Information Equals Amortized Communication*

Mark Braverman† Anup Rao‡

January 30, 2013

Abstract

We show how to efficiently simulate the sending of a *single* message M to a receiver who has partial information about the message, so that the expected number of bits communicated in the simulation is close to the amount of additional information that the message reveals to the receiver. This is a generalization and strengthening of the Slepian-Wolf theorem, which shows how to carry out such a simulation with low *amortized* communication in the case that M is a deterministic function of X. A caveat is that our simulation is interactive.

As a consequence, we prove that the internal information cost (namely the information revealed to the parties) involved in computing any relation or function using a two party interactive protocol is *exactly* equal to the amortized communication complexity of computing independent copies of the same relation or function. We also show that the only way to prove a strong direct sum theorem for randomized communication complexity is by solving a particular variant of the pointer jumping problem that we define. Our work implies that a strong direct sum theorem for communication complexity holds if and only if efficient compression of communication protocols is possible.

*The paper has been previously presented in part at the FOCS'11 conference [BR11].

†Princeton University, mbraverm@cs.princeton.edu. This material is based upon work partially supported by the National Science Foundation under Grant No. 1149888, and the Alfred P. Sloan Research Fellowship.

‡University of Washington, anuprao@cs.washington.edu. Supported by the National Science Foundation under agreement CCF-1016565.

1 Introduction

Suppose a sender wants to transmit a message M that is correlated with an input X to a receiver that has some information Y about X. What is the best way to carry out the communication in order to minimize the expected number of bits transmitted? A natural lower bound for this problem is the mutual information between the message and X, given Y: $I(M; X|Y)$, i.e. the amount of new information M reveals to the receiver about X. In this work, we give an interactive protocol that has the same effect as sending M, yet the expected number of bits communicated is asymptotically close to optimal — it is the same as the amount of new information that the receiving party learns from M, up to a sublinear additive term[1].

Our result is a generalization of classical data compression, where Y is empty (or constant), and M is a deterministic function of X. In this case, the information learnt by the receiver is equal to the entropy $H(M)$, and the compression result above corresponds to classical results on data compression first considered by Shannon [Sha48] — M can be encoded so that the expected number of bits required to transmit M is $H(M) + 1$ (see for example the text [CT91]).

Typical work in information theory focuses on the easier problem of communicating n independent copies $M_1, \ldots, M_n$, where each M_i has an associated dependent X_i, Y_i. Here n is viewed as a growing parameter, and the average communication is measured. Indeed, any solution simulating a single message can be applied to simulate the transmission of n messages, but there is no clear way to use an asymptotically good solution to compress a single message. By the asymptotic equipartition property of the entropy function, taking independent copies essentially forces most of the probability mass of the distributions to be concentrated on sets of the "right" size, which simplifies this kind of problem significantly. The Slepian-Wolf theorem [SW73] addresses the case when M is determined by X. The theorem states that there is a way to encode many independent copies $M_1, \ldots, M_n$ using roughly $I(M; X|Y)$ on average, as n tends to infinity. The theorem and its proof do not immediately give any result for communicating a single message. Other work has focused on the problem of generating two correlated random variables with minimal communication [Cuf08], and understanding the minimal amount of information needed to break the dependence between X, Y [Wyn75], neither of which seem useful to the problem we are interested in here.

Motivated by questions in computer science, prior works have considered the problem of encoding a single message where M is not necessarily determined by X (see [JRS03, HJMR07] and the references there), but these works do not handle the case above, where the receiver has some partial information about the sender's message.

2 Consequences in Communication Complexity

Given a function $f(x, y)$, and a distribution μ on inputs to f, there are several ways to measure the complexity of a communication protocol that computes f.

- The communication complexity D^{μ}_{ρ}, namely the maximum number of bits communicated by a protocol that computes f correctly except with probability ρ.

[1] Observe that if X, Y, M are arbitrary random variables, and the two parties are tasked with sampling M efficiently (as opposed to one party transmitting and the other receiving), it is impossible to succeed in communication comparable to the information revealed by M. For example, if $M = f(X, Y)$, where f is a boolean function with high communication complexity on average for X, Y, M reveals only one bit of information about the inputs, yet cannot be cheaply sampled.

- The amortized communication complexity, $\lim_{n\to\infty} D_\rho^{\mu,n}/n$, where here $D_\rho^{\mu,n}$ denotes the communication involved in the best protocol that computes f on n independent pairs of inputs drawn from μ, getting the answer correct except with probability ρ in each coordinate.

Let $\pi(X,Y)$ denote the public randomness and messages exchanged when the protocol π is run with inputs X,Y drawn from μ. Another set of measures arises when one considers exactly how much information is revealed by a protocol that computes f.

- The minimum amount of information that must be learnt about the inputs by an observer who watches an execution of any protocol ($I(XY;\pi(X,Y))$) that compute f except with probability of failure ρ, called the *external information cost* in [BBCR10].

- The minimum amount of new information that the parties learn about each others input by executing any protocol ($I(X;\pi(X,Y)|Y) + I(Y;\pi(X,Y)|X)$) that computes f except with probability of failure ρ, called the *internal information cost* in [BBCR10]. In this paper we denote this quantity $\mathsf{IC}^{\mathsf{i}}_\mu(f,\rho)$.

- The amortized versions of the above measures, namely the average external/internal information cost of a protocol that computes f on n independent inputs correctly except with probability ρ in each coordinate.

Determining the exact relationship between the amortized communication complexity and the communication complexity of the function is usually referred to as the *direct sum* problem, which has been the focus of much work [CSWY01, Sha01, JRS03, HJMR07, BBCR10, Kla10]. For randomized and average case complexity, we know that n copies must take approximately (at least) $\sqrt{n}$ times the communication of one copy, as shown by the authors with Barak and Chen [BBCR10]. For worst case (deterministic) communication complexity, Feder, Kushilevitz, Naor, and Nisan [FKNN91] showed that if a single copy of a function f requires C bits of communication, then n copies require $\Omega(\sqrt{C}n)$ bits. In the rest of the discussion in this paper, we focus on the average case and randomized communication complexity.

The proofs of the results above for randomized communication complexity have a lot to do with the information theory based measures for the complexity of communication protocols. Chakrabarti, Shi, Wirth and Yao [CSWY01] were the first to define the external information cost, and prove that if the inputs are independent in μ, then the external information cost of f is at most the amortized communication complexity of f. This sparked an effort to relate the amortized communication complexity to the communication complexity. If one could compress any protocol so that the communication in it is bounded by the external information cost, then, at least for product distributions μ, one would show that the two measures of communication complexity are the same.

For the case of general distributions μ, it was shown in [BYJKS04, BBCR10] that the amortized communication complexity can only be larger than the internal information cost. In fact, the internal and external information costs are the same when μ is a product distribution, so the internal information cost appears to be the appropriate measure for this purpose. [BBCR10] gave a way to compress protocols so that the communication is reduced to the geometric mean of the internal information and the communication in the protocol, which gave the direct sum result discussed above.

The main challenge that remains is to find a more efficient way to compress protocols whose internal information cost is small. Indeed, as we discuss below, in this paper we show that this

is essentially the *only* way to make progress on the direct sum question, in the sense that if some protocol cannot be compressed well, then it can be used to define a function whose amortized communication complexity is significantly smaller than its communication complexity.

2.1 Our Results

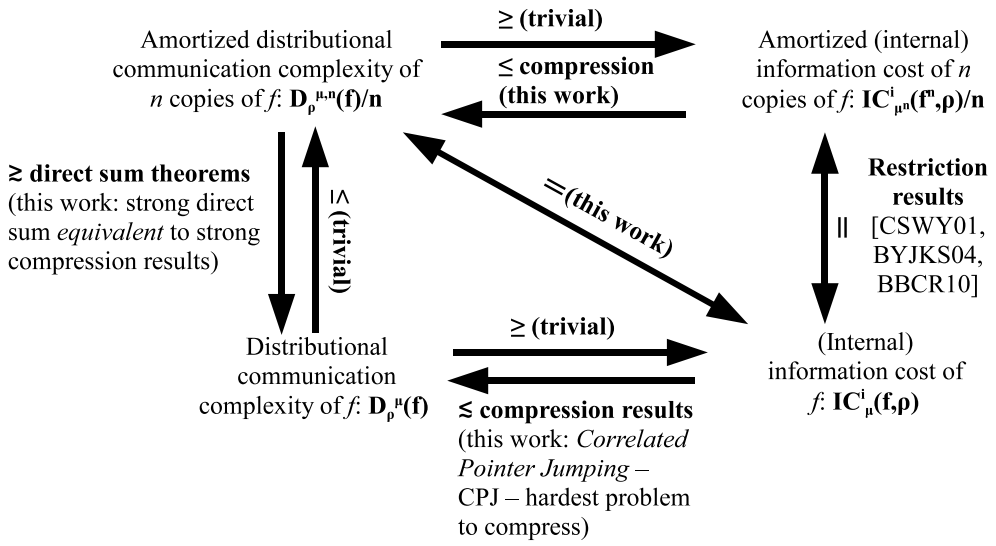

Figure 1: The relationships between different measures of complexity for communication problems, with the new results highlighted. The present work collapses the upper-right triangle in the diagram, showing that amortized communication complexity is equal to the internal information cost of any functionality. We further show three equivalent characterizations that would lead to a collapse in the lower-left triangle: strong direct sum theorems, near-optimal protocol compression and solving the Correlated Pointer Jumping efficiently.

