THEORY OF COMPUTING, Volume 14 (6), 2018, pp. 1–73
www.theoryofcomputing.org

SPECIAL ISSUE: CCC 2017

Trading Information Complexity for Error

Yuval Dagan Yuval Filmus* Hamed Hatami† Yaqiao Li

Received June 2, 2017; *Revised* April 16, 2018; *Published* May 15, 2018

Abstract: We consider the standard two-party communication model. The central problem studied in this article is how much one can save in information complexity by allowing an error of ε.

- For arbitrary functions, we obtain lower bounds and upper bounds indicating a gain that is of order $\Omega(h(\varepsilon))$ and $O(h(\sqrt{\varepsilon}))$, respectively. Here h denotes the binary entropy function.
- We analyze the case of the two-bit AND function in detail to show that for this function the gain is $\Theta(h(\varepsilon))$. This answers a question of Braverman et al. (STOC'13).
- We obtain sharp bounds for the set disjointness function of order n. For the case of the distributional error, we introduce a new protocol that achieves a gain of $\Theta(\sqrt{h(\varepsilon)})$ provided that n is sufficiently large. We apply these results to answer another question of Braverman et al. regarding the randomized communication complexity of the set disjointness function.

A conference version of this paper appeared in the Proceedings of the 32nd IEEE Conference on Computational Complexity, 2017 [13].

*This research was supported by the Israel Science Foundation (grant no. 1337/16).

†Supported by NSERC.

ACM Classification: F.1.3

AMS Classification: 68Q17, 68W20

Key words and phrases: communication complexity, information complexity

 DOI: 10.4086/toc.2018.v014a006

- Answering a question of Braverman (STOC'12), we apply our analysis of the set disjointness function to establish a gap between the two different notions of the prior-free information cost. In the light of Braverman (STOC'12), this implies that the amortized randomized communication complexity is not necessarily equal to the amortized distributional communication complexity with respect to the hardest distribution.

As a consequence, we show that the ε-error randomized communication complexity of the set disjointness function of order n is $n[C_{\mathrm{DISJ}} - \Theta(h(\varepsilon))] + o(n)$, where $C_{\mathrm{DISJ}} \approx 0.4827$ is the constant found by Braverman et al. (STOC'13).

Contents

1 Introduction

Communication complexity studies the amount of communication needed to compute a function whose inputs are spread among several parties. It has many applications to different areas of complexity theory and beyond, mostly as a technical tool used for proving lower bounds. Traditionally, communication complexity has been studied through a combinatorial lens. Recently, *information complexity*, the approach to study communication complexity via information theory, has successfully been used to resolve some of the major problems of this field [11, 2, 3]. While communication complexity is concerned with minimizing the amount of communication required for two players to evaluate a function, information complexity is concerned with the amount of information that the communicated bits reveal about the players' inputs.

The study of information complexity is motivated by fundamental questions regarding compressing communication [3, 7, 4, 16] that extend the seminal work of Shannon [29] to the setting where interaction is allowed. Moreover, it has important applications to communication complexity, and in particular to the study of the direct-sum problem [2, 11, 18, 9, 8, 14, 11, 20, 17, 3, 22, 18, 19, 9, 8]. For example, the only known direct-sum result for general randomized communication complexity is proven via information-theoretic techniques in [3].

One of the most spectacular applications of information complexity, due to Braverman et al. [5], is determining the exact first order asymptotics of the communication complexity of set disjointness. Our goal in this paper is to determine the rate of growth of the second-order term (see Equation (1.1) below). Set disjointness, introduced in [31], is one of the most important functions in communication complexity, and as a result it has been studied extensively in the past four decades (see the surveys [12, 30] and the references therein). In this communication problem, which is denoted by DISJ_n, Alice and Bob each receives a subset of $\{1,\dots,n\}$ and they wish to determine whether their sets are disjoint. The goal, first studied in [1], is to determine the asymptotic rate of growth of the *randomized communication complexity* $R_\varepsilon(\mathrm{DISJ}_n)$ of set disjointness, defined as the smallest number of bits exchanged by the two players in a protocol which computes the function correctly with probability at least $1-\varepsilon$ on every input. The correct asymptotic $R_\varepsilon(\mathrm{DISJ}_n)=\Theta(n)$ was first proved by Kalyanasundaram and Schnitger [21]. Although later Razborov [27] gave a shorter proof, still, despite several decades of research in this area, all known proofs of this fact are intricate. It was thus a great breakthrough when Braverman et al. determined the exact constant in the asymptotics of $R_\varepsilon(\mathrm{DISJ}_n)$ as $\varepsilon\to 0$ by employing several recent results from the area of information complexity. They proved that as the error parameter ε tends to 0, the quantity $\lim_{n\to\infty} R_\varepsilon(\mathrm{DISJ}_n)/n$ tends to a constant $C_{\mathrm{DISJ}}\approx 0.4827$.

Our main result determines the asymptotic rate of growth of the error term $C_{\mathrm{DISJ}}-R_\varepsilon(\mathrm{DISJ}_n)$ as $\varepsilon\to 0$: we show that for all $\varepsilon>0$ and all sufficiently large n (depending on ε) we have

$$C_{\mathrm{DISJ}}-\frac{R_\varepsilon(\mathrm{DISJ}_n)}{n}=\Theta(h(\varepsilon)). \tag{1.1}$$

Here the constants implied by the Θ notation are absolute.

As in the work of Braverman et al., we obtain our result by analyzing the information complexity of the 2-bit AND function (in which each player gets one bit). Roughly speaking, the *information complexity* $\mathrm{IC}_\mu(f,\varepsilon)$ of a function f with respect to a distribution μ on the inputs is the minimal amount of information that the players need to leak in any protocol that computes f correctly with probability at least $1-\varepsilon$ on every input.[1] The asymptotic estimate on $R_\varepsilon(\mathrm{DISJ}_n)$ follows by analyzing $\mathrm{IC}^0(\mathrm{AND},\varepsilon) := \min \mathrm{IC}_\mu(\mathrm{AND},\varepsilon)$, where the minimum is taken over all distributions μ such that $\mu(1,1)=0$. Specifically, we prove the following bound:

$$C_{\mathrm{DISJ}} - \mathrm{IC}^0(\mathrm{AND},\varepsilon) = \Theta(h(\varepsilon)). \tag{1.2}$$

The upper bound on $\mathrm{IC}^0(\mathrm{AND},\varepsilon)$ is attained by a protocol having one-sided error, which is only allowed to make a mistake on the input $(1,1)$. It follows from a black-box modification of the optimal protocol for AND found by Braverman et al. The lower bound is significantly harder, requiring several new ideas which could have wider applicability. We sketch these ideas later on in the introduction.

It is natural to ask whether a bound of the form (1.2) holds for arbitrary functions f. Braverman et al. [5] considered this question in the context of distributional information complexity.[2] The *distributional information complexity* $\mathrm{IC}_\mu(f,\mu,\varepsilon)$ of a function f with respect to a distribution μ on the inputs is the minimal amount of information that the players need to leak in any protocol that computes f correctly with probability at least $1-\varepsilon$ *when the inputs are drawn according to* μ. They showed that $\mathrm{IC}_\mu(f,\mu,0) - \mathrm{IC}_\mu(f,\mu,\varepsilon) \le C(f,\mu)h(\varepsilon^{1/8})$, where $C(f,\mu)$ denotes a positive constant which depends only on f and μ. We significantly improve this lower bound on $\mathrm{IC}_\mu(f,\mu,\varepsilon)$, and obtain the first non-trivial upper and lower bounds for general functions:

$$C_1(f,\mu)h(\varepsilon) \le \mathrm{IC}_\mu(f,\mu,0) - \mathrm{IC}_\mu(f,\mu,\varepsilon) \le C_2(f,\mu)h(\sqrt{\varepsilon}), \tag{1.3}$$
$$C_1(f,\mu)h(\varepsilon) \le \mathrm{IC}_\mu(f,0) - \mathrm{IC}_\mu(f,\varepsilon) \le C_2(f,\mu)h(\sqrt{\varepsilon}), \tag{1.4}$$

for positive $C_1(f,\mu)$ and $C_2(f,\mu)$.

Our results hold in both the non-distributional and distributional settings, as well as in the prior-free settings explained below. The upper bounds on $\mathrm{IC}_\mu(f,\mu,\varepsilon)$ and $\mathrm{IC}_\mu(f,\varepsilon)$ use the same black-box technique used to prove the upper bound in (1.2). The lower bounds use *protocol completion*, a new technique which also figures in the proof of the lower bound in (1.2).

In classical communication complexity, the distributional setting arises from an application of Yao's minimax principle: $R_\varepsilon(f)$ is the maximum over μ of the communication complexity of deterministic protocols which compute f correctly with probability at least $1-\varepsilon$ when the inputs are drawn according to μ. This connection suggests searching for an analog of $R_\varepsilon(f)$ in the setting of information complexity. Braverman [4] defined two such notions of *prior-free information complexity*: $\mathrm{IC}(f,\varepsilon) = \max_\mu \mathrm{IC}_\mu(f,\varepsilon)$, and $\mathrm{IC}^D(f,\varepsilon) = \max_\mu \mathrm{IC}_\mu(f,\mu,\varepsilon)$. Using the minimax theorem, he showed that the two notions coincide when $\varepsilon = 0$. He conjectured that the two notions coincide for all ε, but he could only prove the following bound, for $0<\alpha<1$:

$$\mathrm{IC}^D(f,\varepsilon) \le \mathrm{IC}(f,\varepsilon) \le \frac{\mathrm{IC}^D(f,\alpha\varepsilon)}{1-\alpha}. \tag{1.5}$$

[1] There are two different ways to measure information leakage. The usual notion, internal information complexity, measures how much each player learns about the other player's input. External information complexity, studied in this paper only in passing, measures how much an external observer learns about the players' input.

[2] Information complexity and distributional information complexity are often confused in the literature. One reason might be that they are the same in the zero-error prior-free setting, as shown by Braverman [4] and explained further below.

We separate the two notions of prior-free information complexity, thus showing that this tradeoff is essentially optimal for set disjointness:

$$\limsup_{n\to\infty} \frac{\mathrm{IC}^D(\mathrm{DISJ}_n,\varepsilon)}{n} \le C_{\mathrm{DISJ}} - \Omega\left(\sqrt{h(\varepsilon)}\right), \tag{1.6}$$

$$\frac{\mathrm{IC}(\mathrm{DISJ}_n,\varepsilon)}{n} \ge C_{\mathrm{DISJ}} - O(h(\varepsilon)). \tag{1.7}$$

The upper bound on $\mathrm{IC}^D(\mathrm{DISJ}_n,\varepsilon)$ follows from a new protocol for set disjointness which is asymptotically optimal in the distributional prior-free setting, while the lower bound on $\mathrm{IC}(\mathrm{DISJ}_n)$ follows from the proof of (1.1).

Since information complexity is amortized communication complexity, we can also state our separation in terms of communication complexity. Let $R^m_\varepsilon(f^m)$ denote the randomized communication complexity of computing m copies of f with an error of at most ε on each of the m inputs. Similarly, let $D^{\mu,m}_\varepsilon(f^m)$ denote the corresponding distributional notion, where the error is measured when the inputs are drawn according to μ. Braverman [4] showed that

$$\mathrm{IC}(f,\varepsilon) = \lim_{m\to\infty} R^m_\varepsilon(f^m)/m \qquad \text{and} \qquad \mathrm{IC}^D(f,\varepsilon) = \lim_{m\to\infty} \max_\mu D^{\mu,m}_\varepsilon(f^m)/m,$$

and so our separation of $\mathrm{IC}(\mathrm{DISJ}_n,\varepsilon)$ and $\mathrm{IC}^D(\mathrm{DISJ}_n,\varepsilon)$ also separates

$$\max_\mu D^{\mu,m}_\varepsilon(\mathrm{DISJ}^m_n) \qquad \text{and} \qquad R^m_\varepsilon(\mathrm{DISJ}^m_n).$$

Finally, given a function f we characterize all measures μ such that $\mathrm{IC}_\mu(f,0)=0$. We also prove a few results about external information complexity IC^{ext} (which we do not define here). Given a function f we characterize all measures μ such that $\mathrm{IC}^{ext}_\mu(f,0)=0$. We also show that the upper bound $\mathrm{IC}_\mu(f,\varepsilon) \le \mathrm{IC}_\mu(f,0) - \Omega(h(\varepsilon))$ fails for external information complexity:

$$\mathrm{IC}^{ext}_\mu(\mathrm{XOR},\varepsilon) \ge \mathrm{IC}^{ext}_\mu(\mathrm{XOR},0) - 3\varepsilon,$$

where the distribution μ is given by $\mu(0,0)=\mu(1,1)=1/2$.

1.1 Techniques

1.1.1 Stability for the buzzer protocol

At the heart of the lower bound $\mathrm{IC}^0(\mathrm{AND},\varepsilon) \ge C_{\mathrm{DISJ}} - O(h(\varepsilon))$ lies a stability result for almost-optimal protocols for AND.

Braverman et al. [5] gave an optimal protocol for the AND function, which they call the *buzzer protocol*. They also showed that this protocol is essentially the *unique* optimal protocol for the AND function. We prove a stability version of this result: *any ε-error protocol for AND whose information cost is close to that of the buzzer protocol must be similar to the buzzer protocol.*

There are many possible notions of similarity, and ours (for reasons that will become clear below) focuses on the *leaf distribution* of the protocol, which is the distribution of the terminal point of the protocol. Our stability result roughly states that any ε-error protocol for AND whose information cost is close to that of the buzzer protocol must have a leaf distribution which shares some similarity with the leaf distribution of the buzzer protocol.

We prove our stability result by strengthening the technique of *local concavity constraints* introduced by Braverman et al. On the way, we also simplify the arguments of Braverman et al. by replacing the discrete second derivatives used by Braverman et al. with their continuous counterparts.

1.1.2 The buzzer protocol as a random walk

One of our main insights is an alternative description of the buzzer protocol as a random walk.

As part of their analysis of the AND function, Braverman et al. [5] introduced a new perspective on communication protocols, viewing a communication protocol as a random walk on the space of distributions. Given an initial distribution over the inputs, they associate with each node in the protocol tree the a posteriori distribution of the inputs, which is the distribution of the inputs given that the protocol arrives at the node. Instead of walking down the protocol tree, we can think of the protocol as a random walk on these a posteriori input distributions.

Such a random walk has two nice properties: firstly, it is a martingale. Intuitively, this means that if the a posteriori distribution on the inputs is μ at some point in time, then the expected a posteriori distribution after one more step and conditioned on all the history of the protocol is μ. Secondly, one can define a potential function on these distributions such that the information complexity of the protocol is the difference between the potential at the input distribution and the expected potential at the a posteriori distribution where the protocol terminates.

Braverman et al. describe the buzzer protocol as a continuous time protocol which ends abruptly when one of the players buzzes. We give an alternative description of the buzzer protocol, as a random walk on the space of distributions. Consider the case in which the input distribution μ is a product distribution given by $\Pr[X=1]=p$ and $\Pr[Y=1]=q$, where X,Y are the input bits of Alice and Bob, respectively; we denote this distribution succinctly by (p,q). The buzzer protocol is the limit $\varepsilon\to 0$ of a random walk which starts at (p,q), and at each step moves either vertically or horizontally depending on the current distribution (a,b): if $a\geq b$ it moves to $(a,b+\varepsilon)$ or to $(a,b-\varepsilon)$, with probability $1/2$ each, and if $a<b$ it moves to $(a+\varepsilon,b)$ or to $(a-\varepsilon,b)$, with probability $1/2$ each. In both cases we clip the protocol to $[0,1]^2$. The random walk terminates when $a=0$ or $b=0$, in which case it outputs 0, and when $a=b=1$, in which case it outputs 1.

Our description of the buzzer protocol has two main advantages over the original one. First, the a posteriori distribution varies continuously in our protocol. In contrast, in the original description the a posteriori distribution "collapses" when one of the players presses the buzzer. Second, our protocol is the same for all distributions, whereas the original buzzer protocol has an additional symmetrization step to handle asymmetric initial distributions. Both of these properties simplify our analysis.

1.1.3 Product parametrization

Our most important technical innovation is a way of analyzing non-product distributions as if they were product distributions. Since product distributions are often much easier to analyze, we believe this idea could have many further applications, which we hope to explore in future work.

If the input of a protocol is distributed as a product distribution, all the a posteriori input distributions are also product distributions, hence they can be characterized by the marginals corresponding to the distributions on the inputs of Alice and Bob. Additionally, any bit sent by a player changes only the marginal of the a posteriori distribution corresponding to its input while the marginal corresponding to the input of the other player does not change. This is not the case for non-product input distributions and particularly, a bit sent by a player can also change the other marginal.

We show how one can view a protocol with a non-product input distribution as a random walk over the set of product distributions (each product distribution in this random walk can be extracted from the corresponding a posteriori input distributions). This random walk retains the nice properties described above and in Section 1.1.2: it is a martingale where any bit sent by a player changes only the marginal of the a posteriori distribution corresponding to their input. Furthermore, the information complexity of the protocol can be explained in terms of a potential difference between the potential on the initial distribution of the random walk and the expected potential when it terminates.

1.1.4 Protocol completion

We prove the lower bounds on $\mathrm{IC}_\mu(f,\varepsilon)$ and on $\mathrm{IC}^0(\mathrm{AND},\varepsilon)$ using the technique of *protocol completion*. Given an ε-error protocol for f, we complete it to a zero-error protocol for f in a natural way: when the protocol terminates at a posterior distribution ν (which is the distribution of the inputs given the transcript of the protocol and the initial distribution μ), we run a zero-error protocol for f which is information-efficient for the distribution ν. Using the buzzer protocol, we give a protocol for f whose information cost is $O(h(\sqrt{\alpha}))$, where $1-\alpha$ is the probability of the most probable output given ν. Since $\mathbb{E}[\alpha]=\varepsilon$, this shows that we can complete the given ε-error protocol to a zero-error protocol for f at a cost of $O(h(\sqrt{\varepsilon}))$ in the information cost, implying the bound $\mathrm{IC}_\mu(f,\varepsilon)+O(h(\sqrt{\varepsilon}))\ge \mathrm{IC}_\mu(f,0)$.

For the case $f=\mathrm{AND}$, we are able to improve on this result, tightening the gap from $O(h(\sqrt{\varepsilon}))$ to $O(h(\varepsilon))$, using the stability result for the buzzer protocol. The product parametrization allows us to consider the posterior distribution ν as a product distribution (a,b). If $\max(a,b)=\Omega(1)$ then the buzzer protocol has information cost $O(h(\alpha))$ rather than just $O(h(\sqrt{\alpha}))$ (recall that $1-\alpha$ is the probability of the most probable output given ν). Suppose now that we are given an ε-error protocol π for AND. Our goal is to prove that $\mathrm{IC}_\mu(\pi)\ge \mathrm{IC}_\mu(\mathrm{AND},0)-C_\mu h(\varepsilon)$ for some $C_\mu>0$ (here $\mathrm{IC}_\mu(\pi)$ is the information cost of π). We can complete π to a zero-error protocol π_0 at a cost of $O(h(\sqrt{\varepsilon}))$. We can assume that $\mathrm{IC}_\mu(\pi_0)\le \mathrm{IC}_\mu(\mathrm{AND},0)-C_\mu h(\varepsilon)+O(h(\sqrt{\varepsilon}))$, and so π_0 is an almost-optimal protocol for AND. Our stability result shows that a random leaf (a,b) of π_0 satisfies $\max(a,b)\ge c_\mu$ with high probability, for some $c_\mu>0$. It follows that the same holds for π, and so the cost of completion is only $O(h(\varepsilon))$.

1.1.5 Black-box modification

We prove the upper bounds on $\mathrm{IC}_\mu(f,\varepsilon)$ and (as a special case) on $\mathrm{IC}^0(\mathrm{AND},\varepsilon)$ using a simple black-box argument, which modifies an optimal zero-error protocol to a slightly more information-efficient ε-error protocol. Given a zero-error protocol π for f, one way to create an ε-error protocol for f is to run π with probability $1-\varepsilon$, and output some constant value with probability ε. However, this only saves $O(\varepsilon)$ bits of information. Our modification is different: we identify a player P and two inputs z_0,z_1, and run the following protocol π':

- With probability ε (sampled privately by P), if the input of P is z_1 then P changes its input to z_0.
- The players run π on their possibly modified inputs.

This is also an ε-error protocol, and for a suitable choice of the parameters, it turns out that it saves $\Omega(h(\varepsilon))$ bits of information compared to π.

When the input distribution μ has full support, it is easy to choose the parameters, by finding two inputs $(x_0,y_0),(x_1,y_1)$ which differ on a single coordinate such that $f(x_0,y_0) \neq f(x_1,y_1)$. Such a choice might not exist when μ doesn't have full support, and instead we rely on a rather delicate binary search argument on the set of transcripts.

We can apply this argument to the AND function, showing that $\mathrm{IC}_\mu(\mathrm{AND},\varepsilon) \leq C_{\mathrm{DISJ}} - \Omega(h(\varepsilon))$. However, when using this result to obtain a protocol for set disjointness, we encounter a difficulty: in order to obtain an ε-error protocol for DISJ_n, it seems at first that we need a protocol for AND having error ε/n. This would result in a saving of $O(h(\varepsilon/n))$ rather than $O(h(\varepsilon))$ per coordinate. A similar difficulty was encountered by Molinaro et al. [26] in a similar context, and they overcame it using protocols that abort. In our case there is a simpler solution: we consider ε-error protocols for AND which only make one-sided error, outputting 0 when the correct answer is 1 (the black-box argument can be modified to produce such protocols). If we apply such a protocol coordinatewise to compute the intersection of X,Y, then we always compute the intersection correctly when X,Y are disjoint, and we mistakenly compute the intersection to be empty when X,Y are not disjoint with probability at most $\varepsilon^{|X\cap Y|} \leq \varepsilon$. The resulting protocol thus computes set disjointness correctly with probability at least $1-\varepsilon$ on every input.

1.1.6 Computing set disjointness with error

The lower bound $\mathrm{IC}^0(\mathrm{AND},\varepsilon) \geq C_{\mathrm{DISJ}} - O(h(\varepsilon))$ implies a similar lower bound on the information complexity of set disjointness: $\mathrm{IC}(\mathrm{DISJ}_n,\varepsilon)/n \geq C_{\mathrm{DISJ}} - O(h(\varepsilon))$. In contrast, we can save more than $h(\varepsilon)$ in the distributional prior-free setting:

$$\limsup_{n\to\infty} \frac{\mathrm{IC}^D(\mathrm{DISJ}_n,\varepsilon)}{n} \leq C_{\mathrm{DISJ}} - \Omega\left(\sqrt{h(\varepsilon)}\right). \tag{1.8}$$

A minimax argument of Braverman [4] shows that this bound is tight. We prove this upper bound on $\mathrm{IC}^D(\mathrm{DISJ}_n,\varepsilon)$ using a new protocol for set disjointness. Given a distribution μ, we describe a protocol π which has error ε with respect to μ, whose information cost satisfies

$$\mathrm{IC}_\mu(\pi) \leq n\left[C_{\mathrm{DISJ}} - \Omega\left(\sqrt{h(\varepsilon)}\right)\right] + O(\log n). \tag{1.9}$$

Let p be the probability the input sets X,Y are not disjoint, when $(X,Y) \sim \mu$. The protocol proceeds as follows:

- Using public randomness, Alice and Bob sample a permutation σ on $1,\ldots,n$.
- For $i = 1,\ldots,n$, Alice and Bob run a protocol for AND on $X_{\sigma(i)},Y_{\sigma(i)}$ which has one-sided error $\varepsilon/2p$ with respect to the conditional distribution of $X_{\sigma(i)},Y_{\sigma(i)}$, declaring X,Y to be not disjoint (and halting the protocol) if the AND protocol answers $X_{\sigma(i)} = Y_{\sigma(i)} = 1$.

- Declare X,Y to be disjoint.

The protocol only makes an error when the inputs are not disjoint, and in that case it makes an error with probability $(\varepsilon/2p)^{|X\cap Y|} \leq \varepsilon/2p$. Since the inputs are non-disjoint with probability p, the overall error probability is $\varepsilon/2 < \varepsilon$. A tricky but standard argument shows that this protocol saves roughly $\Omega(n\sqrt{h(\varepsilon)})$ bits of information.

2 Preliminaries

In this section we introduce some basic notation and facts, and review the necessary background for the paper.

2.1 Notation and basic estimates

We typically denote random variables by capital letters (e. g., A,B,C,Π). For the sake of brevity, we shall write $A_1 \dots A_n$ to denote the random variable $(A_1,\dots,A_n)$ and *not* the product of the A_i's. We use $[n]$ to denote the set $\{1,\dots,n\}$, and $\operatorname{supp}\mu$ to denote the support of a measure μ.

For a finite set Ω, we denote by $\Delta(\Omega)$, the set of all discrete probability distributions on Ω. For $\mu,\nu \in \Delta(\Omega)$, we denote their *total variation distance* with

$$|\mu-\nu| := \frac{1}{2}\sum_{x\in\Omega}|\mu(x)-\nu(x)|. \tag{2.1}$$

For every $\varepsilon \in [0,1]$, $h(\varepsilon) = -\varepsilon\log\varepsilon - (1-\varepsilon)\log(1-\varepsilon)$ denotes the *binary entropy*, where here and throughout the paper $\log(\cdot)$ is in base 2, and $0\log 0 = 0$.

2.2 Communication complexity

The notion of two-party communication complexity was introduced by Yao [31] in 1979. In this model there are two players (with unlimited computational power), often called Alice and Bob, who wish to collaboratively perform a task such as computing a given function $f\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$. Alice receives an input $x\in\mathcal{X}$ and Bob receives $y\in\mathcal{Y}$. Neither of them knows the other player's input, and they wish to communicate in accordance with an agreed-upon protocol π to compute $f(x,y)$. The protocol π specifies as a function of (only) the transmitted bits whether the communication is over, and if not, who sends the next bit. Furthermore π specifies what the next bit must be as a function of the transmitted bits, and the input of the player who sends the bit. We will assume that when the protocol terminates Alice and Bob agree on a value as the output of the protocol. We denote this value by $\pi(x,y)$. The *communication cost* of π is the total number of bits transmitted on the worst case input. The *transcript* of an execution of π is a string Π consisting of a list of all the transmitted bits during the execution of the protocol. As protocols are defined using protocol trees, transcripts are in one-to-one correspondence with the leaves of this tree.

In the randomized communication model, the players might have access to a shared random string (*public randomness*), and their own private random strings (*private randomness*). These random strings are independent, but they can have any desired distributions individually. In the randomized model the *transcript* also includes the public random string in addition to the transmitted bits. Similar to the case of deterministic protocols, the *communication cost* is the total number of bits transmitted on the worst case input and random strings. The *average communication cost* of the protocol is the expected number of bits transmitted on the worst case input.

For a function $f\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$ and a parameter $\varepsilon>0$, we denote by $R_\varepsilon(f)$ the communication cost of the best randomized protocol that computes the value of $f(x,y)$ correctly with probability at least $1-\varepsilon$ for *every* (x,y).

2.3 Information complexity

The setting is the same as in communication complexity, where Alice and Bob (having infinite computational power) wish to mutually compute a function $f\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$. To be able to measure information, we also need to assume that there is a prior distribution μ on $\mathcal{X}\times\mathcal{Y}$.

For the purpose of communication complexity, once we allow public randomness, it makes no difference whether we permit the players to have private random strings or not. This is because the private random strings can be simulated by parts of the public random string. However, for information complexity, it is crucial to permit private randomness, and once we allow private randomness, public randomness becomes inessential. Indeed, one of the players can use her private randomness to generate the public random string, and then transmit it to the other player. Although this might have very large communication cost, it has no information cost, as it does not reveal any information about the players' inputs.

Probably the most natural way to define the information cost of a protocol is to consider the amount of information that is revealed about the inputs X and Y to an external observer who sees the transmitted bits and the public randomness. This is called the *external information cost* and is formally defined as the mutual information between XY and the transcript of the protocol (recall that the transcript also contains the public random string). While this notion is interesting and useful, it turns out there is a different way of defining the information cost that enjoys certain desirable properties that the external information cost lack. This is called the *internal information cost* or just the *information cost* for short, and is equal to the amount of information that Alice and Bob learn about each other's inputs from the communication. Note that Bob knows Y, the public randomness R, and his own private randomness R_B, and thus what he learns about X *from the communication* can be measured by the conditional mutual information $I(X;\Pi\mid YRR_B)$. Similarly, what Alice learns about Y from the communication can be measured by $I(Y;\Pi\mid XRR_A)$ where R_A is Alice's private random string. It is not difficult to see [3] that conditioning on the public and private randomness does not affect these quantities. In other words $I(X;\Pi\mid YRR_B)=I(X;\Pi\mid Y)$ and $I(Y;\Pi\mid XRR_A)=I(Y;\Pi\mid X)$. We summarize these in the following definition.

Definition 2.1. The *internal information cost* and the *external information cost* of a protocol π with respect to a distribution μ on inputs from $\mathcal{X}\times\mathcal{Y}$ are defined as

$$\mathrm{IC}_\mu(\pi)=I(\Pi;X\mid Y)+I(\Pi;Y\mid X), \tag{2.2}$$

and

$$\mathrm{IC}_\mu^{ext}(\pi)=I(\Pi;XY), \tag{2.3}$$

respectively, where $\Pi=\Pi_{XY}$ is the transcript of the protocol when it is executed on $XY\sim\mu$.

We will be interested in certain *communication tasks*. Let $[f,\varepsilon]$ denote the task of computing the value of $f(x,y)$ correctly with probability at least $1-\varepsilon$ for *every* (x,y). Thus a protocol π performs this task if

$$\Pr[\pi(x,y) \neq f(x,y)] \leq \varepsilon, \quad \forall\, (x,y) \in \mathcal{X} \times \mathcal{Y}. \tag{2.4}$$

Given another distribution ν on $\mathcal{X} \times \mathcal{Y}$, let $[f,\nu,\varepsilon]$ denote the task of computing the value of $f(x,y)$ correctly with probability at least $1-\varepsilon$ if the input (x,y) is sampled from the distribution ν. A protocol π performs this task if

$$\Pr_{(x,y)\sim\nu}[\pi(x,y) \neq f(x,y)] \leq \varepsilon. \tag{2.5}$$

Note that a protocol π performs $[f,0]$ if it computes f correctly on *every* input while performing $[f,\nu,0]$ means computing f correctly on the inputs that belong to the support of ν.

We will also need a one-sided version of the task $[f,\varepsilon]$. Let $[f,\varepsilon,z_1 \to z_0]$ denote the task of computing the value of $f(x,y)$ correctly with probability at least $1-\varepsilon$ for *every* (x,y), allowing the protocol to err only if it outputs z_0 instead of z_1. Thus a protocol π performs this task if it performs the task $[f,\varepsilon]$, and additionally

$$\pi(x,y) \neq f(x,y) \Longrightarrow f(x,y) = z_1 \text{ and } \pi(x,y) = z_0. \tag{2.6}$$

The *information complexity* of a communication task T with respect to a measure μ is defined as

$$\mathrm{IC}_\mu(T) = \inf_{\pi:\ \pi \text{ performs } T} \mathrm{IC}_\mu(\pi). \tag{2.7}$$

It is essential here that we use infimum rather than minimum as there are tasks for which there is no protocol that achieves $\mathrm{IC}_\mu(T)$ while there is a sequence of protocols whose information cost converges to $\mathrm{IC}_\mu(T)$. The *external information complexity* of a communication task T is defined similarly. We will abbreviate $\mathrm{IC}_\mu(f,\varepsilon) = \mathrm{IC}_\mu([f,\varepsilon])$, $\mathrm{IC}_\mu(f,\nu,\varepsilon) = \mathrm{IC}_\mu([f,\nu,\varepsilon])$, etc. It is important to note that when μ does not have full support, $\mathrm{IC}_\mu(f,\mu,0)$ can be strictly smaller than $\mathrm{IC}_\mu(f,0)$.

