Electronic Colloquium on Computational Complexity, Report No. 123 (2011)

Interactive information complexity

Mark Braverman*

September 14, 2011

Abstract

The primary goal of this paper is to define and study the interactive information complexity of functions. Let $f(x, y)$ be a function, and suppose Alice is given x and Bob is given y. Informally, the interactive information complexity $\mathsf{IC}(f)$ of f is the least amount of information Alice and Bob need to reveal to each other to compute f. Previously, information complexity has been defined with respect to a prior distribution on the input pairs (x, y). Our first goal is to give a definition that is independent of the prior distribution. We show that several possible definitions are essentially equivalent.

We establish some basic properties of the interactive information complexity $\mathsf{IC}(f)$. In particular, we show that $\mathsf{IC}(f)$ is equal to the amortized (randomized) communication complexity of f. We also show a direct sum theorem for $\mathsf{IC}(f)$ and give the first general connection between information complexity and (non-amortized) communication complexity. We explore the information complexity of two specific problems – Equality and Disjointness. We conclude with a list of open problems and research directions.

*Princeton University and the University of Toronto, mbraverm@cs.princeton.edu. Partially supported by an NSERC Discovery Grant.

ISSN 1433-8092

1 Introduction

1.1 Information theory for one-way communication

Shannon [Sha48] originally developed his information theory in order to understand the *one way* data transmission problem over a channel. For noiseless channel, an important early result in information theory is that the cost of transmitting a (random) message X over a channel is closely related to the entropy $H(X)$ of the message. Recall that $H(X)$ is defined as $H(X) := \mathbf{E}_{x\sim X}[-\log_2 P_X(x)]$. Shannon's noiseless coding theorem, also known as Shannon's source coding theorem states that the cost of sending a signal distributed according to X is essentially $H(X)$:

Theorem 1.1 (Shannons' noiseless coding theorem)**.** *Let $C(X)$ denote the number of bits that the optimal lossless scheme for sending a sample distributed according to X needs to send across the channel. Then for all X, $H(X) \leq C(X) < H(X) + 1$.*

As a corollary, if we demote by $C_n := C(X^n)$ – the cost of sending n independent samples distributed according to X, then we see that

$$\lim_{n\to\infty} \frac{C_n}{n} = H(X). \tag{1}$$

Thus the entropy approximately measures the cost of sending a single message, and exactly measures the amortized cost of sending many independent messages from the distribution. The noiseless coding theorem connects the communication cost of sending a message from one player to another to the information content of the message.

An example of a more subtle one-way transmission scenario where the information-theoretic bound is asymptotically tight is given by the Slepian-Wolf Theorem [SW73]. We give one of the interpretations of the result here. Consider the following task: Alice is given an input A which she needs to transmit to Bob. Bob has an input B which is correlated with A, thus giving Bob partial information about A. Clearly this task is not harder than the task of transmitting A on its own, thus $H(A)$ is an upper bound on the amortized transmission cost of this task (i.e. the cost of this task when it is repeated many times). Can the players do better? For example, in the extreme case where A is a deterministic function $A = F(B)$, clearly the number of bits that need to be sent is 0. Turns out that information theory again gives the precise answer to this question:

Theorem 1.2 (Slepian-Wolf coding theorem [SW73])**.** *The amortized communication cost of transmitting the message A to a player that has prior information B is given by the conditional entropy $H(A|B) := H(AB) - H(B)$.*

Remark 1.3. Unlike Shannon's Theorem, Theorem 1.2 only gives the upper bound in the amortized sense – when one needs to solve multiple copies of the transmission problem, and the number of copies goes to infinity. In [BR10] a one-shot version of the theorem was proved – i.e. the transmission problem is solved in $\approx H(A|B) + o(H(A|B))$ bits of transmission – but the protocol solving it is interactive, rather than a transmission protocol.

These two examples – Shannon's Noiseless Coding Theorem and the Slepian-Wolf Theorem – demonstrate the tight connection between the communication cost of *one-way* transmission problems and their inherent information costs. Information-theoretic quantities are the "right" tools to study the transmission complexity of these problems. The goal of this paper is to develop

the "right" information-theoretic notions for *two-way* interactive communication problems. Most importantly, we develop and study the *information complexity* of two-party problems.

Communication complexity studies the communication cost of functions. In the two player setting, the main question it tries to answer is: "How many bits do Alice and Bob need to transmit to each other in order to solve a given problem." Informally, information complexity tries to answer the question: "How much information do Alice and Bob need to reveal to each other in order to solve a give problem?". While – as discussed below – information theory has been used as a tool in the study of communication complexity, it is only recently that the information complexity of interactive problems has been considered in its full generality. The main goal of this paper is to work out the definition of the (two-party) interactive information complexity, and to initiate the study of its properties.

1.2 Information tools in communication complexity

Over the past decade, information theory has been an important tool in the study of communication complexity lower bounds, and their applications to lower bounds for data structures. Chakrabarti, Shi, Wirth and Yao [CSWY01] were the first to define the external information cost. In [CSWY01] information theory is used to prove a direct sum theorem for problems with one simultaneous round of communication. Direct sum theorems (and the related direct product theorems) are theorems giving lower bounds for the complexity of n copies of a certain problem in terms of the complexity of one copy. In the context of communication complexity, direct sum theorems in various contexts have been the focus of much work [FKNN95, CSWY01, Sha03, JRS03, HJMR07, BBCR10, Kla10, Jai11].

Information theoretic tools have also been useful in contexts where the problem is not merely a repetition of multiple copies but some other aggregation function. A notable example is set disjointness $DISJ_n$, where the players are given two subsets of $\{1,\ldots,n\}$ and need to determine whether the sets are disjoint or not. If we represent the sets by their indicator strings $(x_1,\ldots,x_n)$ and $(y_1,\ldots,y_n)$, then the disjointness function can be written as

$$DISJ_n((x_1,\ldots,x_n),(y_1,\ldots,y_n)) := \neg \bigvee_{i=1}^{n} (x_i \wedge y_i). \tag{2}$$

Here the basic function is the conjunction in $x_i \wedge y_i$, and n copies of the conjunction are combined by the big disjunction over all i. A linear communication lower bound for the problem is known [KS92, Raz92]. While $DISJ_n$ is not a direct sum over the $(x_i \wedge y_i)$'s, similar analysis can be adapted, and information-theoretic techniques yield a linear lower bound on this problem [BYJKS04] – we will see an extension of these techniques to information complexity in Section 7. Razborov's simplified proof can also be cast in this framework. The information-theoretic techniques have recently been applied to more complicated composition functions rather than just the AND function [JKR09, LS10, JKZ10]. Information-theoretic reasoning has also been successfully applied in the simpler deterministic setting [DW07].

The information cost for protocols over distributions of inputs was defined implicitly in [BYJKS04] and explicitly in [BBCR10]. It is the tool that allows one to obtain the only known direct sum results for the general randomized communication complexity. If F is a function and $R(F)$ is its randomized communication complexity, it has been shown in [BBCR10] that the randomized communication complexity of n copies of F satisfies $R(F^n) = \tilde{\Omega}(\sqrt{n} \cdot R(F))$. In the follow-up work [BR10] a tight relationship between the amortized distributional communication complexity of a

function and its internal information cost has been established. We further extend this relationship to the non-distributional randomized setting in Section 5.2.

1.3 Main contributions

Interactive information complexity

In [BYJKS04, BBCR10], the information cost $\mathsf{IC}_\mu(\pi)$ of executing a protocol π over an *a-priori* distribution μ of inputs is defined. This immediately yields a definition of the information cost of a function F with respect to a distribution μ: we simply look for a protocol π for F that minimizes $\mathsf{IC}_\mu(\pi)$. It turns out that there is a tight connection between the distributional information cost of F over μ and the amortized *distributional* communication complexity of F^n over μ^n [BR10].

Our first goal is to obtain a definition of the information cost of a function that does not depend on the *a-priori* distribution μ. Let f be a two-party function and $\rho < 1/3$ be an error parameter. Consider the following three quantities:

1. $I_1 = I_1(f, \rho)$ is such that for each $I > I_1$ there is a protocol π_I that on each input computes f except with error $\leq \rho$, and for each prior distribution μ reveals at most I bits of information about the inputs *to the players*. We later denote this quantity by $\mathsf{IC}(f, \rho)$.

2. $I_2 = I_2(f, \rho)$ is such that for each $I > I_2$ and for each prior distribution μ on inputs there is a protocol $\pi_{\mu,I}$ that computes f correctly except with probability $\leq \rho$ with respect to the distribution μ, and that reveals at most I bits of information about the inputs to the players. We later denote this quantity by $\mathsf{IC}_{\mathsf{D}}(f, \rho)$.

3. $I_3 = I_3(f, \rho)$ is the amortized randomized communication complexity of f as the number of copies goes to ∞: $I_3 = \lim_{n\to\infty} \frac{R^n_\rho(f^n)}{n}$. Here $R^n_\rho(f^n)$ denotes the communication complexity of computing n copies of f, where the protocol is allowed to err at most a ρ-fraction of the time on each input.

Note that the difference between I_1 and I_2 is in the order of quantifiers, and clearly $I_2 \leq I_1$. We prove that $I_1 = \Theta(I_2)$ (Theorem 3.5), more precisely, we show that

$$I_2(f, \rho) \leq I_1(f, \rho) \leq 2 \cdot I_2(f, \rho/2).$$

In the special case when $\rho = 0$ we show that $I_1(f, 0) = I_2(f, 0)$ (Theorem 3.6). We then establish the equality $I_3(f, \rho) = I_1(f, \rho)$ (Theorem 6.7), showing that the information complexity in the non-distributional case is equal to the amortized communication complexity. These equivalences establish I_1 as the "right" notion of the interactive information complexity of f, which we denote by $\mathsf{IC}(f, \rho) := I_1$.

Properties of the interactive information complexity

We establish some properties of the interactive information complexity. The first one is that the interactive information complexity is *additive* (Theorem 4.2). The theorem holds for tasks. A task can be, for example, computing a function f with an error bounded by ρ. We show that if T_1 and T_2 are two tasks, and $T := T_1 \times T_2$ is the task comprised of performing one of copy of each of the two tasks, then the information complexity satisfies $\mathsf{IC}(T) = \mathsf{IC}(T_1) + \mathsf{IC}(T_2)$.

Next, we establish a lower bound for the information complexity of any problem in terms of the communication complexity of the problem. The bound is given by Theorem 5.3. For a constant ρ the bound is of the form $\mathsf{IC}(f) = \Omega(\log R(f))$. This is a fairly weak bound, as one may expect $\mathsf{IC}(f) = \Theta(R(f))$ to hold in many cases. Nonetheless, this is the first general bound on the information revealed by *any* protocol for a problem in terms of its communication complexity.

The interactive information complexity of specific functions

We consider two specific functions that have been studied extensively in the literature: the equality function $EQ(x, y) = \mathbf{1}_{x=y}$ and the disjointness function $DISJ_n$ that we discussed earlier. It turns out that the information complexity or EQ is constant, even when one does not allow the protocol to err at all: $\mathsf{IC}(EQ, 0) = O(1)$ (Proposition 3.21). This result is in sharp contrast with the fact that the zero-error randomized communication complexity of the equality function over n-bit strings is $\Omega(n)$.

On the other hand, the information complexity of $DISJ_n$ turns out to be linear (Theorem 7.2):

$$\mathsf{IC}(DISJ_n, 1/2 - \varepsilon) = \Omega(n) \text{ for all } 0 < \varepsilon < 1/2.$$

We give two proofs for this fact. One is through a reduction to the result about the communication complexity of disjointness, while the second one is a more direct information-theoretic proof.

Directions and open problems

Finally, in Section 8 we outline several research directions and many open problems surrounding the interactive information complexity.

Acknowledgments

I would like to thank Boaz Barak, Stephen Cook, Denis Pankratov, Toni Pitassi, Anup Rao, Alexander Razborov, Michael Saks and Avi Wigderson for the insightful conversations and comments on earlier drafts of this paper.

2 Preliminaries

In an effort to make the paper as self-contained as possible, we provide some background on information theory and communication complexity here. Additional details may be found in [BR10]. A more thorough treatment of the subject may be found in the textbooks on the respective subjects [CT91, KN97]. We note that throughout the paper our protocols will make use of both public and private randomness.

Notation. We reserve capital letters for random variables and distributions, calligraphic letters for sets, and small letters for elements of sets. Throughout this paper, we often use the notation $|b$ to denote conditioning on the event $B = b$. Thus $A|b$ is shorthand for $A|B = b$.

We use the standard notion of *statistical/total variation* distance between two distributions.

Definition 2.1. Let D and F be two random variables taking values in a set $\mathcal{S}$. Their *statistical distance* is

$$|D - F| \stackrel{def}{=} \max_{\mathcal{T} \subseteq \mathcal{S}} (|\Pr[D \in \mathcal{T}] - \Pr[F \in \mathcal{T}]|) = \frac{1}{2} \sum_{s \in \mathcal{S}} |\Pr[D = s] - \Pr[F = s]|$$

If $|D - F| \leq \varepsilon$ we shall say that D is ε-*close* to F. We shall also use the notation $D \stackrel{\varepsilon}{\approx} F$ to mean D is ε-close to F.

2.1 Information Theory

Definition 2.2 (Entropy)**.** The *entropy* of a random variable X is

$$H(X) \stackrel{def}{=} \sum_x \Pr[X = x] \log(1/\Pr[X = x]).$$

The *conditional entropy* $H(X|Y)$ is defined to be $\mathbf{E}_{y \in_R Y} [H(X|Y = y)]$.

Fact 2.3. $H(AB) = H(A) + H(B|A)$.

Definition 2.4 (Mutual Information)**.** The *mutual information* between two random variables A, B, denoted $I(A; B)$ is defined to be the quantity

$$H(A) - H(A|B) = H(B) - H(B|A).$$

The *conditional mutual information* $I(A; B|C)$ is $H(A|C) - H(A|BC)$.

In analogy with the fact that $H(AB) = H(A) + H(B|A)$,

Proposition 2.5 (Chain Rule)**.** *Let C_1, C_2, D, B be random variables. Then*

$$I(C_1C_2; B|D) = I(C_1; B|D) + I(C_2; B|C_1D).$$

We also use the notion of *divergence* (also known as Kullback-Leibler distance or relative entropy), which is a different way to measure the distance between two distributions:

Definition 2.6 (Divergence)**.** The informational divergence between two distributions is

$$\mathbf{D}(A||B) \stackrel{def}{=} \sum_x A(x) \log(A(x)/B(x)).$$

For example, if B is the uniform distribution on $\{0,1\}^n$ then $\mathbf{D}(A||B) = n - H(A)$.

Proposition 2.7. *Let A, B, C be random variables in the same probability space. For every a in the support of A and c in the support of C, let B_a denote $B|A = a$ and B_{ac} denote $B|A = a, C = c$. Then $I(A; B|C) = \mathbf{E}_{a,c \in_R A,C} [\mathbf{D}(B_{ac}||B_c)]$*

Lemma 2.8.

$$\mathbf{D}(P_1 \times P_2||Q_1 \times Q_2) = \mathbf{D}(P_1||Q_1) + \mathbf{D}(P_2||Q_2).$$

We will use the following two simple corollaries of the Chain Rule many times throughout the paper:

Proposition 2.9. *Let A, B, C, D be four random variables such that $I(B; D|AC) = 0$. Then*

$$I(A; B|C) \geq I(A; B|CD).$$

Proof. We apply the chain rule twice:

$$\begin{aligned} I(A; B|CD) = I(AD; B|C) - I(D; B|C) = I(A; B|C) + I(D; B|AC) - I(D; B|C) \\ = I(A; B|C) - I(D; B|C) \leq I(A; B|C). \end{aligned}$$

□

Proposition 2.10. *Let A, B, C, D be four random variables such that $I(B; D|C) = 0$. Then*

$$I(A; B|C) \leq I(A; B|CD).$$

Proof. Once again, we apply the chain rule twice:

$$I(A; B|CD) = I(AD; B|C) - I(D; B|C) = I(AD; B|C) = I(A; B|C) + I(D; B|AC) \geq I(A; B|C).$$

□

2.2 Communication Complexity

Let $\mathcal{X}, \mathcal{Y}$ denote the set of possible inputs to the two players, who we name A and B. We view a *private coins protocol* for computing a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{Z}_K$ as a rooted tree with the following structure:

- Each non-leaf node is *owned* by A or by B.
- Each non-leaf node owned by a particular player has a set of children that are owned by the other player. Each of these children is labeled by a binary string, in such a way that this coding is prefix free: no child has a label that is a prefix of another child.
- Every node is associated with a function mapping $\mathcal{X}$ to distributions on children of the node and a function mapping $\mathcal{Y}$ to distributions on children of the node.
- The leaves of the protocol are labeled by output values.

On input x, y, the protocol π is executed as in Figure 1.

A public coin protocol is a distribution on private coins protocols, run by first using shared randomness to sample an index r and then running the corresponding private coin protocol π_r. Every private coin protocol is thus a public coin protocol. The protocol is called deterministic if all distributions labeling the nodes have support size 1.

Definition 2.11. The *communication cost* (or communication complexity) of a public coin protocol π, denoted $\mathsf{CC}(\pi)$, is the maximum number of bits that can be transmitted in any run of the protocol.

Definition 2.12. The *number of rounds* of a public coin protocol is the maximum depth of the protocol tree π_r over all choices of the public randomness.

Generic Communication Protocol

1. Set v to be the root of the protocol tree.
2. If v is a leaf, the protocol ends and outputs the value in the label of v. Otherwise, the player owning v samples a child of v according to the distribution associated with her input for v and sends the label to indicate which child was sampled.
3. Set v to be the newly sampled node and return to the previous step.