Our main technical result is a way to compress one round protocols according to the internal information cost, which corresponds to the problem of efficiently communicating information when the receiver has some partial information, discussed in the introduction. In fact, we design a protocol that solves a harder problem, that we describe next. We give a way for two parties to efficiently sample from a distribution P that only one of them knows, by taking advantage of a distribution Q known only to the other. We obtain a protocol whose communication complexity can be bounded in terms of the informational divergence $\mathbf{D}(P||Q) = \sum_x P(x)\log(P(x)/Q(x))$.

Theorem 2.1. *Suppose that player A is given a distribution (described by the probabilities assigned to each point) P and player B is given a distribution Q, both over the same universe $\mathcal{U}$. There is a public coin protocol that uses an expected*

$$\mathbf{D}(P||Q) + 2\log(1/\varepsilon) + O\left(\sqrt{\mathbf{D}(P||Q)} + 1\right)$$

bits of communication such that at the end of the protocol:

- *Player A outputs an element a distributed according to P;*
- *Player B outputs b such that for each $x \in \mathcal{U}$, $\mathbf{P}[b = x|\ a = x] > 1 - \varepsilon$.*

As a corollary, we obtain the formulation discussed earlier. For any distribution X, Y and message M that is independent of Y once X is fixed, we can have the sender set P to be the distribution of M conditioned on her input x, and the receiver set Q to be the distribution of M conditioned on her input y. The expected divergence $\mathbf{D}(P||Q)$ turns out to be equal to the mutual information $I(M; X|Y)$. Indeed, applying Theorem 2.1 to each round of communication in a multiround protocol, gives the following corollary, where setting $r = 1$ gives the formulation discussed in the introduction. The proof appears in Section 6.1.

Corollary 2.2. *Let X, Y be inputs to a k round communication protocol π whose internal information cost is I. 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 $\mathbf{P}[G] > 1 - k\varepsilon$ such that conditioned on G, the expected communication of τ is $I + O(\sqrt{kI} + k) + 2k\log(1/\varepsilon)$, and both parties output the same transcript distributed exactly according to $\pi(X, Y)$.*

This compression scheme significantly clarifies the relationship between the various measures of complexity discussed in the introduction. In particular, it allows us to prove that the internal information cost of computing a function f according to a fixed distribution is *exactly* equal to the amortized communication complexity of computing many copies of f.

Theorem 2.3. *For any f, μ, and ρ,*

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

This result seems surprising to us, since it characterizes the information cost in terms of a quantity that at first seems to have no direct connection to information theory. The proof appears in Section 6.2. It proves that if a function's information cost is smaller than its communication complexity, then multiple copies of the function can be computed more efficiently in parallel than sequentially. Observe that the naive sequential protocol for computing multiple copies would only give a bound on the error in each copy separately (exactly as in our definition of amortized communication complexity). The consequences to the various measures discussed earlier are summarized in Figure 1.

In Section 6.3, we define a communication problem we call Correlated Pointer Jumping – $\mathsf{CPJ}(C, I)$ – that is parametrized by two parameters C and I such that $C \gg I$. $\mathsf{CPJ}(C, I)$ is designed in a way that the randomized communication complexity cost $I \leq R(\mathsf{CPJ}(C, I)) \leq C$. We show that determining the worst case randomized communication complexity $R(\mathsf{CPJ}(C, I))$ for $I = C/n$ is equivalent (up to poly-logarithmic factors) to determining the best parameter $k(n)$ for which a direct sum theorem $R(f^n) = \Omega(k(n) \cdot R(f))$ holds. For simplicity, we formulate only part of the result here.

Theorem 2.4. *If $\mathsf{R}(\mathsf{CPJ}(C, C/n)) = \tilde{O}(C/n)$ for all C, then a near optimal direct sum theorem holds: $\mathsf{R}(f^n) = \tilde{\Omega}(n \cdot \mathsf{R}(f))$ for* all *f.*

On the other hand, if $\mathsf{R}(\mathsf{CPJ}(C, C/n)) = \Omega((C \log^a C)/n)$ for all $a > 0$, then direct sum is violated by $\mathsf{CPJ}(C, C/n)$:

$$\mathsf{R}(\mathsf{CPJ}(C, C/n)^n) = O(C \log C) = o(n \cdot \mathsf{R}(\mathsf{CPJ}(C, C/n))/\log^a C),$$

for all a.

Finally, letting f^n denote the function that computes n copies of f on n different inputs, our protocol compression yields the following direct sum theorem:

Corollary 2.5 (Direct Sum for Bounded Rounds)**.** *Let C be the communication complexity of the best protocol for computing f with error ρ on inputs drawn from μ. Then any r round protocol computing f^n on the distribution μ^n with error $\rho - \varepsilon$ must involve at least $\Omega(n(C - r\log(1/\varepsilon) - O(\sqrt{C \cdot r})))$ communication.*

2.2 Techniques

The key technical contribution of our work is a sampling protocol that proves Theorem 2.1. The sampling method we show is different from the "Correlated Sampling" technique used in work on parallel repetition [Hol07, Rao08] and in the previous paper on compression [BBCR10]. In those contexts it was guaranteed that the input distributions P, Q are close in *statistical distance*. In this case, the sampling can be done without any communication. In our case, all interesting inputs P, Q are very far from each other in statistical distance, and not only that, but the ratios of the probabilities $P(x)/Q(x)$ may vary greatly with the choice of x. It is impossible to solve this problem without communication, and we believe it is unlikely that it can be solved without interaction.

Indeed, our sampling method involves interaction between the parties, and for good reasons. In the case that the sample is x for which $P(x)/Q(x)$ is very large, one would expect that a lot of communication is needed to sample x, since the second party would be surprised with this sample, while if $P(x)/Q(x)$ is small, then one would expect that a small amount of communication is sufficient. Our protocol operates in rounds, gradually increasing the number of bits that are communicated until the sample is correctly determined.

To illustrate our construction, consider the baby case of the problem where the issue of high variance in $P(x)/Q(x)$ does not affect us. Recall that the informational divergence $\mathbf{D}(P||Q)$ is equal to $\sum_x P(x)\log\frac{P(x)}{Q(x)}$. Suppose Q is the uniform distribution on some subset S_Q of the universe $\mathcal{U}$, and P is the uniform distribution on some subset $S_P \subset S_Q$. Then the informational divergence $\mathbf{D}(P||Q)$ is exactly $\log(|S_Q|/|S_P|)$.

In this case, the players use an infinite public random tape that samples an infinite sequence of elements $a_1, a_2, \ldots$ uniformly at random from the universe $\mathcal{U}$. Player A then picks the first element x that lies in S_P to be his sample. Next the players use the public randomness to sample a sequence of uniformly random boolean functions on the universe. A then sends a stream of these functions evaluated at x. At each round i, player B finds the first element y_i on the tape that belongs to S_Q and is consistent with the values Player A has sent so far. Player B uses y_i as his working hypothesis for the element Player A is trying to communicate. Player B lets Player A know (and outputs y_i) if the element y_i stays the same for some interval $i = [j..2j + \log 1/\varepsilon]$. That is, when the hypothesis for the element x stops changing. For the analysis, one has to note that the (expected) number of elements that B will have to reject before converging to x is bounded in terms of $\log|S_Q|/|S_P|$ – the divergence between P and Q in this case.