Remark 2.2 (A warning regarding notation)**.** In the literature of information complexity it is common to use "$\mathrm{IC}_\mu(f,\varepsilon)$" to denote the distributional error case, i. e., what we denote by $\mathrm{IC}_\mu(f,\mu,\varepsilon)$. Unfortunately this has become the source of some confusions in the past, as sometimes "$\mathrm{IC}_\mu(f,\varepsilon)$" is used to denote both of the distributional error and the point-wise error cases. To avoid ambiguity we distinguish the two cases by using the different notations $\mathrm{IC}_\mu(f,\mu,\varepsilon)$ and $\mathrm{IC}_\mu(f,\varepsilon)$.

Similar to the fact that the maximal distributional communication complexity over all measures equals the public coin randomized communication complexity (see, e. g., [23, Section 3.4]), below we prove a lemma that establishes a similar relation between $\mathrm{IC}_\mu(f,\nu,\varepsilon)$ and $\mathrm{IC}_\mu(f,\varepsilon)$.

Lemma 2.3. $\mathrm{IC}_\mu(f,\varepsilon) = \max_\nu \mathrm{IC}_\mu(f,\nu,\varepsilon)$ *holds for all* $\varepsilon \geq 0$.

Note that the maximum exists due to continuity of $\mathrm{IC}_\mu(f,\nu,\varepsilon)$ with respect to ν, a fact that is discussed later in Section 2.4. (For $\varepsilon = 0$ one can take any full-support ν.)

Proof. We only need to show $\mathrm{IC}_\mu(f,\varepsilon) \leq \max_\nu \mathrm{IC}_\mu(f,\nu,\varepsilon)$ as the other direction is obvious. The proof is an application of von Neumann's minimax theorem.

Pick a small $\delta > 0$, let

$$C_\delta = \{\pi : \mathrm{IC}_\mu(\pi) \le \mathrm{IC}_\mu(f,\varepsilon) - \delta\}.$$

Although C_δ is an infinite set, we can approximate it by a finite set by considering only the protocols with bounded communication cost that use only a bounded number of unbiased random bits. This process does not affect the validity of the proof, and hence the minimax theorem is still applicable.

Consider a two-player zero-sum game in which Alice chooses a protocol $\pi \in C_\delta$ and Bob chooses an input $(x,y) \in \mathcal{X} \times \mathcal{Y}$, and define the utility for Alice to be $\Pr[\pi(x,y) = f(x,y)]$. Note that a mixed strategy for Alice is still just a protocol, and a mixed strategy for Bob corresponds to a probability measure on $\mathcal{X} \times \mathcal{Y}$. By our definition of C_δ and the minimax theorem, we have

$$\min_\nu \max_\pi \mathop{\mathbb{E}}_{(x,y)\sim\nu} \Pr[\pi(x,y) = f(x,y)] = \max_\pi \min_\nu \mathop{\mathbb{E}}_{(x,y)\sim\nu} \Pr[\pi(x,y) = f(x,y)] = 1 - \varepsilon - t(\delta) < 1 - \varepsilon, \quad (2.8)$$

where $t(\delta) > 0$ is a positive quantity. This means that there exists a measure ν_δ^* such that for all $\pi \in C_\delta$,

$$\mathop{\mathbb{E}}_{(x,y)\sim\nu_\delta^*} \Pr[\pi(x,y) \ne f(x,y)] > \varepsilon.$$

Letting $\delta \to 0$ gives $\max_\nu \mathrm{IC}_\mu(f,\nu,\varepsilon) \ge \mathrm{IC}_\mu(f,\varepsilon)$ as desired. □

Finally let us recall the two definitions of the prior-free notions of information complexity introduced in [4]. The *max-distributional information complexity* of a function $f\colon \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ is defined as

$$\mathrm{IC}^D(f,\varepsilon) = \max_\mu \mathrm{IC}_\mu(f,\mu,\varepsilon). \quad (2.9)$$

The information complexity of f with error ε is defined as

$$\mathrm{IC}(f,\varepsilon) = \inf_\pi \max_\mu \mathrm{IC}_\mu(\pi), \quad (2.10)$$

where the infimum is over all protocols π that perform the task $[f,\varepsilon]$. It is possible [4] to use a minimax argument and the concavity of $\mathrm{IC}_\mu(\pi)$ with respect to μ to show that

$$\mathrm{IC}(f,\varepsilon) = \inf_\pi \max_\mu \mathrm{IC}_\mu(\pi) = \max_\mu \inf_\pi \mathrm{IC}_\mu(\pi) = \max_\mu \mathrm{IC}_\mu(f,\varepsilon) = \max_{\mu,\nu} \mathrm{IC}_\mu(f,\nu,\varepsilon), \quad (2.11)$$

where the last equality follows from Lemma 2.3.

2.4 The continuity of information complexity

It is shown in [6, Lemma 4.4] that for every communication task T, $\mathrm{IC}_\mu(T)$ is uniformly continuous with respect to μ. More precisely, for every two measures μ_1 and μ_2 with $|\mu_1 - \mu_2| \le \delta$ (the distance is in total variation distance), we have

$$|\mathrm{IC}_{\mu_1}(T) - \mathrm{IC}_{\mu_2}(T)| \le 2\log(|\mathcal{X} \times \mathcal{Y}|)\delta + 2h(2\delta). \quad (2.12)$$

The information complexity functions $\mathrm{IC}_\mu(f,\varepsilon)$ and $\mathrm{IC}_\mu(f,\nu,\varepsilon)$ are both continuous with respect to ε. The following simple lemma from [4] proves continuity for $\varepsilon \in (0,1]$. The continuity at 0 is more complicated and is proven in [5]. (See also Theorem 3.5 and Theorem 3.6 below.)

Lemma 2.4. *[4] For every $f\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$, $\varepsilon_2>\varepsilon_1>0$ and measures μ,ν on $\mathcal{X}\times\mathcal{Y}$, we have*

$$\mathrm{IC}_\mu(f,\nu,\varepsilon_1)-\mathrm{IC}_\mu(f,\nu,\varepsilon_2)\le(1-\varepsilon_1/\varepsilon_2)\log|\mathcal{X}\times\mathcal{Y}|, \tag{2.13}$$

and

$$\mathrm{IC}_\mu(f,\varepsilon_1)-\mathrm{IC}_\mu(f,\varepsilon_2)\le(1-\varepsilon_1/\varepsilon_2)\log|\mathcal{X}\times\mathcal{Y}|. \tag{2.14}$$

Proof. Consider a protocol π with information cost I, and error $\varepsilon_2>0$. Here we can consider the distributional error as in (2.13) or the point-wise error as in (2.14). Set $\delta=1-\varepsilon_1/\varepsilon_2$, and let τ be the protocol that with probability $1-\delta$ runs π, and with probability δ Alice and Bob exchange their inputs and compute $f(x,y)$ correctly. The theorem follows as the new protocol has error at most $(1-\delta)\varepsilon_2=\varepsilon_1$, and information cost at most $I+\delta\log|\mathcal{X}\times\mathcal{Y}|$. □

Note that $\mathrm{IC}_\mu(f,\mu,0)$ is not always continuous with respect to μ. For example, let the matrices

$$\mu_\varepsilon=\begin{pmatrix}\frac{1-\varepsilon}{3} & \frac{1-\varepsilon}{3}\\ \frac{1-\varepsilon}{3} & \varepsilon\end{pmatrix},\qquad \mu=\lim_{\varepsilon\to0}\mu_\varepsilon=\begin{pmatrix}\frac13 & \frac13\\ \frac13 & 0\end{pmatrix} \tag{2.15}$$

represent distributions on $\{0,1\}^2$. Here the entry at the i-th row and j-th column corresponds to the measure of the point $(i-1,j-1)\in\{0,1\}^2$. Now for the 2-bit AND function, we have $\mathrm{IC}_\mu(\mathrm{AND},\mu,0)=0$, while $\mathrm{IC}_{\mu_\varepsilon}(\mathrm{AND},\mu_\varepsilon,0)=\mathrm{IC}_{\mu_\varepsilon}(\mathrm{AND},0)$ as μ_ε has full support. Thus

$$\lim_{\varepsilon\to0}\mathrm{IC}_{\mu_\varepsilon}(\mathrm{AND},\mu_\varepsilon,0)=\lim_{\varepsilon\to0}\mathrm{IC}_{\mu_\varepsilon}(\mathrm{AND},0)=\mathrm{IC}_\mu(\mathrm{AND},0), \tag{2.16}$$

which is known to be bounded away from 0.

Finally, note that Lemma 2.4 also implies the continuity of $\mathrm{IC}_\mu(f,\nu,\varepsilon)$ with respect to ν when $\varepsilon>0$. Indeed if $|\nu_1-\nu_2|\le\delta\le\varepsilon$, then a protocol that has distributional error ε with respect to ν_2, will have error at most $\varepsilon+\delta$ and at least $\varepsilon-\delta$ with respect to ν_1. Thus

$$\mathrm{IC}_\mu(f,\nu_1,\varepsilon+\delta)\le\mathrm{IC}_\mu(f,\nu_2,\varepsilon)\le\mathrm{IC}_\mu(f,\nu_1,\varepsilon-\delta), \tag{2.17}$$

which establishes the desired continuity. A similar example to (2.15) shows that $\mathrm{IC}_\mu(f,\nu,0)$ is not necessarily continuous with respect to ν.

2.5 Communication protocols as random walks on $\Delta(\mathcal{X}\times\mathcal{Y})$

Recall that $\Delta(\mathcal{X}\times\mathcal{Y})$ denotes the set of probability distributions on $\mathcal{X}\times\mathcal{Y}$. Consider a protocol π and a prior distribution μ on the set of inputs $\mathcal{X}\times\mathcal{Y}$. Suppose that in the first round Alice sends a random signal B to Bob. We can interpret this as a random update of the prior distribution μ to a new distribution $\mu_0=\mu|_{B=0}$ or $\mu_1=\mu|_{B=1}$ depending on the value of B. It is not difficult to see that $\mu_b(x,y)=p_b(x)\mu(x,y)$ for $b=0,1$, where $p_b(x)=\Pr[B=b\mid x]/\Pr[B=b]$. In other words, μ_b is obtained by multiplying the rows of μ by non-negative numbers. From the law of total expectation,

$$\mu=\mathop{\mathbb{E}}_B[\mu\mid B]=\Pr[B=0]\mu_0+\Pr[B=1]\mu_1. \tag{2.18}$$

Similarly if Bob is sending a message, then μ_b is obtained by multiplying the columns of μ by the numbers $p_b(y) = \Pr[B=b \mid y]/\Pr[B=b]$. That is $\mu_b(x,y) = \mu(x,y)p_b(y)$.

The opposite direction is also true: given a distribution μ, distributions μ_0, μ_1, and $0 \le p_0, p_1 \le 1$ such that

- $p_0 + p_1 = 1$,
- μ_0 and μ_1 are obtained from μ by scaling its rows,
- $\mu = p_0\mu_0 + p_1\mu_1$,

one can define a random bit B that can be sent by Alice such that μ_b is μ conditioned on $B = b$ for $b \in \{0,1\}$, and $p_b = \Pr[B=b]$. A similar statement holds for the case where μ_0 and μ_1 are obtained from μ by scaling its columns and B is a signal that will be sent by Bob.

Therefore, we can think of a protocol as a random walk on $\Delta(\mathcal{X}\times\mathcal{Y})$ that starts at μ, and every time that a player sends a message, it moves to a new distribution. Equation (2.18) implies that this random walk is without drift.

Let Π denote the transcript of the protocol. Note that when the protocol terminates, the random walk stops at $\mu_\Pi := \mu|_\Pi$. Since Π itself is a random variable, μ_Π is a random variable that takes values in $\Delta(\mathcal{X}\times\mathcal{Y})$. Interestingly, both the internal and external information costs of the protocol depend only on the distribution of μ_Π (this is a distribution on the set $\Delta(\mathcal{X}\times\mathcal{Y})$, which itself is a set of distributions) [10]. It does not matter how different the steps of two protocols are, and as long as they both yield the same distribution on $\Delta(\mathcal{X}\times\mathcal{Y})$, they have the same internal and external information cost. Consequently, one can directly work with this random walk, instead of working with the actual protocols.

In order to study the relation between the information complexity and the distribution of μ_Π, define the *concealed information* and *external concealed information* of a protocol π with respect to μ, respectively, as

$$\mathrm{CI}_\mu(\pi) = H(X \mid \Pi Y) + H(Y \mid \Pi X) = H(X \mid Y) + H(Y \mid X) - \mathrm{IC}_\mu(\pi), \tag{2.19}$$

and

$$\mathrm{CI}^{ext}_\mu(\pi) = H(XY \mid \Pi) = H(XY) - \mathrm{IC}^{ext}_\mu(\pi). \tag{2.20}$$

With this definition it is easy to see that the information cost of a protocol π with transcript Π only depends on the distribution of μ_Π. Indeed

$$\mathrm{CI}_\mu(\pi) = H_{XY\sim\mu}(X \mid \Pi Y) + H_{XY\sim\mu}(Y \mid \Pi X) = \mathop{\mathbb{E}}_{\Pi} H_{XY\sim\mu_\Pi}(X \mid Y) + \mathop{\mathbb{E}}_{\Pi} H_{XY\sim\mu_\Pi}(Y \mid X). \tag{2.21}$$

Another nice property of concealed information is that if π_0 and π_1 are the two branches of the protocol π corresponding respectively to $B=0$ and $B=1$ where B is the first bit sent, then

$$\mathrm{CI}_\mu(\pi) = \Pr[B=0]\,\mathrm{CI}_{\mu|B=0}(\pi_0) + \Pr[B=1]\,\mathrm{CI}_{\mu|B=1}(\pi_1). \tag{2.22}$$

Thus, the expected value of CI is preserved throughout the execution of the protocol. Similar results hold for $\mathrm{CI}^{ext}_\mu(\pi)$.

3 Main results

In this section, we state and discuss our main results in full detail. Simpler proofs are presented in this section, but the proofs of the more involved results are postponed to later sections.

We will use the following simple estimate:

$$x \in [0,1/2] \Longrightarrow x \log \frac{1}{x} \le h(x) \le 2x \log \frac{1}{x}, \tag{3.1}$$

which holds since in that range $-x \log x \ge -(1-x)\log(1-x)$.

Denote

$$\overline{h}(x) = h(\min(x,1/2)). \tag{3.2}$$

It satisfies $\overline{h}(x) \ge h(x)$ and $x \le \overline{h}(x)$. It is easy to see that h is concave. Therefore, $\overline{h}$ is also concave as it is piecewise differentiable with non increasing derivative. Additionally, $h(0) = \overline{h}(0) = 0$. We will next show how to utilize these two properties of h and $\overline{h}$: for any concave function $g \colon \mathbb{R}^+ \to \mathbb{R}$ for which $g(0) = 0$, and for any $x > 0$ and $0 < q < 1$, it holds that

$$g(qx) \ge qg(x) + (1-q)g(0) = qg(x). \tag{3.3}$$

This implies the subadditivity of g: for all $a_1, a_2 > 0$, $g(a_1 + a_2) \le g(a_1) + g(a_2)$, as

$$g(a_i) \ge \frac{a_i}{a_1 + a_2} g(a_1 + a_2),$$

for all $i = 1, 2$.

3.1 Information complexity with point-wise error

Consider a communication problem $f \colon \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, and a distribution μ. How close can $\mathrm{IC}_\mu(f,\varepsilon)$ be to $\mathrm{IC}_\mu(f,0)$? A simple argument shows that $\mathrm{IC}_\mu(f,\varepsilon) \le \mathrm{IC}_\mu(f,0) - \Omega(\varepsilon)$.

Proposition 3.1. *Let $f \colon \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, and let μ be a measure on $\mathcal{X} \times \mathcal{Y}$. Denoting $c = \mathrm{IC}_\mu(f,0)$, we have*

$$\mathrm{IC}_\mu(f,\varepsilon) \le (1-\varepsilon)\,\mathrm{IC}_\mu(f,0) = \mathrm{IC}_\mu(f,0) - c\varepsilon. \tag{3.4}$$

Proof. Let π be a zero-error protocol for f. Consider a protocol π' in which Alice and Bob use their public randomness to run with probability $1-\varepsilon$ the protocol π, or to terminate with an arbitrary output with probability ε. Let Π and Π' be respectively the transcripts of π and π' on the random input (X,Y). We have

$$I(X;\Pi' \mid Y) = H(X \mid Y) - H(X \mid \Pi' Y) = H(X \mid Y) - \varepsilon H(X \mid Y) - (1-\varepsilon) H(X \mid \Pi Y) = (1-\varepsilon) I(X;\Pi \mid Y). \tag{3.5}$$

The same holds for $I(Y;\Pi' \mid X)$, and the statement follows. □

Our first main theorem shows that this trivial bound can be improved to

$$\mathrm{IC}_\mu(f,\varepsilon) \le \mathrm{IC}_\mu(f,0) - \Omega(h(\varepsilon)).$$

Theorem 3.2. *Consider a function $f\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$ and a probability measure μ on $\mathcal{X}\times\mathcal{Y}$ such that $\mathrm{IC}_\mu(f,0)>0$. There exist positive constants τ,ε_0, depending on f and μ (and thus on $|\mathcal{X}|,|\mathcal{Y}|,|\mathcal{Z}|$), such that for every $\varepsilon\le\varepsilon_0$,*

$$\mathrm{IC}_\mu(f,\varepsilon)\le \mathrm{IC}_\mu(f,0)-\tau h(\varepsilon). \tag{3.6}$$

Moreover:

Non-constant case: *Suppose that $f(a)\neq f(b)$ for two points a,b in the support of μ, and on the same row or column. Then one can take $\tau=\mu(a)^2\mu(b)/32$, and ε_0 depends only on $\min(\mu(a),\mu(b))$ and $|\mathcal{X}\times\mathcal{Y}|$.*

AND case: *Let $x_0,x_1\in\mathcal{X}$ and $y_0,y_1\in\mathcal{Y}$. Suppose that $f(x_0y_0)=f(x_0y_1)=f(x_1y_0)=z_0$ and $f(x_1y_1)=z_1\neq z_0$, and that $x_0y_0,x_0y_1,x_1y_0\in\operatorname{supp}\mu$. Then one can take*

$$\tau=\frac{\mu(x_0y_0)^2}{64}\min(\mu(x_0y_1),\mu(x_1y_0)),$$

and ε_0 depends only on $|\mathcal{X}\times\mathcal{Y}|$ and the minimum of $\mu(x_0y_0)$, $\mu(x_0y_1)$, $\mu(x_1y_0)$.

Proof. See Section 4.1.1. □

Remark 3.3. We prove Theorem 3.2 by taking a zero-error protocol for f, and turning it into an ε-error protocol that has an $\Omega(h(\varepsilon))$ gain in the information cost over the original protocol. The high-level idea is that one of the players checks her/his input and if it is equal to a certain value x_1, then with probability ε changes to a different value x_0. This obviously creates an error of at most ε. In the Non-constant case of Theorem 3.2, the points a and b are used to determine x_0 and x_1, and in the AND case, the same x_0 and x_1 as they are described in the statement of the theorem can be used. Note that this modification can only create errors that erroneously output $f(x_0,y)$ instead of $f(x_1,y)$ for some values of y. This allows us to obtain a one-sided error for many functions. We shall use this later in Corollary 3.9 to obtain an upper bound on the information complexity of the AND function when only one-sided error is allowed.

Despite the simplicity of the idea described in the above remark, the proof is rather involved, and uses some of our other results such as characterization of internal-trivial measures. The heart of the proof is of course showing the existence of appropriate values of x_0 and x_1 that can lead to the desired gain of $\Omega(h(\varepsilon))$.

Let XOR denote the 2-bit XOR function. The next result shows that the analogue of Theorem 3.2 does not hold for the external information complexity.

Proposition 3.4. *Let μ be the distribution defined as*

$$\mu=\begin{array}{|c|c|}\hline 1/2 & 0\\ \hline 0 & 1/2\\ \hline\end{array}\,. \tag{3.7}$$

Then $\mathrm{IC}^{ext}_\mu(\mathrm{XOR},\varepsilon)\ge \mathrm{IC}^{ext}_\mu(\mathrm{XOR},0)-3\varepsilon$.

Proof. See Section 4.1.3. □

For the lower bound we prove the following result.

Theorem 3.5. *For all f,μ,ε, we have*

$$\mathrm{IC}_\mu(f,\varepsilon) \geq \mathrm{IC}_\mu(f,0) - 4|\mathcal{X}||\mathcal{Y}|\bar{h}(\sqrt{\varepsilon}). \tag{3.8}$$

Proof. See Section 4.1.2. □

Theorem 3.5 is obtained by taking an ε-error protocol and completing it to a zero-error protocol. Here Alice and Bob first run the protocol that performs $[f,\varepsilon]$, but when this protocol terminates, instead of returning the output, they continue their interaction to verify that the value that they have obtained is correct. We will be able to show that these additional interactions can be performed at a small information cost, and thus the total information cost of the new protocol is not going to be much larger than that of the original protocol. This method, that we call *protocol completion*, is used in the proofs of other results such as Theorem 3.7 as well.

Finally let us remark that we do not know whether the bound in Theorem 3.5 is tight. In fact we are not aware of any examples of f and μ that refutes the possibility that there exists $C(f,\mu)>0$ such that $\mathrm{IC}_\mu(f,\varepsilon) \geq \mathrm{IC}_\mu(f,0) - C(f,\mu)h(\varepsilon)$.

3.2 Information complexity with distributional error

In Section 3.1 we considered the amount of gain one can obtain by allowing point-wise error. Next we turn to distributional error. How much can one gain in information cost by allowing a distributional error of ε? Small modifications in the proofs of Theorem 3.2 and Theorem 3.5 imply the following bounds.

Theorem 3.6. *Let μ be a probability measure on $\mathcal{X}\times\mathcal{Y}$, and let $f\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$ satisfy $\mathrm{IC}_\mu(f,\mu,0)>0$. We have*

$$\mathrm{IC}_\mu(f,\mu,0) - 4|\mathcal{X}||\mathcal{Y}|\bar{h}(\sqrt{\varepsilon/\alpha}) \leq \mathrm{IC}_\mu(f,\mu,\varepsilon) \leq \mathrm{IC}_\mu(f,\mu,0) - \frac{\alpha^2}{4}h(\varepsilon\alpha/4) + 3\varepsilon\log|\mathcal{X}\times\mathcal{Y}|, \tag{3.9}$$

where $\alpha = \min_{xy\in\operatorname{supp}\mu}\mu(x,y)$.

Proof. See Section 4.2. □

It is also possible to prove the upper bound of Theorem 3.6 using a different approach by "truncating" a zero-error protocol. Unfortunately this approach requires some assumptions on the support of μ. Nevertheless we sketch this proof, as the idea seems to be new, and it might have other applications.

Let $\Delta_0 \subseteq \Delta(\mathcal{X}\times\mathcal{Y})$ be the set of all measures ν such that $\mathrm{IC}_\nu(f,\nu,\varepsilon)=0$. Consider a protocol π that performs $[f,\mu,0]$. First we simulate π with another protocol π' such that no signal of π' jumps from outside of Δ_0 to the interior of Δ_0. In other words if some partial transcript t satisfies $\mu_t \notin \Delta_0$, then when the next signal B is sent, μ_{tB} is either still outside of Δ_0 or it is on the boundary $\partial\Delta_0$. The simulation can be done in a perfect manner so that if Π and Π' denote, respectively, the transcripts of π and π', then $\mu_{\Pi'}$ has the same distribution as μ_Π. The new protocol π' might not necessarily have bounded communication, but it will terminate with probability 1. We refer the reader to [15, Signal Simulation Lemma] and [5, Claim 7.14] for more details on such simulations.

We will truncate π' in the following manner to obtain a new protocol π_0 that performs $[f,\mu,\varepsilon]$. Whenever the corresponding random walk of π' reaches a distribution ν that is on the boundary $\partial\Delta_0$, the two players stop the random walk, and use $\mathrm{IC}_\nu(f,\nu,\varepsilon)=0$ to output a value that creates a distributional error of at most ε with respect to ν at no information cost. Obviously the distributional error of the protocol π_0 is at most ε. To analyze its information cost, denote the transcript of π_0 by P, and note that P is a partial transcript for π'. Let π'_P be the continuation of π' when one starts at this partial transcript. It is not difficult to see that

$$\mathrm{IC}_\mu(\pi)=\mathrm{IC}_\mu(\pi')=\mathrm{IC}_\mu(\pi_0)+\mathop{\mathbb{E}}_P[\mathrm{IC}_{\mu_P}(\pi'_P)].$$

Since π' performs $[f,\mu,0]$, the tail protocol π_P must perform $[f,\mu_P,0]$. Hence in order to finish the proof, it suffices to show that $\mathrm{IC}_\nu(f,\nu,0)=\Omega(h(\varepsilon))$ for every $\nu\in\partial\Delta_0$, as this would imply the desired $\mathrm{IC}_\mu(\pi)\ge\mathrm{IC}_\mu(\pi_0)+\Omega(h(\varepsilon))$. This can be proven with some work when μ is of full support, however it is not true for general measures. For example, consider the AND function, and let μ be the distribution on $\{0,1\}^2$ defined as $\mu(0,0)=1-2\varepsilon$ and $\mu(1,0)=\mu(1,1)=\varepsilon$. Note that although μ is on the boundary of Δ_0, we have $\mathrm{IC}_\mu(\mathrm{AND},\mu,0)\le 2\varepsilon$. Indeed, since $\mu(0,1)=0$, Bob with probability 1 knows the correct output by looking at his own input Y, and so if he sends his bit to Alice, they will both know the correct output. This will have information cost at most $H(Y\mid X)=\Pr[X=1]H(Y\mid X=1)=2\varepsilon$.

3.3 Information complexity of the AND function with error

Building upon the previous works of Ma and Ishwar [24, 25], Braverman et al. [5] developed a method for proving the optimality of information complexity and applied it to determine the internal and external information complexity of the two-bit AND function. They introduced a "continuous-time" protocol for this task, and proved that it has optimal internal and external information cost for any underlying distribution. Although this protocol is not a conventional communication protocol as it has access to a continuous clock, it can be approximated by conventional communication protocols through dividing the time into finitely many discrete units. Then in [5, Problem 1.1] they considered the case where error is allowed, and conjectured a gain of $\mathrm{IC}(\mathrm{AND})-\mathrm{IC}(\mathrm{AND},\varepsilon)=\Theta(h(\varepsilon))$. In this section, we conduct a thorough analysis of the information complexity of the AND function when error is permitted, and among other results, prove the aforementioned conjecture.

Applying our general bounds from Section 3.1 and Section 3.2 (i. e., Theorems 3.2, 3.5, and 3.6) we already obtain the following for positive functions $C_1(\mu)$ and $C_2(\mu)$.

(i). For every distribution μ satisfying $\mathrm{IC}_\mu(\mathrm{AND},0)>0$, we have

$$\mathrm{IC}_\mu(\mathrm{AND},0)-C_1(\mu)\overline{h}(\sqrt{\varepsilon})\le\mathrm{IC}_\mu(\mathrm{AND},\varepsilon)\le\mathrm{IC}_\mu(\mathrm{AND},0)-C_2(\mu)h(\varepsilon). \tag{3.10}$$

(ii). For every distribution μ satisfying $\mathrm{IC}_\mu(\mathrm{AND},\mu,0)>0$, we have

$$\mathrm{IC}_\mu(\mathrm{AND},\mu,0)-C_1(\mu)\overline{h}(\sqrt{\varepsilon})\le\mathrm{IC}_\mu(\mathrm{AND},\mu,\varepsilon)\le\mathrm{IC}_\mu(\mathrm{AND},\mu,0)-C_2(\mu)h(\varepsilon). \tag{3.11}$$

We show that under some conditions on the support of μ, the above lower bounds can be improved to match the upper bounds.

Theorem 3.7. *The following hold:*

(i). For every distribution μ *which is full support, except perhaps for* $\mu(1,1)$*, we have*

$$\mathrm{IC}_\mu(\mathrm{AND},0) - \mathrm{IC}_\mu(\mathrm{AND},\varepsilon) \le C(\mu)\bar{h}(\varepsilon), \tag{3.12}$$

where $C(\mu)$ *is bounded whenever* $\mu(0,0), \mu(0,1), \mu(1,0)$ *are bounded away from* 0*.*

(ii). For every distribution μ *of full support, we have*

$$\mathrm{IC}_\mu(\mathrm{AND},\mu,0) - \mathrm{IC}_\mu(\mathrm{AND},\mu,\varepsilon) \le C(\mu)\bar{h}(\varepsilon), \tag{3.13}$$

where $C(\mu)$ *is a positive function.*

Note that for every distribution μ of full support, we have

$$\mathrm{IC}_\mu(\mathrm{AND},\mu,0) = \mathrm{IC}_\mu(\mathrm{AND},0) > 0,$$

and

$$\mathrm{IC}_\mu(\mathrm{AND},\varepsilon/\alpha) \le \mathrm{IC}_\mu(\mathrm{AND},\mu,\varepsilon) \le \mathrm{IC}_\mu(\mathrm{AND},\varepsilon)$$

where $\alpha = \min_{xy} \mu(xy)$. Thus Theorem 3.7 (ii) follows from (i).

From a technical point of view, Theorem 3.7 is perhaps our most involved result in this article, and its proof occupies the bulk of Section 6. The first idea that facilitates the proof substantially shows that it is possible to parametrize the space of the distributions $\Delta(\mathcal{X} \times \mathcal{Y})$ so that the changes that occur in the prior distribution by the players' interactions can be captured by product measures. This idea, that is discussed in details in Section 5, allows us to first prove the lower bound of Theorem 3.7 for the product measures, and then add minor adjustments to adopt it for non-product distributions. The second component of the proof is a stability result. Recall from Section 2.5 that the information cost of every protocol π depends only on its "leaf distribution," i. e., the distribution of μ_Π, where Π is the transcript of π or equivalently μ_ℓ where ℓ is a random leaf of the protocol tree. Our stability result, Theorem 6.2, shows that the leaf distribution of every almost optimal protocol π for $[\mathrm{AND},0]$ shares certain similarities with that of the buzzer protocol. Note that since π does not make any errors, by the end of the protocol, either both players know that the input is $(1,1)$, or one of them has revealed that her input is 0. Theorem 6.2 formalizes the intuition that in this latter case, the other player must not have revealed that his input is very likely to be 0. This is achieved through defining a potential function that depends only on the distribution of μ_Π and proving that it is bounded by the so called information wastage $\mathrm{IC}_\mu(\pi) - \mathrm{IC}_\mu(\mathrm{AND},0)$. With these results in hand, in order to complete the lower bound of Theorem 3.7, we start with a protocol π performing $[\mathrm{AND},\varepsilon]$ with almost optimal information complexity. First we show that π can be completed to a protocol that performs $[\mathrm{AND},0]$ at a small additional information cost, though possibly larger than the desired $O(h(\varepsilon))$. Then we apply the stability result to deduce certain properties for the leaf distribution of π. This will imply that one indeed needs only an additional cost of $O(h(\varepsilon))$ to extend π to a protocol that solves $[\mathrm{AND},0]$.

Braverman et al. [5] showed that $\mathrm{IC}(\mathrm{AND},0) = \max_\mu \mathrm{IC}_\mu(\mathrm{AND},0)$ is attained on a distribution having full support. This enables us to derive the following corollary on prior-free information complexity.

Corollary 3.8. *The following hold:*

(i). $C_1\overline{h}(\varepsilon) \leq \mathrm{IC}(\mathrm{AND},0) - \mathrm{IC}(\mathrm{AND},\varepsilon) \leq C_2\overline{h}(\varepsilon)$*;*

(ii). $C_1\overline{h}(\varepsilon) \leq \mathrm{IC}(\mathrm{AND},0) - \mathrm{IC}^D(\mathrm{AND},\varepsilon) \leq C_2\overline{h}(\varepsilon)$*;*

where C_1 and C_2 are positive constants.