Figure 1: A communication protocol.

Given a protocol π, $\pi(x, y)$ denotes the concatenation of the public randomness with all the messages that are sent during the execution of π. We call this the *transcript* of the protocol. When referring to the random variable denoting the transcript, rather than a specific transcript, we will use the notation $\Pi(x, y)$, thus $\pi(x, y) \in_R \Pi(x, y)$. When x and y are random variables themselves, we will denote the transcript by $\Pi(X, Y)$, or just Π. We shall use the notation $\pi(x, y)_j$ or π_j to refer to the j'th transmitted message in the protocol. We write $\pi(x, y)_{\leq j}$ to denote the concatenation of the public randomness in the protocol with the first j message bits that were transmitted in the protocol. Given a transcript, or a prefix of the transcript, v, we write $\mathsf{CC}(v)$ to denote the number of message bits in v (i.e. the length of the communication).

Definition 2.13 (Communication Complexity notation)**.** For a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{Z}_K$, a distribution μ supported on $\mathcal{X} \times \mathcal{Y}$, and a parameter $\rho > 0$, $D^{\mu}_{\rho}(f)$ denotes the communication complexity of the cheapest deterministic protocol for computing f on inputs sampled according to μ with error ρ. $R_{\rho}(f)$ denotes the cost of the best randomized public coin protocol for computing f with error at most ρ on *every* input.

We should mention the following theorem due to Yao, which we will extend in this paper to information cost:

Theorem 2.14 (Yao's Min-Max)**.** $R_{\rho}(f) = \max_{\mu} D^{\mu}_{\rho}(f)$.

2.3 Information + communication: the information cost of a protocol

In the sections that follow we will develop the information complexity theory for interactive communication. The most basic definition is that of the information cost of a protocol. This notion is implicit in [BYJKS04], and was explicitly defined in [BBCR10] (see also [BR10]).

Definition 2.15. The information cost of a protocol π over inputs from $\mathcal{X} \times \mathcal{Y}$ is given by:

$$\mathsf{IC}_{\mu}(\pi) := I(\Pi; X|Y) + I(\Pi; Y|X).$$

Intuitively, Definition 2.15 captures what the two parties learn about each other's inputs from the execution transcript of the protocol π. The first term captures what the second player learns about X from Π – the mutual information between the input X and the transcript Π given the input Y. Similarly, the second term captures what the first player learns about Y from Π.

Note that the information of a protocol π depends on the prior distribution μ, as the mutual information between the transcript Π and the inputs depends on the prior distribution on the inputs. To give an extreme example, if μ is a singleton distribution, i.e. one with $\mu(\{(x,y)\}) = 1$ for some $(x,y) \in \mathcal{X} \times \mathcal{Y}$, then $\mathsf{IC}_\mu(\pi) = 0$ for all possible π, as no protocol can reveal anything to the players about the inputs that the do not already know *a-priori*. Similarly, $\mathsf{IC}_\mu(\pi) = 0$ if $\mathcal{X} = \mathcal{Y}$ and μ is supported on the diagonal $\{(x,x) : x \in \mathcal{X}\}$.

As expected, one can show that the communication cost $\mathsf{CC}(\pi)$ of π is an upper bound on its information cost over *any* distribution μ:

Lemma 2.16. *[BR10] For any distribution μ, $\mathsf{IC}_\mu(\pi) \leq \mathsf{CC}(\pi)$.*

On the other hand, as demonstrated by the simple examples above, the converse need not hold. Moreover, while $\mathsf{CC}(\pi)$ is a combinatorial property that depends only on the protocol, $\mathsf{IC}_\mu(\pi)$ depends on μ. One of our fist goals it to make the study of information cost independent of the prior distribution.

3 The prior-free information complexity of a problem

Given a distribution μ on a two-player input space $\mathcal{X} \times \mathcal{Y}$, a function $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ and an error parameter ε, the *information complexity* $\mathsf{IC}_\mu(f,\varepsilon)$ is defined to be the *infimum* of the information cost over all (randomized) protocols π that achieve an error of $\leq \varepsilon$ with respect to μ.

Remark 3.1. We state all our results for boolean functions. In fact we do not need f to be binary or even a function. All our results hold if the goal is to implement a general relation F that accepts inputs in $\mathcal{X} \times \mathcal{Y}$ and for which a protocol is said to succeed if in the end the two players reach an acceptable pair of states. In the case of a boolean function f there is only one acceptable pair of states in which each player outputs "the answer is $f(x,y)$".

Definition 3.2.

$$\mathsf{IC}_\mu(f,\varepsilon) := \inf_{\pi:\ \mathbf{P}_{(x,y)\sim\mu}[\pi(x,y)\neq f(x,y)]\leq\varepsilon} \mathsf{IC}_\mu(\pi).$$

Clearly, $\mathsf{IC}_\mu(f,\varepsilon)$ is monotone non-decreasing in ε – i.e. a lower error ε requires higher information cost. We note that the definition of $\mathsf{IC}_\mu(f,0)$ – the zero-error information complexity also makes sense. By Lemma 2.16, it is easy to see that the information costs are smaller than the corresponding (distributional) communication costs.

Information-theoretic quantities, such as the mutual information between the parties' inputs and the protocol's transcript only make sense with respect to a prior distribution on inputs. Nonetheless, at least syntactically, we can factor out the distribution by taking the maximum over all possible distributions:

Definition 3.3. The max-distributional information complexity of a function f with error ε is

$$\mathsf{IC}_{\mathsf{D}}(f,\varepsilon) := \max_{\mu \text{ a distribution on } \mathcal{X}\times\mathcal{Y}} \mathsf{IC}_\mu(f,\varepsilon).$$

Note that we are justified in using a max in place of inf since the set of all possible distributions over $X \times Y$ is compact and $\mathsf{IC}_\mu(f,\varepsilon)$ can be shown to be continuous in μ (and ε).

It is not immediately apparent that the definition of $\mathsf{IC}_{\mathsf{D}}(f,\varepsilon)$ is an interesting one. While it gives an upper bound on the information cost for each distribution, this upper bound may be attained by a different protocol $\pi = \pi(\mu)$ for each distribution μ. We will show that the quantifiers may "almost" be reversed: there is a single (randomized) protocol π that achieves a low information cost *and* a low error with respect to *any* distribution μ. In fact, we define the *prior-free* information complexity of a problem in terms of such a protocol.

Definition 3.4. The information complexity of a function f with error ε is

$$\mathsf{IC}(f,\varepsilon) := \inf_{\pi \text{ is a protocol with } \mathbf{P}[\pi(x,y)\neq f(x,y)]\leq\varepsilon \text{ for all } (x,y)} \max_{\mu} \mathsf{IC}_{\mu}(\pi).$$

Clearly, $\mathsf{IC}(f,\varepsilon) \geq \mathsf{IC}_{\mathsf{D}}(f,\varepsilon)$. The opposite direction is our main result connecting the prior-free information cocomplexity with the distributional information complexity:

Theorem 3.5. *Let $f : \mathcal{X}\times\mathcal{Y}\to\{0,1\}$ be any function, and $\varepsilon \geq 0$ be an error parameter. For each value of the parameter $0<\alpha<1$ we have*

$$\mathsf{IC}\left(f,\frac{\varepsilon}{\alpha}\right) \leq \frac{\mathsf{IC}_{\mathsf{D}}(f,\varepsilon)}{1-\alpha}$$

In other words, there is a protocol π such that:

1. *for each $(x,y)\in\mathcal{X}\times\mathcal{Y}$, $\mathbf{P}[\pi(x,y)\neq f(x,y)]\leq\frac{\varepsilon}{\alpha}$, i.e. the protocol π makes an error of at most ε/α on each input;*
2. *for each distribution μ on $\mathcal{X}\times\mathcal{Y}$, $\mathsf{IC}_{\mu}(\pi)\leq\frac{\mathsf{IC}_{\mathsf{D}}(f,\varepsilon)}{1-\alpha}$, i.e. for* every *distribution the protocol π reveals not too much information to the participants.*

By selecting $\alpha=\frac{1}{2}$ we can ensure that both the error and the information cost increase by a factor of at most 2.

For the zero-error case, we similarly obtain the following zero-error version of Theorem 3.5:

Theorem 3.6. *Let $f : \mathcal{X}\times\mathcal{Y}\to\{0,1\}$ be any function. Then we have*

$$\mathsf{IC}(f,0) = \mathsf{IC}_{\mathsf{D}}(f,0).$$

In other words, there is a protocol π such that:

1. *for each $(x,y)\in\mathcal{X}\times\mathcal{Y}$, $\pi(x,y)=f(x,y)$, i.e. the protocol π always works correctly;*
2. *for each distribution μ on $\mathcal{X}\times\mathcal{Y}$, $\mathsf{IC}_{\mu}(\pi)\leq\mathsf{IC}(f,0)$.*

3.1 Proof of Theorems 3.5 and 3.6

Proof of Theorem 3.5. We prove the theorem using a minimax argument. For the remainder of the proof fix f and ε, and denote $I := \mathsf{IC}_{\mathsf{D}}(f,\varepsilon)$. Define the following two-player zero-sum game. Player A will come up with a (randomized) two-party protocol $\pi(x,y)$ taking inputs in $\mathcal{X}\times\mathcal{Y}$. Player B will come up with a distribution μ on inputs (x,y). Player B's payoff is given by:

$$P_B(\pi,\mu) := (1-\alpha)\cdot\frac{\mathsf{IC}_{\mu}(\pi)}{I} + \alpha\cdot\frac{\mathbf{P}_{\mu}[\pi(x,y)\neq f(x,y)]}{\varepsilon}.$$

As the game is a zero-sum game, the payoff of player A is given by $P_A(\pi, \mu) := -P_B(\pi, \mu)$. Thus player A comes up with a protocol for f, and player B comes up with a distribution that tries to highlight the mistakes of π, and also make it reveal information about the inputs to the participants. We denote the game by $\mathcal{G}$. Our first goal is to show that the value of $\mathcal{G}$ to player B is bounded by 1.

Claim 3.7. *The value $V_B(\mathcal{G}) \leq 1$.*

Proof. Let ν_B be a probability distribution representing a mixed strategy for player B. Thus ν_B is a distribution on probability distributions μ over $X \times Y$. We will actually show that $V_B(\mathcal{G}) < 1 + \delta$ for each $\delta > 0$. To show that $V_B(\mathcal{G}) < 1 + \delta$ it suffices to show that there is a protocol π such that $\mathbf{E}_{\mu \sim \nu_B}[P_B(\pi, \mu)] < 1 + \delta$. Let $\bar{\mu}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$ that is obtained by taking the average of $\mu \sim \nu_B$. Formally,

$$\bar{\mu}(x, y) := \mathbf{E}_{\mu \sim \nu_B} \mu(x, y).$$

By the definition of $I = \mathsf{IC}(f, \varepsilon)$, we know that there is a protocol π such that $\mathbf{P}_{\bar{\mu}}[\pi(x, y) \neq f(x, y)] \leq \varepsilon$ and $\mathsf{IC}_{\bar{\mu}}(\pi) < I \cdot (1 + \delta)$. We claim that

$$\mathbf{E}_{\mu \sim \nu_B} \left[I_{(X,Y) \sim \mu}(\pi(X, Y); X | Y) \right] \leq I_{(X,Y) \sim \bar{\mu}}(\pi(X, Y); X | Y). \tag{3}$$

In other words, the average amount of information revealed by π with respect to the different distributions $\mu \sim \nu_B$ is smaller or equal to the amount of information revealed with respect to $\bar{\mu}$ – the concavity works in the "right" direction.

To establish (3), consider the following four random variables. Let M be a random variable representing the distribution μ. Then M is distributed according to ν_B. Let X and Y be the inputs to the two parties in π such that (X, Y) is distributed according to μ. Finally, let $\Pi = \pi(X, Y)$ be the transcript of the protocol executed on X and Y. Π is randomized even conditioned on (X, Y) due to the public and private randomness used in the execution of the protocol. In this language, we have:

$$\mathbf{E}_{\mu \sim \nu_B} \left[I_{(X,Y) \sim \mu}(\pi(X, Y); X | Y) \right] = I(\Pi; X | Y M),$$

and

$$I_{(X,Y) \sim \bar{\mu}}(\pi(X, Y); X | Y) = I(\Pi; X | Y).$$

Since the distribution of Π only depends on X and Y, we have $I(\Pi; M | XY) = 0$. By substituting $A = X$, $B = \Pi$, $C = Y$, and $D = M$ into Proposition 2.9 we get

$$I(X; \Pi | Y) \geq I(X; \Pi | Y M), \tag{4}$$

proving (3). Similarly to (3) the following symmetric inequality is established:

$$\mathbf{E}_{\mu \sim \nu_B} \left[I_{(X,Y) \sim \mu}(\pi(X, Y); Y | X) \right] \leq I_{(X,Y) \sim \bar{\mu}}(\pi(X, Y); Y | X). \tag{5}$$

Together, (3) and (5) imply

$$\mathbf{E}_{\mu \sim \nu_B} \left[\mathsf{IC}_{\mu}(\pi) \right] \leq \mathsf{IC}_{\bar{\mu}}(\pi). \tag{6}$$

Using (6) we obtain

$$\begin{aligned}\mathbf{E}_{\mu\sim\nu_B}\left[P_B(\pi,\mu)\right] &= \mathbf{E}_{\mu\sim\nu_B}\left[(1-\alpha)\cdot\frac{\mathsf{IC}_\mu(\pi)}{I}+\alpha\cdot\frac{\mathbf{P}_\mu[\pi(x,y)\neq f(x,y)]}{\varepsilon}\right]\\ &= (1-\alpha)\cdot\mathbf{E}_{\mu\sim\nu_B}\left[\frac{\mathsf{IC}_\mu(\pi)}{I}\right]+\alpha\cdot\frac{\mathbf{P}_{\bar\mu}[\pi(x,y)\neq f(x,y)]}{\varepsilon}\\ &\leq (1-\alpha)\cdot\frac{\mathsf{IC}_{\bar\mu}(\pi)}{I}+\alpha\cdot\frac{\mathbf{P}_{\bar\mu}[\pi(x,y)\neq f(x,y)]}{\varepsilon} < (1-\alpha)\cdot(1+\delta)+\alpha\cdot 1\leq 1+\delta.\end{aligned}$$

This proves Claim 3.7. □

By the Minimax Theorem, Claim 3.7 implies that there is a mixed strategy for player A such that for each response by player B, the value of the game for player B is at most 1. A mixed strategy for player A is a distribution ν_A on protocols. In other words,

$$\mathbf{E}_{\pi\sim\nu_A}P_B(\pi,\mu)\leq 1,\ \text{for all } \mu. \tag{7}$$

Let $\bar\pi$ be the randomized protocol obtained by publicly sampling $\pi\sim\nu_A$, and then applying π to the inputs. We claim that $\bar\pi$ is the protocol we are looking for. In other words, the randomized protocol $\bar\pi$ has the desired payoff properties that will translate into information cost/error properties.

Claim 3.8. *For each distribution μ, $P_B(\bar\pi,\mu)\leq 1$.*

Proof. The proof proceeds similarly to the proof of Claim 3.7. We will prove first that

$$I_{(X,Y)\sim\mu}(\bar\pi(X,Y);X|Y)\leq\mathbf{E}_{\pi\sim\nu_A}\left[I_{(X,Y)\sim\mu}(\pi(X,Y);X|Y)\right]. \tag{8}$$

In other words, the amount of information revealed by $\bar\pi$ is bounded by the average amount of information revealed by π that is drawn according to ν_A – once again, the concavity works in the "right" direction.

