{(X,Y)\sim\rho}[\Pi(X,Y)=\pi, (X,Y)\text{ in a heavy rectangle of }\pi] \\ &= \frac{(1+\varepsilon')^{-2}}{\eta(\pi)}\left(\rho(\pi) - \Pr_{(X,Y)\sim\rho}[\Pi(X,Y)=\pi, (X,Y)\text{ in light rectangle}]\right).\end{aligned} \tag{18}$$

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

$$\Pr_{(X,Y)\sim\rho}[\Pi(X,Y)=\pi, (X,Y)\text{ in a light rectangle of }\pi] = \sum_{(i,j)\notin H_\pi}\sum_{(x,y)\in R_\pi[i,j]}\rho(x,y)\rho|_{x,y}(\pi),$$

and by our earlier notations we have $\rho|_{x,y}(\pi) = P_A^x(\pi)P_B^y(\pi) \leqslant 2^{2M}\eta_A(\pi)\eta_B(\pi) = 2^{2M}\eta(\pi)$, so

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

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

$$\Pr_{(x,y)\sim\mu}[\Pi(x,y)=\pi, f(x,y)=b] \geqslant \left(\frac{1}{2}-\varepsilon\right)(1+\varepsilon')^{-2}\left(\rho(\pi)-\frac{2^{4M}}{\varepsilon'^2}\delta\eta(\pi)\right). \tag{19}$$

We note that summing this up over all π, we get that

$$\Pr_{(x,y)\sim\mu}[f(x,y)=b] \geqslant \left(\frac{1}{2}-\varepsilon\right)(1+\varepsilon')^{-2}(1-\frac{2^{4M}}{\varepsilon'^2}) \geqslant \frac{1}{2}-\varepsilon-\varepsilon'-\frac{2^{4M}}{\varepsilon'^2}. \tag{20}$$

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 $A\subseteq\mathsf{Supp}(\Pi)$ such that this statical distance is equal to

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

Denote $p=\Pr_{(X,Y)\sim\mu}[f(X,Y)=1]$, and note that by (20) we have that $\left|p-\frac{1}{2}\right| \leqslant \varepsilon+\varepsilon'+\frac{2^{4M}}{\varepsilon'^2}$. We get from the above that the statistical distance is equal to

$$\begin{aligned}
&2\left(\frac{1}{2}-\frac{1}{2p}\sum_{\pi\in A}\Pr[f(X,Y)=1,\Pi(X,Y)=\pi]-\frac{1}{2(1-p)}\sum_{\pi\notin A}\Pr[f(X,Y)=0,\Pi(X,Y)=\pi]\right)\\
\leqslant&2\left(\frac{1}{2}-\sum_{\pi\in A}\Pr[f(X,Y)=1,\Pi(X,Y)=\pi]-\sum_{\pi\notin A}\Pr[f(X,Y)=0,\Pi(X,Y)=\pi]\right)\\
&+8(\varepsilon+\varepsilon'+\frac{2^{4M}}{\varepsilon'^2}),
\end{aligned} \tag{21}$$

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

$$\sum_{\pi}\Pr_{(X,Y)\sim\mu}[f(X,Y)=b_\pi,\Pi(X,Y)=\pi] \geqslant \sum_{\pi}\left(\frac{1}{2}-\varepsilon\right)(1+\varepsilon')^{-2}\left(\rho(\pi)-\frac{2^{4M}}{\varepsilon'^2}\delta\eta(\pi)\right),$$

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 $\left(\frac{1}{2}-\varepsilon\right)(1+\varepsilon')^{-2}(1-\frac{2^{4M}}{\varepsilon'^2}\delta) \geqslant \frac{1}{2}-\varepsilon-\varepsilon'-\frac{2^{4M}}{2\varepsilon'^2}\delta$, and plugging this into (21) yields the result. □

3.4 Proof of Theorem 3.2

We can now combine Lemmas 3.7 and 3.12 to deduce Theorem 3.2, as outlined below.

Suppose (f,μ) has (ε,δ) relative-discrepancy with respect to ρ. Let the error of Π be denoted by ε'; if $\varepsilon' \geqslant \frac{1}{2}-2000\varepsilon$ we are done, so assume otherwise. Denote by $\eta=\frac{1}{2}-\varepsilon'$ the advantage of Π. Using Lemma 3.7 (choosing the ε there to be $\eta/160$), we get from Π a protocol Π' whose error is at most $\varepsilon'+\eta/4$ and has universal external information at most $M'=\frac{400M}{\eta}$. Thus, the advantage of Π is at least $\frac{1}{2}-(\varepsilon'+\eta/4)=3\eta/4$. Using Lemma 3.12 (with ε' there to be $\eta/40$), we get that the advantage of Π' is most $20\varepsilon+\frac{1}{2}\eta+\frac{20\cdot 40^2\cdot 2^{4M'}}{\eta^2}\delta$. Combining the upper and lower bound on the advantage of Π', we get

$3\eta/4 \leqslant 20\varepsilon + \frac{1}{2}\eta + \frac{20\cdot 40^2\cdot 2^{4M'}}{\eta^2}\delta$. Simplifying and using $\eta \geqslant 2000\varepsilon$, we get that $\frac{\eta}{10} \leqslant \frac{20\cdot 40^2\cdot 2^{4M'}}{\eta^2}\delta$, and so $\delta \geqslant 2^{-4M'}\eta^3/(10\cdot 20\cdot 40^2) \geqslant 2^{-5M'}$. Taking logarithm gives $5M' \geqslant \log(1/\delta)$ and so by definition of M' we get that $\eta \leqslant \frac{2000M}{\log(1/\delta)}$. □

4 Separating amortized zero-error communication complexity and external information

In this section we prove Theorem 1.3, restated below.

Theorem 1.3 (Restated) . *For large enough m, there is a function $H\colon \{0,1\}^N \times \{0,1\}^N \to \{0,1\}$ and a distribution $\mathcal{D}_H$ over inputs such that $\lim_{q\to\infty}\frac{1}{q}\mathsf{CC}_{\mathcal{D}_H^q}(H^q,0) \leqslant O(\sqrt{m}\log^2 m)$ and $\mathrm{IC}_\nu^{\mathsf{external}}[H,1/16] \geqslant \Omega(m)$.*

We begin by presenting our construction, and then analyze it. In particular, Lemmas 4.4 and 4.8 imply the properties asserted by Theorem 1.3.

4.1 The construction

4.1.1 AND-OR trees

Let $I_1,\ldots,I_{\sqrt{m}}$ be the partition of $[m]$ into equal sized sets given by $I_i = \{\, i\cdot\sqrt{m}+j \mid j=1,\ldots,\sqrt{m}\}$. We define $h\colon \{0,1\}^m \to \{0,1\}$ by

$$h(z) = \bigwedge_{i=1}^{\sqrt{m}} \bigvee_{j\in I_i} z_j.$$

We will need the following easy fact.

Fact 4.1. *For each $z\in\{0,1\}^m$, there exists a certificate for $h(z)$ of size $C(m)$, where $C(m)=O(\sqrt{m})$.*

Consider the function $h_\wedge\colon \{0,1\}^m\times\{0,1\}^m \to \{0,1\}$ whose input is (u,v), and it is defined by $h_\wedge(u,v) = h(u_1\wedge v_1,\ldots,u_m\wedge v_m)$. We will need the following result due to Jayram, Kumar and Sivakumar:

Theorem 4.1. *[JKS03] There exists a distribution $\mathcal{D}_h$ over $\{0,1\}^m\times\{0,1\}^m$, such that $\mathrm{IC}_{\mathcal{D}_h}^{\mathsf{internal}}[h_\wedge,1/8] \geqslant m/100$.*

4.1.2 Our construction: AND-OR trees with a hint

Fix a function h as defined above, and let $\mathcal{D}_h$ be the distribution from Theorem 4.1.