2.3 Subsequent and related work

Here we briefly discussed follow-up and related works that appeared since the publication of the preliminary version of this result. The most closely related is the paper by Braverman [Bra12a], which further generalized the information = amortized communication connection to *prior-free*

problems. It turns out that the natural prior-distribution-free analogue of information complexity is equal to the amortized randomized communication complexity (i.e. the complexity of protocols that are required to solve the problem with low error on *all* inputs). The compression protocol from our work was an ingredient in obtaining a strong direct product theorem for randomized bounded-round communication complexity [JPY12]. More recently, compression tools from the present paper were used in obtaining tight bounds on the communication complexity of Disjointness and other related functions [BGPW12].

The connection established in the present paper is between information complexity and the amortized communication complexity when some (possibly negligible) amount of error is allowed. The relation is no longer valid if the communication is not allowed to make any errors: while information complexity is continuous in the error parameter [BGPW12], the communication complexity may increase dramatically if errors are not allowed (see e.g. [Orl90]). It is conjectured that, in fact, the *external* information complexity captures the amortized zero-error communication complexity. Further discussion on this conjecture can be found in [Bra12b].

Coding of messages with a shared prior distribution in the one-shot setting has been recently considered in the context of explaining ambiguity in natural language [JKKS11]. While the high-level goal of this line of work is similar, the technical content is quite different and roughly corresponds to one-way worst-case compression (as opposed to multi-round average case). Another interesting set of questions arises from considering the role randomness plays in low-information protocols and in compression. Our compression scheme suggests that private randomness is not necessary to achieve a low-information protocol. The role of private randomness in low-information protocols has been recently studied [BBK$^+$12]. The extent to which shared public randomness is needed is unknown and has been recently considered in [HS12].

3 Preliminaries

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 3.1. 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.

3.1 Information Theory

Definition 3.2 (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 3.3. $H(AB) = H(A) + H(B|A)$.

Definition 3.4 (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 3.5 (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 3.6 (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 3.7. *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 3.8.

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

3.2 Communication Complexity

Let $\mathcal{X}, \mathcal{Y}$ denote the set of possible inputs to the two players, who we name A and B. In this paper[2], we view a *private coins protocol* for computing a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbf{Z}_K$ 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 2.

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.

[2]The definitions we present here are equivalent to the classical definitions and are more convenient for our proofs.

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 2: A communication protocol.

Definition 3.9. The *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 3.10. 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.

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. We shall use the notation $\pi(x, y)_j$ 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 3.11 (Communication Complexity notation)**.** For a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbf{Z}_K$, 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.

For ease of notations, we shall sometimes use the shorthand $R(f)$ to denote $R_{1/3}(f)$.
We shall use the following theorem due to Yao:

Theorem 3.12 (Yao's Min-Max)**.** $R_{\rho}(f) = \max_{\mu} D^{\mu}_{\rho}(f)$.

Recall that the internal information cost $\mathsf{IC}^{\mathsf{i}}{}_{\mu}(\pi)$ of a protocol π is defined to be $I(\pi(X, Y); X|Y) + I(\pi(X, Y); Y|X)$.

Lemma 3.13. *Let R be the public randomness used in the protocol π. Then $\mathsf{IC}^{\mathsf{i}}{}_{\mu}(\pi) = \mathbf{E}_R \left[\mathsf{IC}^{\mathsf{i}}{}_{\mu}(\pi_R)\right]$*

Proof. By the chain rule (Proposition 3.5),

$$\begin{aligned}
\mathsf{IC}^{\mathsf{i}}{}_{\mu}(\pi) &= I(\pi(X, Y); X|Y) + I(\pi(X, Y); Y|X) \\
&= I(R; X|Y) + I(R; Y|X) + I(\pi(X, Y); X|YR) + I(\pi(X, Y); Y|XR) \\
&= I(\pi(X, Y); X|YR) + I(\pi(X, Y); Y|XR) \\
&= \mathop{\mathbf{E}}_{R} \left[\mathsf{IC}^{\mathsf{i}}{}_{\mu}(\pi_R)\right]
\end{aligned}$$

□

A priori, one might believe that the internal information cost can be as large as twice the communication in a protocol. However, we can use the fact that each transmission only reveals information to one of the parties to bound it by the communication in the protocol:

Lemma 3.14. $\mathsf{IC}^{\mathsf{i}}_{\mu}(\pi) \leq \mathsf{CC}(\pi)$.

Proof. First, let us assume that π is a private coin protocol. Let π_i denote the i'th bit transmitted in the protocol. Then, by the chain rule,

$$\begin{aligned}\mathsf{IC}^{\mathsf{i}}_{\mu}(\pi) &= I(\pi(X,Y);X|Y) + I(\pi(X,Y);Y|X) \\ &= \sum_{i=1}^{\mathsf{CC}(\pi)} I(\pi_i;X|\pi_1\pi_2\ldots\pi_{i-1}Y) + I(\pi_i;Y|\pi_1\pi_2\ldots\pi_{i-1}X)\end{aligned}$$

Given any prefix $\gamma = \pi_1 \ldots \pi_{i-1}$, let E_γ denote the event that the first $i-1$ bits of the transcript are equal to γ. Then we have

$$\mathsf{IC}^{\mathsf{i}}_{\mu}(\pi) = \sum_{i=1}^{\mathsf{CC}(\pi)} \mathop{\mathbf{E}}_{\gamma \in_{\mathrm{R}} \pi_1\ldots\pi_{i-1}} \left[I(\pi_i;X|E_\gamma Y) + I(\pi_i;Y|E_\gamma X) \right].$$

Now we claim that $I(\pi_i;X|E_\gamma Y) + I(\pi_i;Y|E_\gamma X) \leq 1$. Each of these terms is individually bounded by 1 since π_i contains only one bit. If γ is such that it is the first party's turn to transmit π_i, then for every fixing of X, π_i is independent of Y, so $I(\pi_i;Y|E_\gamma X) = 0$. On the other hand, if γ is such that it is the second party's turn to transmit π_i, then for every fixing of Y, π_i is independent of X, so $I(\pi_i;X|E_\gamma Y) = 0$. Thus,

$$\mathsf{IC}^{\mathsf{i}}_{\mu}(\pi) = \sum_{i=1}^{\mathsf{CC}(\pi)} \mathop{\mathbf{E}}_{\tau \in_{\mathrm{R}} \pi_1\ldots\pi_{i-1}} \left[I(\pi_i;X|E_\tau Y) + I(\pi_i;Y|E_\tau X) \right] \leq \mathsf{CC}(\pi).$$

If π involves public randomness, then by Lemma 3.13, we have that $\mathsf{IC}^{\mathsf{i}}_{\mu}(\pi) = \mathbf{E}_R \left[\mathsf{IC}^{\mathsf{i}}_{\mu}(\pi_R) \right] \leq \mathsf{CC}(\pi)$, where R denotes the public randomness of π.

□

A version of the following theorem was proved in [BYJKS04]. Here we need a slightly stronger version (alluded to in a remark in [BBCR10]):

Theorem 3.15. *For every μ, f, ρ there exists a protocol τ computing f on inputs drawn from μ with probability of error at most ρ and communication at most $\mathsf{D}^{\mu^n}_{\rho}(f^n)$ such that $\mathsf{IC}^{\mathsf{i}}_{\mu}(\tau) \leq \frac{D^{\mu^n}_{\rho}(f^n)}{n}$.*

Since this theorem is subsumed by Theorem 3.17 below, we do not give the details of its proof.

For our results on amortized communication complexity, we need the following definition: we shall consider the problem of computing n copies of f, with error ρ in each coordinate of the

computation, i.e. the computation must produce the correct result in any single coordinate with probability at least $1-\rho$. We denote the communication complexity of this problem by $D_\rho^{\mu,n}(f) \leq \mathsf{D}_\rho^{\mu^n}(f^n)$. Formally,

Definition 3.16. Let μ be a distribution on $X \times Y$ and let $0 < \rho < 1$. We denote by $D_\rho^{\mu,n}(f)$ the distributional complexity of computing f on each of n independent pairs of inputs drawn from μ, with probability of failure at most μ on each of the inputs.