Proof. The measure μ that maximizes $\mathrm{IC}_\mu(\mathrm{AND},0)$ has full support [5], and thus

$$\mathrm{IC}(\mathrm{AND},0) = \mathrm{IC}_\mu(\mathrm{AND},0) = \mathrm{IC}_\mu(\mathrm{AND},\mu,0).$$

By Theorem 3.7 (ii),

$$\mathrm{IC}(\mathrm{AND},\varepsilon) \geq \mathrm{IC}^D(\mathrm{AND},\varepsilon) \geq \mathrm{IC}_\mu(\mathrm{AND},\mu,\varepsilon) \geq \mathrm{IC}_\mu(\mathrm{AND},\mu,0) - C_2\overline{h}(\varepsilon) = \mathrm{IC}(\mathrm{AND},0) - C_2\overline{h}(\varepsilon). \tag{3.14}$$

Moreover by a general upper bound that we prove later in Theorem 3.15, we have

$$\mathrm{IC}^D(\mathrm{AND},\varepsilon) \leq \mathrm{IC}(\mathrm{AND},\varepsilon) \leq \mathrm{IC}(\mathrm{AND},0) - C_1\overline{h}(\varepsilon). \tag{3.15}$$

Both items in the corollary follow. □

Since the difficult distributions for the set disjointness function are the ones in which the inputs typically have small or no intersections at all, the distributions for the AND function that assign a very small or 0 mass to the point $(1,1)$ are of particular importance. Let

$$\mathrm{IC}^\delta(\mathrm{AND},\varepsilon,1\to 0) = \sup_{\mu:\ \mu(1,1)\leq\delta} \mathrm{IC}_\mu(\mathrm{AND},\varepsilon,1\to 0). \tag{3.16}$$

The following corollary is used in Section 3.4 to analyze the information complexity of the set disjointness problem.

Corollary 3.9. *There exist positive constants C_1, C_2 and C_3 such that*

(i). $C_1\overline{h}(\varepsilon) \leq \mathrm{IC}^0(\mathrm{AND},0) - \mathrm{IC}^0(\mathrm{AND},\varepsilon) \leq C_2\overline{h}(\varepsilon)$*;*

(ii). $C_1\overline{h}(\varepsilon) \leq \mathrm{IC}^0(\mathrm{AND},0) - \mathrm{IC}^0(\mathrm{AND},\varepsilon,1\to 0) \leq C_2\overline{h}(\varepsilon)$*;*

(iii). for every $\varepsilon,\delta > 0$,

$$\mathrm{IC}^\delta(\mathrm{AND},\varepsilon,1\to 0) \leq \mathrm{IC}^0(\mathrm{AND},0) - C_1\overline{h}(\varepsilon) + C_3\overline{h}(\delta). \tag{3.17}$$

Proof. Assume that $\varepsilon \leq 1/2$. Let μ be the distribution maximizing $\mathrm{IC}_\mu(\mathrm{AND},0)$ under the constraint $\mu(1,1) = 0$; This measure, which is described in [5], has full support except for $\mu(1,1) = 0$. Thus by Theorem 3.7 (i),

$$\mathrm{IC}^0(\mathrm{AND},\varepsilon) \geq \mathrm{IC}_\mu(\mathrm{AND},\varepsilon) \geq \mathrm{IC}_\mu(\mathrm{AND},0) - C_2\overline{h}(\varepsilon) = \mathrm{IC}^0(\mathrm{AND},0) - C_2 h(\varepsilon). \tag{3.18}$$

Consequently, since $\mathrm{IC}^0(\mathrm{AND},\varepsilon) \le \mathrm{IC}^0(\mathrm{AND},\varepsilon,1\to 0)$, both (i) and (ii) will follow if we prove $\mathrm{IC}^0(\mathrm{AND},\varepsilon,1\to 0) \le \mathrm{IC}^0(\mathrm{AND},0) - C_1\overline{h}(\varepsilon)$. To prove this, we would like to apply the AND case of Theorem 3.2, however to be able to obtain a uniform upper bound on $\mathrm{IC}^0(\mathrm{AND},\varepsilon,1\to 0)$, we need to have a uniform lower bound on the probabilities $\mu(0,0),\mu(0,1),\mu(1,0)$. Let $\alpha > 0$ to be determined later, and consider any distribution μ with $\mu(1,1) = 0$ and $\mu(a) < \alpha$ for some input $a \ne (1,1)$. Pick $b \in \{0,1\}^2 \setminus \{a,(1,1)\}$, and obtain the distribution μ' from μ by transferring all the probability mass on a to b. That is $\mu'(b) = \mu(a)+\mu(b)$ and $\mu'(a) = 0$, and otherwise μ and μ' are identical. Obviously $|\mu-\mu'| \le \alpha$. Now (2.12) implies

$$\mathrm{IC}_\mu(\mathrm{AND},\varepsilon,1\to 0) \le \mathrm{IC}_\mu(\mathrm{AND},0) \le \mathrm{IC}_{\mu'}(\mathrm{AND},0) + 4\alpha + 2h(2\alpha) = 4\alpha + 2h(2\alpha) \le 4h(2\alpha), \tag{3.19}$$

where we used the fact that $\mathrm{IC}_{\mu'}(\mathrm{AND},0) = 0$ as $\operatorname{supp}\mu'$ contains only two points. Setting $\alpha = 0.001$ for example yields $\mathrm{IC}_\mu(\mathrm{AND},0) \le 4h(2\alpha) < 0.1 < \mathrm{IC}^0(\mathrm{AND},0) \approx 0.4827$. It remains to prove the statement for the distributions μ with $\mu(0,0),\mu(0,1),\mu(1,0) \ge \alpha$. In this case Theorem 3.2 (see Remark 3.3 regarding the one-sidedness) implies that exists a constant $C_1 > 0$ such that $\mathrm{IC}_\mu(\mathrm{AND},\varepsilon,1\to 0) \le \mathrm{IC}^0(\mathrm{AND},0) - C_1\overline{h}(\varepsilon)$. This finishes the proof of (i) and (ii).

To prove (iii), consider an arbitrary distribution μ with $\mu(1,1) \le \delta$, and let μ' be the distribution that is obtained from μ by moving the probability mass on $(1,1)$ to a different point so that $\mu'(1,1) = 0$ and $|\mu-\mu'| \le \delta$. Similar to (3.19), we obtain

$$\mathrm{IC}_\mu(\mathrm{AND},\varepsilon,1\to 0) \le \mathrm{IC}_{\mu'}(\mathrm{AND},\varepsilon,1\to 0) + 4h(2\delta) \le \mathrm{IC}^0(\mathrm{AND},\varepsilon,1\to 0) + 4h(2\delta),$$

and thus (iii) follows from (ii). □

3.4 Set disjointness function with error

In this section we focus on the set disjointness function. Firstly it is not hard to obtain the following result.

Corollary 3.10. *There exists a constant $C > 0$ such that the following holds:*

$$\mathrm{IC}(\mathrm{DISJ}_n,\varepsilon) \ge n\left[\mathrm{IC}^0(\mathrm{AND},0) - C\overline{h}(\varepsilon)\right]. \tag{3.20}$$

Proof sketch. By the argument that proves the additivity of information complexity (see, e. g., [7]), one can prove that $\mathrm{IC}(\mathrm{DISJ}_n,\varepsilon) \ge n\,\mathrm{IC}^0(\mathrm{AND},\varepsilon)$. Then apply Corollary 3.9. The essential idea is the following. Consider a distribution μ on $\{0,1\}^2$ with $\mu(1,1) = 0$, and let $(a,b) \in \{0,1\}^2$ be an input for the AND function. Let $XY \in \{0,1\}^n \times \{0,1\}^n$ be such that for some randomly selected $J \in \{1,\dots,n\}$ we have $(X_j,Y_j) = (a,b)$, and for $i \in \{1,\dots,n\}\setminus\{J\}$, the pairs (X_i,Y_i) are i.i.d. random variables, each with distribution μ. Since $\mu(1,1) = 0$, we have $\mathrm{DISJ}_n(X,Y) = 1 - \mathrm{AND}(a,b)$ with probability 1. Thus one can take a protocol π for DISJ_n and use it to solve $\mathrm{AND}(a,b)$ correctly for every (a,b). By sampling XY in a clever way, using both public and private randomness, one can guarantee that the information cost of the new protocol that solves $\mathrm{AND}(a,b)$ will be the information cost of π divided by n. □

As a result one also obtains that $R_\varepsilon(\mathrm{DISJ}_n) \ge n[\mathrm{IC}^0(\mathrm{AND},0) - \Omega(h(\varepsilon))]$. It turns out that by using techniques from [5] and [4], one can prove the following upper bound.

Theorem 3.11. *For the set disjointness function* DISJ_n *on inputs of length n, we have*

$$\limsup_{n\to\infty} \frac{R_\varepsilon(\mathrm{DISJ}_n)}{n} \leq \mathrm{IC}^0(\mathrm{AND},0) - Ch(\varepsilon), \tag{3.21}$$

for some constant $C > 0$.

Proof. See Section 7.1. □

We conjecture that in fact the exact constant is given by $\mathrm{IC}^0(\mathrm{AND},\varepsilon,1\to 0)$. In other words:

Conjecture 3.12. *For the set disjointness function* DISJ_n *on inputs of length n, we have*

$$R_\varepsilon(\mathrm{DISJ}_n) = n\,\mathrm{IC}^0(\mathrm{AND},\varepsilon,1\to 0) + o(n). \tag{3.22}$$

Braverman [4] proved that for all $0 < \alpha < 1$ and for all functions f,

$$\mathrm{IC}^D(f,\varepsilon) \geq (1-\alpha)\,\mathrm{IC}\left(f,\frac{\varepsilon}{\alpha}\right). \tag{3.23}$$

When $f = \mathrm{DISJ}_n$, Corollary 3.10 gives

$$\frac{\mathrm{IC}^D(\mathrm{DISJ}_n,\varepsilon)}{n} \geq (1-\alpha)(\mathrm{IC}^0(\mathrm{AND},0) - \Theta(h(\varepsilon/\alpha))) \geq \mathrm{IC}^0(\mathrm{AND},0) - \Theta(\alpha + h(\varepsilon/\alpha)). \tag{3.24}$$

Substituting $\alpha = \sqrt{\varepsilon\log(1/\varepsilon)}$ yields

$$\frac{\mathrm{IC}^D(\mathrm{DISJ}_n,\varepsilon)}{n} \geq \mathrm{IC}^0(\mathrm{AND},0) - \Theta(\sqrt{h(\varepsilon)}). \tag{3.25}$$

In Theorem 3.13 below, which is one of our main contributions, we show that this bound is sharp. The proof relies on introducing a new protocol for set disjointness problem, and analyzing its information cost.

Theorem 3.13. *For the set disjointness function* DISJ_n *on inputs of length n, we have*

$$\mathrm{IC}^D(\mathrm{DISJ}_n,\varepsilon) \leq n\left(\mathrm{IC}^0(\mathrm{AND},0) - C_1\sqrt{h(\varepsilon)}\right) + C_2\log n, \tag{3.26}$$

for constants $C_1, C_2 > 0$.

Proof. See Section 7.2. □

3.5 Prior-free information cost

Theorem 3.13 shows that for $\alpha = \sqrt{\varepsilon\log(1/\varepsilon)} = \Theta(\sqrt{h(\varepsilon)})$, and sufficiently large n, we have

$$\frac{\mathrm{IC}^D(\mathrm{DISJ}_n,\varepsilon)}{1-\Theta(\alpha)} = \mathrm{IC}(\mathrm{DISJ}_n,\varepsilon/\alpha) < \mathrm{IC}(\mathrm{DISJ}_n,\varepsilon), \tag{3.27}$$

and thus proves a separation between distributional and non-distributional prior-free information complexity. As we discussed in the introduction this has the important implication that amortized randomized communication complexity is not necessarily equal to the amortized distributional communication complexity with respect to the hardest distribution. More precisely, there are examples for which $\max_\mu D^{\mu,n}_\varepsilon(f^n) \neq R^n_\varepsilon(f^n)$.

Next we turn to proving general lower bounds and upper bounds for the prior-free information complexity. Theorem 3.5 immediately implies a lower bound for non-distributional prior-free information complexity.

Corollary 3.14 (corollary of Theorem 3.5)**.** *For every function f, we have*

$$\mathrm{IC}(f,\varepsilon) \geq \mathrm{IC}(f,0) - 4|\mathcal{X}\times\mathcal{Y}|\bar{h}(\sqrt{\varepsilon}). \tag{3.28}$$

Since unless μ satisfies certain conditions, Theorem 3.2 does not provide an upper bound on $\mathrm{IC}_\mu(f,\varepsilon)$ that is uniform on μ, we cannot apply it directly to bound $\mathrm{IC}(f,\varepsilon)$. However, we will get around this problem by proving that the "difficult distributions" satisfy these conditions and hence we obtain the desired upper bound.

Theorem 3.15. *If $f\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$ is non-constant, then*

$$\mathrm{IC}(f,\varepsilon) \leq \mathrm{IC}(f,0) - C(f)\bar{h}(\varepsilon), \tag{3.29}$$

where $C(f) > 0$.

Proof. See Section 4.3. □

The same upper bound and lower bound hold for $\mathrm{IC}^D(f,\varepsilon)$.

Theorem 3.16. *If $f\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$ is non-constant, then*

$$\mathrm{IC}^D(f,0) - C_1(f)\bar{h}(\sqrt{\varepsilon}) \leq \mathrm{IC}^D(f,\varepsilon) \leq \mathrm{IC}^D(f,0) - C_2(f)h(\varepsilon), \tag{3.30}$$

where $C_1(f), C_2(f) > 0$.

Proof. It is shown in [4] that $\mathrm{IC}^D(f,0) = \mathrm{IC}(f,0)$, and thus the upper bound follows from Theorem 3.15 as $\mathrm{IC}^D(f,\varepsilon) \leq \mathrm{IC}(f,\varepsilon)$.

To prove the lower bound, choose a measure μ that maximizes $\mathrm{IC}_\mu(f,\mu,0)$, and let

$$\alpha = \min_{xy\in\operatorname{supp}\mu} \mu(x,y).$$

Applying Theorem 3.6, we get

$$\mathrm{IC}^D(f,\varepsilon) \geq \mathrm{IC}_\mu(f,\mu,\varepsilon) \geq \mathrm{IC}_\mu(f,\mu,0) - 4|\mathcal{X}||\mathcal{Y}|\bar{h}(\sqrt{\varepsilon/\alpha}) = \mathrm{IC}^D(f,0) - C_1(f)\bar{h}(\sqrt{\varepsilon}). \tag{3.31}$$

□

3.6 A characterization of trivial measures

We start with a few definitions. Let $f: \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be an arbitrary function, and μ a distribution on $\mathcal{X} \times \mathcal{Y}$. We say that μ is *external-trivial* if $\mathrm{IC}^{ext}_{\mu}(f,\mu,0) = 0$. We say that μ is *strongly external-trivial* if there exists a protocol π computing f correctly on all inputs satisfying $\mathrm{IC}^{ext}_{\mu}(\pi) = 0$. We say that μ is *structurally external-trivial* if f is constant on $S_A \times S_B$, where S_A is the support of the marginal of μ on Alice's input and S_B is the support of the marginal of μ on Bob's input.

Similarly we say that μ is *internal-trivial* if $\mathrm{IC}_{\mu}(f,\mu,0) = 0$. We say that μ is *strongly internal-trivial* if there exists a protocol π computing f correctly on all inputs satisfying $\mathrm{IC}_{\mu}(\pi) = 0$. We say that μ is *structurally internal-trivial* if the marginals of μ can be partitioned as $S_A = \bigcup_i \mathcal{X}_i$ and $S_B = \bigcup_i \mathcal{Y}_i$ so that the support of μ is contained in $\bigcup_i \mathcal{X}_i \times \mathcal{Y}_i$, and f is constant on each $\mathcal{X}_i \times \mathcal{Y}_i$.

Theorem 3.17 below shows that all our definitions of internal triviality are equivalent. In particular, if $\mathrm{IC}_{\mu}(f,0) = 0$, then the infimum in the definition of IC_{μ} is achieved by a finite protocol.

Theorem 3.17. *Let $f: \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be an arbitrary function, and μ a distribution on $\mathcal{X} \times \mathcal{Y}$. The distribution μ is internal-trivial iff it is strongly internal-trivial iff it is structurally internal-trivial.*

Proof. See Section 4.4. □

In order to prove Theorem 3.17, we first obtain a characterization of measures that are not structurally internal-trivial, by defining a graph G_μ on the support of every distribution μ on $\mathcal{X} \times \mathcal{Y}$.

Definition 3.18. Let G be the graph whose vertex set is $\mathcal{X} \times \mathcal{Y}$, and two vertices are connected if they agree on one of their coordinates. That is, xy, xy' are connected for every $x \in \mathcal{X}$ and $y \neq y' \in \mathcal{Y}$, and $xy, x'y$ are connected for every $x \neq x' \in \mathcal{X}$ and $y \in \mathcal{Y}$. In short, G is the Cartesian product of the complete graphs $K_{\mathcal{X}}$ and $K_{\mathcal{Y}}$. Let G_μ be the subgraph of G induced by the support of μ. For every connected component C of G_μ, define

$$\begin{aligned} C_A &= \{x \in \mathcal{X} : xy \in C \text{ for some } y \in \mathcal{Y}\}, \\ C_B &= \{y \in \mathcal{Y} : xy \in C \text{ for some } x \in \mathcal{X}\}. \end{aligned}$$

The following lemma shows that if μ is not structurally internal-trivial, then there exists a connected component C of G_μ such that f is not constant on $C_A \times C_B$. We will use this fact later in Section 4.1.1 in the proof of Theorem 3.2.

Lemma 3.19. *Let $f: \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be an arbitrary function, and μ a distribution on $\mathcal{X} \times \mathcal{Y}$. Then the distribution μ is structurally internal-trivial iff for every connected component C of G_μ, the function f is constant on $C_A \times C_B$.*

Proof. Suppose first that μ is structurally internal-trivial. Thus there exist partitions $S_A = \bigcup_i \mathcal{X}_i$ and $S_B = \bigcup_i \mathcal{Y}_i$ such that the support of μ is contained in $\bigcup_i \mathcal{X}_i \times \mathcal{Y}_i$ and f is constant on $\mathcal{X}_i \times \mathcal{Y}_i$ on each i. Any connected component C of G_μ must lie in some $\mathcal{X}_i \times \mathcal{Y}_i$, hence $C_A \times C_B \subseteq \mathcal{X}_i \times \mathcal{Y}_i$ and f is constant on $C_A \times C_B$ for every connected component C.

Conversely, suppose that for every connected component C of G_μ, the function f is constant on $C_A \times C_B$. If C, C' are two different connected components then C_A, C'_A are disjoint: otherwise, if (say) $xy \in$

C and $xy' \in C'$ then xy is connected to xy' and so $C = C'$. Thus $\{C_A : C \text{ a connected component of } G_\mu\}$ partitions the support of $\mathcal{X}$. Similarly, $\{C_B : C \text{ a connected component of } G_\mu\}$ partitions the support of $\mathcal{Y}$. These partitions serve as a witness that μ is structurally internal-trivial. $\square$

Finally we note that the analogue of Theorem 3.17 holds for the external case as well.

Theorem 3.20. *Let $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be an arbitrary function, and μ a distribution on $\mathcal{X} \times \mathcal{Y}$. The distribution μ is external-trivial iff it is strongly external-trivial iff it is structurally external-trivial.*

Proof. See Section 4.4. $\square$

4 Proofs for general functions

In this section we present the proofs of the main results on general functions presented in Section 3.

4.1 Information complexity with point-wise error

4.1.1 Proof of Theorem 3.2

We discuss some notation before the proof. Consider a protocol π. For an input xy, let Π_{xy} denote the random variable corresponding to the transcript of π when it is executed on the input xy. Let Π denote the random variable for transcripts of π, whose distribution is given as

$$\Pr[\Pi = t] = \mathop{\mathbb{E}}_{xy} \Pr[\Pi_{xy} = t] = \sum_{xy} \Pr[xy] \Pr[\Pi_{xy} = t], \tag{4.1}$$

where $\Pr[\Pi_{xy} = t] = \Pr[\Pi = t \mid XY = xy]$. As usual we abbreviate $\Pr[xy] = \Pr[XY = xy]$, and $\Pr[x \mid y] = \Pr[X = x \mid Y = y]$, and so on.

The next lemma shows that under some conditions, if we modify a protocol π to a new protocol π' according to Figure 1, then the information cost will have a significant drop.

On input XY:

- Alice privately samples a Bernoulli random variable B with parameter ε.
- If $X = x_1$ and $B = 1$, Alice sets $X' = x_0$, otherwise she sets $X' = X$.
- The players run π on $X'Y$.

Figure 1: The protocol π' is obtained from a protocol π using $x_0, x_1 \in \mathcal{X}$.

Lemma 4.1. *Let μ be a distribution on $\mathcal{X} \times \mathcal{Y}$, and π be a protocol with input set $\mathcal{X} \times \mathcal{Y}$. Suppose there is a set $\mathcal{L}$ of transcripts of π that satisfies, for some $C_1 \in [0,1]$,* *(1)* $\Pr[\Pi \in \mathcal{L}] \geq C_1$*;*

and there are $x_0\bar{y}, x_1\bar{y}$*, both in the support of* μ*, and* $C_2 \in (0,1], \delta \in [0,1]$ *with* $C_2 > 2\delta$*, such that for every* $t \in \mathcal{L}$*,*

(2) $\Pr[XY = x_0\bar{y} \mid \Pi = t] \geq C_2$*;*

(3) $\Pr[XY = x_1\bar{y} \mid \Pi = t] \leq \delta$*.*

Let $K = \log|\mathcal{X} \times \mathcal{Y}|$*. Then for sufficiently small* $\varepsilon > 0$ *(depending on* μ, C_2, δ*), the protocol* π' *defined in Figure 1 satisfies*

$$\mathrm{IC}_\mu(\pi') \leq \mathrm{IC}_\mu(\pi) - C_1 C_2 h\left(\frac{\varepsilon}{2}\min\left\{1, C_2\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]}\right\}\right) + 3\varepsilon K + \bar{h}(\delta/C_2). \tag{4.2}$$

Explicitly, the upper bound holds as long as

$$\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]}\varepsilon + (1-\varepsilon)\delta/C_2 \leq 1/2.$$

Intuitively, this condition says that π has a set of transcripts $\mathcal{L}$ that happen with significant probability, and every transcript in $\mathcal{L}$ probabilistically differentiates between $x_0\bar{y}$ and $x_1\bar{y}$. In other words, if we we see a transcript in $\mathcal{L}$, then we know that the input was much more likely to be $x_0\bar{y}$ than to be $x_1\bar{y}$. One point to note here is that we require the two points $x_0\bar{y}$ and $x_1\bar{y}$ to be in the same column. By symmetry, if there are two points in the same row satisfying the same properties, then the claim of Lemma 4.1 also holds.

Proof. Consider the protocols π and π' as described in Figure 1. Note that $\Pi_{X'Y}$ is the transcript of π'. We shorthand $\Pi' = \Pi_{X'Y}$. The information cost of π' is given by

$$\mathrm{IC}_\mu(\pi') = I(X;\Pi' \mid Y) + I(Y;\Pi' \mid X) = H(X \mid Y) + H(Y \mid X) - H(X \mid \Pi'Y) - H(Y \mid \Pi'X), \tag{4.3}$$

while

$$\mathrm{IC}_\mu(\pi) = I(X;\Pi \mid Y) + I(Y;\Pi \mid X) = H(X \mid Y) + H(Y \mid X) - H(X \mid \Pi Y) - H(Y \mid \Pi X). \tag{4.4}$$

Hence

$$\mathrm{IC}_\mu(\pi) - \mathrm{IC}_\mu(\pi') = H(X \mid \Pi'Y) - H(X \mid \Pi Y) + H(Y \mid \Pi'X) - H(Y \mid \Pi X). \tag{4.5}$$

Note that

$$H(Y \mid \Pi'X) \geq H(Y \mid \Pi'XB) \geq (1-\varepsilon)H(Y \mid \Pi'X,(B=0)) = (1-\varepsilon)H(Y \mid \Pi X) \geq H(Y \mid \Pi X) - \varepsilon K. \tag{4.6}$$

Similarly, for every $y \in \mathcal{Y}$ and every possible transcript t, we have

$$H(X \mid \Pi'Y = ty) \geq H(X \mid \Pi Y = ty) - \varepsilon K. \tag{4.7}$$

We will show that for $Y = \bar{y}$ and every transcript $t \in \mathcal{L}$,

$$H(X \mid \Pi'Y = t\bar{y}) \geq H(X \mid \Pi Y = t\bar{y}) + h\left(\frac{\varepsilon}{2}\min\left\{1, C_2\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]}\right\}\right) - \bar{h}(\delta/C_2) - \varepsilon K. \tag{4.8}$$

Note that Condition (2) implies that for $t \in \mathcal{L}$,

$$\Pr[\Pi Y = t\bar{y}] \geq \Pr[\Pi XY = tx_0\bar{y}] = \Pr[XY = x_0\bar{y} \mid \Pi = t] \Pr[\Pi = t] \geq C_2 \Pr[\Pi = t]. \tag{4.9}$$

Hence

$$\Pr[\Pi \in \mathcal{L}, Y = \bar{y}] \geq C_2 \Pr[\Pi \in \mathcal{L}] \geq C_1 C_2. \tag{4.10}$$

This together with (4.7) and (4.8) would show that

$$\begin{aligned}
H(X \mid \Pi' Y) &= \sum_t \sum_{y \in \mathcal{Y}} \Pr[\Pi' Y = ty] H(X \mid \Pi' Y = ty) \geq \sum_t \sum_{y \in \mathcal{Y}} (1-\varepsilon) \Pr[\Pi Y = ty] H(X \mid \Pi' Y = ty) \\
&\geq \sum_t \sum_{y \in \mathcal{Y}} \Pr[\Pi Y = ty] H(X \mid \Pi Y = ty) \\
&\quad + \Pr[\Pi \in \mathcal{L}, Y = \bar{y}] \left(h\left(\frac{\varepsilon}{2} \min\left\{ 1, C_2 \frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]} \right\} \right) - \bar{h}(\delta/C_2) \right) - 2\varepsilon K \\
&\geq H(X \mid \Pi Y) + C_1 C_2 h\left(\frac{\varepsilon}{2} \min\left\{ 1, C_2 \frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]} \right\} \right) - 2\varepsilon K - \bar{h}(\delta/C_2).
\end{aligned}$$

Applying (4.6) would immediately give the claimed bound.

Our aim, then, is to show (4.8). From now on we consider exclusively $t \in \mathcal{L}$.

The idea is to consider the indicator variable $C := 1_{[X \neq x_1]}$. Since C is a deterministic function of X, we have

$$H(X \mid \Pi' Y = t\bar{y}) = H(XC \mid \Pi' Y = t\bar{y}) = H(X \mid C, (\Pi' Y = t\bar{y})) + H(C \mid \Pi' Y = t\bar{y}). \tag{4.11}$$

We obtain by Conditions (2) and (3) that

$$\Pr[X = x_1 \mid \Pi Y = t\bar{y}] = \frac{\Pr[XY = x_1\bar{y} \mid \Pi = t]}{\Pr[Y = \bar{y} \mid \Pi = t]} \leq \frac{\Pr[XY = x_1\bar{y} \mid \Pi = t]}{\Pr[XY = x_0\bar{y} \mid \Pi = t]} \leq \frac{\delta}{C_2}. \tag{4.12}$$

Hence using (4.12), the first term in (4.11) can be bounded as

$$\begin{aligned}
H(X \mid C, (\Pi' Y = t\bar{y})) &\geq (1-\varepsilon) H(X \mid C, (B\Pi' Y = 0t\bar{y})) && (4.13)\\
&\geq H(X \mid C, (\Pi Y = t\bar{y})) - \varepsilon K && (4.14)\\
&= H(XC \mid \Pi Y = t\bar{y}) - H(C \mid \Pi Y = t\bar{y}) - \varepsilon K && (4.15)\\
&\geq H(X \mid \Pi Y = t\bar{y}) - \bar{h}(\delta/C_2) - \varepsilon K. && (4.16)
\end{aligned}$$

Next, we bound the second term $H(C \mid \Pi' Y = t\bar{y})$ in (4.11):

$$\begin{aligned}
\Pr[C = 0 \mid \Pi' Y = t\bar{y}] &= \Pr[X = x_1 \mid \Pi' Y = t\bar{y}] && (4.17)\\
&= \frac{\Pr[\Pi' XY = tx_1\bar{y}]}{\Pr[\Pi' Y = t\bar{y}]} && (4.18)\\
&= \frac{\Pr[\Pi' = t \mid XY = x_1\bar{y}] \Pr[x_1\bar{y}]}{\Pr[\Pi' Y = t\bar{y}]} && (4.19)\\
&= (\varepsilon \Pr[\Pi = t \mid XY = x_0\bar{y}] + (1-\varepsilon) \Pr[\Pi = t \mid XY = x_1\bar{y}]) \frac{\Pr[x_1\bar{y}]}{\Pr[\Pi' Y = t\bar{y}]} && (4.20)\\
&= \frac{\varepsilon \Pr[\Pi XY = tx_0\bar{y}] \frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]} + (1-\varepsilon) \Pr[\Pi XY = tx_1\bar{y}]}{\Pr[\Pi' Y = t\bar{y}]}. && (4.21)
\end{aligned}$$

We start by upper bounding (4.21):

$$(4.21) \leq \frac{\varepsilon \Pr[\Pi XY = tx_0\bar{y}]\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]} + (1-\varepsilon)\Pr[\Pi XY = tx_1\bar{y}]}{\Pr[\Pi' XY = tx_0\bar{y}]} \tag{4.22}$$

$$= \frac{\varepsilon \Pr[\Pi XY = tx_0\bar{y}]\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]} + (1-\varepsilon)\Pr[\Pi XY = tx_1\bar{y}]}{\Pr[\Pi XY = tx_0\bar{y}]} \tag{4.23}$$

$$= \frac{\varepsilon \Pr[XY = x_0\bar{y} \mid \Pi = t]\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]} + (1-\varepsilon)\Pr[XY = x_1\bar{y} \mid \Pi = t]}{Pr[XY = x_0\bar{y} \mid \Pi = t]} \tag{4.24}$$

$$\leq \frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]}\varepsilon + (1-\varepsilon)\delta/C_2. \tag{4.25}$$

Next, we lower bound (4.21). First, obtain an upper bound on its denominator:

$$\Pr[\Pi' Y = t\bar{y}] = \sum_{x \neq x_1} \Pr[\Pi' XY = tx\bar{y}] + \Pr[\Pi' XY = tx_1\bar{y}] \tag{4.26}$$

$$= \sum_{x \neq x_1} \Pr[\Pi XY = tx\bar{y}] + \varepsilon \Pr[\Pi XY = tx_0\bar{y}]\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]} + (1-\varepsilon)\Pr[\Pi XY = tx_1\bar{y}] \tag{4.27}$$

$$\leq \sum_{x} \Pr[\Pi XY = tx\bar{y}] + \varepsilon \Pr[\Pi XY = tx_0\bar{y}]\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]} \tag{4.28}$$

$$= \Pr[\Pi Y = t\bar{y}] + \varepsilon \Pr[\Pi XY = tx_0\bar{y}]\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]} \tag{4.29}$$

$$\leq 2\max\left\{\Pr[\Pi Y = t\bar{y}], \Pr[\Pi XY = tx_0\bar{y}]\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]}\right\}, \tag{4.30}$$

where (4.27) follows from the calculations between (4.18) and (4.21). Hence,

$$(4.21) \geq \frac{\varepsilon \Pr[\Pi XY = tx_0\bar{y}]\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]}}{2\max\{\Pr[\Pi Y = t\bar{y}], \Pr[\Pi XY = tx_0\bar{y}]\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]}\}} \tag{4.31}$$

$$= \frac{\varepsilon}{2}\min\left\{\frac{\Pr[\Pi XY = tx_0\bar{y}]}{\Pr[\Pi Y = t\bar{y}]}\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]}, 1\right\}. \tag{4.32}$$

$$= \frac{\varepsilon}{2}\min\left\{\Pr[X = x_0 \mid \Pi Y = t\bar{y}]\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]}, 1\right\}. \tag{4.33}$$

$$= \frac{\varepsilon}{2}\min\left\{\frac{\Pr[XY = x_0\bar{y} \mid \Pi = t]}{\Pr[Y = \bar{y} \mid \Pi = t]}\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]}, 1\right\}. \tag{4.34}$$

$$\geq \frac{\varepsilon}{2}\min\left\{\Pr[XY = x_0\bar{y} \mid \Pi = t]\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]}, 1\right\}. \tag{4.35}$$

$$\geq \frac{\varepsilon}{2}\min\left\{C_2\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]}, 1\right\}. \tag{4.36}$$

Combining (4.21), (4.25) and (4.36):

$$\frac{\varepsilon}{2}\min\left\{1, C_2\frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]}\right\} \le \Pr[C = 0 \mid \Pi' Y = t\bar{y}] \le \frac{\Pr[x_1\bar{y}]}{\Pr[x_0\bar{y}]}\varepsilon + \frac{(1-\varepsilon)\delta}{C_2}. \tag{4.37}$$

This together with (4.11) and (4.16) gives (4.8) as desired, as long as $\varepsilon > 0$ is small enough such that the upper bound in (4.37) is at most $1/2$. □

Theorem 3.2 (restated). *Consider a function $f\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$ and a probability measure μ on $\mathcal{X}\times\mathcal{Y}$ such that $\mathrm{IC}_\mu(f,0) > 0$. There exist positive constants τ, ε_0, depending on f and μ, such that for every $\varepsilon \le \varepsilon_0$,*

$$\mathrm{IC}_\mu(f,\varepsilon) \le \mathrm{IC}_\mu(f,0) - \tau h(\varepsilon). \tag{4.38}$$

Moreover:

Non-constant case: *Suppose that $f(a) \ne f(b)$ for two points a, b in the support of μ, and on the same row or column. Then one can take $\tau = \mu(a)^2\mu(b)/32$, and ε_0 depends only on $\min(\mu(a),\mu(b))$ and $|\mathcal{X}\times\mathcal{Y}|$.*

AND case: *Let $x_0, x_1 \in \mathcal{X}$ and $y_0, y_1 \in \mathcal{Y}$. Suppose that $f(x_0y_0) = f(x_0y_1) = f(x_1y_0) = z_0$ and $f(x_1y_1) = z_1 \ne z_0$, and that $x_0y_0, x_0y_1, x_1y_0 \in \operatorname{supp}\mu$. Then one can take*

$$\tau = \frac{\mu(x_0y_0)^2}{64}\min(\mu(x_0y_1),\mu(x_1y_0)),$$

and ε_0 depends only on $|\mathcal{X}\times\mathcal{Y}|$ and the minimum of $\mu(x_0y_0)$, $\mu(x_0y_1)$, $\mu(x_1y_0)$.