Let $f_{\mathsf{hint}}\colon \{0,1\}^n\times\{0,1\}^n\to\{0,1\}$ be a function, and $\mu = \frac{1}{2}\mu_0+\frac{1}{2}\mu_1$ be a distribution over inputs such that (f_{hint},μ) has high relative discrepancy (we encourage the reader to think of the bursting noise function for sufficiently large n). Here, for each $b\in\{0,1\}$, μ_b is supported on $f_{\mathsf{hint}}^{-1}(b)$. We define the function $H\colon \{0,1\}^m\times\{0,1\}^m\times(\{0,1\}^n\times\{0,1\}^n)^{C(m)\log m}\to\{0,1\}$ as follows. View the input as $(u,x,v,y)\in\{0,1\}^m\times\{0,1\}^{nC(m)\log m}\times\{0,1\}^m\times\{0,1\}^{nC(m)\log m}$, and define

$$H(u,x,v,y) = h_\wedge(u,v)$$

At first glance, one may wonder what is the role of the x, y-part of the input in the function H, as the definition of the function ignores them altogether. The idea is that in the input distribution we consider, the inputs x, y will be used as a pointer to a certificate of h on the input $z = (u_1 \wedge v_1, \ldots, u_m \wedge v_m)$, and we will be able to compute this pointer with low amortized communication complexity (but with error). Once the players compute the certificate set $I[z]$ of h on z, they are able to communicate $C(m) \log m$ bits in order to reveal learn z_i for each $i \in I[z]$. Thus, if the players were successful in computing the certificate set $I[z]$, they above protocol would compute $H(u, x, v, y)$ with no error.

The input distribution $\mathcal{D}_H$. A sample according to the distribution $\mathcal{D}_H$ is drawn in the following way. First, sample $(u, v) \sim \mathcal{D}_h$, and consider the point $z = (u_1 \wedge v_1, \ldots, u_m \wedge v_m)$. By Fact 4.1, there is a certificate for $h(z)$ of size at most $C(m)$. Choose such certificate (in some canonical way), and let $\alpha(u, v) = (\alpha_1, \ldots, \alpha_{C(m)\log m})$ be a binary encoding of it. For each $i = 1, \ldots, C(m) \log m$ independently, we choose $(x^i, y^i) \sim \mu_{\alpha_i}$. The sample of $\mathcal{D}_H$ is now (u, x, v, y).

Choice of the parameters. Let $k \in \mathbb{N}$ be a large parameter, and choose f_{hint} to be the bursting noise function on the tree of height 2^{4^k} and n accordingly. Finally, we choose $m = 2^{k/10}$.

4.2 The zero-error protocol

In this section, we show a zero-error protocol of low amortized communication complexity as asserted in Theorem 1.3. To show that, we will need the following result from [BR14].

Theorem 4.2. *[[BR14]]* $\lim_{q\to\infty} \mathrm{CC}_{\mu^q}[f^q, \varepsilon]/q = \mathrm{IC}^{\mathsf{internal}}_{\mu}[f, \varepsilon]$.

We also need the following easy fact.

Fact 4.3. *Suppose μ, μ_0, μ_1 are distributions such that $\mu = \frac{1}{2}\mu_0 + \frac{1}{2}\mu_1$, and let ν be any convex combination of μ_0, μ_1. Then* $\mathrm{IC}^{\mathsf{internal}}_{\nu}[f, 2\varepsilon] \leqslant 2\mathrm{IC}^{\mathsf{internal}}_{\mu}[f, \varepsilon] + 6$.

Proof. Let Π be a protocol for (f, μ) with ε-error. Note that when executed on μ_0 (or on μ_1), the protocol has at most 2ε error, and hence its error on ν is also at most 2ε. We next upper bound the internal information cost of Π when executed on ν.

Consider the random variables D, X, Y, where we first sample $D \in \{0, 1\}$ uniformly, then sample $(X, Y) \sim \mu_D$. Then $\mathrm{I}[\Pi; X|Y] \geqslant \mathrm{I}[\Pi; X|Y, D] - 1$ and similarly $\mathrm{I}[\Pi; Y|X] \geqslant \mathrm{I}[\Pi; Y|X, D] - 1$, so we get that

$$\mathrm{I}^{\mathsf{internal}}_{\mu}[\Pi] \geqslant \frac{1}{2}\mathrm{I}^{\mathsf{internal}}_{\mu_0}[\Pi] + \frac{1}{2}\mathrm{I}^{\mathsf{internal}}_{\mu_1}[\Pi] - 2,$$

hence $\max(\mathrm{I}^{\mathsf{internal}}_{\mu_0}[\Pi], \mathrm{I}^{\mathsf{internal}}_{\mu_1}[\Pi]) \leqslant 2\mathrm{I}^{\mathsf{internal}}_{\mu}[\Pi] + 4$.

Since ν is a convex combination of μ_0 and μ_1, we may write $\nu = \lambda\mu_0 + (1 - \lambda)\mu_1$ and define random variables D', X', Y' where: $D' = 0$ with probability λ and otherwise $D' = 1$, and then we sample $(X', Y') \sim \mu_D$. We thus have

$$\mathrm{I}[\Pi(X', Y'); X'|Y'] \leqslant \mathrm{I}[\Pi(X', Y'); X'|Y', D'] + 1, \;\; \mathrm{I}[\Pi(X', Y'); Y'|X'] \leqslant \mathrm{I}[\Pi(X', Y'); Y'|X', D'] + 1,$$

and so

$$\mathrm{I}^{\mathsf{internal}}_{\nu}[\Pi] \leqslant \lambda\mathrm{I}^{\mathsf{internal}}_{\mu_0}[\Pi] + (1-\lambda)\mathrm{I}^{\mathsf{internal}}_{\mu_1}[\Pi] + 2 \leqslant \max(\mathrm{I}^{\mathsf{internal}}_{\mu_0}[\Pi], \mathrm{I}^{\mathsf{internal}}_{\mu_1}[\Pi]) + 2 \leqslant 2\mathrm{I}^{\mathsf{internal}}_{\mu}[\Pi] + 6. \;\square$$

Lemma 4.4. $\lim_{q\to\infty} \frac{1}{q}\mathrm{CC}_{\mathcal{D}_H^q}(H^q, 0) \leqslant O(\sqrt{m}\log^2 m)$

Proof. Let $q \in \mathbb{N}$ be the number of copies of H we wish to compute (thought of as very large), and denote the q-inputs to H by $(u(1), x(1), v(1), y(1)), \ldots, (u(q), x(q), v(q), y(q))$. Set $r = C(m)\log m$; we now define r distinct q-tuples of inputs for f. For each $i = 1, \ldots, r$, consider the q-tuple of inputs for f resulting from taking the ith input from each one of $(x(1), y(1)), \ldots, (x(q), y(q))$, i.e. $a(i) = (x(1)^i, \ldots, x(q)^i)$ and $b(i) = (y(1)^i, \ldots, y(q)^i)$. We note that for each i, the distribution of $(a(i), b(i))$ is a product distribution ν_i^q where ν_i is some convex combination of μ_0, μ_1.

Combining Lemma 3.3 and Fact 4.3 we get that for each i, $\mathrm{IC}^{\mathsf{internal}}_{\nu_i}[f, 2^{1-k}] = O(k)$, and so by Theorem 4.2, there is $q(i)$ such that for all $q \geqslant q(i)$, we may find a protocol Π_i communicating $O(qk)$ bits in expectation whose error on each copy of f^q on ν_i^q is at most 2^{1-k}. For the rest of the proof, we take $q \geqslant \max_{i=1,\ldots,r} q(i)$.