The result above can actually be strengthened:

Theorem 3.17. *For every μ, f, ρ, let π be a protocol realizing $D_\rho^{\mu,n}(f)$. Then there exists a protocol τ computing f on inputs drawn from μ with probability of error at most ρ such that $\mathsf{CC}(\tau) = \mathsf{CC}(\pi)$ and $\mathsf{IC}^{\mathsf{i}}_\mu(\tau) \leq \frac{\mathsf{IC}^{\mathsf{i}}_{\mu^n}(\pi)}{n} \leq \frac{D_\rho^{\mu,n}(f)}{n}$.*

Proof. First let us assume that π only uses private randomness. The protocol $\tau(x,y)$ is defined as follows.

1. The parties publicly sample J, a uniformly random element of the set $\{1,2,3,\ldots,n\}$.
2. The parties publicly sample $X_1,\ldots,X_{J-1}$ and $Y_{J+1},\ldots,Y_n$.
3. The first party privately samples $X_{J+1},\ldots,X_n$ conditioned on the corresponding Y's. Similarly, the second party privately samples $Y_1,\ldots,Y_{J-1}$.
4. The parties set $X_J = x$, $Y_J = y$, and run the protocol π on inputs $X_1,\ldots,X_n,Y_1,\ldots,Y_n$. They output the result computed for the J'th coordinate.

Observe that $\mathsf{CC}(\tau) = \mathsf{CC}(\pi)$, and the probability of making an error in τ is bounded by ρ. It only remains to bound $\mathsf{IC}^{\mathsf{i}}_\mu(\tau) = I(X;\tau|Y) + I(Y;\tau|X)$. Let us bound the first term.

$$\begin{aligned} I(X;\tau|Y) &\leq I(X;\tau Y_1,\ldots,Y_n|Y) \\ &= I(X;JX_1\ldots X_{J-1}Y_1\ldots Y_n\pi|Y) \\ &= I(X;JX_1\ldots X_{J-1}Y_1\ldots Y_n|Y) + I(X_J;\pi|JX_1\ldots X_{J-1}Y_1\ldots Y_n) \\ &= I(X_J;\pi|JX_1\ldots X_{J-1}Y_1\ldots Y_n) \end{aligned}$$

where the final equality is from the fact that $J, X_1,\ldots,X_{J-1},Y_1,\ldots,Y_n$ are all independent of X,Y, conditioned on every fixing of Y.

Expanding the expectation according to J, we get by the Chain Rule:

$$\begin{aligned} I(X;\tau|Y) &\leq (1/n)\sum_{j=1}^{n} I(X_j;\pi|X_1\ldots X_{j-1}Y_1\ldots Y_n) \\ &= I(X_1\ldots X_n;\pi|Y_1\ldots Y_n)/n \end{aligned}$$

Similarly, we can bound $I(Y;\tau|X) \leq I(Y_1\ldots Y_n;\pi|X_1\ldots X_n)/n$, and thus $\mathsf{IC}^{\mathsf{i}}_\mu(\tau) \leq \mathsf{IC}^{\mathsf{i}}_{\mu^n}(\pi)/n \leq \mathsf{CC}(\pi)/n$, by Lemma 3.14.

If π uses public randomness R, then denote by τ_R the protocol induced for each fixing of R. Then $\mathsf{IC}^{\mathsf{i}}_\mu(\tau) = \mathbf{E}_R\left[\mathsf{IC}^{\mathsf{i}}_\mu(\tau_R)\right] \leq \mathbf{E}_R\left[\mathsf{IC}^{\mathsf{i}}_{\mu^n}(\pi_R)/n\right] \leq \mathsf{CC}(\pi)/n$.

□

4 Proof of Theorem 2.1

We shall prove a stronger version of Theorem 2.1.

Theorem 4.1. *Suppose that player A is given a distribution P and player B is given a distribution Q over a universe $\mathcal{U}$. There is a protocol such that at the end of the protocol:*

- *player A outputs an element a distributed according to P;*
- *player B outputs an element b such that for each x, $\mathbf{P}[b = a|\ a = x] > 1 - \varepsilon$.*
- *the communication in the protocol is bounded by $\log P(a)/Q(a) + \log 1/\varepsilon + \log\log 1/\varepsilon + 5\sqrt{\log P(a)/Q(a)} + 9$.*

Note that the second condition implies in particular that player B outputs an element b such that $b = a$ with probability $> 1 - \varepsilon$. The protocol requires no prior knowledge or assumptions on $\mathbf{D}\,(P||Q)$.

Proof. The protocol runs as follows. Both parties interpret the shared random tape as a sequence of uniformly selected elements $\{a_i\}_{i=1}^{\infty} = \{(x_i, p_i)\}_{i=1}^{\infty}$ from the set $\mathcal{A} := \mathcal{U} \times [0, 1]$. Denote the subset

$$\mathcal{P} := \{(x, p)\ :\ P(x) > p\}$$

of $\mathcal{A}$ as the set of points under the histogram of the distribution P. Similarly, define

$$\mathcal{Q} := \{(x, p)\ :\ Q(x) > p\}.$$

For a constant $C \geq 1$ we will define the C-multiple of $\mathcal{Q}$ as

$$C \cdot \mathcal{Q} := \{(x, p) \in \mathcal{A}\ :\ (x, p/C) \in \mathcal{Q}\}.$$

We will also use a different part of the shared random tape to obtain a sequence of random hash functions $h_i : \mathcal{U} \to \{0, 1\}$ so that for any $x \neq y \in \mathcal{U}$, $\mathbf{P}[h_i(x) = h_i(y)] = 1/2$.

We are now ready to present the protocol:

1. Player A selects the first index i such that $a_i = (x_i, p_i) \in \mathcal{P}$, and outputs x_i;
2. Player A uses $1 + \lceil \log\log 1/\varepsilon \rceil$ bits to send Player B the binary encoding of $k := \lceil i/|\mathcal{U}| \rceil$ (if k is too large, Player A sends an arbitrary string);
3. For all t, set parameters $C_t := 2^{t^2}$, $s_t = 1 + \lceil \log 1/\varepsilon \rceil + (t + 1)^2$;
4. Repeat, until Player B produces an output, beginning with iteration $t = 0$:
   (a) Player A sends the values of all hash functions $h_j(x_i)$ for $1 \leq j \leq s_t$, that have not previously been sent.
   (b) if there is an $a_r = (y_r, q_r)$ with $r \in \{(k - 1) \cdot |\mathcal{U}| + 1, \ldots, k \cdot |\mathcal{U}|\}$ in $C_t \cdot \mathcal{Q}$ such that $h_j(y_r) = h_j(x_i)$ for $1 \leq j \leq s_t$, Player B responds 'success' and outputs y_r; if there is more than one such a_r, player B selects the first one;
   (c) otherwise, Player B responds 'failure', and the parties increment t and repeat.

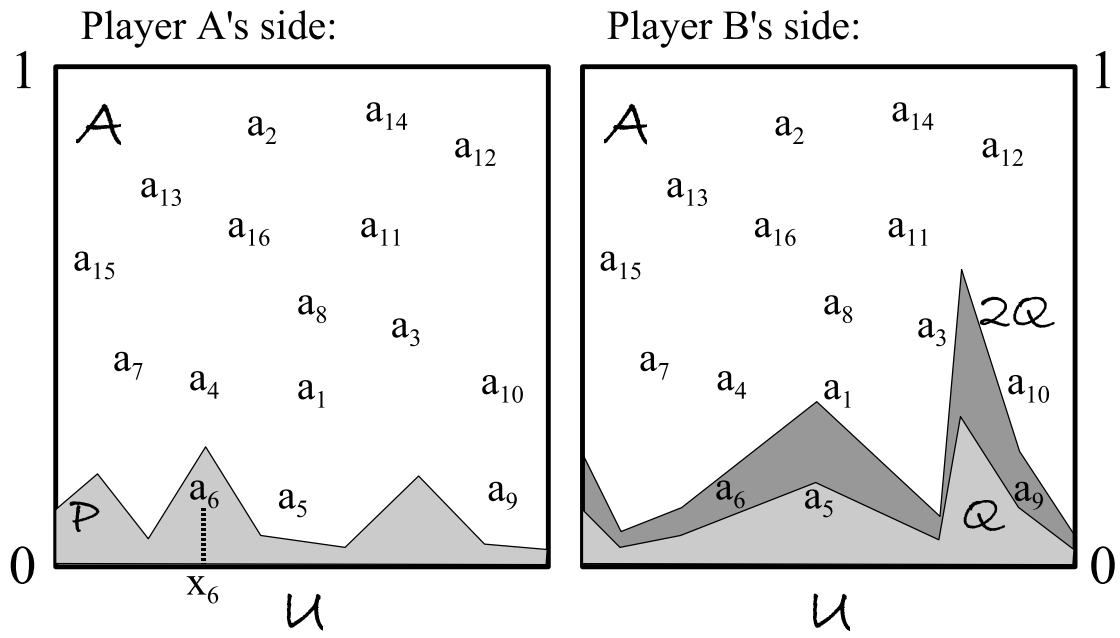

Figure 3: An illustration on the execution of the protocol. The elements a_i are selected uniformly from $\mathcal{A} = \mathcal{U} \times [0,1]$. The first a_i to fall in $\mathcal{P}$ is a_6, and thus player A outputs x_6. Player A sends hashes of a_6, which do not match the hashes of a_5, the only a_i in $\mathcal{Q}$. Player B responds 'failure', and considers surviving elements in $2\mathcal{Q}$, which are a_6 and a_9. After a few more hashes from A, a_6 is selected by B with high probability.