To establish (8), consider the following four random variables. Let S be a "selector" random variable, that picks the protocol π to run according to the distribution ν_A. Let X and Y be inputs distributed according to μ independently of S. Finally, let $\Pi=\pi(X,Y)$ be the transcript of the selected protocol executed on X and Y. We have:

$$\mathbf{E}_{\pi\sim\nu_A}\left[I_{(X,Y)\sim\mu}(\pi(X,Y);X|Y)\right]=I(\Pi;X|YS),$$

and

$$I_{(X,Y)\sim\mu}(\bar\pi(X,Y);X|Y)=I(\Pi;X|Y).$$

Since the protocol π is selected independently of the inputs, we have $I(X;S|Y)=0$. By substituting $A=\Pi$, $B=X$, $C=Y$, and $D=S$ into Proposition 2.10 we get

$$I(\Pi;X|Y)\leq I(\Pi;X|YS), \tag{9}$$

establishing (8). Similarly to (8) the following symmetric inequality is established:

$$I_{(X,Y)\sim\mu}(\bar\pi(X,Y);Y|X)\leq\mathbf{E}_{\pi\sim\nu_A}\left[I_{(X,Y)\sim\mu}(\pi(X,Y);Y|X)\right]. \tag{10}$$

Together, (8) and (10) imply

$$\mathsf{IC}_\mu(\bar\pi)\leq\mathbf{E}_{\pi\sim\nu_A}\left[\mathsf{IC}_\mu(\pi)\right]. \tag{11}$$

Finally, (11) implies that

$$\begin{aligned} P_B(\bar{\pi},\mu) &= (1-\alpha)\cdot\frac{\mathsf{IC}_\mu(\bar{\pi})}{I} + \alpha\cdot\frac{\mathbf{P}_\mu[\bar{\pi}(x,y)\neq f(x,y)]}{\varepsilon} \\ &= (1-\alpha)\cdot\frac{\mathsf{IC}_\mu(\bar{\pi})}{I} + \mathbf{E}_{\pi\sim\nu_A}\left[\alpha\cdot\frac{\mathbf{P}_\mu[\pi(x,y)\neq f(x,y)]}{\varepsilon}\right] \\ \leq (1-\alpha)\cdot\mathbf{E}_{\pi\sim\nu_A}\left[\frac{\mathsf{IC}_\mu(\pi)}{I}\right] &+ \mathbf{E}_{\pi\sim\nu_A}\left[\alpha\cdot\frac{\mathbf{P}_\mu[\pi(x,y)\neq f(x,y)]}{\varepsilon}\right] = \mathbf{E}_{\pi\sim\nu_A}\left[P_B(\pi,\mu)\right] \leq 1, \end{aligned}$$

completing the proof of Claim 3.8. □

To complete the proof of Theorem 3.5, we observe that the randomized protocol $\bar{\pi}$ satisfies the conclusions of the theorem. For each distribution μ we know that $P_B(\bar{\pi},\mu)\leq 1$, and thus

$$\mathsf{IC}_\mu(\bar{\pi}) \leq \frac{I}{1-\alpha} \quad \text{and} \quad \mathbf{P}_\mu[\bar{\pi}(x,y)\neq f(x,y)] \leq \frac{\varepsilon}{\alpha}.$$

The first inequality is exactly the second requirement in Theorem 3.5. We obtain the second requirement by letting μ be the atomic distribution on $\{(x,y)\}$. □

Remark 3.9. Note that the statement of Theorem 3.5 contains a gap in the following sense. We are unable to simultaneously achieve the goals of low information and low error with the same parameters as in the distributional setting. Instead we lose factors of $\frac{1}{1-\alpha}$ and $\frac{1}{\alpha}$. It is an interesting open problem whether this analysis is actually tight, i.e. whether $\mathsf{IC}(f,\varepsilon) = \mathsf{IC}_{\mathsf{D}}(f,\varepsilon)$. It is quite possible that there is, in fact, a gap between the two quantities.

Proof of Theorem 3.6. The proof is very similar to the proof of Theorem 3.5. The key difference is that now we only deal with zero-error protocols. Denote $I := \mathsf{IC}_{\mathsf{D}}(f,0)$. We define the following zero-sum two player game $\mathcal{G}_0$. Player A will come up with a (randomized) two-party protocol $\pi(x,y)$ taking inputs in $\mathcal{X}\times\mathcal{Y}$. The protocol is required to always be correct: for any pair of inputs (x,y), $\pi(x,y)=f(x,y)$ with probability 1 with respect to the random coin tosses within π. Player B will come up with a distribution μ on inputs (x,y). Note that the protocol π has to be correct even on inputs outside of the support of μ. Player B's payoff is given by:

$$P_B(\pi,\mu) := \frac{\mathsf{IC}_\mu(\pi)}{I}.$$

Once again, we first need to establish that the value of the game for player B is bounded by 1.

Claim 3.10. *The value $V_B(\mathcal{G}_0)\leq 1$.*

Proof. The proof is quite similar to the proof of Claim 3.7 with one minor twist: by the assumption, for each distribution μ there is a zero-error protocol π_μ whose information cost with respect to μ is $< (1+\delta)\cdot I$. The problem is that π_μ only has to be zero-error *on the support of* μ. This problem is easily overcome by modifying μ very slightly to ensure that it has full support.

We will actually show that $V_B(\mathcal{G})<1+\delta$ for each $\delta>0$. Let ν_B be a probability distribution representing a mixed strategy for player B. Thus ν_B is a distribution on probability distributions μ over $\mathcal{X}\times\mathcal{Y}$. We modify ν_B to select the distribution μ_u uniform on $\mathcal{X}\times\mathcal{Y}$ with probability $\delta/4$ to obtain ν'_B. Player B's payoff is always non-negative. Hence, if there is a strategy π for player A

that guarantees that B's payoff with respect to ν'_B is $< 1 + \delta/2$, then using the same strategy will guarantee that B's payoff with respect to ν_B is $< 1 + \delta$. Otherwise, B's payoff with respect to ν'_B would be at least

$$(1 + \delta) \cdot (1 - \delta/4) = 1 + 3\delta/4 - \delta^2/4 > 1 + \delta/2.$$

Thus we need to show that there is a zero-error protocol π such that $\mathbf{E}_{\mu \sim \nu'_B}[\mathsf{IC}_\mu(\pi)] < 1 + \delta/2$. Let $\bar{\mu}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$ that is obtained by taking the average of $\mu \sim \nu'_B$. Formally,

$$\bar{\mu}(x, y) := \mathbf{E}_{\mu \sim \nu_B} \mu(x, y).$$

Note that since one of the distributions in ν'_B is the uniform distribution on $\mathcal{X} \times \mathcal{Y}$, $\bar{\mu}$ has full support.

By the definition of $I = \mathsf{IC}(f, 0)$, we know that there is a protocol π that achieves zero-error with respect to $\bar{\mu}$ such that $\mathsf{IC}_{\bar{\mu}}(\pi) < I \cdot (1 + \delta/2)$. Since $\bar{\mu}$ has full support, π is a feasible strategy for player A. Following the analysis in Claim 3.7 we obtain that (6) still holds here:

$$\mathbf{E}_{\mu \sim \nu'_B} [\mathsf{IC}_\mu(\pi)] \leq \mathsf{IC}_{\bar{\mu}}(\pi).$$

Thus

$$\mathbf{E}_{\mu \sim \nu_B} [P_B(\pi, \mu)] = \mathbf{E}_{\mu \sim \nu'_B} [\mathsf{IC}_\mu(\pi)] / I \leq \mathsf{IC}_{\bar{\mu}}(\pi)/I < 1 + \delta/2,$$

completing the proof of Claim 3.10. □

The rest of the proof of Theorem 3.6 is identical to the proof of Theorem 3.5. There is a distribution ν_A of strategies for player A such that each protocol in ν_A is a zero-error protocol, and for each distribution μ the expected payoff $\mathbf{E}_{\pi \sim \nu_A}[P_B(\pi, \mu)] \leq 1$. If we let $\bar{\pi}$ be the randomized protocol obtained by sampling π according to ν_A and then running it, then exactly as in Claim 3.8 above we obtain

$$\mathsf{IC}_\mu(\bar{\pi}) \leq \mathbf{E}_{\pi \sim \nu_A}[\mathsf{IC}_\mu(\bar{\pi})] = I \cdot \mathbf{E}_{\pi \sim \nu_A}[P_B(\pi, \mu)] \leq I.$$

□

3.2 External information complexity

We next turn our attention to the *external information complexity* of a problem. Recall that the regular (internal) information cost of a problem is the amount of information the participating parties have to learn about the inputs to solve the problem. By analogy, the external information cost is the amount of information an observer necessarily has to learn when the parties perform the computation.

We start by defining the external information cost of a protocol π.

Definition 3.11. Let π be a two-party communication protocol over inputs in $\mathcal{X} \times \mathcal{Y}$. Let μ be a distribution over $\mathcal{X} \times \mathcal{Y}$. We denote by (X, Y) the (random) pair of inputs given to the players that are distributed according to μ. Let $\Pi = \Pi(X, Y)$ denote the random variable that is the transcript of the protocol. Then the external information cost of π with respect to μ is given by:

$$\mathsf{IC}^{\mathsf{ext}}_\mu(\pi) := I(XY; \Pi).$$

We note that the external information cost is always greater or equal to the internal information cost of a protocol:

Proposition 3.12. *For each protocol π and distribution μ,*

$$\mathsf{IC}^{\mathsf{ext}}_{\mu}(\pi) \geq \mathsf{IC}_{\mu}(\pi).$$

Moreover, the two quantities are equal if $\mu = \mu_{\mathcal{X}} \times \mu_{\mathcal{Y}}$ is a product distribution on $\mathcal{X} \times \mathcal{Y}$.

Proof. The proposition is only true because π is a protocol: it is not hard to see that it would fail if Π was a general random variable that is correlated with the inputs X and Y. Denote by $\Pi_1, \Pi_2, \ldots, \Pi_N$ the messages sent in the protocol π. We assume that Π_i is sent by Alice (who holds X) for odd i's and by Bob for even i's. N itself is also a random variable here.

By the chain rule, we have:

$$\mathsf{IC}_{\mu}(\pi) = I(X;\Pi|Y) + I(Y;\Pi|X) = \sum_{i=1}^{N} \left[I(X;\Pi_i|\Pi_1,\ldots,\Pi_{i-1}Y) + I(Y;\Pi_i|\Pi_1,\ldots,\Pi_{i-1}X) \right],$$

and

$$\mathsf{IC}^{\mathsf{ext}}_{\mu}(\pi) = I(XY;\Pi) = \sum_{i=1}^{N} I(XY;\Pi_i|\Pi_1,\ldots,\Pi_{i-1}).$$

We will prove the inequality term-wise. Let i be an index. Without loss of generality assume that i is odd, so that Π_i is a message sent by Alice. This means that conditioned on prior communications and on X, Π_i is independent from Y:

$$I(\Pi_i; Y|\Pi_1,\ldots,\Pi_{i-1}X) = 0. \tag{12}$$

We can now write:

$$I(XY;\Pi_i|\Pi_1,\ldots,\Pi_{i-1}) = I(X;\Pi_i|\Pi_1,\ldots,\Pi_{i-1}) + I(Y;\Pi_i|\Pi_1,\ldots,\Pi_{i-1}X) \geq \\ I(X;\Pi_i|\Pi_1,\ldots,\Pi_{i-1}Y) + I(Y;\Pi_i|\Pi_1,\ldots,\Pi_{i-1}X),$$

where the last inequality follows from (12) and Proposition 2.9 by taking $A = X$, $B = \Pi_i$, $C = \Pi_1,\ldots,\Pi_{i-1}$ and $D = Y$.

It remains to prove the converse inequality in the case when μ is a product distribution. This direction actually holds when Π is any random variable. If μ is a product distribution, then X and Y are independent, and thus $I(X;Y) = 0$. We have

$$\mathsf{IC}^{\mathsf{ext}}_{\mu}(\pi) = I(XY;\Pi) = I(X;\Pi) + I(Y;\Pi|X) \leq I(X;\Pi|Y) + I(Y;\Pi|X) = \mathsf{IC}_{\mu}(\pi),$$

where the inequality holds by Proposition 2.10 with $A = \Pi$, $B = X$, $C = \emptyset$, and $D = Y$. □

We can now define the natural analogues of Definitions 3.2, 3.3, and 3.4 for external information, starting with the distributional external information complexity.

Definition 3.13.

$$\mathsf{IC}^{\mathsf{ext}}_{\mu}(f,\varepsilon) := \inf_{\pi:\ \mathbf{P}_{(x,y)\sim\mu}[\pi(x,y)\neq f(x,y)]\leq\varepsilon} \mathsf{IC}^{\mathsf{ext}}_{\mu}(\pi).$$

Definition 3.14. The max-distributional external information complexity of a function f with error ε is

$$\mathsf{IC}^{\mathsf{ext}}_{\mathsf{D}}(f,\varepsilon) := \max_{\mu \text{ a distribution on } \mathcal{X}\times\mathcal{Y}} \mathsf{IC}^{\mathsf{ext}}_{\mu}(f,\varepsilon).$$

Definition 3.15. The external information complexity of a function f with error ε is

$$\mathsf{IC}^{\mathsf{ext}}(f,\varepsilon) := \inf_{\pi \text{ is a protocol with } \mathbf{P}[\pi(x,y)\neq f(x,y)]\leq \varepsilon \text{ for all } (x,y)} \max_{\mu} \mathsf{IC}^{\mathsf{ext}}_{\mu}(\pi).$$

Proposition 3.12 implies that for all μ, $\mathsf{IC}^{\mathsf{ext}}_{\mu}(f,\varepsilon) \geq \mathsf{IC}_{\mu}(f,\varepsilon)$, and also $\mathsf{IC}^{\mathsf{ext}}_{\mathsf{D}}(f,\varepsilon) \geq \mathsf{IC}_{\mathsf{D}}(f,\varepsilon)$ and $\mathsf{IC}^{\mathsf{ext}}(f,\varepsilon) \geq \mathsf{IC}(f,\varepsilon)$.

Finally, the analogues of Theorems 3.5 and 3.6 for external information cost are proved in the exact same way. We formulate the theorems here:

Theorem 3.16. *Let $f : \mathcal{X}\times\mathcal{Y}\to\mathbb{Z}_K$ be any function, and $\varepsilon\geq 0$ be an error parameter. For each value of the parameter $0<\alpha<1$ we have*

$$\mathsf{IC}^{\mathsf{ext}}\left(f,\frac{\varepsilon}{\alpha}\right) \leq \frac{\mathsf{IC}^{\mathsf{ext}}_{\mathsf{D}}(f,\varepsilon)}{1-\alpha}$$

In other words, there is a protocol π such that:

1. *for each $(x,y)\in\mathcal{X}\times\mathcal{Y}$, $\mathbf{P}[\pi(x,y)\neq f(x,y)]\leq\frac{\varepsilon}{\alpha}$, i.e. the protocol π makes an error of at most ε/α on each input;*
2. *for each distribution μ on $\mathcal{X}\times\mathcal{Y}$, $\mathsf{IC}^{\mathsf{ext}}_{\mu}(\pi)\leq\frac{\mathsf{IC}^{\mathsf{ext}}_{\mathsf{D}}(f,\varepsilon)}{1-\alpha}$, i.e. for* every *distribution the protocol π reveals not too much information to the observer.*

Theorem 3.17. *Let $f : \mathcal{X}\times\mathcal{Y}\to\mathbb{Z}_K$ be any function. Then we have*

$$\mathsf{IC}^{\mathsf{ext}}(f,0) = \mathsf{IC}^{\mathsf{ext}}_{\mathsf{D}}(f,0).$$

In other words, there is a protocol π such that:

1. *for each $(x,y)\in\mathcal{X}\times\mathcal{Y}$, $\pi(x,y)=f(x,y)$, i.e. the protocol π always works correctly;*
2. *for each distribution μ on $\mathcal{X}\times\mathcal{Y}$, $\mathsf{IC}^{\mathsf{ext}}_{\mu}(\pi)\leq\mathsf{IC}^{\mathsf{ext}}(f,0)$.*

The only difference between the proofs of Theorem 3.5 and of its external version 3.16 is that equations (4) and (9) are slightly different. The external version of equation (4) is

$$I(XY;\Pi)\geq I(XY;\Pi|M), \tag{13}$$

which still follows from Proposition 2.9 with $A=XY$, $B=\Pi$, $C=\emptyset$, and $D=M$, and the fact that $I(\Pi;M|XY)=0$. The external version of equation (9) is

$$I(\Pi;XY)\leq I(\Pi;XY|S), \tag{14}$$

which follows from Proposition 2.10 with $A=\Pi$, $B=XY$, $C=\emptyset$, and $D=S$, and the fact that $I(XY;S)=0$.

3.3 The information complexity is convex in the error parameter

In this section we prove that the information complexity $\mathsf{IC}(f,\varepsilon)$ is convex in the error parameter ε. Note that this function is trivially non-increasing.

Theorem 3.18. *For any f, the function $I(\varepsilon) := \mathsf{IC}(f,\varepsilon)$ is convex on the interval $\varepsilon \in [0,1]$.*

Since a convex function on an interval must be continuous, the following is an immediate corollary of Theorem 3.18:

Corollary 3.19. *For any f, the function $I(\varepsilon) := \mathsf{IC}(f,\varepsilon)$ is continuous on the interval $\varepsilon \in [0,1]$.*

Remark 3.20. A statement analogous to Theorem 3.18, with a very similar proof, holds for the max-distributional information complexity $\mathsf{IC}_{\mathsf{D}}(f,\varepsilon)$.

Proof of Theorem 3.18. Let $0 \leq \varepsilon_1 < \varepsilon_2 \leq 1$ be two values, and let $\alpha \in (0,1)$ be a parameter. Set $\varepsilon := \alpha \cdot \varepsilon_1 + (1-\alpha)\cdot \varepsilon_2$. Our goal is to show that

$$I(\varepsilon) \leq \alpha \cdot I(\varepsilon_1) + (1-\alpha)\cdot I(\varepsilon_2).$$

Let $\delta > 0$ be a parameter that tends to 0. By the definition of $\mathsf{IC}(f,\varepsilon)$, there is a pair of protocols π_1 and π_2 that attain an error of at most ε_1 and ε_2, respectively, on each input (x,y), and such that for each distribution μ,

$$\mathsf{IC}_\mu(\pi_1) < I(\varepsilon_1) + \delta, \quad \text{and} \quad \mathsf{IC}_\mu(\pi_2) < I(\varepsilon_2) + \delta.$$

Let π be a protocol that publicly tosses a coin to select between π_1 and π_2 and then runs the selected protocol. π_1 is selected with probability α and π_2 is selected with probability $(1-\alpha)$. For each pair of inputs (x,y) we have

$$\mathbf{P}[\pi(x,y) \neq f(x,y)] = \alpha \cdot \mathbf{P}[\pi_1(x,y) \neq f(x,y)] + (1-\alpha)\cdot \mathbf{P}[\pi_2(x,y) \neq f(x,y)] \leq \varepsilon.$$

Following the same proof as the proof of (11) in Claim 3.8 we can obtain for each μ:

$$\mathsf{IC}_\mu(\pi) \leq \alpha \cdot \mathsf{IC}_\mu(\pi_1) + (1-\alpha)\cdot \mathsf{IC}_\mu(\pi_2) < \alpha \cdot I(\varepsilon_1) + (1-\alpha)\cdot I(\varepsilon_2) + \delta.$$

This implies that $I(\varepsilon) < \alpha \cdot I(\varepsilon_1) + (1-\alpha)\cdot I(\varepsilon_2) + \delta$ for each $\delta > 0$, and hence

$$I(\varepsilon) \leq \alpha \cdot I(\varepsilon_1) + (1-\alpha)\cdot I(\varepsilon_2).$$

□

3.4 Example: the information complexity of equality

As an instructive illustration let us consider the information complexity of the equality function. The equality function $EQ : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ is given by $EQ(x,y) = 1$ if and only if $x = y$. It is well known that the deterministic communication complexity of EQ is $n+1$, while the randomized communication complexity if one allows error ε is bounded by $O(\log 1/\varepsilon)$ (assuming the parties have access to public randomness).