We are now ready to present the protocol for H. For each i, the players use the protocol Π_i in order to compute $f^q(a(i), b(i))$. Thus, the players now have a candidate answer for $f(x(j)^i, y(j)^i)$ for each $j = 1, \ldots, q$ and $i = 1, \ldots, r$. The players now check for each $j = 1, \ldots, q$ the subset of coordinates that $(f(x(j)^i, y(j)^i))_{i=1,\ldots,r}$ encodes, call it A_i, and then communicate all of the bits in $u(j), v(j)$ that are in it, i.e. $u(j)^\ell, v(j)^\ell$ for $\ell \in A_i$. If this partial assignment to $h_\wedge$ is indeed a certificate – the players declare the copy i to be successful and thus know the value $H(u(j), x(j), v(j), y(j))$. Otherwise, if copy j is not successful, the players use the trivial protocol for H and exchange $u(j), v(j)$ fully to compute H on that copy (that simply exchanges the players' inputs).

Correctness. It is easy to see that the players are always correct on each copy. They are correct on a "successful copy" by the definition of certificates, and are correct on "unsuccessful copy" since they exchange all of the relevant input bits needed to compute $h_\wedge$ on such copy.

Average communication complexity. Let E_i be the event that copy i is successful, and let $Q = \sum_{i=1}^{q} 1_{E_i}$ be the number of successful copies. Also, denote by A the total number of bits communicated by the simulation of Π_i for $i = 1, \ldots, r$. With these notations, we note that the random variable $A + (q - Q)2m$ bounds the total number of bits communicated by the protocol.

First, note that by choice of Π_i we have that $\mathbb{E}[A] \leqslant O(qkr)$. Secondly we lower bound the expectation of Q. Note that for each i, j, the probability the players computed $f(x(j)^i, y(j)^i)$ correctly is at least $1 - 2^{1-k}$. Thus, for each j, the probability that they computed $f(x(j)^i, y(j)^i)$ correctly for all $i = 1, \ldots, r$ is at least $1 - r2^{1-k} \geqslant 1 - 2^{-9k/10}$. Thus, $\mathbb{E}[Q] \geqslant (1 - 2^{-9k/10})q$.

Overall, we get that the expected number of bits communicated is at most

$$\mathbb{E}[A + (q - Q)2m] \leqslant O(qkr) + 2mq2^{-9k/10} = O(qkr),$$

where in the last inequality we used the fact that $m = 2^{k/10}$. Therefore, we get that

$$\lim_{q\to\infty} \frac{1}{q}\mathrm{CC}_{\mathcal{D}_H^q}(H^q, 0) = O(kr) = O(\sqrt{m}\log^2 m). \qquad \square$$

4.3 Lower bounding the external information of H

In this section, we prove the lower bound on the external information of $(H, \mathcal{D}_H)$ as asserted in Theorem 1.3. The main step in this proof is Lemma 4.5, which asserts that any protocol for $(H, \mathcal{D}_H)$ with low external

information can be converted into a protocol for $(h_{\mathsf{final}}, \mathcal{D}_h)$ with low external information. The second step of the proof is to argue that the latter is impossible, and is essentially the content of Theorem 4.1 (up to the choice of the parameters).

4.3.1 Dropping the hints

Lemma 4.5. *Let $M \in \mathbb{N}$, $\varepsilon, \varepsilon', \xi, \delta > 0$, let $h_\wedge \colon \{0,1\}^m \times \{0,1\}^m \to \{0,1\}$ be as above and let $f \colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ be a function with (ε, δ) relative discrepancy.*

Then any protocol Π for $(H, \mathcal{D}_H)$ with error ε' and $\mathrm{I}^{\mathsf{external}}_{\mathcal{D}_H}[\Pi] \leqslant M$, can be converted into a protocol Π' for $(h_\wedge, \mathcal{D}_h)$ satisfying:

1. *The protocol Π' has universal external information at most $\frac{4M}{\xi}$.*
2. *The error of Π' is at most $\varepsilon' + 61\xi + 20C(m)\log m \cdot \varepsilon + 2^{33C(m)\log mM/\xi^2} \cdot \delta$.*

The rest of this section is devoted to the proof of Lemma 4.5. Denote $r = C(m)\log m$.

Recall the input distribution $\mathcal{D}_H$ of the function H, and consider the distribution $\tilde{\mathcal{D}}_H$ defined as follows. To draw a sample, we take $(u, v) \sim \mathcal{D}_h$, and for each $i = 1, \ldots, r$ take $(x^i, y^i) \sim \mu$ independently, and output (u, x, v, y).

By Lemma 3.7, we may convert Π into a protocol Γ that has error at most $\varepsilon'' = \varepsilon' + 40\xi$ on $(H, \mathcal{D}_H)$ and has universal external information at most $M' = 4M/\xi$. We would like to show that the distribution over the transcript $\Gamma(U, X, V, Y)$ when $(U, X, V, Y) \sim \mathcal{D}_H$, is statistically close to the distribution of the transcript $\Gamma(U, \tilde{X}, V, \tilde{Y})$ where $(U, \tilde{X}, V, \tilde{Y}) \sim \tilde{\mathcal{D}}_H$.

To do so, we consider the hybrid ensembles of random variables. That is, sample $(U, X, V, Y) \sim \mathcal{D}_H$ as well as $\tilde{X}, \tilde{Y} \sim \mu^{C(m)\log m}$, so that the distribution of $(U, \tilde{X}, V, \tilde{Y})$ is $\tilde{\mathcal{D}}_H$. Denote

$$\mathcal{X}^i = (\tilde{X}_1, \ldots, \tilde{X}_i, X_{i+1}, X_{i+2}, \ldots, X_r), \qquad \mathcal{Y}^i = (\tilde{Y}_1, \ldots, \tilde{Y}_i, Y_{i+1}, Y_{i+2}, \ldots, Y_r)$$

and define $\Gamma_i = \Gamma(U, \mathcal{X}^i, V, \mathcal{Y}^i)$. We show that the statistical distance between Γ_i and Γ_{i+1} is small.

Claim 4.6. *For all $i \in \{0, \ldots, r-1\}$, we have that $\mathsf{SD}(\Gamma_i, \Gamma_{i+1}) \leqslant 20\varepsilon + 21\frac{\xi}{r} + 20\frac{r^2 2^{8M'r/\xi}}{\xi^2}\delta$.*

Proof. Since Γ has universal external information at most M', it follows from Lemma 3.6 that the external information of Γ when ran on $(U, \mathcal{X}^{i+1}, V, \mathcal{Y}^{i+1})$ is at most $2M'$, i.e.

$$\mathrm{I}[\Gamma(U, \mathcal{X}^{i+1}, V, \mathcal{Y}^{i+1}); \mathcal{X}^{i+1}, \mathcal{Y}^{i+1}, U, V] \leqslant 2M'.$$

By Fact 2.6, the left hand side is at least

$$\mathrm{I}[\Gamma(U, \mathcal{X}^{i+1}, V, \mathcal{Y}^{i+1}); \tilde{X}_{i+1}, \tilde{Y}_{i+1} \mid \mathcal{X}^{i+1}_{\neq i+1}, Y^{i+1}_{\neq i+1}, U, V].$$

For each $x_{\neq i+1}, y_{\neq i+1}, u, v$ we define

$$M[x_{\neq i}, y_{\neq i}, u, v] = \mathrm{I}[\Gamma(U, \mathcal{X}^{i+1}, V, \mathcal{Y}^{i+1}); \tilde{X}_{i+1}, \tilde{Y}_{i+1} \mid \mathcal{X}^{i+1}_{\neq i+1} = x_{\neq i+1}, \tilde{Y}^{i+1}_{\neq i+1} = y_{\neq i+1}, U = u, V = v].$$

Then we have that $\mathbb{E}\left[M[\mathcal{X}^{i+1}_{\neq i+1}, \tilde{Y}^{i+1}_{\neq i+1}, U, V]\right] \leqslant 2M'$, and thus by Markov's inequality we have that $M[\mathcal{X}^{i+1}_{\neq i+1}, \tilde{Y}^{i+1}_{\neq i+1}, U, V] \leqslant 2rM'/\xi = M''$ with probability at least $1 - \xi/r$; denote this event by E.