The output of Player A is distributed according to the distribution P, and further, the output is independent of k. To see this, note that the output is independent of whether or not $k > s$, for every s.

Fix a choice of i and the pair (x_i, p_i) by Player A. Step 4 of the protocol is guaranteed to terminate when $t^2 \geq \log P(x_i)/Q(x_i)$ since a_i belongs to $\frac{P(x_i)}{Q(x_i)} \cdot \mathcal{Q}$. Denote $T := \left\lceil \sqrt{\log P(x_i)/Q(x_i)} \right\rceil$. By iteration T, Player A will have sent s_T bits in Step 4, and Player B will have sent $T+1$ bits. Thus the amount of communication in Step 4 is bounded by

$$\begin{aligned} s_T + T + 1 &= 1 + \lceil \log 1/\varepsilon \rceil + (T+1)^2 + T + 1 \\ &\leq (\sqrt{\log P(x_i)/Q(x_i)} + 2)^2 + \sqrt{\log P(x_i)/Q(x_i)} + \lceil \log 1/\varepsilon \rceil + 3 \\ &= \log P(x_i)/Q(x_i) + 5\sqrt{\log P(x_i)/Q(x_i)} + \lceil \log 1/\varepsilon \rceil + 7, \end{aligned}$$

which shows that the total communication is at most

$$\log P(x_i)/Q(x_i) + 5\sqrt{\log P(x_i)/Q(x_i)} + \log 1/\varepsilon + \log\log 1/\varepsilon + 9.$$

It only remains to show that Player B outputs the same x_i with probability $> 1 - \varepsilon$. We start with the following claim.

Claim 4.2. *For each n, $\mathbf{P}[k > n] < e^{-n}$.*

Proof. For each n, we have

$$\mathbf{P}[k > n] = \mathbf{P}[a_i \notin \mathcal{P} \text{ for } i = 1, \ldots, n \cdot |\mathcal{U}|] = (1 - 1/|\mathcal{U}|)^{|\mathcal{U}| \cdot n} < e^{-n}.$$

□

Thus the probability that the binary encoding of k exceeds $1 + \lceil \log \log 1/\varepsilon \rceil$ bits is less than $e^{-2 \cdot 2^{\lceil \log \log 1/\varepsilon \rceil}} \leq \varepsilon/2$. It remains to analyze Step 4 of the protocol. We say that an element $a = (x, p)$ *survives* iteration t if $a \in 2^{t^2} \cdot \mathcal{Q}$ and it satisfies $h_j(x) = h_j(x_i)$ for all $j = 1, \ldots, s_t$ for this t.

Note that the "correct" element a_i survives iteration t if and only if $2^{t^2} \geq P(x_i)/Q(x_i)$.

Claim 4.3. *Let E_{a_i} be the event that the element selected by player A is a_i, which is the i-th element on the tape. Denote $k := \lceil i/|\mathcal{U}| \rceil$. Conditioned on E_{a_i}, the probability that a different element a_j with $j \in \{(k-1) \cdot |\mathcal{U}| + 1, \ldots, k \cdot |\mathcal{U}|\}$ survives iteration t is bounded by $\varepsilon/2^{t+1}$.*

Proof. Without loss of generality we can assume that $|\mathcal{U}| \geq 2$, since for a singleton universe our sampling protocol will succeed trivially. This implies that for a uniformly selected $a \in \mathcal{A}$, $\mathbf{P}[a \notin \mathcal{P}] \geq 1/2$, so:

$$\mathbf{P}[a \in C_t \cdot \mathcal{Q} |\ a \notin \mathcal{P}] \leq \mathbf{P}[a \in C_t \cdot \mathcal{Q}]/\mathbf{P}[a \notin \mathcal{P}] \leq 2 \cdot \mathbf{P}[a \in C_t \cdot \mathcal{Q}] \leq 2C_t/|\mathcal{U}|.$$

Denote $K := k \cdot |\mathcal{U}|$. Conditioning on E_{a_i}, the elements $a_{K-|\mathcal{U}|+1}, \ldots, a_{i-1}$ are distributed uniformly on $\mathcal{A} \setminus \mathcal{P}$, and $a_{i+1}, \ldots, a_K$ are distributed uniformly on $\mathcal{A}$. For any such $j = K - |\mathcal{U}| + 1, \ldots, i-1$, and for any $C > 0$,

$$\mathbf{P}[a_j \in C \cdot \mathcal{Q}] \leq 2C/|\mathcal{U}|.$$

For such a j, surviving round t means a_j belonging to $2^{t^2} \cdot \mathcal{Q}$ and agreeing with a_i on $s_t = 1 + \lceil \log 1/\varepsilon \rceil + (t+1)^2$ random hashes $h_1, \ldots, h_{s_t}$. The probability of this event is thus bounded by

$$\begin{aligned}
\mathbf{P}[a_j \text{ survives round } t] &\leq \mathbf{P}[a_j \in 2^{t^2} \cdot \mathcal{Q}] \cdot 2^{-s_t} \\
&\leq 2 \cdot 2^{t^2} \cdot 2^{-s_t}/|\mathcal{U}| \\
&\leq 2^{1+t^2-s_t}/|\mathcal{U}| \\
&\leq 2^{-2t-1}\varepsilon/|\mathcal{U}|.
\end{aligned}$$

By taking a union bound over all $j = K - |\mathcal{U}| + 1, \ldots, K$, $j \neq i$, we obtain the bound of $\varepsilon/2^{2t+1} \leq \varepsilon/2^{t+1}$. □

Thus for any E_{a_i}, the probability of Player B to output anything other than x_i conditioned on E_{a_i} is $< \sum_{t=0}^{\infty} \varepsilon/2^{t+1} = \varepsilon$.

□

To get a bound on the expected amount of communication in the protocol, as in Theorem 2.1, note that

$$\begin{aligned}
\mathbf{E}_{x_i \sim P}\left[\log P(x_i)/Q(x_i) + 2 + 5\sqrt{\log P(x_i)/Q(x_i)}\right] &= \mathbf{D}\left(P||Q\right) + 9 + 5 \cdot \mathbf{E}_{x_i \sim P}\sqrt{\log P(x_i)/Q(x_i)} \\
&\leq \mathbf{D}\left(P||Q\right) + 2 + 5 \cdot \sqrt{\mathbf{E}_{x_i \sim P} \log P(x_i)/Q(x_i)} = \mathbf{D}\left(P||Q\right) + O(\mathbf{D}\left(P||Q\right)^{1/2} + 1),
\end{aligned}$$

where the inequality is by the concavity of $\sqrt{\ }$. This completes the proof.

Remark 4.4. Note that if the parties are trying to sample many independent samples from distributions $P_1, P_2, \ldots$, with the receiving party knowing $Q_1, \ldots$, as in the setting of the Slepian-Wolf theorem, the analysis of the above protocol can easily be strengthened to show that the communication is the right amount with high probability. This is because the central limit theorem can be used to show that samples x_i with higher than expected $P_i(x_i)/Q_i(x_i)$ are rare and do not contribute much to the communication on average (see for example Section 6.2).

Remark 4.5. The sampling in the proof of Theorem 4.1 may take significantly more than one round. In fact, the expected number of rounds is $\Theta(\sqrt{\mathbf{D}\left(P||Q\right)})$. We suspect that the dependence of the number of rounds in the simulation on the divergence cannot be eliminated, since $\mathbf{D}\left(P||Q\right)$ is not known to the players ahead of time, and the only way to "discover" it (and thus to estimate the amount of communication necessary to perform the sampling task) is through interactive communication. By increasing the expected communication by a constant multiplicative factor, it is possible to decrease the expected number of rounds to $O(\log \mathbf{D}\left(P||Q\right))$.