We will demonstrate a gap between the information complexity and the communication complexity of EQ. Specifically, we show that the information complexity of EQ is bounded by a constant that is independent of the error parameter ε. Moreover, we show that the zero-error information cost of equality is also constant. Thus there is a major gap between the deterministic communication complexity and the zero-error information cost (linear vs. constant).

Proposition 3.21. $\mathsf{IC}(EQ, 0) = O(1)$.

Discussion: Proposition 3.21, together with our other results, sheds interesting light onto the results by Feder, Kushilevitz, Naor, and Nisan [FKNN95] on the (violation of) direct sum for randomized communication complexity. [FKNN95] (and related constructions) give EQ as an example that appears to violate direct sum for randomized communication complexity. It is not hard to see that the randomized communication complexity of EQ with error ε, $R_\varepsilon(EQ) = \Omega(\log 1/\varepsilon)$. On the other hand, the (non-trivial) algorithm from [FKNN95] shows that the amortized communication complexity of EQ, $R_\varepsilon(EQ^n)/n = O(1)$, even for relatively small $n = \log^2 1/\varepsilon$. This shows that there is no hope for a direct sum theorem to hold if one regards the error ε as a parameter. Combining Proposition 3.21 with the '$\geq$' direction of the general Theorem 6.7 gives an alternative proof of this fact for any $\varepsilon > 0$. Thus, intuitively, the gap between the randomized and the amortized communication complexity of equality for small values of ε is caused by the gap between the randomized communication complexity (which depends on ε) and the inherent information cost of the problem (which is constant, even in the extreme case when $\varepsilon = 0$). At the same time, in light of this explanation, it is not clear whether this gap shed any light on the situation with constant ε. We will revisit these points in the discussion of Theorem 6.7.

Proof. We present a zero-error protocol π for equality that will have low information cost with respect to any prior distribution. Note that to prove the proposition it would have been sufficient to produce such a protocol π_μ for each given prior distribution μ, as Theorem 3.6 would guarantee the existence of a protocol π that simultaneously has low information cost with respect to all distributions μ. The protocol is presented on Figure 2.

A zero-error constant-information complexity protocol for EQ

1. The parties use public randomness to sample a uniformly random non-singular matrix $A \in \mathbb{F}_2^n$. Denote the rows of A by $a_1, a_2, \ldots, a_n$.
2. At step i, the first player sends $a_i \cdot x$ to the second player. If $a_i \cdot x \neq a_i \cdot y$, the second player responds "not equal" and the protocol terminates with $EQ(x, y) = 0$.
3. If the protocol hasn't terminated after n steps, the protocol terminates with $EQ(x, y) = 1$.

Figure 2: The protocol for Equality

Correctness. The protocol terminates on each input after at most n rounds. Both parties always output the same answer. If the protocol terminates and returns 0, then there is an a_i with $a_i \cdot x \neq a_i \cdot y$, and thus $x \neq y$. If the protocol terminates and outputs 1, then $Ax = Ay$. Since A is always chosen to be non-singular, this implies $x = y$.

Information cost. First of all, note that the protocol is essentially symmetric with respect to the two players: at each round the two players learn whether $a_i \cdot x = a_i \cdot y$, and only continue if this is the case. Let μ be any distributions on inputs (x, y). It suffices to show that $I_\mu(X; \pi(X, Y)|Y) = O(1)$ where $(X, Y) \sim \mu$. Let R be the public randomness used by the protocol to sample the matrix A, and let π_R be the deterministic protocol that uses this randomness. We know that (see e.g.

[BravermanRaoAmortized]) $I_\mu(X;\pi|Y) = \mathbf{E}_R[I_\mu(X;\pi_R|Y)]$. In fact, we will show that

$$\mathbf{E}_R[H_\mu(\pi|Y)] = O(1),$$

which would suffice since $\mathbf{E}_R[I_\mu(X;\pi_R|Y)] \leq \mathbf{E}_R[H_\mu(\pi|Y)]$.

Given an input y to the second player, and given the matrix A that is publicly sampled, the possible protocol transcripts are of two kinds:

1. If $x \neq y$ then the protocol may last anywhere from 1 to n rounds, and the transcript will only depend on the number i of rounds. Let $p_i = p_i(R, y, \mu)$ denote the probability that the protocol lasts exactly i rounds and $x \neq y$.

2. if $x = y$ then the protocol lasts n rounds, and its transcript is completely determined by A and y. Let $p_0 = p_0(R, y, \mu)$ denote the probability that the protocol outputs $x = y$.

Denote by $\bar{p}_i(y, \mu) := \mathbf{E}_R p_i(R, y, \mu)$. We can now calculate $\mathbf{E}_R[H_\mu(\pi|Y)]$:

$$\mathbf{E}_R[H_\mu(\pi|Y)] = \mathbf{E}_R \mathbf{E}_{y\sim\mu_y} \sum_{i=0}^{n} p_i(R, y, \mu) \log \frac{1}{p_i(R, y, \mu)} =$$

$$\mathbf{E}_{y\sim\mu_y} \mathbf{E}_R \sum_{i=0}^{n} p_i(R, y, \mu) \log \frac{1}{p_i(R, y, \mu)} \leq \mathbf{E}_{y\sim\mu_y} \sum_{i=0}^{n} \bar{p}_i(y, \mu) \log \frac{1}{\bar{p}_i(R, y, \mu)},$$

where the last inequality follows from the concavity of the $x \log(1/x)$ function. Next, note that for each $i \geq 2$, $\bar{p}_i(y, \mu) \leq 2^{-i+1}$, since the probability over R that the protocol will last for at least i rounds for each fixed pair $x \neq y$ is bounded by 2^{-i+1}. Thus for each y, and for each $i \geq 3$, $\bar{p}_i(y, \mu) \log \frac{1}{\bar{p}_i(R,y,\mu)} \leq 2^{-i+1} \cdot (i-1)$. In addition, it is always the case that $\bar{p}_i(y, \mu) \log \frac{1}{\bar{p}_i(R,y,\mu)} \leq 1$, and thus:

$$\mathbf{E}_R[H_\mu(\pi|Y)] \leq \mathbf{E}_{y\sim\mu_y} \sum_{i=0}^{n} \bar{p}_i(y, \mu) \log \frac{1}{\bar{p}_i(R, y, \mu)} < 3 + \sum_{i=3}^{\infty} 2^{-i+1} \cdot (i-1) = 4.5 = O(1).$$

□

4 The additivity of information complexity

In this section we will show that information complexity is additive. That is, the information complexity of performing two independent tasks is the sum of the information complexities of each individual task. This is a more general statement, though closely related, than a direct sum theorem that states that n copies of *the same* task cost n times as much as one copy [BR10]. In contrast with the information complexity, we do not know such a statement to be true for communication complexity, and it is known to be false in other settings. See [BBCR10] for a further discussion on direct sum problems.

We formulate the result in terms of tasks. A task is just a (possibly partial) relation $R(x, y, O_x, O_y)$, along with a required success criterion. Here O_x and O_y correspond to the outputs of the two players. For example, if the task is to compute a function f with probability $> 1-\varepsilon$ then $R(x, y, O_x, O_y) = 1$ if and only if $O_x = O_y = f(x, y)$, and the success criterion is that $R(x, y, O_x, O_y) = 1$ with probability $> 1 - \varepsilon$. Another example of a task is computing n independent copies of a function f such

that the probability of success on each copy is $> 1 - \varepsilon$. In this way, a task is essentially "anything that can be solved by a communication protocol".

The information cost of a task $T(x, y)$ is defined similarly to the information cost of a function. To be concrete and to avoid cumbersome notation, we assume that the inputs (x, y) belong to $\mathcal{U} := \{0,1\}^n$. Let $\mathcal{M}$ be a set of distributions on $\mathcal{U} \times \mathcal{U}$. Then the information cost of T with respect to $\mathcal{M}$ is defined as:

Definition 4.1. $\mathsf{IC}(T, \mathcal{M}) := \inf_{\pi \text{ succeeds at } T} \sup_{\mu \in \mathcal{M}} \mathsf{IC}_\mu(\pi)$, where the infimum is taken over protocols that successfully perform the task T.

Let $T_1(x_1, y_1)$ and $T_2(x_2, y_2)$ be two tasks. Let $T(x_1, x_2, y_1, y_2) := T_1 \times T_2$ be the task that consists of performing T_1 and T_2 in parallel on two pairs of inputs. A protocol is successful at T if it is successful at each of the two sub-tasks separately.

Next, we consider two products of sets of distributions. Let $\mathcal{M}_1$ be a set of distributions of pairs (x_1, y_1), and let $\mathcal{M}_2$ be a set of distributions of pairs (x_2, y_2). Denote

$$\mathcal{M}_1 \times \mathcal{M}_2 := \{\mu_1 \times \mu_2 : \ \mu_1 \in \mathcal{M}_1, \ \mu_2 \in \mathcal{M}_2\}.$$

We also define a bigger class of distributions which are not necessary product distributions, but whose projections fall into the sets $\mathcal{M}_1, \mathcal{M}_2$:

$$\mathcal{M}_1 \otimes \mathcal{M}_2 := \{\mu : \ \mu|_{(X_1,Y_1)} \in \mathcal{M}_1, \ \mu|_{(X_2,Y_2)} \in \mathcal{M}_2\} \supset \mathcal{M}_1 \times \mathcal{M}_2.$$

Theorem 4.2. *Let $T_1(x_1, y_1)$ and $T_2(x_2, y_2)$ be two tasks, let $\mathcal{M}_1$ and $\mathcal{M}_2$ be any sets of distributions over (x_1, y_1) and (x_2, y_2), respectively, and let $T := T_1 \times T_2$. Then:*

$$\mathsf{IC}(T, \mathcal{M}_1 \times \mathcal{M}_2) = \mathsf{IC}(T, \mathcal{M}_1 \otimes \mathcal{M}_2) = \mathsf{IC}(T_1, \mathcal{M}_1) + \mathsf{IC}(T_2, \mathcal{M}_2).$$

Proof. We prove the theorem by establishing three non-strict inequalities.

$\mathsf{IC}(T, \mathcal{M}_1 \times \mathcal{M}_2) \leq \mathsf{IC}(T, \mathcal{M}_1 \otimes \mathcal{M}_2)$. This is obvious, since $\mathcal{M}_1 \otimes \mathcal{M}_2 \supset \mathcal{M}_1 \times \mathcal{M}_2$, and thus the sup in Definition 4.1 is taken over a larger set.

$\mathsf{IC}(T, \mathcal{M}_1 \otimes \mathcal{M}_2) \leq \mathsf{IC}(T_1, \mathcal{M}_1) + \mathsf{IC}(T_2, \mathcal{M}_2)$. Let $\varepsilon > 0$ be an arbitrarily small parameter. Let π_1 and π_2 be two protocols that succeed at tasks T_1 and T_2, respectively, such that

$$\mathsf{IC}_{\mu_1}(\pi_1) < \mathsf{IC}(T_1, \mathcal{M}_1) + \varepsilon \text{ and } \mathsf{IC}_{\mu_2}(\pi_2) < \mathsf{IC}(T_2, \mathcal{M}_2) + \varepsilon,$$

for all $\mu_1 \in \mathcal{M}_1$ and $\mu_2 \in \mathcal{M}_2$. Let π be the protocol that on (the random) inputs (X_1, X_2), (Y_1, Y_2) independently runs π_1 on the pair (X_1, Y_1) and π_2 on the pair (X_2, Y_2). Then clearly the protocol π succeeds at the task $T = T_1 \times T_2$. It remains to analyze π's information cost. Let μ be a distribution in $\mathcal{M}_1 \otimes \mathcal{M}_2$. We will show that

$$\mathsf{IC}_\mu(\pi) < \mathsf{IC}(T_1, \mathcal{M}_1) + \mathsf{IC}(T_2, \mathcal{M}_2) + 2\varepsilon,$$

since $\varepsilon > 0$ is arbitrary, this will complete the proof. Denote $\mu_1 := \mu|_{(X_1,Y_1)} \in \mathcal{M}_1$, and $\mu_2 := \mu|_{(X_2,Y_2)} \in \mathcal{M}_2$. Let Π_1 be the random variable denoting the transcript of π_1, and similarly let Π_2 denote the transcript of π_2. We know that

$$I(X_1; \Pi_1 | Y_1) + I(Y_1; \Pi_1 | X_1) = \mathsf{IC}_{\mu_1}(\pi_1) < \mathsf{IC}(T_1, \mathcal{M}_1) + \varepsilon,$$

and similarly

$$I(X_2;\Pi_2|Y_2)+I(Y_2;\Pi_2|X_2)=\mathsf{IC}_{\mu_2}(\pi_2)<\mathsf{IC}\,(T_2,\mathcal{M}_2)+\varepsilon.$$

Next we note that the execution of π_1 only depends on the inputs X_1, Y_1, and the public/private randomness pertaining to the execution of π_1. This implies

$$\begin{aligned}&I(\Pi_1;X_2|X_1Y_1)=I(\Pi_1;Y_2|X_1Y_1)=I(\Pi_1;\Pi_2|X_1Y_1)=\\&I(\Pi_1;X_2Y_2|X_1Y_1)=I(\Pi_1;Y_2\Pi_2|X_1Y_1)=I(\Pi_1;X_2\Pi_2|X_1Y_1)=I(\Pi_1;X_2Y_2\Pi_2|X_1Y_1)=0,\end{aligned}\quad(15)$$

and similar equalities hold for Π_2.

By Proposition 2.9 with $A=X_1$, $B=\Pi_1$, $C=Y_1$, $D=Y_2$, and the fact that $I(\Pi_1;Y_2|X_1Y_1)=0$ we get

$$I(\Pi_1;X_1|Y_1Y_2)\le I(\Pi_1;X_1|Y_1).\quad(16)$$

Similarly, by Proposition 2.9 with $A=X_2$, $B=\Pi_2$, $C=Y_2$, $D=\Pi_1Y_1$, and the fact that $I(\Pi_2;\Pi_1Y_1|X_2Y_2)=0$ we get

$$I(\Pi_2;X_2|\Pi_1Y_1Y_2)\le I(\Pi_2;X_2|Y_2).\quad(17)$$

Putting these and (15) together, we obtain

$$\begin{aligned}&I(\Pi;X_1X_2|Y_1Y_2)=I(\Pi_1\Pi_2;X_1X_2|Y_1Y_2)=I(\Pi_1;X_1X_2|Y_1Y_2)+I(\Pi_2;X_1X_2|Y_1Y_2\Pi_1)=\\&I(\Pi_1;X_1|Y_1Y_2)+I(\Pi_1;X_2|X_1Y_1Y_2)+I(\Pi_2;X_2|Y_1Y_2\Pi_1)+I(\Pi_2;X_1|X_2Y_1Y_2\Pi_1)\overbrace{=}^{\text{by (15)}}\\&I(\Pi_1;X_1|Y_1Y_2)+I(\Pi_2;X_2|Y_1Y_2\Pi_1)\overbrace{\le}^{\text{by (16) and (17)}}I(\Pi_1;X_1|Y_1)+I(\Pi_2;X_2|Y_2).\end{aligned}$$

Similarly, $I(\Pi;Y_1Y_2|X_1X_2)\le I(\Pi_1;Y_1|X_1)+I(\Pi_2;Y_2|X_2)$. Thus

$$\begin{aligned}\mathsf{IC}_\mu(\pi)=I(\Pi;X_1X_2|Y_1Y_2)+I(\Pi;Y_1Y_2|X_1X_2)\le\\I(\Pi_1;X_1|Y_1)+I(\Pi_2;X_2|Y_2)+I(\Pi_1;Y_1|X_1)+I(\Pi_2;Y_2|X_2)=\\\mathsf{IC}_{\mu_1}(\Pi_1)+\mathsf{IC}_{\mu_2}(\Pi_2)<\mathsf{IC}\,(T_1,\mathcal{M}_1)+\mathsf{IC}\,(T_2,\mathcal{M}_2)+2\varepsilon.\end{aligned}$$

$\underline{\mathsf{IC}\,(T_1,\mathcal{M}_1)+\mathsf{IC}\,(T_2,\mathcal{M}_2)\le\mathsf{IC}\,(T,\mathcal{M}_1\times\mathcal{M}_2)}$. Let $\mu_1\in\mathcal{M}_1$ and $\mu_2\in\mathcal{M}_2$ be two distributions, and let $\varepsilon>0$ be a parameter. We will show that there are protocols π_1 and π_2 that succeed at tasks T_1 and T_2 respectively such that

$$\mathsf{IC}_{\mu_1}(\pi_1)+\mathsf{IC}_{\mu_2}(\pi_2)<\mathsf{IC}\,(T,\mathcal{M}_1\times\mathcal{M}_2)+\varepsilon.$$

By the definition of $\mathsf{IC}\,(T,\mathcal{M}_1\times\mathcal{M}_2)$, there is a protocol π that succeed at the task $T_1\times T_2$ such that

$$\mathsf{IC}_{\mu_1\times\mu_2}(\pi)<\mathsf{IC}\,(T,\mathcal{M}_1\times\mathcal{M}_2)+\varepsilon.$$

Define the protocols π_1 and π_2 as in Figure 3. From the definition of $T=T_1\times T_2$ it follows that the protocols π_1 and π_2 succeed at T_1 and T_2 respectively. It remains to analyze their information complexity.