Proof. In order to apply the assumption $\mathrm{IC}_\mu(f,0) > 0$, we will need to use our characterization of internal-trivial measures. Consider the graph G_μ defined on $\mathcal{X}\times\mathcal{Y}$ as given in Definition 3.18. By Theorem 3.17 and Lemma 3.19, the assumption $\mathrm{IC}_\mu(f,0) > 0$ implies the existence of a connected component C of G_μ such that f is not constant on $C_A \times C_B$. Note that $C \subseteq \operatorname{supp}\mu$, and $C_A \times C_B$ is the corresponding rectangle given by C.

Case I: f is not constant on C.

As C is connected, there must be two adjacent points $a, b \in C$ such that $f(a) \ne f(b)$. By our definition of adjacency in Definition 3.18, without loss of generality we can assume that a, b are in the same column. Now consider any protocol π that solves $[f,0]$. Let $\mathcal{L}_0$ be the set of the transcripts that can occur when π runs with input a; formally,

$$\mathcal{L}_0 = \{t : \Pr[\Pi_a = t] > 0\}. \tag{4.39}$$

Clearly $\Pr[\Pi \in \mathcal{L}_0] \ge \mu(a)$. As $f(a) \ne f(b)$ and π has no error, for every $t \in \mathcal{L}_0$,

$$\Pr[XY = b \mid \Pi = t] = 0. \tag{4.40}$$

Let

$$\mathcal{L} = \{t \in \mathcal{L}_0 : \Pr[XY = a \mid \Pi = t] \ge \mu(a)/2\}. \tag{4.41}$$

We claim

$$\Pr[\Pi \in \mathcal{L}] \geq \mu(a)/2. \tag{4.42}$$

Indeed, note

$$\sum_{t \in \mathcal{L}_0} \Pr[\Pi = t] \Pr[XY = a \mid \Pi = t] = \sum_t \Pr[\Pi = t] \Pr[XY = a \mid \Pi = t] = \mu(a). \tag{4.43}$$

Using the trivial bound $\Pr[XY = a \mid \Pi = t] \leq 1$, we have

$$\begin{aligned}\mu(a) &= \sum_{t \in \mathcal{L}} \Pr[\Pi = t] \Pr[XY = a \mid \Pi = t] + \sum_{t \in \mathcal{L}_0 \setminus \mathcal{L}} \Pr[\Pi = t] \Pr[XY = a \mid \Pi = t] \\ &\leq \sum_{t \in \mathcal{L}} \Pr[\Pi = t] + \frac{\mu(a)}{2} \sum_{t \in \mathcal{L}_0 \setminus \mathcal{L}} \Pr[\Pi = t] \leq \Pr[\Pi \in \mathcal{L}] + \frac{\mu(a)}{2}(1 - \Pr[\Pi \in \mathcal{L}]),\end{aligned}$$

which gives $\Pr[\Pi \in \mathcal{L}] \geq \mu(a)/(2 - \mu(a)) \geq \mu(a)/2$, as claimed. For small enough ε, the set $\mathcal{L}$ and the points a, b satisfy the three conditions in Lemma 4.1 with $C_1 = C_2 = \mu(a)/2$ and $\delta = 0$, respectively from (4.42), (4.41) and (4.40). We conclude that

$$\mathrm{IC}_\mu(f, \varepsilon) \leq \mathrm{IC}_\mu(f, 0) - \frac{\mu(a)^2}{4} h\left(\frac{\mu(b)}{4}\varepsilon\right) + 3\varepsilon K \qquad \text{whenever} \qquad \frac{\mu(b)}{\mu(a)}\varepsilon \leq 1/2, \tag{4.44}$$

where $K = \log|\mathcal{X} \times \mathcal{Y}|$. Hence when $\varepsilon \leq 1/2$, by (3.3) we have

$$\mathrm{IC}_\mu(f, \varepsilon) \leq \mathrm{IC}_\mu(f, 0) - \frac{\mu(a)^2 \mu(b)}{16} h(\varepsilon) + 3\varepsilon K. \tag{4.45}$$

We can thus find $\varepsilon_0 > 0$, depending only on $\mu(a), \mu(b), K$, such that for $\varepsilon \leq \varepsilon_0$,

$$\mathrm{IC}_\mu(f, \varepsilon) \leq \mathrm{IC}_\mu(f, 0) - \frac{\mu(a)^2 \mu(b)}{32} h(\varepsilon). \tag{4.46}$$

Case II: f is constant on C but not on $C_A \times C_B$.

We first make a simple observation:

Property A: For any protocol π that performs $[f, 0]$, and for every transcript t of π, there exists at least one point $b \in C$ (which can depend on t) such that $\Pr[XY = b \mid \Pi = t] = 0$.

Indeed, otherwise f would be constant on $C_A \times C_B$ by the rectangle property of protocols (i. e.,

$$\Pr[\Pi = t \mid x_1 y_1] \Pr[\Pi = t \mid x_2 y_2] = \Pr[\Pi = t \mid x_1 y_2] \Pr[\Pi = t \mid x_2 y_1]$$

for all x_1, x_2, y_1, y_2).

Given a protocol π that performs $[f, 0]$ and a point $a \in C$, let the set $\mathcal{L}(\pi, a)$ of transcripts be defined as

$$\mathcal{L}(\pi, a) = \{t : \Pr[XY = a \mid \Pi = t] \geq \mu(a)/2\}. \tag{4.47}$$

The same argument as in **Case I** shows that $\Pr[\Pi \in \mathcal{L}(\pi,a)] \geq \mu(a)/2$. For any other point $b \in C$, define

$$\mathcal{L}(\pi,a,b) = \{t \in \mathcal{L}(\pi,a) : \Pr[XY = b \mid \Pi = t] = 0\}. \tag{4.48}$$

Let $k := |C|$; necessarily $k \geq 3$. By **Property A**, we have

$$\mathcal{L}(\pi,a) = \bigcup_{b \in C} \mathcal{L}(\pi,a,b). \tag{4.49}$$

This implies the existence of a point $b \in C$ with

$$\Pr[\Pi \in \mathcal{L}(\pi,a,b)] \geq \Pr[\Pi \in \mathcal{L}(\pi,a)]/k \geq \mu(a)/2k\,.$$

To sum up, we have shown that there exist two different points $a,b \in C \subseteq \operatorname{supp}\mu$ such that the set of transcripts $\mathcal{L}(\pi,a,b)$ satisfies the following properties:

(1') $\Pr[\Pi \in \mathcal{L}(\pi,a,b)] \geq \mu(a)/2k$;

(2') $\Pr[XY = a \mid \Pi = t] \geq \mu(a)/2$ for every $t \in \mathcal{L}(\pi,a,b)$;

(3') $\Pr[XY = b \mid \Pi = t] = 0$ for every $t \in \mathcal{L}(\pi,a,b)$.

Now consider a sequence of protocols π_n that all perform $[f,0]$ and $\lim_{n\to\infty} \mathrm{IC}_\mu(\pi_n) = \mathrm{IC}_\mu(f,0)$. Fix (arbitrarily) a point $a \in C$. For every protocol π_n we construct $\mathcal{L}(\pi_n,a,b_{\pi_n})$ as above. Since there are only $k-1$ different values of b, by picking a subsequence of π_n if necessary, without loss of generality, we may assume that for some point $b \in C$, $b_{\pi_n} = b$ for all π_n. Hence for every π_n we have a set of transcripts $\mathcal{L}(\pi_n,a,b)$ such that properties (1'), (2') and (3') are all satisfied.

If we compare these three conditions with the conditions in Lemma 4.1, we find that the only issue is that we do not know whether a and b are in the same row or column (in terms of the graph G_μ, whether a and b are adjacent).

Case IIa: a,b are adjacent in G_μ. As we expand on below, we can guarantee that this case happens in the *AND case* (see theorem statement) by choosing $a = x_0y_0$.

For small enough ε, the set $\mathcal{L}(\pi,a,b)$ and the points a,b satisfy the three conditions in Lemma 4.1 with $C_1 = \mu(a)/2k$, $C_2 = \mu(a)/2$ and $\delta = 0$, respectively from (1'), (2') and (3'). We conclude that

$$\mathrm{IC}_\mu(f,\varepsilon) \leq \mathrm{IC}_\mu(f,0) - \frac{\mu(a)^2}{4k} h\left(\frac{\mu(b)}{4}\varepsilon\right) + 3\varepsilon K \qquad \text{whenever} \qquad \frac{\mu(b)}{\mu(a)}\varepsilon \leq 1/2, \tag{4.50}$$

where $K = \log|\mathcal{X} \times \mathcal{Y}|$. Repeating the calculations of Case I, we can find $\varepsilon_0 > 0$, depending only on $\mu(a), \mu(b), K$, such that for $\varepsilon \leq \varepsilon_0$,

$$\mathrm{IC}_\mu(f,\varepsilon) \leq \mathrm{IC}_\mu(f,0) - \frac{\mu(a)^2\mu(b)}{32k} h(\varepsilon). \tag{4.51}$$

Suppose now that we are in the AND case. Choosing $a = x_0y_0$, we see that Property A must hold for some $b \in \{x_0y_1, x_1y_0\}$, since a transcript having positive probability on both x_0y_1 and x_1y_0 also has positive probability on x_1y_1, whereas $f(x_0y_1) \neq f(x_1y_1)$ by assumption. Property (1') thus holds with $k = 2$, and we conclude that for $\varepsilon \leq \varepsilon_0$,

$$\mathrm{IC}_\mu(f,\varepsilon) \leq \mathrm{IC}_\mu(f,0) - \frac{\mu(a)^2\mu(b)}{64} h(\varepsilon). \tag{4.52}$$

Case IIb: a,b are not adjacent in G_μ. To handle this case, we run a *binary search* along a shortest path connecting a and b in C.

Pick an arbitrary point $c \in C$ in some shortest path connecting a and b. For every π_n, sort the transcripts in $\mathcal{L}(\pi_n,a,b)$ according to

$$p_{n,t,c} := \Pr[XY = c \mid \Pi_n = t]$$

in increasing order, where Π_n is the random variable representing the transcript of π_n. Let m_n be the median of the sequence $p_{n,t,c}$ according to the conditional probability measure $\nu_n(t) := \Pr[\Pi_n = t \mid t \in \mathcal{L}(\pi_n,a,b)]$, i. e.,

$$\nu_n(\{t \in \mathcal{L}(\pi_n,a,b) : p_{n,t,c} \leq m_n\}), \nu_n(\{t \in \mathcal{L}(\pi_n,a,b) : p_{n,t,c} \geq m_n\}) \geq 1/2. \tag{4.53}$$

Such a median always exists: if m_n is the smallest value such that $\nu_n(\{t \in \mathcal{L}(\pi_n,a,b) : p_{n,t,c} \leq m_n\}) \geq 1/2$ then

$$\nu_n(\{t \in \mathcal{L}(\pi_n,a,b) : p_{n,t,c} \geq m_n\}) = 1 - \nu_n(\{t \in \mathcal{L}(\pi_n,a,b) : p_{n,t,c} < m_n\}) \geq 1/2\,.$$

As trivially $m_n \in [0,1]$, the sequence m_n must have a convergent subsequence. Again by picking a subsequence from m_n if necessary, we may assume that the sequence m_n itself is convergent, say $\lim_{n\to\infty} m_n = m$; moreover, if $m > 0$, by picking another subsequence we can assume that $m_n \geq m/2$ for all n. The *binary search* algorithm is then given as:

- If $m = 0$, update the set of transcripts to

  $$\mathcal{L}(\pi_n,a,c) := \{t \in \mathcal{L}(\pi_n,a,b) : p_{n,t,c} \leq m_n\}, \tag{4.54}$$

  and continue the algorithm with b replaced by c;

- If $m > 0$, update the set of transcripts to

  $$\mathcal{L}(\pi_n,c,b) := \{t \in \mathcal{L}(\pi_n,a,b) : p_{n,t,c} \geq m_n\}, \tag{4.55}$$

  and continue the algorithm with a replaced by c.

We argue that the three properties are roughly preserved. In the case $m = 0$, Property (2’) is kept, while Properties (1’) and (3’) change to

$$\Pr[\Pi_n \in \mathcal{L}(\pi_n,a,c)] \geq \mu(a)/4k \quad \text{and} \quad \Pr[XY = c \mid \Pi_n = t] \leq m_n,\ \forall\, t \in \mathcal{L}(\pi_n,a,c), \tag{4.56}$$

respectively. In the case $m > 0$, Property (3’) is preserved while Properties (1’) and (2’) change to

$$\Pr[\Pi_n \in \mathcal{L}(\pi_n,c,b)] \geq \mu(a)/4k \quad \text{and} \quad \Pr[XY = c \mid \Pi_n = t] > m/2,\ \forall\, t \in \mathcal{L}(\pi_n,c,b). \tag{4.57}$$

In either case, we have seen that the new set of transcripts $\mathcal{L}(\pi_n,a,b)$ together with the new two points a and b satisfy Condition (1), (2) and (3) in Lemma 4.1 with proper constants (e. g., δ_n in Condition (3) is at most m_n for protocol π_n, and $m_n \to 0$). After finitely many steps, the binary search algorithm has to stop and return two adjacent points a and b. Suppose that it stops after s steps; note that $s \leq \lceil \log k \rceil$. Lemma 4.1 then gives the upper bound

$$\mathrm{IC}_\mu(f,\varepsilon) \leq \mathrm{IC}_\mu(\pi_n) - \frac{\mu(a)}{2^{s+1}k} C_2 h\left(\frac{\varepsilon}{2}\min\{1, C_2 R\}\right) + 3\varepsilon K + \bar{h}(\delta_n/C_2) \tag{4.58}$$

for some $C_2, R, K > 0$ (where C_2, R depend on μ) and a sequence δ_n tending to zero, assuming that

$$R\varepsilon + (1-\varepsilon)\delta_n/C_2 \leq 1/2 \text{ and } \delta_n/C_2 \leq 1/2. \tag{4.59}$$

By picking a subsequence, we can assume that $\delta_n \leq C_2/4$ for all n. Lemma 4.1 then applies for all $\varepsilon \leq 1/(4R)$. Taking the limit of the right-hand side of (4.58) as $n \to \infty$, we obtain

$$\mathrm{IC}_\mu(f,\varepsilon) \leq \mathrm{IC}_\mu(f,0) - \frac{\mu(a)}{2^{s+1}k} C_2 h\left(\frac{\varepsilon}{2}\min\{1, C_2 R\}\right) + 3\varepsilon K = \mathrm{IC}_\mu(f,0) - \Omega(h(\varepsilon)). \tag{4.60}$$

□

4.1.2 Proof of Theorem 3.5

Theorem 3.5 (restated). *For all f, μ, ε, we have*

$$\mathrm{IC}_\mu(f,\varepsilon) \geq \mathrm{IC}_\mu(f,0) - 4|\mathcal{X}||\mathcal{Y}|\bar{h}(\sqrt{\varepsilon}). \tag{4.61}$$

Proof of Theorem 3.5. Without loss of generality assume that μ is a full-support distribution as otherwise we can approximate it by a sequence of full-support distributions and appeal to the continuity of $\mathrm{IC}_\nu(f,\varepsilon)$ with respect to ν. Consider a protocol π that performs $[f,\varepsilon]$. For every leaf ℓ of π, let z_ℓ and μ_ℓ respectively denote the output of the leaf, and the distribution of the inputs conditioned on the leaf ℓ. We will complete it into a protocol π' that performs $[f,0]$, as follows.

On input (X,Y):

- Alice and Bob run the protocol π and reach a leaf ℓ;
- For every $(x,y) \in \Omega_\ell := \{(x,y) : f(x,y) \neq z_\ell\}$, Alice and Bob verify whether $XY = xy$, as follows:
  - If $\mu_\ell(x) \leq \mu_\ell(y)$, Alice reveals whether $X = x$ to Bob, and if yes, Bob reveals whether $Y = y$ to Alice. If $XY = xy$, they terminate.
  - If $\mu_\ell(x) > \mu_\ell(y)$, Bob initiates the verification process.

Clearly, in the end, either both Alice and Bob already revealed their inputs to each other, or otherwise they know $XY \notin \Omega_\ell$, and hence z_ℓ is the correct output. Therefore π' performs the task $[f,0]$.

Next we analyze $\mathrm{IC}_\mu(\pi')$. Let $\pi_{\ell,xy}$ denote the sub-protocol that starts with the distribution μ_ℓ and verifies whether $XY = xy$. In the case when Alice initiates the verification procedure, we have

$$\mathrm{IC}_{\mu_\ell}(\pi_{\ell,xy}) \le h(\mu_\ell(x)) + \mu_\ell(x) h\left(\frac{\mu_\ell(x,y)}{\mu_\ell(x)}\right) \le h(\mu_\ell(x)) + \mu_\ell(x) \le 2\overline{h}(\mu_\ell(x)), \tag{4.62}$$

where by an abuse of notation we are denoting by $\mu_\ell(x)$ the marginal of μ_ℓ on x. We can obtain a similar bound for the case where Bob initiates the process, and hence

$$\begin{aligned}\mathrm{IC}_{\mu_\ell}(\pi_{\ell,xy}) &\le 2\min\{\overline{h}(\mu_\ell(x)), \overline{h}(\mu_\ell(y))\} \\ &= 2\overline{h}\Big(\mu_\ell(x,y) + \min\{\Pr_{\mu_\ell}[X \ne x, Y = y], \Pr_{\mu_\ell}[X = x, Y \ne y]\}\Big) \\ &\le 2\overline{h}(\mu_\ell(x,y)) + 2\overline{h}\Big(\min\{\Pr_{\mu_\ell}[X \ne x, Y = y], \Pr_{\mu_\ell}[X = x, Y \ne y]\}\Big)\end{aligned}$$

by the subadditivity of $\overline{h}$. Using the monotonicity of $\overline{h}$ together with $\min\{a,b\} \le \sqrt{ab}$, we obtain that

$$\mathrm{IC}_{\mu_\ell}(\pi_{\ell,xy}) \le 2\overline{h}(\mu_\ell(x,y)) + 2\overline{h}\left(\sqrt{\Pr_{\mu_\ell}[X = x, Y \ne y]\Pr_{\mu_\ell}[X \ne x, Y = y]}\right) \tag{4.63}$$

holds for every leaf ℓ and $(x,y) \in \Omega_\ell$. Let $\Pi_{\ell,xy}$ denote the transcript of $\pi_{\ell,xy}$. Since $\pi_{\ell,xy}$ is a deterministic protocol, we have $H_{\mu_\ell}(\Pi_{\ell,xy} \mid XY) = 0$, and thus

$$\mathrm{IC}_{\mu_\ell}(\pi_{\ell,xy}) = I(\Pi_{\ell,xy}; Y \mid X) + I(\Pi_{\ell,xy}; X \mid Y) = H_{\mu_\ell}(\Pi_{\ell,xy} \mid X) + H_{\mu_\ell}(\Pi_{\ell,xy} \mid Y). \tag{4.64}$$

Thus the sub-additivity of entropy implies that the information cost of running all the protocols $\pi_{\ell,xy}$ (for all $x,y \in \Omega_\ell$) is bounded by the sum of their individual information cost. Let ℓ be a leaf of π sampled by running π on a random input. By (4.63),

$$\begin{aligned}\mathrm{IC}_\mu(\pi') - \mathrm{IC}_\mu(\pi) &\le \mathbb{E}_\ell \sum_{xy\in\Omega_\ell} \mathrm{IC}_{\mu_\ell}(\pi_{\ell,xy}) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mathbb{E}_\ell 1_{z_\ell \ne f(x,y)} \mathrm{IC}_{\mu_\ell}(\pi_{\ell,xy}) && (4.65)\\ &\le \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} 2\,\mathbb{E}_\ell 1_{z_\ell \ne f(x,y)} \overline{h}(\mu_\ell(x,y)) + && (4.66)\\ &\quad \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} 2\,\mathbb{E}_\ell 1_{z_\ell \ne f(x,y)} \overline{h}\left(\sqrt{\Pr_{\mu_\ell}[X = x, Y \ne y]\Pr_{\mu_\ell}[X \ne x, Y = y]}\right) && (4.67)\\ &\le \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} 2\overline{h}\left(\mathbb{E}_\ell 1_{z_\ell \ne f(x,y)} \mu_\ell(x,y)\right) + && (4.68)\\ &\quad \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} 2\overline{h}\left(\mathbb{E}_\ell \sqrt{1_{z_\ell \ne f(x,y)} \Pr_{\mu_\ell}[X = x, Y \ne y]\Pr_{\mu_\ell}[X \ne x, Y = y]}\right) && (4.69)\end{aligned}$$

where we used the concavity of $\overline{h}$ in the last step.

For the first summand, we have that for every (x,y),

$$\mathop{\mathbb{E}}_{\ell} 1_{z_\ell \neq f(x,y)} \mu_\ell(x,y) = \sum_\ell \Pr[XY = xy, \pi \text{ reaches } \ell] 1_{z_\ell \neq f(x,y)} \tag{4.70}$$

$$= \sum_\ell \Pr[\pi \text{ reaches } \ell \mid XY = xy] \mu(xy) 1_{z_\ell \neq f(x,y)} \tag{4.71}$$

$$= \mu(xy) \sum_\ell \Pr[\pi_{x,y} \text{ reaches } \ell] 1_{z_\ell \neq f(x,y)} = \mu(xy) \Pr[\pi(x,y) \neq f(x,y)] \tag{4.72}$$

$$\leq \mu(xy)\varepsilon \leq \varepsilon, \tag{4.73}$$

where we used that by definition $\mu_\ell(xy) = \Pr[XY = xy \mid \pi \text{ reaches } \ell]$, and the fact that the protocol π performs the task $[f,\varepsilon]$.

Now consider the second summand in (4.69). Since μ_ℓ is obtained by scaling rows and columns of μ, for every $x' \neq x, y' \neq y$, we have (recall that we assumed μ is of full support)

$$\frac{\Pr_\mu[X = x, Y = y] \Pr_\mu[X = x', Y = y']}{\Pr_\mu[X = x', Y = y] \Pr_\mu[X = x, Y = y']} = \frac{\Pr_{\mu_\ell}[X = x, Y = y] \Pr_{\mu_\ell}[X = x', Y = y']}{\Pr_{\mu_\ell}[X = x', Y = y] \Pr_{\mu_\ell}[X = x, Y = y']}, \tag{4.74}$$

hence

$$\begin{aligned}
&\Pr_{\mu_\ell}[X = x, Y \neq y] \Pr_{\mu_\ell}[Y = y, X \neq x] = \left(\sum_{y' \neq y} \Pr_{\mu_\ell}[X = x, Y = y'] \right) \cdot \left(\sum_{x' \neq x} \Pr_{\mu_\ell}[X = x', Y = y] \right) \\
&= \sum_{x' \neq x, y' \neq y} \Pr_{\mu_\ell}[X = x, Y = y'] \Pr_{\mu_\ell}[X = x', Y = y] \\
&= \sum_{x' \neq x, y' \neq y} \frac{\Pr_{\mu_\ell}[X = x, Y = y] \Pr_{\mu_\ell}[X = x', Y = y']}{\Pr_\mu[X = x, Y = y] \Pr_\mu[X = x', Y = y']} \Pr_\mu[X = x', Y = y] \Pr_\mu[X = x, Y = y'] \\
&= \frac{\Pr_{\mu_\ell}[X = x, Y = y]}{\Pr_\mu[X = x, Y = y]} \sum_{x' \neq x, y' \neq y} \frac{\Pr_{\mu_\ell}[X = x', Y = y']}{\Pr_\mu[X = x', Y = y']} \Pr_\mu[X = x', Y = y] \Pr_\mu[X = x, Y = y'].
\end{aligned}$$

Define

$$a_\ell := 1_{z_\ell \neq f(x,y)} \frac{\Pr_{\mu_\ell}[X = x, Y = y]}{\Pr_\mu[X = x, Y = y]} \tag{4.75}$$

and

$$b_\ell := \sum_{x' \neq x, y' \neq y} \frac{\Pr_{\mu_\ell}[X = x', Y = y']}{\Pr_\mu[X = x', Y = y']} \Pr_\mu[X = x', Y = y] \Pr_\mu[X = x, Y = y']. \tag{4.76}$$

Then

$$1_{z_\ell \neq f(x,y)} \Pr_{\mu_\ell}[X = x, Y \neq y] \Pr_{\mu_\ell}[Y = y, X \neq x] = a_\ell b_\ell. \tag{4.77}$$

Now for a_ℓ, by (4.73) we have

$$\mathop{\mathbb{E}}_{\ell} a_\ell = \frac{1}{\mu(xy)} \mathop{\mathbb{E}}_{\ell} 1_{z_\ell \neq f(x,y)} \mu_\ell(x,y) = \Pr[\pi(x,y) \neq f(x,y)] \leq \varepsilon, \tag{4.78}$$

and for b_ℓ, by the fact $\mathbb{E}_\ell \Pr_{\mu_\ell}[X=x',Y=y'] = \Pr_\mu[X=x',Y=y']$ we have

$$\begin{aligned}\mathbb{E}_\ell b_\ell &= \sum_{x'\neq x, y'\neq y} \frac{\mathbb{E}_\ell \Pr_{\mu_\ell}[X=x',Y=y']}{\Pr_\mu[X=x',Y=y']} \Pr_\mu[X=x',Y=y]\Pr_\mu[X=x,Y=y'] \\ &= \sum_{x'\neq x, y'\neq y} \Pr_\mu[X=x',Y=y]\Pr_\mu[X=x,Y=y'] \\ &= \Pr_\mu[X=x,Y\neq y]\Pr_\mu[Y=y,X\neq x] \leq 1.\end{aligned}$$

Hence we can bound the second summand in (4.69) using the Cauchy-Schwarz inequality

$$\mathbb{E}_\ell \sqrt{a_\ell b_\ell} \leq \sqrt{\mathbb{E}_\ell a_\ell \mathbb{E}_\ell b_\ell} \leq \sqrt{\varepsilon}. \tag{4.79}$$

Using (4.69), (4.73), (4.79), and the monotonicity of $\overline{h}$, we have

$$\mathrm{IC}_\mu(f,0) - \mathrm{IC}_\mu(\pi) \leq \mathrm{IC}_\mu(\pi') - \mathrm{IC}_\mu(\pi) \leq 2|\mathcal{X}\times\mathcal{Y}|\overline{h}(\varepsilon) + 2|\mathcal{X}\times\mathcal{Y}|\overline{h}(\sqrt{\varepsilon}) \leq 4|\mathcal{X}\times\mathcal{Y}|\overline{h}(\sqrt{\varepsilon}). \qquad \square$$

4.1.3 Proof of Proposition 3.4

Proposition 3.4 (restated). *Let μ be the distribution defined as*

$$\mu = \begin{array}{|c|c|}\hline 1/2 & 0 \\ \hline 0 & 1/2 \\ \hline\end{array}. \tag{4.80}$$

Then $\mathrm{IC}^{ext}_\mu(\mathrm{XOR},\varepsilon) \geq \mathrm{IC}^{ext}_\mu(\mathrm{XOR},0) - 3\varepsilon$.

Proof of Proposition 3.4. The distribution μ is supported on the inputs $(0,0),(1,1)$, on which the output is 0. It is easy to check (and follows from the analysis below) that $\mathrm{IC}^{ext}_\mu(\mathrm{XOR},0) = 1$, since at the end of any protocol that performs $[\mathrm{XOR},0]$, we know whether the input is $(0,0)$ or $(1,1)$.

Consider a protocol π having at most ε error on every input, where $\varepsilon \leq 1/3$. Let $\mathcal{L}_z$ be the set of transcripts on which the output is z; Every transcript is either in $\mathcal{L}_0$ or $\mathcal{L}_1$.