5 Correlated Pointer Jumping

Here we define the correlated pointer jumping problem, that is at the heart of several of our results. The input in this problem is a rooted tree such that

- Each non-leaf node is *owned* by Player A or by Player 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.
- Each node v is associated with two distributions on its children: $\mathsf{child}_v^{\mathsf{A}}$ known to Player A and $\mathsf{child}_v^{\mathsf{B}}$ known to Player B.
- The leaves of the tree are labeled by output values.

The number of rounds in the instance is the depth of the tree.

The goal of the problem is for the players to sample the leaf according to the distribution that is obtained by sampling each child according to the distribution specified by the owner of the parent. We call the distribution of this path, the *correct* distribution. We give a way to measure the correlation between the knowledge of the two parties in the problem.

For every non-root vertex w in the tree whose parent is v, define the *divergence cost* of w as

$$\mathbf{D}\left(w\right) = \begin{cases} \log\left(\frac{\mathsf{child}_v^{\mathsf{A}}(w)}{\mathsf{child}_v^{\mathsf{B}}(w)}\right) & \text{if v is owned by Player A} \\ \\ \log\left(\frac{\mathsf{child}_v^{\mathsf{B}}(w)}{\mathsf{child}_v^{\mathsf{A}}(w)}\right) & \text{if v is owned by Player B} \end{cases}$$

The divergence cost of the root is set to 0.

Given a path T that goes from the root to a leaf in the tree, the divergence cost of the path, denoted $\mathbf{D}\left(T\right)$ is the sum of the divergence costs of the nodes encountered on this path. Finally, the divergence cost of the instance F, denoted $\mathbf{D}\left(F\right)$ is the expected sum of divergence costs of the vertices encountered in the correct distribution on paths.

We can use our sampling lemma to solve the correlated pointer jumping problem, with communication bounded by the divergence cost:

Theorem 5.1. *There is a protocol that when given a k-round correlated pointer jumping instance F, can sample a path T such that there is an event E, with $\mathbf{P}[E] > 1 - k\varepsilon$, and conditioned on E,*

- *the parties both output the same sampled path T that has the correct distribution*
- *the communication in the protocol is bounded by $\mathbf{D}(T) + 2k\log(1/\varepsilon) + 5\sqrt{k\mathbf{D}(T)} + 9k$.*

Proof. The protocol for sampling the path is obtained simply by repeatedly running the protocol from Theorem 4.1. In each step, the parties sample the correct child. For each round i let E_i denote the event that the parties are consistent after round i. When E_i occurs, the sampled vertex has the correct distribution, and $\Pr[E_i] > 1-\varepsilon$. Define E to be the intersection of the events E_i. Then $\Pr[E] > 1 - k\varepsilon$. Conditioned on E, the sampled path has the correct distribution. Moreover, the if the sampled path is $T = v_0, v_1, \ldots, v_k$, then by Theorem 4.1, the communication in the protocol is at most

$$\begin{aligned}
&\sum_{i=1}^{k}\left(\mathbf{D}(v_i) + \log(1/\varepsilon) + \log\log(1/\varepsilon) + 5\sqrt{\mathbf{D}(v_i)} + 9\right)\\
&\leq \sum_{i=1}^{k}\left(\mathbf{D}(v_i)\right) + 2k\log(1/\varepsilon) + 5\sqrt{k\cdot\sum_{i=1}^{k}\mathbf{D}(v_i)} + 9k\\
&= \mathbf{D}(T) + 2k\log(1/\varepsilon) + 5\sqrt{k\cdot\mathbf{D}(T)} + 9k,
\end{aligned}$$

where the inequality is by the Cauchy-Schwartz inequality.

□

A key fact is that both the internal and external information cost of a protocol can be used to bound the expected divergence cost of an associated distribution on correlated pointer jumping instances. Since, in this work, we only require the connection to internal information cost, we shall restrict our attention to it.

Given a public coin protocol with inputs X, Y and public randomness R, for every fixing of x, y, r, we obtain an instance of correlated pointer jumping. The tree is the same as the protocol tree with public randomness r. If a node v at depth d is owned by Player A, let M be the random variable denoting the child of v that is picked. Then define $\mathsf{child}_v^{\mathsf{A}x}$ so that it has the same distribution as $M|\ X = x, \pi(X,Y)_{\leq d} = rv$, and $\mathsf{child}_v^{\mathsf{B}y}$ so it has the same distribution as $M|\ Y = y, \pi(X,Y)_{\leq d} = rv$. We denote this instance of correlated sampling by $F_\pi(x,y,r)$. Let μ denote the distribution on X, Y. Next we relate the average divergence cost of this instance to the internal information cost of π:

Lemma 5.2. $\mathbf{E}_{X,Y,R}\left[\mathbf{D}\left(F_\pi(x,y,r)\right)\right] = \mathsf{IC}^{\mathsf{i}}{}_\mu(\pi)$

Proof. We shall prove that for every r, $\mathbf{E}_{X,Y}\left[\mathbf{D}\left(F_\pi(x,y,r)\right)\right] = \mathsf{IC}^{\mathsf{i}}{}_\mu(\pi_r)$. The proof can then be completed by Lemma 3.13.

So without loss of generality, assume that π is a private coin protocol, and write $F(x,y)$ to denote the corresponding divergence cost. We proceed by induction on the depth of the protocol

tree of π. If the depth is 0, then both quantities are 0. For the inductive step, without loss of generality, assume that Player A owns the root node v of the protocol. Let M denote the child of the root that is sampled during the protocol, and let $F(x,y)_m$ denote the divergence cost of the subtree rooted at m. Then

$$\mathop{\mathbf{E}}_{X,Y}\left[\mathbf{D}\left(F(x,y)\right)\right] = \mathop{\mathbf{E}}_{x,y,m\in_{\mathrm{R}}X,Y,M}\left[\log(\mathsf{child}_{\mathsf{v}}^{\mathsf{A}\,x}(m)/\mathsf{child}_{\mathsf{v}}^{\mathsf{B}\,y}(m)) + \mathbf{D}\left(F(x,y)_m\right)\right] \tag{1}$$

Since for every x, y, $M|xy$ has the same distribution as $M|x$, Proposition 3.7 gives that the first term in Equation 1 is exactly equal to $I(X;M|Y) = I(X;M|Y) + I(Y;M|X)$. The second term is $\mathbf{E}_M\left[\mathbf{E}_{X,Y|M}\left[\mathbf{D}\left(F(X,Y)_M\right)\right]\right]$. For each fixing of $M = m$, the inductive hypothesis shows that the inner expectation is equal to $I(X;\pi(X,Y)|Ym) + I(Y;\pi(X,Y)|Xm)$. Together, these two bounds imply that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{X,Y}\left[\mathbf{D}\left(F(x,y)\right)\right] \\
&= I(X;M|Y) + I(Y;M|X) + I(X;\pi(X,Y)|YM) + I(Y;\pi(X,Y)|XM) \\
&= \mathsf{IC}^{\mathsf{i}}{}_{\mu}(\pi)
\end{aligned}$$

□

6 Applications

In this section, we use Theorem 5.1 to prove a few results about compression and direct sums.

6.1 Compression and Direct sum for bounded-round protocols

Here we prove our result about compressing bounded round protocols (Corollary 2.2).

Proof of Corollary 2.2. The proof follows by applying our sampling procedure to the correlated pointer jumping instance $F_\pi(x,y,r)$. For each fixing of x, y, r, define the event $G_{x,y,r}$ to be the event E from Theorem 5.1. Then we have that $\mathbf{P}[G] > 1 - k\varepsilon$. Conditioned on G, we sample from exactly the right distribution, and the expected communication of the protocol is

$$\begin{aligned}
&\mathop{\mathbf{E}}_{X,Y,R}\left[\mathbf{D}\left(F_\pi(X,Y,R)\right) + 2k\log(1/\varepsilon) + O(\sqrt{k\mathbf{D}\left(F_\pi(X,Y,R)\right)} + k)\right] \\
&\leq \mathop{\mathbf{E}}_{X,Y,R}\left[\mathbf{D}\left(F_\pi(X,Y,R)\right)\right] + 2k\log(1/\varepsilon) + O\left(\sqrt{\mathop{\mathbf{E}}_{X,Y,R}\left[k\mathbf{D}\left(F_\pi(X,Y,R)\right)\right]} + k\right),
\end{aligned}$$

where the inequality follows from the concavity of the square root function. By Lemma 5.2, this proves that the expected communication conditioned on G is $\mathsf{IC}^{\mathsf{i}}{}_{\mu}(\pi) + 2k\log(1/\varepsilon) + O\left(\sqrt{k\mathsf{IC}^{\mathsf{i}}{}_{\mu}(\pi)} + k\right)$.