Protocol $\pi_1(x_1, y_1)$

1. The parties jointly and publicly sample Y_2 according to $\mu_2|_{Y_2}$.
2. The first party privately samples $X_2 = X_2|Y_2$.
3. The parties run $\pi((x_1, X_2), (y_1, Y_2))$ and output the output of the task T_1.

Protocol $\pi_2(x_2, y_2)$

1. The parties jointly and publicly sample X_1 according to $\mu_1|_{X_1}$.
2. The second party privately samples $Y_1 = Y_1|X_1$.
3. The parties run $\pi((X_1, x_2), (Y_1, y_2))$ and output the output of the task T_2.

Figure 3: The protocols π_1 and π_2

We have

$$\mathsf{IC}_{\mu_1}(\pi_1) = I(\pi_1; X_1|Y_1) + I(\pi_1; Y_1|X_1) = I(\pi; X_1|Y_1Y_2) + I(\pi; Y_1|X_1X_2Y_2) \tag{18}$$

and

$$\mathsf{IC}_{\mu_2}(\pi_2) = I(\pi_2; X_2|Y_2) + I(\pi_2; Y_2|X_2) = I(\pi; X_2|X_1Y_1Y_2) + I(\pi; Y_1|X_1X_2). \tag{19}$$

Putting the two equations together we get:

$$\begin{aligned}
&\mathsf{IC}_{\mu_1}(\pi_1) + \mathsf{IC}_{\mu_2}(\pi_2) = \\
&\quad I(\pi; X_1|Y_1Y_2) + I(\pi; Y_1|X_1X_2Y_2) + I(\pi; X_2|X_1Y_1Y_2) + I(\pi; Y_1|X_1X_2) = \\
&\quad I(\pi; X_1|Y_1Y_2) + I(\pi; X_2|X_1Y_1Y_2) + I(\pi; Y_1|X_1X_2) + I(\pi; Y_1|X_1X_2Y_2) = \\
&\quad\quad I(\pi; X_1X_2|Y_1Y_2) + I(\pi; Y_1Y_2|X_1X_2) = \mathsf{IC}_{\mu_1\times\mu_2}(\pi) < \mathsf{IC}\left(T, \mathcal{M}_1 \times \mathcal{M}_2\right) + \varepsilon.
\end{aligned}$$

□

Theorem 4.2 can be now used in many different ways. Let us use it to show exact direct sum theorems for both the information cost and the distributional information cost.

Theorem 4.3. *Let $f(x, y)$ be any function, and $\rho \geq 0$ an error parameter. Let f^n be the problem of computing f on n pairs of inputs such that when one considers each coordinate separately, the error is bounded by ρ. Then*

1. $\mathsf{IC}_{\mathsf{D}}(f^n, \rho) = n \cdot \mathsf{IC}_{\mathsf{D}}(f, \rho)$.
2. $\mathsf{IC}\left(f^n, \rho\right) = n \cdot \mathsf{IC}\left(f, \rho\right)$.

Proof. We note that the proofs of the two parts are quite similar; the reader may want to only read the proof of the second part, which is more important (and also simpler).

$\underline{\mathsf{IC}_{\mathsf{D}}(f^n, \rho) \geq n \cdot \mathsf{IC}_{\mathsf{D}}(f, \rho)}$. Let μ be the distribution that realizes $\mathsf{IC}_{\mathsf{D}}(f, \rho)$. Let T^n be the task of computing n copies of f such that the error on each copy when measured against the distribution μ is $\leq \rho$. Further, let $\mathcal{M}^n = \{\mu^n\}$ be the set consisting of only one product distribution. Then by the definition of μ,

$$\mathsf{IC}_{\mathsf{D}}(f, \rho) = \mathsf{IC}\left(\mathcal{M}^1, T^1\right).$$

By Theorem 4.2 we have:

$$\mathsf{IC}_{\mathsf{D}}(f^n, \rho) \geq \mathsf{IC}\left(\mathcal{M}^n, T^n\right) = \mathsf{IC}\left(\mathcal{M}^{n-1}, T^{n-1}\right) + \mathsf{IC}\left(\mathcal{M}^1, T^1\right) = \ldots$$
$$= n \cdot \mathsf{IC}\left(\mathcal{M}^1, T^1\right) = n \cdot \mathsf{IC}_{\mathsf{D}}(f, \rho).$$

$\underline{\mathsf{IC}_{\mathsf{D}}(f^n, \rho) \leq n \cdot \mathsf{IC}_{\mathsf{D}}(f, \rho)}$. Let μ be the distribution on n-tuples of inputs that realizes $\mathsf{IC}_{\mathsf{D}}(f^n, \rho)$, and let μ_i be the restriction of μ to the i-th coordinate. In other words, any protocol that fails with probability $\leq \rho$ with respect to μ_i on the i-th coordinate must reveal at least $\mathsf{IC}_{\mathsf{D}}(f^n, \rho)$ information with respect to μ. Let T_i be the task of computing f correctly with error $\leq \rho$ with respect to the distribution μ_i. Let $T := T_1 \times T_2 \times \ldots \times T_n$. By the definition of the task we have

$$\mathsf{IC}\left(\{\mu\}, T\right) \geq \mathsf{IC}_{\mathsf{D}}(f^n, \rho).$$

Let $\mathcal{M}_i := \{\mu_i\}$ and $\mathcal{M} := \mathcal{M}_1 \otimes \mathcal{M}_2 \otimes \ldots \otimes \mathcal{M}_n$. Then $\mu \in \mathcal{M}$. By Theorem 4.2 we have:

$$\mathsf{IC}_{\mathsf{D}}(f^n, \rho) \leq \mathsf{IC}\left(\{\mu\}, T\right) \leq \mathsf{IC}\left(\mathcal{M}, T\right) = \mathsf{IC}\left(\mathcal{M}_1, T_1\right) + \ldots + \mathsf{IC}\left(\mathcal{M}_n, T_n\right) \leq n \cdot \mathsf{IC}_{\mathsf{D}}(f^n, \rho).$$

$\underline{\mathsf{IC}\left(f^n, \rho\right) = n \cdot \mathsf{IC}\left(f, \rho\right)}$. Let T^n be the task of computing n copies of f such that the (worst case) probability of error on each copy is $\leq \rho$. Let $\mathcal{M}^n$ be the set of all possible distributions over n-tuples of inputs. Then, by definition

$$\mathsf{IC}\left(f, \rho\right) = \mathsf{IC}\left(\mathcal{M}^1, T^1\right),$$

and

$$\mathsf{IC}\left(f^n, \rho\right) = \mathsf{IC}\left(\mathcal{M}^n, T^n\right).$$

Note that $\mathcal{M}^n = \underbrace{\mathcal{M}^1 \otimes \ldots \otimes \mathcal{M}^1}_{n \text{ times}}$ and $T^n = \underbrace{T^1 \times \ldots \times T^1}_{n \text{ times}}$, and thus by Theorem 4.2,

$$\mathsf{IC}\left(f^n, \rho\right) = \mathsf{IC}\left(\mathcal{M}^n, T^n\right) = n \cdot \mathsf{IC}\left(\mathcal{M}^1, T^1\right) = n \cdot \mathsf{IC}\left(f, \rho\right).$$

□

5 Information complexity vs. communication complexity

5.1 A new sampling lemma

In this section we prove a new sampling lemma. The lemma is then used to establish a new connection between the information and the communication complexity of any problem. We start with the following claim about the information divergence of distributions:

Claim 5.1. *Supposed that* $\mathbf{D}(\mu||\nu) \leq I$. *Let* ε *be any parameter. Then*

$$\mu\left\{x:\ 2^{(I+1)/\varepsilon}\cdot\nu(x) < \mu(x)\right\} < \varepsilon.$$

Proof. Recall that $\mathbf{D}(\mu||\nu) = \sum_{x\in\mathcal{U}}\mu(x)\log\frac{\mu(x)}{\nu(x)}$. Denote by $\mathcal{N} = \{x:\ \mu(x) < \nu(x)\}$ – the terms that contribute a negative amount to $\mathbf{D}(\mu||\nu)$. First we observe that for all $0 < x < 1$, $x\log x > -1$, and thus

$$\sum_{x\in\mathcal{N}}\mu(x)\log\frac{\mu(x)}{\nu(x)} = \sum_{x\in\mathcal{N}}\nu(x)\cdot\frac{\mu(x)}{\nu(x)}\log\frac{\mu(x)}{\nu(x)} \geq \sum_{x\in\mathcal{N}}\nu(x)\cdot(-1) > -1.$$

Denote by $\mathcal{L} = \left\{x:\ 2^{(I+1)/\varepsilon}\cdot\nu(x) < \mu(x)\right\}$; we need to show that $\mu(\mathcal{L}) < \varepsilon$. For each $x\in\mathcal{L}$ we have $\log\frac{\mu(x)}{\nu(x)} > (I+1)/\varepsilon$. Thus

$$I \geq \mathbf{D}(\mu||\nu) \geq \sum_{x\in\mathcal{L}}\mu(x)\log\frac{\mu(x)}{\nu(x)} + \sum_{x\in\mathcal{N}}\mu(x)\log\frac{\mu(x)}{\nu(x)} > \mu(\mathcal{L})\cdot(I+1)/\varepsilon - 1,$$

implying $\mu(\mathcal{L}) < \varepsilon$. □

We are now ready to state and prove our main sampling lemma.

Lemma 5.2. *Let* μ *be any distribution over a universe* $\mathcal{U}$ *and let* $I \geq 0$ *be a parameter that is known to both* A *and* B. *Further, let* ν_A *and* ν_B *be two distributions over* $\mathcal{U}$ *such that* $\mathbf{D}(\mu||\nu_A) \leq I$ *and* $\mathbf{D}(\mu||\nu_B) \leq I$. *The players are each given a real function* $p_A, q_A, p_B, q_B : \mathcal{U} \to [0,1]$ *such that for all* $x\in\mathcal{U}$, $\mu(x) = p_A(x)\cdot p_B(x)$, $\nu_A(x) = p_A(x)\cdot q_A(x)$, *and* $\nu_B(x) = p_B(x)\cdot q_B(x)$. *Let* $\varepsilon > 0$ *be an error parameter. Then there is a sampling protocol* $\Pi = \Pi(p_A, p_B, q_A, q_B, I, \varepsilon)$ *that communicates* $2^{O(1+I/\varepsilon)}$ *bits such that the following hold:*

1. *at the end of the protocol, the players output* $x_A\in\mathcal{U}$ *and* $x_B\in\mathcal{U}$, *respectively;*
2. *there is an event* $\mathcal{E}$ *such that* $\neg\mathcal{E} \Rightarrow x_A = x_B$ *and* $\mathbf{P}[\mathcal{E}] < \varepsilon$;
3. *let* μ' *is the distribution of* x_A *conditioned on* $\neg\mathcal{E}$, *then* $|\mu - \mu'| < \varepsilon$.

Proof. Firstly, Alice and Bob interpret the shared random tape as a source of points (x_i, α_i, β_i) uniformly distributed in $\mathcal{U}\times[0,1]\times[0,1]$. Alice and Bob consider $T = 2|\mathcal{U}|\ln 1/\varepsilon$ such points. Their goal will be to discover the first index τ such that $\alpha_\tau \leq p_A(x_\tau)$ and $\beta_\tau \leq p_B(x_\tau)$. Note that the probability of each x to be such x_τ is proportional to $p_A(x_\tau)\cdot p_B(x_\tau) = \mu(x_\tau)$. Thus the distribution of x_τ is correct.

Denote $B_A := \{x:\ 2^{8(I+1)/\varepsilon}\cdot\nu_A(x) < \mu(x)\}$ and $B_B := \{x:\ 2^{8(I+1)/\varepsilon}\cdot\nu_B(x) < \mu(x)\}$. Then by Claim 5.1, $\mu(B_A), \mu(B_B) < \varepsilon/8$. Next, we note that the probability that an index t satisfies $\alpha_t \leq p_A(x_t)$ and $\beta_t \leq p_B(x_t)$ is exactly $1/|\mathcal{U}|$. Hence the probability that $\tau > T$ (i.e. that x_τ is not among the T points considered) is bounded by

$$(1 - 1/|\mathcal{U}|)^T < e^{-T/|\mathcal{U}|} = e^{-2\ln 1/\varepsilon} = \varepsilon^2 < \varepsilon/16.$$

Denote by $\mathcal{A}$ the set of indices

$$\mathcal{A} := \{i \leq T:\ \alpha_i \leq p_A(x_i) \text{ and } \beta_i \leq 2^{8(I+1)/\varepsilon}\cdot q_A(x_i)\}.$$

$\mathcal{A}$ is the set of indices that are candidates to be τ from A's viewpoint. Similarly,

$$\mathcal{B} := \{i \le T : \ \alpha_i \le 2^{8(I+1)/\varepsilon} \cdot q_B(x_i) \text{ and } \beta_i \le p_B(x_i)\}.$$

Assuming $x_\tau \notin B_A \cup B_B$, we have that $\tau \in \mathcal{A} \cap \mathcal{B}$, because $x_\tau \notin B_A$ implies

$$\frac{p_B(x_\tau)}{q_A(x_\tau)} = \frac{\mu(x_\tau)}{\nu_A(x_\tau)} \le 2^{8(I+1)/\varepsilon},$$

and thus $\beta_\tau < p_B(x_\tau) \le 2^{8(I+1)/\varepsilon} \cdot q_A(x_\tau)$, and hence $\tau \in \mathcal{A}$. In fact, τ is the first element in $\mathcal{A} \cap \mathcal{B}$. Note that for each t, $\mathbf{P}[t \in \mathcal{A}] \le 2^{8(I+1)/\varepsilon}/|\mathcal{U}|$. Thus $\mathbf{E}[|\mathcal{A}|] \le 2^{8(I+1)/\varepsilon} \cdot 2 \ln 1/\varepsilon < 2^{9I/\varepsilon}$. Thus, by Chernoff bound,

$$\mathbf{P}[|\mathcal{A}| > 2^{10I/\varepsilon}] \ll \varepsilon/16.$$

Let the event $\mathcal{E}_1 := [x_\tau \in B_A \cup B_B]$, and $\mathcal{E}_2 := [\tau > T \text{ or } |\mathcal{A}| > 2^{10I/\varepsilon} \text{ or } |\mathcal{B}| > 2^{10I/\varepsilon}]$. Then by union bound, setting $\mathcal{E} := \mathcal{E}_1 \cup \mathcal{E}_2$, $\mathbf{P}[\mathcal{E}] < 2 \cdot \varepsilon/8 + 3 \cdot \varepsilon/16 < \varepsilon/2$. The distribution μ' conditioned on on $\neg(\mathcal{E}_1 \cup \mathcal{E}_2)$ satisfies $|\mu' - \mu| < \varepsilon/2$, because it is the distribution μ restricted to the set $\mathcal{U} \setminus (B_A \cup B_B)$ (note that only the restriction by $\neg\mathcal{E}_1$ biases the distribution). We will show a protocol, such that assuming the event $\neg(\mathcal{E}_1 \cup \mathcal{E}_2)$ the parties succeed at outputting the same correct value of x_τ with probability $> 1 - \varepsilon/2$, thus completing the proof.

We have reduced the problem to the problem of finding and outputting the first element in $\mathcal{A} \cap \mathcal{B}$, where $|\mathcal{A}|, |\mathcal{B}| \le 2^{10I/\varepsilon}$. The communication complexity of the protocol will be $2^{O(1+I/\varepsilon)}$. We note that the protocol can be organized in such a way that only Alice needs to communicate that many bits, while Bob communicates only $O(1 + I/\varepsilon)$ bits.

Information-cost only sampling protocol

1. Alice computes the set $\mathcal{A}$. If $|\mathcal{A}| > 2^{10I/\varepsilon}$ the protocol fails.
2. Bob computes the set $\mathcal{B}$. If $|\mathcal{B}| > 2^{10I/\varepsilon}$ the protocol fails.
3. For each $a \in \mathcal{A}$, Alice computes $d = \lceil 20I/\varepsilon + \log 1/\varepsilon + 2 \rceil$ random hash values $h_1(a), \ldots, h_d(a)$, where the hash functions are evaluated using public randomness.
4. Alice sends the values $\{h_j(a_i)\}_{a_i \in \mathcal{A},\ 1 \le j \le d}$ to Bob.
5. Bob finds the first index i such that there is a $b \in \mathcal{B}$ for which $h_j(b) = h_j(a_i)$ for $j = 1..d$ (if such an i exists). Bob outputs x_b and sends the index i to Alice.
6. Alice outputs x_{a_i}.