For each transcript t achievable from the initial distribution, the distribution of $XY \mid t$ is of the form

$$\begin{array}{|c|c|}\hline p & 0 \\ \hline 0 & 1-p \\ \hline\end{array}$$

for some $p = p(t)$. Bayes' law shows that

$$\Pr[t \mid 00] = \frac{\Pr[00 \mid t]\Pr[t]}{\Pr[00]} = 2p(t)\Pr[t], \quad \Pr[t \mid 11] = \frac{\Pr[11 \mid t]\Pr[t]}{\Pr[11]} = 2(1-p(t))\Pr[t]. \tag{4.81}$$

For each transcript t, the rectangle property says $\Pr[t \mid 00]\Pr[t \mid 11] = \Pr[t \mid 10]\Pr[t \mid 01]$. Hence

$$\frac{\Pr[t \mid 01] + \Pr[t \mid 10]}{2} \geq \sqrt{\Pr[t \mid 01]\Pr[t \mid 10]} = \sqrt{\Pr[t \mid 00]\Pr[t \mid 11]} = 2\sqrt{p(t)(1-p(t))}\Pr[t]. \tag{4.82}$$

The protocol π has distributional error at most ε, and so

$$\Pr[\mathcal{L}_1] = \sum_{t\in\mathcal{L}_1} \Pr[t] \le \varepsilon, \quad \text{and} \quad \Pr[\mathcal{L}_0] = \sum_{t\in\mathcal{L}_0} \Pr[t] \ge 1-\varepsilon. \tag{4.83}$$

On the other hand, since π has point-wise error at most ε, we have

$$\sum_{t\in\mathcal{L}_0} \sqrt{p(t)(1-p(t))}\Pr[t] \le \frac{1}{2}\sum_{t\in\mathcal{L}_0} \frac{\Pr[t \mid 01]+\Pr[t \mid 10]}{2} \le \frac{\varepsilon}{2}. \tag{4.84}$$

Finally,

$$I(XY;\Pi) = H(XY) - H(XY \mid \Pi) = 1 - \sum_t \Pr[t]h(p(t)). \tag{4.85}$$

Let T be a random transcript conditioned on belonging to $\mathcal{L}_0$, and consider the random variable $P := p(T)$. On the one hand,

$$1 - I(XY;\Pi) = \sum_t \Pr[t]h(p(t)) \le \Pr[\mathcal{L}_0]\,\mathbb{E}[h(P)] + \Pr[\mathcal{L}_1] \le \mathbb{E}[h(P)] + \varepsilon. \tag{4.86}$$

On the other hand, by (4.84)

$$\mathbb{E}[\sqrt{P(1-P)}] \le \frac{\varepsilon}{2\Pr[\mathcal{L}_0]} \le \frac{\varepsilon}{2(1-\varepsilon)} \le \varepsilon, \tag{4.87}$$

as we assumed $\varepsilon \le 1/3$. Thus it suffices to verify that $\mathbb{E}[h(P)] \le 2\varepsilon$ for any random variable P that takes values in $[0,1]$ and satisfies $\mathbb{E}[\sqrt{P(1-P)}] \le \varepsilon$. Indeed this would imply

$$1 - I(XY;\Pi) \le \mathbb{E}[h(P)] + \varepsilon \le 3\varepsilon, \tag{4.88}$$

alternatively, $\mathrm{IC}^{ext}_\mu(\mathrm{XOR},\varepsilon) \ge 1-3\varepsilon$ for all $\varepsilon \le 1/3$, which in turn shows that $\mathrm{IC}^{ext}_\mu(\mathrm{XOR},0) = 1$.

Apply the change of variable $Q = \sqrt{P(1-P)}$, so that the assumption simplifies to $\mathbb{E}[Q] \le \varepsilon$; note that $0 \le Q \le 1/2$, and $P = (1 \pm \sqrt{1-4Q^2})/2$. Since $h(P) = h(1-P)$, we conclude that

$$\mathbb{E}[h(P)] = \mathbb{E}[\phi(Q)], \text{ where } \phi(Q) = h\left(\frac{1+\sqrt{1-4Q^2}}{2}\right). \tag{4.89}$$

It is routine to check that the function ϕ is monotonically increasing and strictly convex. Since ϕ is continuous and the domain of Q is restricted to $[0,1/2]$, the maximum of $\mathbb{E}[\phi(Q)]$ under the constraint $\mathbb{E}[Q] \le \varepsilon$ is achieved.[3] Since ϕ is increasing, the maximum value of $\mathbb{E}[\phi(Q)]$ is achieved when $\mathbb{E}[Q] = \varepsilon$. Since ϕ is strictly convex, the maximum value of $\mathbb{E}[\phi(Q)]$ is achieved on a measure supported on the endpoints $0, 1/2$. Thus this measure must be $\Pr[Q = 1/2] = 2\varepsilon$ and $\Pr[Q = 0] = 1-2\varepsilon$. So

$$\mathbb{E}[h(P)] = \mathbb{E}[\phi(Q)] \le (1-2\varepsilon)\phi(0) + 2\varepsilon\phi(1/2) = 2\varepsilon. \tag{4.90}$$

□

[3] This follows from Prokhorov's theorem, which implies that the set of probability measures over $[0,1/2]$ is compact with respect to the weak-* topology. The same result also follows from the Riesz representation theorem [28].

4.2 Information complexity with distributional error

Theorem 3.6 (restated). *Let μ be a probability measure on $\mathcal{X} \times \mathcal{Y}$, and let $f\colon \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ satisfy $\mathrm{IC}_\mu(f,\mu,0) > 0$. We have*

$$\mathrm{IC}_\mu(f,\mu,0) - 4|\mathcal{X}||\mathcal{Y}|\overline{h}(\sqrt{\varepsilon/\alpha}) \le \mathrm{IC}_\mu(f,\mu,\varepsilon) \le \mathrm{IC}_\mu(f,\mu,0) - \frac{\alpha^2}{4}h(\varepsilon\alpha/4) + 3\varepsilon \log|\mathcal{X}\times\mathcal{Y}|, \quad (4.91)$$

where $\alpha = \min_{xy\in\operatorname{supp}\mu} \mu(x,y)$.

Proof of Theorem 3.6. **Lower bound:** The proof is almost identical to the proof of Theorem 3.5, however now we start from a distribution μ that possibly does not have full support. Consider a protocol π that performs $[f,\mu,\varepsilon]$, and define z_ℓ and μ_ℓ as in the proof of Theorem 3.5. Now the new protocol π' that performs $[f,\mu,0]$, is defined similar to the one in the proof of Theorem 3.5 with the only difference that the verification is only performed on the set

$$\Omega'_\ell := \{(x,y) : f(x,y) \neq z_\ell\} \cap \operatorname{supp}\mu. \quad (4.92)$$

Obviously π' solves $[f,\mu,0]$. Note that π has point-wise error at most ε/α on every point in $\operatorname{supp}\mu$. Thus the same analysis of Theorem 3.5 shows

$$\mathrm{IC}_\mu(f,\mu,0) - \mathrm{IC}_\mu(\pi) \le \mathrm{IC}_\mu(\pi') - \mathrm{IC}_\mu(\pi) \le 4|\mathcal{X}\times\mathcal{Y}|\overline{h}(\sqrt{\varepsilon/\alpha}). \quad (4.93)$$

Upper bound: For every $z \in \mathcal{Z}$, let $\mathcal{X}_z$ denote the set of all $x \in \mathcal{X}$ such that for some $xy \in \operatorname{supp}\mu$, we have $f(x,y) = z$. Similarly let $\mathcal{Y}_z$ denote the set of all $y \in \mathcal{Y}$ such that for some $xy \in \operatorname{supp}\mu$, we have $f(x,y) = z$. The assumption $\mathrm{IC}_\mu(f,\mu,0) > 0$ implies the existence of distinct $z_1, z_2 \in \mathcal{Z}$ such that either $\mathcal{X}_{z_1} \cap \mathcal{X}_{z_2} \neq \emptyset$ or $\mathcal{Y}_{z_1} \cap \mathcal{Y}_{z_2} \neq \emptyset$, otherwise, Alice and Bob can exchange the unique values of z determined by their inputs, and since with probability 1, these two values coincide, they can perform $[f,\mu,0]$ with zero information cost. Hence without loss of generality assume there exists $x_0\overline{y}, x_1\overline{y} \in \operatorname{supp}\mu$ such that $f(x_0,\overline{y}) \neq f(x_1,\overline{y})$ and $\mu(x_0\overline{y}) \ge \mu(x_1\overline{y})$. We will apply Lemma 4.1. Consider a protocol π with transcript Π that performs $[f,\mu,0]$, and define the set of transcripts

$$\mathcal{L} := \{t \mid \Pr[x_0\overline{y} \mid t] \ge \Pr[x_0\overline{y}]/2\},$$

and note that

$$\Pr[x_0\overline{y}] = \sum_t \Pr[x_0\overline{y} \mid t]\Pr[\Pi = t] \le \Pr[\Pi \in \mathcal{L}] + \Pr[\Pi \notin \mathcal{L}]\frac{\Pr[x_0\overline{y}]}{2},$$

which implies

$$\Pr[\Pi \in \mathcal{L}] \ge \frac{\Pr[x_0\overline{y}]}{2} \ge \frac{\alpha}{2}.$$

Note that the protocol π' defined in Figure 1 performs $[f,\mu,\varepsilon]$. Furthermore we can set $C_1 = C_2 = \alpha/2$ and $\delta = 0$, to obtain

$$\mathrm{IC}_\mu(\pi') \le \mathrm{IC}_\mu(\pi) - \frac{\alpha^2}{4}h\left(\frac{\varepsilon\alpha}{4}\right) + 3\varepsilon\log|X\times Y|,$$

for $\varepsilon \le 1/2$. As

$$-\frac{\alpha^2}{4}h(\varepsilon\alpha/4) + 3\varepsilon\log|\mathcal{X}\times\mathcal{Y}| \ge 0$$

for $\varepsilon \ge 1/2$, this finishes the proof for all $0 \le \varepsilon \le 1$. □

4.3 Non-distributional prior-free information cost

In this section we prove Theorem 3.15, that is

$$\mathrm{IC}(f,\varepsilon) \leq \mathrm{IC}(f,0) - \Omega(h(\varepsilon)). \tag{4.94}$$

First we present some lemmas, and the proof of Theorem 3.15 will appear at the end of this section.

While Theorem 3.2 does not give a uniform bound on the parameters C, ε_0 for every distribution μ, it does for distributions in which there exist two elements with different outputs, that are in the same row or column and whose probabilities are $\Omega(1)$. We will show that for any non-constant function, the worst distribution is of this form; this might be of independent interest.

We start with the following simple lemma.

Lemma 4.2. *Let $f\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$. Suppose that $\operatorname{supp}\mu \subseteq \bigcup_i \mathcal{X}_i\times\mathcal{Y}_i$, where the $\mathcal{X}_i$ are disjoint and the $\mathcal{Y}_i$ are disjoint. Then*

$$\mathrm{IC}_\mu(f,0) = \sum_i \mu(\mathcal{X}_i\times\mathcal{Y}_i)\,\mathrm{IC}_{\mu|_{\mathcal{X}_i\times\mathcal{Y}_i}}(f|_{\mathcal{X}_i\times\mathcal{Y}_i}). \tag{4.95}$$

Proof. The upper bound is easy to see: the players exchange which block they are in, and assuming that they are in the same block, they run an almost optimal protocol for that block. If they are not in the same block, then they exchange inputs, but this happens with probability zero.

In the other direction, let J be the block in which Alice's input lies. Since the value of J is determined by the value of X, for a protocol π with transcript Π, we have

$$I(Y;\Pi\mid X) = I(Y;\Pi\mid XJ) = \sum_j \Pr[J=j]I(Y;\Pi\mid X,J=j) = \sum_j \mu(\mathcal{X}_j\times\mathcal{Y}_j)I(Y;\Pi\mid X,J=j). \tag{4.96}$$

With probability 1, J is also the block in which Bob's input lies, and so

$$\mathrm{IC}_\mu(\pi) = \sum_j \mu(\mathcal{X}_j\times\mathcal{Y}_j)[I(X;\Pi\mid Y,J=j)+I(Y;\Pi\mid X,J=j)] \geq \sum_j \mu(\mathcal{X}_j\times\mathcal{Y}_j)\,\mathrm{IC}_{\mu|_{\mathcal{X}_j\times\mathcal{Y}_j}}(f|_{\mathcal{X}_j\times\mathcal{Y}_j}). \tag{4.97}$$

□

We can therefore restrict our attention (for now) to distributions based on a single block. The crucial observation is the following.

Lemma 4.3. *Let $f\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$, and let μ be a distribution such that f is constant on its support, each atom in the support has probability at least α, and the marginals of the support are $\mathcal{X},\mathcal{Y}$. If f is not constant then there is a distribution ν such that $\mathrm{IC}_\nu(f,0) \geq \mathrm{IC}_\mu(f,0)+C(\alpha)$, where $C(\alpha)>0$ depends only on $\alpha, |\mathcal{X}|, |\mathcal{Y}|$.*

Proof. Let (x_0,y_0) be any point not in the support of μ such that $f(x_0y_0)$ is different from the constant value of f on $\operatorname{supp}\mu$. Since the marginals of the support are $\mathcal{X},\mathcal{Y}$ and every atom in the support has probability at least α, we see that $\Pr[X=x_0],\Pr[Y=y_0]\geq\alpha$.

Let $\nu = \varepsilon\delta_{x_0y_0} + (1-\varepsilon)\mu$, where ε is a parameter to be determined later, and $\delta_{x_0y_0}$ denotes the Dirac measure concentrated on the point (x_0, y_0). Note that $X'Y' \sim \nu$ can be sampled in the following manner. First we pick $XY \sim \mu$ and an independent Bernoulli random variable B with $\Pr[B = 1] = \varepsilon$. Then

$$X'Y' = \begin{cases} XY & \text{if } B = 0, \\ x_0y_0 & \text{if } B = 1. \end{cases} \tag{4.98}$$

Let π be a protocol that performs the task $[f, 0]$, and let Π_{xy} denote the transcript of this protocol when it is run on the input xy. Note that with probability 1, the value of B is determined by the value of $X'Y'$, and thus

$$\begin{aligned} I(X';\Pi_{X'Y'} \mid Y') &= I(X'B;\Pi_{X'Y'} \mid Y') = I(B;\Pi_{X'Y'} \mid Y') + I(X';\Pi_{X'Y'} \mid Y'B) \\ &= I(B;\Pi_{X'Y'} \mid Y') + (1-\varepsilon)I(X;\Pi_{XY} \mid Y). \end{aligned}$$

Moreover, since $f(x_0, y_0)$ is different from the constant value of f on the support of μ, the value of B is determined by $\Pi_{X'Y'}$. Thus $I(B;\Pi_{X'Y'} \mid Y') = H(B \mid Y')$, and

$$I(X';\Pi_{X'Y'} \mid Y') = H(B \mid Y') + (1-\varepsilon)I(X;\Pi_{XY} \mid Y). \tag{4.99}$$

To lower-bound $H(B \mid Y')$, note that

$$\Pr[B = 1 \mid Y' = y_0] = \frac{\Pr[B = 1, Y' = y_0]}{\Pr[Y' = y_0]} = \frac{\varepsilon}{(1-\varepsilon)\Pr[Y = y_0] + \varepsilon} \geq \varepsilon, \tag{4.100}$$

and on the other hand,

$$\Pr[B = 1 \mid Y' = y_0] \leq \frac{\varepsilon}{(1-\varepsilon)\alpha + \varepsilon}, \tag{4.101}$$

which for $\varepsilon \leq \sqrt{\alpha}/2$ will be at most $1-\varepsilon$. Since $\Pr[Y' = y_0] = (1-\varepsilon)\Pr[Y = y_0] + \varepsilon \geq \alpha$, we conclude that $H(B \mid Y') \geq \alpha h(\varepsilon)$. We deduce that

$$I(X';\Pi_{X'Y'} \mid Y') \geq \alpha h(\varepsilon) + (1-\varepsilon)I(X;\Pi_{XY} \mid Y) \geq I(X;\Pi_{XY} \mid Y) + \alpha h(\varepsilon) - \varepsilon \log |\mathcal{X} \times \mathcal{Y}|. \tag{4.102}$$

The gain is

$$I(X';\Pi_{X'Y'} \mid Y') - I(X;\Pi_{XY} \mid Y) \geq \alpha\varepsilon \log \frac{1}{\varepsilon} - \varepsilon \log |\mathcal{X} \times \mathcal{Y}| = \left(\alpha \log \frac{1}{\varepsilon} - \log |\mathcal{X} \times \mathcal{Y}|\right)\varepsilon, \tag{4.103}$$

and so when $\varepsilon \leq \varepsilon_0 := |\mathcal{X} \times \mathcal{Y}|^{-2/\alpha}$, the gain is at least $\varepsilon \log |\mathcal{X} \times \mathcal{Y}|$. Taking $\varepsilon = \min(\varepsilon_0, \sqrt{\alpha}/2)$, we obtain a constant $C(\alpha) > 0$, depending on $|\mathcal{X} \times \mathcal{Y}|$, such that

$$I(X';\Pi_{X'Y'} \mid Y') \geq I(X;\Pi_{XY} \mid Y) + C(\alpha), \tag{4.104}$$

and similarly

$$I(Y';\Pi_{X'Y'} \mid X') \geq I(X;\Pi_{XY} \mid Y) + C(\alpha).$$

This shows that

$$\mathrm{IC}_\nu(f, 0) \geq \mathrm{IC}_\mu(f, 0) + 2C(\alpha). \tag{4.105}$$

□

We obtain the following important consequence.

Lemma 4.4. *Let $f\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$ be a non-constant function. There exist constants $c,\delta>0$, depending only on the function f such that if for all μ, $\mathrm{IC}_\mu(f,0)\geq \mathrm{IC}(f,0)-\delta$ then there exist points P,Q, on the same row or column, such that $\mu(P),\mu(Q)\geq c$ and $f(P)\neq f(Q)$.*

Proof. Call a distribution ν on $\mathcal{X}\times\mathcal{Y}$ *optimal* if $\mathrm{IC}(f,0)=\mathrm{IC}_\nu(f,0)$. Braverman et al. [6] showed that $\mathrm{IC}_\nu(f,0)$ is continuous in ν, and this implies that optimal distributions exist, and moreover the set of optimal distributions is closed. It is also convex, due to the concavity of $\mathrm{IC}_\nu(f,0)$ (see [5]).

For a distribution ν, let $\beta(\nu)$ be the maximal value β such that there exist two points P,Q, on the same row or column, such that $\nu(P),\nu(Q)\geq\beta$ and $f(P)\neq f(Q)$. Note that $\beta(\nu)$ is continuous in ν.

Suppose that $\beta(\nu)=0$. For $z\in\mathcal{Z}$, let $\mathcal{X}_z$ be the set of rows on which some point $P\in\operatorname{supp}\nu$ satisfies $f(P)=z$, and define $\mathcal{Y}_z$ analogously. We claim that the sets $\mathcal{X}_z$ for $z\in\mathcal{Z}$ are disjoint, similarly $\mathcal{Y}_z$ are disjoint. Indeed, if $x\in\mathcal{X}_{z_1}\cap\mathcal{X}_{z_2}$, then the row x contains two points P,Q in the support such that $f(P)\neq f(Q)$, and so $\beta(\nu)>0$. Next we show that $\operatorname{supp}\nu\subseteq\bigcup_z\mathcal{X}_z\times\mathcal{Y}_z$. Indeed if $P\in\mathcal{X}_{z_1}\times\mathcal{Y}_{z_2}$ is in the support of ν, and $f(P)\neq z_1$, then there exists some point Q on the same row as P which is in the support and satisfies $f(Q)=z_1$, showing that $\beta(\nu)>0$; a similar conclusion is reached if $f(P)\neq z_2$.

We will show that ν is not optimal. Since f is non-constant, $\mathrm{IC}(f,0)>0$, hence we can assume that $\mathrm{IC}_\mu(f,0)>0$. From the characterization of trivial measures (Theorem 3.17) there exists a block $\mathcal{X}_z\times\mathcal{Y}_z$ on which f is not constant. Lemma 4.3 shows that we can modify the component of ν on that block so as to increase the information complexity, and Lemma 4.2 shows that this increases the information complexity over the entire domain. This implies ν is not optimal, as required.

For $\rho\geq 0$, let $O_\rho=\{\nu:\mathrm{IC}_\nu(f,0)\geq\mathrm{IC}(f,0)-\rho\}$. Continuity of $\mathrm{IC}_\nu(f,0)$ shows that O_ρ is closed. We define $b(\rho)=\inf\{\beta(\nu):\nu\in O_\rho\}$; since β is continuous and O_ρ is closed, the infimum is achieved. In view of the preceding paragraph, $b(0)>0$. Continuity of $\beta(\nu)$ and $\mathrm{IC}_\nu(f,0)$ shows that $b(\rho)$ is continuous as well, and so $b(\delta)>0$ for some $\delta>0$. The proof is complete by taking $c=b(\delta)$. ☐

We can now apply Theorem 3.2 to deduce that $\mathrm{IC}(f,\varepsilon)\leq\mathrm{IC}(f,0)-\Omega(h(\varepsilon))$.

Theorem 3.15 (restated). *If $f\colon\mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$ is non-constant then*

$$\mathrm{IC}(f,\varepsilon)\leq\mathrm{IC}(f,0)-C(f)h(\varepsilon), \tag{4.106}$$

where $C(f)>0$.

Proof. Let c,δ be the parameters from Lemma 4.4. For a distribution μ, either $\mathrm{IC}_\mu(f,0)\leq\mathrm{IC}(f,0)-\delta$ or Theorem 3.2 shows that

$$\mathrm{IC}_\mu(f,\varepsilon)\leq\mathrm{IC}_\mu(f,0)-(c^3/32)h(\varepsilon)\leq\mathrm{IC}(f,0)-(c^3/32)h(\varepsilon)$$

for all $\varepsilon\leq\varepsilon_0$ where ε_0 depends only on c and $|\mathcal{X}\times\mathcal{Y}|$. Choose ε sufficiently small such that $(c^3/32)h(\varepsilon)\leq\delta$ and $\varepsilon\leq\varepsilon_0$; we conclude in both cases that $\mathrm{IC}_\mu(f,\varepsilon)\leq\mathrm{IC}(f,0)-C(f)h(\varepsilon)$. ☐

4.4 A characterization of trivial measures

First we present the proof of the external case, i. e., Theorem 3.20, as it is simpler.

Theorem 3.20 (restated). *Let $f\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$ be an arbitrary function, and μ a distribution on $\mathcal{X}\times\mathcal{Y}$. The distribution μ is external-trivial iff it is strongly external-trivial iff it is structurally external-trivial.*

Proof of Theorem 3.20. **If μ is external-trivial then μ is structurally external-trivial.** Suppose that μ is external-trivial but not structurally external-trivial. We will reach a contradiction.

We start by showing that if μ is external-trivial then f has to be constant on the support of μ. Indeed, suppose that the protocol π computes f correctly, and denote by Π the transcript of π. The data processing inequality shows that

$$I(\Pi;XY)\geq I(\Pi;f(XY))=H(f(XY))-H(f(XY)\mid\Pi)=H(f(XY)). \tag{4.107}$$

This shows that μ can only be external-trivial if $H(f(XY))=0$, that is, if f is constant on the support of μ. From now, we assume that this is indeed the case.

Let ab be an arbitrary point in the support of μ, and let $c=f(ab)$. Since μ is not structurally external-trivial, there must be some input $x_0y_0\in S_A\times S_B$ for which $f(x_0y_0)\neq c$. Note that x_0y_0 is not in the support of μ. Since $x_0\in S_A$, x_0y_1 is in the support of μ for some $y_1\in S_B$. Similarly, x_1y_0 is in the support of μ for some $x_1\in S_A$.

Since μ is external-trivial, there is a sequence π_n of protocols computing f correctly on every input such that $I(XY;\Pi_n)\to 0$, where $XY\sim\mu$. We think of π_n also as a distribution over transcripts t. Since $f(XY)=c$ with probability 1, if $\pi_n(t)>0$ then the transcript t indicates that the output is c. Let p_n be the joint distribution of X,Y,t. Recall that

$$D(p_n(x,y,t)\parallel\mu(x,y)\pi_n(t))=I(XY;\Pi_n),$$

hence $D(p_n(x,y,t)\parallel\mu(x,y)\pi_n(t))\to 0$.

For two distributions μ and ν on a finite space, Pinsker's inequality states that

$$D(\mu\parallel\nu)\geq\frac{1}{2}\|\mu-\nu\|_1^2.$$

This implies that $\|p_n(x,y,t)-\mu(x,y)\pi_n(t)\|_1\to 0$. On the other hand, for every transcript t appearing with positive probability, either $p_n(x_0,y_1,t)=0$ or $p_n(x_1,y_0,t)=0$: otherwise $p_n(x_0,y_0,t)>0$ (due to the rectangular property of protocols), contradicting the correctness of π_n (since $f(x_0y_0)\neq c$). Therefore

$$|\mu(x_0,y_1)\pi_n(t)-p_n(x_0,y_1,t)|+|\mu(x_1,y_0)\pi_n(t)-p_n(x_1,y_0,t)|\geq\pi_n(t)\min(\mu(x_0,y_1),\mu(x_1,y_0)). \tag{4.108}$$

Summing over all transcripts having positive probability, we deduce that

$$\|p_n(x,y,t)-\mu(x,y)\pi_n(t)\|_1\geq\sum_t\pi_n(t)\min(\mu(x_0,y_1),\mu(x_1,y_0))=\min(\mu(x_0,y_1),\mu(x_1,y_0)), \tag{4.109}$$

contradicting our assumption that

$$\|p_n(x,y,t)-\mu(x,y)\pi_n(t)\|_1\to 0.$$

If μ is structurally external-trivial then μ is strongly external-trivial. Consider the following protocol. Alice tells Bob whether her input is in S_A. Bob tells Alice whether his input is in S_B. If the input is in $S_A \times S_B$, then the output is known. Otherwise, the players reveal their inputs (but this happens with probability zero). It's not difficult to check that this protocol has zero external information cost.

If μ is strongly external-trivial then μ is external-trivial. This is obvious. □

We comment that our proof gives an explicit lower bound on $\mathrm{IC}_\mu^{ext}(f,0)$ whenever μ is not external-trivial.

Next we present the proof of Theorem 3.17, showing that all our definitions of internal triviality are equivalent. As before, we can get an explicit lower bound on $\mathrm{IC}_\mu(f,0)$ whenever μ is not internal-trivial.

Theorem 3.17 (restated). *Let $f\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$ be an arbitrary function, and μ a distribution on $\mathcal{X}\times\mathcal{Y}$. The distribution μ is internal-trivial iff it is strongly internal-trivial iff it is structurally internal-trivial.*

Proof of Theorem 3.17. **If μ is internal-trivial then μ is structurally internal-trivial.** Suppose that μ is internal-trivial but not structurally internal-trivial. We will reach a contradiction.

Since μ is internal-trivial, there is a sequence of protocols π_n such that $I(X;\Pi_n \mid Y)+I(Y;\Pi_n \mid X)\to 0$. In particular,

$$I(X;\Pi_n \mid Y), I(Y;\Pi_n \mid X)\to 0.$$

Moreover, for every $x\in S_A$ and for every $y\in S_B$,

$$I(X;\Pi_n \mid Y=y), I(Y;\Pi_n \mid X=x)\to 0.$$

Let $p_n(x,y,t)$ be the joint probability of the input and of the transcript of π_n being t. We also think of π_n as a distribution over transcripts. As in the proof of Theorem 3.20, using Pinsker's inequality we deduce that for all $y\in S_B$,

$$\|p_n(x,t \mid y)-\mu(x \mid y)\pi_n(t \mid y)\|_1 \to 0,$$

and so for all $y\in S_B$,

$$B_y := \sum_{x,t} |p_n(x,y,t)-\mu(x,y)\pi_n(t \mid y)| \to 0. \tag{4.110}$$

Similarly, for all $x\in S_A$ we have

$$A_x := \sum_{y,t} |p_n(x,y,t)-\mu(x,y)\pi_n(t \mid x)| \to 0. \tag{4.111}$$

According to Lemma 3.19, there exists a connected component C of G_μ such that f is not constant on $C_A\times C_B$. Suppose first that there is an edge (P,Q) on which f is not constant. Without loss of generality, assume $P=(a,y_0)$ and $Q=(a,y_1)$. Thus

$$\sum_t |p_n(a,y_0,t)-\mu(a,y_0)\pi_n(t \mid a)|+|p_n(a,y_1,t)-\mu(a,y_1)\pi_n(t \mid a)| \to 0. \tag{4.112}$$

On the other hand, for each transcript t either $p_n(a,y_0,t)=0$ or $p_n(a,y_1,t)=0$, since $f(ay_0)\neq f(ay_1)$. Thus

$$\sum_t |p_n(a,y_0,t)-\mu(a,y_0)\pi_n(t\mid a)|+|p_n(a,y_1,t)-\mu(a,y_1)\pi_n(t\mid a)| \geq \sum_t \pi_n(t\mid a)\min(\mu(a,y_0),\mu(a,y_1)) = \min(\mu(a,y_0),\mu(a,y_1)),$$

contradicting the assumption that the left-hand side tends to zero.

Suppose next that f is constant across all edges (and so on the entire connected component), say $f(x,y)=c$ for all $(x,y)\in C$. Since f is not monochromatic on $C_A\times C_B$, there must exist a point $P\in C_A\times C_B$ such that $f(P)\neq c$. There must be points $P_A,P_B\in\operatorname{supp}\mu$ with the same row and column (respectively) as P. Since P_A,P_B are in the same connected component, there is some path $P_A=Q_0,Q_1,\dots,Q_m=P_B$ connecting them: for every $i<m$, Q_i,Q_{i+1} are either in the same row or in the same column. We can assume that $m\leq M:=|\mathcal{X}|+|\mathcal{Y}|$. No transcript can have positive probability for both Q_0 and Q_m, since otherwise it would have positive probability for P as well, and this cannot happen since $f(Q_0)=f(Q_m)=c$ while $f(P)\neq c$.

Let t be any transcript satisfying $p_n(Q_0,t)>0$. Since $p_n(Q_m,t)=0$, there must be an index i such that

$$p_n(t\mid Q_i)-p_n(t\mid Q_{i+1})\geq p_n(t\mid Q_0)/m\geq p_n(t\mid Q_0)/M\,.$$

Assume without loss of generality that $Q_i=(a,y_0)$ and $Q_{i+1}=(a,y_1)$. The contribution of t to A_a is

$$\begin{aligned}|\mu(a,y_0)\pi_n(t\mid a)-p_n(a,y_0,t)|+|\mu(a,y_1)\pi_n(t\mid a)-p_n(a,y_1,t)| &= \\ \mu(a,y_0)|\pi_n(t\mid a)-p_n(t\mid a,y_0)|+\mu(a,y_1)|\pi_n(t\mid a)-p_n(t\mid a,y_1)| &\geq \\ \frac{\min(\mu(a,y_0),\mu(a,y_1))}{M}p_n(t\mid Q_0) \geq \frac{\min(\mu(a,y_0),\mu(a,y_1))}{M}p_n(Q_0,t)&,\end{aligned}$$

using the triangle inequality in the form $|\alpha-\gamma|+|\gamma-\beta|\geq|\alpha-\beta|$.

Denoting by δ the minimum of $\mu(x,y)$ over the support of μ, we conclude that $\sum_x A_x+\sum_y B_y$ is at least

$$\sum_t \frac{\delta}{M}p_n(Q_0,t)=\frac{\delta}{M}\mu(Q_0)\geq\frac{\delta^2}{M}, \tag{4.113}$$

contradicting our assumption that

$$\sum_x A_x+\sum_y B_y\to 0.$$

If μ is structurally internal-trivial then μ is strongly internal-trivial. Consider the following protocol. Alice tells Bob which block $\mathcal{X}_i$ her input belongs to. Bob tells Alice which block $\mathcal{Y}_i$ his input belongs to. If the input is in $\mathcal{X}_i\times\mathcal{Y}_i$, then the output is known. Otherwise, the players reveal their inputs (but this happens with probability zero). It's not difficult to check that this protocol has zero internal information cost.

If μ is strongly internal-trivial then μ is internal-trivial. This is obvious. □

5 Parametrization of all distributions as product distributions

In Section 2.5 we discussed how a communication protocol can be interpreted as a random walk on the set of distributions on $\mathcal{X} \times \mathcal{Y}$. Every time a player sends a signal, we update the underlying distribution based on the information provided by the sent signal. These updates are by scaling either the $\mathcal{X}$ marginal or the $\mathcal{Y}$ marginal of the distribution. This restricted way in which the underling distribution can be updated will allow us to parametrize the set of all reachable distributions from a specific distribution $\overline{\mu}$ in such a way that the changes are captured by product measures. First note that each reachable distribution $\overline{\mu}'$ can be identified by the constants that multiplied $\overline{\mu}$ to obtain $\overline{\mu}'$.