□

6.2 Information = amortized communication

In this section we will show that Theorem 5.1 reveals a tight connection between the amount of information that has to be revealed by a protocol computing a function f and the amortized communication complexity of computing many copies of f. Recall that $\mathsf{IC}^{\mathsf{i}}_{\mu}(f,\rho)$ denotes the smallest possible internal information cost of any protocol computing f with probability of failure at most ρ when the inputs are drawn from the distribution μ. Observe that $\mathsf{IC}^{\mathsf{i}}_{\mu}(f,\rho)$ is an infimum over all possible protocols and may not be achievable by any individual protocol. It is also clear that $\mathsf{IC}^{\mathsf{i}}_{\mu}(f,\rho)$ may only increase as ρ decreases.

We first make the following simple observation.

Claim 6.1. *For each f, ρ and μ,*

$$\lim_{\alpha\to\rho} \mathsf{IC}^{\mathsf{i}}_{\mu}(f,\alpha) = \mathsf{IC}^{\mathsf{i}}_{\mu}(f,\rho)$$

Proof. The idea is that if we have any protocol with internal information cost I, error δ and input length ℓ, for every ε we can decrease the error to $(1-\varepsilon)\delta$ at the cost of increasing the information by at most $\varepsilon\cdot\ell$ just by using public randomness to run the original protocol with probability $1-\varepsilon$, and with probability ε, run the trivial protocol where the players simply exchange their inputs. Thus as α tends to ρ, the information cost of the best protocols must tend to each other. □

Next we define the amortized communication complexity of f. We define it to be the cost of computing n copies of f with error ρ in each coordinate, divided by n. Note that computing n copies of f with error ρ in each coordinate is in general an easier task than computing n copies of f with probability of success $1-\rho$ for all copies. We use the notation $D^{\mu,n}_{\rho}(f)$ to denote the communication complexity for this task, when the inputs for each coordinate are sampled according to μ. $D^{\mu,n}_{\rho}(f)$ was formally defined in Definition 3.16.

It is trivial to see in this case that $D^{\mu,n}_{\rho}(f) \le n\cdot D^{\mu}_{\rho}(f)$. The amortized communication complexity of f with respect to μ is the limit

$$\mathsf{AC}(f^{\mu}_{\rho}) := \lim_{n\to\infty} D^{\mu,n}_{\rho}(f)/n,$$

when the limit exists. We prove an exact equality between amortized communication complexity and the information cost:

Theorem 6.2.

$$\mathsf{AC}(f^{\mu}_{\rho}) = \mathsf{IC}^{\mathsf{i}}_{\mu}(f,\rho).$$

Proof. There are two directions in the proof:

$\underline{\mathsf{AC}(f^{\mu}_{\rho}) \ge \mathsf{IC}^{\mathsf{i}}_{\mu}(f,\rho).}$ This is a direct consequence of Theorem 3.17.

$\underline{\mathsf{AC}(f^{\mu}_{\rho}) \le \mathsf{IC}^{\mathsf{i}}_{\mu}(f,\rho).}$ Let $\delta > 0$. We will show that $D^{\mu,n}_{\rho}(f)/n < \mathsf{IC}^{\mathsf{i}}_{\mu}(f,\rho)+\delta$ for all sufficiently large n.

By Claim 6.1 there is an $\alpha < \rho$ such that $\mathsf{IC}^{\mathsf{i}}_{\mu}(f,\alpha) < \mathsf{IC}^{\mathsf{i}}_{\mu}(f,\rho)+\delta/4$. Thus there is a protocol π that computes f with error $< \alpha$ with respect to μ and that has an internal information cost bounded by $I := \mathsf{IC}^{\mathsf{i}}_{\mu}(f,\rho)+\delta/4$.

For every n, denote by π^n the protocol that takes n pairs of inputs X^n, Y^n and executes in parallel, sending the first bits of each copy in the first round, and then the second bits in the second

round and so on. Thus π^n has $\mathsf{CC}(\pi)$ rounds, and communication complexity $n\mathsf{CC}(\pi)$. Further, π^n computes n copies of f as per Definition 3.16 with error bounded by α.

We shall obtain our results by compressing π^n.

Let ${}^i\pi$ denote the transcript of the i'th copy, and let X_i, Y_i denote the i'th inputs. Then observe that for all i, $(X_i, Y_i, {}^i\pi)$ are mutually independent of each other. Indeed, this implies that $\mathsf{IC}^{\mathsf{i}}_{\mu^n}(\pi^n) = \sum_{i=1}^n \mathsf{IC}^{\mathsf{i}}_{\mu}({}^i\pi) = n\mathsf{IC}^{\mathsf{i}}_{\mu}(\pi)$. On the other hand, compressing π^n incurs a per round overhead that is still dependent only on $\mathsf{CC}(\pi)$.

Let T^n denote the random variable of the path sampled in π^n. Let $T_1, \ldots, T_n$ denote the random variables of the n paths sampled in the individual copies of π. Then, since each protocol runs independently, $\mathbf{E}[\mathbf{D}(T^n)] = \sum_{i=1}^n \mathbf{E}[\mathbf{D}(T)]$. Indeed, each vertex in the protocol tree of π^n corresponds to an n-tuple of vertices of π, and if w corresponds to the vertices $({}^1w, \ldots, {}^n w)$, with parents $v = ({}^1v, \ldots, {}^n v)$ owned by Player A, then

$$\mathbf{D}(w) = \log\left(\frac{\mathsf{child}^{\mathsf{A}}_{\mathsf{v}}(w)}{\mathsf{child}^{\mathsf{B}}_{\mathsf{v}}(w)}\right) = \log\left(\frac{\prod_{i=1}^n \mathsf{child}^{\mathsf{A}}_{\mathsf{v_i}}(w_i)}{\prod_{i=1}^n \mathsf{child}^{\mathsf{B}}_{\mathsf{v_i}}(w_i)}\right) = \sum_{i=1}^n \log\left(\frac{\mathsf{child}^{\mathsf{A}}_{\mathsf{v_i}}(w_i)}{\mathsf{child}^{\mathsf{B}}_{\mathsf{v_i}}(w_i)}\right) = \sum_{i=1}^n \mathbf{D}(w_i).$$

By Lemma 5.2, $\mathbf{E}[\mathbf{D}(T)] = \mathsf{IC}^{\mathsf{i}}_{\mu}(\pi)$. Thus, by the central limit theorem, for n large enough,

$$\Pr[\mathbf{D}(T^n) \geq n \cdot (\mathsf{IC}^{\mathsf{i}}_{\mu}(\pi) + \delta/4)] < (\rho - \alpha)/2.$$

We use Theorem 5.1 to simulate π^n, with error parameter $\varepsilon = (\rho - \alpha)/2$ and truncate the protocol after

$$n \cdot (\mathsf{IC}^{\mathsf{i}}_{\mu}(\pi) + \delta/4) + 5\sqrt{\mathsf{CC}(\pi) \cdot n \cdot (\mathsf{IC}^{\mathsf{i}}_{\mu}(\pi) + \delta/4)} + 2\log(1/\varepsilon) + 9 \cdot \mathsf{CC}(\pi)$$

bits of communication. The new protocol thus has error $< \alpha + \rho - \alpha = \rho$. On the other hand, for n large enough, the per copy communication of this protocol is at most $\mathsf{IC}^{\mathsf{i}}_{\mu}(\pi) + \delta/2$ as required. □

6.3 A complete problem for direct sum

Let f^n denote the function mapping n inputs to n outputs according to f. We will show that the promise version of the correlated pointer jumping problem is complete for direct sum. In other words, if near-optimal protocols for correlated pointer jumping exist, then direct sum holds for all promise problems. On the other hand, if there are no near-optimal protocols for correlated pointer jumping, then direct sum fails to hold, with the problem itself as the counterexample. Thus any proof of direct sum for randomized communication complexity must give (or at least demonstrate existence) of near-optimal protocols for the problem.