Figure 4: The main sampling protocol from Lemma 5.2

First note that the number of bits communicated by Alice is bounded by $2^{10I/\varepsilon} \cdot d = 2^{O(1+I/\varepsilon)}$. The number of bits communicated by Bob is bounded by $\log |\mathcal{A}| \le 10I/\varepsilon$. To see that the protocol works, observe that for $a \in \mathcal{A}$ and $b \in \mathcal{B}$ such that $a \ne b$, the probability (over possible choises of the hash functions) that $h_j(a) = h_j(b)$ for $j = 1..d$ is bounded by $2^{-d} < \frac{\varepsilon}{4|\mathcal{A}| \cdot |\mathcal{B}|}$. Thus, by union

bound, the probability that there is an $a \in cA$, $b \in \mathcal{B}$ such that $a \neq b$ but the hashes match, is bounded by $\varepsilon/4$. Assuming there are no such a and b, and there is a $\tau \in \mathcal{A} \cap \mathcal{B}$, the protocol is guaranteed to find it, completing the proof. □

5.2 Information vs. communication

In the language of communication complexity, Lemma 5.2 implies the following:

Theorem 5.3. *Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{Z}_K$ be any function, and let $\rho, \varepsilon > 0$ be error parameters:*

1. *For any distribution μ, $D^{\mu}_{\rho+\varepsilon}(f) = 2^{O(1+\mathsf{IC}_{\mu}(f,\rho)/\varepsilon^2)}$.*
2. *$R_{\rho+\varepsilon}(f) = 2^{O(1+\mathsf{IC}(f,\rho)/\varepsilon^2)}$.*

Theorem 5.3 can be used as a generic (if weak) tool to obtain unconditional lower bounds on the information complexity of problems. For example, the fact that $R_{1/3}(DISJ_n) = \Omega(n)$ immediately implies that $\mathsf{IC}(DISJ_n, 1/6) = \Omega(\log n)$. In section 7 we will see that in fact $\mathsf{IC}(DISJ_n, 1/6) = \Omega(n)$.

Proof. We begin with the first part of the theorem. Let μ be any distribution. Let π be a protocol that realizes the value $I_{\mu} := \mathsf{IC}_{\mu}(f, \rho)$. In other words, π has an error rate of at most ρ and information cost of at most I_{μ} with respect to μ. Denote by π_{xy} the random variable that represents that transcript π given the inputs (x, y), and by π_x (resp. π_y) the protocol conditioned on only the input x (resp. y). We denote by π_{XY} the transcripts where (X, Y) are also a pair of random variables. We know that

$$I_{\mu} = I(X; \pi_{XY}|Y) + I(Y; \pi_{XY}|X) = \mathbf{E}_{(x,y)\sim\mu}[\mathbf{D}(\pi_{xy}||\pi_x) + \mathbf{D}(\pi_{xy}||\pi_y)].$$

Thus by Markov inequality, with probability at least $1 - \varepsilon/2$ (with respect to μ) we will have $\mathbf{D}(\pi_{xy}||\pi_x) \leq 2I_{\mu}/\varepsilon$ and $\mathbf{D}(\pi_{xy}||\pi_y) \leq 2I_{\mu}/\varepsilon$. Let us now run the sampling algorithm from Lemma 5.2. With the distribution μ in Lemma 5.2 taken to be π_{xy}, the distribution ν_A taken to be π_x, the distribution ν_B taken to be π_y, I taken to be $2I_{\mu}/\varepsilon$, and the error parameter is taken to be $\varepsilon/4$.

At each node v of the protocol tree that is owned by player X let $p_0(v)$ and $p_1(v) = 1 - p_0(v)$ denote the probabilities that the next bit sent by X is 0 and 1, respectively. For nodes owned by player Y, let $q_0(v)$ and $q_1(v) = 1 - q_0(v)$ denote the probabilities that the next bit sent by Y is 0 and 1, respectively, *as estimated by player X given the input x*. For each leaf ℓ let $p_X(\ell)$ be the product of all the values of $p_b(v)$ from the nodes that are owned by X along the path from the root to ℓ; let $q_X(\ell)$ be the product of all the values of $q_b(v)$ from the nodes that are owned by Y along the path from the root to ℓ. The values $p_Y(\ell)$ and $q_Y(\ell)$ are defined similarly. For each ℓ we have $\mathbf{P}[\pi_{xy} = \ell] = p_X(\ell) \cdot p_Y(\ell)$, $\mathbf{P}[\pi_x = \ell] = p_X(\ell) \cdot q_X(\ell)$, and $\mathbf{P}[\pi_y = \ell] = p_Y(\ell) \cdot q_Y(\ell)$. Thus we can apply Lemma 5.2, to successfully obtain a sample transcript T such that the statistical distance $|T - \pi_{xy}| < \varepsilon/2$. T is obtained using $2^{O(1+I_{\mu}/\varepsilon^2)}$ communication. Let T_{out} be the final output of the transcript T, and π_{out} be the final output of the original protocol. $|T - \pi_{xy}| < \varepsilon/2$ implies that $\mathbf{P}[T_{out} \neq \pi_{out}] < \varepsilon/2$.

It remains to bound the error probability of our new protocol. In other words, we need to show that

$$\mathbf{P}_{(x,y)\sim\mu,\text{protocol randomness}}[T_{out} \neq f(x, y)] < \rho + \varepsilon.$$

The protocol above only works for pairs (x, y) where $\mathbf{D}(\pi_{xy}||\pi_x) \leq 2I_\mu/\varepsilon$ and $\mathbf{D}(\pi_{xy}||\pi_y) \leq 2I_\mu/\varepsilon$. Call such pairs "good". We saw that $\mathbf{P}_{(x,y)\sim\mu}[(x, y) \text{ is good}] > 1 - \varepsilon/2$. Thus we have:

$$\mathbf{P}[T_{out} \neq f(x, y)] \leq \mathbf{P}[(x, y) \text{ not good}] + \mathbf{P}[T_{out} \neq \pi_{out}|(x, y) \text{ is good}] + \mathbf{P}[\pi_{out} \neq f(x, y)] < \varepsilon/2 + \varepsilon/2 + \rho = \rho + \varepsilon.$$

To prove the second part of the theorem recall that by Yao's minimax theorem there is a distribution μ such that $D^\mu_{\rho+\varepsilon}(f) = R_{\rho+\varepsilon}(f)$. Since by definition $\mathsf{IC}_\mu(f, \rho) \leq \mathsf{IC}(f, \rho)$, we obtain

$$R_{\rho+\varepsilon}(f) = D^\mu_{\rho+\varepsilon}(f) = 2^{O(1+\mathsf{IC}_\mu(f,\rho)/\varepsilon^2)} \leq 2^{O(1+\mathsf{IC}(f,\rho)/\varepsilon^2)}.$$

□

6 Information complexity and amortized communication

6.1 Preliminaries: the distributional case

We next turn our attention to the relationship between the information cost of a problem and its amortized communication cost – the communication cost of solving n copies of the problem as n goes to ∞. In [BR10] such a connection has been established for the distributional setting. Here we extend the connection to the distribution-free setting. The extension is very similar to the original proof. In addition, we formulate a more careful statement of the result. This statement may be more useful in applications. For example, it will be useful in one of the proofs establishing the linear lower bound on the information cost of disjointness in Section 7. We will first repeat some definitions and theorems from prior works, before stating our main results.

Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{Z}_K$ be any function[1]. The distributional communication cost of k copies of f has been defined in [BR10] as follows:

Definition 6.1. Let μ be a distribution on $\mathcal{X} \times \mathcal{Y}$ and let $0 < \rho < 1$. We denote by $D^{\mu,n}_\rho(f^n)$ the distributional complexity of computing f on each of n independent pairs of inputs drawn from μ, with probability of failure at most ρ on each of the inputs.

Note that trivially $D^{\mu,n}_\rho(f^n) \leq n \cdot D^\mu_\rho(f)$ – this can be achieved by just running n copies of the protocol for f independently.

Remark 6.2. In Definition 6.1 we deliberately do not require the n-copy protocol to succeed with probability $1 - \rho$ on *all* n copies. This latter task may be qualitatively more difficult. Consider, for example, a scenario where f and μ are such that $D^\mu_{1/10}(f) \ll D^\mu_{1/100}(f)$. This may happen, for example, if 90% of the input pairs are "very easy", and thus $D^\mu_{1/10}(f)$ is very small, while the remaining 10% are "very hard", and thus $D^\mu_{1/100}(f)$ is very large. Then $D^{\mu,n}_{1/10}(f^n)$ as in Definition 6.1 will still be small. At the same time, computing f correctly on *all* n copies simultaneously (except with probability 10%) would require one so solve a lot of "very hard" instances, and would have a much higher communication complexity that cannot possibly be related to $D^\mu_{1/10}(f)$.

A version of the following theorem, giving a connection between (distributional) information cost and amortized communication complexity was proved in [BYJKS04]. The sharper statement presented below is from [BR10] (cf. also Theorem 4.3 above).

[1] As before, our results also hold for tasks and not just functions.

Theorem 6.3. *For every μ, f, ρ there exists a protocol τ computing f on inputs drawn from μ with probability of error at most ρ on each input and communication at most $D_\rho^{\mu,n}(f^n)$ such that $\mathsf{IC}_\mu(\tau) \leq \frac{D_\rho^{\mu,n}(f^n)}{n}$, and thus $\mathsf{IC}_\mu(f,\rho) \leq \frac{D_\rho^{\mu,n}(f^n)}{n}$.*

Thus the amortized communication complexity gives an upper bound on the information cost of the problem. For each fixed n, we do not know whether the inequality in Theorem 6.3 is tight or not, and it probably isn't: for $n = 1$ this is equivalent to the question whether $\mathsf{IC}_\mu(f,\rho) = D_\rho^\mu(f)$ – which remains open. However, in [BR10] it has been shown that *in the limit* the inequality is tight:

Theorem 6.4. *The distributional information cost is equal to amortized communication:*

$$\mathsf{IC}_\mu(f,\rho) = \lim_{n\to\infty} \frac{D_\rho^{\mu,n}(f^n)}{n}.$$

To prove Theorem 6.4 one needs to use a low-information protocol for one copy of f to produce a low-amortized communication cost protocol for n copies, when n is large. The main technical ingredient in the proof is the following protocol compression lemma:

Lemma 6.5. *[BR10] Let (x,y) be inputs to an r-round communication protocol π whose (internal) information cost is $I := \mathsf{IC}_\mu(\pi)$. Then for every $\varepsilon > 0$ there exists a protocol τ such that at the end of the protocol, each party outputs a transcript for π. Furthermore, there is an event G with $\mathbf{P}[G] > 1 - r\varepsilon$ such that conditioned on G, the expected communication of τ is $I + O(\sqrt{rI}+1) + 2r\log(1/\varepsilon)$, and both parties output the same transcript distributed according to $\pi(x,y)$.*

The lemma is proved in [BR10] using an iterated correlated sampling argument. We will return to some of the proof details in the next section. Our goal will be to establish the analogue of Theorem 6.4 for the randomized communication complexity.

6.2 The non-distributional case

Recall that $R_\rho(f)$ is the randomized communication complexity of f with error ρ. We know, by Yao's minimax theorem that $R_\rho(f) = \max_\mu D_\rho^\mu(f)$. We now give a definition of the randomized communication complexity of n copies of f. It is the distribution-free analogue of Definition 6.1.

Definition 6.6. Let $0 < \rho < 1$. We denote by $R_\rho^n(f^n)$ the randomized communication complexity of computing f so that on each set of n inputs the probability of failure on each of the inputs is at most ρ.

Our main result states that the prior-free information cost of f captures precisely f's amortized communication complexity:

Theorem 6.7. *For $\rho > 0$,*

$$\mathsf{IC}(f,\rho) = \lim_{n\to\infty} \frac{R_\rho^n(f^n)}{n}.$$

Remark 6.8. We note that Theorem 6.7 does not hold when $\rho = 0$, i.e. when the protocol is not allowed to err. Recall (Proposition 3.21) that $\mathsf{IC}(EQ,0) = O(1)$. At the same time, it is not hard to see that if the function EQ is on the space $\{0,1\}^m \times \{0,1\}^m$ then $R_\rho^n(f^n) = \Omega(m\cdot n)$ and thus the right-hand-side in Theorem 6.7 is $\Omega(m)$. The '$\leq$' direction of Theorem 6.7 holds even for $\rho = 0$. The other direction, however, requires the use of compression, such as Lemma 6.5. Such lemmas necessarily introduce a small amount of additional error. For each fixed $\rho > 0$, the additional error introduced can be made negligible, but this does not work for $\rho = 0$.

The proof of Theorem 6.7 will consist of proving inequalities in two directions. We will actually prove a slightly stronger statement in the '$\geq$' direction – showing that low information cost implies low amortized communication complexity. Here is the precise statement that we prove:

Theorem 6.9. *Let $f : X \times Y \to \{0,1\}$, and let $I := \mathsf{IC}(f,\rho)$ then for each $\delta_1, \delta_2 > 0$ there is an $N = N(f,\rho,\delta_1,\delta_2)$ such that for each $n \geq N$ there is a protocol $\pi_n = \pi_n((x_1,\ldots,x_n),(y_1,\ldots,y_n))$ for computing n instances of f. The protocol π_n will have communication complexity $< n \cdot I \cdot (1+\delta_1)$, and will have error $\lesssim \rho$ on each copy. Moreover, except with probability $< \delta_2$, the protocol errors will behave as if the n evaluations were independent.*

More precisely, let $Q : \{0,1\}^n \to \{0,1\}$ be any monotone function. Fix the inputs $(x_1,\ldots,x_n)$, $(y_1,\ldots,y_n)$, let $(f_1,\ldots,f_n) := (f(x_1,y_1),\ldots,f(x_n,y_n))$ and let $(p_1,\ldots,p_n)$ be the random variable representing π_n's output. Let $\mathbf{e} = (e_1,\ldots,e_n)$ be the "errors vector" – $e_i = \chi_{p_i \neq f_i}$. Let $\mathbf{b} = (b_1,\ldots,b_n)$ be a vector of independent Bernoulli variables $b_i \sim B_\rho$. Then

$$\mathbf{P}[Q(\mathbf{e}) = 1] \leq \mathbf{P}[Q(\mathbf{b}) = 1] + \delta_2.$$

Clearly, Theorem 6.9 (applied with Q being one coordinate indicator functions) implies that $R^n_{\rho+\delta_2}(f^n) \leq n \cdot I \cdot (1+\delta_1)$. Taking $\delta_1 \to 0$ and $n \to \infty$, for each $\delta_2 > 0$ we get:

$$\mathsf{IC}(f,\rho) \geq \lim_{n\to\infty} \frac{R^n_{\rho+\delta_2}(f^n)}{n}.$$

Substituting $\rho - \delta_2$ for ρ we get:

$$\mathsf{IC}(f,\rho-\delta_2) \geq \lim_{n\to\infty} \frac{R^n_{\rho}(f^n)}{n}.$$

Finally, by the continuity of $\mathsf{IC}(f,\rho)$ in ρ (Corollary 3.19), we have $\mathsf{IC}(f,\rho) = \lim_{\delta_2\to 0} \mathsf{IC}(f,\rho-\delta_2)$, and thus

$$\mathsf{IC}(f,\rho) \geq \lim_{n\to\infty} \frac{R^n_{\rho}(f^n)}{n},$$

which proves the '$\geq$' direction of Theorem 6.7. It remains to prove the '$\leq$' direction of Theorem 6.7 and Theorem 6.9.

Proof of the '$\leq$' direction of Theorem 6.7. It follows immediately from Part 2 of Theorem 4.3 that

$$\mathsf{IC}(f,\rho) = \frac{\mathsf{IC}(f^n,\rho)}{n} \leq \frac{R^n_\rho(f^n)}{n}.$$

The second inequality holds because the amount of communication is always an upper bound on information cost. □

The proof of the other direction of Theorem 6.7, i.e. Theorem 6.9, is unfortunately fairly complicated. It is easier to give a proof of the weaker statement that $\lim_{n\to\infty} \frac{R^n_\rho(f^n)}{n} = O(\mathsf{IC}(f,\rho))$ – if we didn't care to prove that the constant is 1. It would be interesting to see whether this proof can be simplified.

Proof of Theorem 6.9. Recall that f is a function on the space $X \times Y$. Let K be such that $|X|, |Y| \leq 2^K$. By the continuity of $\mathsf{IC}(f, \rho)$, there is a $\delta_3 > 0$ such that

$$\mathsf{IC}(f, \rho - \delta_3) < \mathsf{IC}(f, \rho) \cdot (1 + \delta_1/3).$$

By the definition of $\mathsf{IC}(f, \rho - \delta_3)$, there is a protocol π that on each pair of inputs succeeds except with probability $\leq \rho - \delta_3$, and such that for each distribution μ,

$$\mathsf{IC}_\mu(\pi) \leq \mathsf{IC}(f, \rho) \cdot (1 + \delta_1/3).$$

Our overall strategy will be to try and execute n copies of π in a communication-efficient way. We claim that it is possible. Set $\delta_4 := \min(\delta_2, \delta_3)/2$.

Claim 6.10. *For each sufficiently large n there is a protocol π_n that takes n instances of f as an input and has the following properties:*

1. *For each set of inputs, the statistical distance between the output of π_n and the output of π^n (i.e. an execution of n independent copies of π) is $< \delta_4/(2n^2)$.*
2. *The expected communication cost of π_n is $< n \cdot \mathsf{IC}(f, \rho) \cdot (1 + 2\delta_1/3)$.*
3. *The worst case communication cost of π_n is $< 100nK/\delta_1$.*

The bulk of the effort will go into proving Claim 6.10. Assuming Claim 6.10, note that the execution of π_n will be very similar to the execution of π^n. In turn, the errors in the execution of π^n are dominated by n independent Bernoulli random variables $B_{\rho-\delta_3} < B_\rho$, thus almost completing the proof of the theorem.

The reason the proof is not complete is the additional complication stemming from π^n using $< n \cdot \mathsf{IC}(f, \rho) \cdot (1 + 2\delta_1/3)$ communication *in expectation*, while we would like an upper bound on the worst-case communication complexity of f^n. We are able to prove the theorem with n replaced with n^3 by taking n^2 independent copies of π_n. The following claim thus completes the proof of the theorem.