Take $(u, v, x_{\neq i+1}, y_{\neq i+1}) \in E$ and condition on $\mathcal{X}^{i+1}_{\neq i+1} = x_{\neq i+1}$, $\mathcal{Y}^{i+1}_{\neq i+1} = y_{\neq i+1}$, $U = u$ and $V = v$. Note that the distribution of $(\tilde{X}_{i+1}, \tilde{Y}_{i+1})$ is μ, and that $\Gamma(U, \mathcal{X}^{i+1}, V, \mathcal{Y}^{i+1})$ is a protocol whose input is $\tilde{X}_{i+1}, \tilde{Y}_{i+1}$ and has mutual information $M[x_{\neq i}, y_{\neq i}, u, v]$ with its the inputs. Therefore, by Lemma 3.12 (applied with $\varepsilon' = \xi/r$) we get that

$$\mathsf{SD}\left(\Gamma(U, \mathcal{X}^{i+1}, V, \mathcal{Y}^{i+1})|_{f(\tilde{X}_{i+1}, \tilde{Y}_{i+1})=0}, \Gamma(U, \mathcal{X}^{i+1}, V, \mathcal{Y}^{i+1})|_{f(\tilde{X}_{i+1}, \tilde{Y}_{i+1})=1}\right) \leqslant 20\left(\varepsilon + \frac{\xi}{r} + \frac{r^2 2^{4M''}}{\xi^2}\delta\right).$$

Let $\alpha_1, \ldots, \alpha_r$ be the encoding of the certificate chosen for $h(u, v)$, and let $b = \alpha_{i+1}$. Then it follows that

$$\mathsf{SD}\left(\Gamma(U, \mathcal{X}^{i+1}, V, \mathcal{Y}^{i+1})|_{f(\tilde{X}_{i+1}, \tilde{Y}_{i+1})=b}, \Gamma(U, \mathcal{X}^{i+1}, V, \mathcal{Y}^{i+1})\right) \leqslant 20\left(\varepsilon + \frac{\xi}{r} + \frac{r^2 2^{4M''}}{\xi^2}\delta\right).$$

Taking average over $\mathcal{X}^{i+1}_{\neq i+1}, \mathcal{Y}^{i+1}_{\neq i+1}, U, V$ and noting that the event E fails with probability at most ξ/r (and then the statistical distance is at most 1), we get that

$$\mathsf{SD}\left(\Gamma(U, \mathcal{X}^{i+1}, V, \mathcal{Y}^{i+1})|_{f(\tilde{X}_{i+1}, \tilde{Y}_{i+1})=\alpha_{i+1}(U,V)}, \Gamma(U, \mathcal{X}^{i+1}, V, \mathcal{Y}^{i+1})\right) \leqslant 20\varepsilon + 21\frac{\xi}{r} + 20\frac{r^2 2^{4M''}}{\xi^2}\delta.$$

The statement of the claim now follows since the distribution $\Gamma(U, \mathcal{X}^{i+1}, V, \mathcal{Y}^{i+1})|_{f(\tilde{X}_{i+1}, \tilde{Y}_{i+1})=\alpha_{i+1}(U,V)}$ is precisely Γ_i, and the distribution of $\Gamma(U, \mathcal{X}^{i+1}, V, \mathcal{Y}^{i+1})$ is precisely Γ_{i+1}. □

Claim 4.7. $\mathsf{SD}(\Gamma(U, X, V, Y), \Gamma(U, \tilde{X}, V, \tilde{Y})) \leqslant 20r\varepsilon + 21\xi + 20\frac{r^3 2^{8M' r/\xi}}{\xi^2}\delta$.

Proof. Since $\Gamma_0 = \Gamma(U, X, V, Y)$ and $\Gamma_r = \Gamma(U, \tilde{X}, V, \tilde{Y})$, the statement of the claim follows by summing Claim 4.6 and using the triangle inequality. □

Set $\eta = 20r\varepsilon + 21\xi + 20\frac{r^3 2^{8M' r/\xi}}{\xi^2}\delta$. Using Claim 4.7, since the error of Γ_0 in computing $(H, \mathcal{D}_H)$ is at most ε'', it follows that Γ_r has error at most $\varepsilon'' + \eta$ in computing $(H, \tilde{\mathcal{D}}_H)$. Note that the distribution of $\tilde{X}, \tilde{Y}$ is completely independent of U, V, and (U, V) is distributed according to $\mathcal{D}_h$. Thus, we may find $\tilde{x}, \tilde{y}$ such that the protocol $\Pi'(U, V) \stackrel{def}{=} \Gamma(U, \tilde{x}, V, \tilde{y})$ has error at most $\varepsilon'' + \eta$ in computing $(h_\wedge, \mathcal{D}_h)$. Bounding $\eta \leqslant 20r\varepsilon + 21\xi + 2^{33rM/\xi^2} \cdot \delta$ and plugging in r gives the claimed bound. □

4.3.2 Concluding the external information lower bound

Lemma 4.8. $\mathrm{IC}^{\mathsf{external}}_{\mathcal{D}_H}[H, 1/16] \geqslant \Omega(m)$.

Proof. Let $\xi = 1/1000$, and suppose we have a protocol Π for $(H, \mathcal{D}_H)$ with external information at most $\xi m/800$ and error ε'. Using Lemma 4.5, we find a protocol Π' for $(h_\wedge, \mathcal{D}_h)$ with external information at most $m/200$, whose error is upper bounded as in Lemma 4.5. However, by Fact 2.10 and Theorem 4.1 we have that $\mathrm{IC}^{\mathsf{external}}_{\mathcal{D}_h}[h_\wedge, 1/8] \geqslant \mathrm{IC}^{\mathsf{internal}}_{\mathcal{D}_h}[h_\wedge, 1/8] \geqslant m/100$, so Π' must have error at least $1/8$. Combining the upper and lower bounds on the error of Π yields that

$$\frac{1}{8} \leqslant \varepsilon' + 61\xi + 20C(m)\log m \cdot \varepsilon + 2^{33m \cdot C(m)\log m/\xi^2} \cdot \delta$$

By the choice of parameters, we have that $m = 2^{k/10}$ and that f has (ε, δ) relative-discrepancy for $\varepsilon = 2^{-k}$, $\delta = \varepsilon/2^{2^k}$ (by Lemma 3.4), so $20C(m)\log m \cdot \varepsilon + 2^{33C(m)\log m M/\xi^2} \cdot \delta \leqslant 2^{-k/2} + 2^{m^2} \cdot \delta \leqslant 2^{-k/4}$. Thus we get that $\varepsilon' \geqslant \frac{1}{8} - 61\xi - 2^{-k/4} \geqslant \frac{1}{16}$, as desired. □

5 Tightness and additional implications

In this section, we prove Theorem 1.5, which asserts that Theorem 1.3 is nearly-tight for protocols with constant error. We then prove that for protocols with zero-error, a better, arbitrarily large, separation holds in the form of Corollary 1.4.