We define the $\mathsf{CPJ}(C, I)$ promise problem as follows.

Definition 6.3. The $\mathsf{CPJ}(C, I)$ is a promise problem, where the players are provided with a *binary* instance[3] F of a C-round pointer jumping problem, i.e. player A is provided with the distributions $\mathsf{child}(v)_x$ and player B is provided with the distributions $\mathsf{child}(v)_y$ for each v, with the following additional guarantees:

- the divergence cost $\mathbf{D}(F) \leq I$;

[3]Each vertex has degree 2.

- let μ_F be the correct distribution on the leafs of F; each leaf z of F are labeled with $\ell(z) \in \{0,1\}$ so that there is a value $g = g(F)$ such that $\mathbf{P}_{z \in_R \mu_F}[\ell(z) = g(F)] > 1 - \varepsilon$, for some small ε. The goal of the players is to output $g(F)$ with probability $> 1 - 2\varepsilon$.

Note that players who know how to sample from F can easily solve the CPJ problem. It follows from [BBCR10] that:

Theorem 6.4. *If $\mathsf{CPJ}(C, I)$ has a randomized protocol that uses $T(C, I) := \mathsf{R}(\mathsf{CPJ}(C, I))$ communication, so that $T(C, C/n) < C/k(n)$, then for each f,*

$$\mathsf{R}(f^n) = \Omega(k(n) \cdot \mathsf{R}(f)).$$

In [BBCR10] a bound of $T(C, I) = \tilde{O}(\sqrt{C \cdot I})$ is shown, which implies $\mathsf{R}(f^n) = \tilde{\Omega}(\sqrt{n} \cdot \mathsf{R}(f))$ for any f. Using Theorem 5.1 we are able to prove the converse direction.

Theorem 6.5. *For any $C > I > 0$, set $n := \lfloor C/I \rfloor$, then*

$$\mathsf{R}(\mathsf{CPJ}(C, I)^n) = O(C \log(nC/\varepsilon)).$$

Thus, if there are parameters C and n such that $\mathsf{CPJ}(C, C/n)$ cannot be solved using $I = C/n$ communication, i.e. $T(C, C/n) > C/k(n) \gg C/n$, then $\mathsf{CPJ}(C, C/n)$ is a counterexample to direct sum, i.e.

$$\mathsf{R}(\mathsf{CPJ}(C, I)^n) = O(C \log nC/\varepsilon) = \tilde{O}(C) = \tilde{O}(k(n)\mathsf{R}(\mathsf{CPJ}(C, C/n))) = o(n \cdot \mathsf{R}(\mathsf{CPJ}(C, C/n))).$$

Proof. (of Theorem 6.5) We solve $\mathsf{CPJ}(C, I)^n$ by taking $m := n \log n$ copies of the $\mathsf{CPJ}(C, I)$ problem representing $\log n$ copies of each of the n instances. The players will compute all the copies in parallel with error $< 2\varepsilon$, and then take a majority of the $\log n$ copies for each instance. For a sufficiently large n this guarantees the correct answer for all n instances except with probability $< \varepsilon$. Thus our goal is to simulate m copies of $\mathsf{CPJ}(C, I)$. We view $\mathsf{CPJ}(C, I)^m$ as a degree-2^m, C-round correlated pointer jumping problem in the natural way. Each node represents a vector $V = (v_1, \ldots, v_m)$ of m nodes in the m copies of $\mathsf{CPJ}(C, I)$. The children of V are the 2^m possible combinations of children of $\{v_1, \ldots, v_m\}$. The distribution on the children is the product distribution induced by the distributions in $v_1, \ldots, v_m$. We claim that

$$\mathbf{D}\left(\mathsf{CPJ}(C, I)^n_{v_1,\ldots,v_m}\right) = \sum_{i=1}^{m} \mathbf{D}\left(\mathsf{CPJ}(C, I)_{v_i}\right). \tag{2}$$

This follows easily by induction on the tree, since the distribution on each node is a product distribution, and for each independent pairs $(P_1, Q_1), \ldots, (P_m, Q_m)$ we have

$$\mathbf{D}\left(P_1 \times P_2 \times \ldots \times P_m || Q_1 \times Q_2 \times \ldots \times Q_m\right) = \mathbf{D}\left(P_1 || Q_1\right) + \ldots + \mathbf{D}\left(P_m || Q_m\right),$$

by Lemma 3.8. By applying (2) to the root of the tree we see that $\mathbf{D}\left(\mathsf{CPJ}(C, I)^m\right) \leq m \cdot I \leq C \log n$. Thus Theorem 5.1 implies that $\mathsf{CPJ}(C, I)^n$ can be solved with an additional error of $\varepsilon/2$ using an expected

$$C \log n + C \log C/\varepsilon + o(C \log n)$$

bits of communication. □

7 Acknowledgments

We thank Boaz Barak and Xi Chen for useful discussions.

References

[BBCR10] B. Barak, M. Braverman, X. Chen, and A. Rao. How to compress interactive communication. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing*, 2010.

[BBK+12] J. Brody, H. Buhrman, M. Koucky, B. Loff, and F. Speelman. Towards a reverse Newman's theorem in interactive information complexity. *Electronic Colloquium on Computational Complexity (ECCC)*, (TR12-179), 2012.

[BGPW12] M. Braverman, A. Garg, D. Pankratov, and O. Weinstein. From information to exact communication. 2012.

[BR11] M. Braverman and A. Rao. Information equals amortized communication. In *FOCS*, pages 748–757, 2011.

[Bra12a] M. Braverman. Interactive information complexity. In *Proceedings of the 44th symposium on Theory of Computing*, STOC '12, pages 505–524, New York, NY, USA, 2012. ACM.

[Bra12b] M. Braverman. Coding for interactive computation: progress and challenges. In *50th Annual Allerton Conference on Communication, Control, and Computing.*, 2012. to appear, available at `http://www.cs.princeton.edu/~mbraverm/`.

[BYJKS04] Z. Bar-Yossef, T. S. Jayram, R. 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.

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

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

[Cuf08] P. Cuff. Communication requirements for generating correlated random variables. *CoRR*, abs/0805.0065, 2008. informal publication.

[FKNN91] T. Feder, E. Kushilevitz, M. Naor, and N. Nisan. Amortized communication complexity. *SIAM Journal on Computing*, 24(4):736–750, 1995. Prelim version by Feder, Kushilevitz, Naor FOCS 1991.

[HJMR07] P. Harsha, R. Jain, D. A. McAllester, and J. Radhakrishnan. The communication complexity of correlation. In *IEEE Conference on Computational Complexity*, pages 10–23. IEEE Computer Society, 2007.

[Hol07] T. Holenstein. Parallel repetition: Simplifications and the no-signaling case. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, 2007.

[HS12] E. Haramaty and M. Sudan. Deterministic compression with uncertain priors. *arXiv preprint arXiv:1211.5718*, 2012.

[JKKS11] B. Juba, A. Kalai, S. Khanna, and M. Sudan. Compression without a common prior: An information-theoretic justification for ambiguity in language. 2011.

[JPY12] R. Jain, A. Pereszlényi, and P. Yao. A direct product theorem for bounded-round public-coin randomized communication complexity. *CoRR*, abs/1201.1666, 2012.

[JRS03] R. Jain, J. Radhakrishnan, and P. Sen. A direct sum theorem in communication complexity via message compression. In J. C. M. Baeten, J. K. Lenstra, J. Parrow, and G. J. Woeginger, editors, *Automata, Languages and Programming, 30th International Colloquium, ICALP 2003, Eindhoven, The Netherlands, June 30 - July 4, 2003. Proceedings*, volume 2719 of *Lecture Notes in Computer Science*, pages 300–315. Springer, 2003.

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

[Orl90] A. Orlitsky. Worst-case interactive communication. i. two messages are almost optimal. *Information Theory, IEEE Transactions on*, 36(5):1111–1126, 1990.

[Rao08] A. Rao. Parallel repetition in projection games and a concentration bound. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, 2008.

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

[Sha01] R. Shaltiel. Towards proving strong direct product theorems. *Computational Complexity*, 12(1-2):1–22, 2003. Prelim version CCC 2001.

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

[Wyn75] A. D. Wyner. The common information of two dependent random variables. *IEEE Transactions on Information Theory*, 21(2), March 1975.