Claim 6.11. *Let Π be the protocol that runs on n^3 pairs of inputs by dividing them into n^2 blocks of n pairs each and running π_n on each block. Further, the protocol Π is truncated (and fails) if it does not terminated after $< n^3 \cdot \mathsf{IC}(f, \rho) \cdot (1 + \delta_1)$ communication. Then for each set of inputs, the statistical distance between the output of Π and the output of π^{n^3} is $< \delta_4$.*

Proof. Fix a set of n^3 inputs. There are two sources of statistical distance between the output of Π and the output of π^{n^3}: the first one is due to one of the n^2 blocks being different under π_n than under π^n; the second is from the probability that the protocol Π fails altogether by not terminating in the allotted time. The probability of the first event is bounded by $n^2 \cdot \delta_4/(2n^2) = \delta_4/2$. It remains to bound probability of the second event by $\delta_4/2$.

Let T_i for $i = 1, \ldots, n^2$ denote the random variable representing the amount of communication used by the i-th copy of π_n during the execution of Π. Denote $T := \sum_{i=1}^{n^2} T_i$. Our goal is to show that

$$\mathbf{P}\left[T \geq n^3 \cdot \mathsf{IC}(f, \rho) \cdot (1 + \delta_1)\right] < \delta_4/2.$$

We know that T_i are i.i.d., $\mathbf{E}[T_i] < n \cdot \mathsf{IC}(f, \rho) \cdot (1 + 2\delta_1/3)$ and

$$Var(T_i) < \mathbf{E}[T_i] \cdot 100nK/\delta_1 < 200n^2 K \cdot \mathsf{IC}(f, \rho)/\delta_1.$$

Hence $\mathbf{E}[T] < n^3 \cdot \mathsf{IC}(f,\rho) \cdot (1 + 2\delta_1/3)$ and $Var(T) < 200n^4 K \cdot \mathsf{IC}(f,\rho)/\delta_1$. Thus, by Chebyshev's inequality, we get

$$\mathbf{P}\left[T \geq n^3 \cdot \mathsf{IC}(f,\rho) \cdot (1+\delta_1)\right] < \mathbf{P}\left[T > \mathbf{E}[T] + n^3 \cdot \mathsf{IC}(f,\rho) \cdot \delta_1/3\right] < \frac{200n^2 K \cdot \mathsf{IC}(f,\rho)/\delta_1}{(n^3 \cdot \mathsf{IC}(f,\rho) \cdot \delta_1/3)^2} < \delta_4/2,$$

for a sufficiently large n. □

Claim 6.11 implies that for each sufficiently large n,

$$\frac{R^n_\rho(f^{n^3})}{n^3} < \mathsf{IC}(f,\rho) \cdot (1+\delta_1),$$

and, further, that the other conclusions of Theorem 6.9 hold.

It remains to prove Claim 6.10

Proof of Claim 6.10. We first give a protocol π^n that satisfies the first condition and that has a slightly lower expected communication cost of $< n \cdot \mathsf{IC}(f,\rho) \cdot (1 + 3\delta_1/5)$. We then show how to modify it to satisfy the third condition without increasing the expected communication cost by too much.

Let r be the number of rounds in the protocols π, and let $\alpha > 0$ be a (small) parameter that we will set later. We let $\mathcal{G}$ be the following zero-sum game. The first player M produces a distribution μ on n-tuples of pairs of inputs. The second player T produces a (randomized) protocol τ. The payoff for the first player is the sum

$$P_M(\mu,\tau) := (1-\alpha) \cdot \frac{\mathbf{E}_{\mathbf{x},\mathbf{y}\sim\mu}|\tau(\mathbf{x},\mathbf{y})|}{n \cdot I \cdot (1+\delta_1/2)} + \frac{\mathbf{E}_{\mathbf{x},\mathbf{y}\sim\mu}|\tau(\mathbf{x},\mathbf{y}) - \pi^n(\mathbf{x},\mathbf{y})|}{\delta_4/(2n^2)}. \tag{20}$$

Here $|\tau(\mathbf{x},\mathbf{y})|$ denotes the expected communication cost of τ on inputs $(\mathbf{x},\mathbf{y})$, and $|\tau(\mathbf{x},\mathbf{y}) - \pi^n(\mathbf{x},\mathbf{y})|$ denotes the statistical distance between the two protocols on a given input. We first establish that for any large enough n the value of the game is bounded by 1:

$\underline{Val_M(\mathcal{G}) < 1}$: Let ν be any mixed strategy for player M. Denote by $\bar{\mu}$ the average distribution in ν: $\bar{\mu}(\mathbf{x},\mathbf{y}) = \mathbf{E}_{\mu\sim\nu}\mu(\mathbf{x},\mathbf{y})$. Since the payoff function is calculated in terms of expectations over $(\mathbf{x},\mathbf{y}) \sim \mu$, for any τ we have:

$$\mathbf{E}_{\mu\sim\nu} P_M(\mu,\tau) = P_M(\bar{\mu},\tau).$$

Thus it is enough to show that for each distribution μ there is a protocol τ such that $P_M(\mu,\tau) < 1$.

Fix a distribution μ. Let $\mu_1,\ldots,\mu_n$ be the projections of μ onto its n coordinates. As in the proof of the second part of Theorem 4.2, one can see that the information revealed by the protocol π^n that applies π to each copy independently has bounded information cost:

$$\mathsf{IC}_\mu(\pi^n) \leq \sum_{i=1}^n \mathsf{IC}_{\mu_i}(\pi) \leq n \cdot I \cdot (1+\delta_1/3). \tag{21}$$

Furthermore, as π, the protocol π^n is still an r-round protocol, independently of n. Let

$$\varepsilon := \frac{\alpha \cdot \delta_4}{2n^2 r}.$$

By Lemma 6.5 the protocol π^n can be simulated by a protocol τ so that the expected communication $\mathbf{E}_{\mathbf{x},\mathbf{y}\sim\mu}|\tau(\mathbf{x},\mathbf{y})|$ is bounded by

$$n \cdot I \cdot (1+\delta_1/3) + O(\sqrt{r \cdot n \cdot I \cdot (1+\delta_1/3)} + 1) + 2r\log(1/\varepsilon) < n \cdot I \cdot (1+\delta_1/2),$$

so that the statistical distance between the execution of π^n and the execution of τ is $< r \cdot \varepsilon$. Putting the pieces together, we get

$$P_M(\mu,\tau) < (1-\alpha) \cdot \frac{n \cdot I \cdot (1+\delta_1/2)}{n \cdot I \cdot (1+\delta_1/2)} + \frac{r \cdot \varepsilon}{\delta_4/(2n^2)} = (1-\alpha) + \alpha = 1.$$

By the Minimax Theorem, there is a distribution ν on protocols τ, such that for each distribution μ, $\mathbf{E}_{\tau\sim\nu}[P_M(\mu,\tau)] < 1$. This implies that the randomized protocol π_n obtained by executing a protocol τ that is distributed according to ν also satisfies $P_M(\mu,\pi_n) < 1$ for all μ. This is in particular true for each singleton $\mu = 1_{(\mathbf{x},\mathbf{y})}$. Let

$$\alpha < 1 - \frac{1+\delta_1/2}{1+3\delta_1/5}.$$

Then for each $(\mathbf{x},\mathbf{y})$, $P_M(1_{(\mathbf{x},\mathbf{y})},\pi_n) < 1$ implies that the statistical distance between the output of π_n and the output of π^n is bounded by

$$P_M(1_{(\mathbf{x},\mathbf{y})},\pi_n) \cdot \frac{\delta_4}{2n^2} < \frac{\delta_4}{2n^2}.$$

At the same time, the expected running time of π_n on $(\mathbf{x},\mathbf{y})$ is bounded by

$$P_M(1_{(\mathbf{x},\mathbf{y})},\pi_n) \cdot \frac{n \cdot I \cdot (1+\delta_1/2)}{1-\alpha} < n \cdot I \cdot (1+3\delta_1/5).$$

To finish the proof, it remains to show how to modify π_n so that its worst-case communication cost is bounded by $100nK/\delta_1$. We modify π_n as follows. We let it run as usual for $80nK/\delta_1$ bits. If the protocol π_n does not terminate in this many steps, the players will use $2nK$ bits of communication to completely exchange their inputs $(\mathbf{x},\mathbf{y})$. They then finish the execution of the protocol π_n using private randomness but without communicating. They can do this since now they can simulate each other's internal state.

The bound on the worst-case communication cost clearly holds. To bound the expected communication, note that by Markov's inequality the probability of the protocol reaching $80nK/\delta_1$ bits of communication is bounded by

$$\frac{\textbf{Expected Communication}}{80nK/\delta_1} < \frac{2nI}{80nK/\delta_1} = \frac{\delta_1 I}{40K}.$$

Thus the additional contribution to the communication cost caused by the modification is bounded by

$$\frac{\delta_1 I}{40K} \cdot (2nK) = \frac{\delta_1 nI}{20}.$$

Overall, the modified protocol has expected communication complexity of less than

$$n \cdot I \cdot (1+3\delta_1/5) + \frac{\delta_1 nI}{20} < n \cdot I \cdot (1+2\delta_1/3),$$

thus still satisfying the second condition. □

This concludes the proof of Theorem 6.9. □

7 The information complexity of disjointness

7.1 A linear lower bound, proof 1: using self-reducibility

Our goal in this section is to prove a linear lower bound on the information complexity of the disjointness function $DISJ_n : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ given by

$$DISJ_n((x_1,\ldots,x_n),(y_1,\ldots,y_n)) := \neg \bigvee_{i=1}^{n} (x_i \wedge y_i).$$

In our first proof, we will rely on the following classical result on the *communication complexity* of $DISJ_n$:

Theorem 7.1. *[KS92, Raz92]* $R_{1/6}(DISJ_n) = \Omega(n)$.

We do not have a general bound that would show that $\mathsf{IC}(DISJ_n, \rho) = \Omega(R_{1/6}(DISJ_n)) = \Omega(n)$ for some constant ρ. However, we are able to exploit the self-similar structure of the disjointness function to prove the linear lower bound directly.

Theorem 7.2. $\mathsf{IC}(DISJ_n, 1/2 - \varepsilon) = \Omega(n)$ *for all* $0 < \varepsilon < 1/2$.

Proof idea of Theorem 7.2. In the proof we combine theorems 6.9 and 7.1. By Theorem 7.1, we know that $R_{1/6}(DISJ_n) = \Omega(n)$. Let us say $R_{1/6}(DISJ_n) \geq n/50$ to be concrete. Suppose for contradiction that $DISJ_n$ had sublinear information cost. By repeating the protocol for disjointness a constant number of time, we obtain that $\mathsf{IC}(DISJ_n, 1/100) = o(n)$. Next we use the fact (Theorem 6.9) that information is equal to amortized communication to obtain a protocol π_N for N copies of disjointness on n bits each such that the probability of π_N to be wrong on each n-tuple is $< 1/100$. We use π_N to derive a protocol for $DISJ_{n \cdot N}$ by first running π_N (using communication $o(n \cdot N)$), and then only focus on the n-tuples where π_N returned 0 (i.e. which were determined to be non-disjoint). We expect there to be $\sim N/100$ mistakes, so we should expect to go over fewer than $N/75$ n-tuples that were incorrectly classified as non-disjoint before either terminating (and returning 1) or discovering an n-tuple that is not disjoint (and returning 0). Thus the total communication required to solve $DISJ_{n \cdot N}$ would be $< n \cdot N/75 + o(n \cdot N) < n \cdot N/50$, contradicting Theorem 7.1. □

Proof of Theorem 7.2. By Theorem 7.1 we know that there is a constant $0 < \alpha < 1$ such that

$$R_{1/6}(DISJ_n) > \alpha n \tag{22}$$

for all sufficiently large n. Assume, for the sake of contradiction that $\mathsf{IC}(DISJ_n, 1/2 - \varepsilon) = o(n)$ for some $0 < \varepsilon < 1/2$. By repeating the same protocol a constant number of times and taking the majority we obtain $\mathsf{IC}(DISJ_n, \gamma) = o(n)$ for all constant $\gamma > 0$. We fix $\gamma := \alpha/10$. Thus for a sufficiently large $n \geq n_0$ we have $\mathsf{IC}(DISJ_n, \alpha/10) < (\alpha/10) \cdot n$.

Fix an $n \geq n_0$. We apply Theorem 6.9 to $DISJ_n$ with $\rho = \alpha/10$, $\delta_1 = 1/4$, and $\delta_2 = 1/40$. There is an N_0 such that for all $N > N_0$ there is a protocol π_N for computing N copies of $DISJ_n$. π_N has communication complexity $< N \cdot (\alpha/10) \cdot n \cdot (1 + 1/4) = (\alpha/8) \cdot (N \cdot n)$. The probability of π_N being wrong on each input is $\leq \rho + \delta_2 < 1/8$. Moreover, the stronger property in Theorem 6.9 holds. We use the protocol π_N to give a very efficient protocol for $DISJ_{N \cdot n}$, which will contradict

The protocol $\Pi((x_1,\ldots,x_{N\cdot n}),(y_1,\ldots,y_{N\cdot n}))$

The players fist interpret the input as N blocks of n bits each. Denote the blocks by $(X_1,\ldots,X_N)$, $(Y_1,\ldots,Y_N)$.

1. The players run $\pi_N((X_1,\ldots,X_N),(Y_1,\ldots,Y_N))$ to obtain an output $(a_1,\ldots,a_N)$.
2. Let $i_1,\ldots,i_k$ be the indexes where $a_{i_j}=0$.
3. If $k \geq \alpha N/4$, **return** 0 ("not disjoint").
4. Otherwise, for each $j = 1..k$:
   (a) Run the brute force disjointness protocol on (X_{i_j},Y_{i_j}).
   (b) If it returns 0, **return** 0 ("not disjoint").
5. If none of the executions on (X_{i_j},Y_{i_j}) returned 0, **return** 1 ("disjoint").

Figure 5: The protocol Π

Theorem 7.1. The protocol Π is given on Figure 5. We next analyze the protocol's complexity and failure probability.

The communication complexity of Π. The first step of Π consists of an execution of π_N, and thus uses $< (\alpha/8)\cdot(N\cdot n)$ communication. The fourth step is only executed if $k < \alpha N/4$, in which case it is executed fewer than $\alpha N/4$ times to a total communication cost of $< (\alpha N/4)\cdot(n+1)$. Thus the total communication cost of Π is bounded by

$$(\alpha/8)\cdot(N\cdot n)+(\alpha N/4)\cdot(n+1) < (\alpha/2)\cdot(N\cdot n).$$

The failure probability of Π. Let $((x_1,\ldots,x_{N\cdot n}),(y_1,\ldots,y_{N\cdot n})) = (X_1,\ldots,X_N),(Y_1,\ldots,Y_N)$ be an input to the protocol. We need to show that the probability that Π computes $DISJ_{N\cdot n}$ incorrectly on this input is $< 1/6$. There are two cases to consider.

Case 1: $DISJ_{n\cdot N}((x_1,\ldots,x_{N\cdot n}),(y_1,\ldots,y_{N\cdot n}))=0$. In this case there is a coordinate ℓ such that $DISJ_n(X_\ell,Y_\ell)=0$. By the property of π_N we will have the first-step output $a_\ell=0$ except with probability $< 1/8$. Assuming $a_\ell=0$ note that the protocol Π is guaranteed to output 0, either in step 3 or in step 4 after considering the pair (X_ℓ,Y_ℓ). Thus the probability of an error in this case is $< 1/8 < 1/6$.

Case 2: $DISJ_{n\cdot N}((x_1,\ldots,x_{N\cdot n}),(y_1,\ldots,y_{N\cdot n}))=1$. Let $Q:\{0,1\}^N\to\{0,1\}$ be the monotone function that returns 1 if and only if the Hamming weight of the input is at least $\alpha N/4$. Let B_ρ be a Bernoulli variable with the probability of 1 being $\rho=\alpha/10$. By Chernoff bound, for a sufficiently large N,

$$\mathbf{P}[Q(B_\rho,\ldots,B_\rho)=1] < 1/10.$$

Thus by Theorem 6.9, the probability of π_N making at least $\alpha N/4$ errors is bounded by

$$\mathbf{P}[Q(B_\rho,\ldots,B_\rho)=1]+\delta_2 < 1/8.$$

Thus the probability of Π outputting 0 in step 3 is $< 1/8$. Note that if Π does not output 0 in step 3 it is guaranteed to output 1, since all the pairs (X_i, Y_i) are disjoint. Thus the probability of error in this case is also $< 1/8 < 1/6$.

We proved that $R_{1/6}(DISJ_{N \cdot n}) < (\alpha/2) \cdot (N \cdot n)$ for all sufficiently large N. This contradicts our assumption (22), and completes the proof of the theorem. □

7.2 Proof 2: using information cost direct sum (sketch)

The second proof will be much shorter and does not use the communication complexity lower bound for disjointness. Rather, it will use the information cost direct sum techniques as in Theorem 4.2. It will resemble the proof [Raz92, BYJKS04] that disjointness has linear communication complexity, but will work for information cost.

Let us define the task $T(x, y)$ on two bits $x, y \in \{0, 1\}$ of computing the AND function $x \wedge y$. Let μ be the distribution on $\{0, 1\}^2$ that is uniform on the set $\{(0, 0), (0, 1), (1, 0)\}$. Let μ' be the restriction of μ to one coordinate. In other words, μ' is a distribution on $\{0, 1\}$ such that $\mu'(0) = 2/3$ and $\mu'(1) = 1/3$. Suppose, for contradiction, that there was a protocol Π for solving disjointness on n-bit long strings with error δ and in $< \varepsilon n$ *information* cost, where δ and ε are very small constants.