To formalize this intuition, we have the following definition.

Definition 5.1. For two distributions $\mu, \nu \in \Delta(\mathcal{X}, \mathcal{Y})$, define

$$\mu \odot \nu := \frac{\mu \cdot \nu}{\langle \mu, \nu \rangle}, \tag{5.1}$$

where $\mu \cdot \nu$ is the usual point-wise product of the two measures.

Clearly, $\mu \odot \nu \in \Delta(\mathcal{X}, \mathcal{Y})$ unless $\langle \mu, \nu \rangle = 0$, in which case the product is undefined. For our purposes, we will consider decompositions of the form $\overline{\mu} = \nu \odot \mu$, where μ is a *product measure*. The statement "$\overline{\mu}$ is a distribution obtained from ν by scaling its rows and columns" is equivalent to "there exists a product measure μ such that $\overline{\mu} = \nu \odot \mu$." Note that if μ is the uniform distribution, then $\nu = \mu \odot \nu$ for all distributions ν.

Let $\overline{\mu}$ be the prior distribution on $\mathcal{X} \times \mathcal{Y}$ in a communication protocol. We fix a decomposition $\overline{\mu} = \nu \odot \mu$, where μ is a product distribution. For every distribution $\overline{\mu}'$ reachable from $\overline{\mu}$ there is a product distribution μ' such that $\overline{\mu}' = \nu \odot \mu'$, for the same distribution ν. This follows from the fact that $\overline{\mu}'$ is obtained from $\overline{\mu}$ by scaling its rows and columns; therefore if we scale the rows and columns of μ by the same constants and then normalize it, we obtain the desired μ'. In such a decomposition $\overline{\mu} = \nu \odot \mu$, $\overline{\mu}$ is called the *real* distribution, ν the *reference distribution* and μ the *pretend distribution*.

We would like to work with product distributions since they are simpler, and easier to analyze, as we will demonstrate in Section 6. Therefore, we define a *pretend random walk*, which is a random walk on pretend distributions, as opposed to the normal random walk presented in Section 2.5, which we call the *real random walk* to distinguish it from the pretend one. It starts from a product measure $\mu = (\mu^{\mathcal{X}}, \mu^{\mathcal{Y}})$, where $\mu^{\mathcal{X}}$ and $\mu^{\mathcal{Y}}$ are the $\mathcal{X}$ and $\mathcal{Y}$ marginals of μ. At each step we either move by scaling the $\Delta(\mathcal{X})$ marginal or the $\Delta(\mathcal{Y})$ marginal. The transition in $\Delta(\mathcal{X})$ is performed by moving with probability λ_0 to $(\mu_0, \mu^{\mathcal{Y}})$ and with probability λ_1 to $(\mu_1, \mu^{\mathcal{Y}})$, where $0 < \lambda_0, \lambda_1 < 1$, $\lambda_0 + \lambda_1 = 1$ and

$$\sum_{b=0,1} \lambda_b \mu_b = \mu^{\mathcal{X}}.$$

A step in the $\Delta(\mathcal{Y})$ direction is performed similarly.

Every pretend random walk corresponds to a real random walk performed by some protocol. Given such a pretend random walk, and a reference distribution ν, if we replace every distribution μ encountered in the random walk by $\nu \odot \mu$, and scale the transition probabilities, we obtain a real random walk performed by some protocol. Here ν can be any distribution such that $\nu \odot \mu$ is defined for every μ

encountered in the protocol (e. g., if supp ν includes the support of the initial distribution). The inverse transformation is also possible.

To formalize this idea, consider a pretend random walk step, from μ to μ_0 and μ_1 with transition probabilities λ_0 and λ_1, respectively. Fix a reference distribution ν. Then

$$\nu \odot \mu = \frac{\nu \cdot \mu}{\langle \nu, \mu \rangle} = \sum_{b=0,1} \lambda_b \frac{\nu \cdot \mu_b}{\langle \nu, \mu \rangle} = \sum_{b=0,1} \frac{\langle \nu, \mu_b \rangle}{\langle \nu, \mu \rangle} \lambda_b (\nu \odot \mu_b) = \sum_{b=0,1} \overline{\lambda_b} (\nu \odot \mu_b) \tag{5.2}$$

for the values

$$\overline{\lambda_b} = \frac{\langle \nu, \mu_b \rangle}{\langle \nu, \mu \rangle} \lambda_b. \tag{5.3}$$

A calculation shows

$$\sum_{b=0,1} \overline{\lambda_b} = \sum_{b=0,1} \frac{\langle \nu, \mu_b \rangle}{\langle \nu, \mu \rangle} \lambda_b = \frac{\langle \nu, \sum_{b=0,1} \lambda_b \mu_b \rangle}{\langle \nu, \mu \rangle} = \frac{\langle \nu, \mu \rangle}{\langle \nu, \mu \rangle} = 1. \tag{5.4}$$

Furthermore, if the pretend random walk step is performed in the $\Delta(\mathcal{X})$ direction, then $\nu \odot \mu_b$ is obtained by scaling the rows of μ, and if in the $\Delta(\mathcal{Y})$ direction, then by scaling the columns. Therefore, there exists a real random walk step where we move from $\nu \odot \mu$ to $\nu \odot \mu_0$ and $\nu \odot \mu_1$ with probabilities $\overline{\lambda_0}$ and $\overline{\lambda_1}$ respectively. The conversion in the opposite direction, from the real world to the pretend world, is possible due to essentially the same calculations.

Let π_0 and π_1 be the two branches of the protocol π corresponding to the value of the first bit that was sent. Let $\overline{\mu}$ be an input distribution that moves either to $\overline{\mu}_0$ or to $\overline{\mu}_1$ with probabilities $\overline{\lambda_0}$ and $\overline{\lambda_1}$, respectively. Recall the notion of concealed information we discussed in Section 2.5, the following equation regarding the concealed information,

$$\mathrm{CI}_{\overline{\mu}}(\pi) = \sum_{b=0,1} \overline{\lambda_b} \, \mathrm{CI}_{\overline{\mu}_b}(\pi_b) \tag{5.5}$$

translates to

$$\mathrm{CI}_{\nu \odot \mu}(\pi) = \sum_{b=0,1} \frac{\langle \nu, \mu_b \rangle}{\langle \nu, \mu \rangle} \lambda_b \, \mathrm{CI}_{\nu \odot \mu_b}(\pi_b). \tag{5.6}$$

Multiplying by $\langle \nu, \mu \rangle$ we get

$$\mathrm{CI}_{\nu \odot \mu}(\pi) \langle \nu, \mu \rangle = \sum_{b=0,1} \lambda_b \langle \nu, \mu_b \rangle \, \mathrm{CI}_{\nu \odot \mu_b}(\pi_b). \tag{5.7}$$

This motivates the following definition.

Definition 5.2. Let ν be a fixed reference distribution. Define the *scaled information* of a protocol π with respect to a product distribution μ as

$$\mathrm{SIM}_\mu(\pi) := \langle \nu, \mu \rangle \, \mathrm{CI}_{\nu \odot \mu}(\pi). \tag{5.8}$$

Equation (5.8) allows us to write

$$\mathrm{SIM}_\mu(\pi) = \lambda_0\,\mathrm{SIM}_{\mu_0}(\pi_0) + \lambda_1\,\mathrm{SIM}_{\mu_1}(\pi_1). \tag{5.9}$$

Recall that CI is the expected amount of entropy that the players have concealed from each other by the end of the protocol. To formally state this, let $\overline{\mu}$ be a distribution over the inputs, π some protocol and Π the random variable representing the transcript of the protocol. Let $\overline{\mu}_\Pi$ be the random variable that represents the distribution over the inputs given the transcript Π, as defined in Section 2.5. Then

$$\mathrm{CI}_{\overline{\mu}}(\pi) = \mathop{\mathbb{E}}_{\Pi}\left[H_{\overline{\mu}_\Pi}(X \mid Y) + H_{\overline{\mu}_\Pi}(Y \mid X)\right]. \tag{5.10}$$

We will translate (5.10) to a formula involving the pretend random walk. Let $\overline{\mu} = \nu \odot \mu$, and denote by μ_Π the pretend distribution where the pretend random walk ends if its associated protocol has the transcript Π. Or, in a more formal way, μ_Π is the distribution such that $\nu \odot \mu_\Pi = \overline{\mu}_\Pi$. Equation (5.8) implies

$$\mathrm{SIM}_\mu(\pi) = \mathop{\mathbb{E}}_{\Pi}\langle \nu, \mu_\Pi\rangle \left[H_{(\nu\odot\mu)_\Pi}(X \mid Y) + H_{(\nu\odot\mu)_\Pi}(Y \mid X)\right], \tag{5.11}$$

where the probability for each transcript Π is according to the pretend random walk rather than the real one.

One should ask: What is the probability of a transcript t in the pretend random walk, given its probability $\overline{\lambda}$ in the real world? The answer turns out to be very simple. Let $\overline{\mu}^0, \ldots, \overline{\mu}^k$ be the real distributions encountered in the real random walk, where $\overline{\mu}^0$ is the input distribution and $\overline{\mu}^k = \overline{\mu}_t$ is the last distribution encountered. For all $1 \le i \le k$, let $\overline{\lambda^i}$ be the transition probability from $\overline{\mu}^{i-1}$ to $\overline{\mu}^i$ in the real random walk, so that $\overline{\lambda} = \overline{\lambda^1} \cdots \overline{\lambda^k}$. Let μ^i be the pretend distribution associated with $\overline{\mu}^i$ such that $\overline{\mu}^i = \nu \odot \mu^i$ for all i . Then, the transition probability from μ^{i-1} to μ^i in the pretend world equals

$$\lambda^i = \frac{\langle \nu, \mu^{i-1}\rangle}{\langle \nu, \mu^i\rangle}\overline{\lambda^i}, \tag{5.12}$$

using the conversion in (5.3). Multiplying all together, we get that the probability of t in the pretend world is

$$\lambda = \prod_{i=1}^{k} \lambda^i = \prod_{i=1}^{k} \frac{\langle \nu, \mu^{i-1}\rangle}{\langle \nu, \mu^i\rangle}\overline{\lambda^i} = \frac{\langle \nu, \mu^0\rangle}{\langle \nu, \mu^k\rangle}\overline{\lambda}. \tag{5.13}$$

This equation also shows how one can derive (5.11) from (5.10) by multiplying the equation by $\langle \nu, \mu^0\rangle$.

6 The analysis of the AND function

This section is mainly devoted to proving the only remaining case of Theorem 3.7, i. e., the lower bound on $\mathrm{IC}_\mu(\mathrm{AND}, \varepsilon)$. This is presented below separately as Theorem 6.5. Our general strategy for this proof was sketched in Section 3.3 following Theorem 3.7.

Preliminaries and notations. The section relies strongly on the parametrization of distributions as product distributions, as presented in Section 5. A real distribution is usually denoted as $\overline{\mu}$, and it is usually decomposed as $\overline{\mu} = \nu \odot \mu$, where ν is a symmetric reference distribution and μ a pretend distribution. Pretend distributions are always product ones. We will use the shorthand notation $\mu = (p,q)$ for the product distribution in which $p = \mu(1,0) + \mu(1,1)$ and $q = \mu(0,1) + \mu(1,1)$. The distribution $\overline{\mu}$ will usually be assumed to be of full support, which in turn forces ν and μ to be so too.

We are usually going to be working in a pretend world, dealing with the pretend distributions, and keeping the reference distributions in the background. Furthermore, reference distributions are usually kept fixed. We regard protocols as pretend random walks, as presented in Section 5.

Suppose that we run a protocol π starting at a distribution $\overline{\mu} = \nu \odot \mu$. As we explained in Section 5, for each transcript t of the protocol, there is a product distribution μ_t such that $\nu \odot \mu_t$ is the distribution of the players' inputs conditioned on the protocol terminating at the leaf t. Let Π be the random transcript of the pretend random walk associated with an execution of π on input distribution $\overline{\mu}$. Therefore, for any transcript t, $\Pr[\Pi = t]$ is the probability for the transcript t in the pretend random walk, which might be different than the corresponding probability in the real random walk. Throughout this section our view of the protocol is only by the pretend random walk, therefore all random variables that correspond to Π are assumed to be distributed according to the pretend random walk. Since μ_Π, the pretend distribution on the random transcript Π, is a product distribution, it can be written as $\mu_\Pi = (\mathbf{p},\mathbf{q})$, where $\mathbf{p},\mathbf{q}$ are random variables. We call $(\mathbf{p},\mathbf{q})$ the *leaf distribution* of π. We define a crucial random variable, $\ell = \max(\mathbf{p},\mathbf{q})$.

If π is a zero-error protocol, then the leaf distribution is supported on product distributions of the form $(p,0)$, $(0,q)$ or $(1,1)$, since in order to know the AND of the two players' inputs we need to know that one of the players has input 0, or that both inputs are 1.

Since we are concerned with almost-optimal protocol, we would like to quantify optimality. Given a protocol π, define its *wastage* with respect to a distribution $\overline{\mu}$ by

$$\mathrm{IW}_{\overline{\mu}}(\pi) = \mathrm{IC}_{\overline{\mu}}(\pi) - \mathrm{IC}_{\overline{\mu}}(\mathrm{AND},0) = \mathrm{CI}_{\overline{\mu}}(\mathrm{AND},0) - \mathrm{CI}_{\overline{\mu}}(\pi). \tag{6.1}$$

6.1 Stability results

Braverman et al. [5], studying the complexity of the AND function, suggested a continuous protocol whose information complexity equals $\mathrm{IC}_{\overline{\mu}}(\mathrm{AND},0)$, called the *buzzer protocol*. This protocol is defined differently for any input distribution $\overline{\mu}$. Here we denote this protocol by π^*. The buzzer protocol is not a conventional communication protocol as it has access to a continuous clock, however, it can be viewed as a limit of a sequence of genuine protocols. The information complexity of the protocols in that sequence converge to that of the buzzer protocol, and their leaf distribution converges in distribution.

We start by presenting the leaf distribution of the buzzer protocol. We assume that the input reference distribution is symmetric; its importance will become apparent later on.

Table 1: The leaf distribution of the buzzer protocol starting from (p,q), where $p \geq q$.

Distribution μ_Π	$(p,0)$	$(\ell,0)$, $(0,\ell)$ $(p < \ell < 1)$	$(1,1)$
The probability to reach that distribution	$1-q/p$	$pq/\ell^3\,\mathrm{d}\ell$	pq

As it can be seen in Table 1, this is a mix of discrete probabilities and a continuous density. To verify that the above formulas are correct, we can convert the leaf distribution of the buzzer protocol as it is calculated in [5] for the real random walk to its corresponding leaf distribution in the pretend random walk. The formulas that are discussed in Section 5 can be used to calculate the appropriate scaling of the probabilities as we convert the real random walk to the pretend one.

There is also a second and more intuitive way to obtain these formulas. This is done by considering a sequence of protocols that converges to the buzzer protocol. We describe the protocols in that sequence by their pretend random walk. The initial distribution in the pretend world of a protocol in that sequence is (p,q), where $p,q \in \{0, 1/n, 2/n, \dots, 1\}$. In each step, the pretend random walk moves to one of two adjacent grid points, each with probability half. If we are currently in a distribution $(a/n, b/n)$ where $a \geq b$, then the step moves to one of $(a/n, (b+1)/n)$ and $(a/n, (b-1)/n)$. Otherwise, the protocol moves to one of $((a+1)/n, b/n)$ and $((a-1)/n, b/n)$.

Therefore, starting at the point $(a/n, b/n)$ where $a \geq b$, the random walk moves in the y axis, until it ends up either at $(a/n, 0)$ or at $(a/n, (a+1)/n)$. Since this walk is balanced, the probabilities to get to these points are $1 - b/(a+1)$ and $b/(a+1)$, respectively. Then, from that point the random walk moves in the x axis, until it either gets to the point $(0, (a+1)/n)$ or to $((a+1)/n, (a+1)/n)$, with probabilities $1/(a+1)$ and $a/(a+1)$ respectively. Then again, it ends up either at $((a+1)/n, 0)$ or at $((a+1)/n, (a+2)/n)$, then at $(0, (a+2)/n)$ or $((a+2)/n, (a+2)/n)$ and continues this way, until it either gets to the point $(1,1)$, or to a point of the form $(0, i/n)$ or $(i/n, 0)$. Calculating the leaf distribution of each pretend random walk in that sequence, and taking the limit as $n \to \infty$, results in a leaf distribution, which equals that of the buzzer protocol, as will be explained below.

The buzzer protocol can also be defined similarly as a sequence of converging protocols, where for each protocol in the sequence, the real-world analogue of moving in the y direction is performed whenever $\Pr[X=1] \geq \Pr[Y=1]$, while the analogue of moving in the x direction is performed otherwise. In order for our limit protocol to behave identical to the buzzer protocol, we would like the region $\Pr[X=1] \geq \Pr[Y=1]$ to correspond to the region $p \geq q$. This is done by using a symmetric reference distribution.

Next, we would like to show a stability result, proving that every protocol performing the task $[\mathrm{AND}, 0]$ with nearly optimal information complexity is similar to the buzzer protocol. We measure similarity in terms of the leaf distribution $(\mathbf{p}, \mathbf{q})$, and define the following potential function:

Definition 6.1. Given a protocol π for $[\mathrm{AND}, 0]$, a constant $0 < c < 1$, and a pretend distribution μ, let

$$\Phi_{c,\mu}(\pi) = \mathbb{E}\left[((c - \ell)_+)^2\right], \tag{6.2}$$

where $(\cdot)_+ = \max\{\cdot, 0\}$, and $\ell = \max(\mathbf{p}, \mathbf{q})$. Denote $\Phi_{c,\mu} = \Phi_{c,\mu}(\pi^*)$, where π^* is the buzzer protocol.

The following theorem shows that the value of the potential function is small for nearly optimal protocols.

Theorem 6.2. *Let $\overline{\mu}$ be a full support distribution, and $\overline{\mu} = \nu \odot \mu$ be its decomposition, where ν is a symmetric reference distribution and $\mu = (p,q)$ is the product pretend distribution. Assume that $c \leq \max\{p,q\}$. Let π be a protocol performing* $[\mathrm{AND}, 0]$. *Then*

$$\Phi_{c,\mu}(\pi) = O(\mathrm{IC}_{\overline{\mu}}(\pi) - \mathrm{IC}_{\overline{\mu}}(\mathrm{AND}, 0)) = O(\mathrm{IW}_{\overline{\mu}}(\pi)). \tag{6.3}$$

The constant in the $O(\cdot)$ is uniform whenever $\nu(0,0), \nu(0,1), \nu(1,0), p, q$ are bounded away from 0 *and* 1.

In order to prove this theorem, we measure how each performed step contributes both to the wastage and to the potential function. To measure the wastage, we work with SIM instead of IC, as it is a more natural measure for this task.

Lemma 6.3. *Let $\overline{\mu}$ be a full support distribution, and $\overline{\mu} = \nu \odot \mu$ be its decomposition, where ν is a symmetric reference distribution and μ is the pretend distribution. Let $0 < c < 1$, and let π be the protocol which behaves as follows:*

1. *One step of a pretend random walk is performed, which corresponds to one bit that is sent in the protocol.*

2. *The pretend random walk that corresponds to the buzzer protocol is simulated from that point: assuming that after the first bit was sent the pretend distribution is (p,q), let $\pi^*_{(p,q)}$ be the buzzer protocol for the input distribution $\nu \odot (p,q)$. Then, the pretend random walk that corresponds to $\pi^*_{(p,q)}$ is simulated (the value of (p,q) is different for the case that the first bit equals 1, and when it equals 0).*

Then

$$\Phi_{c,\mu}(\pi) - \Phi_{c,\mu} = O_\nu(\mathrm{SIM}_\mu(\mathrm{AND}, 0) - \mathrm{SIM}_\mu(\pi)). \tag{6.4}$$

The constant in the $O(\cdot)$ is uniform whenever $\nu(0,0), \nu(0,1), \nu(1,0), c$ are bounded away from 0 and 1.

The potential function of Definition 6.1 is defined in that manner so that Lemma 6.3 holds. Let us elaborate on this: assume that a protocol π is defined as in this lemma, with a pretend input distribution of (p,q). Assume that the first step moves from (p,q) either to $(p+\delta)$ or to $(p-\delta)$ with equal probability. Then

$$\begin{aligned}
&\mathrm{SIM}_{(p,q)}(\pi) - \mathrm{SIM}_{(p,q)}(\mathrm{AND}, 0) \\
&= \frac{1}{2}\mathrm{SIM}_{(p+\delta,q)}(\mathrm{AND}, 0) + \frac{1}{2}\mathrm{SIM}_{(p-\delta,q)}(\mathrm{AND}, 0) - \mathrm{SIM}_{(p,q)}(\mathrm{AND}, 0) \\
&\approx \frac{\delta^2}{2}\frac{\partial^2}{\partial p^2}\mathrm{SIM}_{(p,q)}(\mathrm{AND}, 0).
\end{aligned}$$

Thus, this difference has the same order of magnitude as δ^2. We would like the change in the potential function to have the same order. Looking at the function x^2, it holds that

$$\frac{1}{2}(x+\delta)^2 + \frac{1}{2}(x-\delta)^2 - x^2 = \frac{\delta^2}{2}. \tag{6.5}$$

If a protocol π moves according to the direction of the buzzer protocol, then π is the same as π^* and both differences are zero. Therefore, assume that $p > q$, and π moves in the x direction, whereas the buzzer protocol would have moved in the y direction. Roughly speaking, the leaf distribution of π is obtained from the leaf distribution of π^* by splitting some of the mass around $\ell \approx p$ between $\ell \approx p-\delta$ and $\ell \approx p+\delta$. Thus, $\Phi_{c,\mu}(\pi) - \Phi_{c,\mu}$ approximately has the order of magnitude of

$$\frac{1}{2}(c-p-\delta)^2 + \frac{1}{2}(c-p+\delta)^2 - (c-p)^2 = \frac{\delta^2}{2}. \tag{6.6}$$

We chose $(c-p)_+^2$ instead of $(c-p)^2$ since Lemma 6.8 requires the buzzer protocol to have a value of zero. Indeed, by choosing c carefully we can achieve this.

We will prove Lemma 6.3 using the following criterion.

Lemma 6.4. *Let ν be a symmetric reference distribution, and $C > 0$ a constant. Define*

$$F(p,q) = C\,\mathrm{SIM}_{(p,q)}(\mathrm{AND},0) + \Phi_{c,(p,q)}.$$

If for every q, $F(p,q)$ is concave as a function of p, and for every p, $F(p,q)$ is concave as a function of q, then Lemma 6.3 holds, and the constant in the $O(\cdot)$ can be taken to be equal to C.

Proof. Let π be the protocol defined in Lemma 6.3, and let μ be its pretend input distribution. Assume that the pretend random walk of π first moves from μ either to μ_0 or to μ_1, with probabilities λ_0 and λ_1. We assume this step is on the x-direction, thus, the first step is from (p,q) to (p_0,q) or (p_1,q). The analysis for the case that this step is in the y-direction is similar. Let $0 < c < 1$. Then

$$\mathrm{SIM}_{(p,q)}(\pi) = \sum_b \lambda_b\,\mathrm{SIM}_{(p_b,q)}(\mathrm{AND},0),$$

and

$$\Phi_{c,(p,q)} = \sum_b \lambda_b \Phi_{c,(p_b,q)}.$$

From concavity,

$$\begin{aligned} C\,\mathrm{SIM}_{(p,q)}(\mathrm{AND},0) + \Phi_{c,(p,q)} &= F(p,q) \geq \sum_b \lambda_b F(p_b,q) \\ &= \sum_b \lambda_b (C\,\mathrm{SIM}_{(p_b,q)}(\mathrm{AND},0) + \Phi_{c,(p_b,q)}) \\ &= C\,\mathrm{SIM}_{(p,q)}(\pi) + \Phi_{c,(p,q)}(\pi). \end{aligned}$$

□

Thus, our focus would be proving that these concavity conditions hold for some value C. We proceed by calculating $\Phi_{c,(p,q)}$, assuming without loss of generality that $p \geq q$. One can see that whenever $p \geq c$, with probability 1 the leaf distribution of the buzzer protocol satisfies $\ell \geq p \geq c$, and thus the potential function evaluates to 0. Consider the case $p < c$. Using the leaf distribution, we obtain the formula

$$\Phi_{c,(p,q)} = (1-q/p)(c-p)^2 + 2\int_{\ell=p}^{c} \frac{pq}{\ell^3}(c-\ell)^2 d\ell. \tag{6.7}$$

Thus, the general definition is as follows:

$$\Phi_{c,(p,q)} = \begin{cases} 0 & \text{if } \max\{p,q\} \geq c, \\ (1-q/p)(c-p)^2 + 2\int_{\ell=p}^{c} \frac{pq}{\ell^3}(c-\ell)^2 d\ell & \text{if } q \leq p \leq c, \\ (1-p/q)(c-q)^2 + 2\int_{\ell=q}^{c} \frac{pq}{\ell^3}(c-\ell)^2 d\ell & \text{if } p \leq q \leq c. \end{cases} \tag{6.8}$$

In order to apply Lemma 6.4, we start by showing that the function $\Phi_{c,(p,q)}$ is differentiable for all p (in the direction of p) given a fixed value of q, and for all q given a fixed value of p. This is done by calculating the two one-sided derivatives in the points suspected of non-differentiability: $p = q$ and $\max\{p,q\} = c$. To state it into more detail, for any fixed q, we calculate both

$$\left.\frac{\partial \Phi_{c,(p,q)}}{\partial p}\right._{+} = \lim_{h \to 0^+} \frac{\Phi_{c,(p+h,q)} - \Phi_{c,(p,q)}}{h}, \tag{6.9}$$

and

$$\left.\frac{\partial \Phi_{c,(p,q)}}{\partial p}\right._{-} = \lim_{h \to 0^-} \frac{\Phi_{c,(p+h,q)} - \Phi_{c,(p,q)}}{h}, \tag{6.10}$$

and verify that both values are equal in all suspected points. We do the same switching the roles of p and q (though it is not required as this potential function is symmetric, since we assume the reference distribution to be symmetric). Additionally, we calculate its second derivatives whenever they are defined. If $\max\{p,q\} > c$, then they are trivially zero. For $q < p < c$, we get:

$$\frac{\partial^2 \Phi_{c,(p,q)}}{\partial p^2} = 2(1 - q/p) \tag{6.11}$$

and

$$\frac{\partial^2 \Phi_{c,(p,q)}}{\partial q^2} = 0. \tag{6.12}$$

Actually, there is a reason why this second derivative with respect to q is zero. For any $0 < \delta \leq \min\{p-q, q\}$, consider a protocol π that first moves to $(p, q-\delta)$ or to $(p, q+\delta)$, each with probability $1/2$, and then simulates the buzzer protocol. It has the same leaf distribution as the buzzer protocol (in the pretend world). Both the buzzer protocol and π either get to the point $(p,0)$ or to the point (p,p), with probabilities $1 - q/p$ and q/p, respectively. From that point on, both continue the same way, resulting in the same leaf distribution. This validates the equality

$$\Phi_{c,(p,q)} = \frac{1}{2}\Phi_{c,(p,q+\delta)} + \frac{1}{2}\Phi_{c,(p,q-\delta)} \tag{6.13}$$

for all q and δ sufficiently small, which implies linearity in the region $q \in [0,p]$ (given a fixed p).

Similar calculations will now be performed with regard to

$$\mathrm{SIM}_{p,q}(\mathrm{AND}, 0).$$

Denote $x = \nu(0,0), y = \nu(1,0) = \nu(0,1), z = \nu(1,1)$. It is possible to extract the value of this function from the equations in [5], using the conversion from SIM to CI (5.8) and from CI to IC (2.19). Nevertheless, we calculate it using the formula (5.11), which is an expectation over a value obtained in the leafs of the protocol. Let $p \geq q$, and let Π correspond to the buzzer protocol, which starts at distribution (p,q). Then,

$$\begin{aligned}
\mathrm{SIM}_{p,q}(\mathrm{AND},0) &= \mathop{\mathbb{E}}_{\Pi}[\langle \nu, \mu_\Pi\rangle (H_{\mu_\Pi}(X \mid Y) + H_{\mu_\Pi}(Y \mid X))] \\
&= \left(1 - \frac{q}{p}\right)((1-p)x + py)h\left(\frac{py}{(1-p)x+py}\right) + \\
&\quad \int_p^1 \frac{2pq}{\ell^3}((1-\ell)x + \ell y)h\left(\frac{y\ell}{x(1-\ell)+y\ell}\right) \mathrm{d}\ell \\
&= -\left[q(1-p)y + (1-p)(1-q)x \log\frac{(1-p)x}{(1-p)x+py} + \right. \\
&\quad \left. \left(\frac{pqy^2}{x} + (p+q-2pq)y\right)\log\frac{py}{(1-p)x+py}\right].
\end{aligned}$$

Calculating the second derivative, we get for $p > q$,

$$\frac{\partial^2 \,\mathrm{SIM}_{(p,q)}(\mathrm{AND},0)}{\partial p^2} = -2(1-q/p)\frac{xy}{2(1-p)p^2((1-p)x+py)}, \tag{6.14}$$

and

$$\frac{\partial^2 \,\mathrm{SIM}_{(p,q)}(\mathrm{AND},0)}{\partial q^2} = 0. \tag{6.15}$$

The reason that the second derivative is zero is the same as explained for the potential function. For proving differentiability (on each direction separately), the only suspected point is $p = q$. Comparing the two one-sided derivatives implies the result.

Now we are almost ready to apply Lemma 6.4. Define

$$C = \max_{0 \le p \le 1} \frac{2(1-p)p^2((1-p)x+py)}{xy}, \tag{6.16}$$

and $F(p,q) = C\,\mathrm{SIM}_\mu(\pi^*) + \Phi_{c,\mu}$. For any fixed q,

$$\frac{\partial F(p,q)}{\partial p}$$

is continuous, piecewise differentiable, and its derivative,

$$\frac{\partial}{\partial p}\frac{\partial F(p,q)}{\partial p}$$

is non-positive wherever it is defined. Thus,

$$\frac{\partial F(p,q)}{\partial p}$$

is non-increasing, and $F(p,q)$ is concave as a function of p. The same holds when switching the roles of p and q, thus the conditions in Lemma 6.4 are satisfied, which concludes the proof of Lemma 6.3. Finally, we are able to prove Theorem 6.2.

Proof of Theorem 6.2. Let T be the protocol tree of π. This is a directed binary tree with two children for each internal node. Each node corresponds to a state of the protocol when some communication has taken place, and its children are the two consecutive states, chosen according to the bit sent by the player owning the node.

We can construct T using a sequence of trees, $T_1, T_2, \dots, T_k = T$. The tree T_1 contains only the root of T, and for all i, T_i is obtained from T_{i-1} by adding the children of a leaf of T_{i-1} which is not a leaf of T.