5.1 Proof of Theorem 1.5

Fix (f,μ) as in the Theorem, denote $A = \lim_{q\to\infty} \frac{1}{q}\mathrm{CC}_{\mu^q}(f^q,0)$ and take q such that $\mathrm{CC}_{\mu^q}(f^q,0) \leqslant 2Aq$. Thus, there is a zero-error protocol Π for f^q, whose average communication cost on μ^q is at most $2Aq$. Let $(X_1,Y_1),\ldots,(X_q,Y_q) \sim \mu$ be independent. Then the independence implies that $\mathrm{I}[\Pi;X_i,Y_i] \leqslant \mathrm{I}[\Pi;X_i,Y_i \mid X_{<i},Y_{<i}]$, and so by Fact 2.5

$$\mathop{\mathbb{E}}_{i\in[q]}\left[\mathrm{I}[\Pi;X_i,Y_i]\right] \leqslant \frac{1}{q}\sum_{i=1}^{q}\mathrm{I}[\Pi;X_i,Y_i \mid X_{<i},Y_{<i}] = \frac{1}{q}\mathrm{I}[\Pi;X,Y] \leqslant \frac{1}{q}\mathrm{H}[\Pi] \leqslant \frac{1}{q}\mathbb{E}\left[|\Pi|\right] \leqslant 2A.$$

It follows that there is an $i\in[q]$ such that $\mathrm{I}[\Pi;X_i,Y_i] \leqslant 2A$, and without loss of generality assume $i=1$ is such copy. We first handle the simple case in which $A \leqslant \varepsilon\log(1/\varepsilon)/10$. In this case, by the Data Processing inequality we have $\mathrm{I}[\Pi;f(X_1,Y_1)] \leqslant 2A$, and since Π has zero-error we get that $\mathrm{H}[f(X_1,Y_1)] \leqslant 2A$. Thus, $f(X_1,Y_1)$ is close to constant, i.e. there is $b\in\{0,1\}$ such that $\Pr_{(X_1,Y_1)\sim\mu}\left[f(X_1,Y_1)=b\right] \geqslant 1-\varepsilon$, and we have a trivial protocol for f (in particular $\mathrm{I}^{\mathsf{external}}_{\mu}[f,\varepsilon]=0$). Therefore, we may assume for the rest of the proof that $A \geqslant \varepsilon\log(1/\varepsilon)/10$.

Define the set of good tuples $G = \left\{(x_1,y_1) \mid \mathrm{D}_{\mathrm{KL}}\left(\Pi_{X_1=x_1,Y_1=y_1} \parallel \Pi\right) \leqslant 2A/\varepsilon\right\}$, and note that using Fact 2.4 and Markov's inequality yields that $\mu(G) \geqslant 1-\varepsilon$.

Denote $X=(X_1,\ldots,X_q)$, $Y=(Y_1,\ldots,Y_q)$. For a transcript $\pi=(a,b)$ where $a=(a_1,\ldots,a_m)$ are the messages of Alice and $b=(b_1,\ldots,b_m)$ are the messages of Bob, and inputs x,y define

$$P^x_A(a,b) = \prod_{j<m}\Pr_{(X,Y)\sim\mu}\left[A_j=a_j \mid A_{<j}=a_{<j},B_{<j}=b_{<j},X=x\right],$$

and similarly

$$P^y_B(a,b) = \prod_{j<m}\Pr_{(X,Y)\sim\mu}\left[B_j=b_j \mid A_{\leqslant j}=a_{\leqslant j},B_{<j}=b_{<j},Y=y\right].$$

We will use the protocol Π to construct a protocol Π' for (f,μ), whose input is (x_1,y_1), but first let us introduce some terminology and make some observations. We say a transcript $\pi=(a,b)$ of Π is compatible with Alice, if there exists $(x_2,\ldots,x_q)$ such that $P^x_A(\pi)>0$ for $x=(x_1,\ldots,x_n)$, and analogously for Bob. Note that if π is compatible with both Alice and Bob, then there are extensions x of x_1, and y of y_1, such that the probability that $\Pi(x,y)=\pi$ is $P^x_A(\pi)P^y_B(\pi)>0$. Since Π has zero-error, it means that in that case the value of $f(x_1,y_1)$ is computed correctly in transcript π.

For each π, let R_π be the set of (x_1,y_1) for which π is compatible with both Alice and Bob, and note that R_π is a monochromatic rectangle of f. We say a transcript π is a 0-transcript if the value of $f(x_1,y_1)$ on R_π is 0, and say it is a 1-transcript if the value of $f(x_1,y_1)$ on R_π is 1. Note that if π_0 is a 0-transcript, and π_1 is a 1-transcript, then R_{π_0} and R_{π_1} are disjoint.

We need the following claim, asserting that a randomly chosen transcript is compatible with both players with noticeable probability.

Claim 5.1. *Let $(x_1, y_1) \in G$. Then*

$$\Pr_{\pi\sim\Pi(X_1,Y_1)}[\pi \textit{ is compatible with both Alice and Bob on inputs } (x_1, y_1)] \geqslant 2^{-2A/\varepsilon-17}.$$

Proof. Define $p(\pi) = \Pr_{(X_1,Y_1)\sim\mu}[\Pi(X_1, Y_1) = \pi]$ and $p(\pi \mid x_1, y_1) = \Pr[\Pi(x_1, y_1) = \pi]$, and consider

$$H = \left\{ \pi \in \mathsf{Supp}(\Pi) \mid 0 < p(\pi \mid x_1, y_1) \leqslant 2^{2(A/\varepsilon+8)} p(\pi) \right\}.$$

By Fact 2.13 and the definitions of G and H,

$$\begin{aligned}\sum_{\pi\notin H} p(\pi \mid x_1, y_1) \cdot 2\left(\frac{A}{\varepsilon} + 8\right) \leqslant \sum_{\pi} p(\pi \mid x_1, y_1) \left|\log\left(\frac{p(\pi \mid x_1, y_1)}{p(\pi)}\right)\right| &\leqslant \mathrm{D}_{\mathrm{KL}}\left(\Pi_{X_1=x_1,Y_1=y_1} \parallel \Pi\right) + 8 \\ &\leqslant \frac{A}{\varepsilon} + 8,\end{aligned}$$

so $\sum_{\pi\in H} p(\pi \mid x_1, y_1) \geqslant 1/2$. It follows that

$$\begin{aligned}\Pr_{\pi\sim\Pi(X_1,Y_1)}[\pi \text{ is compatible with both Alice and Bob on inputs } (x_1, y_1)] &\geqslant \sum_{\pi\in H} p(\pi) \\ &\geqslant \sum_{\pi\in H} 2^{-2(A/\varepsilon+8)} p(\pi \mid x_1, y_1) \\ &\geqslant 2^{-2A/\varepsilon-17}.\square\end{aligned}$$

The idea of Π' is to consider a long enough list of possible transcripts of Π, such that almost all input tuples $(x, y) \in G$ have a compatible transcript in the list. Thus, we will have a collection of L monochromatic rectangles that cover most of the mass of μ, and we may invoke the classical argument from [AUY83] that constructs a protocol from monochromatic rectangles. We outline the argument below for completeness.

Using Claim 5.1, there is a list of $L = 2^{2A/\varepsilon+17}\log(1/\varepsilon) = 2^{O(A/\varepsilon)}$ possible transcripts of Π, $\pi_1, \ldots, \pi_L$, and $G' \subseteq G$ with $\mu(G') \geqslant \mu(G) - \varepsilon \geqslant 1 - 2\varepsilon$, such that for each $(x_1, y_1) \in G'$ there is $i \in [L]$ such that π_i is compatible with both Alice and Bob on (x_1, y_1). We now describe $\Pi'(X_1, Y_1)$. The players will try to convince themselves that $f(x_1, y_1) = 0$, and for that they will try to eliminate from the list all 1-transcripts. Formally, at each step of the protocol there is an active set of 1-transcripts, $S \subseteq [L]$, and the goal of the players at each step is either to shrink the size of S by factor 2, or learn the value of $f(x_1, y_1)$.

Suppose that both players are compatible with a 0-transcript τ_0 from the list, and consider all 1-transcripts τ_1 in S. Note that since R_{τ_0} and R_{τ_1} are disjoint, either their x-range is disjoint, or their y-range is disjoint. In particular, it follows that either for the x-range or y-range – say x-range, at least half of the R_{τ_1}'s are disjoint from R_{τ_0} in it. In this case, we say that τ_0 eliminates half of the τ_1's from the point of view of Alice.

Thus, at each step, the player considers all 0-transcripts from the list that are compatible with their input, and checks whether there is at least one that eliminates half of the 1-transcripts in S from their view.

1. If there is, the player chooses one such τ_0 arbitrarily, and sends the index of that transcript in the list. All of the 1-transcripts that are inconsistent with τ_0 from that player's point of view are discarded from S. If τ_0 is also compatible with the other player, the players declare the output to be 0 and terminate, and otherwise the players continue in the protocol.