In particular we have $\mathsf{IC}_{\mu^n}(\Pi) < \varepsilon n$. Let π be the protocol obtained by restricting Π to one coordinate as in Figure 6 (cf. [BR10]).

Protocol $\pi(x, y)$

1. The parties jointly and publicly sample a uniformly selected index $J \in \{1, \ldots, n\}$.
2. The parties publicly sample $X_1, \ldots, X_{J-1}, Y_{J+1}, \ldots, Y_n$ according to μ'.
3. The first party privately samples $X_{J+1}, \ldots, X_n$ and the second party privately samples $Y_1, \ldots, Y_{J-1}$ conditioned on the corresponding publicly sampled variables, so that each (X_i, Y_i) is distributed according to μ.
4. The parties run $\Pi((X_1, \ldots, X_{J-1}, x, X_{J+1}, \ldots, X_n), (Y_1, \ldots, Y_{J-1}, y, Y_{J+1}, \ldots, Y_n))$ and output the output its output.

Figure 6: The protocol $\pi(x, y)$, $x, y \in \{0, 1\}$

The proof from [BR10] (and also the proof of Theorem 4.2) shows that the information cost

$$\mathsf{IC}_{\mu}(\pi) = \frac{1}{n} \cdot \mathsf{IC}_{\mu^n}(\Pi) < \varepsilon.$$

At the same time we note that π can be used to solve the task $T(x, y)$. For each pair of inputs (x, y), and for each possible randomized selections in the protocol π, we have:

$$T(x, y) = x \wedge y = Disj_n((X_1, \ldots, X_{J-1}, x, X_{J+1}, \ldots, X_n), (Y_1, \ldots, Y_{J-1}, y, Y_{J+1}, \ldots, Y_n)) = \pi(x, y),$$

where the last equality holds with probability $> 1 - \delta$. We note that we measure the information revealed by π only against the distribution μ, whereas π will perform correctly on all four possible pairs of inputs. This is important to achieve a contradiction, since computing $x \wedge y$ correctly on inputs in the support of μ is trivial, as the function is identically 0 on this set.

It remains to show that a protocol that reveals almost no information over the distribution μ cannot be computing $x \wedge y$ correctly. This part is very similar to previous proofs of the communication lower bounds for disjointness, and we omit the details here.

8 Directions and open problems

In this section we outline some open problems surrounding the interactive information complexity. We group these problems by topic. Some of the problems are very concrete, while others take the form of a potential research direction.

8.1 Properties of the interactive information complexity

The first set of problems has to do with the properties of $\mathsf{IC}(f, \varepsilon)$, and its relationship with other communication complexity measures. First and foremost, we would like to know whether interactive computation can be compressed. In other words, whether the interactive information complexity is equal to the communication complexity of any function:

Problem 1. *Is it true that for all f, $\mathsf{IC}(f, \varepsilon) = \Omega(R(f, \varepsilon))$?*

Note that we know that $\mathsf{IC}(f, \varepsilon) \leq R(f, \varepsilon)$. An affirmative answer to Problem 1 would prove a strong direct sum theorem for communication complexity. If would also mean that it is impossible to solve problems that have high communication complexity without violating the (information-theoretic) privacy of the participants' inputs. A negative answer to Problem 1 would give an example of a problem that violate the direct sum conjecture for randomized communication complexity [BR10]. The only general result in the direction of Problem 1 that we have is Theorem 5.3, and it only gives a lower bound of the form $\mathsf{IC}(f, \varepsilon/2) = \Omega(\log R(f, \varepsilon))$ for constant $\varepsilon > 0$.

A less ambitious problem is compressing communication to the *external* information cost of the problem:

Problem 2. *Is it true that for all f, $\mathsf{IC}^{\mathsf{ext}}(f, \varepsilon) = \Omega(R(f, \varepsilon))$?*

As has been observed in [BBCR10], compressing a protocol to the external information cost can be much easier than compressing to the internal information cost. While an affirmative answer to Problem 2 would have no direct-sum implications, it would still mean that any protocol for a distribute function f that has high communication complexity must reveal a lot of information to an observer, and thus cannot be information-theoretically secure.

Also of interest is the relationship between $\mathsf{IC}(f, \varepsilon)$ and other quantities related to the communication complexity of f. One notable problem here is the relationship between the quantum communication complexity $Q(f, \varepsilon)$ and $\mathsf{IC}(f, \varepsilon)$. While we know that there is an exponential gap between $R(f, \varepsilon)$ and $Q(f, \varepsilon)$ [Raz99, KR11], it is not clear whether it carries over to the information complexity.

Problem 3. *What is the relationship between $Q(f, \varepsilon)$ and $\mathsf{IC}(f, \varepsilon)$? In particular are there problems for which $Q(f, \varepsilon) = O(\mathrm{polylog}(\mathsf{IC}(f, \varepsilon)))$?*

Note that answering Problem 1 may provide an answer to Problem 3.

Another interesting problem is whether the log-rank conjecture holds for the interactive information complexity. Such a conjecture would be easier to prove than the regular log-rank conjecture since the information complexity is always bounded by the communication complexity:

Problem 4. *Let $f(x,y)$ be a binary function, and let M_f be the corresponding matrix. Is it true that*

$$\mathsf{IC}(f,0) = O(\text{polylog}\, rank(M_f))?$$

More generally, for any ε, is it the case that

$$\mathsf{IC}(f,\varepsilon) = O(\text{polylog} \min_{\|M-M_f\|_\infty \leq \varepsilon} rank(M))?$$

Finally, we turn our attention to the following surprising problem. Given a truth table of a function f and the error parameter ε, it is not clear how to compute $\mathsf{IC}(f,\varepsilon)$. Indeed, it is not even clear that when viewed as a function, $\mathsf{IC}(f,\varepsilon)$ is computable:

Problem 5. *What is the computational complexity of the problem of evaluating the information complexity of a function f? In other words, given a truth table of $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$, and a rational $\varepsilon \geq 0$, what is the complexity of evaluating $\mathsf{IC}(f,\varepsilon)$? Is this function computable?*

While it is clear that successive estimates of $\mathsf{IC}(f,\varepsilon)$ from above can be constructed by enumerating all two-party protocols, it is not clear how to estimate the rate of convergence of these estimates so as to compute $\mathsf{IC}(f,\varepsilon)$.

8.2 The information complexity of specific problems

To improve our understanding of the information complexity of problems we need to understand which communication complexity techniques can be adapted to the more challenging information setting. This may reveal additional insights about the functions in question, and also help us make progress towards Problem 1 above. Here we mention three specific such problems. The first one is the Gap Hamming Distance problem, that has received much attention recently, and has been shown to have linear communication complexity [CR11, Vid11, She11]. Given two vectors x, y in $\{-1,+1\}^n$, the GHD_n problem is distinguishing between the case when $\langle x,y\rangle > \sqrt{n}$ and the case when $\langle x,y\rangle < -\sqrt{n}$:

Problem 6. *Is it true that $\mathsf{IC}(GHD_n, 1/3) = \Omega(n)$?*

The second problem is closely related to Problem 3 above. The Vector in Subspace Problem (VSP) is the (promise) problem that yields an exponential separation between the one-way quantum and randomized communication complexity given by Klartag and Regev [KR11]. VSP_n is defined as follows. Alice is given a unit vector $u \in \mathbf{R}^n$, and Bob is given a subspace $H \subset \mathbf{R}^n$ of dimension $n/2$ with the promise that either $u \in H$ or $u \in H^\perp$. Their goal is to decide which is the case. [KR11] showed an $\Omega(n^{1/3})$ lower bound for this problem, we ask whether a polynomial lower bound on the information cost of the problem holds.

Problem 7. *Is it true that $\mathsf{IC}(VSP_n, 1/3) = n^{\Omega(1)}$?*

Finally, we turn our attention to the zero-error information cost of concrete functions. The simplest functions to consider are 2-bit functions $f : \{0,1\} \times \{0,1\} \to \{0,1\}$. It is not too difficult to see that the information complexity of the two-bit exclusive-OR function is $\mathsf{IC}(XOR_2, 0) = 2$. We do not know, however, the precise information complexity even for the two-bit AND function AND_2:

Problem 8. *What is the value of* $\mathsf{IC}(AND_2, 0)$*?*

We know that the value is in the interval $(1, 2)$ – in fact, better estimates can be proved – but the precise value is unknown. Note that this value would also give the precise value of the amortized communication complexity of the two-bit AND. In addition, a solution to Problem 8 would probably shed some light on the computability of the information complexity (Problem 5).

8.3 Quantum information complexity

We next turn our attention to directions of interest. These are more open-ended than the problems above, and are less well-posed. The first one has to do with extending interactive information complexity into the quantum communication setting.

Problem 9. *What is the correct quantum analogue of* $\mathsf{IC}(f, \varepsilon)$*? What is its relationship with the quantum communication complexity? Is it an interesting quantity? In particular, is it always bounded by a constant?*

We note that it is quite possible that the quantum analogue of the interactive information complexity of boolean functions collapses to a constant due to the existence of reversible computing in the quantum world.

8.4 Multiparty information complexity

Perhaps the most natural extension to the present work one should consider is the extension to the communication setting with more than two players. There are several models for (randomized) multiparty communication complexity. They can be broadly split into the number-in-hand and the number-on-the-forehead models.

In the number-in-hand (NIH) model, each of the k players is given an input x_i, and the players need to compute a function $f(x_1, \ldots, x_k)$. The players may be communicating either through a blackboard that is visible to all players, or through private messages, depending on the model. Note that if k is constant, e.g. $k = 3$, then the two settings are equivalent up to a multiplicative constant. The NIH setting is usually very similar to the two-party setting. One communication complexity lower bound technique in this case is to partition the k players into two groups and show that a large number of bits must be exchanged between the two groups. This brings the natural question of extending the interactive information cost to the multiparty NIH setting:

Problem 10. *What is the "right" definition of the multiparty NIH information complexity, and what are its properties?*

It is quite plausible that a straight generalization of the information cost would work. One source of difficulty is that now there are potentially different kinds of "public" randomness – e.g. randomness that is accessible by some but not all players.

The number-on-the-forehead (NOF) model is notoriously more difficult. In this model there are still k players and k inputs $x_1, \ldots, x_k$, but each player i has access to *all* inputs but x_i – as if x_i was written on her forehead. Lower bounds in this model are notoriously more difficult. There is no straightforward reduction to the two-party case. Even the disjointness problem over $[n]$ with fewer than $\log n$ players proves to be a major challenge that has been resolved only recently [CA08, LS09]. Moreover, any explicit lower bound on the multiparty communication complexity in the NOF model with $(\log n)^{\omega(1)}$ players, would yield to superpolynomial circuit lower bounds for the class $\mathbf{ACC^0}$ of bounded-depth circuits with modulo gates [Yao90, BT91, HG90]. Thus proving lower bounds in the NOF model is a well-motivated problem.

Problem 11. *Is there a way to define information complexity in the multiparty NOF model, and what are its properties?*

Note that in this model one has to be particularly careful since, at least when there is randomness-on-the-forehead available, one can design protocols that reveal no information to the participants except for the output of the computation.

8.5 Continuous complexity measures

Finally, we take a higher level view on the relationship between the (two-party) information complexity of problems and their communication complexity. One way to look at $\mathsf{IC}(f, \varepsilon)$ is as a continuous relaxation of the communication complexity $R(f, \varepsilon)$: $\mathsf{IC}(f, \varepsilon) \leq R(f, \varepsilon)$, since communication is always an upper bound of information. On the other hand, as far as we know, $\mathsf{IC}(f, \varepsilon)$ may be smaller than $R(f, \varepsilon)$. The advantage comes from the fact that when considering *information*, each communication round may carry a non-integral amount of information – potentially much smaller than 1 bit. Thus if during the execution of a protocol a bit is sent that is not very "informative", $\mathsf{IC}(\cdot, \cdot)$ will measure the amount of information correctly, while the communication complexity of sending this bit is "rounded" to cost 1. One can argue that if $\mathsf{IC}(\cdot, \cdot)$ has nicer properties than $R(\cdot, \cdot)$ – such as an exact direct sum theorem – it is because $\mathsf{IC}(\cdot, \cdot)$ is non-integral and measures the interaction between the player more accurately.

This raises the open-ended question on whether there are reasonable (and useful) continuous "relaxations" of other classically integral complexity measures.

Problem 12. *Is there a good way to define continuous relaxations of the classically integral complexity measures such as circuit complexity and branching program size?*

Such continuous measures, may potentially lead to lower bounds that are currently obfuscated by the integrality of the complexity measures involved. More broadly, information-theoretic considerations may inspire lower bounds in these models. For example, the derandomization technique in [BRRY10] has been inspired by information-theoretic considerations.

References

[BBCR10] Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to compress interactive communication. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing*, 2010.

[BR10] Mark Braverman and Anup Rao. Information equals amortized communication. *CoRR*, abs/1106.3595, 2010.

[BRRY10] Mark Braverman, Anup Rao, Ran Raz, and Amir Yehudayoff. Pseudorandom generators for regular branching programs. In *FOCS*, pages 40–47, 2010.

[BT91] Richard Beigel and Jun Tarui. On acc. In *FOCS*, pages 783–792, 1991.

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

[CA08] Arkadev Chattopadhyay and Anil Ada. Multiparty communication complexity of disjointness. *Electronic Colloquium on Computational Complexity (ECCC)*, 15(002), 2008.

[CR11] Amit Chakrabarti and Oded Regev. An optimal lower bound on the communication complexity of gap-hamming-distance. In *STOC*, pages 51–60, 2011.

[CSWY01] Amit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In Bob Werner, editor, *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, pages 270–278, Los Alamitos, CA, October 14–17 2001. IEEE Computer Society.

[CT91] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley series in telecommunications. J. Wiley and Sons, New York, 1991.

[DW07] M. Dietzfelbinger and H. Wunderlich. A characterization of average case communication complexity. *Information processing letters*, 101(6):245–249, 2007.

[FKNN95] Tomàs Feder, Eyal Kushilevitz, Moni Naor, and Noam Nisan. Amortized communication complexity. *SIAM Journal on Computing*, 24(4):736–750, 1995. Prelim version by Feder, Kushilevitz, Naor FOCS 1991.

[HG90] Johan Håstad and Mikael Goldmann. On the power of small-depth threshold circuits. In *FOCS*, pages 610–618, 1990.

[HJMR07] Prahladh Harsha, Rahul Jain, David A. McAllester, and Jaikumar Radhakrishnan. The communication complexity of correlation. In *IEEE Conference on Computational Complexity*, pages 10–23. IEEE Computer Society, 2007.

[Jai11] Rahul Jain. New strong direct product results in communication complexity. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:24, 2011.

[JKR09] T. S. Jayram, Swastik Kopparty, and Prasad Raghavendra. On the communication complexity of read-once ac^0 formulae. In *IEEE Conference on Computational Complexity*, pages 329–340, 2009.

[JKZ10] Rahul Jain, Hartmut Klauck, and Shengyu Zhang. Depth-independent lower bounds on the communication complexity of read-once boolean formulas. In *COCOON*, pages 54–59, 2010.

[JRS03] Rahul Jain, Jaikumar Radhakrishnan, and Pranab Sen. A direct sum theorem in communication complexity via message compression. In Jos C. M. Baeten, Jan Karel Lenstra, Joachim Parrow, and Gerhard J. Woeginger, editors, *ICALP*, volume 2719 of *Lecture Notes in Computer Science*, pages 300–315. Springer, 2003.

[Kla10] Hartmut Klauck. A strong direct product theorem for disjointness. In Leonard J. Schulman, editor, *STOC*, pages 77–86. ACM, 2010.

[KN97] Eyal Kushilevitz and Noam Nisan. *Communication complexity*. Cambridge University Press, Cambridge, 1997.

[KR11] Bo'az Klartag and Oded Regev. Quantum one-way communication can be exponentially stronger than classical communication. In *STOC*, pages 31–40, 2011.

[KS92] Bala Kalyanasundaram and Georg Schnitger. The probabilistic communication complexity of set intersection. *SIAM Journal on Discrete Mathematics*, 5(4):545–557, November 1992.

[LS09] Troy Lee and Adi Shraibman. Disjointness is hard in the multiparty number-on-the-forehead model. *Computational Complexity*, 18(2):309–336, 2009.

[LS10] Nikos Leonardos and Michael Saks. Lower bounds on the randomized communication complexity of read-once functions. *Computational Complexity*, 19(2):153–181, 2010.

[Raz92] Alexander Razborov. On the distributed complexity of disjointness. *TCS: Theoretical Computer Science*, 106, 1992.

[Raz99] Ran Raz. Exponential separation of quantum and classical communication complexity. In *STOC*, pages 358–367, 1999.

[Sha48] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27, 1948. Monograph B-1598.

[Sha03] Ronen Shaltiel. Towards proving strong direct product theorems. *Computational Complexity*, 12(1-2):1–22, 2003. Prelim version CCC 2001.

[She11] Alexander A. Sherstov. The communication complexity of gap hamming distance. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:63, 2011.

[SW73] David Slepian and Jack K. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19(4):471–480, July 1973.

[Vid11] Thomas Vidick. A concentration inequality for the overlap of a vector on a large set, with application to the communication complexity of the gap-hamming-distance problem. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:51, 2011.

[Yao90] Andrew Chi-Chih Yao. On acc and threshold circuits. In *FOCS*, pages 619–627, 1990.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de