Given a tree T_i, construct a protocol π_i, that whenever it reaches a state represented by node v which is not a leaf of T_i, the protocol behaves as π for the next bit sent, and if the state is represented by a leaf of T_i, then the buzzer protocol is simulated from that point on. Let D be the constant in the $O(\cdot)$ guaranteed from Lemma 6.3. The lemma implies that for all i,

$$\Phi_{c,\mu}(\pi_i) - \Phi_{c,\mu}(\pi_{i-1}) \leq D(\mathrm{SIM}_\mu(\pi_{i-1}) - \mathrm{SIM}_\mu(\pi_i)).$$

Summing over i, we get a telescopic summation that results in

$$\Phi_{c,\mu}(\pi) = \Phi_{c,\mu}(\pi_k) - \Phi_{c,\mu}(\pi_1) \leq D(\mathrm{SIM}_\mu(\pi_1) - \mathrm{SIM}_\mu(\pi_k)) = D(\mathrm{SIM}_\mu(\mathrm{AND}, 0) - \mathrm{SIM}_\mu(\pi)). \quad (6.17)$$

We used the fact that

$$\Phi_{c,\mu}(\pi_1) = \Phi_{c,\mu} = 0,$$

which hold since we assumed that $c \leq \max\{p, q\}$, and the leaf distribution of the buzzer protocol has zero mass on $\ell < \max\{p, q\}$, therefore its potential cost is zero. This finishes the proof as

$$\mathrm{SIM}_\mu(\mathrm{AND}, 0) - \mathrm{SIM}_\mu(\pi) = \langle \nu, \mu \rangle (\mathrm{CI}_{\overline{\mu}}(\mathrm{AND}, 0) - \mathrm{CI}_{\overline{\mu}}(\pi)) = \langle \nu, \mu \rangle \, \mathrm{IW}_{\overline{\mu}}(\pi) \leq \mathrm{IW}_{\overline{\mu}}(\pi). \quad (6.18)$$

□

6.2 Lower bound on the information complexity of $\mathrm{IC}_\mu(\mathrm{AND}, \varepsilon)$

In this section, we prove Theorem 3.7 by showing that every distribution $\overline{\mu}$ which is of full support, except perhaps for $\overline{\mu}(1,1)$, satisfies $\mathrm{IC}_{\overline{\mu}}(\mathrm{AND}, \varepsilon) \geq \mathrm{IC}_{\overline{\mu}}(\mathrm{AND}, 0) - O(\overline{h}(\varepsilon))$. Recall that Theorem 3.7 (ii) follows from Part (i) and we have already established the upper bound of Theorem 3.7 (i) in Theorem 3.2. Hence it remains to prove the following theorem.

Theorem 6.5 (The remaining case of Theorem 3.7)**.** *Let $\overline{\mu}$ be a full-support distribution, except perhaps for $\overline{\mu}(1,1)$. For all $\varepsilon \geq 0$,*

$$\mathrm{IC}_{\overline{\mu}}(\mathrm{AND}, \varepsilon) \geq \mathrm{IC}_{\overline{\mu}}(\mathrm{AND}, 0) - O_{\overline{\mu}}(\overline{h}(\varepsilon)). \quad (6.19)$$

The hidden constant can be fixed if $\overline{\mu}(0,0), \overline{\mu}(0,1), \overline{\mu}(1,0)$ are bounded away from 0*.*

The proof uses the idea of *protocol completion*: given a protocol π performing $[\mathrm{AND}, \varepsilon]$, we can create a protocol π_0, which we call the zero-error *completion* of π. Such a protocol π_0 takes the following steps:

- First Alice and Bob simulate π until it terminates.
- Afterwards they run a protocol that solves the AND function with zero error.

The *cost of completion* is the amount of information revealed in the second step, and it is equal to $\mathrm{IC}_{\overline{\mu}}(\pi_0) - \mathrm{IC}_{\overline{\mu}}(\pi)$. We have shown in the proof of Theorem 3.5 that for general functions, this cost is bounded by $O(\overline{h}(\sqrt{\varepsilon}))$, but here we would like to prove a stronger bound of $O(\overline{h}(\varepsilon))$ for protocols that are almost optimal for the AND function. This obviously would yield the desired lower bound, and prove Theorem 6.5. This completion cost can be arbitrarily close to

$$\mathop{\mathbb{E}}_{\Pi}[\mathrm{IC}_{\overline{\mu}_\Pi}(\mathrm{AND}, 0)].$$

In order to bound this quantity, we first bound the information complexity of the AND function.

Lemma 6.6. *Consider a reference distribution ν with $\nu(0,0) = x, \nu(1,0) = \nu(0,1) = y, \nu(1,1) = z$, such that $x,y,z > 0$. Let $\mu = (p,q)$ be a pretend distribution. Let $\overline{\mu} = \nu \odot \mu$, and $\overline{\mu}(1,1) = \delta$. Let $0 < C < 1$ be an arbitrary constant.*

Firstly $\mathrm{IC}_{\overline{\mu}}(\mathrm{AND}, 0) \leq 2\overline{h}(1-\delta)$. Secondly

$$\mathrm{IC}_{\overline{\mu}}(\mathrm{AND}, 0) \leq \begin{cases} O(\overline{h}(\delta/z)) & \text{if } \max(p,q) \geq C, \\ O(\overline{h}(\sqrt{\delta/z})) & \text{if } p,q < C. \end{cases} \tag{6.20}$$

The hidden constants can be fixed if x,y,C are bounded away from both 0 and 1.

Proof. First we prove that $\mathrm{IC}_{\overline{\mu}}(\mathrm{AND}, 0) \leq 2\overline{h}(1-\delta)$. Assume that $\delta \geq 1/2$, as otherwise the inequality trivially follows. The information complexity is achieved by a protocol where both Alice and Bob send their inputs. The cost of that protocol is at most $H(XY) \leq H(X) + H(Y) \leq 2h(\delta)$.

For proving the other bounds, assume that $\delta < 1/2$, since otherwise the lemma trivially follows. If $p,q > 1/2$, then

$$\delta = \frac{\nu(1,1)pq}{\langle \nu, \mu \rangle} \geq \nu(1,1) = z,$$

as

$$\langle \nu, \mu \rangle = (1-p)(1-q)x + [p(1-q) + (1-p)q]y + pqz \leq (x+2y+z)pq = pq. \tag{6.21}$$

In this case, the lemma follows.

Assume that either $p \leq 1/2$ or $q \leq 1/2$. Without loss of generality, $p \leq q$. We will analyze the protocol in which Alice first sends her input to Bob, and if $X = 1$ then Bob sends his input to Alice. This protocol has a cost of

$$H(X \mid Y) + \Pr[X=1] H(Y \mid X=1) \leq H(X) + \Pr[X=1] \leq \overline{h}(\Pr[X=1]) + \Pr[X=1]. \tag{6.22}$$

The obtained bound is monotonic in $\Pr[X=1]$, a fact that we will use.

Now

$$\Pr[X=1] = \frac{p(1-q)y + pqz}{\langle \nu, \mu \rangle} \leq \frac{p(y+z)}{\langle \nu, \mu \rangle} = \frac{\delta(y+z)}{zq}. \tag{6.23}$$

Thus, if $q \geq C$, then the cost of completion is at most

$$\overline{h}\left(\frac{\delta(y+z)}{zC}\right) + \frac{\delta(y+z)}{zC} \leq \frac{(y+z)\delta}{Cz} + \begin{cases} \overline{h}(\delta/z) & \text{if } \frac{y+z}{C} < 1, \\ \frac{y+z}{C} 2\overline{h}(\delta/z) & \text{otherwise}, \end{cases} \tag{6.24}$$

using the bound $\overline{h}(cx) \le 2c\overline{h}(x)$ for all $c > 1$, from (3.3).

If $q \le C$, $\Pr[X=1]$ is maximized at $q = p$. Assume indeed that $p = q$. We will bound its value from below. The equation

$$\frac{q^2 z}{\langle \nu, \mu \rangle} = \frac{q^2 z}{\langle \nu, \mu \rangle} = \delta$$

implies

$$q = \sqrt{\frac{\delta \langle \nu, \mu \rangle}{z}}. \tag{6.25}$$

Now since

$$\langle \nu, \mu \rangle \ge \nu(0,0)\mu(0,0) = (1-p)(1-q)x \ge (1-C)^2 x, \tag{6.26}$$

we have

$$\Pr[X=1] \le \frac{\delta(y+z)}{zq} \le \sqrt{\frac{\delta}{z}} \frac{y+z}{(1-C)\sqrt{x}}. \tag{6.27}$$

The proof concludes applying similar calculations as in (6.24). □

Next, we use this bound to show that if the probability that $\max\{\mathbf{p},\mathbf{q}\}$ does not exceed some constant is very small, then one can get an improvement over $\overline{h}(\sqrt{\varepsilon})$ for the completion cost.

Lemma 6.7. *Let ν be a symmetric reference distribution with $\nu(0,0) = x$, $\nu(0,1) = \nu(1,0) = y$ and $\nu(1,1) = z > 0$. Let $\mu = (p,q)$ be a pretend distribution, and let $\overline{\mu} = \nu \odot \mu = \nu$.*

Let π be a protocol performing $[\mathrm{AND}, \varepsilon]$. Let $0 < C < 1$ be an arbitrary constant,

$$\kappa = \Pr[\max\{\mathbf{p},\mathbf{q}\} \le C].$$

The protocol π can be completed to a zero-error protocol using an additional information cost of

$$O\left(\kappa \overline{h}(\sqrt{\varepsilon/\kappa}) + (1-\kappa)\overline{h}(\tfrac{\varepsilon}{1-\kappa})\right), \tag{6.28}$$

where the cost is according to the distribution $\overline{\mu}$, and the hidden constant in $O(\cdot)$ can be fixed if x,y,p,q,C are all bounded away from both 0 and 1.

Proof. First, note that

$$\overline{\mu}(1,1) = \frac{zpq}{\langle \nu, \mu \rangle} \le \frac{zpq}{x(1-p)(1-q)} = O(z). \tag{6.29}$$

Let ψ be the random variable denoting the completion cost as a function of Π. Let $1_{o=b}$ be the indicator of whether π outputs b given the transcript Π, for $b = 0,1$. The total completion cost is

$$\mathbb{E}[\psi] = \sum_{b=0,1} \mathbb{E}[\psi 1_{o=b}]. \tag{6.30}$$

We start by bounding $\mathbb{E}[\psi 1_{o=1}]$. Let δ be the random variable which equals $\overline{\mu}_\Pi(1,1)$.

$$\mathbb{E}[(1-\delta)1_{o=1}] = \Pr[(X,Y) \neq (1,1), \pi \text{ outputs } 1] \le \varepsilon. \tag{6.31}$$

From Lemma 6.6, the completion cost ψ is at most $2\overline{h}(1-\delta)$. From the concavity of $\overline{h}$,

$$\mathbb{E}[\psi 1_{o=1}] = \mathbb{E}\, O(\overline{h}(1-\delta))1_{o=1} = \mathbb{E}\, O(\overline{h}((1-\delta)1_{o=1})) \le O(\overline{h}(\mathbb{E}[(1-\delta)1_{o=1}])) \le O(\overline{h}(\varepsilon)). \quad (6.32)$$

This can be bounded as desired since in both cases of $\kappa > 1/2$ and $\kappa \le 1/2$, we have

$$\overline{h}(\varepsilon) = O\left(\kappa\overline{h}(\sqrt{\varepsilon/\kappa}) + (1-\kappa)\overline{h}(\tfrac{\varepsilon}{1-\kappa})\right). \quad (6.33)$$

Next we bound $\mathbb{E}[\psi 1_{o=0}]$.

$$\mathbb{E}[\delta 1_{o=0}] = \Pr[(X,Y) = (1,1), \pi \text{ outputs } 0] \le \varepsilon\overline{\mu}(1,1) \le \varepsilon O(z). \quad (6.34)$$

Let S be the event that $\max\{\mathbf{p},\mathbf{q}\} \le C$. Then,

$$\mathbb{E}[\delta 1_{o=0} \mid S] \le \varepsilon O(z)/\Pr[S] = \varepsilon O(z)/\kappa. \quad (6.35)$$

$$\mathbb{E}[\delta 1_{o=0} \mid \overline{S}] \le \varepsilon O(z)/(1-\kappa). \quad (6.36)$$

From Lemma 6.6, the completion cost is of order of $\overline{h}\left(\sqrt{\delta/z}\right)$ when S happens, and $\overline{h}(\delta/z)$ otherwise.

$$\begin{aligned}
\mathbb{E}[\psi 1_{o=0}] &= \Pr[S]\,\mathbb{E}[\psi 1_{o=0} \mid S] + \Pr[\overline{S}]\,\mathbb{E}[\psi 1_{o=0} \mid \overline{S}] && (6.37)\\
&= O\left(\kappa\mathbb{E}\left[\overline{h}\left(\sqrt{\delta 1_{o=0}/z}\right) \mid S\right] + (1-\kappa)\mathbb{E}[\overline{h}(\delta 1_{o=0}/z) \mid \overline{S}]\right) && (6.38)\\
&\le O\left(\kappa\overline{h}\left(\sqrt{\mathbb{E}[\delta 1_{o=0} \mid S]/z}\right) + (1-\kappa)\overline{h}(\mathbb{E}[\delta 1_{o=0} \mid \overline{S}]/z)\right) && (6.39)\\
&\le O\left(\kappa\overline{h}\left(\sqrt{O(\varepsilon)/\kappa}\right) + (1-\kappa)\overline{h}(O(\varepsilon)/(1-\kappa))\right) && (6.40)\\
&\le O\left(\kappa\overline{h}\left(\sqrt{\varepsilon/\kappa}\right) + (1-\kappa)\overline{h}\left(\frac{\varepsilon}{1-\kappa}\right)\right), && (6.41)
\end{aligned}$$

where (6.39) follows from the concavity of $\overline{h}(\cdot/z)$ and $\overline{h}(\sqrt{\cdot/z})$, and (6.41) follows from (3.3). □

Consider an almost optimal protocol π_0 so that $\mathrm{IC}_{\overline{\mu}}(\pi_0) - \mathrm{IC}_{\overline{\mu}}(\mathrm{AND}, 0)$ is small. Our stability result, Theorem 6.2, translates this to a bound on the potential function introduced in Definition 6.1. The next lemma uses this to show that for such a protocol π_0, one can obtain a strong bound on the value of κ in Lemma 6.7.

Lemma 6.8. *Let $\overline{\mu}$ be full-support distribution and let $\overline{\mu} = \nu \odot \mu$ be its decomposition, where ν is a symmetric reference distribution, and μ is the pretend distribution. Let $c = \max\left\{\Pr_\mu[X=1], \Pr_\mu[Y=1]\right\}$. Let π be an arbitrary protocol, and π_0 be the completion of π to a protocol performing* $[\mathrm{AND}, 0]$. *Then*

$$\Pr\left[\max\{\mathbf{p},\mathbf{q}\} \le \frac{c}{4}\right] = O_{c,\mu,\nu}(\mathrm{IC}_{\overline{\mu}}(\pi_0) - \mathrm{IC}_{\overline{\mu}}(\mathrm{AND}, 0)). \quad (6.42)$$

The hidden constant can be fixed if $p, q, \mu(0,0), \mu(0,1), \mu(1,0)$ are all bounded away from both 0 *and* 1*, where $\mu = (p,q)$.*

Proof. Let $\ell_{p,q}$ be the distribution of ℓ that corresponds to the buzzer protocol when it is invoked from a pretend distribution parametrized by (p,q).

We start by showing that for any $0 < p,q < 1$,

$$\Pr[\ell_{p,q} \leq 2\max\{p,q\}] \geq \frac{3}{4}. \tag{6.43}$$

Assume without loss of generality that $p \geq q$. Using the leaf distribution from Section 6.1,

$$\Pr[p \leq \ell \leq 2p] = 2\int_p^{2p} \frac{pq}{\ell^3}\,\mathrm{d}\ell + \left(1 - \frac{q}{p}\right) > \frac{3}{4}. \tag{6.44}$$

This implies

$$\begin{aligned}
\Pr\left[\ell_{\pi_0} \leq \frac{c}{2}\right] &= \Pr\left[\ell_{\pi_0} \leq 2\frac{c}{4}\right] \\
&\geq \Pr\left[\max\{\mathbf{p},\mathbf{q}\} \leq \frac{c}{4}\right] \Pr\left[\ell_{\mathbf{p},\mathbf{q}} \leq 2\max\{\mathbf{p},\mathbf{q}\}\right] \\
&\geq \frac{3}{4}\Pr\left[\max\{\mathbf{p},\mathbf{q}\} \leq \frac{c}{4}\right].
\end{aligned}$$

Markov's inequality and Theorem 6.2 imply

$$\Pr\left[\ell_{\pi_0} \leq \frac{c}{2}\right] = \Pr\left[(c-\ell_{\pi_0})_+^2 \geq \frac{c^2}{4}\right] \leq \frac{\mathbb{E}[(c-\ell_{\pi_0})_+^2]}{c^2/4} = \frac{\Phi_{c,\mu}(\pi_0)}{c^2/4} = O(\mathrm{IC}_{\overline{\mu}}(\pi_0) - \mathrm{IC}(\mathrm{AND},0)). \tag{6.45}$$

□

Now we are ready to prove Theorem 6.5, and thus complete the proof of Theorem 3.7.

Proof of Theorem 6.5. We first prove the theorem for the full-support distributions. Consider such a distribution $\overline{\mu}$. Let π be a protocol performing $[\mathrm{AND},\varepsilon]$. We can assume that

$$\mathrm{IC}_{\overline{\mu}}(\pi) \leq \mathrm{IC}_{\overline{\mu}}(\mathrm{AND},0),$$

and let

$$C = \max\left\{\Pr_\mu[X=1], \Pr_\mu[Y=1]\right\}/4, \qquad \kappa = \Pr[\max\{\mathbf{p},\mathbf{q}\} \leq C].$$

Lemma 6.7 constructs a zero-error protocol π_0 whose wastage w is at most

$$w = O\left(\kappa\overline{h}\left(\sqrt{\frac{\varepsilon}{\kappa}}\right) + (1-\kappa)\overline{h}\left(\frac{\varepsilon}{1-\kappa}\right)\right). \tag{6.46}$$

Lemma 6.8 states that $\kappa = O(w)$, and so

$$\kappa = O\left(\kappa\overline{h}\left(\sqrt{\frac{\varepsilon}{\kappa}}\right) + (1-\kappa)\overline{h}\left(\frac{\varepsilon}{1-\kappa}\right)\right). \tag{6.47}$$

If $\varepsilon/(1-\kappa) \le 1/2$, then (3.3) shows that

$$\kappa = O\left(\kappa\overline{h}\left(\sqrt{\frac{\varepsilon}{\kappa}}\right) + \overline{h}(\varepsilon)\right). \tag{6.48}$$

Otherwise, $\kappa \ge 1-2\varepsilon \ge 1/2$ (assuming $\varepsilon \le 1/4$), and so

$$\kappa = O(h(\sqrt{\varepsilon}) + (1-\kappa)) = O(h(\sqrt{\varepsilon}) + \varepsilon), \tag{6.49}$$

which contradicts $\kappa \ge 1/2$ for small enough ε.

Denoting the hidden constant in (6.48) by M, we get

$$\left(1 - Mh\left(\sqrt{\frac{\varepsilon}{\kappa}}\right)\right)\kappa \le Mh(\varepsilon). \tag{6.50}$$

We will show that for small ε, this forces $\kappa \le 2Mh(\varepsilon)$. Indeed, suppose that $\kappa > 2Mh(\varepsilon)$, which implies that $\kappa > 2M\varepsilon\log(1/\varepsilon)$. Then

$$\frac{\varepsilon}{\kappa} < \frac{1}{2M\log(1/\varepsilon)}, \tag{6.51}$$

and so for small enough ε, $Mh(\sqrt{\varepsilon/\kappa}) < 1/2$. This shows that

$$\left(1 - Mh\left(\sqrt{\frac{\varepsilon}{\kappa}}\right)\right)\kappa > \frac{\kappa}{2} > Mh(\varepsilon), \tag{6.52}$$

contradicting the inequality above. We conclude that for small ε we have $\kappa = O(h(\varepsilon))$.

Applying Lemma 6.7 again, we see that

$$\mathrm{IC}_{\overline{\mu}}(\pi_0) - \mathrm{IC}_{\overline{\mu}}(\pi) \le \kappa O\left(\overline{h}\left(\sqrt{\frac{\varepsilon}{\kappa}}\right)\right) + O(\overline{h}(\varepsilon)) \le O(\kappa) + O(\overline{h}(\varepsilon)) = O(\overline{h}(\varepsilon)). \tag{6.53}$$

Since $\mathrm{IC}_{\overline{\mu}}(\pi_0) \ge \mathrm{IC}_{\overline{\mu}}(\mathrm{AND}, 0)$, we conclude that $\mathrm{IC}_{\overline{\mu}}(\pi) \ge \mathrm{IC}_{\overline{\mu}} - O(\overline{h}(\varepsilon))$.

Next consider a distribution $\overline{\mu}$ with $\overline{\mu}(1,1) = 0$, that assigns a strictly positive probability for every other input. There is a series of full support distributions, $\overline{\mu}_1, \overline{\mu}_2, \ldots$ that converge to $\overline{\mu}$, and assume without loss of generality that for every input $a \in \{0,1\}^2$ and for every $n \in \mathbb{N}$, $\overline{\mu}_n(a) \ge \overline{\mu}(a)/2$. From the continuity of information complexity with respect to the tasks $[\mathrm{AND}, 0]$ and $[\mathrm{AND}, \varepsilon]$,

$$\lim_{n\to\infty} \mathrm{IC}_{\overline{\mu}_n}(\mathrm{AND}, 0) = \mathrm{IC}_{\overline{\mu}}(\mathrm{AND}, 0), \tag{6.54}$$

and

$$\lim_{n\to\infty} \mathrm{IC}_{\overline{\mu}_n}(\mathrm{AND}, 0) = \mathrm{IC}_{\overline{\mu}}(\mathrm{AND}, 0). \tag{6.55}$$

Assume that $\overline{\mu}(0,0), \overline{\mu}(0,1), \overline{\mu}(1,0)$ are bounded from below. It is possible to decompose $\overline{\mu}$ into $\nu \odot (p,q)$, where ν is symmetric and $p, q, \nu(0,0), \nu(0,1)$ and $\nu(1,0)$ are bounded. This is done by considering a decomposition where $p = 1/2$ and q is chosen such that ν is symmetric. Therefore, there is a constant $C > 0$ such that

$$\mathrm{IC}_{\mu_n}(\mathrm{AND}, \varepsilon) \ge \mathrm{IC}_{\mu_n}(\mathrm{AND}, \varepsilon) - C\overline{h}(\varepsilon). \tag{6.56}$$

Thus,

$$\mathrm{IC}_{\mu}(\mathrm{AND}, \varepsilon) \ge \mathrm{IC}_{\mu}(\mathrm{AND}, \varepsilon) - C\overline{h}(\varepsilon). \tag{6.57}$$

□

7 The set disjointness function with error

In this section we present the proofs of the results concerning the set disjointness function. It will be convenient to switch the roles of 0 and 1 in the range of the function, and redefine DISJ_n as

$$\mathrm{DISJ}_n(X,Y) = \vee_{i=1}^{n}(X_i \wedge Y_i)\,,$$

i. e., $\mathrm{DISJ}_n(X,Y) = 0$ if the inputs are disjoint and it is equal to 1 otherwise. Obviously, this will not affect the correctness of our results.

7.1 Proof of Theorem 3.11

Theorem 3.11 (restated). *For the set disjointness function* DISJ_n *on inputs of length* n*, we have*

$$\limsup_{n\to\infty} \frac{R_\varepsilon(\mathrm{DISJ}_n)}{n} \le \mathrm{IC}^0(\mathrm{AND},0) - C\,h(\varepsilon), \tag{7.1}$$

for some constant $C > 0$.

We will prove the following lemma, from which Theorem 3.11 follows using Corollary 3.8.

Lemma 7.1.

$$\limsup_{n\to\infty} \frac{R_\varepsilon(\mathrm{DISJ}_n)}{n} \le \mathrm{IC}^0(\mathrm{AND},\varepsilon,1\to 0). \tag{7.2}$$

Intuitively, an upper bound like Lemma 7.1 is essentially a compression result. Besides, as DISJ_n has a self-reducible structure (see [6]), one can make use of this fact together with the Braverman–Rao [7] compression. A difficulty is that what we want to solve is $[\mathrm{DISJ}_n,\varepsilon]$, that is, the error allowed is non-distributional, while the error unavoidably introduced in the compression phase is distributional. Fortunately, this can be salvaged by a minimax argument introduced in Section 6.2 of [4].

In order to use self-reducibility and compression, one first needs to have a control on the information cost of solving $[\mathrm{DISJ}_n,\varepsilon]$.

Lemma 7.2.

$$\limsup_{n\to\infty} \frac{\mathrm{IC}(\mathrm{DISJ}_n,\varepsilon,1\to 0)}{n} \le \mathrm{IC}^0(\mathrm{AND},\varepsilon,1\to 0), \tag{7.3}$$

where $\mathrm{IC}(\mathrm{DISJ}_n,\varepsilon,1\to 0) := \max_\mu \mathrm{IC}_\mu(\mathrm{DISJ}_n,\varepsilon,1\to 0)$.

The proof is a direct adaptation of the proof of Lemma 8.5 in [5].

Proof. Let Ω_0 denote the set of all measures μ on $\{0,1\}^2$ with $\mu(1,1) = 0$. Let π be a protocol that computes $[\mathrm{AND},\varepsilon,1\to 0]$ and satisfies

$$\max_{\mu\in\Omega_0} \mathrm{IC}_\mu(\pi) \le \mathrm{IC}^0(\mathrm{AND},\varepsilon,1\to 0) + \delta$$

for some small $\delta > 0$. Consider the following protocol τ that computes DISJ_n with error.

- Alice and Bob exchange (with replacement using public randomness) $n^{2/3}$ random coordinates. Denote this set of random coordinates by J. If for some $j \in J$, $x_j = 1$ and $y_j = 1$, then they output 1 and terminate.
- For each coordinate outside J, Alice and Bob run the protocol π and output 1 if π outputs 1 on some coordinate. Otherwise they output 0.

The intuition of the protocol is as follows: fix some μ. If with high probability for many of the coordinates i, $x_i = y_i = 1$, then with high probability only the first step of the protocol is performed and

$$\mathrm{IC}_\mu(\pi) \lessapprox n^{2/3}.$$

Otherwise, with high probability, for most coordinates i, $(x_i, y_i) \neq (1,1)$. Hence, for most coordinates i, the information complexity of evaluating $\pi(x_i, y_i)$ is not much higher than its information complexity on some distribution which gives zero mass to $x = y = 1$, which is not much higher than $\mathrm{IC}^0(\mathrm{AND}, \varepsilon, 1 \to 0)$.

As π has one-sided $1 \to 0$ error, obviously τ has only one-sided $1 \to 0$ error too, and this error happens with probability at most $\varepsilon^d \leq \varepsilon$, where d is the number of coordinates outside J which satisfy $x_j = y_j = 1$ (if $x_j = y_j = 1$ for some coordinate in J, there is no error). In particular, τ computes $[\mathrm{DISJ}_n, \varepsilon, 1 \to 0]$.

A direct inspection shows that the remaining proof of Lemma 8.5 in [5] depends only on the protocol but not on the specific problem, hence the proof works for our problem too, and the lemma can be proved similarly. □

Next we prove an amortized upper bound for DISJ_n.

Lemma 7.3. *For every $\varepsilon, \delta > 0$, there exists a constant $C > 0$ that depends on n, ε, δ, such that as long as $N \geq C(n, \varepsilon, \delta)$, we have*

$$\frac{R_\varepsilon(\mathrm{DISJ}_{nN})}{N} \leq (1+\delta)\,\mathrm{IC}(\mathrm{DISJ}_n, \varepsilon, 1 \to 0). \tag{7.4}$$

We will use the following claim [4, Claim 6.11]. Note that we state it slightly different from its original form in order to adapt to our purpose.

We introduce some notation that will be used in the statement of the following claim and the proof of Lemma 7.3: let π be a protocol getting its inputs from $\mathcal{X} \times \mathcal{Y}$. For every nonzero integer $k \in \mathbb{N}$, let π^k denote the protocol that gets inputs from $\mathcal{X}^k \times \mathcal{Y}^k$ and executes π on each pair (x_i, y_i) $(i = 1, 2, \ldots, k)$ independently. Define $\mathrm{IC}(\pi) := \max_\mu \mathrm{IC}_\mu(\pi)$ where the maximum is over all distributions $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$.

Claim 7.4 ([4]). *Let π be a protocol getting its inputs from $\mathcal{X} \times \mathcal{Y}$. For any numbers $\delta_1, \delta_2 > 0$, there exists $M \in \mathbb{N}$ such that for any integer $m \geq M$, there exists a protocol π_{m^3} that gets inputs from $\mathcal{X}^{m^3} \times \mathcal{Y}^{m^3}$ and:*

- *For each input in $\mathcal{X}^{m^3} \times \mathcal{Y}^{m^3}$, the total variation distance between the output of π_{m^3} and the output of π^{m^3} is at most δ_1.*

- *The worst-case communication complexity of* π_{m^3} *is at most* $m^3\,\mathrm{IC}(\pi)(1+\delta_2)$.

Proof of Lemma 7.3. Denote $I := \mathrm{IC}(\mathrm{DISJ}_n, \varepsilon, 1 \to 0)$. By the continuity of information complexity with respect to ε (Lemma 2.4, which holds for one-sided error with the same proof), there exists some $\xi > 0$ such that

$$\mathrm{IC}(\mathrm{DISJ}_n, \varepsilon - \xi, 1 \to 0) \le I(1+\delta)^{1/4}.$$

Let π be a protocol for the task $[\mathrm{DISJ}_n, \varepsilon - \xi, 1 \to 0]$, such that $\mathrm{IC}(\pi) \le I(1+\delta)^{1/2}$. Applying Claim 7.4 to π with $\delta_1 = \xi$ and $\delta_2 = (1+\delta)^{1/4} - 1$, we obtain a number M such that for any $m \ge M$ there is a protocol π_{m^3} with communication complexity at most

$$m^3\,\mathrm{IC}(\pi)(1+\delta)^{1/4} \le m^3 I(1+\delta)^{3/4}$$

and for each input (x,y) the total variation distance between the outputs of π_{m^3} and π^{m^3} is at most ξ.

Let τ_{nm^3} be the protocol for DISJ_{nm^3} where the parties split their input into m^3 blocks of n bits, simulate π_{m^3} and output "disjoint" if π_{m^3} outputs "disjoint" for all m^3 sub-problems. Similarly, define τ^{nm^3} with respect to π^{m^3}. Note that the communication complexity of τ_{nm^3} is the same as the communication complexity of π_{m^3}, which is bounded by $m^3 I(1+\delta)^{3/4}$.

We will show that τ_{nm^3} has a worst-case error of at most ε: for any input $(x,y) \in \{0,1\}^{nm^3}$, let $\mathrm{T}_{nm^3}(x,y)$ be the random variable denoting the transcript of τ_{nm^3} conditioned on the input being (x,y) and similarly define $\Pi_{m^3}(x,y)$ and $\Pi^{m^3}(x,y)$ with respect to π_{m^3} and π^{m^3}, respectively. Let $\varphi\colon \{0,1\}^{m^3} \to \{0,1\}$ be the function such that τ_{nm^3} outputs $\varphi(v)$ whenever the simulated π_{m^3} outputs v. The probability of τ_{nm^3} to err on input (x,y) equals

$$\begin{aligned}
&\Pr[\mathrm{T}_{nm^3}(x,y) \ne \mathrm{DISJ}_{nm^3}(x,y)] && (7.5)\\
&= \Pr\left[\Pi_{m^3}(x,y) \in \left\{v \in \{0,1\}^{m^3} : \varphi(v) \ne \mathrm{DISJ}_{nm^3}(x,y)\right\}\right] && (7.6)\\
&\le \Pr\left[\Pi^{m^3}(x,y) \in \left\{v \in \{0,1\}^{m^3} : \varphi(v) \ne \mathrm{DISJ}_{nm^3}(x,y)\right\}\right] + \xi && (7.7)\\
&\le \varepsilon && (7.8)
\end{aligned}$$

where (7.7) follows from the fact that given any (x,y), the transcripts of π_{m^3} and π^{m^3} have a total variation distance of at most ξ, and (7.8) follows from the fact that π^{m^3} simulates m^3 copies of DISJ_n, each with a one-sided error of $\varepsilon - \xi$. This implies that for all $m \ge M$,

$$\frac{R_\varepsilon(\mathrm{DISJ}_{nm^3})}{m^3} \le I(1+\delta)^{3/4}. \tag{7.9}$$

To complete the proof, by

$$\lim_{N\to\infty} \frac{(N^{1/3}+1)^3}{N} = 1,$$

there exists $C' > 0$ such that for all $N \ge C'$ one has

$$\frac{(N^{1/3}+1)^3}{N} \le (1+\delta)^{1/4}.$$

Set $C(n,\varepsilon,\delta) := \max(M^3, C')$. Then for any $N \ge C(n,\varepsilon,\delta)$,

$$\frac{R_\varepsilon(\mathrm{DISJ}_{nN})}{N} \le \frac{R_\varepsilon\left(\mathrm{DISJ}_{n\lceil N^{1/3}\rceil^3}\right)}{\lceil N^{1/3}\rceil^3} \cdot \frac{\lceil N^{1/3}\rceil^3}{N} \le I(1+\delta)^{3/4}\frac{\lceil N^{1/3}\rceil^3}{N} \le I(1+\delta)^{3/4}\frac{(N^{1/3}+1)^3}{N} \le I(1+\delta) \tag{7.10}$$

as desired. □

Now we prove the upper bound.