2. If there is no such τ_0, the player indicates so and passes the turn to the other player.

If both players passed as in item 2 above in consecutive turns, the players declare the output of the function to be 1. Otherwise, the protocol continues until S becomes empty, in which case the players declare the output to be 0.

Analyzing the communication complexity. Note that in each turn, the number of bits communicated is at most $O(A/\varepsilon)$, so to bound the communication complexity of Π' we need to bound the number of turns. Since the list S starts off at size $2^{O(A/\varepsilon)}$ and shrinks by factor 2 at least once every 3 turns, it follows that the number of rounds is $O(A/\varepsilon)$, and so the communication complexity of Π' is at most $O\left(\frac{A^2}{\varepsilon^2}\right)$. In particular, we get that $\mathrm{I}_{\mu}^{\mathsf{external}}[\Pi'] \leqslant O\left(\frac{A^2}{\varepsilon^2}\right)$.

Correctness. We claim the $\Pi'(x_1, y_1)$ is correct whenever $(x_1, y_1) \in G'$, and since $\mu(G') \geqslant 1 - 2\varepsilon$ it follows that it has error at most 2ε when ran on μ. Indeed, note that if $(x_1, y_1) \in G'$ and $f(x_1, y_1) = 0$, then there is τ_0 in the list that is compatible with both of them, hence the players will never pass in two consecutive turns and the output of the protocol will be 0. If $f(x_1, y_1) = 1$, then there is τ_1 a 1-transcript that is compatible with both players, hence it will never be removed from S. Thus, the output of the protocol will be 1.

5.2 Proof of Corollary 1.4

Fix k, and pick (f, μ) from Theorem 1.3. Define $f' \colon \{0,1\}^{n+1} \times \{0,1\}^{n+1} \to \{0,1\}$ in the following way: view the input (x', y') as $x' = (x, a)$, $y' = (y, b)$ where $(x, y) \in \{0,1\}^n \times \{0,1\}^n$, and define $f'(x, y) = f(x, y)$ if $a = b = 1$ and 0 otherwise. Set $p = \frac{1}{\sqrt{k}\log^2 k}$ and consider the distribution ν that puts $(1-p)$ mass uniformly on $x' = (x, a), y' = (y, b)$ such that $a = b = 0$, and puts the rest of its mass on x', y' where $a = b = 1$, i.e. $\nu(x', y') = p\mu(x, y)$ for any such x', y'. We claim that (f, ν) satisfies the properties asserted by Corollary 1.4.

Let Π be a zero-error protocol for (f, ν), and consider the random variables $(D, X' = (X, A), Y' = (Y, B))$ where $D = 0$ with probability p and otherwise $D = 1$. If $D = 0$ we sample $(X, Y) \sim \mu$ and set $A = B = 1$, and otherwise we sample X and Y uniformly form $\{0,1\}^n$ and set $A = B = 0$. Then

$$\mathrm{I}_{\nu}^{\mathsf{external}}[\Pi] = \mathrm{I}[\Pi; X, Y] \geqslant \mathrm{I}[\Pi; X', Y' \mid D] - 1 \geqslant p\mathrm{I}[\Pi; X', Y' \mid D = 0] - 1,$$

so $\mathrm{IC}_{\nu}^{\mathsf{external}}[f', 0] \geqslant p\mathrm{IC}_{\mu}^{\mathsf{external}}[f, 1/16] - 1 \geqslant \Omega(\sqrt{k}/\log^2 k)$.

As for a zero-error protocol with $O(1)$ amortized communication on ν, by assumption there is a protocol Π_q solving (f^m, μ^m) with zero-error and expected communication complexity $O(m\sqrt{k}\log^2 k) + o(m) = O(m/p) + \alpha(m)$, where $\alpha(m)$ is monotone and $\alpha(m)/m \to 0$.

Consider the following protocol Π for (f^q, ν^q). Denote the inputs by $(x'^1, y'^1), \ldots, (x'^q, y'^q)$, and write $x'^i = (x^i, a^i)$ and $y'^i = (y^i, b^i)$. Alice first identifies all copies i such that $a^i = 0$, and communicates them to Bob, so that the answer to that copy is 0. Similarly, Bob identifies all copies i where $b^i = 0$, and communicates them to Alice. Let A be the set of i's such that $a^i = b^i = 1$ (note that both Alice and Bob know A), and set $m = |A|$. The players use the protocol Π_m to solve the copies of f corresponding to A.

Correctness. It is clear that the protocol Π is correct and has zero error.

Amortized communication cost. Let $Q = |A|$ be a random variable. Note that the number of bits transmitted in the phase of the protocol in which trivial copies are identified, is $O(q)$ (corresponding to encodings of two subset of $[q]$). Thus, the expected communication cost of Π is at most $O(q) + \mathbb{E}\left[\mathrm{CC}_{\mu'^Q}(\Pi_Q)\right]$. Conditioning on Q, we have that $\mathrm{CC}_{\mu'^Q}(\Pi_Q) = O(|Q|/p) + \alpha(|Q|) \leqslant O(|Q|/p) + \alpha(q)$. Note that $\mathbb{E}[|Q|] = pq$, so we get that

$$\mathbb{E}\left[\mathrm{CC}_{\mu'^Q}(\Pi_Q)\right] \leqslant \mathbb{E}\left[O(|Q|/p) + \alpha(q)\right] \leqslant O(q) + \alpha(q).$$

In conclusion, the expected communication cost of Π is $O(q) + \alpha(q)$, and since $\alpha(q)/q \to 0$ we conclude that the amortized communication cost of Π is $O(1)$.

References

[ABB+16] Anurag Anshu, Aleksandrs Belovs, Shalev Ben-David, Mika Göös, Rahul Jain, Robin Kothari, Troy Lee, and Miklos Santha. 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*, pages 555–564, 2016.

[ABK16] Scott Aaronson, Shalev Ben-David, and Robin Kothari. 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*, pages 863–876, 2016.

[AC94] Rudolf Ahlswede and Ning Cai. On communication complexity of vector-valued functions. *IEEE Transactions on Information Theory*, 40(6):2062–2067, 1994.

[AUY83] Alfred V. Aho, Jeffrey D. Ullman, and Mihalis Yannakakis. On notions of information transfer in VLSI circuits. In *Proceedings of the 15th Annual ACM Symposium on Theory of Computing, 25-27 April, 1983, Boston, Massachusetts, USA*, pages 133–139, 1983.

[BBCR13] Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to compress interactive communication. *SIAM J. Comput.*, 42(3):1327–1363, 2013.

[BGPW13] Mark Braverman, Ankit Garg, Denis Pankratov, and Omri Weinstein. From information to exact communication. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 151–160, 2013.

[BR14] Mark Braverman and Anup Rao. Information equals amortized communication. *IEEE Trans. Inf. Theory*, 60(10):6058–6069, 2014.

[Bra12] Mark Braverman. Coding for interactive computation: progress and challenges. In *50th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2012*, 2012.

[Bra15] Mark Braverman. Interactive information complexity. *SIAM Journal on Computing*, 44(6):1698–1739, 2015.

[BW15] Mark Braverman and Omri Weinstein. 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*, pages 341–350, 2015.

[BYCKO93] Reuven Bar-Yehuda, Benny Chor, Eyal Kushilevitz, and Alon Orlitsky. Privacy, additional information and communication. *IEEE Transactions on Information Theory*, 39(6):1930–1943, 1993.

[BYJKS04] Ziv Bar-Yossef, Thathachar S Jayram, Ravi Kumar, and D Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004.

[Cov99] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[CSWY01] Amit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 270–278. IEEE, 2001.

[FKNN95] Tomas Feder, Eyal Kushilevitz, Moni Naor, and Noam Nisan. Amortized communication complexity. *SIAM Journal on computing*, 24(4):736–750, 1995.