Proof of Lemma 7.1. Fix $\varepsilon > 0$. By Lemma 7.2, there exists $T(\varepsilon)$ depending on ε such that

$$\mathrm{IC}(\mathrm{DISJ}_n, \varepsilon, 1 \to 0) \le n\,\mathrm{IC}^0(\mathrm{AND}, \varepsilon, 1 \to 0) + o(n) \tag{7.11}$$

whenever $n \ge T(\varepsilon)$. For every such sufficiently large n, choose $\delta = 1/n$. Lemma 7.3 states that

$$\frac{R_\varepsilon(\mathrm{DISJ}_{nN})}{N} \le \left(1 + \frac{1}{n}\right)\mathrm{IC}(\mathrm{DISJ}_n, \varepsilon, 1 \to 0) \tag{7.12}$$

whenever $N \ge C(n,\varepsilon)$ for some constant $C(n,\varepsilon)$. Since $\mathrm{IC}(\mathrm{DISJ}_n, \varepsilon, 1 \to 0) \le n$,

$$\frac{R_\varepsilon(\mathrm{DISJ}_{nN})}{nN} \le \mathrm{IC}^0(\mathrm{AND}, \varepsilon, 1 \to 0) + \frac{1}{n} + o(1) \tag{7.13}$$

for $N \ge C(n,\varepsilon)$. It follows that

$$\frac{R_\varepsilon(\mathrm{DISJ}_M)}{M} \le \mathrm{IC}^0(\mathrm{AND}, \varepsilon, 1 \to 0) + o(1) \tag{7.14}$$

where $o(1) \to 0$ as $M \to \infty$, completing the proof. □

7.2 A protocol for Set-Disjointness

Theorem 3.13 (restated). *For the set disjointness function* DISJ_n *on inputs of length* n*, we have*

$$\mathrm{IC}^D(\mathrm{DISJ}_n, \varepsilon) \le n\left(\mathrm{IC}^0(\mathrm{AND}, 0) - C_1\sqrt{h(\varepsilon)}\right) + C_2 \log n, \tag{7.15}$$

for constants $C_1, C_2 > 0$.

Proof. We already established the lower bound in (3.25), it remains to prove the upper bound.

Let μ be an input distribution for DISJ_n, and let $p = \Pr_\mu[\mathrm{DISJ}_n(X,Y) = 1]$. We can assume that $p \ge \varepsilon$ as otherwise $\mathrm{IC}_\mu(\mathrm{DISJ}_n, \mu, \varepsilon) = 0$, and the upper bound trivially holds. Below we introduce a protocol π in Figure 2 that solves $[\mathrm{DISJ}_n, \mu, \varepsilon]$ and has the desired information cost. In fact, our protocol is stronger in the sense that it has only one-sided error: the protocol π always outputs 0 correctly if the correct output is 0, and on the other hand, if there are $t \ge 1$ coordinates satisfying $X_i = Y_i = 1$, then π will erroneously output 0 with probability at most $(\varepsilon/2p)^t \le \varepsilon/2p$. Thus the distributional error of π is at most $p \cdot \varepsilon/(2p) < \varepsilon$, and π indeed solves $[\mathrm{DISJ}_n, \mu, \varepsilon, 1 \to 0]$.

On input (X,Y):

- Alice and Bob, using public randomness, jointly sample a permutation σ on the set $\{1,2,\ldots,n\}$ uniformly at random; and they run the following sub-protocol π^σ:
- For $i=1,2,\ldots,n$ repeat:
  - Alice and Bob run a protocol π_i^σ that is (almost) optimal for $\mathrm{IC}_{\nu_i}(\mathrm{AND},\varepsilon/2p,1\to 0)$ on input $(X_{\sigma(i)},Y_{\sigma(i)})$, where ν_i is the distribution of $(X_{\sigma(i)},Y_{\sigma(i)})$ conditioned on the event that the protocol has not yet terminated;
  - if the protocol π_i^σ outputs 1, then terminate and output 1;
- If the "for-loop" ends without outputting 1, output 0 and terminate.

Figure 2: The protocol π that solves $[\mathrm{DISJ}_n,\mu,\varepsilon,1\to 0]$.

We now analyze the information cost. We start by analyzing the information cost of the sub-protocol π^σ. Let Π^σ be the transcript of π^σ, and write

$$\Pi^\sigma=\Pi_1^\sigma\ldots\Pi_n^\sigma$$

where Π_i^σ denotes the transcript of the protocol π_i^σ for $i=1,\ldots,n$. As usual let

$$\Pi_{<i}^\sigma=\Pi_1^\sigma\ldots\Pi_{i-1}^\sigma$$

be the partial transcript. Let μ_i denote the distribution of $X_{\sigma(i)}Y_{\sigma(i)}$, and ν_i denote the distribution of $X_{\sigma(i)}Y_{\sigma(i)}$ conditioned on $\Pi_{<i}^\sigma$. Corollary 3.9 (iii) gives a bound on the information exchanged in each round: there exist constants $C_1,C_2>0$ such that for any distribution ν,

$$\mathrm{IC}_\nu(\mathrm{AND},\varepsilon/2p,1\to 0)\le \mathrm{IC}^0(\mathrm{AND},0)+C_1\overline{h}(\nu(1,1))-C_2\overline{h}(\varepsilon/p). \tag{7.16}$$

Note that $(\Pi_i^\sigma\mid XY\Pi_{<i}^\sigma)$ has the same distribution as $(\Pi_i^\sigma\mid X_{\sigma(i)}Y_{\sigma(i)}\Pi_{<i}^\sigma)$, and thus

$$\begin{aligned}I(Y;\Pi^\sigma\mid X)&=\sum_{i=1}^n I(Y;\Pi_i^\sigma\mid X,\Pi_{<i}^\sigma)=\sum_{i=1}^n[H(\Pi_i^\sigma\mid X,\Pi_{<i}^\sigma)-H(\Pi_i^\sigma\mid XY,\Pi_{<i}^\sigma)]\\&\le\sum_{i=1}^n[H(\Pi_i^\sigma\mid X_{\sigma(i)},\Pi_{<i}^\sigma)-H(\Pi_i^\sigma\mid X_{\sigma(i)}Y_{\sigma(i)},\Pi_{<i}^\sigma)]\\&=\sum_{i=1}^n I(Y_{\sigma(i)};\Pi_i^\sigma\mid X_{\sigma(i)},\Pi_{<i}^\sigma).\end{aligned}$$

Thus, denoting by T^σ the number of AND protocols executed before the termination of π^σ, the above inequality implies (note that ν_i is a random variable, and π_i^σ depends on ν_i)

$$\begin{aligned}
\mathrm{IC}_\mu(\pi^\sigma) &\le \sum_{i=1}^n \mathbb{E}\,\mathrm{IC}_{\nu_i}(\pi_i^\sigma) \le \sum_{i=1}^n \Pr[T^\sigma \ge i]\,\mathbb{E}\,[\mathrm{IC}_{\nu_i}(\pi_i^\sigma) \;\mid\; T^\sigma \ge i] \\
&\le \sum_{i=1}^n \Pr[T^\sigma \ge i]\,\mathbb{E}\left[\mathrm{IC}^0(\mathrm{AND},0) + C_1\overline{h}(\nu_i(1,1)) - C_2\overline{h}(\varepsilon/p) \;\mid\; T^\sigma \ge i\right] \\
&\le \left(\mathrm{IC}^0(\mathrm{AND},0) - C_2\overline{h}(\varepsilon/p)\right)\mathbb{E}[T^\sigma] + C_1\sum_{i=1}^n \Pr[T^\sigma \ge i]\,\mathbb{E}\left[\overline{h}(\nu_i(1,1)) \mid T^\sigma \ge i\right].
\end{aligned}$$

We want to bound the second term. Note since $p \ge \varepsilon$,

$$\Pr[T^\sigma = i \mid T^\sigma \ge i, X_{\sigma(i)} = Y_{\sigma(i)} = 1] = \Pr[\pi_i^\sigma(X_{\sigma(i)}Y_{\sigma(i)}) = 1 \mid T^\sigma \ge i, X_{\sigma(i)} = Y_{\sigma(i)} = 1] \ge 1 - \frac{\varepsilon}{2p} \ge \frac{1}{2}. \tag{7.17}$$

Hence, applying (3.3) twice and using the concavity of $\overline{h}$, we get

$$\begin{aligned}
\Pr[T^\sigma \ge i]\,\mathbb{E}\left[\overline{h}(\nu_i(1,1)) \mid T^\sigma \ge i\right] &\le \Pr[T^\sigma \ge i]\,\overline{h}\left(\mathbb{E}\left[\nu_i(1,1) \mid T^\sigma \ge i\right]\right) \\
&= \Pr[T^\sigma \ge i]\,\overline{h}(\Pr[X_{\sigma(i)} = Y_{\sigma(i)} = 1 \mid T^\sigma \ge i]) \\
&\le \overline{h}(\Pr[X_{\sigma(i)} = Y_{\sigma(i)} = 1 \mid T^\sigma \ge i]\Pr[T^\sigma \ge i]) \\
&= \overline{h}(\Pr[T^\sigma \ge i, X_{\sigma(i)} = Y_{\sigma(i)} = 1]) \\
&\le 2\Pr[T^\sigma = i \mid T^\sigma \ge i, X_{\sigma(i)} = Y_{\sigma(i)} = 1] \\
&\qquad \cdot\, \overline{h}(\Pr[T^\sigma \ge i, X_{\sigma(i)} = Y_{\sigma(i)} = 1]) \\
&\le 2\overline{h}(\Pr[T^\sigma = i, X_{\sigma(i)} = Y_{\sigma(i)} = 1]) \\
&\le 2\overline{h}(\Pr[T^\sigma = i, \pi(X,Y) = 1]).
\end{aligned}$$

Using concavity of $\overline{h}$ again,

$$\frac{1}{n}\sum_{i=1}^n \overline{h}(\Pr[T^\sigma = i, \pi(X,Y) = 1]) \le \overline{h}(\Pr[\pi(X,Y) = 1]/n) = \overline{h}(p/n). \tag{7.18}$$

Therefore

$$\sum_{i=1}^n \Pr[T^\sigma \ge i]\,\mathbb{E}\left[\overline{h}(\nu_i(1,1)) \mid T^\sigma \ge i\right] \le 2n\overline{h}(p/n). \tag{7.19}$$

That is, we have shown

$$\mathrm{IC}_\mu(\pi^\sigma) \le \left(\mathrm{IC}^0(\mathrm{AND},0) - C_2\overline{h}(\varepsilon/p)\right)\mathbb{E}[T^\sigma] + 2C_1 n\overline{h}(p/n). \tag{7.20}$$

Taking the expectation with respect to σ, we obtain

$$\mathrm{IC}_\mu(\pi) = \mathop{\mathbb{E}}_{\sigma}\mathrm{IC}_\mu(\pi^\sigma) = \left(\mathrm{IC}^0(\mathrm{AND},0) - C_2\overline{h}(\varepsilon/p)\right)\mathop{\mathbb{E}}_{\sigma,XY}[T^\sigma] + 2C_1 n\overline{h}(p/n). \tag{7.21}$$

Hence it remains to bound $\mathbb{E}[T^\sigma]$ where the expectation is over σ and the input XY.

Let x,y be such that $\mathrm{DISJ}(x,y)=1$, and let j be an index such that $\mathrm{AND}(x_j,y_j)=1$. Then

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\sigma,XY}[T^\sigma \mid XY=xy] = \sum_{i=1}^{n}\Pr[\sigma(i)=j]\,\mathbb{E}[T^\sigma \mid XY=xy,\sigma(i)=j]\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[T^\sigma \mid XY=xy,\sigma(i)=j]\\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{b=0,1}\mathbb{E}[T^\sigma \mid XY=xy,\sigma(i)=j,\pi_i^\sigma(X,Y)=b]\Pr[\pi_i^\sigma(X,Y)=b \mid XY=xy,\sigma(i)=j]\\
&\le \frac{1}{n}\sum_{i=1}^{n}\Big(i\Pr[\pi_i^\sigma(X,Y)=1 \mid XY=xy,\sigma(i)=j]+n\Pr[\pi_i^\sigma(X,Y)=0 \mid XY=xy,\sigma(i)=j]\Big)\\
&\le \frac{1}{n}\sum_{i=1}^{n}\left(i(1-\frac{\varepsilon}{2p})+n\frac{\varepsilon}{2p}\right) = (1-\frac{\varepsilon}{2p})\frac{n+1}{2}+\frac{\varepsilon}{2p}n \le \frac{n+1}{2}+\frac{\varepsilon}{4p}n.
\end{aligned}$$

This allows us the next bound:

$$\mathop{\mathbb{E}}_{\sigma,XY}[T^\sigma] = \Pr[\mathrm{DISJ}(X,Y)=1]\,\mathbb{E}[T \mid \mathrm{DISJ}(X,Y)=1]+\Pr[\mathrm{DISJ}(X,Y)=0]\,\mathbb{E}[T \mid \mathrm{DISJ}(X,Y)=0] \tag{7.22}$$

$$\le p\left(\frac{n+1}{2}+\frac{\varepsilon}{4p}n\right)+(1-p)n \le \frac{2p}{3}n+\frac{\varepsilon}{4}n+(1-p)n = (1-p/3+\varepsilon/4)n. \tag{7.23}$$

Combine (7.21) and (7.23) we get

$$\begin{aligned}
\mathrm{IC}_\mu(\pi) &\le n(1-p/3+\varepsilon/4)\left(\mathrm{IC}^0(\mathrm{AND},0)-C_2\overline{h}(\varepsilon/p)\right)+C_1 2n\overline{h}(p/n)\\
&= n(\mathrm{IC}_0(\mathrm{AND},0)-C_3\overline{h}(\varepsilon/p)+p)+C_4 n\overline{h}(p/n),
\end{aligned}$$

for constants $C_3,C_4>0$.

It remains to optimize over p. We start by minimizing $p+\overline{h}(\varepsilon/p)$. Up to a constant multiple, the minimum is attained at the point satisfying $p=\overline{h}(\varepsilon/p)$. A simple calculation shows that $p\approx\sqrt{h(\varepsilon)}$, and so $p+\overline{h}(\varepsilon/p)=\Omega(\sqrt{h(\varepsilon)})$. Thus

$$\mathrm{IC}_\mu(\pi) \le n[\mathrm{IC}^0(\mathrm{AND},0)-\Omega(\sqrt{h(\varepsilon)})]+O(n\overline{h}(p/n)). \tag{7.24}$$

The value of the error term $O(n\overline{h}(p/n))$ is at most $O(n\overline{h}(1/n))=O(n\log(n)/n)=O(\log n)$, and the theorem follows. □

8 Open problems and concluding remarks

- In Conjecture 3.12 we speculated that the exact asymptotics of $R_\varepsilon(\mathrm{DISJ}_n)$ is given by the information complexity of the AND function when only one-sided error is allowed:

$$R_\varepsilon(\mathrm{DISJ}_n) = n\,\mathrm{IC}^0(\mathrm{AND},\varepsilon,1\to 0)\pm o(n). \tag{8.1}$$

The set disjointness function has a "self-reducible" structure in the sense that it is possible to solve an instance of the corresponding communication problem by dividing the input into blocks and solving the same problem on each block separately. This structure allows us to relate the communication complexity of the problem to its amortized communication complexity, and thus to its information complexity via the fundamental result of Braverman and Rao [7]. Applying such ideas we showed (the lower bound is obvious)

$$\mathrm{IC}(\mathrm{DISJ}_n,\varepsilon) \le R_\varepsilon(\mathrm{DISJ}_n) \le m\,\mathrm{IC}(\mathrm{DISJ}_{\frac{n}{m}},\varepsilon,1\to 0)+o(n), \tag{8.2}$$

for an appropriate choice of $m = m(n)$ that tends to infinity as $n \to \infty$. In Theorem 3.11 we combined this with our analysis of the information complexity of the set disjointness to prove $R_\varepsilon(\mathrm{DISJ}_n) = n[\mathrm{IC}^0(\mathrm{AND},0) - \Theta(h(\varepsilon))]$. More precisely we showed

$$n\,\mathrm{IC}^0(\mathrm{AND},\varepsilon) \le \mathrm{IC}(\mathrm{DISJ}_n,\varepsilon) \le \mathrm{IC}(\mathrm{DISJ}_n,\varepsilon,1\to 0) \le n\,\mathrm{IC}^0(\mathrm{AND},\varepsilon,1\to 0)+o(n), \tag{8.3}$$

and combined it with our results regarding the information complexity of the AND function. We believe that the upper bound is the truth; that is

$$\mathrm{IC}(\mathrm{DISJ}_n,\varepsilon) \ge n\,\mathrm{IC}^0(\mathrm{AND},\varepsilon,1\to 0)-o(n), \tag{8.4}$$

which would imply Conjecture 3.12.

- The example of the AND function shows that the $\Omega(h(\varepsilon))$ gain in the information cost, appearing in our upper bounds in Theorems 3.2, 3.6, 3.15 and 3.16 is tight. However we do not know whether the $O(h(\sqrt{\varepsilon}))$ gain appearing in the lower bounds in Theorems 3.5 and 3.6, Corollary 3.14 and Theorem 3.16 is sharp. In fact we are not aware of any example that exhibits a gain that is not $\Theta(h(\varepsilon))$. Is it true that for every function $f\colon \mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$, and measure μ on $\mathcal{X}\times\mathcal{Y}$ with $\mathrm{IC}_\mu(f,0) > 0$, we have $\mathrm{IC}_\mu(f,\varepsilon) = \mathrm{IC}_\mu(f,0) - \Theta(h(\varepsilon))$? One can ask a similar question for $\mathrm{IC}_\mu(f,\mu,\varepsilon)$, $\mathrm{IC}(f,\varepsilon)$, and $\mathrm{IC}^D(f,\varepsilon)$.

- Recall the *inner product function* $\mathrm{IP}_n\colon \{0,1\}^n\times\{0,1\}^n\to\{0,1\}$ is defined as

$$\mathrm{IP}_n\colon (x,y)\mapsto \sum_{i=1}^n x_i y_i \mod 2. \tag{8.5}$$

Let ν denote the uniform distribution on $\{0,1\}^n\times\{0,1\}^n$. In [6, Theorem 1.3], Braverman et al. exploited the self-reducibility property of IP_n and showed that for any $\delta > 0$ there exists $\varepsilon > 0$ such that

$$\liminf_{n\to\infty}\frac{\mathrm{IC}_\nu(\mathrm{IP}_n,\nu,\varepsilon)}{n} \ge 1-\delta. \tag{8.6}$$

This in particular implies

$$\lim_{n\to\infty}\frac{\mathrm{IC}_\nu(\mathrm{IP}_n,\nu,0)}{n} = 1.$$

In the same paper they also showed $\mathrm{IC}_\nu(\mathrm{IP}_n,\nu,\varepsilon) \le (1-2\varepsilon)(n+1)$ for any $0<\varepsilon<1/2$, alternatively,

$$\limsup_{n\to\infty}\frac{\mathrm{IC}_\nu(\mathrm{IP}_n,\nu,\varepsilon)}{n} \le 1-2\varepsilon. \tag{8.7}$$

In [6, Problem 1.4] they asked whether this upper bound is tight, possibly with the coefficient 2 be replaced by some other constant α? Formally, does there exist a constant $\alpha \geq 2$ such that

$$\liminf_{n\to\infty} \frac{\mathrm{IC}_\nu(\mathrm{IP}_n, \nu, \varepsilon)}{n} \geq 1 - \alpha\varepsilon - o(\varepsilon)\ ? \tag{8.8}$$

A positive answer would imply that, by allowing ε-error for IP_n, comparing to 0-error protocols one can save roughly $n\Theta(\varepsilon)$ information as $n \to \infty$. Note that our bound $\mathrm{IC}_\nu(f,\nu,\varepsilon) \leq \mathrm{IC}_\nu(f,\nu,0) - \Omega(h(\varepsilon))$ in Theorem 3.6 does not refute (8.8) as the constant in $\Omega(\cdot)$ equals to $o(n)$ when $f = \mathrm{IP}_n$.

- The focus of this paper has been on the internal information complexity since this notion has nice properties such as the direct sum property (see the introduction for references) comparing to external information complexity. However, it would still be interesting to study external information complexity. Considering that external information complexity is typically simpler than internal information complexity, we believe that the analogues of many of our results, specially those about the AND function, can be proven for this case as well. We defer this to future research.

References

[1] LÁSZLÓ BABAI, PETER FRANKL, AND JANOS SIMON: Complexity classes in communication complexity theory. In *Proc. 27th FOCS*, pp. 337–347. IEEE Comp. Soc. Press, 1986. [doi:10.1109/SFCS.1986.15] 4

[2] ZIV BAR-YOSSEF, T. S. JAYRAM, RAVI KUMAR, AND D. SIVAKUMAR: An information statistics approach to data stream and communication complexity. *J. Comput. System Sci.*, 68(4):702–732, 2004. Preliminary version in FOCS'02. [doi:10.1016/j.jcss.2003.11.006] 4

[3] BOAZ BARAK, MARK BRAVERMAN, XI CHEN, AND ANUP RAO: How to compress interactive communication. *SIAM J. Comput.*, 42(3):1327–1363, 2013. Preliminary version in STOC'10. [doi:10.1137/100811969] 4, 11

[4] MARK BRAVERMAN: Interactive information complexity. *SIAM J. Comput.*, 44(6):1698–1739, 2015. Preliminary version in STOC'12. [doi:10.1137/130938517] 4, 5, 6, 9, 13, 14, 22, 23, 24, 61, 62

[5] MARK BRAVERMAN, ANKIT GARG, DENIS PANKRATOV, AND OMRI WEINSTEIN: From information to exact communication (extended abstract). In *Proc. 45th STOC*, pp. 151–160. ACM Press, 2013. [doi:10.1145/2488608.2488628] 4, 5, 6, 7, 13, 18, 19, 20, 21, 22, 42, 49, 50, 53, 61, 62

[6] MARK BRAVERMAN, ANKIT GARG, DENIS PANKRATOV, AND OMRI WEINSTEIN: Information lower bounds via self-reducibility. *Theory Comput. Syst.*, 59(2):377–396, 2016. Preliminary version in CSR'13. [doi:10.1007/s00224-015-9655-z] 13, 42, 61, 68, 69

[7] MARK BRAVERMAN AND ANUP RAO: Information equals amortized communication. *IEEE Trans. Inform. Theory*, 60(10):6058–6069, 2014. Preliminary version in FOCS'11. [doi:10.1109/TIT.2014.2347282, arXiv:1106.3595] 4, 22, 61, 68

[8] MARK BRAVERMAN, ANUP RAO, OMRI WEINSTEIN, AND AMIR YEHUDAYOFF: Direct product via round-preserving compression. In *Proc. 40th Internat. Colloq. on Automata, Languages and Programming (ICALP'13)*, pp. 232–243. Springer, 2013. [doi:10.1007/978-3-642-39206-1_20] 4

[9] MARK BRAVERMAN, ANUP RAO, OMRI WEINSTEIN, AND AMIR YEHUDAYOFF: Direct products in communication complexity. In *Proc. 54th FOCS*, pp. 746–755. IEEE Comp. Soc. Press, 2013. [doi:10.1109/FOCS.2013.85] 4

[10] MARK BRAVERMAN AND JON SCHNEIDER: Information complexity is computable. In *Proc. 43rd Internat. Colloq. on Automata, Languages and Programming (ICALP'16)*. DROPS, 2016. [doi:10.4230/LIPIcs.ICALP.2016.87, arXiv:1502.02971] 15

[11] AMIT CHAKRABARTI, YAOYUN SHI, ANTHONY WIRTH, AND ANDREW YAO: Informational complexity and the direct sum problem for simultaneous message complexity. In *Proc. 42nd FOCS*, pp. 270–278. IEEE Comp. Soc. Press, 2001. [doi:10.1109/SFCS.2001.959901] 4

[12] ARKADEV CHATTOPADHYAY AND TONIANN PITASSI: The story of set disjointness. *SIGACT News*, 41(3):59–85, 2010. [doi:10.1145/1855118.1855133] 4

[13] YUVAL DAGAN, YUVAL FILMUS, HAMED HATAMI, AND YAQIAO LI: Trading information complexity for error. In *Proc. 32nd IEEE Conf. on Computational Complexity (CCC'17)*, pp. 16:1–16:59. DROPS, 2017. [doi:10.4230/LIPIcs.CCC.2017.16, arXiv:1611.06650] 1

[14] TOMÁS FEDER, EYAL KUSHILEVITZ, MONI NAOR, AND NOAM NISAN: Amortized communication complexity. *SIAM J. Comput.*, 24(4):736–750, 1995. Preliminary version in CCC'17. [doi:10.1137/S0097539792235864] 4

[15] YUVAL FILMUS, HAMED HATAMI, YAQIAO LI, AND SUZIN YOU: Information complexity of the AND function in the two-party, and multiparty settings. In *Proc. 23rd Ann. Internat. Computing and Combinatorics Conf. (COCOON'17)*, pp. 200–211. Springer, 2017. [doi:10.1007/978-3-319-62389-4_17, arXiv:1703.07833] 18

[16] 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. Preliminary version in STOC'15. [doi:10.1145/2907939] 4

[17] PRAHLADH HARSHA, RAHUL JAIN, DAVID A. MCALLESTER, AND JAIKUMAR RADHAKRISHNAN: The communication complexity of correlation. *IEEE Trans. Inform. Theory*, 56(1):438–449, 2010. [doi:10.1109/TIT.2009.2034824] 4

[18] RAHUL JAIN: New strong direct product results in communication complexity. *J. ACM*, 62(3):20:1–20:27, 2015. Preliminary version in ECCC'11. [doi:10.1145/2699432] 4

[19] RAHUL JAIN, ATTILA PERESZLÉNYI, AND PENGHUI YAO: A direct product theorem for bounded-round public-coin randomized communication complexity. *Algorithmica*, 76(3):720–748, 2016. Preliminary version in FOCS'12. [doi:10.1007/s00453-015-0100-0, arXiv:1201.1666] 4

[20] RAHUL JAIN, JAIKUMAR RADHAKRISHNAN, AND PRANAB SEN: A direct sum theorem in communication complexity via message compression. In *Proc. 30th Internat. Colloq. on Automata, Languages and Programming (ICALP'03)*, pp. 300–315. Springer, 2003. [doi:10.1007/3-540-45061-0_26, arXiv:cs/0304020] 4

[21] BALA KALYANASUNDARAM AND GEORG SCHNITGER: The probabilistic communication complexity of set intersection. *SIAM J. Discrete Math.*, 5(4):545–557, 1992. Preliminary version in SCT'87. [doi:10.1137/0405044] 4

[22] HARTMUT KLAUCK: A strong direct product theorem for disjointness. In *Proc. 42nd STOC*, pp. 77–86. ACM Press, 2010. [doi:10.1145/1806689.1806702, arXiv:0908.2940] 4

[23] EYAL KUSHILEVITZ AND NOAM NISAN: *Communication complexity*. Cambridge University Press, 1997. 12

[24] NAN MA AND PRAKASH ISHWAR: Some results on distributed source coding for interactive function computation. *IEEE Trans. Inform. Theory*, 57(9):6180–6195, 2011. [doi:10.1109/TIT.2011.2161916] 19

[25] NAN MA AND PRAKASH ISHWAR: The infinite-message limit of two-terminal interactive source coding. *IEEE Trans. Inform. Theory*, 59(7):4071–4094, 2013. Preliminary version in Allerton'09. [doi:10.1109/TIT.2013.2251412, arXiv:0908.3512] 19

[26] MARCO MOLINARO, DAVID P. WOODRUFF, AND GRIGORY YAROSLAVTSEV: Beating the direct sum theorem in communication complexity with implications for sketching. In *Proc. 24th Ann. ACM-SIAM Symp. on Discrete Algorithms (SODA'13)*, pp. 1738–1756. ACM Press, 2013. [doi:10.1137/1.9781611973105.125] 9

[27] ALEXANDER A. RAZBOROV: On the distributional complexity of disjointness. *Theoret. Comput. Sci.*, 106(2):385–390, 1992. Preliminary version in ICALP'90. [doi:10.1016/0304-3975(92)90260-M] 4

[28] BYRON SCHMULAND: On the compacity of the space of probability measures. Mathematics Stack Exchange, January 18, 2014. 38

[29] CLAUDE E. SHANNON: A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. [doi:10.1002/j.1538-7305.1948.tb01338.x] 4

[30] ALEXANDER A. SHERSTOV: Communication complexity theory: Thirty-five years of set disjointness. In *Proc. 39th Internat. Symp. on Mathematical Foundations of Computer Science (MFCS'14)*, pp. 24–43. Springer, 2014. [doi:10.1007/978-3-662-44522-8_3] 4

[31] ANDREW CHI-CHIH YAO: Some complexity questions related to distributive computing (Preliminary report). In *Proc. 11th STOC*, pp. 209–213. ACM Press, 1979. [doi:10.1145/800135.804414] 4, 10

AUTHORS

Yuval Dagan
M. Sc. student
Technion — Israel Institute of Technology
Haifa, Israel
yuval.dagan@cs.technion.ac.il

Yuval Filmus
Assistant professor
Technion — Israel Institute of Technology
Haifa, Israel
yuvalfi@cs.technion.ac.il
http://www.cs.toronto.edu/~yuvalf/

Hamed Hatami
Associate professor
McGill University
Montreal, QC, Canada
hatami@cs.mcgill.ca
http://www.cs.mcgill.ca/~hatami/

Yaqiao Li
Ph. D. student
McGill University
Montreal, QC, Canada
yaqiao.li@mail.mcgill.ca
http://www.cs.mcgill.ca/~yli252/

ABOUT THE AUTHORS

YUVAL DAGAN studied at the Technion between 2011-2017; his advisor was Yuval Filmus. His thesis was in theoretical computer science, focusing on information complexity and learning theory. He lives in Tel Aviv, Israel.

YUVAL FILMUS graduated from the University of Toronto in 2013; his advisor was Toni Pitassi. His thesis focused on spectral aspects of extremal combinatorics, but he has also maintained a keen interest in computational complexity. In his free time, he enjoys helping people on Computer Science Stack Exchange.

HAMED HATAMI graduated from the University of Toronto in 2009; his advisors were Michael Molloy and Balazs Szegedy. His work is focused on the applications of mathematical analysis to theoretical computer science and discrete mathematics.

YAQIAO LI is a Ph. D. student at McGill university, supervised by Hamed Hatami. His current research focuses on communication and information complexity. He grew up in Sichuan where spicy food rules. He has a lot of fun playing badminton and fishing.