[GKR16] Anat Ganor, Gillat Kol, and Ran Raz. Exponential separation of information and communication for boolean functions. *J. ACM*, 63(5):46:1–46:31, 2016.

[GKR19] Anat Ganor, Gillat Kol, and Ran Raz. Exponential separation of communication and external information. *SIAM Journal on Computing*, (0):STOC16–236, 2019.

[HJMR10] Prahladh Harsha, Rahul Jain, David A. McAllester, and Jaikumar Radhakrishnan. The communication complexity of correlation. *IEEE Trans. Inf. Theory*, 56(1):438–449, 2010.

[JKS03] T. S. Jayram, Ravi Kumar, and D. Sivakumar. Two applications of information complexity. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing, June 9-11, 2003, San Diego, CA, USA*, pages 673–682, 2003.

[KMSY16] Gillat Kol, Shay Moran, Amir Shpilka, and Amir Yehudayoff. Direct sum fails for zero-error average communication. *Algorithmica*, 76(3):782–795, 2016.

[Wei15] Omri Weinstein. Information complexity and the quest for interactive compression. *ACM SIGACT News*, 46(2):41–64, 2015.

A Missing proofs

A.1 Upper bounding amortized zero-error communication complexity

Theorem A.1. *For every total function $f\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$, it holds that*

$$\lim_{q\to\infty} \frac{1}{q}\mathrm{CC}_{\mu^q}(f^q,0) \leqslant \mathrm{IC}^{\mathsf{external}}_{\mu}[f,0].$$

For the proof, we need the following lemma due to [HJMR10].

Lemma A.2. *Let P, Q be distributions such that $\mathrm{D}_{KL}(P \parallel Q) < \infty$. Then there exists a sampling procedure $\mathcal{P}$, that on an input $(q_1, q_2, \ldots)$ consisting of a list of independent samples from Q, outputs an index $r^\star$ such that the distribution of $q_{r^\star}$ is P, and the expected length of $r^\star$ is at most*

$$\mathrm{D}_{KL}(P \parallel Q) + 2\log(\mathrm{D}_{KL}(P \parallel Q) + 1) + O(1).$$

Proof of Theorem A.1. Let Π be a zero-error protocol for f. We show that there is a sequence of zero-error protocols $(\Gamma_k)_{k\in\mathbb{N}}$ for $(f^k)_{k\in\mathbb{N}}$, such that $\mathrm{CC}(\Gamma_k) = k\mathrm{I}^{\mathsf{external}}_{\mu}[\Pi] + o(k)$, from which the theorem clearly follows.

Let $M = \mathrm{CC}_{\mu}(\Pi) < \infty$. Let k be large, let $(X_1,Y_1),\ldots,(X_k,Y_k) \sim \mu$ be independently sampled (i.e. an input for f^k), and denote by $(x_1,y_1),\ldots,(x_k,y_k) \sim \mu$ a realization of them. The protocol Γ_k we

construct works by rounds, wherein in round t, the player that speaks in that round in Π also speaks. The goal of that player is to communicate to the other player all of the bits that would be sent on that round in $\Pi(X_1, Y_1), \ldots, \Pi(X_k, Y_k)$ conditioned on the transcripts so far and the input of the speaking player.

We describe Γ_k more precisely now. For each $i \in [k]$ and step t, the players maintain a transcript of $\Pi(X_i, Y_i)$ up to step t, which we denote by $\pi_{i,t}$. Thus, denoting $\pi_t = (\pi_{1,t}, \ldots, \pi_{k,t})$, the goal of the players is to correctly sample π_{t+1} conditioned on π_t, and to do so with low expected communication complexity low. Towards this end, we denote by $\Pi(X_i, Y_i)_j$ the jth bit exchanged in the protocol Π on X_i, Y_i, $X = (X_1, \ldots, X_k)$ and $Y = (Y_1, \ldots, Y_k)$. Denote

$$D(\pi_t, t, x, y) = \mathrm{D}_{\mathrm{KL}}\Big(\big(\Pi(X_1, Y_1)_{t+1}, \ldots, \Pi(X_k, Y_k)_{t+1}\big)|_{X=x, Y=y, \Pi(X,Y)_{\leqslant t}=\pi_t} \;\|\; \big(\Pi(X_1, Y_1)_{t+1}, \ldots, \Pi(X_1, Y_1)_{t+1}\big)|_{\Pi(X,Y)_{\leqslant t}=\pi_t}\Big).$$

We now argue that the players can sample π_{t+1} conditioned on $\pi_{i,t}$ by expectedly communicating at most $O(1) + D(\pi_t, t) + 2\log(D(\pi_t, t) + 1)$ bits. Note that regardless of what player speaks, the distribution $\big(\Pi(X_1, Y_1)_{t+1}, \ldots, \Pi(X_k, Y_k)_{t+1}\big)|_{\Pi(X,Y)_{\leqslant t}=\pi_t}$ is known to both players at the $(t+1)$-step, and they therefore can think of their shared string of randomness as a list of samples from it. Assume without loss of generality that Alice speaks. Since her next message only depends on her input and the transcript so far, the distribution $\Pi(X_1, Y_1)_{t+1}, \ldots, \Pi(X_k, Y_k)_{t+1}\big)|_{X=x, Y=y, \Pi(X,Y)_{\leqslant t}=\pi_t}$ is identical to the distribution $\Pi(X_1, Y_1)_{t+1}, \ldots, \Pi(X_k, Y_k)_{t+1}\big)|_{X=x, \Pi(X,Y)_{\leqslant t}=\pi_t}$, and in particular she knows it. Therefore, Alice can use the sampling procedure from Lemma A.2 to pick an index $r^\star$ in their string of randomness that refers to the $r^\star$th sample in their shared randomness string, such that this sample is distributed according to $\Pi(X_1, Y_1)_{t+1}, \ldots, \Pi(X_k, Y_k)_{t+1})|_{X=x, Y=y, \Pi(X,Y)_{\leqslant t}=\pi_t}$. Alice can communicate $r^\star$ to Bob, and then they can continue. We note that the correctness of the sampling, as well as the expected communication cost of the $(t+1)$th step, trivially follow from Lemma A.2.

Thus, Γ_k is a zero-error protocol for f^k, and we next upper bound its expected communication complexity. Let $Z_1, \ldots, Z_k$ denote the number of rounds in the execution of Π on $(X_1, Y_1), \ldots, (X_k, Y_k)$ respectively, and let $T = \max(Z_1, \ldots, Z_k)$. By our analysis of each round, we have that

$$\mathrm{CC}_{\mu^k}(\Gamma_k) \leqslant \sum_{t=0}^{\infty} \mathop{\mathbb{E}}_{x,y}\left[\mathop{\mathbb{E}}_{\pi_t}\left[1_{T \geqslant t}\left(O(1) + D(\pi_t, t, , x, y) + 2\log(D(\pi_t, t, , x, y) + 1)\right)\right]\right]. \tag{22}$$

Clearly, the first term contributes at most $O(\mathbb{E}[T])$, which by Claim A.3 is $o(k)$. Next, we upper bound the contribution from the second term, which will also allow us to bound the last term using Jensen's inequality. To upper bound the contribution of the second term in (22), note that by the chain-rule for conditional KL-divergence we have that

$$\begin{aligned}&\mathop{\mathbb{E}}_{x_i, y_i}\left[\sum_{t=0}^{\infty} \mathop{\mathbb{E}}_{\pi_{i,t}}\left[D(\pi_t, t, x, y)\right]\right] \\ &\quad = \mathop{\mathbb{E}}_{\vec{x}, \vec{y}}\left[\mathrm{D}_{\mathrm{KL}}\left(\big(\Pi(X_1, Y_1), \ldots, \Pi(X_k, Y_k)\big)|_{X=x, Y=y} \;\|\; \Pi(X_1, Y_1), \ldots, \Pi(X_k, Y_k)\big)\right].\end{aligned}$$

By Fact 2.4, this is equal to $\mathrm{I}[X, Y; \Pi(X_1, Y_1), \ldots, \Pi(X_k, Y_k)]$, which by independence between the (X_i, Y_i) for different i's, is equal to $k\mathrm{I}[X_1, Y_1; \Pi(X_1, Y_1)] = k\mathrm{I}_\mu^{\mathsf{external}}[\Pi]$.

For the second term, we note that $\log(1+z) \leqslant \sqrt{z}$ and so

$$\sum_{t=0}^{\infty} \mathop{\mathbb{E}}_{x,y}\left[\mathop{\mathbb{E}}_{\pi_t}\left[2 \cdot 1_{T \geqslant t} \log(D(\pi_t,t,x,y))\right]\right] \leqslant 2\sum_{t=0}^{\infty} \mathop{\mathbb{E}}_{x,y}\left[\mathop{\mathbb{E}}_{\pi_t}\left[1_{T\geqslant t}\sqrt{D(\pi_t,t,x,y)}\right]\right]$$
$$\leqslant 2\sqrt{\sum_{t=0}^{\infty} \mathop{\mathbb{E}}_{x_i,y_i}\left[\mathop{\mathbb{E}}_{\pi_t}\left[1_{T\geqslant t}\right]\right]}\sqrt{\sum_{t=0}^{\infty} \mathop{\mathbb{E}}_{x,y}\left[\mathop{\mathbb{E}}_{\pi_t}\left[D(\pi_t,t,x,y)\right]\right]},$$

where the last inequality is by Cauchy-Schwarz. The first term in this product is $\sqrt{\mathbb{E}[T]} = o(\sqrt{k})$, and the second term in this product is $\sqrt{k\mathrm{I}_{\mu}^{\mathsf{external}}[\Pi]}$, so overall this expression is $o(k)$.

Plugging everything into (22), we see that the expected communication complexity of Γ_k is at most $k\mathrm{I}_{\mu}^{\mathsf{external}}[\Pi] + o(k)$, as desired. □

Claim A.3. *Suppose Z is a non-negative, integer random variable with finite expectation, and let $Z_1, \ldots, Z_k$ be independent copies of Z. Then*

$$\lim_{k\to\infty} \frac{1}{k}\mathbb{E}\left[\max(Z_1,\ldots,Z_k)\right] = 0.$$

Proof. The assumption implies that $\mathbb{E}[Z] = \sum_{r\geqslant 1} \Pr\left[Z \geqslant r\right] < \infty$, so $\lim_{R\to\infty} \sum_{r\geqslant R} \Pr\left[Z\geqslant r\right] = 0$. Let $\varepsilon > 0$. Then there is R (that depends on the distribution of Z) such that $\sum_{r\geqslant R} \Pr\left[Z\geqslant r\right] \leqslant \varepsilon$.

Computing, we get that

$$\frac{1}{k}\mathbb{E}\left[\max(Z_1,\ldots,Z_k)\right] = \frac{1}{k}\sum_{r\geqslant 1}\Pr\left[\max(Z_1,\ldots,Z_k)\geqslant r\right] = \frac{1}{k}\sum_{r\geqslant 1} 1 - \Pr\left[Z<r\right]^k.$$

We split the sum into $r < R$ and $r \geqslant R$. The contribution from $r < R$ is clearly at most R/k, and by our earlier observation the contribution from $r \geqslant R$ is at most

$$\frac{1}{k}\sum_{r\geqslant R} 1 - \Big(1-\Pr\left[Z\geqslant r\right]\Big)^k \leqslant \frac{1}{k}\sum_{r\geqslant R} k\Pr\left[Z\geqslant r\right] \leqslant \varepsilon.$$

It follows that $\limsup \frac{1}{k}\mathbb{E}\left[\max(Z_1,\ldots,Z_k)\right] \leqslant \varepsilon$, and since this is true for all $\varepsilon > 0$ the claim is proved. □

A.2 Non-deterministic external information complexity

In this section pe prove Theorem 1.1:

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

$$\mathrm{IC}_{\mu}^{\mathsf{external},1}[f,0] \leqslant \lim_{q\to\infty} \mathrm{CC}_{\mu^q}^{1^q}[f^q,0]/q.$$

Before proceeding, let us formally define the quantities in the theorem. For a function f and an output a, we say that a distribution of messages $M = M(X,Y)$ along with acceptance functions $Acc_A : (X,M) \mapsto \{0,1\}$, $Acc_B : (Y,M) \mapsto \{0,1\}$ is an a-proof for f, if the following properties hold for all (x,y):

- If $f(x,y) = a$, then for all m with $\Pr[M = m|(x,y)] > 0$, $Acc_A(x,m) = 1$ and $Acc_B(y,m) = 1$;
- if $f(x,y) \neq a$, then for all m, $Acc_A(x,m) = 0$ or $Acc_B(y,m) = 0$.

Then the (average case) non-deterministic amortized communication complexity of a boolean $f(x,y)$ is defined as:

$$\mathrm{CC}^{1^q}_{\mu^q}[f^q, 0] := \inf_{M\colon M \text{ is a } 1^q\text{-proof for } f^q} \mathop{\mathbb{E}}_{(x,y)\sim\mu; m\sim M|_{(x,y)}} |m|. \tag{23}$$

The non-deterministic external information complexity of f is given by:

$$\mathrm{IC}^{\mathsf{external},1}_{\mu}[f, 0] := \inf_{M\colon M \text{ is a 1-proof for } f} \mathrm{I}_{(x,y)\sim\mu; m\sim M|_{(x,y)}}(XY; M). \tag{24}$$

In other words, it's the smallest amount of information a proof that definitively convinces Alice and Bob that $f(x,y) = 1$ can reveal about (x,y) to an outside observer. With the definitions in place, we are ready to prove the theorem.

Proof of Theorem 1.1. We will prove the inequality for any fixed q. Theorem 1.1 then follows by taking $q \to \infty$. For a fixed q, let M^q be a 1^q-proof for f^q such that $\mathbb{E}\,|M^q|$ realizes $\mathrm{CC}^{1^q}_{\mu^q}[f^q, 0]$.

Fix an index $i \in [q]$. Let $M^q_i(x,y)$ be obtained as follows: (1) set $(x_i, y_i) = (x,y)$; (2) pick values $(x_j, y_j) \sim \mu$ for $j \neq i$; (3) sample M^q conditioned on $(x_j, y_j)_{j=1}^q$.

Note that since M^q is a 1^q-proof for f^q, M^q_i is a 1-proof for $f(x_i, y_i)$ for all i. Thus each i gives rise to a valid 1-proof for f. Importantly, to verify that M^q_i is a valid proof that $f(x,y) = f(x_i, y_i) = 1$ the players do not need to know the values of (x_j, y_j) for $j \neq i$.

To complete the proof we only need to show that there exists an i such that

$$\mathrm{I}(M^q_i; XY) = \mathrm{I}(M^q; X_i Y_i) \leqslant \mathbb{E}\,|M^q|/q. \tag{25}$$

To see this, observe that

$$\begin{aligned}\mathbb{E}\,|M^q| \geqslant \mathrm{H}(M^q) \geqslant \mathrm{I}(M^q; X_1Y_1 \ldots X_qY_q) &= \sum_{i=1}^q \mathrm{I}(M^q; X_iY_i | X_{<i}Y_{<i}) \\ = \sum_{i=1}^q (\mathrm{I}(M^q X_{<i}Y_{<i}; X_iY_i) - \mathrm{I}(X_{<i}Y_{<i}; X_iY_i)) &= \sum_{i=1}^q \mathrm{I}(M^q X_{<i}Y_{<i}; X_iY_i) \geqslant \sum_{i=1}^q \mathrm{I}(M^q; X_iY_i).\end{aligned}$$

Therefore, there must exist and $i \in [q]$ such that (25) holds. □